The Discrete Brier and Ranked Probability Skill Scores

ANDREAS P. WEIGEL, MARK A. LINIGER, AND CHRISTOF APPENZELLER

Federal Office of Meteorology and Climatology (MeteoSwiss), Zurich, Switzerland

(Manuscript received 23 February 2006, in final form 14 April 2006)

ABSTRACT

The Brier skill score (BSS) and the ranked probability skill score (RPSS) are widely used measures to describe the quality of categorical probabilistic forecasts. They quantify the extent to which a forecast strategy improves predictions with respect to a (usually climatological) reference forecast. The BSS can thereby be regarded as the special case of an RPSS with two forecast categories. From the work of Müller et al., it is known that the RPSS is negatively biased for ensemble prediction systems with small ensemble sizes, and that a debiased version, the RPSS_D, can be obtained quasi empirically by random resampling from the reference forecast. In this paper, an analytical formula is derived to directly calculate the RPSS bias correction for any ensemble size and combination of probability categories, thus allowing an easy implementation of the RPSS_D. The correction term itself is identified as the "intrinsic unreliability" of the ensemble prediction system. The performance of this new formulation of the RPSS_D is illustrated in two examples. First, it is applied to a synthetic random white noise climate, and then, using the ECMWF Seasonal Forecast System 2, to seasonal predictions of near-surface temperature in several regions of different predictability. In both examples, the skill score is independent of ensemble size while the associated confidence thresholds decrease as the number of ensemble members and forecast/observation pairs increase.

1. Introduction

Probabilistic forecasts with ensemble prediction systems (EPSs) have found a wide range of applications in weather and climate risk management, and their importance grows continuously. For example, the European Centre for Medium-Range Weather Forecasts (ECMWF) meanwhile operationally applies a 51-member EPS for medium-range weather predictions (e.g., Buizza et al. 2005) and a 40-member system for seasonal climate forecasts (Anderson et al. 2003). The rationale behind the ensemble method used is to approximate the expected probability density function (PDF) of a quantity by a finite set of forecast realizations. While the predicted probabilities account for the intrinsic uncertainties of atmospheric and ocean evolution, it is not trivial to verify them such that the full information content is considered. In particular, a predicted probability cannot be verified by a single observation (Candille and Talagrand 2005).

Several scores have been developed to quantify the performance of probabilistic prediction systems, for example, the rank histogram (Anderson 1996) and the relative operating characteristic (ROC; Mason 1982). The focus of this paper is on two other widely used scores, namely the Brier score and the ranked probability score (BS and RPS, respectively). The RPS (Epstein 1969; Murphy 1969, 1971) is a squared measure that compares the cumulative density function (CDF) of a probabilistic forecast with the CDF of the corresponding observation over a given number of discrete probability categories. The BS (Brier 1950) is a special case of an RPS with two categories.

The RPS can be formulated as a skill score (RPSS) by relating the score of the forecasts to the score of a reference forecast. For short-range climate predictions, "climatology" is often used as a reference forecast, that is, a forecast strategy that is based on the climatological probabilities of the forecast categories. The RPSS is a favorable skill score in that it considers both the shape and overall tendency of the forecast PDF. It is sensitive to distance, that is, a forecast is increasingly penalized the more its CDF differs from the actual outcome (Wilks 1995). Finally, the RPSS is a strictly proper skill score, meaning that it cannot be increased by hedging

Corresponding author address: Dr. Andreas Weigel, MeteoSwiss, Krähbühlstrasse 58, P.O. Box 514, CH-8044 Zurich, Switzerland.
E-mail: andreas.weigel@meteoswiss.ch

DOI: 10.1175/MWR3280.1

the probabilistic forecasts toward other values against the forecaster's true belief. A major flaw of the RPSS is its strong dependence on ensemble size (noticed, e.g., by Buizza and Palmer 1998), at least for ensembles with less than 40 members (Déque 1997). Indeed, the RPSS is negatively biased (Richardson 2001; Kumar et al. 2001; Mason 2004). Müller et al. (2005) investigated the reasons for this negative bias of the RPSS, and they suggested that it arises from sampling errors (properties inherent to the discretization and squaring measure used in the RPSS definition). As a strategy to overcome this deficiency, Müller et al. (2005) proposed to artificially introduce the same sampling error on the reference score. Using this "compensation of errors" approach, they defined a new "debiased" ranked probability skill score (RPSS_D) that is independent of ensemble size while retaining the favorable properties, in particular the strict propriety, of the RPSS. However, a deficiency of their approach is that the bias correction is determined quasi empirically by frequent random resampling from climatology.

In this paper, we overcome this deficiency and introduce an improved analytical version of the RPSS_D that does not require a resampling procedure. A simple formula is presented directly relating RPSS and RPSS_D and thus allowing for an analytical straightforward correction of the ensemble size dependent bias. The equation can be applied for any ensemble size and any choice of categories. It is derived and discussed in section 2. In section 3, the performance of the new RPSS_D formulation is demonstrated in a synthetic and a real case example. The last section provides concluding remarks.

2. The RPSS bias correction

a. Definition of RPS and RPSS

Let K be the number of forecast categories to be considered. For a given probabilistic forecast–observation pair, the ranked probability score is defined as

$$\mathrm{RPS} = \sum_{k=1}^{K} (Y_k - O_k)^2 = (\mathbf{Y} - \mathbf{O})^2. \tag{1}$$

Here, Y_k and O_k denote the kth component of the cumulative forecast and observation vectors $\mathbf{Y}$ and $\mathbf{O}$, respectively. That is, $Y_k = \Sigma_{i=1}^{k} y_i$, with y_i being the probabilistic forecast for the event to happen in category i, and $O_k = \Sigma_{i=1}^{k} o_i$ with $o_i = 1$ if the observation is in category i and $o_i = 0$ if the observation falls into a category $j \neq i$. Note that the RPS is zero for a perfect forecast and positive otherwise.

Let p_i be the climatological probability of the event falling into category i, and let P_i be the corresponding cumulative probability. If climatology is chosen as a reference strategy (as is done henceforth), then the RPS for the reference forecast, $\mathrm{RPS}_{\mathrm{Cl}}$, is

$$\mathrm{RPS}_{\mathrm{Cl}} = \sum_{k=1}^{K} (P_k - O_k)^2 = (\mathbf{P} - \mathbf{O})^2. \tag{2}$$

The ranked probability skill score relates RPS and $\mathrm{RPS}_{\mathrm{Cl}}$ such that a positive value of RPSS indicates forecast benefit with respect to the climatological forecast (Wilks 1995):

$$\mathrm{RPSS} = 1 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathrm{RPS}_{\mathrm{Cl}} \rangle}. \tag{3}$$

The angle brackets $\langle \cdot \rangle$ denote the average of the scores over a given number of forecast–observation pairs. In the special case of forecasts with exactly two categories, the RPS becomes the well-known BS and the RPSS becomes the Brier skill score (BSS), respectively:

$$\mathrm{BSS} = 1 - \frac{\langle \mathrm{BS} \rangle}{\langle \mathrm{BS}_{\mathrm{Cl}} \rangle}. \tag{4}$$

b. The discrete ranked probability skill score $RPSS_D$

As mentioned above, the RPSS is subject to a negative bias that is strongest for small ensemble size. Müller et al. (2005) have shown that this bias can be removed when the score of the reference forecast is calculated by repeated random resampling from climatology, with the number of samples being equal to the ensemble size. As a new reference, the mean is taken from a sufficient number of randomly resampled scores $\mathrm{RPS}_{\mathrm{ran}}$ [see Eq. (11) in Müller et al. 2005]. In other words, $\mathrm{RPS}_{\mathrm{Cl}}$ in Eq. (3) needs to be replaced by the expectation $\mathcal{E}$ of the scores $\mathrm{RPS}_{\mathrm{ran}}$, which the very ensemble prediction system under consideration would produce in the case of merely random climatological resamples. A debiased ranked probability skill score can then be formulated as

$$\mathrm{RPSS}_D = 1 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathcal{E}(\mathrm{RPS}_{\mathrm{ran}}) \rangle}. \tag{5}$$

In the following, we derive an expression for the difference D between $\mathcal{E}(\mathrm{RPS}_{\mathrm{ran}})$ and $\mathrm{RPS}_{\mathrm{Cl}}$, that is, for the term that apparently is responsible for the negative bias of the RPSS. We start by considering the process of randomly sampling one M-member ensemble from climatology, leading to one realization of $\mathrm{RPS}_{\mathrm{ran}}$. As above, consider K forecast categories with climatological probabilities $p_1, p_2, \ldots, p_K$. Let m_k denote the number of ensemble members drawn from the kth category

(such that $\Sigma_{k=1}^{K} m_k = M$), and let $y_k = m_k/M$ be the corresponding probability forecast issued. The probability that the sampling process yields a forecast $\mathbf{y} = (y_1, y_2, \ldots, y_K)$, that is, the probability to simultaneously draw $m_1, m_2, \ldots, m_K$ ensemble members from category $1, 2, \ldots, K$ with climatological probabilities $p_1, p_2, \ldots, p_K$, is described by the multinomial distribution $\mathcal{M}$ (e.g., Linder 1960):

$$\mathcal{M}(m_1, m_2, \ldots, m_K | p_1, p_2, \ldots, p_K) = \frac{M!}{m_1! m_2! \ldots m_K!} p_1^{m_1} p_2^{m_2} \ldots p_K^{m_K}. \quad (6)$$

Note that $\mathcal{M}$ is the generalization of the binomial distribution to multievent situations. Using $\mathcal{M}$, the expectation of the number of ensemble members drawn from the kth category, $\mathcal{E}[m_k]$, can be calculated by summing over all possible sampling results $\{m_1, m_2, \ldots, m_K\}$, yielding (Linder 1960):

$$\begin{aligned}\mathcal{E}(m_k) &= \sum_{\{m_1, m_2, \ldots, m_K\}} m_k \\ &\quad \times \mathcal{M}(m_1, m_2, \ldots, m_K | p_1, p_2, \ldots, p_K) \\ &= M p_k. \end{aligned} \quad (7)$$

Similarly, variance and covariance can be determined to (Linder 1960):

$$\mathrm{var}(m_k) = M p_k (1 - p_k), \quad (8)$$

$$\mathrm{cov}(m_k, m_l) = -M p_k p_l, \quad (9)$$

where $k, l \in \{1, 2, \ldots, K\}$ and $k \neq l$. Using the multinomial distribution and Eq. (2), the difference D between $\mathcal{E}(\mathrm{RPS}_{\mathrm{ran}})$ and $\mathrm{RPS}_{\mathrm{Cl}}$ can now be calculated for a given arbitrary observation with cumulative density $\mathbf{O}$:

$$D = \mathcal{E}(\mathrm{RPS}_{\mathrm{ran}}) - \mathrm{RPS}_{\mathrm{Cl}} \quad (10)$$

$$= \mathcal{E}[(\mathbf{Y} - \mathbf{O})^2] - (\mathbf{P} - \mathbf{O})^2. \quad (11)$$

From Eq. (7), it follows that $p_k = \mathcal{E}(m_k/M) = \mathcal{E}(y_k)$. Thus:

$$\begin{aligned} D &= \mathcal{E}[(\mathbf{Y} - \mathbf{O})^2] - [\mathcal{E}(\mathbf{Y}) - \mathbf{O}]^2 \\ &= \mathcal{E}(\mathbf{Y}^2) - [\mathcal{E}(\mathbf{Y})]^2 \\ &= \sum_{k=1}^{K} \{\mathcal{E}(Y_k^2) - [\mathcal{E}(Y_k)]^2\} \\ &= \sum_{k=1}^{K} \mathrm{var}(Y_k) \\ &= \sum_{k=1}^{K} \mathrm{var}\left(\sum_{i=1}^{k} \frac{m_i}{M}\right). \end{aligned} \quad (12)$$

Equation (12) can be further simplified by applying the law for the variance of a sum and Eqs. (8) and (9):

$$\begin{aligned} D &= \sum_{k=1}^{K} \mathrm{var}\left(\sum_{i=1}^{k} \frac{m_i}{M}\right) \\ &= \frac{1}{M^2} \sum_{k=1}^{K} \left[\sum_{i=1}^{k} \mathrm{var}(m_i) + 2 \sum_{i=1}^{k} \sum_{j=i+1}^{k} \mathrm{cov}(m_i, m_j)\right] \\ &= \frac{1}{M} \sum_{k=1}^{K} \sum_{i=1}^{k} \left[p_i (1 - p_i - 2 \sum_{j=i+1}^{k} p_j)\right]. \end{aligned} \quad (13)$$

Note that D only depends on ensemble size M and on the climatological probabilities $p_1, p_2, \ldots, p_k$ of the K categories, but not on the observation $\mathbf{O}$. Thus, the RPSS_D can be simply formulated as

$$\mathrm{RPSS}_D = 1 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathrm{RPS}_{\mathrm{Cl}} \rangle + D}, \quad (14)$$

with D given by Eq. (13). For very large ensemble size M, the correction term D converges toward zero and the RPSS_D toward the RPSS. On the other extreme, D is the maximum for "ensembles" with one member only.

If the K categories are equiprobable, that is, if $p_k = 1/K$ for all $k \in \{1, \ldots, K\}$, then Eq. (13) can be further simplified and D now only depends on ensemble size M and on the number of categories K:

$$D = \frac{1}{M} \frac{K^2 - 1}{6K}. \quad (15)$$

Note that increasing the number of categories also increases D. In the limit of large K, the correction D becomes proportional to K/M.

c. The discrete Brier skill score BSS_D

Finally, consider the special case of two categories with probabilities p and $(1 - p)$, respectively. Then Eq. (13) becomes

$$D_{\mathrm{bin}} \equiv D = \frac{1}{M} p(1 - p) \quad (16)$$

and a debiased version of the Brier skill score is given by

$$\mathrm{BSS}_D = 1 - \frac{\langle \mathrm{BS} \rangle}{\langle \mathrm{BS}_{\mathrm{Cl}} \rangle + D_{\mathrm{bin}}}. \quad (17)$$

The D_{bin} is maximum for $p = 0.5$ and minimum for $p = 0$ or $p = 1$. Thus, the impact of the skill score correction is negligible when rare (i.e., extreme) events are con-

sidered and is strongest when occurrence and nonoccurence of the event are equiprobable.

To understand the meaning of the correction term D_{bin}, we decompose the average Brier score $\langle \text{BS} \rangle$ into its reliability (REL), resolution (RES), and uncertainty (UNC) components (for details of this decomposition, see Murphy 1973; Wilks 1995):

$$\begin{aligned}\langle \text{BS} \rangle &= \text{REL} - \text{RES} + \text{UNC} \\ &= \frac{1}{N}\sum_{j=1}^{J} N_j(y_j - \overline{o}_j)^2 \\ &\quad - \frac{1}{N}\sum_{j=1}^{J} N_j(\overline{o}_j - \overline{o})^2 + \overline{o}(1 - \overline{o}),\end{aligned} \tag{18}$$

where N is the total number of forecast–observation pairs considered, J is the number of distinct forecast probabilities issued, and N_j denotes the number of times each forecast probability y_j is used in the collection of forecasts being verified. Here, $\overline{o}$ is the relative frequency of the observations (note that $\overline{o} = p$ for sufficient N), and $\overline{o}_j$ is the subsample relative frequency. The average Brier score for a climatological forecast, $\langle \text{BS}_{\text{Cl}} \rangle$, equals the uncertainty term, as both resolution and reliability are zero:

$$\langle \text{BS}_{\text{Cl}} \rangle = \text{UNC}. \tag{19}$$

For the average Brier score for a climatological forecast based on random resampling from climatology, $\langle \text{BS}_{\text{ran}} \rangle$, the reliability term does not vanish since the probability forecasts issued (y_j) are generally different from $\overline{o}$. The resolution term is still zero because it does not depend on the forecast probabilities. We get

$$\langle \text{BS}_{\text{ran}} \rangle = \text{UNC} + \text{REL}_{\text{ran}}. \tag{20}$$

Using Eqs. (19) and (20), the correction term D_{bin} can be identified as

$$D_{\text{bin}} = \langle \text{BS}_{\text{ran}} \rangle - \langle \text{BS}_{\text{Cl}} \rangle = \text{REL}_{\text{ran}}. \tag{21}$$

Thus, D_{bin} is the reliability term of the EPS under consideration for the case of randomly resampled forecasts. The correction term can therefore be interpreted as the EPS's intrinsic reliability—or better yet, "intrinsic unreliability"—and is due to the finite number of ensemble members. This means that the classical BSS, which does not consider D_{bin}, compares intrinsically unreliable forecasts with a perfectly reliable reference forecast. However, given that reliability deficits can be removed a posteriori by calibration (Toth et al. 2003), a skill assessment based on such a comparison is unfair against the EPS. This conceptual deficiency of the BSS is overcome by adding D_{bin} to the climatological reference, ensuring that two equally (un)reliable forecast systems are compared. This interpretation holds analogously for the multicategorical RPSS, because the RPS can be formulated as a sum of Brier scores (Toth et al. 2003).

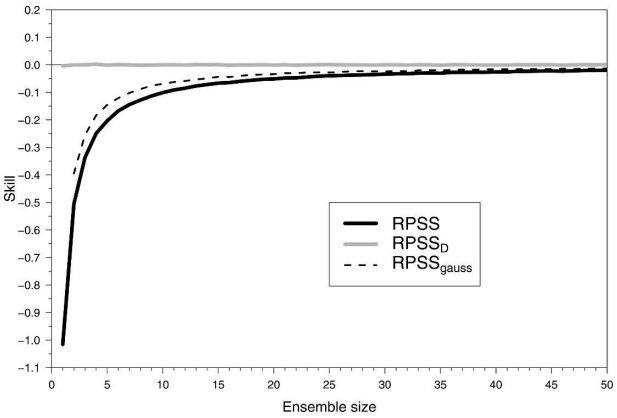

FIG. 1. Expectation $\mathcal{E}$ of ranked probability skill score (RPSS) and debiased ranked probability skill score (RPSS_D) for random white noise forecasts as a function of ensemble size (solid lines). The dashed line ($\text{RPSS}_{\text{gauss}}$) is the RPSS derived from Gaussian distributions fitted to the forecast ensembles. The skill scores are based on three equiprobable categories.

3. Examples

In this section, the advantages of the new RPSS_D formulation for discrete probability forecasts are illustrated with two examples.

a. RPSS_D applied to synthetic white noise forecasts

A Gaussian white noise climate is used to explore the performance of the derived analytical bias correction for different ensemble sizes. Similar to the procedure of Müller et al. (2005), 15 forecast ensembles together with 15 observations are randomly chosen from this synthetic "climate" and are classified into three equiprobable categories. From these, the $\langle \text{RPS} \rangle$ and $\langle \text{RPS}_{\text{Cl}} \rangle$ values are calculated. Using Eqs. (3) and (14), the RPSS and RPSS_D (in its new formulation) are determined. This procedure is repeated 10 000 times for ensemble sizes ranging from 1 to 50 members, yielding 10 000 RPSS and RPSS_D values each. From these, the means are obtained as plotted in Fig. 1 (solid lines) in the function of different ensemble sizes. While the RPSS reveals a negative bias that becomes stronger with decreasing ensemble size (ranging from −0.02 for a 50-member ensemble to −0.50 for a 2-member ensemble), the RPSS_D remains constantly near zero for all ensemble sizes; that is, the negative bias is removed. These results are equivalent to those shown by Müller

et al. (2005, their Fig. 1), who determined the $RPSS_D$ with Eq. (5), that is, by frequent resampling.

In an alternative approach, we fit Gaussian distributions to the randomly chosen ensembles from above and calculate the RPSS values on the basis of these distributions, rather than using the raw ensemble output. This method is considered to yield more accurate estimates of the forecast quantiles, provided that the parent distribution is Gaussian (Wilks 2002). The negative bias of the RPSS can indeed be slightly reduced, though not removed, by such a smoothing procedure (dashed line in Fig. 1). This is consistent with the notion of increasing the effective ensemble size by such a fit (Wilks 2002). In the terminology of section 2c, the intrinsic unreliability of an EPS can be reduced, but not eliminated, by smoothing the forecast ensembles with an appropriate distribution.

Müller et al. (2005) pointed out that despite the bias elimination provided by the $RPSS_D$, there remains a signal-to-noise detection limit, or in other words, there is always a chance that a positive skill score is due to coincidence and that the predictions are, in fact, not any better than the reference. This is the null hypothesis one wants to reject. Therefore, it is necessary to know the confidence intervals of the $RPSS_D$ distribution that the very EPS under consideration would produce in the case of merely random climatological forecasts. An estimation of this distribution can be obtained by the procedure described above, that is, by explicitly sampling a large number of $RPSS_D$ values from a white noise climate for a given number of ensemble members and forecast–observation pairs. In Fig. 2, the 95% quantiles of such $RPSS_D$ distributions are displayed. Obviously, the confidence threshold decreases both with growing ensemble size and the number of forecast–observation pairs. Thus, while the expectation $\mathcal{E}$ of the $RPSS_D$ is independent of ensemble size, large ensembles are nevertheless favorable, as the confidence threshold moves closer toward zero and smaller skill score values become statistically significant. For example, if an EPS with five ensemble members is given and five forecast–observation pairs are available, then one cannot make the statistically significant conclusion that the EPS outperforms the reference, unless the $RPSS_D$ is larger than 0.42. This confidence threshold can be halved, for instance, by increasing the ensemble size to 27.

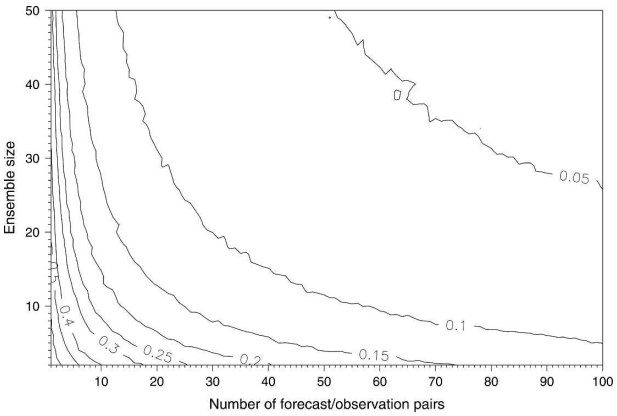

FIG. 2. Upper 95% confidence levels (contour lines) for the debiased ranked probability skill score ($RPSS_D$) for random white noise forecasts as a function of ensemble size and number of forecast–observation pairs. Three equiprobable forecast categories are used. The confidence levels are estimated from 10 000 $RPSS_D$ values.

b. BSS_D applied to real forecast data

In a second example, the correction formula for the RPSS [Eq. (13)] is applied to the problem of quantifying the gain in prediction skill due to increasing ensemble size—a topic that is also relevant for comparing single versus multimodel approaches in seasonal forecasting (e.g., Palmer et al. 2004). Here, the widely used Brier skill score is employed with two equiprobable categories. Forecasts of mean near-surface (2 m) temperature for March with 4-months lead time are used as hindcast data. They are obtained from the ECMWF System 2 seasonal prediction model with 40 ensemble members (Anderson et al. 2003) and are verified gridpoint-wise against the corresponding "observations" from the 40-yr ECMWF Re-Analysis (ERA-40) dataset (Uppala et al. 2005). The threshold separating the two probability categories is determined from the 15 yr of hindcast and observation data, respectively. Three regions are considered: (i) the Niño-3.4 region (5°S–5°N, 120°–170°W) with enhanced seasonal predictability (e.g., Latif et al. 1998), (ii) a region over southern Africa (0°–40°S, 10°–50°E) with "intermediate" seasonal predictability (Klopper et al. 1998), and (iii) a region over central Europe (40°–55°N, 10°–30°E) where seasonal predictability is considered low (e.g., Müller et al. 2004). For these three regions, both BSS [Eq. (4)] and BSS_D [Eq. (17)] are determined by spatial and temporal averaging over all BS and BS_{Cl} scores obtained at each grid point for each hindcast year. This procedure is then repeated for smaller ensemble sizes (down to 2 members), which are constructed by randomly resampling from the 40 ensemble members available in the ECMWF forecasts. The results are shown in Fig. 3, where BSS and BSS_D are plotted against ensemble size for each of the regions. Consistent with what has been shown for the white noise climate above, in all three cases, the value of the classical BSS increases with ensemble size. This is also consistent with earlier studies (see, e.g., Palmer et al. 2004, their Fig. 5). On the other

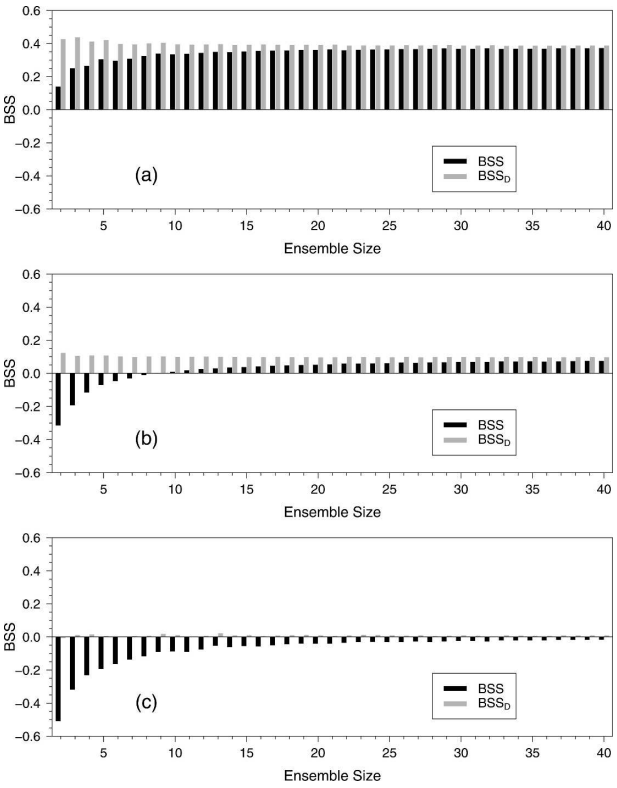

FIG. 3. Brier skill score (BSS) and debiased Brier skill score (BSS_D) as a function of ensemble size for near-surface temperature predictions for March with a lead time of 4 months. Data are averaged over 15 yr (1988–2002) over (a) the Niño-3.4 region, (b) southern Africa, and (c) central Europe. Two equiprobable classes are used, that is, the events are "temperature below normal" and "temperature above normal." The forecasts are obtained from the ECMWF Seasonal Forecast System 2 and the verifying observations are from ERA-40 data.

hand, with the new formulation of the Brier skill score, the negative bias disappears for all three geographical regions considered. Figure 3 suggests that the value of the BSS_D becomes independent of ensemble size for any prediction skill. Strictly speaking, in section 2, this has only been proven for uncorrelated ensemble members. Note that in this example, the calculation of the 95% confidence intervals is not trivial because spatial averages over statistically dependent grid points are involved and thus the number of independent forecast–observation pairs is not known. This problem could be overcome, for example, by estimating the effective number of spatial degrees of freedom, as suggested by Bretherton et al. (1999).

4. Conclusions

This study has addressed the negative bias of the widely used Brier and ranked probability skill scores (BSS and RPSS, respectively). A simple analytical formula has been derived quantifying the bias dependent upon category probabilities and ensemble size. Using this expression, an easy-to-implement formulation for a debiased discrete ranked probability skill score ($RPSS_D$) has been obtained. Formally, the $RPSS_D$ is calculated by adding a correction term D to the "classical" reference score of the RPSS. This correction term is indirectly proportional to the number of ensemble members and can therefore be neglected for systems with large ensemble size.

The performance of the new $RPSS_D$ formulation has been illustrated in two examples. They show that the expectation of the $RPSS_D$ for a given forecast strategy is indeed independent of ensemble size but not of its variance. This means that increasing the number of ensemble members does not increase prediction skill, per se; rather, the statistical significance of the skill score is enhanced.

The actual reason for the negative bias of the RPSS is the fact that ensemble prediction systems (EPSs) are inherently unreliable, that is, have a positive reliability term. Their "intrinsic unreliability" grows as ensemble size decreases. Given that reliability deficits can be corrected by calibration a posteriori, it may be misleading to base a skill score on the comparison of two prediction systems with different ensemble size and thus different intrinsic unreliability. This is exactly what happens in the case of the conventional RPSS, because its reference score implies—in the $RPSS_D$ view—an infinitely large ensemble size [such that D is zero in Eq. (14)]. In other words, the RPSS could be regarded as ill conditioned because it compares the score of an intrinsically unreliable forecast system of finite ensemble size with a reference score of a forecast system of infinite ensemble size and perfect reliability. This problem is resolved and the skill score debiased by adding the EPS's intrinsic unreliability, D, to the climatological reference.

In summary, the $RPSS_D$ provides a powerful and easily implemented tool for the evaluation of probabilistic forecasts with small ensembles and for the comparison of different EPSs of different ensemble size. The bias correction is particularly important when multimodel ensembles are to be evaluated, where the benefit of increasing ensemble size needs to be clearly distinguished from the benefits of multimodel combination.

Acknowledgments. Thanks are expressed to Wolfgang Müller for his helpful comments on this manuscript. This study was supported by the Swiss National Science Foundation through the National Centre for Competence in Research (NCCR Climate) and by ENSEMBLES EC Contract GOCE-CT-2003-505539.

REFERENCES

Anderson, D. L. T., and Coauthors, 2003: Comparison of the ECMWF seasonal forecast systems 1 and 2, including the relative performance for the 1997/8 El Niño. ECMWF Tech. Memo. 404, 93 pp.

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

Bretherton, C. S., M. Widmann, V. P. Dymnikov, J. M. Wallace, and I. Bladé, 1999: The effective number of spatial degrees of freedom of a time-varying field. *J. Climate,* **12,** 1990–2009.

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

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.,* **126,** 2508–2518.

——, P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.,* **133,** 1076–1097.

Candille, G., and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.,* **131,** 2131–2150.

Déque, M., 1997: Ensemble size for numerical weather forecasts. *Tellus,* **49A,** 74–86.

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

Klopper, E., W. A. Landmann, and J. van Heerden, 1998: The predictability of seasonal maximum temperature in South Africa. *Int. J. Climatol.,* **18,** 741–758.

Kumar, A., A. G. Barnston, and M. R. Hoerling, 2001: Seasonal predictions, probabilistic verifications, and ensemble size. *J. Climate,* **14,** 1671–1676.

Latif, M., D. Anderson, M. Cane, R. Kleeman, A. Leetmaa, J. O'Brien, A. Rosati, and E. Schneider, 1998: A review of the predictability and prediction of ENSO. *J. Geophys. Res.,* **103,** 14 375–14 393.

Linder, A., 1960: *Statistische Methoden.* 3d ed. Birkhäuser Verlag, 484 pp.

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

Mason, S. J., 2004: On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.,* **132,** 1891–1895.

Müller, W. A., C. Appenzeller, and C. Schär, 2004: Probabilistic seasonal prediction of the winter North Atlantic Oscillation and its impact on near surface temperature. *Climate Dyn.,* **24,** 213–226.

——, ——, F. J. Doblas-Reyes, and M. A. Liniger, 2005: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *J. Climate,* **18,** 1513–1523.

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

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

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

Palmer, T. N., and Coauthors, 2004: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.,* **127,** 2473–2489.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification–A Practitioner's Guide in Atmospheric Science,* I. T. Joliffe and D. B. Stephenson, Eds., John Wiley & Sons, 137–163.

Uppala, S. M., and Coauthors, 2005: The ERA-40 re-analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* International Geophysics Series, Vol. 59, Academic Press, 467 pp.

——, 2002: Smoothing forecast ensembles with fitted probability distributions. *Quart. J. Roy. Meteor. Soc.,* **128,** 2821–2836.