METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **15**: 19–24 (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/met.45

RMetS
ROYAL METEOROLOGICAL SOCIETY

On the effect of ensemble size on the discrete and continuous ranked probability scores

Christopher A. T. Ferro,[a]* David S. Richardson[b] and Andreas P. Weigel[c]
[a] *School of Engineering, Computing and Mathematics, University of Exeter, UK*
[b] *European Centre for Medium-Range Weather Forecasts, Reading, UK*
[c] *Federal Office of Meteorology and Climatology (MeteoSwiss), Zurich, Switzerland*

ABSTRACT: Four recent papers have investigated the effects of ensemble size on the Brier score (BS) and discrete ranked probability score (RPS) attained by ensemble-based probabilistic forecasts. The connections between these papers are described and their results are generalized. In particular, expressions, explanations and estimators for the expected effect of ensemble size on the RPS and continuous ranked probability score (CRPS) are obtained. Copyright © 2008 Royal Meteorological Society

KEY WORDS Brier score; forecast verification; reliability; sharpness; skill scores

Received 16 August 2007; Revised 23 October 2007; Accepted 22 November 2007

1. Introduction

Ensemble predictions are often used to construct probabilistic forecasts, such as the proportion of ensemble members forecasting the occurrence of an event. Ensemble size affects the quality of such forecasts, and this fact raises several questions for forecast developers, providers and users: what is the expected effect on verification scores of changing ensemble size; what is the explanation for the effect and how can it be estimated; which reference forecasts should be used for skill scores; and how should forecasts be compared between systems with different ensemble sizes? We answer these questions by deriving expressions, explanations and estimators for the expected effect of ensemble size on verification scores. With these tools, developers of forecasting systems can estimate the impact of increasing or decreasing their ensemble size before implementing the changes in their operational system. The potential forecast quality that would be reached with an infinite ensemble size can also be estimated. Furthermore, differences in the qualities of forecasting systems with unequal ensemble sizes can be diagnosed. For example, the scores that would be achieved by the systems were their ensemble sizes equal can be estimated and compared, perhaps revealing that the apparent superiority of one system is due only to its larger ensemble size rather than any fundamental difference in the forecasting model. In a similar way, reference forecasts that account for ensemble size can be used in skill scores to provide a fair baseline.

These topics have been addressed recently in four papers: Richardson (2001), Müller *et al.* (2005), Weigel *et al.* (2007), and Ferro (2007), which we shall refer to as R01, M05, W07, and F07. These papers consider different scores [the Brier score (BS) and discrete ranked probability score (RPS)], make different assumptions, use different notations, and propose different interpretations of their results. Nevertheless, there are strong connections between the papers. We clarify these connections in Section 2 and discuss their interpretation in Section 3 before presenting a generalization in Section 4 that includes results from all four papers as special cases. We present similar results for the continuous ranked probability score (CRPS) in Section 5. We adopt a notation similar to F07; a comparison of notations from the four papers is contained in Table I.

2. Connections

Suppose that we observe which one of K possible categories occurs at each time $t = 1, \ldots, n$. Let $I_{t,k} = 1$ if category k is observed at time t, and $I_{t,k} = 0$ otherwise. Suppose also that we issue probabilistic forecasts $\hat{Q}_{t,k}$ equal to the proportion of m ensemble members forecasting category k at time t. A popular verification score for such forecasts due to Brier (1950) is

$$\hat{B}_{K,m} = \frac{1}{n} \sum_{t=1}^{n} \sum_{k=1}^{K} (\hat{Q}_{t,k} - I_{t,k})^2 \qquad (1)$$

This is known as the RPS (Epstein, 1969; Murphy, 1971) if the categories are nested and category K is certain to occur. For example, if the categories correspond to

* Correspondence to: Christopher A. T. Ferro, School of Engineering, Computing and Mathematics, University of Exeter, Harrison Building, North Park Road, Exeter, EX4 4QF, UK.
E-mail: c.a.t.ferro@exeter.ac.uk

Table I. A comparison of some notation and assumptions used by F07, R01, M05, and W07. IID means 'independent and identically distributed'.

	F07	R01	M05/W07
Observation	I	E	o
Probabilistic forecast	$\hat{Q}$	P_f	y
Ensemble probabilities	Q	P_∞	p
Climatology	$E(P)$	$\overline{o}$	p
Brier score	$\hat{B}_{M,n}$	–	$\langle \mathrm{BS} \rangle$
Expected Brier score	B_M	b_M	–
Ensemble members	Exchangeable	IID	IID
Ensemble probabilities	–	Reliable	Climatological

temperature not exceeding thresholds $u_1 < \cdots < u_K = \infty$ then $\hat{Q}_{t,1} \leq \cdots \leq \hat{Q}_{t,K} = 1$ can be thought of as cumulative probabilities. When $K = 2$, the RPS is also known as the BS, which we write as

$$\hat{B}_m = \frac{1}{n} \sum_{t=1}^{n} (\hat{Q}_t - I_t)^2 \tag{2}$$

where $\hat{Q}_t = \hat{Q}_{t,1}$ and $I_t = I_{t,1}$. We denote by p_k the long-run frequency with which category k is observed to occur, and assume stationarity throughout.

M05 and W07 consider the ranked probability skill score, $1 - \mathrm{RPS}/\mathrm{RPS}_{\mathrm{ref}}$, in which the RPS achieved by the forecasts is compared with the expected RPS ($\mathrm{RPS}_{\mathrm{ref}}$) that would be obtained were a reference forecast issued for the available observations. A positive score indicates forecast skill. The climatology $(p_1, \ldots, p_K)$ is a typical reference forecast and is equivalent to a random ensemble forecast with infinite ensemble size in which each member independently forecasts category k with probability p_k. M05 and W07 show that this reference forecast attains a better expected RPS than the equivalent forecast with only m members, and therefore that a better-than-random, m-member ensemble forecast can score negative skill if climatology is used as the reference. M05 recommend the m-member random forecast as a fairer reference and obtain the corresponding value of $\mathrm{RPS}_{\mathrm{ref}}$ by simulation: W07 derive an analytical expression.

In the case of the BS, when $K = 2$, the expression derived by W07 becomes

$$B_m = B_\infty + \frac{p(1-p)}{m} \tag{3}$$

where $p = p_1$ and $B_m = E(\hat{B}_m) \to B_\infty$ as $m \to \infty$. Equation (3) gives the expected BS for m-member ensemble forecasts in which each member independently forecasts category 1 with climatological probability p. This is sufficient for defining fair reference values in skill scores, but more general expressions for B_m with wider implications are obtained by R01 and F07.

R01 also assumes that ensemble members are independent and identically distributed but places no restriction on the probabilities with which they forecast the two categories. In this case,

$$B_m = B_\infty + \frac{E\{Q(1-Q)\}}{m} \tag{4}$$

where Q represents the probability with which any ensemble member will forecast category 1 at a randomly chosen time. Equation (3) is recovered when $Q = p$ at all times.

F07 relaxes the assumption of independent ensemble members to one of exchangeability. Exchangeability says that the labelling of the ensemble members has no impact on our prior beliefs about the values they will assume: their joint distribution is invariant to relabelling the members. Unlike independence, exchangeability permits dependence between members; for example, members can be correlated with one another, as long as all correlations are equal. Exchangeability is justifiable for most initial-condition ensembles, with possible exceptions arising when differences among initial states are selected systematically and are expected to persist over the integration time. Without exchangeability, generic ensemble-size effects such as Equation (4) do not exist because we know nothing about the statistical properties of potential new members, and therefore nothing about how they might affect verification scores.

Assuming exchangeability, F07 shows that Equation (4) generalizes to

$$B_m = B_\infty + \frac{E(Q-R)}{m} \tag{5}$$

where R is the probability with which any two distinct ensemble members at a randomly chosen time will both forecast category 1. Since $Q \geq R$, the expected BS decreases as the ensemble size increases. Independent members implies $R = Q^2$ and recovers Equation (4). We return to the interpretation of Equation (5) in Section 3.

Ensemble probabilities Q are perfectly reliable if, for all q, category 1 is observed with frequency q on those occasions when $Q = q$. Substituting Equation (10) of R01 into Equation (15) of R01 reveals that, under

DOI: 10.1002/met

the assumptions of perfect reliability and independent ensemble members,

$$B_m = \frac{m+1}{m} B_\infty = \frac{M(m+1)}{m(M+1)} B_M \tag{6}$$

for any M. This also holds under the alternative assumption of a 'perfect' ensemble in which not only the ensemble members but also the observation are exchangeable at each time. Given an original ensemble size m, the expected value of the BS that would be obtained for any ensemble size $M \neq m$ can then be estimated without bias by

$$\frac{m(M+1)}{M(m+1)} \hat{B}_m \tag{7}$$

F07 also obtains an unbiased estimator for B_M when the original ensemble size is m, but assumes that only the ensemble members are exchangeable. This more widely applicable estimator is

$$\hat{B}_m - \frac{M-m}{M(m-1)n} \sum_{t=1}^{n} \hat{Q}_t (1 - \hat{Q}_t) \tag{8}$$

This estimator is undefined when $m = 1$, in which case F07 shows that an unbiased estimator for B_M does not exist and the stricter assumptions behind Equation (7) must be employed.

Plotting such estimators for B_M against M shows how the BS is expected to change with ensemble size. Consider also comparing two forecasting systems with different ensemble sizes. As well as comparing the raw BSs, we can use our estimators to compare the BSs after adjusting for the difference in ensemble size. For example, F07 describes how to construct confidence intervals for the difference between the expected BSs that would be obtained were both ensemble sizes equal to M. Such comparisons can help to identify whether or not the raw superiority of one system is due only to its larger ensemble size. Data examples illustrating these and other applications can be found in R01, M05, W07 and F07.

3. Interpretation

Increasing ensemble size merely improves the precision of the forecasts $\hat{Q}_t$ as estimators for the probabilities that would be issued were the ensemble size infinite. Our most general expression (Equation (5)) for B_m shows that the consequent expected improvement in the BS is independent of the observations, as are the adjustment terms in the unbiased estimators (Equations (7) and (8)) for B_M. F07 noted that the latter estimator can be written as

$$\hat{B}_m - \frac{M-m}{M(m-1)} \left(\frac{1}{4} - S \right) \tag{9}$$

where

$$S = \frac{1}{n} \sum_{t=1}^{n} \left(\hat{Q}_t - \frac{1}{2} \right)^2 \tag{10}$$

The effect of changing the ensemble size from m to M therefore depends only on the values of m and M, and a measure of forecast sharpness, S. In particular, the improvement in BS from increasing ensemble size decreases as either m or the forecast sharpness increases.

The BS is often decomposed into reliability, resolution and uncertainty terms:

$$\hat{B}_m = \frac{1}{n} \sum_{k=0}^{m} N_k \left(\frac{k}{m} - \overline{I}_k \right)^2 - \frac{1}{n} \sum_{k=0}^{m} N_k (\overline{I}_k - \overline{I})^2 + \overline{I}(1 - \overline{I}) \tag{11}$$

where $\overline{I}_k$ is the mean of the N_k observations coincident with forecasts equal to k/m, and $\overline{I}$ is the mean of all n observations (Murphy, 1973). Changes in ensemble size can affect both the reliability and resolution terms. For example, we show in Appendix A that, when ensemble members and observations are all independent and Q always equals p, the expected reliability and resolution terms are

$$E(\mathrm{REL}) = \frac{p(1-p)}{n} E(N) + \frac{p(1-p)}{m} \tag{12}$$

$$E(\mathrm{RES}) = \frac{p(1-p)}{n} E(N) - \frac{p(1-p)}{n} \tag{13}$$

where N is the number of possible forecasts k/m that are issued at least once. If $n \gg m$ then $E(\mathrm{REL}) \approx p(1-p)/m$ and $E(\mathrm{RES}) \approx 0$, and most of the ensemble-size effect is restricted to the reliability term. This fits with R01 and W07 who show that a finite-sized ensemble is intrinsically unreliable. In a more general perfect reliability setting, R01 shows that the increase in variance of forecasts relative to Q brought about by smaller ensemble sizes results in over-confidence: forecasts less than p under-forecast and forecasts greater than p over-forecast, inflating the reliability term and giving a clockwise tilt to reliability diagrams. Smaller ensemble sizes also reduce the resolution term because fewer distinct forecasts are possible.

4. Generalization

We have seen that F07 gives the most general results (Equations (5) and (8)) for the BS, but W07 considered the more general RPS (Equation (1)). In this section, we extend the results of F07 to the RPS.

Since $\hat{B}_{K,m}$ is equal to the sum of K BSs, one for each of the K categories, we immediately obtain

$$B_{K,m} = B_{K,\infty} + \frac{1}{m} \sum_{k=1}^{K} E(Q_k - R_k) \tag{14}$$

as a generalization of Equation (5), where $B_{K,m} = E(\hat{B}_{K,m}) \to B_{K,\infty}$ as $m \to \infty$, and Q_k and R_k are the counterparts of Q and R for category k. Since $Q_k \geq R_k$

DOI: 10.1002/met

for all k, the expected RPS decreases as the ensemble size increases. When ensemble members are independent so that $R_k = Q_k^2$, and when $Q_k = \sum_{i=1}^{k} p_i$, the climatology forecast for category k when categories are nested, we recover the expression derived by W07 for the RPS:

$$B_{K,m} = B_{K,\infty} + \frac{1}{m}\left\{\sum_{k=1}^{K}\sum_{i=1}^{k} p_i(1-p_i) - 2\sum_{k=2}^{K}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} p_i p_j\right\} \quad (15)$$

The unbiased estimator (Equation (8)) for B_M also generalizes to

$$\hat{B}_{K,m} - \frac{M-m}{M(m-1)n}\sum_{t=1}^{n}\sum_{k=1}^{K}\hat{Q}_{t,k}(1-\hat{Q}_{t,k}) \quad (16)$$

which can be used as before to estimate how the RPS will change with ensemble size, and to compare forecasting systems with different ensemble sizes.

As with the estimator for B_M given by Equation (8), the foregoing estimator for $B_{K,M}$ is undefined when $m = 1$. Mirroring the derivation of Equation (7) for the BS, however, the estimator

$$\frac{m(M+1)}{M(m+1)}\hat{B}_{K,m} \quad (17)$$

is unbiased for $B_{K,M}$ under the stricter assumptions that either the Q_k are perfectly reliable for all k and the ensemble members are independent, or the ensemble is perfect.

Let us also consider the multi-category BS, which has the same definition (Equation (1)) as the RPS, but requires the K categories to be mutually exclusive and exhaustive rather than nested. The foregoing results of this section for the RPS apply equally to the multi-category BS. However, the unbiased estimator (Equation (16)) for the multi-category BS also equals

$$\hat{B}_{K,m} - \frac{M-m}{M(m-1)}\left(\frac{K-1}{K} - S_K\right) \quad (18)$$

where

$$S_K = \frac{1}{n}\sum_{t=1}^{n}\sum_{k=1}^{K}\left(\hat{Q}_{t,k} - \frac{1}{K}\right)^2 \quad (19)$$

is the sum of the sample variances of forecasts for category k around the uniform forecasts $1/K$. This generalization of the sharpness measure (Equation (10)) for the BS attains its minimum value of zero when $\hat{Q}_{t,k} = 1/K$ for all t and k, and attains its maximum value of $(K-1)/K$ when $\hat{Q}_{t,k} = 1$ for some k at each time t.

5. Continuous ranked probability score

Following a suggestion from an anonymous referee, we now perform a similar analysis of the effect of ensemble size on the CRPS (e.g. Hersbach, 2000). This score extends the RPS from a sum over K categories to an integral over a continuum of categories, as follows.

Let $I_t(u) = 1$ if the observation at time t fails to exceed a threshold u, and $I_t(u) = 0$ otherwise. Let also $\hat{Q}_t(u)$ be the proportion of m ensemble members failing to exceed u at time t. The CRPS is defined as

$$\hat{C}_m = \frac{1}{n}\sum_{t=1}^{n}\int_{-\infty}^{\infty}\{\hat{Q}_t(u) - I_t(u)\}^2 \mathrm{d}u \quad (20)$$

Assuming stationarity and exchangeable ensemble members, we show in Appendix B that the expected CRPS can be written as

$$C_m = C_\infty + \frac{E|X_1 - X_2|}{2m} \quad (21)$$

where $C_m = E(\hat{C}_m) \to C_\infty$ as $m \to \infty$ and $\{X_1, X_2\}$ is any pair of distinct ensemble members at a randomly chosen time. Since $|X_1 - X_2| \geq 0$, the expected CRPS decreases as the ensemble size increases. Replacing the ensemble members in Equation (21) with independent variables following the climatological distribution yields a reference value for skill scores.

We can also write

$$C_M = C_m - \frac{M-m}{2Mm}E|X_1 - X_2| \quad (22)$$

where C_M is the expected CRPS that would be obtained for an ensemble of size M. Therefore, an unbiased estimator for C_M based on ensembles of size m is

$$\hat{C}_m - \frac{M-m}{2Mmn}\sum_{t=1}^{n}\Delta_t \quad (23)$$

where

$$\Delta_t = \frac{1}{m(m-1)}\sum_{i\neq j}|X_{t,i} - X_{t,j}| \quad (24)$$

is Gini's mean difference of the ensemble members $\{X_{t,1}, \ldots, X_{t,m}\}$ at time t. See David (1998) for a history of the mean difference.

As with the estimator for B_M given by Equation (8), the foregoing estimator for C_M is undefined when $m = 1$. As for the BS and RPS, however, an unbiased estimator for C_M does exist under the stricter, perfect ensemble assumption. In that case,

$$C_m = \frac{m+1}{2m}E|X_1 - X_2| \quad (25)$$

which leads to

$$C_M = \frac{m(M+1)}{M(m+1)}C_m \quad (26)$$

DOI: 10.1002/met

and the unbiased estimator

$$\frac{m(M+1)}{M(m+1)}\hat{C}_m \tag{27}$$

As discussed in Section 3, increasing ensemble size improves the precision of the forecasts $\hat{Q}_t(u)$. Equations (21) and (23) show that the consequent expected improvement in the CRPS is independent of the observations and proportional to a measure of average ensemble spread, with smaller improvement when the spread is low. Low spread arises when ensemble members are strongly dependent, so that new members add little independent information, or when the distribution from which the ensemble forms a sample is sharp, that is when confidence is high.

Our analysis of the CRPS also sheds light on our earlier interpretation of the effect of ensemble size on the BS. The mean difference between variables indicating whether or not two distinct members forecast an event A is

$$\begin{aligned} &E|\mathrm{I}(X_1 \in A) - \mathrm{I}(X_2 \in A)| \\ &\quad = \Pr(X_1 \in A, X_2 \notin A) + \Pr(X_1 \notin A, X_2 \in A) \\ &\quad = 2\{\Pr(X_1 \in A) - \Pr(X_1 \in A, X_2 \in A)\} \\ &\quad = 2(Q - R) \end{aligned} \tag{28}$$

Therefore, the effect of ensemble size on the BS given by Equation (5) has the same form as in Equation (21) for the CRPS, where X_1 and X_2 are replaced by indicator variables for the event. The dependence on sharpness of the ensemble-size effect for the BS can thus be restated in terms of the average spread of the ensemble of indicator variables.

6. Conclusion

The expected BS, RPS, CRPS and multi-category BS achieved by probabilistic forecasts constructed as proportions of exchangeable ensemble members decrease as ensemble size increases. This effect can be accounted for when choosing reference forecasts for skill scores. The magnitude of the effects on scores of changing the ensemble size depends on the original ensemble size and measures of average ensemble spread or forecast sharpness. Unbiased estimators for these effects can be used to estimate how scores are expected to change with ensemble size, and to compare forecasting systems with different ensemble sizes.

Acknowledgements

We thank the WWRP/WGNE Joint Working Group on Verification for organizing the 3rd International Workshop on Verification Methods (held at the European Centre for Medium-Range Weather Forecasts in Reading, UK, from 29 January to 2 February 2007) where this work was conceived. CF was funded by the Natural Environment Research Council's National Centre for Atmospheric Science. AW was funded by the Swiss National Centre of Competence in Research Climate and ENSEMBLES EC Contract GOCE-CT-2003-505539.

Appendix A. Reliability and resolution.

We derive Equations (12) and (13) for the expected reliability and resolution when the observations and ensemble members are all mutually independent Bernoulli random variables with event probability equal to p. This implies that the vector $(N_0, \ldots, N_m)$ is multinomial with total n and probabilities

$$\binom{m}{k} p^k (1-p)^{m-k}$$

for $k = 0, \ldots, m$. We find the expected reliability in two steps: first fixing the N_k and taking the expectation with respect to the I_t, then taking the expectation with respect to the N_k. Since

$$\begin{aligned} \overline{I}_k^2 = \frac{1}{N_k^2}\Bigg\{ &\sum_{t=1}^{n} \mathrm{I}(\hat{Q}_t = k/m) I_t \\ &+ \sum_{t=1}^{n}\sum_{s\neq t} \mathrm{I}(\hat{Q}_s = \hat{Q}_t = k/m) I_s I_t \Bigg\} \end{aligned}$$

where $\mathrm{I}(A) = 1$ if A is true and $\mathrm{I}(A) = 0$ otherwise, the conditional expectation of the reliability given the N_k is

$$\begin{aligned} &\frac{1}{n}\sum_{k=0}^{m} \mathrm{I}(N_k > 0) N_k \left[\frac{k^2}{m^2} - \frac{2k}{m} p \right. \\ &\qquad \left. + \frac{1}{N_k^2}\{N_k p + N_k(N_k-1)p^2\} \right] \\ &\quad = \frac{1}{n}\sum_{k=0}^{m} \mathrm{I}(N_k > 0) N_k \left(\frac{k}{m} - p \right)^2 + \frac{p(1-p)}{n} N \end{aligned}$$

Taking the expectation with respect to the N_k, use

$$E\{\mathrm{I}(N_k > 0) N_k\} = E(N_k) = n \binom{m}{k} p^k (1-p)^{m-k}$$

to obtain the expected reliability

$$\begin{aligned} &\frac{1}{m^2}\sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} (k - mp)^2 \\ &\qquad + \frac{p(1-p)}{n} E(N) \end{aligned}$$

The sum in this expression is the variance of a binomial random variable, and equals $mp(1-p)$. The expected uncertainty term in the decomposition of the BS equals $(n-1)p(1-p)/n$ and the expected BS equals

DOI: 10.1002/met

$(m+1)p(1-p)/m$; subtraction yields the expected resolution, $p(1-p)E(N)/n - p(1-p)/n$.

Appendix B. Expected CRPS.

We derive Equation (21) for the expected CRPS,

$$C_m = E\left[\frac{1}{n}\sum_{t=1}^{n}\int_{-\infty}^{\infty}\{\hat{Q}_t(u) - I_t(u)\}^2 \mathrm{d}u\right]$$

By stationarity, the expectation of the integral is the same for all t, so we can write

$$C_m = E\left[\int_{-\infty}^{\infty}\left\{\frac{1}{m}\sum_{i=1}^{m}\mathrm{I}(X_i \le u) - \mathrm{I}(Y \le u)\right\}^2 \mathrm{d}u\right]$$

where $\{X_1, \ldots, X_m\}$ and Y are the ensemble members and observation at an arbitrary time. Writing $\alpha = \min\{X_1, \ldots, X_m, Y\}$ and $\beta = \max\{X_1, \ldots, X_m, Y\}$, the integral becomes

$$\begin{aligned}
&\int_{\alpha}^{\beta}\left\{\frac{1}{m}\sum_{i}\mathrm{I}(X_i \le u) - \mathrm{I}(Y \le u)\right\}^2 \mathrm{d}u \\
&= \frac{1}{m^2}\sum_{i,j}\int_{\alpha}^{\beta}\mathrm{I}(X_i \le u, X_j \le u)\mathrm{d}u \\
&\quad -\frac{2}{m}\sum_{i}\int_{\alpha}^{\beta}\mathrm{I}(X_i \le u, Y \le u)\mathrm{d}u + \int_{\alpha}^{\beta}\mathrm{I}(Y \le u)\mathrm{d}u \\
&= \frac{1}{m^2}\sum_{i,j}(\beta - \max\{X_i, X_j\}) \\
&\quad -\frac{2}{m}\sum_{i}(\beta - \max\{X_i, Y\}) + (\beta - Y) \\
&= -\frac{1}{m^2}\sum_{i\neq j}\max\{X_i, X_j\} - \frac{1}{m^2}\sum_{i}X_i \\
&\quad + \frac{2}{m}\sum_{i}\max\{X_i, Y\} - Y
\end{aligned}$$

By exchangeability, we can then write the expectation as

$$\begin{aligned}
C_m &= -\frac{m-1}{m}E(\max\{X_1, X_2\}) - \frac{1}{m}E(X_1) \\
&\quad +2E(\max\{X_1, Y\}) - E(Y) \\
&= C_\infty + \frac{1}{m}E(\max\{X_1, X_2\} - X_1)
\end{aligned}$$

where $C_\infty = 2E(\max\{X_1, Y\}) - E(Y) - E(\max\{X_1, X_2\})$ is independent of m. Furthermore, since

$$\begin{aligned}
|X_1 - X_2| &= \max\{0, X_1 - X_2\} + \max\{0, X_2 - X_1\} \\
&= (\max\{X_1, X_2\} - X_2) + (\max\{X_1, X_2\} - X_1)
\end{aligned}$$

and similarly for $|X_1 - Y|$, we have

$$\begin{aligned}
E|X_1 - X_2| &= 2E(\max\{X_1, X_2\} - X_1) \\
E|X_1 - Y| &= 2E(\max\{X_1, Y\}) - E(X_1) - E(Y)
\end{aligned}$$

so that $C_\infty = E|X_1 - Y| - E|X_1 - X_2|/2$ and

$$C_m = C_\infty + \frac{E|X_1 - X_2|}{2m}$$

References

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**: 1–3.

David HA. 1998. Early sample measures of variability. *Statistical Science* **13**: 368–377.

Epstein ES. 1969. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology* **8**: 985–987.

Ferro CAT. 2007. Comparing probabilistic forecasting systems with the Brier score. *Weather and Forecasting* **22**: 1076–1088.

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

Müller WA, Appenzeller C, Doblas-Reyes FJ, Liniger MA. 2005. A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *Journal of Climate* **18**: 1513–1523.

Murphy AH. 1971. A note on the ranked probability score. *Journal of Applied Meteorology* **10**: 155–156.

Murphy AH. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology* **12**: 595–600.

Richardson DS. 2001. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society* **127**: 2473–2489.

Weigel AP, Liniger MA, Appenzeller C. 2007. The discrete Brier and ranked probability skill scores. *Monthly Weather Review* **135**: 118–124.

DOI: 10.1002/met