Comment on “The Discrete Brier and Ranked Probability Skill Scores”

MICHAEL K. TIPPETT[*]

International Research Institute for Climate and Society, Palisades, NY, USA

March 10, 2008

[*] *Corresponding author address*: M. K. Tippett, International Research Institute for Climate and Society, The Earth Institute of Columbia University, Lamont Campus / 61 Route 9W, Palisades New York 10964, USA. (tippett@iri.columbia.edu)

1. Introduction

The ranked probability score (RPS) is the sum of the squared differences between cumulative forecast probabilities and cumulative observed probabilities, and measures both forecast reliability and resolution (Murphy 1973). The ranked probability skill score (RPSS) compares the RPS of a forecast to some reference forecast such as climatology, oriented so that RPSS < 0 (RPSS > 0) corresponds to a forecast that is less (more) skillful than climatology.

Categorical forecast probabilities are often estimated from ensembles of numerical model integrations by counting the number of ensemble members in each category. Finite ensemble size introduces sampling error into such probability estimates, and the RPSS of a reliable forecast model with finite ensemble size is an increasing function of ensemble size (Kumar et al. 2001; Tippett et al. 2007). A similar relation exists between correlation and ensemble size (Sardeshmukh et al. 2000). The dependence of RPSS on ensemble size makes it challenging to use RPSS to compare forecast models with different ensemble sizes. For instance, it may be difficult to know whether a forecast system has higher RPSS because it is based on a superior forecast model or because it uses a larger ensemble. This question often arises in the comparison of multi-model and single model forecasts (Hagedorn et al. 2005; Tippett and Barnston 2008). The dependence of RPSS on ensemble size is not a problem when comparing forecast quality. Improved RPSS is associated with improved forecast quality and is desirable whether due to larger ensemble size or a better forecast model.

Recently, Müller et al. (2005) introduced a re-sampling strategy to estimate the infinite-ensemble RPSS from the finite-ensemble RPSS and called this estimate the *debiased RPSS*. Weigel et al. (2007) derived an analytical formula for the debiased RPSS and proved that it is an unbiased estimate of the infinite-ensemble RPSS in the case of uncorrelated ensemble members, that is, forecasts without skill. Here it is proved that the debiased RPSS is an unbiased estimate of the infinite-ensemble RPSS for any reliable forecasts. It is shown that over- or under-confident forecasts introduce a dependence of the debiased RPSS on ensemble size. Simplification of the results Weigel et al. (2007) show that the debiased RPSS is a multi-category generalization of the result of Richardson (2001) for the Brier skill score.

2. RPSS and debiased RPSS

The RPS of a K-category probability forecast is

$$\mathrm{RPS} = \sum_{k=1}^{K} \left[\sum_{i=1}^{k} P_i - O_i \right]^2 , \tag{1}$$

where P_i is the forecast probability assigned to the i-th category and O_i is one when the observation falls into the i-th category and zero otherwise. When forecast probabilities are computed by counting the number of ensemble members in each category, finite ensemble size results in sampling errors which increase RPS.

In the case of two categories, RPS is the Brier score. Richardson (2001) showed the dependence of the Brier score on ensemble size M in a reliable forecast system. Tippett et al. (2007) generalized that result to tercile categories and later (Tippett and Barnston 2008) to arbitrary number of categories as

$$\langle \mathrm{RPS}(M) \rangle = \left(1 + \frac{1}{M} \right) \langle \mathrm{RPS}(\infty) \rangle \, , \tag{2}$$

indicating how decreasing ensemble size increases the expected RPS.

The ranked probability skill score (RPSS) is

$$\mathrm{RPSS} \equiv 1 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathrm{RPS_{Cl}} \rangle}\,, \tag{3}$$

where $\mathrm{RPS_{Cl}}$ is the RPS of a reference forecast consisting of climatological probabilities and $\langle\cdot\rangle$ denotes average over forecasts. Sampling error causes RPSS to decrease. Using (2), the infinite-ensemble RPSS can be expressed in terms of the finite-ensemble RPSS as

$$\mathrm{RPSS}(\infty) = 1 - \frac{\langle \mathrm{RPS}(\infty) \rangle}{\langle \mathrm{RPS_{Cl}} \rangle} = 1 - \frac{\langle \mathrm{RPS}(M) \rangle}{\langle \mathrm{RPS_{Cl}} \rangle + \frac{1}{M}\langle \mathrm{RPS_{Cl}} \rangle}\,. \tag{4}$$

The strategy introduced by Müller et al. (2005) to estimate $\mathrm{RPSS}(\infty)$ from $\mathrm{RPSS}(M)$ was to artificially increase the error in the reference forecast by computing climatological probabilities using the same number of samples as ensemble members and then define a *debiased RPSS* denoted RPSS_D by

$$\mathrm{RPSS}_D \equiv 1 - \frac{\langle \mathrm{RPS}(M) \rangle}{\langle \mathrm{RPS_{Cl}}(M) \rangle}\,. \tag{5}$$

Müller et al. (2005) showed in numerical examples with reliable forecasts and tercile categories that RPSS_D had little if any dependence on ensemble size.

Using (2), one can immediately see that RPSS_D is the same as $\mathrm{RPSS}(\infty)$ and is indeed an unbiased estimate for the infinite-ensemble RPSS for all reliable forecasts since

$$\begin{aligned}\mathrm{RPSS}_D &= 1 - \frac{\langle \mathrm{RPS}(M) \rangle}{\langle \mathrm{RPS_{Cl}}(M) \rangle} \\ &= 1 - \frac{\left(1 + \frac{1}{M}\right)\langle \mathrm{RPS}(\infty) \rangle}{\langle \left(1 + \frac{1}{M}\right) \mathrm{RPS_{Cl}} \rangle} \\ &= 1 - \frac{\langle \mathrm{RPS}(\infty) \rangle}{\langle \mathrm{RPS_{Cl}} \rangle} \\ &= \mathrm{RPSS}(\infty)\,.\end{aligned} \tag{6}$$

The impact of sample size on expected RPS is multiplicative and independent of skill level. Therefore the ratio of the RPSS of two reliable forecasts systems with the same ensemble size is independent of ensemble size.

In Müller et al. (2005) $\langle \mathrm{RPS}_{\mathrm{Cl}}(M) \rangle$ was computed by repeatedly sampling from the historical record. Weigel et al. (2007) computed $\langle \mathrm{RPS}_{\mathrm{Cl}}(M) \rangle$ analytically using properties of the multinomial distribution and expressed RPSS_D as

$$\mathrm{RPSS}_D \equiv 1 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathrm{RPS}_{\mathrm{Cl}} \rangle + D} \, , \tag{7}$$

where

$$D \equiv \frac{1}{M} \sum_{k=1}^{K} \sum_{i=1}^{k} \left[p_i (1 - p_i - 2 \sum_{j=i+1}^{k} p_j) \right] \, , \tag{8}$$

and where p_i is the climatological probability of the i-th category. In light of (4), it must be the case that

$$D = \frac{1}{M} \langle \mathrm{RPS}_{\mathrm{Cl}} \rangle \, . \tag{9}$$

To prove (9) directly, first the expression for D is simplified. From (12) of Weigel et al. (2007),

$$D = \sum_{k=1}^{K} \mathrm{var} \left(\sum_{i=1}^{k} \hat{p}_i \right) \, , \tag{10}$$

where $\hat{p}_i$ is the M-member sample estimate of p_i. Since the M-member sample estimates of the cumulative probabilities are binomially distributed, their means are C_i and their variances are $C_i(1 - C_i)/M$ where the cumulative climatological probability C_i is defined by

$$C_i \equiv \sum_{k=1}^{i} p_k \, . \tag{11}$$

Therefore, D has the simple form

$$D = \frac{1}{M} \sum_{i=1}^{m} C_i (1 - C_i) \, . \tag{12}$$

Next $\langle \mathrm{RPS}_{\mathrm{Cl}} \rangle$ is expressed in term of the climatological categorical probabilities p_i. Explicitly, $\langle \mathrm{RPS}_{\mathrm{Cl}} \rangle$ is

$$\mathrm{RPS}_{\mathrm{Cl}} = \sum_{i=1}^{m} \left[\sum_{j=1}^{i} p_j - O_j \right]^2 . \tag{13}$$

The expected value of $\mathrm{RPS}_{\mathrm{Cl}}$ is simply (13) summed over all possible outcomes of the observations, weighted by the probabilities of each outcome. That is,

$$\langle \mathrm{RPS}_{\mathrm{Cl}} \rangle = \sum_{l=1}^{m} p_l \sum_{i=1}^{m} \left[\sum_{j=1}^{i} p_j - \delta_{jl} \right]^2 , \tag{14}$$

where the Kronecker delta δ_{ij} is defined to be one when $i = j$ and zero otherwise. Direct manipulation of this expression gives

$$\begin{aligned} \langle \mathrm{RPS}_{\mathrm{Cl}} \rangle &= \sum_{l=1}^{m} p_l \sum_{i=1}^{m} \left[\sum_{j=1}^{i} p_j - \delta_{jl} \right] \left[\sum_{k=1}^{i} p_k - \delta_{kl} \right] \\ &= \sum_{l=1}^{m} \sum_{i=1}^{m} \sum_{j=1}^{i} \sum_{k=1}^{i} p_l (p_j - \delta_{jl})(p_k - \delta_{kl}) \\ &= \sum_{l=1}^{m} \sum_{i=1}^{m} \sum_{j=1}^{i} \sum_{k=1}^{i} p_l (p_j p_k - \delta_{kl} p_j - \delta_{jl} p_k + \delta_{jl} \delta_{kl}) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{i} \sum_{k=1}^{i} p_j p_k - p_k p_j - p_j p_k + \delta_{jk} p_j \\ &= \sum_{i=1}^{m} \sum_{j=1}^{i} \sum_{k=1}^{i} \delta_{jk} p_j - p_j p_k \\ &= \sum_{i=1}^{m} \sum_{k=1}^{i} p_k - \left[\sum_{k=1}^{i} p_k \right]^2 = \sum_{i=1}^{m} C_i (1 - C_i) , \end{aligned} \tag{15}$$

thus proving (9).

3. Unreliable forecasts

However, the results above do not give any guidance about the dependence of RPSS on ensemble size when the forecasts are unreliable. Ferro et al. (2008) derive a more general estimator for RPSS that is applicable to under-confident and over-confident ensembles, as long as the ensemble members are "exchangeable." While Müller et al. (2005) states that RPSS_D is an unbiased estimate of the infinite-ensemble RPSS and is independent of ensemble size, there was no explicit examination of the behavior of the RPSS_D for unreliable forecasts. The behavior of RPSS_D is investigated here in an example where the forecasts are unreliable.

A simple univariate example is considered here in which the forecasts and observations are normally distributed. The expected correlation between the ensemble mean and observations is r, while the expected correlation between the ensemble mean and an ensemble member is r_f; r_f measures potential predictability, the ability of the forecast model to predict itself. Explicitly, the observations are normally distributed with mean rs and variance $1-r^2$, denoted $N(rs, 1-r^2)$, and the forecast distribution is $N(r_f s, 1-r_f^2)$; the distribution of s is $N(0,1)$. The forecast is reliable when $r_f = r$ and overconfident (under-confident) when $r_f > r$ ($r_f < r$).

Values of r and r_f were chosen corresponding to reliable, weakly over-confident, very over-confident, weakly under-confident, and very under-confident forecast systems, as indicated in Table 1. The expected values of $\mathrm{RPSS}(M)$ and RPSS_D for tercile-based categorical forecasts were computed from 10^6 simulations of the observations and forecast ensembles. Figure 1 shows the results as a function of ensemble size M. Figure 1a shows that RPSS_D is, as proved, an unbiased estimate of $\mathrm{RPSS}(\infty)$ independent of ensemble size. Figures 1b and 1c show that for over-confident forecasts RPSS_D over estimates $\mathrm{RPSS}(\infty)$, with the discrepancy between RPSS_D and $\mathrm{RPSS}(\infty)$ being greater than that between $\mathrm{RPSS}(M)$ and $\mathrm{RPSS}(\infty)$ for very over-confident forecasts. There is some indication of the tendency of RPSS_D to over estimate $\mathrm{RPSS}(\infty)$ in Figs. 3a and 3b of Weigel et al. (2007), indicating model over-confidence. In the under-confident examples, RPSS_D slightly under estimates $\mathrm{RPSS}(\infty)$.

4. Summary

The ranked probability skill score (RPSS) measures the reliability and resolution of categorical probability forecasts relative to the climatology forecast. (Murphy 1973). When categorical forecast probabilities are estimated from finite ensembles, sampling error negatively impacts RPSS (Kumar et al. 2001; Tippett et al. 2007). Recently, Weigel et al. (2007) derived an analytical formula for the debiased RPSS, an estimate of the infinite-ensemble RPSS in terms of the finite ensemble RPSS, based on the re-sampling strategy of Müller et al. (2005). Here it has been proved that the debiased RPSS is an unbiased estimate of the infinite-ensemble RPSS for reliable forecasts only. Over- or under-confident forecasts introduce dependence of the debiased RPSS on ensemble size. Analysis of the results of Weigel et al. (2007) show that the debiased RPSS is a multi-category generalization of the Brier skill score result of Richardson (2001).

Acknowledgments.

The author thanks Simon Mason, Andreas Weigel and Tony Barnston for their comments and suggestions. The author is supported by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NA05OAR4311004). The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its sub-agencies.

REFERENCES

Ferro, C. A. T., D. S. Richardson, and A. P. Weigel, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, in press.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus A*, **57 (3)**, 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

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

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

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.

Sardeshmukh, P. D., G. P. Compo, and C. Penland, 2000: Changes of probability associated with El Niño. *J. Climate*, **13**, 4268–4286.

Tippett, M. K. and A. G. Barnston, 2008: Skill of multi-model ENSO probability forecasts. *Mon. Wea. Rev.*, accepted.

Tippett, M. K., A. G. Barnston, and A. W. Robertson, 2007: Estimation of seasonal precipitation tercile-based categorical probabilities from ensembles. *J. Climate*, **20**, 2210–2228.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **135**, 118–124.

List of Figures

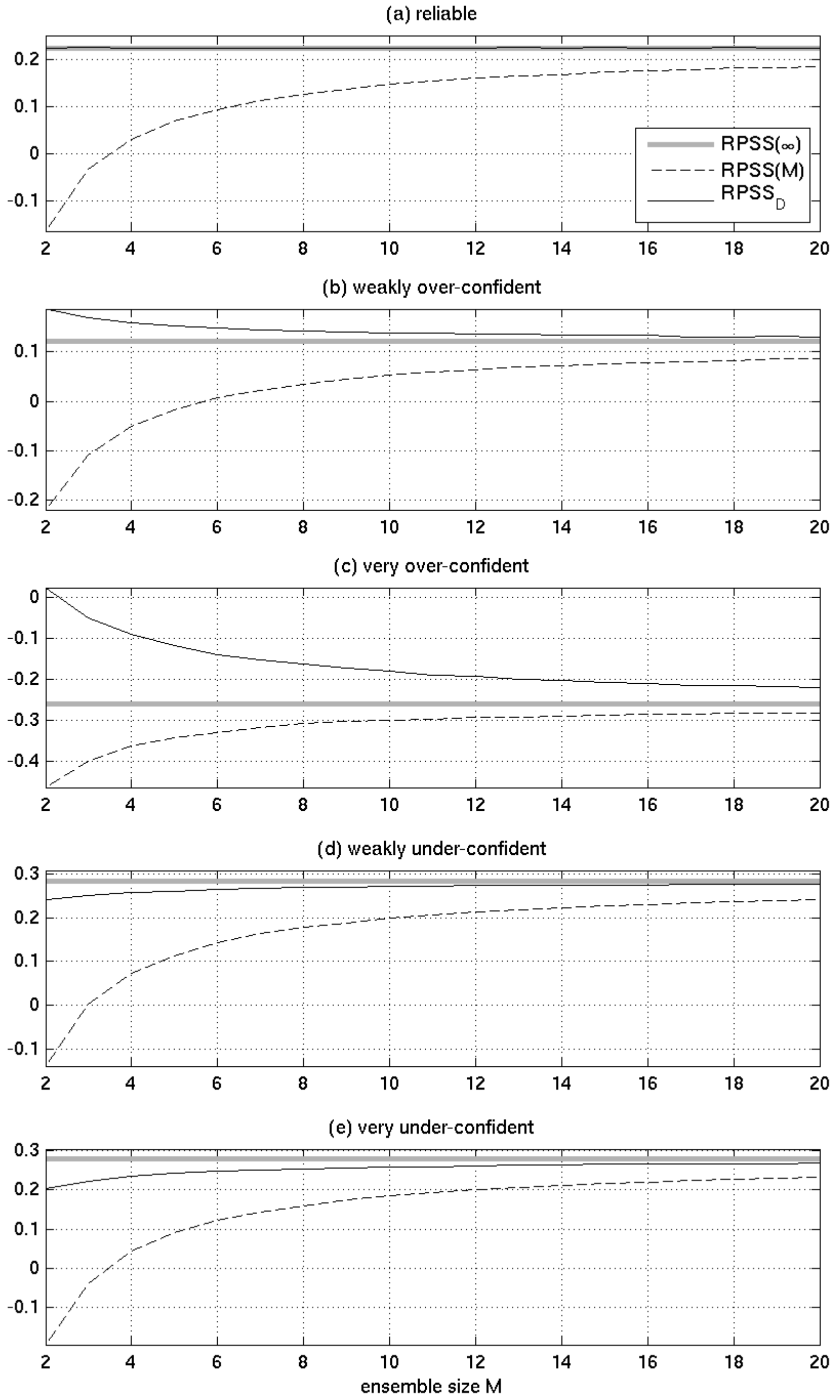

FIG. 1. $\mathrm{RPSS}(\infty)$ (thick gray line), $\mathrm{RPSS}(M)$ (dashed line) and RPSS_D (solid black line) plotted as function of ensemble size M for the cases listed in Table 1.

List of Tables

	reliable	weakly	very	weakly	very
		over-confident	over-confident	under-confident	under-confident
r	0.6000	0.5000	0.3000	0.7000	0.9000
r_f	0.6000	0.7000	0.9000	0.5000	0.3000

TABLE 1. Values of $r = r$ and r_f used in the numerical experiments.