Verification of monthly and seasonal forecasts

MeteoSchweiz – Andreas Weigel, Daniel Baggenstos and Mark Liniger

MeteoSwiss has continued to use and verify seasonal forecasts of prediction System 2 and, since April 2007, System 3. Moreover, seasonal forecasts of the operational EUROSIP multi-model have been evaluated, as well as 32-days forecasts of the ECMWF monthly prediction system. Much effort has been put into the derivation of a suitable and unbiased skill metric for these studies.

3.1 Objective Verifcation

A skill score for small ensembles

An objective verification of probabilistic prediction systems is not trivial, since hindcast ensemble sizes are typically very small (monthly forecasts: 5 members; System 3: 11 members), which leads to inconsistencies of the skill metrics applied. This is particularly apparent for the widely used ranked probability and Brier skill scores (RPSS and BSS) which are strongly negatively biased and therefore severely underestimate predictive skill (Müller 2004, Müller et al., 2005).

A new technique has been developed to overcome this deficiency. By adequately considering the effects of finite ensemble size in the climatologic reference score, the negative bias can be removed analytically and a "debiased" version of the ranked probability skill score (RPSSd) and Brier skill score (BSSd) can be formulated (Weigel et al. 2007a). As an example, Figure 1 shows a comparison of the classical BSS and the new BSSd formulation for probabilistic seasonal forecasts over two regions in dependence of ensemble size. The new technique yields reasonable results even for systems with small ensemble sizes.

Multi-model verification

Motivated from the introduction of the operational EUROSIP multi-model seasonal forecasting system, the RPSSd skill metric has been generalized to weighted multi-model situations. Indeed, when verifying multi-model ensemble forecasts, it is important to distinguish the consequences of varying ensemble size from the true benefits of multi-model combination (Weigel et al. 2007b). Technically, this can be achieved by introducing the concept of an "effective" ensemble size characterizing the multi-model. In terms of multi-model verification, this RPSSd imposes a stricter skill criterion than the classical RPSS, since multi-models are not rewarded any more for their larger ensemble size. Nevertheless, also with the RPSSd metric, in most regions the multi-model outperforms any single-model strategy on average, even a "best-model-approach" (Figure 2).

Verification of ECMWF monthly forecasts

The RPSSd has also been used to verify the ECMWF monthly prediction system with full consideration of all forecast and hindcast data available (Baggenstos 2007a). Figure 3 shows the annual averaged prediction skill of 2-metre-temperature over Europe for forecast weeks 1 to 4. Verification has been carried out against ERA40 (before 2001) and the ECMWF operational analysis (after 2001), respectively. While forecast week 1 is skillful over the entire continent, forecast week 2 reveals significant skill only more in very few regions (notably the Mediterranean and the North Sea). For week 3 and 4, skill significant above 0 cannot be identified. If the four seasons are evaluated individually, some seasons reveal better skill for the second forecast week (Figure 4). For instance, forecasts in the Mediterranean and southern Europe are more skillful in autumn and winter than in spring in summer, while in the northern Europe the opposite is found.

Finally, a verification of 2m temperature with real station data has been carried out for 12 stations in Switzerland, covering all climatological regions of the country (Baggenstos 2007b). The resulting skill values have been compared to those obtained by verification against (re)analysis and lapse rate corrected operational analysis data (Figure 5). The scatterplots reveal a slight deterioration in prediction skill of about 20% if real station data are used. However, given the complex topography of Switzerland, this can be considered as a small reduction, thus justifying the use of ERA40 and operational analysis data as a proxy for real measurements in the present context confirming results from a direct comparison by Kunz et al. (2007).

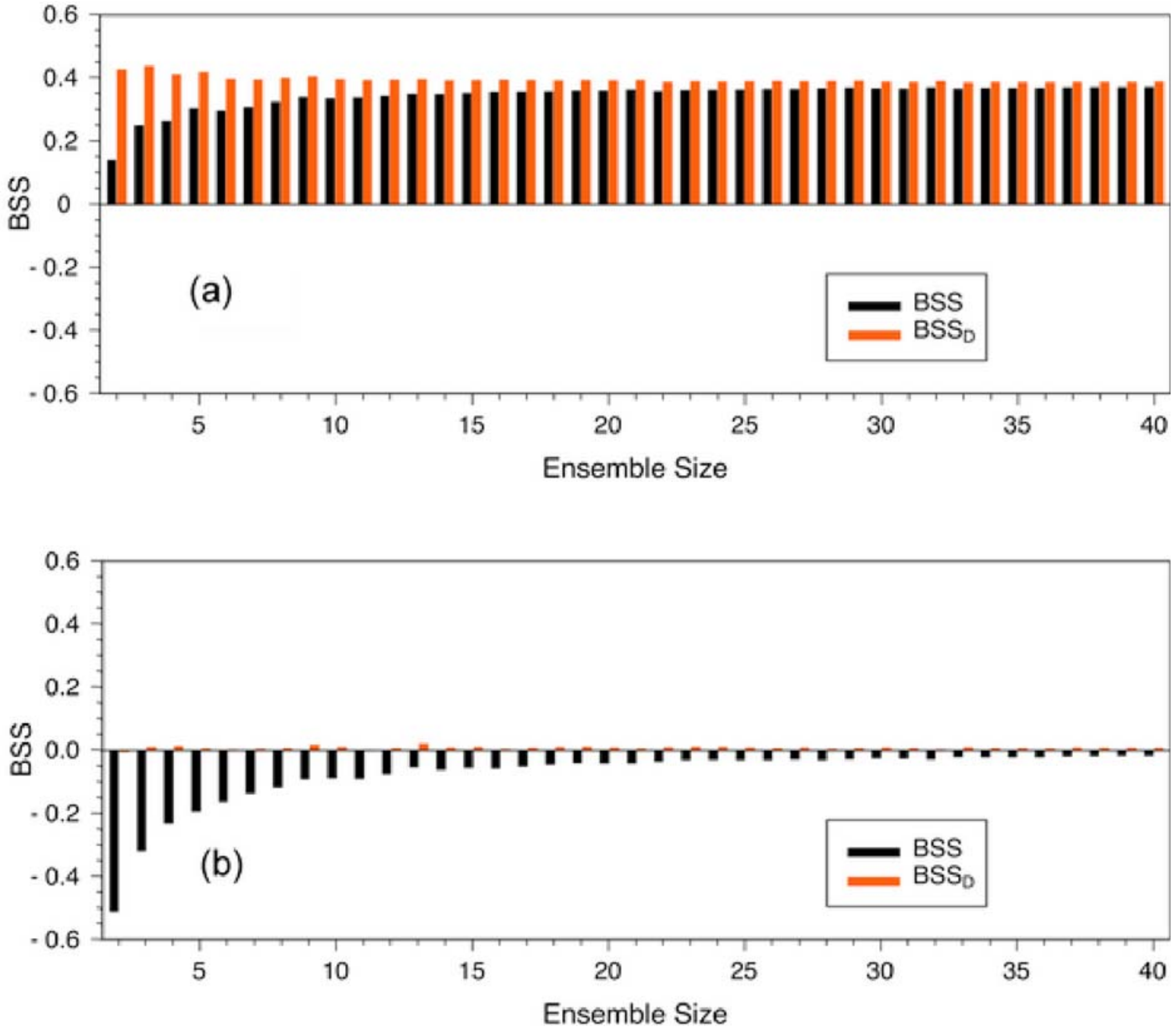

Fig. 1 Conventional Brier skill score (BSS, black) and new debiased Brier skill score (BSSD, red) as a function of ensemble size for near-surface temperature predictions for March with a lead time of 4 months. Scores are averaged over 15 years (1988-2002) over (a) the Nino3.4 region in the equatorial Pacific and (b) over Central Europe. The ECMWF System 2 data are verified against ERA40 re-analysis data.

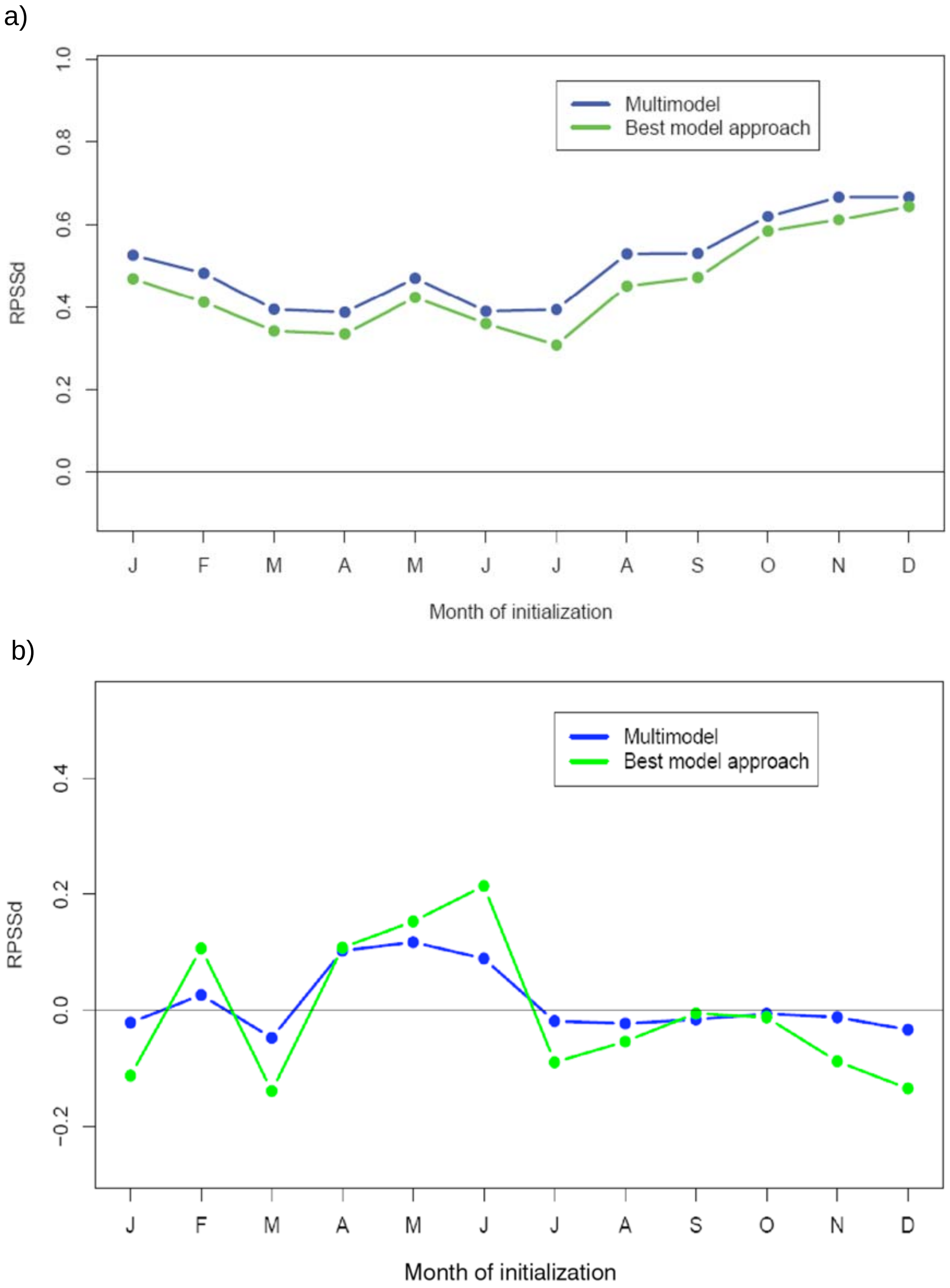

Fig. 2 Skill (RPSSd) of seasonal forecasts of 2m temperature (months 2-3-4) over (a) the Nino3.4 region and (b) Switzerland in function of initialization month. Blue line: Weighted multi-model, consisting of ECMWF System 2, Met Office GloSea2, and climatology. Weights are obtained by minimizing the ignorance score. Green line: Best model approach, i.e. for each forecasting month the respectively best participating single model is used. Verification is against ERA40 re-analysis data (before 2001) and the ECMWF operational analysis (after 2001).

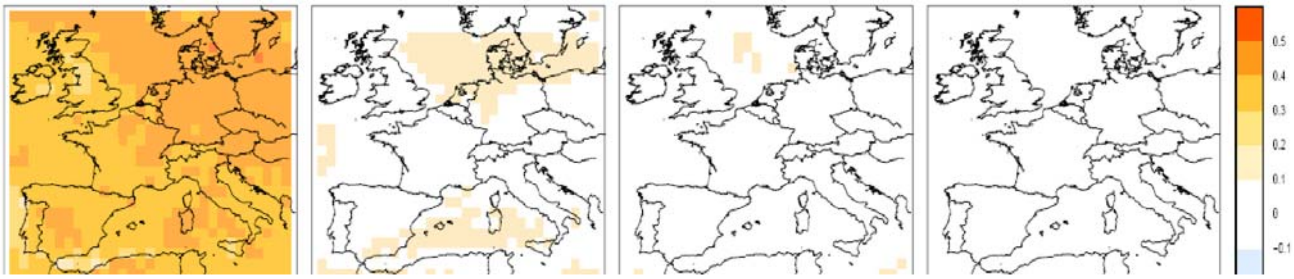

Fig. 3 Annual mean skill (RPSSd) of monthly forecasts of 2m-temperature for weeks 1 to 4 (left to right) in the European domain. Verification is against ERA40 re-analysis data (before 2001) and the ECMWF operational analysis (after 2001). From Baggenstos (2007a).

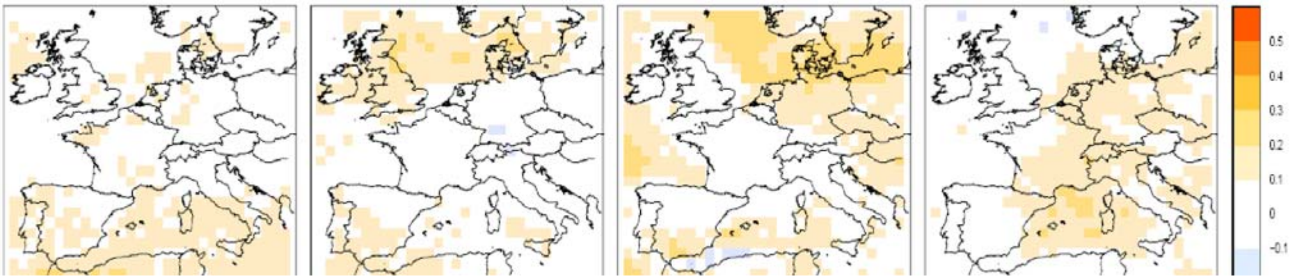

Fig. 4 Annual mean skill (RPSSd) of week 2 forecasts of 2m-temperature for winter, spring, summer and autumn (left to right) in the European domain. Verification is against ERA40 re-analysis data (before 2001) and the ECMWF operational analysis (after 2001). From Baggenstos (2007a)

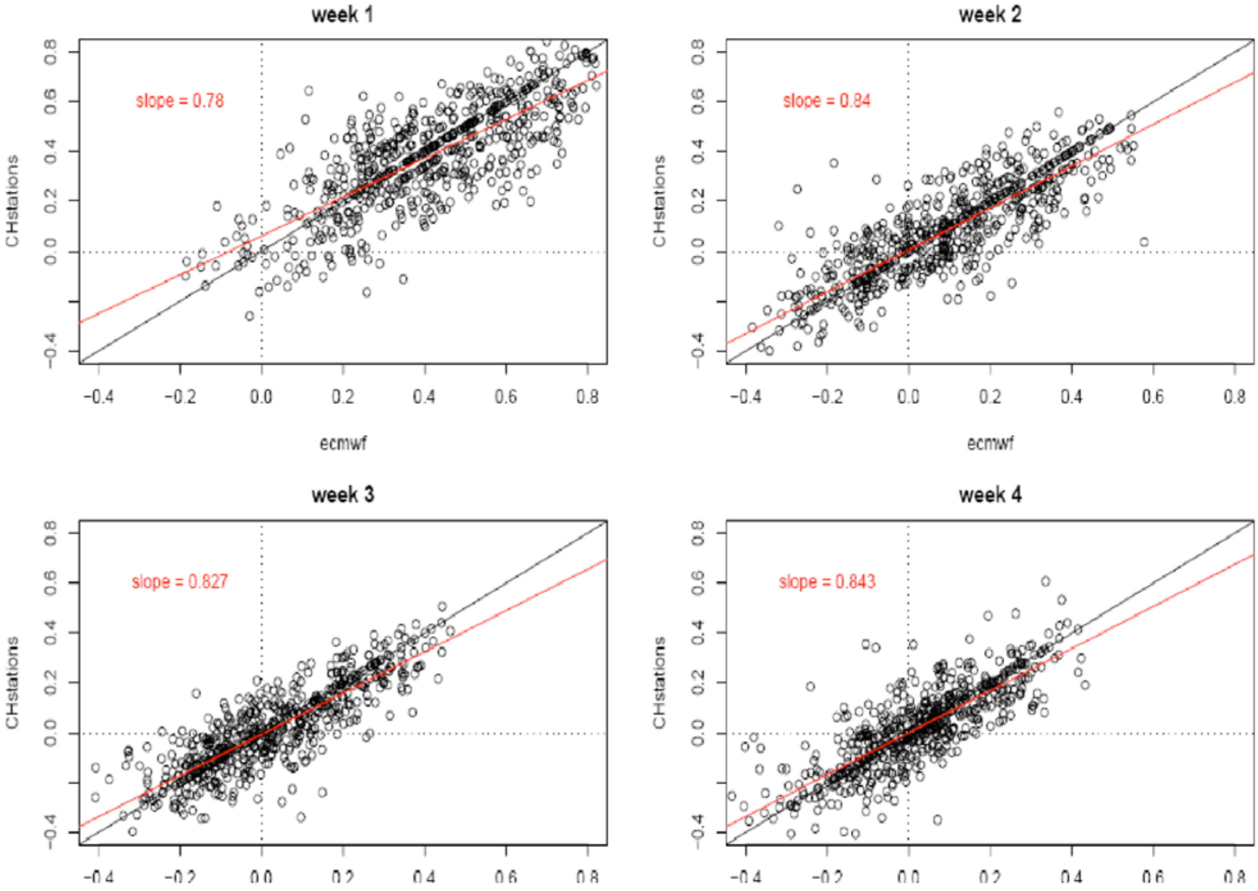

Fig. 5 Annual mean skill (RPSSd) of monthly forecasts of 2m-temperature at 12 representative station sites in Switzerland for weeks 1 to 4. Skill obtained from verification with measured temperatures is plotted against the corresponding skill values obtained from verification with ERA40 re-analysis data (before 2001) and the ECMWF operational analysis (after 2001). Black: Bisecting line. Red: regression line. From Baggenstos (2007b).

4. References

Baggenstos D., 2007a: Probabilistic verification of operational monthly temperature forecasts. *Veröffentlichung MeteoSchweiz* **Nr. 76**

Baggenstos D., 2007b: Verification of monthly temperature forecasts using Swiss observations, *Arbeitsbericht MeteoSchweiz,* in prep.

Müller, W., 2004: Analysis and Prediction of the European Winter Climate, Dissertation, ETH Zürich Nr. 15540, in *Veröffentlichung der MeteoSchweiz*, **Nr. 69**, pp101.

Müller W. A., C. Appenzeller, F. J. Doblas-Reyes, M. A. Liniger, 2005b: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes, *Journal of Climate*, **18** (10), 1513-1523.

Kunz, H.; Scherrer, S. C.; Liniger, M. A. and **Appenzeller, C.,** 2007: The evolution of ERA-40 surface temperatures and total ozone compared to observed Swiss time series. *Meteor. Z.*, **16(2)**, 171-181. DOI: 10.1127/0941-2948/2007/0183

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

Weigel A.P., Liniger M.A. and **C. Appenzeller.** 2007. Generalization of the discrete Brier and ranked probability skill scores for weighted multi-model ensemble forecasts. *Mon. Wea. Rev*. in press