Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Office of Meteorology and Climatology MeteoSwiss

Bias of ranked probability and Brier skill scores

Andreas P. Weigel, Mark A. Liniger, Christof Appenzeller
Federal Office of Meteorology and Climatology (MeteoSwiss), Switzerland

NCCR CLIMATE
Swiss Climate Research

Introduction

Probabilistic forecasts with ensemble prediction systems have found a wide range of applications in the context of weather and climate risk management. Among the most widely used probabilistic skill scores are the Brier and ranked probability skill scores (BSS and RPSS, respectively), which are based on a quadratic metric applied on a finite number of forecast categories. The skill scores quantify the degree to which a given ensemble prediction system outperforms a (typically climatologic) reference strategy. From earlier studies (e.g. Müller et al. 2005) it is known that the BSS/RPSS are substantially negatively biased for small ensemble sizes, imposing major problems on the verification of ensemble predictions. This deficiency can be overcome by considering the effects of finite ensemble size in the climatologic reference score. Here we present an analytical formula which quantifies the bias in dependence of category probabilities and ensemble size, thus allowing an easy-to-implement "bias-less" formulation of these skill scores (BSS_D and $RPSS_D$).

Debiasing the RPSS

The ranked probability score (RPS) and the corresponding skill score (RPSS) are defined as follows (e.g. Wilks 1995):

$$\mathrm{RPS} = \sum_{k=1}^{K} (Y_k - O_k)^2 = (\mathbf{Y} - \mathbf{O})^2 \qquad \mathrm{RPSS} = 1 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathrm{RPS}_{Cl} \rangle}$$

For a given forecast-observation pair, Y_k and O_k denote the k-th component of the cumulative probabilistic forecast and observation vectors **Y** and **O**, respectively. K is the number of forecast categories considered, and RPS_{Cl} is the RPS for the climatologic reference forecast. The brackets <...> denote the average of the scores over a large number of forecast-observation pairs. Note that the BSS is the special case of an RPSS with two forecast categories.

The RPSS is subject to a negative bias that is strongest for small ensemble size (Fig. 1, black line). Müller et al. (2005) showed that this bias can be removed when the score of the reference forecast is estimated by repeated random resampling from climatology, with the sample size being equal to ensemble size.

The process of randomly sampling an ***M*-member ensemble** from ***K* forecast categories** with **climatologic probabilities $p_1,\ldots,p_K$** is a multinomial process. Using the corresponding multinomial distribution, an analytical expression for a debiased version of the RPSS, the $RPSS_D$, can be derived (shown step by step in the **supplementary material** and in Weigel et al. 2006):

$$\mathrm{RPSS}_D = 1 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathrm{RPS}_{Cl} \rangle + D} \quad \textit{with} \quad D = \frac{1}{M} \cdot \sum_{k=1}^{K} \sum_{i=1}^{k} \left[p_i \cdot \left(1 - p_i - 2 \sum_{j=i+1}^{k} p_j \right) \right]$$

Special case 1:

K equiprobable forecast categories

$$D = \frac{1}{M} \cdot \frac{K^2 - 1}{6K}$$

Special case 2:

Brier score, that is two categories with probabilities p and (1-p)

$$D = \frac{1}{M} \cdot p \cdot (1 - p)$$

Example 1: White noise forecasts

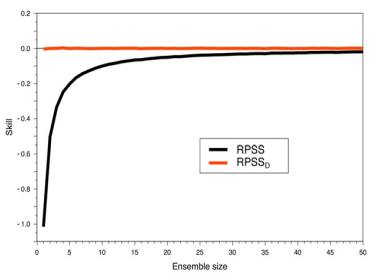

Figure 1: Expectation of the RPSS and the $RPSS_D$ for random synthetic white noise forecasts as a function of ensemble size. The skill scores are based on three equiprobable categories.

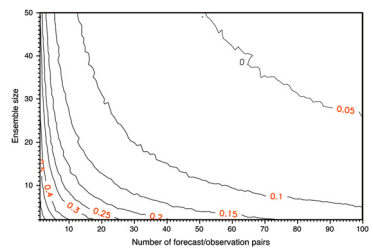

Figure 2: Upper 95% confidence levels for the $RPSS_D$ (estimated from 10000 $RPSS_D$ values) as a function of ensemble size and number of forecast-observation pairs.

Example 2: Debiased Brier skill score applied to real forecast data

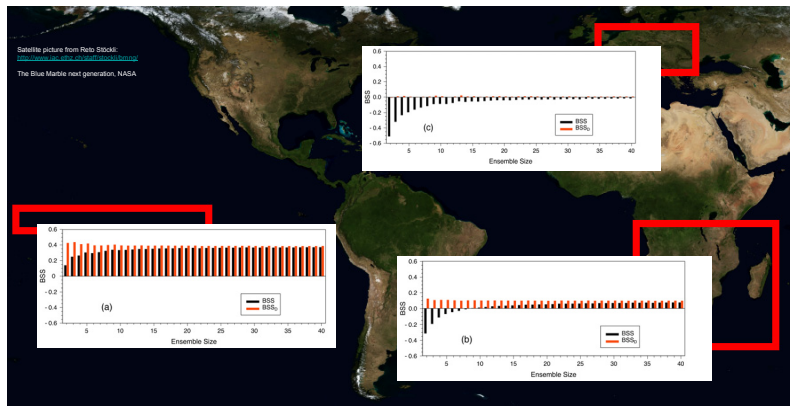

Figure 3: Brier skill score (BSS) and debiased Brier skill score (BSS_D) as a function of ensemble size for near-surface temperature predictions for March with a lead time of 4 months. Scores are averaged over 15 years (1988-2002) over the three regions indicated. Two equiprobable classes are used. The ECMWF Seasonal Forecast System 2 data are verified against ERA40 re-analysis.

Conclusions

RPSS and BSS are negatively biased for small ensemble sizes.

With an easy-to-implement analytical formula, this bias can be corrected and "bias-less" versions of the RPSS and BSS can be formulated.

Increasing ensemble size does not increase skill per se; rather, the statistical significance is enhanced.

References

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. Clim.*, **18**, 1513-1523
Wilks, D.S., 1995: *Statistical methods in the atmospheric sciences*. International Geophysics Series, Vol. 59, Academic Press, 467 pp.
Weigel, A.P., M.A. Liniger, C. Appenzeller, 2006: The discrete Brier and ranked probability scores, *Mon. Wea. Rev., accepted*

Contact

Andreas Weigel, MeteoSwiss
P.O. Box 514, CH-8044 Zürich
andreas.weigel@meteoswiss.ch

Manuscript to this poster: Weigel et al. 2006 (Download from http://www.meteoswiss.ch/nccr/weigel)