Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems

HANS HERSBACH

Koninklijk Nederlands Meteorologisch Instituut, De Bilt, Netherlands

(Manuscript received 30 November 1999, in final form 3 May 2000)

ABSTRACT

Some time ago, the continuous ranked probability score (CRPS) was proposed as a new verification tool for (probabilistic) forecast systems. Its focus is on the entire permissible range of a certain (weather) parameter. The CRPS can be seen as a ranked probability score with an infinite number of classes, each of zero width. Alternatively, it can be interpreted as the integral of the Brier score over all possible threshold values for the parameter under consideration. For a deterministic forecast system the CRPS reduces to the mean absolute error.

In this paper it is shown that for an ensemble prediction system the CRPS can be decomposed into a reliability part and a resolution/uncertainty part, in a way that is similar to the decomposition of the Brier score. The reliability part of the CRPS is closely connected to the rank histogram of the ensemble, while the resolution/uncertainty part can be related to the average spread within the ensemble and the behavior of its outliers. The usefulness of such a decomposition is illustrated for the ensemble prediction system running at the European Centre for Medium-Range Weather Forecasts. The evaluation of the CRPS and its decomposition proposed in this paper can be extended to systems issuing continuous probability forecasts, by realizing that these can be interpreted as the limit of ensemble forecasts with an infinite number of members.

1. Introduction

Appropriate verification tools are essential in understanding the abilities and weaknesses of (probabilistic) forecast systems.

Verification is often focused on specific (weather) events. Such a binary event either occurs, or does not occur, and is forecast to occur or not to occur, with certain probabilities p and $1 - p$ respectively. Examples of such events are more than 10-mm precipitation in 24 h or an anomaly (from a climatological mean) of more than 50 m of the geopotential at 500 hPa. Several well-established tools exist that test how accurately the forecast system is able to describe the occurrence and non-occurrence of the event under consideration, that is, how good the agreement is between the forecasted probabilities and observed states. Examples of scores, which are commonly used by operational centers such as the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction are Brier scores (Brier 1950), the Relative Operating Characteristics (ROC) curves (Mason 1982; Stanski et al. 1989), and economic cost–loss analyses (see, e.g., Katz and Murphy 1997; or Richardson 1998, 2000).

The (half) Brier score is one of the oldest verification tools in use. From its numerical value alone the quality of a forecast system is difficult to assess. An attractive property of the Brier score, however, is that it can be decomposed into a reliability, a resolution, and an uncertainty part (Murphy 1973). The reliability tests whether the forecast system has the correct statistical properties. It can be presented in a graphical way by the so-called reliability diagram. The uncertainty is the Brier score one would obtain when only the climatological frequency for the occurrence of the event is available. The resolution shows the impact obtained by issuing case-dependent probability forecasts (which do not always equal the probability based on climatology). Therefore, the decomposition of the Brier score gives a detailed insight into the performance of the forecast system with respect to the event under consideration.

Binary events only highlight one aspect of the forecast. Such a single aspect may be quite relevant. For instance, certain extreme events can lead to economic losses, which could be avoided with the help of an accurate forecast system. This kind of issue is addresses by the ROC curve and economic cost–loss analyses. However, it may be desirable to obtain a broader overall view of performance. Several tools in this direction exist. It should however be mentioned that the term overall is often still restricted to the behavior of one forecast

Corresponding author address: Dr. Hans Hersbach, KNMI, P.O. Box 201, 3730 AE Utrecht, Netherlands.
E-mail: hersbach@knmi.nl

parameter only, such as precipitation or the geopotential at 500 hPa.

An example is the Talagrand diagram (Talagrand and Vautard 1997), also known as the rank histogram (Hamill and Collucci 1997) or the binned probability ensemble (Anderson 1996). This tool is tailor made for an ensemble system, that is, in case the probability density function (PDF) is represented by an ensemble of forecasts. Given such an ensemble, its N members divide the permissible range of the parameter of interest into $N + 1$ bins. The verifying analysis will be found to be in one of these bins. If all members are assumed to be equally weighted and representative, it is expected that, on average, each bin should be equally populated by the verifying analyses. Deviations from such a flat rank histogram indicate a violation of the above-made assumptions. For instance, a too high frequency of outliers is an indication that the average spread within the ensemble system is too low.

Another example is the ranked probability score (RPS) (see Epstein 1969; Murphy 1969, 1971). It is a generalization of the (half) Brier score. Instead of two options (event occurs or does not occur), the range of the parameter of interest is divided into more classes. In addition, the RPS contains a sense of distance of how far the forecast was found from reality. For a deterministic forecast for instance, the RPS is proportional to the number of classes by which the forecast missed the verifying analysis. Although the choice and number of classes may be prescribed by the specific application, the exact value of RPS will depend on this choice. It is possible to take the limit of an infinite number of classes, each with zero width. This leads to the concept of the continuous ranked probability score (CRPS) (Brown 1974; Matheson and Winkler 1976; Unger 1985; Bouttier 1994). This CRPS has several appealing properties. First of all, it is sensitive to the entire permissible range of the parameter of interest. Second, its definition does not require, such as for the RPS, the introduction of a number of predefined classes, on which results may depend. In addition, it can be interpreted as an integral over all possible Brier scores. Finally, for a deterministic forecast, the CRPS is equal to the mean absolute error (MAE) and, therefore, has a clear interpretation.

Despite these advantages, the CRPS is a single quantity, from which it is difficult to disentangle the detailed behavior of a forecast system. It would be desirable to be able to decompose the CRPS like it is possible for the Brier score. In this paper it is shown how for an ensemble prediction system this indeed can be achieved. In a similar way to the Brier score, the CRPS is shown to be decomposable into a reliability part, an uncertainty part, and a resolution part. The reliability part tests whether for each bin i on average the verifying analysis was found to be with a fraction i/N below this bin. It has a close relation to the rank histogram. The uncertainty part is equal to the CRPS one would receive, in case only a PDF-based on climatology would be available. The resolution finally expresses the improvement gained by issuing probability forecasts that are case dependent. It is shown that the resolution is sensitive to the average ensemble spread and the frequency and magnitude of the outliers. Finally, it is illustrated how the various contributions to the CRPS can be presented in a graphical way, like the reliability diagram of the Brier score.

The paper is organized as follows. In section 2 the CRPS is defined, and some characteristics are mentioned. The uncertainty part of the CRPS is highlighted in section 3. In section 4, the full decomposition for an ensemble system is derived. As an example, the decomposition of the CRPS for total precipitation in the ensemble prediction system (EPS) running at ECMWF is presented in section 5. A summary and some concluding remarks are made in section 6.

2. The continuous ranked probability score

Let the parameter of interest be denoted by x. For instance, x could be the 2-m temperature or 10-m wind speed. Suppose that the PDF forecast by an ensemble system is given by $\rho(x)$ and that x_a is the value that actually occurred. Then the continuous ranked probability score (Brown 1974; Matheson and Winkler 1976; Unger 1985; Bouttier 1994), expressing some kind of distance between the probabilistic forecast ρ and truth x_a, is defined as

$$\text{CRPS} = \text{CRPS}(P, x_a) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 \, dx. \quad (1)$$

Here, P and P_a are cumulative distributions:

$$P(x) = \int_{-\infty}^{x} \rho(y) \, dy \quad \text{and} \quad (2)$$

$$P_a(x) = H(x - x_a), \quad (3)$$

where

$$H(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases} \quad (4)$$

is the well-known Heaviside function. So, $P(x)$ is the forecasted probability that x_a will be smaller than x. Obviously, for any cumulative distribution, $P(x) \in [0, 1]$, $P(-\infty) = 0$, and $P(\infty) = 1$. This is also true for parameters that are only defined on a subdomain of $\Re$. In that case $\rho(x) = 0$ and P constant outside the domain of definition. The CRPS measures the difference between the predicted and occurred cumulative distributions. Its minimal value of zero is only achieved for $P = P_a$, that is, in the case of a perfect deterministic forecast. Note that the CRPS has the dimension of the parameter x (which enters via the integration over dx).

In practice the CRPS is averaged over an area and a number of cases:

$$\overline{\text{CRPS}} = \sum_k w_k \text{CRPS}(P^k, x_a^k), \tag{5}$$

where k labels the considered grid points and cases. The weights w_k may depend on k (for instance proportional to the cosine of latitude).

The CRPS can be seen as the limit of a ranked probability score with an infinite number of classes, each with zero width.

There is a direct relation between the CRPS and the Brier score (Brier 1950). The Brier score (BS) is a verification tool for the prediction of the occurrence of a specific event. Usually, such an event is characterized by a threshold value x_t. The event is said to have happened ($O = 1$) if $x_a \leq x_t$ and not happened ($O = 0$) if $x_a > x_t$. If p is the forecast probability that the event will occur, the Brier score is defined as

$$\text{BS}(x_t) = \sum_k w_k (p^k - O^k)^2. \tag{6}$$

It is not difficult to see that $p^k = P^k(x_t)$ and $O^k = P_a^k(x_t)$ and therefore

$$\overline{\text{CRPS}} = \int_{-\infty}^{\infty} \text{BS}(x_t)\, dx_t. \tag{7}$$

For a deterministic forecast, that is, $x = x_d$ without any specified uncertainty, $P(x) = H(x - x_d)$. In that case, the integrand of Eq. (1) is either zero or one. The nonzero contributions are found in the region where $P(x)$ and $P_a(x)$ differ, which is the interval between x_d and x_a. As a result,

$$\overline{\text{CRPS}} = \sum_k w_k |x_d^k - x_a^k|, \tag{8}$$

which is the MAE.

3. The uncertainty of the CRPS

For an ensemble prediction system, the forecast PDF will in general be case dependent. If instead, only climatological information about the behavior of the quantity x is available, the same probability forecast $P^k = P_{\text{cli}}$ will be made for each situation. In that case,

$$\begin{aligned}\overline{\text{CRPS}} &= \sum_k w_k \int_{-\infty}^{\infty} [P_{\text{cli}}(x) - H(x - x_a^k)]^2\, dx \\ &= \int_{-\infty}^{\infty} \left[\sum_k w_k P_{\text{cli}}^2(x) - 2P_{\text{cli}}(x) \sum_k w_k H(x - x_a^k) \right. \\ &\qquad \left. + \sum_k w_k H^2(x - x_a^k) \right] dx.\end{aligned}$$

Note that $\Sigma_k w_k = 1$ by definition, and $H^2 = H$. If one defines

$$P_{\text{sam}}(x) = \sum_k w_k H(x - x_a^k), \tag{9}$$

the CRPS can be rewritten as

$$\begin{aligned}\overline{\text{CRPS}} &= \int_{-\infty}^{\infty} [P_{\text{cli}}^2(x) - 2P_{\text{cli}}(x)P_{\text{sam}}(x) \\ &\qquad + P_{\text{sam}}(x)]\, dx \\ &= \overline{R} + \overline{U},\end{aligned} \tag{10}$$

where

$$\overline{R} = \int_{-\infty}^{\infty} [P_{\text{cli}}(x) - P_{\text{sam}}(x)]^2\, dx, \quad \text{and} \tag{11}$$

$$\overline{U} = \int_{-\infty}^{\infty} P_{\text{sam}}(x)[1 - P_{\text{sam}}(x)]\, dx. \tag{12}$$

The distribution P_{sam} is the cumulative distribution based on the sample used in the verification. If, for instance, all M weights would be equal, so $w_k = 1/M$, then $P_{\text{sam}}(x)$ is just the fraction of cases in which the verifying analysis was found to be smaller than x. The value of $P_{\text{sam}}(x)$ also equals the sample frequency of occurrence $o(x_t)$ for the Brier score with threshold $x_t = x$.

From Eqs. (10)–(12) it is seen that the CRPS based on climatology is minimal when P_{cli} is equal to P_{sam}. The impact on the CRPS due to a deviation from the sample statistics is expressed by Eq. (11).

The lowest possible value of a CRPS based on climatology is given by Eq. (12). It is solely determined by the climatology within the sample and does not depend on the performance of the forecast model. Expression (12) is equal to the integral of the uncertainty U (Murphy 1973; or see, e.g., Wilks 1995) of the Brier score over all possible thresholds:

$$U(x_t) = o(x_t)[1 - o(x_t)] \Rightarrow \overline{U} = \int_{-\infty}^{\infty} U(x_t)\, dx_t. \tag{13}$$

Here

$$o(x_t) = \sum_k w_k H(x_t - x_a^k) = P_{\text{sam}}(x_t) \tag{14}$$

is the observed frequency that the event $x < x_t$ occurred. Therefore, it is very natural to define $\overline{U}$ as the uncertainty of the CRPS. It is the CRPS based on the sample climatology. It is proportional to the standard deviation of the sample distribution $\rho_{\text{sam}} = dP_{\text{sam}}/dx$, because the main contribution to the integral in Eq. (12) comes from the region in x where P_{sam} is significantly different from 0 and 1. An illustration is given in Fig. 1. To be more exact, the sample distribution ρ_{sam} can always be written as

$$\rho_{\text{sam}}(x) = \frac{1}{\sigma}\rho_0\left(\frac{x}{\sigma}\right) \Rightarrow P_{\text{sam}}(x) = P_0\left(\frac{x}{\sigma}\right), \tag{15}$$

where ρ_0 is a distribution with $\sigma = 1$ (for instance similar to a standardized Gaussian) and P_0 [see Eq. (2)] its cumulative distribution. From the uncertainty

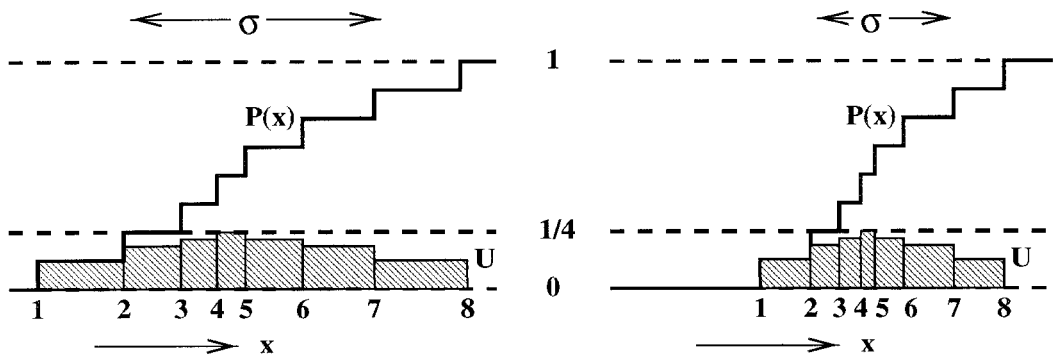

FIG. 1. Sample distribution P_{sam}, as defined in Eq. (9), for two samples of eight cases, all with equal weight. The shaded area represents the corresponding uncertainty U [see Eq. (12)]. It is proportional to the standard deviation σ of the distribution.

$$\overline{U}_0 \equiv \int_{-\infty}^{\infty} P_0(y)[1 - P_0(y)]\, dy \tag{16}$$

of this distribution, it follows that

$$\overline{U} = \int_{-\infty}^{\infty} P_0\left(\frac{x}{\sigma}\right)\left[1 - P_0\left(\frac{x}{\sigma}\right)\right] dx = \sigma\overline{U}_0, \tag{17}$$

so indeed proportional to σ.

It should be noted that the term climatology depends on the degree of desired sophistication. The most crude level would be to assume the same climatological distribution at all grid points and cases. The mean climatological value of x, however, may be quite location and seasonal dependent. The mean 2-m temperature of Norway in January, for instance, is much lower than that of Spain in March. This would result in a very broad sample distribution and, therefore, to a large uncertainty. In order to correct for this, as a first step, the variable x can be redefined as being the anomaly with respect to the local climatology. The definition of the CRPS is invariant for such a shift in the variable x, as is easily seen from Eq. (1). As a consequence, the distribution P_{sam} will change, because for each k in Eq. (9) a different shift may have been applied. This should result in a distribution that is much sharper, so the uncertainty $\overline{U}$ in Eq. (12) should be smaller. For a parameter in which the permissible range is limited, like precipitation or 10-m wind speed, such an approach may not be profitable. The reason for this is that the sample distribution obtained in this way (based on anomalies) will for part of the locations lead to nonvanishing probabilities outside the permissible range.

Finally, the entire climatological distribution (so not just its mean) could be chosen to depend on the location and/or season, so $P^k = P_{\text{cli,location,season}}$. For this, the best achievable distribution would be a location/seasonal-dependent sample distribution, also given by Eq. (9) but in which the sum (and the normalization of the weights) is restricted to all points k that belong to the same location and or season. Again, the resulting uncertainty is expected to become lower. For parameters like precipitation this will also lead to a lower uncertainty.

This section will be concluded by showing how the uncertainty can be evaluated in practice. The most straightforward method is to substitute definition (9) into Eq. (12):

$$\overline{U} = \sum_{k,l} w_k w_l \int_{-\infty}^{\infty} H(x - x_a^k)[1 - H(x - x_a^l)]\, dx. \tag{18}$$

The integrand will only be nonzero when both $H(x - x_a^k) = 1$ and $H(x - x_a^l) = 0$. This condition can only be met when $x_a^k < x_a^l$, in which case the integral is $x_a^l - x_a^k$. As a result,

$$\overline{U} = \sum_{k,l<k} w_k w_l |x_a^k - x_a^l|. \tag{19}$$

Another way to calculate $\overline{U}$ is to realize that Eq. (9) is based on a finite number of verifying analyses. Therefore P_{sam} will be piecewise constant (see, e.g., Fig. 1). It is zero for $x = -\infty$ and each time an x_a^k is passed, it makes a jump of w_k. Beyond the largest verifying analysis in the set, $P_{\text{sam}} = 1$. Now if the x_a^k are ordered from small to large, then

$$\overline{U} = \sum_{k=1}^{N-1} p_k(1 - p_k)[x_a^{\text{sort}(k+1)} - x_a^{\text{sort}(k)}], \tag{20}$$

where

$$p_k = p_{k-1} + w_{\text{sort}(k)} \quad \text{and} \quad p_0 = 0.$$

Evaluation (19) is of order M^2, where M is the size of the sample set. If M becomes on the order of a few thousand, this evaluation becomes time consuming. In addition, roundoff errors are expected to become nonnegligible. Method (20) only involves a sum of order M. The price to be paid is that the x_k^a should be sorted first. However, efficient sorting algorithms, such as quicksort or heapsort (see Press et al. 1989), are of order $M \log(M)$. Therefore, this latter method is still quite feasible and accurate for very large samples.

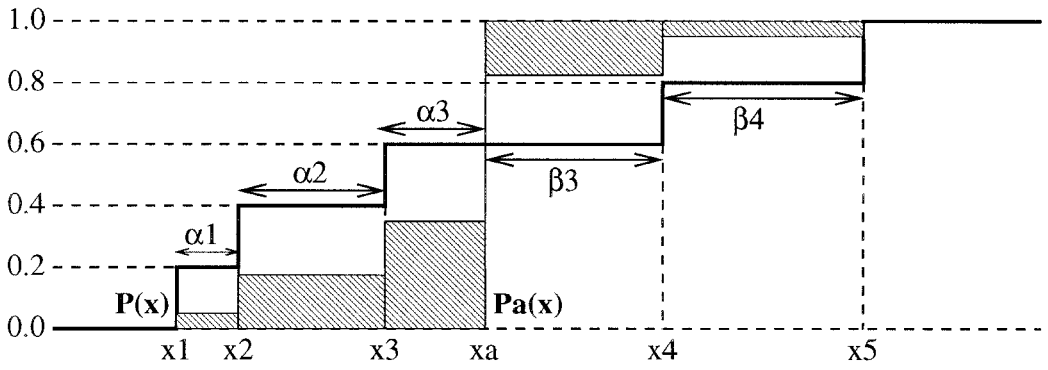

FIG. 2. Cumulative distribution for an ensemble $\{x_1, \ldots, x_5\}$ of five members (thick solid line) and for the verifying analysis x_a (thin solid line). The CRPS is represented by the shaded area. The α_i and β_i are defined in Eq. (26).

4. The CRPS for an ensemble system

a. The cumulative distribution of an ensemble

For an ensemble system, such as EPS, an equal weight is given to each of its members. Therefore, the probability assigned to the occurrence of a certain event is given by the fraction of members that predict the event. Effectively, for the variable x this means that the cumulative distribution forecasted by the ensemble system is given by

$$P(x) = \frac{1}{N}\sum_{i=1}^{N} H(x - x_i), \tag{21}$$

where $x_1, \ldots, x_N$ are the outcomes of the N ensemble members. From now on it is assumed that the members are ordered, that is,

$$x_i \leq x_j, \quad \text{for } i < j. \tag{22}$$

The cumulative distribution P is a piecewise constant function. Transitions occur at the values x_i:

$$P(x) = p_i \equiv \frac{i}{N}, \quad \text{for } x_i < x < x_{i+1}, \tag{23}$$

in which $x_0 = -\infty$ and $x_{N+1} = \infty$ are introduced for convenience. An example of the cumulative distribution for an ensemble of five members is given (thick solid curve) in Fig. 2.

b. Decomposition for a single case

The CRPS, as defined in Eq. (1), can be evaluated as follows:

$$c_i = \int_{x_i}^{x_{i+1}} [p_i - H(x - x_a)]^2\, dx \Rightarrow \text{CRPS} = \sum_{i=0}^{N} c_i. \tag{24}$$

Depending on the position of the verifying analysis x_a, $H(x - x_a)$ will be either 0, or 1, or partly 0, partly 1, in the interval $[x_i, x_{i+1}]$. For each of these three possible situations, c_i can be written as

$$c_i = \alpha_i p_i^2 + \beta_i (1 - p_i)^2, \tag{25}$$

where

$0 < i < N$	α_i	β_i
$x_a > x_{i+1}$	$x_{i+1} - x_i$	0
$x_{i+1} > x_a > x_i$	$x_a - x_i$	$x_{i+1} - x_a$
$x_a < x_i$	0	$x_{i+1} - x_i$

. (26)

Note that the α_i and β_i have the dimension of the parameter x.

For the example given in Fig. 2, the verifying analysis is in between x_3 and x_4. Therefore, for this case $\beta = 0$ for $i = 1$ and 2, and $\alpha = 0$ for $i = 4$. Only for $i = 3$ both α and β are nonzero.

Some care should be taken for $i = 0$ and $i = N$. These concern the intervals $(-\infty, x_1]$ and $[x_N, \infty)$, respectively, and for which $p_i = 0$ and $p_i = 1$, respectively. These two intervals will only contribute to the CRPS in cases when the verifying analysis is an outlier, that is, when it is outside the range of the ensemble. In this situation Eq. (25) can also be used, but with

Outlier	α_i	β_i
$x_a < x_1$	0	$x_1 - x_a$
$x_N < x_a$	$x_a - x_N$	0

. (27)

In Fig. 3 an example is given in which the verifying analysis is found to be below the ensemble (left panel) and above the ensemble (right panel). For the first case, there will be a contribution from β_0, being the difference between x_a and the smallest ensemble member. In the second case, α_N is nonzero and equal to the distance of x_a from the largest ensemble member. Outliers can contribute significantly to the CRPS, because nonzero values of β_0 and α_N are weighted stronger than other α's and β's (see, e.g., the shaded areas in Fig. 3).

c. The average over a set of cases

For M cases and/or grid points, each with a weight w_k, the average CRPS [Eq. (5)] can be found as

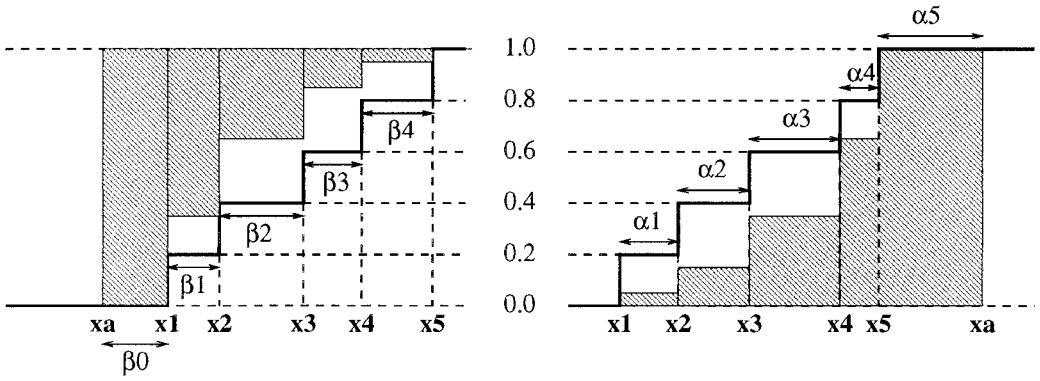

FIG. 3. The same as in Fig. 2 but now for the case that the verifying analysis is outside the ensemble (outlier). Only when x_a is below the ensemble (left panel) is β_0 [see Eq. (27)] nonzero. And only when case x_a is above the ensemble (right panel) is α_N nonzero. The CRPS is given by the shaded area. Note that β_0 and α_N are weighted stronger than other α's and β's.

$$\overline{\text{CPRS}} = \sum_{i=0}^{N} [\overline{\alpha}_i p_i^2 + \overline{\beta}_i (1 - p_i)^2], \tag{28}$$

where

$$\overline{\alpha}_i = \sum_k w_k \alpha_i^k \quad \text{and} \quad \overline{\beta}_i = \sum_k w_k \beta_i^k \tag{29}$$

are the weighted average values of α_i and β_i.

The quantities $\overline{\alpha}_i$ and $\overline{\beta}_i$ can be expressed into two quantities $\overline{g}_i$ and $\overline{o}_i$, which both have a physical interpretation. First the case $0 < i < N$ is considered. Let

$$\overline{g}_i = \overline{\alpha}_i + \overline{\beta}_i \quad \text{and} \tag{30}$$

$$\overline{o}_i = \frac{\overline{\beta}_i}{\overline{\alpha}_i + \overline{\beta}_i}. \tag{31}$$

It can be seen from Eq. (26) that $\overline{g}_i$ is the average width of bin number i:

$$\overline{g}_i = \overline{x_{i+1} - x_i}, \quad \text{for } 0 < i < N. \tag{32}$$

For the moment concentrate on a specific value of i. Then, for most cases, the verifying analysis will not lie in the interval $[x_i, x_{i+1}]$. Therefore, usually, α_i will be zero and β_i is equal to the width of bin number i, or vice versa. The first case applies to the situation in which the verifying analysis was found to be smaller than the ensemble member i, as can be seen from Eq. (26), the second case to which it was found to be larger than member $i + 1$. Taking this in mind, $\overline{o}_i$ can be seen to be closely related to the average frequency that the verifying analysis was found to be below $\frac{1}{2}(x_i + x_{i+1})$. Ideally these observed frequencies should match with the forecasted probability that the verifying analysis is to be found below the ith interval. Such a consistency is closely related to the flatness of the rank histogram [also known as Talagrand diagram or binned probability ensemble; see, e.g., Anderson (1996), Talagrand and Vautard (1997), or Hamill and Collucci (1997)].

For the outliers, $\overline{o}_0$ and $\overline{o}_N$ is defined as the (weighted) frequency that x_a was found to be smaller than x_1 and x_N, respectively. Here, $\overline{g}_{0,N}$ is defined as the average length of the outlier, given that it occurred:

$$\overline{o}_0 = \sum_k w_k H(x_1^k - x_a^k) \qquad \overline{g}_0 = \overline{\beta}_0 / \overline{o}_0 \quad \text{and}$$

$$\overline{o}_N = \sum_k w_k H(x_N^k - x_a^k) \qquad \overline{g}_N = \overline{\alpha}_N / (1 - \overline{o}_N). \tag{33}$$

The user may verify that for all $i = 0, \ldots, N$, so including the outliers

$$\overline{\alpha}_i p_i^2 = \overline{g}_i (1 - \overline{o}_i) p_i^2 \quad \text{and}$$

$$\overline{\beta}_i (1 - p_i)^2 = \overline{g}_i \overline{o}_i (1 - p_i)^2. \tag{34}$$

The average CRPS [see Eq. (5)] can now be decomposed as

$$\begin{aligned} \overline{\text{CRPS}} &= \sum_{i=0}^{N} \overline{g}_i [(1 - \overline{o}_i) p_i^2 + \overline{o}_i (1 - p_i)^2] \\ &= \overline{\text{Reli}} + \text{CRPS}_{\text{pot}}, \end{aligned} \tag{35}$$

where

$$\overline{\text{Reli}} = \sum_{i=0}^{N} \overline{g}_i (\overline{o}_i - p_i)^2, \qquad p_i = \frac{i}{N} \quad \text{and} \tag{36}$$

$$\text{CRPS}_{\text{pot}} = \sum_{i=0}^{N} \overline{g}_i \overline{o}_i (1 - \overline{o}_i). \tag{37}$$

This decomposition looks similar to the decomposition of the Brier score as it was introduced by Murphy (Murphy 1973; or see, e.g., Wilks 1995). The interpretation, however, is somewhat different.

The quantity $\overline{\text{Reli}}$ is identified as the reliability part of the CRPS. For a Brier score the reliability tests whether for all cases in which a certain probability p was forecast, on average, the event occurred with that fraction p. Here, it is tested whether, on average, the frequency $\overline{o}_i$ that the verifying analysis was found to be below the middle of interval number i is proportional to i/n. Therefore, it is tested here whether the ensemble is capable of generating cumulative distributions that

have, on average, this desired statistical property. The reliability (36) is closely connected to the rank histogram, which shows whether the frequency that the verifying analysis was found in bin number i is equal for all bins. The rank histogram does not take care of the width of the ensemble. It only counts how often the verifying analysis was located in a bin, regardless of the width of the bins. The reliability $\overline{\text{Reli}}$ does take this into account, because the larger a bin width (and therefore the larger the spread) the more weight it has in $\overline{\alpha}_i$ and $\overline{\beta}_i$ and therefore $\overline{o}_i$. Note that $\overline{\text{Reli}}$ has a dimension (of x), while the reliability of the Brier score is dimensionless. The term CRPS_{pot} given in Eq. (37) is called the potential CRPS (in analogy with Murphy and Epstein 1989), because it is the CRPS one would obtain after the probabilities p_i would have been retuned, such that the system would become perfectly reliable, that is, for which $\overline{\text{Reli}} = 0$. It is sensitive to the average spread of the ensemble. The narrower the ensemble system, the smaller the g_i and the smaller Eq. (37). The potential CRPS is also sensitive to outliers. Too many and too large outliers will result in large values of $\overline{g}_0\overline{o}_0$ and $\overline{g}_N(1 - \overline{o}_N)$ and therefore affect CRPS_{pot} considerably. Although the small average bin widths $\overline{g}_1, \ldots, \overline{g}_N$ of an ensemble system with a too small spread may have a positive impact on the potential CRPS, the too high frequency of outliers and the large magnitudes of such outliers will have a clear negative impact. Given a certain degree of unpredictability, the optimal value for CRPS_{pot} will be achieved for an ensemble system in which the spread and the statistics of outliers are in balance.

The uncertainty $\overline{U}$ as defined in (12) can be seen as the potential reliability for a forecast system based on the sample climatology. Such a system is by definition, perfectly reliable. To see the relation between Eqs. (12) and (37), the integral over x in Eq. (12) is to be approximated by a sum over intervals Δx_i, each representing an equal part of $1/N$ of integrated probability. The Δx_i may be identified with the widths g_i and the $P_{\text{sam}}(x_i)$ with the observed frequencies o_i. As a result, these approximations lead to Eq. (37). It may be clear that it is desirable for an ensemble system that CRPS_{pot} is smaller than the potential CRPS based on climatology. Therefore, the potential CRPS may, although perhaps somewhat artificially, be further decomposed into

$$\text{CRPS}_{\text{pot}} = \overline{U} - \overline{\text{Resol}}. \tag{38}$$

This gives the following decomposition:

$$\overline{\text{CRPS}} = \overline{\text{Reli}} - \overline{\text{Resol}} + \overline{U}. \tag{39}$$

The resolution $\overline{\text{Resol}}$ is nothing else than the difference between the potential CRPS and the climatological uncertainty. The ensemble system has positive resolution if it performs better than the climatological probabilistic forecast. In the previous section it was discussed that the uncertainty (12) depends on the level of sophistication. Therefore, the same is true for the resolution. Unlike the resolution of the Brier score, the resolution part of the CRPS need not be positive definite.

d. Relation to the decomposition of the Brier score

In section 2 it was shown that the $\overline{\text{CRPS}}$ can be seen as an integral of the Brier score over all possible thresholds [see Eq. (7)]. The question may emerge whether the terms in decomposition (39) are also equal to the reliability, resolution, and uncertainty of the Brier score integrated over all possible thresholds.

The Brier score defined by Eq. (6) (with thresholds x) may be stratified with respect to the set of allowable probabilities $p_i = 0, 1/N, \ldots, 1$:

$$\text{BS}(x) = \sum_{i=0}^{N} g_i(x)\{o_i(x)(1 - p_i)^2 + [1 - o_i(x)]p_i^2\}. \tag{40}$$

Here g_i is the (weighted) fraction of cases in which a probability $p = p_i$ was issued, while o_i is the fraction of such cases in which indeed the event was observed. Note that both quantities depend on the value of the threshold x.

After some algebra, it follows that the Brier score can be decomposed into

$$\begin{aligned}\text{BS}(x) &= \sum_{i=0}^{N} g_i(x)[o_i(x) - p_i]^2 \\ &\quad - \sum_{i=0}^{N} g_i(x)[o_i(x) - o(x)]^2 + o(x)[1 - o(x)] \\ &= \text{Reli}(x) - \text{Resol}(x) + U(x),\end{aligned} \tag{41}$$

where

$$o(x) = \sum_{i=0}^{N} g_i(x)o_i(x) \tag{42}$$

is the (weighted) frequency that the event occurred within the sample. In the appendix, Eq. (42) is shown to be equal to definition (14). There it is also shown [see Eqs. (A8)–(A11)] that the integral of $g_i(x)$ and $g_i(x)o_i(x)$ over x is equal to the $\overline{g_i}$ and $\overline{g}_i\overline{o}_i$, respectively, defined by Eqs. (30)–(33). When integral (7) is performed, the relation between decompositions (39) and (41) can be established:

$$\overline{\text{CRPS}} = \langle\text{Reli}\rangle - \langle\text{Resol}\rangle + \langle U\rangle, \tag{43}$$

where

$$\langle\text{Reli}\rangle \equiv \int_{-\infty}^{\infty} \text{Reli}(x)\, dx = \overline{\text{Reli}} + D, \tag{44}$$

$$\langle\text{Resol}\rangle \equiv \int_{-\infty}^{\infty} \text{Resol}(x)\, dx = \overline{\text{Resol}} + D, \quad \text{and} \tag{45}$$

$$\langle U\rangle \equiv \int_{-\infty}^{\infty} U(x)\, dx = \overline{U}. \tag{46}$$

Here

$$D = \sum_{i=0}^{N-1} \overline{g}_i(\overline{o_i^2} - \overline{o}_i^2) + \overline{g}_N[\overline{(1 - o_N)^2} - (1 - \overline{o}_N)^2], \quad (47)$$

where for $z_i = o_0, o_1, \ldots, (1 - o_N)$

$$\overline{z_i^2} \equiv \frac{1}{\overline{g}_i} \int_{-\infty}^{\infty} g_i(x) z_i^2(x)\, dx. \quad (48)$$

In general, D will be nonzero. Therefore, the integration of the resolution and reliability of the Brier score over all possible thresholds, in general, differs from the reliability and resolution, respectively, of the $\overline{\text{CRPS}}$. Only the integral over all uncertainties $U(x)$ is equal to the uncertainty of the $\overline{\text{CRPS}}$. Using Eqs. (A8) and (A9), it is not difficult to see that for $1 < i < N$,

$$\overline{g}_i(\overline{o_i^2} - \overline{o}_i^2) = \int_{-\infty}^{\infty} g_i(x)[o_i(x) - \overline{o}_i]^2\, dx, \quad (49)$$

from which it follows that these terms in D are positive definite. Only when $o_i(x)$ does not depend on x, they are zero. Therefore this part of ⟨Reli⟩ is stricter than the corresponding part of $\overline{\text{Reli}}$, because ⟨Reli⟩ insists on a perfect reliability for all possible events, while $\overline{\text{Reli}}$ concentrates on the more overall reliability of the system. For the outliers the integral over $g_i(x)$ is infinite, and therefore Eq. (49) is not valid for $i = 0, N$.

The quantities ⟨Reli⟩ and ⟨Resol⟩, as well as D, involve integrals over $g_i o_i^2$. These integrals are, in contrast to integrals over g_i and $g_i o_i$ (see the appendix), difficult to perform analytically. Therefore, in practice, it is a tedious procedure to evaluate ⟨Reli⟩ and ⟨Resol⟩. Besides, ⟨Reli⟩ does not have the same clear relation to the rank histogram as $\overline{\text{Reli}}$ has. For these reasons, decomposition (39) is to be preferred above decomposition (43).

5. Decomposition for the EPS at ECMWF

The ideas developed in the previous sections will be illustrated by the performance of the ensemble prediction system running at ECMWF. This ensemble forecasting system (see Molteni et al. 1996; Buizza and Palmer 1998; Buizza et al. 1999) consists of 50 perturbed forecasts plus a control forecast integrated with the ECMWF $T_L159L31$ primative equation (PE) model up to day 10. For seven cases in the summer of 1999, the $\overline{\text{CRPS}}$ of total precipitation has been evaluated for the European area (30.0°–72.5°N, 22.5°W–42.5°E) using a grid spacing of 2.5° in both the latitudinal and the longitudinal direction (486 grid points). The weights w_k [see Eq. (5)] were chosen to be proportional to the cosine of latitude. As verifying analysis the precipitation accumulated within the first 24 h of the ECMWF operational $T_L319L50$ PE model forecasts was taken [for a discussion on this choice, see the appendix of Buizza et al. (1999)].

Table 1 shows the $\overline{\text{CRPS}}$ and its decomposition (39) between forecast day 2 and 10. It is seen that the continuous ranked probability score gradually grows (although not monotonously) from 0.98 mm (24 h^{-1}) at day 2 to 1.31 mm (24 h^{-1}) at day 10, expressing a decreasing predictability as a function of forecast time. The reliability only forms a small part of the $\overline{\text{CRPS}}$. There is a trend that it decreases. Apparently reliability is less optimal for the first forecast days. The uncertainty shown in Table 1 is based on sample distributions in which no corrections for anomalies or location were applied. It fluctuates somewhat from day to day, expressing differences in the sample distributions (each consisting of 3402 verifying analyses) obtained for the various forecast days. The resolution strongly decreases from 0.322 mm $(24\ h)^{-1}$ at day 2, to 0.086 mm $(24\ h)^{-1}$ at day 10. Therefore, the first days, EPS significantly outperforms a forecast based on climatology, while for longer forecast periods there is an onset of convergence to climatology.

TABLE 1. Continuous ranked probability score and its decomposition into reliability, resolution, and uncertainty [see Eq. (39)] of total precipitation accumulated in the 24 h prior to the displayed forecast day for seven cases in the summer of 1999 for the ECMWF ensemble prediction system. The dimension of these quantities is mm $(24h)^{-1}$.

Day	$\overline{\text{CRPS}}$	$\overline{\text{Reli}}$	$\overline{\text{Resol}}$	$\overline{U}$
2	0.98	0.060	0.322	1.24
3	1.13	0.031	0.317	1.42
4	1.04	0.027	0.229	1.25
5	1.17	0.024	0.209	1.36
6	1.17	0.020	0.179	1.33
7	1.18	0.026	0.110	1.26
8	1.29	0.015	0.123	1.40
9	1.21	0.019	0.073	1.26
10	1.31	0.016	0.086	1.38

In order to be able to understand these trends in more detail, in Figs. 4, 5, and 6 a graphical representation of the reliability, uncertainty, and resolution is displayed for forecast days 3, 6, and 9, respectively. In the top panels the observed frequencies o_i as defined in Eqs. (31) and (33) are plotted as a function of the fraction of members p_i. Any deviation from the diagonal will contribute to the reliability Reli defined in Eq. (36). The lower panels of Figs. 4–6 show (staircase curve) the accumulation of the average bin widths g_i, as defined in Eq. (30). The leftmost and rightmost bins show the average magnitude g_0 and g_N, respectively, of the outliers [see Eq. (33)]. The width of this curve determines the potential CRPS, because CRPS_{pot} can be seen as the integral over this curve with the weight function $o_i(1 - o_i)$. The narrower the staircase curve, the smaller the region for which the weight function is significantly different from zero, and as a result, the smaller CRPS_{pot} is. In addition, the lower panels show the cumulative distribution (''smooth'' curve) of the sample climatology, as defined in Eq. (9). As is illustrated by Fig. 1, for example, the uncertainty $\overline{U}$ is proportional to the width of P_{sam}. In addition (see discussion at the end of

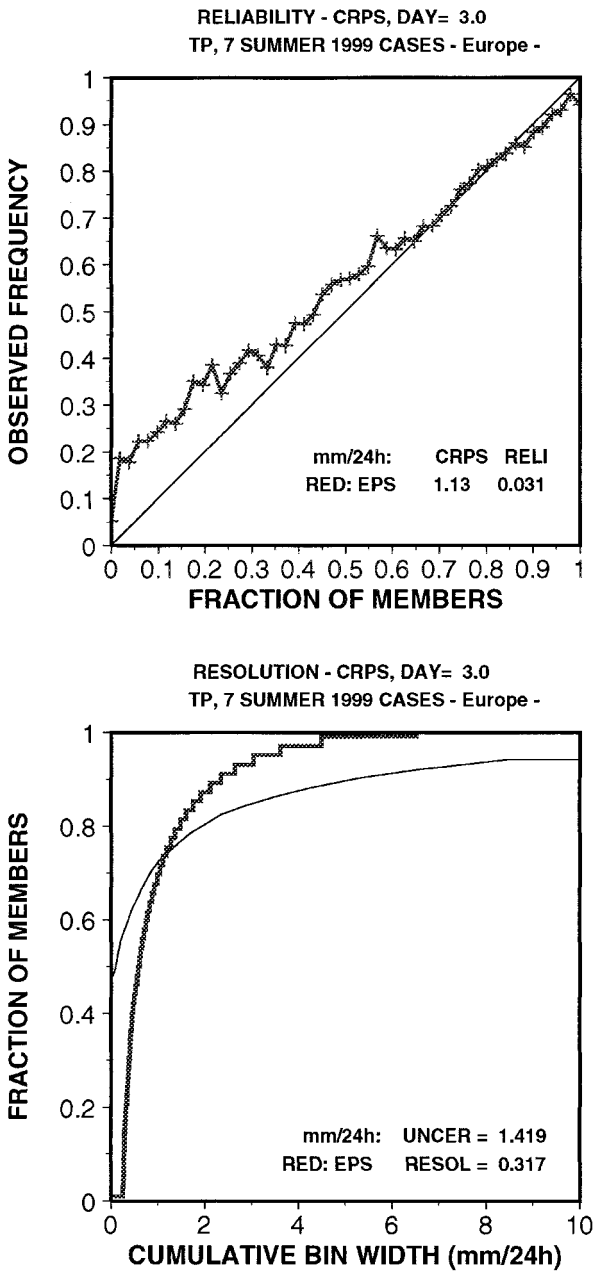

FIG. 4. Decomposition of continuous ranked probability score for total precipitation accumulated between day 2 and day 3 for seven summer cases in 1999 and averaged over the European area.

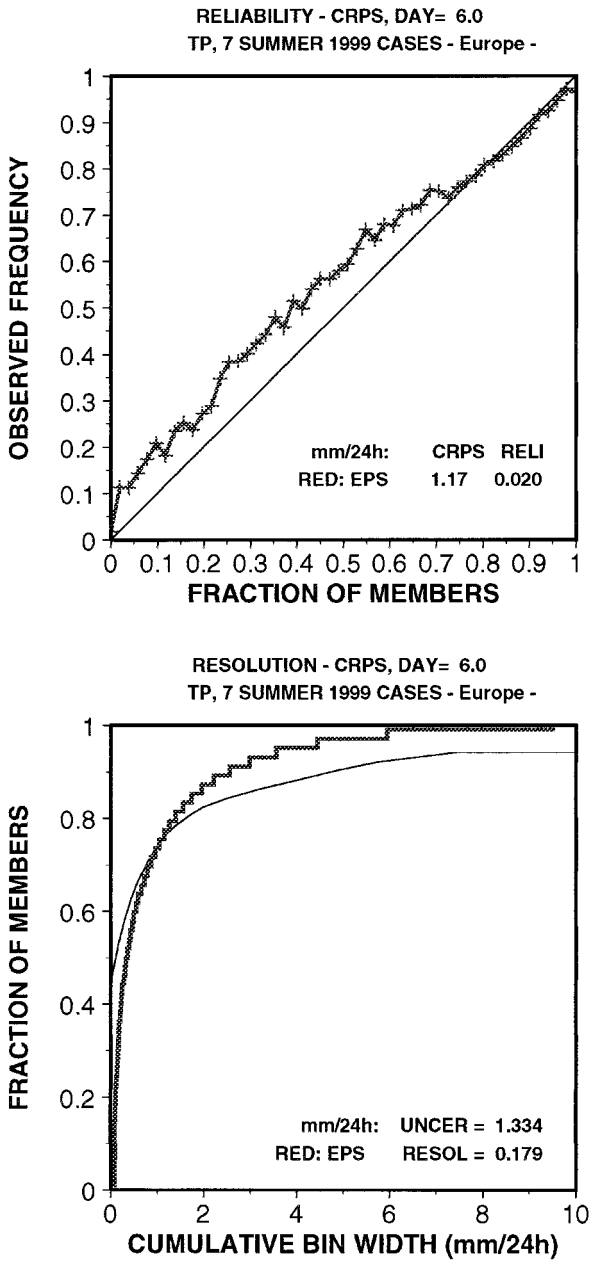

FIG. 5. The same as Fig. 4 but for day 6.

section 4c) it can be seen as the expected CRPS of a forecast system based on the climatology of the sample. The difference in widths between the staircase curve and the cumulative distribution, therefore, is a measure for the resolution (38).

The discrepancy from perfect reliability for the first forecast days is mainly due to the lower bins of the ensembles, as can be seen in Fig. 4 for day 3. The frequency that the verifying analysis is found to be below these bins is too high. It occurs too often that all members predict at least some precipitation, while it remained dry (based on climatology as can be seen from P_{sam} in the lower panel of Fig. 4, the probability that it remains dry is about 50%). However, for these cases, the amount of precipitation of the member with the smallest amount of rain is on average quite small (around 0.3 mm; see g_0 in bottom panel of Fig. 4). Therefore this mild overestimation of precipitation will not contribute very strongly to Reli. Such a delicate analysis would not be visible from the rank histogram. It would only show a too high frequency of outliers.

The high resolution of the EPS for day 3 can clearly

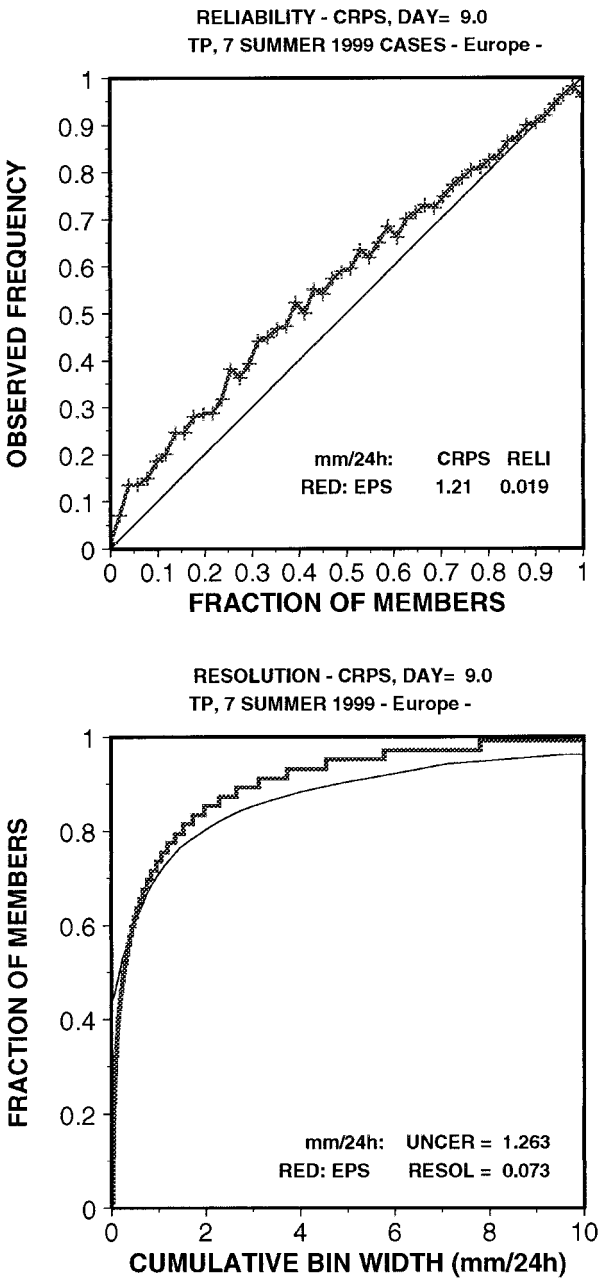

FIG. 6. The same as Fig. 4 but for day 9.

be seen from the bottom panel of Fig. 4. The average bin widths of the ensemble, including the outliers, is, compared to P_{sam}, considerably small. The climatological distribution has a large tail for high amounts of precipitation. Apparently, for such cases, the EPS was capable of generating sharp ensembles with fair amounts of precipitation. This is the reason why the size of the outlier g_N is reasonably small. The reduction of resolution with increasing forecast time is well illustrated by comparing the lower panels of Figs. 4–6. At day 3, the ensemble is much sharper than P_{sam}, while at day 9, it is quite similar to the sample distribution, leaving only a low value of resolution.

6. Concluding remarks

In this paper it was shown how for an ensemble prediction system, the continuous ranked probability score can be decomposed into three parts. This decomposition is very similar to that of the Brier score. The first part, reliability, is closely related to the rank histogram. An important difference, however, is that the reliability of the CRPS is sensitive to the width of the ensemble bins, while the rank histogram gives each forecast the same weight. The reliability should be zero for an ensemble system with the correct statistical properties. The second part, uncertainty, is the best achievable value of the continuous ranked probability score, in case only climatological information is available. It was discussed that in contrast to the uncertainty of the Brier score, the value of uncertainty depends on the degree of sophistication. The third term, the resolution, expresses the superiority of a forecast system with respect to a forecast system based on climatology. The uncertainty/reliability part was found to be both sensitive to the average spread within the ensemble, and to the behavior of the outliers. It was shown that the proposed decomposition is not equal to the integral over the decomposition of the Brier score.

It was illustrated how the reliability part could be presented in a graphical way. In addition, it was shown how the resolution part of the CRPS can be visualized by looking at the difference between the sample climate distribution and the accumulated average bin widths of the ensemble system. As an example the decomposition for total precipitation for seven summer cases in 1999 of the ECMWF ensemble prediction system was considered.

In this paper attention was focused on ensemble forecasts, for which the allowable set of forecasted probabilities is finite. However, in general, a forecast system could issue any probability between 0 and 1. Such systems could be regarded as the limit of $N \rightarrow \infty$, of an N-member ensemble, in which the ith member is positioned at the location where the cumulative distribution has the value $P(x_i) = p_i = i/N$. Therefore, the decomposition of the CRPS, given in section 4, can be extended to any continuous forecast system. As a result, the summations over probabilities p_i in the definitions of reliability, resolution, and uncertainty will transform into integrals (from 0 to 1) over probabilities. In order to evaluate such integrals for continuous systems, it is more sensible to discretize the allowable set of probabilities, than to discretize the variable x. Therefore, in practice, the evaluation of the CRPS and its decomposition for continuous forecast systems exactly reduces to the method proposed in section 4.

The continuous ranked probability score is a verification tool that is sensitive to the overall (with respect

to a certain parameter) performance of a forecast system. By using the decomposition proposed in this paper, it was argued how for an ensemble prediction system, a detailed picture of this overall behavior can be obtained.

Acknowledgments. The author would like to thank François Lalaurette at ECMWF and Kees Kok at KNMI for stimulating discussions.

APPENDIX

Some Technical Details

In this appendix the relation between the various terms of the Brier score defined in Eq. (40) and the terms of the continuous ranked probability score given in Eq. (39) will be determined.

Let the function $I(x, a, b)$ be defined by

$$I(x, a, b) = \begin{cases} 1 & \text{if } a \le x \le b \\ 0 & \text{elsewhere,} \end{cases} \tag{A1}$$

then the densities $g_i(x)$ and frequencies $o_i(x)$ introduced in Eq. (40) can be written as

$$g_i(x) = \sum_k w_k I(x, x_i^k, x_{i+1}^k) \quad \text{and} \tag{A2}$$

$$g_i(x)o_i(x) = \sum_k w_k H(x - x_a^k) I(x, x_i^k, x_{i+1}^k). \tag{A3}$$

By using the property

$$\sum_{i=0}^{N} I(x, x_i^k, x_{i+1}^k) = 1, \tag{A4}$$

it is evident that

$$\sum_{i=0}^{N} g_i(x) = \sum_k w_k = 1, \tag{A5}$$

$$o(x) \equiv \sum_{i=0}^{N} g_i(x)o_i(x) = \sum_k w_k H(x - x_a^k) = P_{\text{sam}}(x). \tag{A6}$$

So the g_i are normalized and $o(x)$ is related to the cumulative distribution of the sample. From the expressions

$$\int_{-\infty}^{\infty} I(x, x_i^k, x_{i+1}^k)\, dx = x_{i+1}^k - x_i^k$$

$$\int_{-\infty}^{\infty} H(x - x_a^k) I(x, x_i^k, x_{i+1}^k)\, dx = \beta_i^k, \tag{A7}$$

$1 < i < N$, where β_i^k is defined by Eq. (26), it follows that

$$\int_{-\infty}^{\infty} g_i(x)\, dx = \sum_k w_k (x_{i+1}^k - x_i^k) = \overline{g}_i \tag{A8}$$

$$\int_{-\infty}^{\infty} g_i(x)o_i(x)\, dx = \sum_k w_k \beta_i^k = \overline{g}_i \overline{o}_i, \tag{A9}$$

where $\overline{g}_i$ and $\overline{g}_i\overline{o}_i$ are defined in Eq. (32).

For the outliers one has to keep in mind that $g_0(-\infty) = 1$ and $g_N(\infty) = 1$, and therefore Eq. (A8) would become infinite. With the help of definitions (27) and (33), the reader may verify that

$$\int_{-\infty}^{\infty} g_0(x)o_0(x)\, dx = \overline{g}_0 \overline{o}_0 \tag{A10}$$

$$\int_{-\infty}^{\infty} g_N(x)[1 - o_N(x)]\, dx = \overline{g}_N(1 - \overline{o}_N). \tag{A11}$$

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9,** 1518–1530.

Bouttier, F., 1994: Sur la prévision de la qualité des prévisions météorologiques. Ph.D. thesis, Université Paul Sabatier, Toulouse, France, 240 pp. [Available from Libray, Université Paul Sabatier, route de Narbonne, Toulouse, France.]

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78,** 1–3.

Brown, T. A., 1974: Admissible scoring systems for continuous distributions. Manuscript P-5235, The Rand Corporation, Santa Monica, CA, 22 pp. [Available from The Rand Corporation, 1700 Main St., Santa Monica, CA 90407-2138.]

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126,** 2503–2518.

——, A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **14,** 168–189.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8,** 985–987.

Hamill, T., and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125,** 1312–1327.

Katz, R. W., and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts.* Cambridge University Press, 222 pp.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30,** 291–303.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.*, **22,** 1087–1095.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart J. Roy. Meteor. Soc.*, **122,** 73–119.

Murphy, A. H., 1969: On the ''ranked probability score.'' *J. Appl. Meteor.*, **8,** 988–989.

——, 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10,** 155–156.

——, 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12,** 595–600.

——, and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117,** 572–581.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, 1989: *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, 818 pp.

Richardson, D., 1998: Obtaining economic value from the EPS. *ECMWF Newsletter,* Vol. 80, 8–12.

——, 2000: Skill and relative economic value of the ECMWF Ensemble Prediction System. *Quart J. Roy. Meteor. Soc.*, **126,** 649–668.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. Atmospheric Environment Service Research Rep. 89-5, 114 pp. [Available from Forecast Research Division, 4905 Dufferin St., Downsview, ON M3H 5T4, Canada.]

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

Unger, D. A., 1985: A method to estimate the continuous ranked probability score. Preprints, *Ninth Conf. on Probability and Statistics in Atmospheric Sciences*, Virginia Beach, VA, Amer. Meteor. Soc., 206–213.

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