RMetS
Royal Meteorological Society

Evaluating raw ensembles with the continuous ranked probability score

Jochen Bröcker*
[a]*Max–Planck–Institut für Physik komplexer Systeme, Dresden, Germany*

*Correspondence to: J. Bröcker, Max–Planck–Institut, Nöthnitzer Strasse 34, 01187 Dresden, Germany.
E-mail: broecker@pks.mpg.de

The continuous ranked probability score (CRPS) is a frequently used scoring rule. In contrast with many other scoring rules, the CRPS evaluates cumulative distribution functions. An ensemble of forecasts can easily be converted into a piecewise constant cumulative distribution function with steps at the ensemble members. This renders the CRPS a convenient scoring rule for the evaluation of 'raw' ensembles, obviating the need for sophisticated ensemble model output statistics or dressing methods prior to evaluation. In this article, a relation between the CRPS score and the quantile score is established. The evaluation of 'raw' ensembles using the CRPS is discussed in this light. It is shown that latent in this evaluation is an interpretation of the ensemble as quantiles but with non-uniform levels. This needs to be taken into account if the ensemble is evaluated further, for example with rank histograms. Copyright © 2012 Royal Meteorological Society

Key Words: probability forecasts; ensemble forecasts; scoring rules; quantile score

Received 16 May 2011; Revised 2 November 2011; Accepted 7 December 2011; Published online in Wiley Online Library 2 February 2012

Citation: Bröcker J. 2012. Evaluating raw ensembles with the continuous ranked probability score. *Q. J. R. Meteorol. Soc.* **138**: 1611–1617. DOI:10.1002/qj.1891

1. Introduction

Providing forecasts in terms of probabilities has become increasingly popular in the meteorological community. By now, a wide range of probabilistic forecast products is available commercially for short- and medium-term weather forecasts. However, the pristine output of atmospheric circulation models are usually not probabilities, but rather *ensembles*. An ensemble is a collection of model trajectories, generated using slightly different initial conditions (and sometimes also perturbed model equations). The different initial conditions and perturbed model equations are supposed to represent the uncertainty about the current state of the atmosphere and the uncertainty about the relevant physics, respectively. Consequently, the individual ensemble members represent likely scenarios of the future atmospheric development, consistent with the currently available (necessarily incomplete) information.

The question arises how an ensemble can be interpreted and evaluated in probabilistic terms. There is a considerable body of literature concerned with the problem of transforming ensembles into probability density functions (or cumulative distribution functions, or similar objects). These techniques, often referred to as ensemble interpretation models, usually include means to compensate and correct for statistical errors, that is, for *de-biasing*. Ensemble interpretation models of considerable sophistication exist (e.g. Jewson, 2003; Raftery *et al.*, 2005; Bröcker and Smith, 2008, and further references therein). Once the ensembles have been converted to probabilities, they are amenable to evaluation with so-called probabilistic scoring rules. There is a wide range of useful scoring rules; section 2 provides further discussion and references.

Nonetheless, it is of interest to evaluate the raw ensembles, with minimum interference from ensemble interpretation models. The raw ensemble might, for example, serve as a benchmark for more sophisticated ensemble interpretation models, or limited computational resources might preclude using the latter. As a further example, forecasters might want to evaluate raw ensembles in order to compare ensemble generation systems (i.e. their atmospheric circulation models and data assimilation systems). Again, to get

meaningful results in such a comparison, the influence of the ensemble interpretation models must be kept to a minimum.

It turns out that a useful scoring rule to evaluate the raw ensemble is the Continuous Ranked Probability Score (CRPS), to be discussed in section 2. Essentially, the CRPS evaluates probability forecasts in the form of cumulative distribution functions; the raw ensemble can be converted to a piecewise constant cumulative distribution function with jumps at the ensemble members, and thus be evaluated with the CRPS. This will be discussed in detail in section 3. In that section, we will make an observation which is central to the discussion in this article, namely that evaluating the raw ensemble with the CRPS is equivalent to evaluating the ensemble members as if they were quantiles with certain levels. In other words, if we optimise the ensemble with respect to its CRPS performance, we can regard the ensemble members as quantile estimators.

This observation has important consequences. Essentially, it imposes a statistical interpretation upon the ensemble members which needs to be taken into account in subsequent statistical analyses of the ensemble. By statistical analysis, we mean checking whether the ensemble is consistent with some statistical hypothesis. This hypothesis might be at variance with the interpretation of the ensemble members as quantiles. In section 3, we will see that this can be the case with rank histograms, a popular test for reliability. This discussion is part of a wider theme (discussed in section 5) of statistical consistency between evaluation methods. This includes scoring rules as well as forecast distributions being picked from a specific model class. Statistical consistency would mean that the statistical hypotheses implied or assumed by the different evaluation methods do not mutually exclude each other.

2. Set-up and the Continuous Ranked Probability Score

We consider a variable Y, referred to as the *verification*, which is to be forecast. We treat Y as a random variable with values on the real line or a semi–infinite or finite interval. The distribution of Y can be specified by a cumulative distribution function F, that is,

$$F(y) = \mathbb{P}(Y < y)\,,$$

where $\mathbb{P}$ denotes the probability. It follows readily from the definitions that F is a monotonically increasing function. In general, F is left continuous and has right limits, that is

$$\lim_{\epsilon\to 0} F(y+\epsilon) \quad \text{exists}, \tag{1}$$

$$\text{and} \quad \lim_{\epsilon\to 0} F(y-\epsilon) = F(y). \tag{2}$$

These two limits are not necessarily equal, whence the function F has an upward jump. Since

$$\lim_{\epsilon\to 0} F(y+\epsilon) - F(y) = \mathbb{P}(Y = y), \tag{3}$$

we see that a jump occurs whenever Y assumes a certain value y with non-zero probability. A cumulative distribution function can have at most countably many jumps.

We assume that probability forecasts are issued for Y in the form of a cumulative distribution function G. A scoring rule is a means to evaluate the forecast G. In the present article, we shall focus on the CRPS, which is defined as

$$S(G,y) = \int \big\{G(x) - H(x-y)\big\}^2\,\mathrm{d}x\,, \tag{4}$$

with H the Heaviside function, which is 1 if the argument is positive and zero otherwise. Note that the CRPS is a function of the real number y as well as a function*al* of the cumulative distribution function G. Heuristically speaking, it measures the difference between the forecast G and a perfect forecast H which puts all mass on the verification y. Given the forecast G and a verification Y, we would assign the score $S(G,Y)$ to the forecast. Note also that a *small* score S indicates a good forecast.

The score is a random number, so we would interprete the average of this number as a measure of average performance of G. We denote the average score as

$$s(G,F) = \int S(G,y)\,\mathrm{d}F. \tag{5}$$

Here we have taken the average over $S(G,Y)$, writing F for the distribution of Y. By straightforward manipulation, we get the following representation:

$$s(G,F) = \int \{F(x) - G(x)\}^2 + F(x)\,\{1 - F(x)\}\;\mathrm{d}x\,.$$

We can define the *divergence*

$$d(G,F) = s(G,F) - s(F,F);$$

using the previous expression for s, we get

$$d(G,F) = \int \{F(x) - G(x)\}^2\,\mathrm{d}x\,, \tag{6}$$

which shows that the divergence d is necessarily positive unless $G = F$. The interpretation of this mathematical fact is that the cumulative distribution function F of y achieves, on average, a better score than any forecast G, unless $F = G$. Scoring rules with this property are referred to as *proper*. Proper scoring rules have been the subject of theoretical studies and are applied widely for evaluation of probabilistic forecasts. The reader is referred to Savage (1971) and more recently Gneiting and Raftery (2007) for mathematical discussion of scoring rules, the latter discussing also the CRPS. Original articles on specific scoring rules are Brier (1950) and Good (1952). The CRPS was apparently introduced by Epstein (1969). The importance of using proper scores to obtain consistent results has been illustrated in Brown (1970) and Bröcker and Smith (2007). The literature abounds in examples of scoring rules being applied to evaluate probability forecasts: Roulston and Smith (2002, 2003); Roulston *et al.* (2003); Gneiting *et al.* (2005); Raftery *et al.* (2005); Roulston *et al.* (2005); Grimit *et al.* (2006); Sloughter *et al.* (2007). This is by no means an exhaustive list.

3. Evaluating raw ensembles using the CRPS

Having introduced the CRPS in the last section, we now discuss how to evaluate raw ensembles using the CRPS. Let $(e_1,\ldots,e_K)$ be an ensemble of forecasts for Y. For all

$k = 1, \ldots, K$, the e_k are confined to the same range of values as Y (all real numbers, a semi–infinite interval, or a finite interval). The initial ordering of the ensemble members is considered insignificant, and we therefore assume that the e_k are in increasing order. Using the ensemble, we can form the following piecewise constant function

$$G_e(x) = \sum_{k=1}^{K} w_k H(x - e_k) \tag{7}$$

with *weights* w_k so that $w_k > 0$ for all k, and $\sum_k w_k = 1$. Clearly, G_e comprises a piecewise constant cumulative distribution function satisfying the regularity properties (1, 2). Furthermore, G_e features exactly K jumps (in the sense of Eq. (3)) at the points $x = e_k$ with jump height w_k. Therefore, due to relation (3), $\mathbb{P}(Y = e_k) = w_k$, while zero probability is assigned to any set that does not contain an ensemble member. Using the CRPS, we can evaluate G_e by $S(G_e, Y)$, which on average is equal to $s(G_e, F)$.

The question arises as to how such an evaluation method should be interpreted. More specifically, we have seen that the minimum of $s(G, F)$ over *all* cumulative distribution functions G obtains if $G = F$. Clearly though, the cumulative distribution functions G_e are of a very special type, and we cannot, in general, expect that for a given F there is an e so that $G_e = F$. There is no 'correct' ensemble; however, there might be an ensemble $\widehat{e}$ which minimises the score $s(G_e, F)$ for a given cumulative distribution function F (and given weights w_k). Concerning such an optimal ensemble $\widehat{e}$, one might reasonably ask the following two questions:

1. The ensemble $\widehat{e}$ minimising the score $s(G_e, F)$ will be a function of F. What exactly does that function look like?
2. Suppose that the ensemble $\widehat{e}$ minimises the score $s(G_e, F)$, where F is *some* distribution (e.g. our forecast probability). How can we test the hypothesis that the verification Y has distribution F, using only the ensemble $\widehat{e}$? For example, will the ensemble $\widehat{e}$ display a flat rank histogram?

We will discuss these two questions.

To answer the first question, we have to determine the minimum of $s(G_e, F)$ with respect to e. Clearly, this is equivalent to minimising $d(G_e, F)$, for which we have the expression (6). However, for illustrative purposes, we will take another route, thereby revealing an interesting connection to the well-known quantile score. We will first present an illuminating expression for the score $S(G_e, y)$ and then compute the integral (5).

Substituting with Eq. (7) in Eq. (4), we obtain, after some algebra which has been relegated to the Appendix, that we may write $S(G_e, y)$ as a weighted sum

$$S(G_e, y) = 2 \sum_j w_j \{\alpha_j (y - e_j)_+ + (1 - \alpha_j)(e_j - y)_+\}, \tag{8}$$

with $\alpha_j = \sum_{k \le j} w_k - (w_j/2)$. Under the sum in (8), the function $\sigma_\alpha(y, x) = \alpha(y - x)_+ + (1 - \alpha)(x - y)_+$ appears; this function is well known as the *quantile score* of level α (Gneiting and Raftery, 2007; Friederichs and Hense, 2008). Thus we see that applying the CRPS to the function G_e amounts to applying the quantile score with certain level α_j to each individual ensemble member e_j. As the name suggests, the quantile score is often used to score quantile estimates. This is motivated by the mathematical fact that the expectation value of $\sigma_\alpha(Y, x)$ is minimal (as a function of x) if $F(x) = \alpha$ holds, that is, if x is the α quantile of F. To see this, write

$$\begin{aligned}
&\int_{-\infty}^{\infty} \sigma_\alpha(y, x)\, \mathrm{d}F(y) \\
&= \int_{-\infty}^{\infty} \alpha(y - x)\, \mathrm{d}F(y) + \int_{-\infty}^{\infty} (x - y)_+\, \mathrm{d}F(y) \\
&= \alpha(\overline{Y} - x) + \int_{-\infty}^{x} (x - y)\, \mathrm{d}F(y) \\
&= \alpha(\overline{Y} - x) + \int_{-\infty}^{x} F(y)\, \mathrm{d}y\,. \qquad \text{(integration by parts)}
\end{aligned}$$

Here, $\overline{Y}$ denotes the expectation value of Y. Further, we have assumed (and will continue to do) that F has no jumps. Setting the derivative of this relation with respect to x to zero, we arrive at the necessary condition

$$\alpha = F(x), \tag{9}$$

that is, x has to be a quantile of F with level α. (If F has a jump that leap-frogs α, then Eq. (9) has no solution; in fact, the necessary condition reads a little differently in that situation.)

If we integrate Eq. (8) over $\mathrm{d}F(y)$ and set the derivatives with respect to the e_j equal to zero, the same reasoning applies, and we obtain the following necessary condition for the optimal ensemble $\widehat{e}$:

$$\alpha_j = F(\widehat{e}_j) \qquad \text{for all } j = 1, \ldots, K, \tag{10}$$

$$\text{with } \alpha_j = \sum_{k \le j} w_k - \frac{w_j}{2}. \tag{11}$$

This result can be described graphically as in Figure 1. On the ordinate, the interval $[0, 1]$ (the 'probability axis') is divided into K subintervals, with the kth interval having length w_k. Due to Eq. (11), α_k is the midpoint of the kth interval. The ensemble member $\widehat{e}_k$ obtains as the pre-image of α_k under F. Raw ensembles are usually evaluated with all weights equal to $1/K$, which gives $\alpha_j = (j - 0.5)/K$. Hence, we can conclude that the optimal ensemble member $\widehat{e}_j$ is a quantile of level $(j - 0.5)/K$. This concludes the discussion of the first question.

Turning to the second question, our hypothesis that Y has distribution F entails that F is reliable, which we

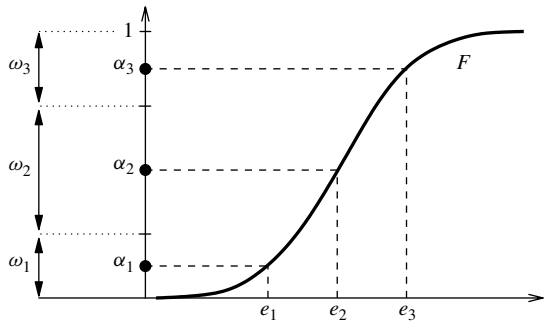

Figure 1. Illustration of the results in Eq. (11). The quantile levels α_k emerge as the midpoints of the intervals of width w_k. The corresponding ensemble members e_k are the pre-images of the α_k under the cumulative distribution function F.

assume hereafter. We are interested in how this bears on the ensemble $\widehat{e}$ computed from Eqs (10) and (11). Clearly, the probability mass between two consecutive ensemble members is

$$F(\widehat{e}_{k+1}) - F(\widehat{e}_k) = \alpha_{k+1} - \alpha_k = \frac{w_{k+1} + w_k}{2}, \qquad (12)$$

while the probability masses below $\widehat{e}_1$ and above $\widehat{e}_K$ are equal to $w_1/2$, and $w_K/2$, respectively. For the case of equal weights, that is all weights equal to $1/K$, the usual choice for evaluating raw ensembles with the CRPS score, we get from Eq. (12) that the probability mass between any two consecutive ensemble members is $1/K$, while the mass below $\widehat{e}_1$ and also above $\widehat{e}_K$ is just $1/2K$. This should be compared with the null hypothesis usually imposed for reliability tests, which states that the probability masses between two consecutive ensemble members as well as below $\widehat{e}_1$ and above $\widehat{e}_K$ are all equal (to $1/(K+1)$). This null hypothesis (referred to as $\mathcal{H}_R$ in the following) can be tested for, using rank histograms. The rank $r(Y)$ of Y is defined as the smallest j so that $Y < e_j$, with $r(Y) = K+1$ if Y exceeds all ensemble members. Under $\mathcal{H}_R$, the rank $r(Y)$ assumes the values $1, \ldots, (K+1)$ with equal probability. In other words, a histogram of a sample of $r(Y)$ should be flat, within statistical fluctuations. However, if the ensemble is assumed to be optimal with respect to the CRPS with equal weights, then the rank $r(Y)$ should assume a value $k \in \{2, \ldots, K\}$ with probability $1/K$, while the extreme values $k = 1$ and $k = K+1$ both have probability $1/2K$. We might thus view Eq. (12) as another hypothesis $\mathcal{H}_C$ about the distribution of ranks. It is a consequence of the fact that the ensemble is optimal with respect to the CRPS. Clearly, the hypothesis $\mathcal{H}_C$ in the case of equal weights does *not* yield equal rank probabilities. Numerical examples in section 4 will further illustrate this point. Hence, $\mathcal{H}_C$ and $\mathcal{H}_R$ are different hypotheses or, in other words, the ensemble we get by optimising the CRPS cannot be expected to produce a flat rank histogram.

Finally, we address the question of whether the two hypotheses $\mathcal{H}_C$ and $\mathcal{H}_R$ can be made to coincide by choosing the weights appropriately. Interestingly, this seems not to be the case, as might already be guessed from Figure 1. In the hypothesis $\mathcal{H}_R$, the α_j are all integer multiples of α_1, that is $\alpha_j = j \cdot \alpha_1$ and, in particular, $\alpha_j - \alpha_{j-1} = \alpha_1$. Comparison with Eq. (11) gives the inductive relation

$$w_j = 2\alpha_1 - w_{j-1}.$$

Furthermore, we know from Eq. (11) that $w_1 = 2\alpha_1$. This yields

$$w_j = \begin{cases} 2\alpha_1 & \text{if } j \text{ is odd,} \\ 0 & \text{if } j \text{ is even.} \end{cases}$$

However, this would mean that, in the actual CRPS (Eq. (8)), the even ensemble members are not evaluated at all, as their corresponding weight in the sum (8) is zero. Note also that if K is even, we would have the contradiction

$$1 = \sum_{j=1}^{K} w_j = \frac{K}{2} 2\alpha_1 = \frac{K}{K+1} \neq 1.$$

As a conclusion, we see that there is no appropriate choice of weights so that $\mathcal{H}_C$ becomes equal to $\mathcal{H}_R$.

4. Numerical experiments

The results of this article are illustrated with a small numerical experiment using artificial data, which are generated as follows. Samples y_n of the verification Y are defined through

$$y_n = u_n(1 + s_1\xi_n) + s_2\zeta_n\,, \qquad (13)$$

where ξ_n and ζ_n, interpreted as multiplicative and additive disturbances, are independent and standard normal random variables, while u_n is interpreted as the 'underlying signal', given as

$$u_n = \{A\sin(\pi\omega_1 n) + B\sin(\pi\omega_2 n)\}^2\,. \qquad (14)$$

The values of the parameters can be found in Table 1.

With these parameters, y_n roughly resembles temperature data scaled to unit standard deviation. A plot of these data (for t spanning two years) is shown in Figure 2. For the subsequent analysis, 3650 days (i.e. roughly ten years) of data were used.

Further, we generate an ensemble $e^{(1)}$ using the same model as for y_n but with independent disturbances. More specifically, if we denote by $e^{(1)}_{k,n}$ the kth member of $e^{(1)}$ at time n, we set

$$e^{(1)}_{k,n} = u_n(1 + s_1\xi_{k,n}) + s_2\zeta_{k,n},$$

where $\xi_{k,n}, \zeta_{k,n}$ are standard normal random variables, independent for different k, n. However all ensemble members are driven by the same underlying signal u_n. We work with $K = 10$ ensemble members.

By construction, this ensemble is consistent with the hypothesis $\mathcal{H}_R$, i.e. a histogram of the ranks of y_n should be flat, within statistical fluctuations. Such a histogram is shown in Figure 3(a). Visual inspection already indicates that this ensemble is consistent with $\mathcal{H}_R$. This is further confirmed by a χ^2 test, which gives a p value of 0.39. There is thus no indication for a deviation from $\mathcal{H}_R$, as expected.

Table 1. Parameter values used for generating the data (Eqs (13) and (14)).

s_1	s_2	A	B	ω_1	ω_2
0.3	0.3	1.68	0.0336	1/365.25	1/11

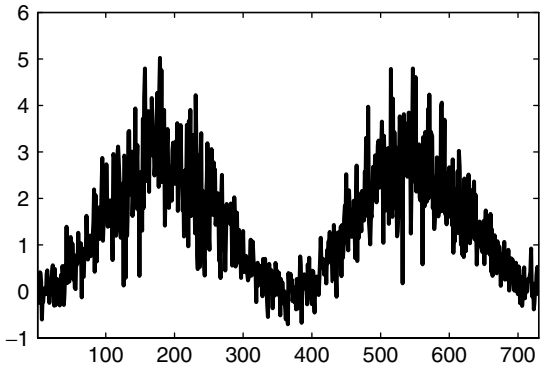

Figure 2. A plot of the artificial data y_n over time. 730 data points are shown. The data are supposed to resemble temperature data, normalised to unit standard deviation.

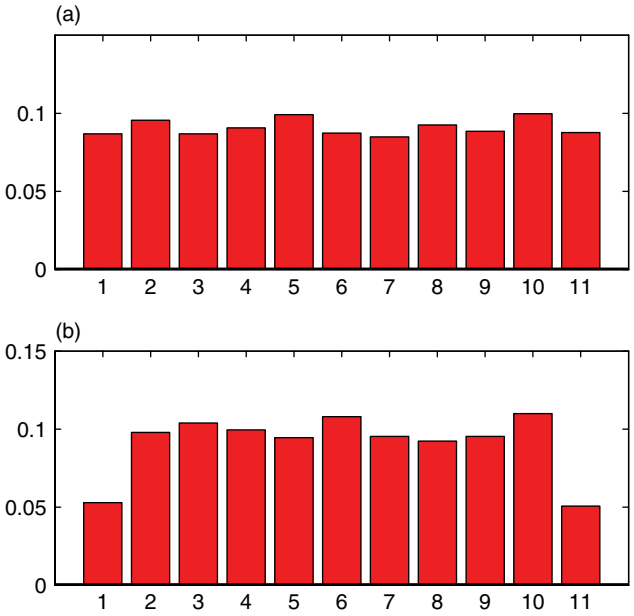

Figure 3. Rank histograms for (a) the ensemble $e^{(1)}$, and (b) the ensemble $e^{(3)}$. Under the null hypothesis of equal rank probability, the χ^2 test gives a p value of 0.39 for histogram (a). There is thus no indication of a deviation from reliability, as expected. Ensemble $e^{(3)}$ optimises the CRPS; the rank probabilities should therefore be $1/K$ for all bins except the extreme ones, which should be $1/(2K)$. Under this null hypothesis, the χ^2 test gives a p value of 0.29 for (b), consistent with our theory. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

We will now construct two further ensembles $e^{(2)}$ and $e^{(3)}$ by optimising the CRPS in a certain sense. It is clear from our definition of y_n in Eq. (13) that the distribution of y_n for given u_n has the form

$$F_n(y) = \Phi\left(\frac{y - u_n}{\sigma_n}\right) \quad \text{with} \quad \sigma_n = \sqrt{u_n^2 s_1^2 + s_2^2}, \qquad (15)$$

and Φ being the standard normal distribution function. Using F_n, we construct ensemble $e^{(2)}$ as in Eqs (10) and (11); more specifically,

$$e_{k,n}^{(2)} = F_n^{-1}(\alpha_k)$$
$$\text{with } \alpha_k = \frac{k}{K} - \frac{1}{2K}, \qquad k = 1, \ldots, K, \qquad K = 10.$$

As discussed in section 3, the ensemble $e^{(2)}$ does *not* display a flat rank histogram. Rather, since the individual ensemble members $e_{k,n}^{(2)}$ represent α_k quantiles with $\alpha_k = (k/K) - (1/2K)$, the rank histogram bars should display a height of roughly $1/K$ except for the extreme ranks, which should be of height $1/2K$; any deviations should arise merely through sampling fluctuations. (The rank histogram did not hold any surprises and is not shown.)

The ensemble $e^{(2)}$ describes a situation in which a forecaster has access to the conditional distribution function F but for some reason wants to cast her forecasts in the form of an ensemble that optimises the CRPS. In practice though, it is rarely the case that the forecaster has access to F explicitely in closed form. We will therefore mimic a more realistic situation in which the ensemble members have to be constructed (or 'debiased') from some information sources through statistical estimation procedures. To be specific, the ensemble members $e_{k,n}^{(3)}$ are constructed as linear functions of the ensemble mean of $e_n^{(1)}$. That is, with

$$\overline{e}_n^{(1)} = \frac{1}{K} \sum_k e_{k,n}^{(1)}$$

denoting the mean of ensemble $e_n^{(1)}$, we will let

$$e_{k,n}^{(3)} = c_{1,k} + c_{2,k} \cdot \overline{e}_n^{(1)}$$

with coefficients $c_{1,k}, c_{2,k}$ for $k = 1, \ldots, K$ to be determined by minimising the sample mean CRPS

$$\text{CRPS}_\text{E} = \sum_n S(G_{e_n^{(3)}}, y'_n),$$

with $S(G_e, y)$ given by Eq. (8). The time series y'_n (the training observations) are yet another realisation from the model (13), again with the same underlying signal u_n but with independent disturbances. Later, all evaluation will be done on the 'test' verifications y_n. As before, we use the standard weights $w_k = 1/K$.

From the discussion in section 3, we gather that the ensemble $e^{(3)}$ will at least approximately satisfy the relation (10) with $\alpha_k = (k/K) - (1/2K)$. That is, the individual ensemble members $e_{k,n}^{(3)}$ should at least approximately represent α_k quantiles of F_n (Eq. (15)) with $\alpha_k = (k/K) - (1/2K)$. Thus, the rank histogram bars should display a height of roughly $1/K$ except for the extreme ranks, which should be of height $1/2K$. (In particular, the rank histogram is *not* expected to be flat). The histogram is shown in Figure 3(b). The predicted behaviour is already apparent from visual inspection. This is further confirmed by a χ^2 test, which gives a p value of 0.29. Thus there is no indication that the observed histogram deviates from our theory. The important message is that the way in which the ensemble was constructed shows through in the reliability analysis. The ensemble members are approximations of certain quantiles with *non*-uniform levels, and hence the rank histogram is expected to be non-uniform as well.

As a small digression, since $e^{(3)}$ is only an approximation to $e^{(2)}$, the hypothesis $\mathcal{H}_C$ is only approximately true for $e^{(3)}$, and the reader might wonder whether we should come up with an *exact* hypothesis for $e^{(3)}$. Unfortunately, it seems that the exact shape of the rank histogram is not universal but depends on the particular distributions involved. However, in the present case, the difference turns out to be small enough so that the hypothesis $\mathcal{H}_C$ is appropriate. What is more important here is the fact that the standard hypothesis $\mathcal{H}_R$ is *not* appropriate; the conclusion that F is unreliable because the rank histogram of $e^{(3)}$ is not flat would be faulty.

Finally, we compare the three ensembles by means of the CRPS. Since $e^{(3)}$ has been trained using the data y'_n, we compute the scores with respect to y_n to allow for a fair comparison. The CRPS values are shown in Table 2.

As expected, $e^{(2)}$ shows the best performance. Further, we can conclude that, even on the test verifications, the CRPS

Table 2. CRPS scores for the three ensembles, with $\pm 2\sigma$ confidence ranges.

$e^{(1)}$	$e^{(2)}$	$e^{(3)}$
0.173 ± 0.005	0.157 ± 0.0044	0.158 ± 0.0044

of $e^{(3)}$ is significantly smaller (i.e. better) than that of $e^{(1)}$. In fact, the CRPS of $e^{(3)}$ is not significantly different from that of $e^{(2)}$. Alternatively, the mathematical expectation of the CRPS of $e^{(2)}$ can be computed analytically (up to numerical evaluation of integrals) and turns out to be 0.1582, which is not significantly different from the sample average CRPS of $e^{(2)}$ and $e^{(3)}$.

The results of this experiment might appear paradoxical at first sight, in that the ensembles $e^{(2)}$ and $e^{(3)}$ achieve a better CRPS score than $e^{(1)}$, despite the fact that the latter ensemble contains the same (or even more) information and, having a flat rank histogram, appears to be more reliable. This would then contradict the results of Hersbach (2000) which imply that, of two forecasts with the same information content, an unreliable forecast cannot have a better score than a reliable one. Of course, the solution of this paradox is that, in some sense, $e^{(2)}$ *is* reliable (and so is $e^{(3)}$, inasmuch as it is an approximation to $e^{(2)}$). In general, if we suppose that F is a reliable forecast distribution, and we construct an ensemble e from it, then the specific construction entails a hypothesis $\mathcal{H}$ about the rank histogram. We might then call the ensemble *reliable* if the actual rank histogram is consistent with $\mathcal{H}$. In the cases studied here, both $e^{(1)}$ and $e^{(2)}$ are based on a reliable forecast distribution F, but they are constructed in slightly different ways. For this reason, the corresponding rank histograms have to be checked against the slightly different hypotheses $\mathcal{H}_R$ and $\mathcal{H}_C$, respectively.

5. Concluding remarks

The continuous ranked probability score (CRPS) is a frequently used scoring rule that evaluates probability forecasts based on cumulative distribution functions. This property renders the CRPS a convenient score to evaluate raw ensemble forecasts (for real-valued verifications), since an ensemble can easily be converted into a piecewise constant cumulative distribution function with steps at the ensemble members. (The steps commonly have equal height.) This approach to ensemble evaluation obviates the need for sophisticated ensemble model output statistics or dressing methods. Thereby, forecasters can evaluate the raw ensemble without confounding potential shortcomings of the ensemble with those of the dressing method.

It has been shown here that, as a function of the raw ensemble members, the CRPS can be written as a sum of quantile scores applied to individual ensemble members. Hence, evaluation of the raw ensemble with the CRPS amounts to interpreting the ensemble members as quantiles. More specifically, the kth ensemble member is evaluated as a quantile with level $\alpha_k = (k - 0.5)/K$, where K is the total number of ensemble members, and equal step heights are used. This needs to be taken into account if the ensemble is evaluated further, for example with rank histograms. Common interpretations of ensembles amount to the null hypothesis of a flat rank histogram (allowing for statistical fluctuations). However, an ensemble optimal with respect to the CRPS is *not* expected to show a flat rank histogram; rather, all ranks are expected to obtain with probability $1/K$, apart from the extreme ones, which have probability $1/2K$.

For ensembles which approximate quantiles, this should still be approximately true. An ensemble which gets, for example, re-calibrated or 'de-biased' using the CRPS can thus, in general, not be expected to display a flat rank histogram; this is obviously of practical relevance. The findings of this article have been illustrated by a simple numerical experiments using artificial data.

In some sense, the simple bottom line of this contribution is that ensembles which are constructed differently might feature different rank histograms, even if they are based on the same reliable forecast distribution. Referee Christopher Ferro pointed out that this article can be considered part of a wider discussion on what has been called *consistency* of performance measures by Murphy (1997). The issue, to which Nau (1985) refers as 'Considerations in Choosing an Admissible Set' in an even more general context, can be described as follows. A rational forecaster will pick her forecast (say F) independent of the scoring rule (which determines her reward), as long as the scoring rule is proper, and as long as no further restrictions limit her choice. In this article, a situation was considered in which the forecast *is* subject to restrictions, namely it had to be a step function, with the only freedom being the location of the steps (and possibly the step height). In the case of the CRPS score, the forecaster will chose the step locations to be certain quantiles of F, but this need not be so for other scores! In other words, the interpretation of the step locations does depend on the employed scoring rule, unlike F, which is the same for all (proper) scoring rules. To give another example, suppose the forecaster is forced to use normal densities, described by mean and variance. Unless F is also normal, the mean and variance chosen by the forecaster will depend on the employed scoring rule; in particular, they will in general not be equal to the mean and variance of F. The bottom line is that, since practical issues usually impose restrictions on possible forecast distributions, the forecast chosen by the forecaster usually *does* depend on the scoring rule.

Acknowledgements

Stimulating discussions with the members of the Time Series Analysis group at the Max Planck Institute for the Physics of Complex Systems are sincerely acknowledged, in particular Stefan Siegert and Holger Kantz. Constructive comments by an anonymous and a known referee were helpful in further improving the paper.

Appendix

Derivation of Eq. (8)

Substituting with Eq. (7) in Eq. (4), we obtain

$$\begin{aligned} S(G_e, y) &= \int \Big\{ \sum_k w_k H(x - e_k) - H(x - y) \Big\}^2 \mathrm{d}x \\ &= \int \Big\{ \sum_k w_k \big[H(x - e_k) - H(x - y) \big] \Big\}^2 \mathrm{d}x, \end{aligned} \tag{A1}$$

since $\sum_k w_k = 1$. We will now analyse

$$I(x; \eta, y) = H(x - \eta) - H(x - y)$$

as a function of x. This function is constant except at $x = \eta$ and $x = y$, where it has jumps of magnitudes 1 and -1, respectively. Therefore,

$$\begin{aligned} &\int I(x; \eta_1, y)\, I(x; \eta_2, y)\, \mathrm{d}x \\ &\quad = \big(\min\{\eta_1, \eta_2\} - y \big)_+ + \big(y - \max\{\eta_1, \eta_2\} \big)_+, \end{aligned} \tag{A2}$$

where $(x)_+ = x$ if $x > 0$ and zero otherwise. In particular,

$$\int I(x; \eta, y)^2 \, \mathrm{d}x = |\eta - y|. \tag{A3}$$

We now expand the squares in Eq. (A1) and use Eqs (A2) and (A3), yielding

$$\begin{aligned}
&S(G_e, y) \\
&= \int \sum_i w_i^2 I(x; e_i, y)^2 \, \mathrm{d}x \\
&\qquad + 2 \int \sum_{j,i<j} w_i w_j I(x; e_i, y) \, I(x; e_j, y) \, \mathrm{d}x \\
&= \sum_i w_i^2 |e_i - y| + 2 \sum_{j,i<j} w_i w_j \big\{ (e_i - y)_+ + (y - e_j)_+ \big\} \\
&= \sum_j w_j^2 |e_j - y| + 2 \sum_j \Big(\sum_{i>j} w_i \Big) w_j (e_j - y)_+ \\
&\qquad + 2 \sum_j \Big(\sum_{i<j} w_i \Big) w_j (y - e_j)_+ \\
&= \sum_j w_j (2 W_{\geq j} - w_j) \cdot (e_j - y)_+ \\
&\qquad + \sum_j w_j (2 W_{\leq j} - w_j) \cdot (y - e_j)_+ \, ,
\end{aligned}$$

with $W_{\leq j}$ and $W_{\geq j}$ being the cumulative sums $\sum_{k \leq j} w_k$ and $\sum_{k \geq j} w_k$, respectively. This expression for $S(G_e, y)$ can be written as the weighted sum in Eq. (8) with $\alpha_j = W_{\leq j} - (w_j/2)$.

References

Brier GW. 1950. Verification of forecasts expressed in terms of probabilities. *Mon. Weather Rev.* **78**: 1–3.

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

Bröcker J, Smith LA. 2008. From ensemble forecasts to predictive distribution functions. *Tellus A* **60**: 663–678.

Brown TA. 1970. *Probabilistic forecasts and reproducing scoring systems*. Tech. Report RM–6299–ARPA, RAND Corporation: Santa Monica, CA.

Epstein ES. 1969. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol.* **8**: 985–987.

Friederichs P, Hense A. 2008. A probabilistic forecast approach for daily precipitation totals. *Weather and Forecasting* **23**: 659–673.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102**: 359–378.

Gneiting T, Raftery AE, Westveld III AH, Goldmann T. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**: 1098–1118.

Good IJ. 1952. Rational decisions. *J. R. Statist. Soc.* **14**: 107–114.

Grimit EP, Gneiting T, Berrocal VJ, Johnson NA. 2006. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Q. J. R. Meteorol. Soc.* **132**: 2925–2942.

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

Jewson S. 2003. Moment based methods for ensemble assessment and calibration. *arXiv:physics*/0309042v1 [physics.ao-ph].

Murphy AH. 1997. Forecast verification. In *Economic value of weather and climate forecasts*, Murphy AH, Katz RW (eds) 19–74. Cambridge University Press: Cambridge, UK.

Nau RF. 1985. Should scoring rules be 'effective'? *Management Sci.* **31**: 527–535.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**: 1155–1174.

Roulston MS, Smith LA. 2002. Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.* **130**: 1653–1660.

Roulston MS, Smith LA. 2003. Combining dynamical and statistical ensembles. *Tellus* **55A**: 16–30.

Roulston MS, Kaplan DT, von Hardenberg J, Smith LA. 2003. Using medium range weather forecasts to improve the value of wind energy production. *Renewable Energy* **28**: 585–602.

Roulston MS, Ellepola J, von Hardenberg J, Smith LA. 2005. Forecasting wave height probabilities with numerical weather prediction models. *Ocean Eng.* **32**: 1841–1863.

Savage LJ. 1971. Elicitation of personal probabilities and expectation. *J. Amer. Statist. Assoc.* **66**: 783–801.

Sloughter JM, Raftery AE, Gneiting T, Fraley C. 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Weather Rev.* **135**: 3209–3220.

