Verification of monthly and seasonal forecasts

MeteoSwiss - Andreas Weigel and Mark Liniger

MeteoSwiss has continued to use and verify seasonal forecasts of prediction System 3 (both raw and recalibrated), seasonal forecasts of the operational EUROSIP multi-model, as well as 32-days forecasts of the ECMWF monthly prediction system. Most of our evaluations are based on the RPSSd, an unbiased probabilistic skill metric that has been specially derived for the evaluation of ensemble prediction systems with small sized hindcast ensembles. Additionally, the so-called "discrimination score", a new verification metric with a special appeal for administrative purposes, has been developed in collaboration with IRI and applied to System 3 forecasts.

1. Objective verification

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), making the forecasts inherently unreliable. This introduces a negative bias to probabilistic skill metrics that are sensitive to reliability, such as the Brier and ranked probability skill scores (BSS and RPSS). 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 (Müller 2004, Müller et al. 2005, Weigel et al 2007a, Weigel et al 2007b, Ferro et al. 2008). If not stated differently, the RPSSd is the skill metric of choice for the evaluations shown below.

Verification of ECMWF monthly forecasts

The RPSSd has been applied to verify the ECMWF monthly prediction system with full consideration of all forecast and hindcast data available (Baggenstos 2007, Weigel et al. 2008a). Figure 1 shows the annually averaged global prediction skill of probabilistic (terciles) 2-metre-temperature for forecast weeks 1 through 4. Verification has been carried out against ERA40 (before 2001) and the orography-adjusted ECMWF operational analysis (after 2001). Terciles of the climatology were smoothed. The evaluation shows that continental prediction skill during forecast week 1 (days 5-11 from initialization) is on the order of 0.4-0.6 in the extratropics and on the order of 0.2-0.3 in the tropics. In forecast week 2 (days 12-18), continental skill drops to much lower values of about 0.1-0.2. Beyond forecast week 2, pronounced prediction skill is predominantly found over the oceans, particularly over the ENSO region and in the central and eastern Pacific. Continental skill, however, essentially vanishes after forecast week 2, apart from some areas in tropical Africa and South America. While this result may appear disappointing, it also has a positive facet in that the RPSSd is generally not negative. In other words, in the worst case the ECMWF monthly forecasts of temperature are essentially identical to the climatological forecasts, but not worse.

In a second evaluation, the impact of the averaging interval has been investigated for Europe and the Nino3.4-region. For these regions, annually averaged prediction skill is now not only calculated for 7-days temperature averages, but also for 3-days and 14-days averages. The results are displayed in Fig. 2 as a function of the number of days between initialization and the end of the respective prediction interval. The picture shows that both for Europe and the Nino3.4-region skill increases as the length of the averaging interval becomes larger. This implies that the limit of predictability can be significantly exceeded into the future for users who do not require information in temporal resolutions as high as weekly.

A more thorough discussion of these findings, together with an in-depth evaluation of the prediction skill in function of lead-time, region and season is given in Weigel et al. (2008a).

Verification of the System 3

Similarly to the monthly forecasts, the RPSSd has been applied to evaluate the seasonal prediction skill of the System 3 seasonal prediction system, using all forecasts and hindcasts from 1987 through 2007. Fig. 3 shows the average prediction skill of seasonal forecasts of 2-metre-temperature in Europe for winter (DJF), spring (MAM), summer (JJA) and autumn (SON). Lead-time has been one month. Again, the verification has been carried out against ERA40 (before 2001) and the ECMWF operational analysis (after 2001), respectively. The evaluation in Fig. 3 shows that significantly positive prediction skill can be found only (i) in spring over central Europe, (ii) in summer over southern Europe and (iii) in autumn over the British Isles. Particularly the negative skill over northern Europe in winter and even more in summer is pronounced and surprising; it implies that the use of System 3 forecasts in these regions can yield worse results than if one simply guessed on the basis of climatology. Such negative skill can be induced if the forecasts are overconfident, i.e. if the forecast ensembles are too sharp while being centered at the wrong value (Weigel et al. 2008b). We have considered two methods to improve the prediction skill, namely (i) multi-model combination and (ii) recalibration, which will be discussed in the following.

The effect of multi-model combination and recalibration

Weigel et al. (2008b) have shown that the main effect of multi-model ensemble combination (MMEC) is to reduce the overconfidence of ensemble forecasts and to reduce the random error of the ensemble mean. We have tested the effect of MMEC and combined forecasts of the ECMWF System 3 model with forecasts of the UK Met Office GloSea2 of the EUROSIP dataset. Summer forecasts as in Fig. 3c have been considered, and the evaluation of the multi-model is shown in Fig. 4. The skill gain with respect to System 3 alone is clearly visible, particularly over the region of negative skill in northern Europe. Adding more models and applying a weighting scheme according to previous performance can further improve the skill (Weigel et al. 2008c).

A second approach to make unreliable forecasts more reliable is the application of an appropriate recalibration scheme, as described in Doblas-Reyes et al. (2005) and Weigel et al. (2008c). In such recalibration approaches, appropriate rescaling factors are derived from a sufficiently long set of hindcasts and corresponding verifications. Using these rescaling factors, overconfident ensemble forecasts can be shifted and inflated such that they become reliable. An application of this technique to seasonal summer temperature forecasts of System 3 (as in Fig. 3c) are shown in Fig. 5. Most notably, the negative skill over Northern Europe is seen to be totally removed or even positive if the forecasts are recalibrated. On the other hand, some of the high skill over Italy is reduced in comparison with the non-recalibrated forecasts of Fig. 3c. This "signal dilution" is a typical effect when recalibration is applied to regions of high predictability (Weigel et al. 2008c).

Application of a new verification concept to seasonal forecasts

In collaboration with Simon Mason of the International Research for Climate and Society (IRI), Columbia University, USA, a new and generic verification framework, the so-called "discrimination score", has been developed. It is motivated from the fact that so far only little attention has been given to scores that may be useful for administrative reasons, such as communicating changes in forecast quality to bureaucrats, and providing indications of forecast quality to the general public. The "discrimination score" can be applied to essentially all forecasting contexts, including both deterministic and probabilistic forecasts. It has a broad intuitive appeal in that the expected skill of an unskilled set of forecasts is 50%, as well as being interpretable as an indication of how often the forecasts are "correct". A detailed description of this verification concept is provided in Mason and Weigel (2008). We have applied the discrimination score on the prediction context of Fig. 3c, i.e. on the seasonal summer temperature forecasts of System 3. The resulting skill map (Fig. 6) has now the advantage that it can be more easily interpreted by non-specialists than if the RPSSd is used. In essence, the numbers quantify the probability, that the observed outcomes can be correctly discriminated on the basis of the forecasts, with 50% being the skill one would obtain by random guessing.

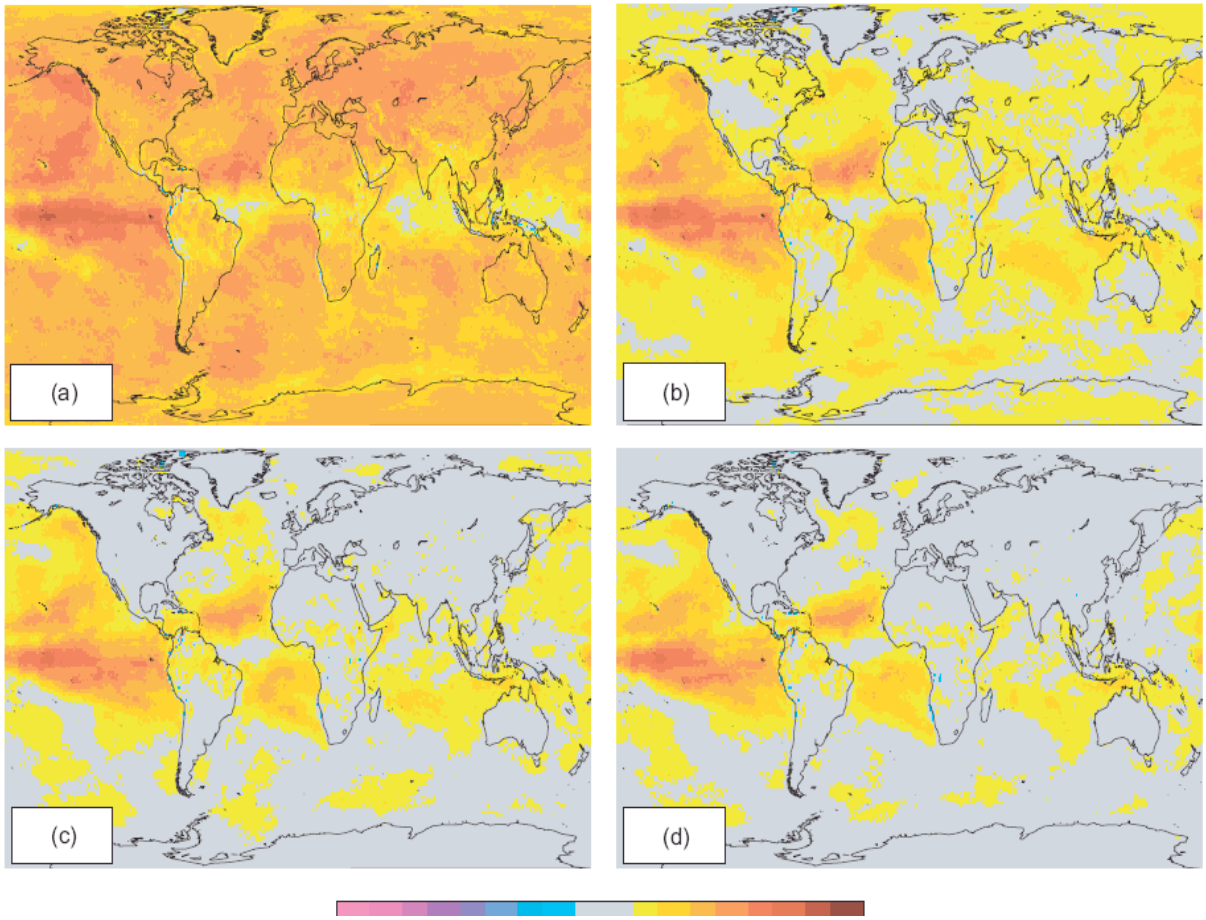

Fig. 1 Annual mean skill (RPSSd) of all available monthly forecasts (and hindcasts) of 2m temperature for forecast weeks 1 to 4 (a-d). Verification is against ERA40 re-analysis data (before 2001) and the ECMWF operational analysis (after 2001). From Weigel et al. (2008a).

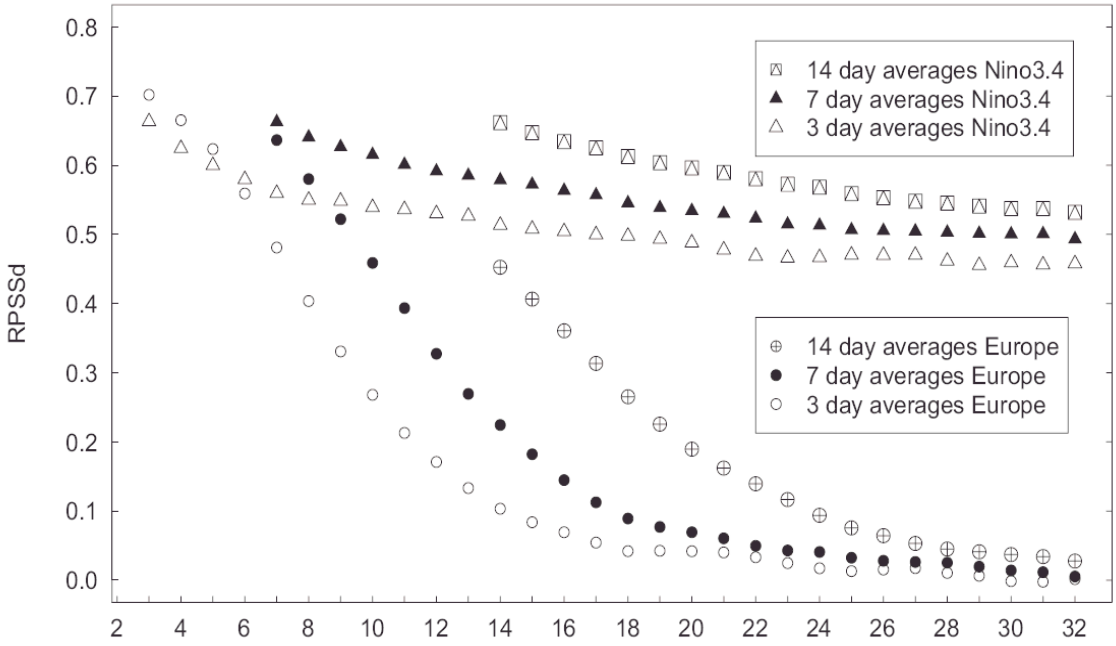

Fig. 2 Dependence of temperature prediction skill (RPSSd) on the length of the forecast averaging interval. Annually averaged skill is plotted against the prediction time, i.e. the time between initialization and the end of the prediction interval, for the Nino3.4 region and continental Europe. Forecast averaging intervals of 3, 4 and 14 days are considered. From Weigel et al. (2008a).

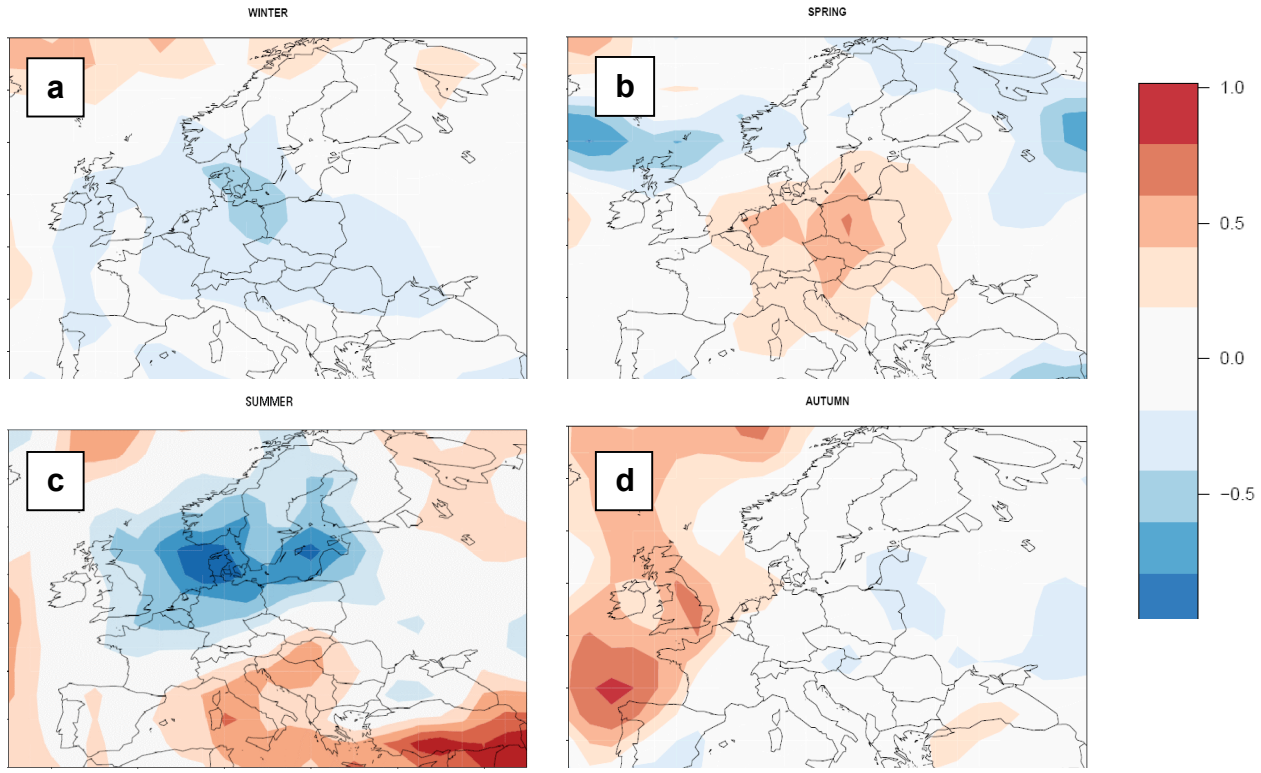

Fig.3 Prediction skill (RPSSd) of System 3 forecasts (lead-time one month) for seasonal temperature averages in Europe in (a) winter, (b) spring, (c) summer and (d) winter. The years 1987 through 2007 have been evaluated.

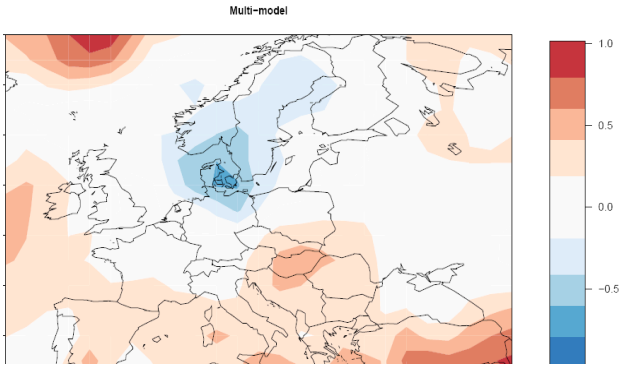

Fig.4 Prediction skill (RPSSd) of a simple multi-model for summer temperature averages in Europe. The multi-model consists of forecasts from the ECMWF's System 3 and the UK Met Office's GloSea 2, obtained from the EUROSIP data-set. The skill gain with respect to the System 3 alone (Fig. 3c) is clearly visible. Adding more models and applying a weighting scheme according to previous performance can further improve the skill (Weigel et al. 2008b).

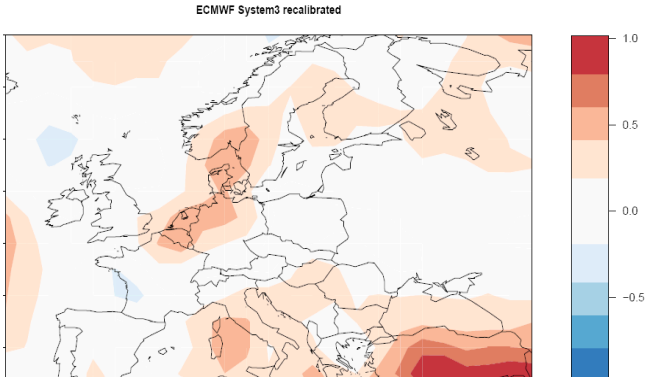

Fig.5 Prediction skill (RPSSd) of recalibrated System 3 summer forecasts in Europe. The average skill gain with respect to non-recalibrated forecasts (Fig. 3c) is evident (see also Weigel et al 2008c).

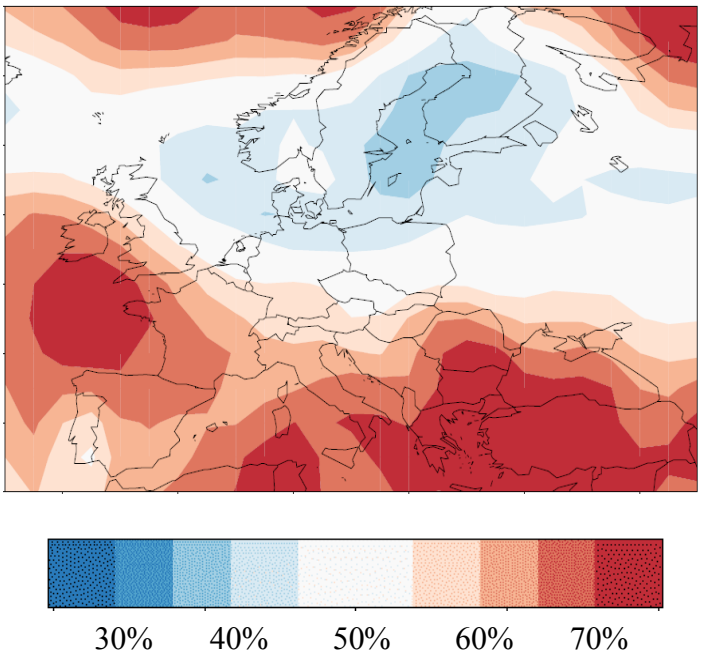

Fig.6 As Fig. 3c, but using the discrimination score of Mason and Weigel (2008) rather than the RPSSd.

2 References to relevant publications

Baggenstos D., 2007: Probabilistic verification of operational monthly temperature forecasts. *Veröffentlichung MeteoSchweiz,* **76**, 52 pp.

Ferro C.A.T., Richardson D.S. and **A.P. Weigel**, 2008. On the effect of ensemble size on the ranked probability and multi-category Brier scores. *Met. Apps.* **15**, 19-24

Mason S.J. and **A.P. Weigel.** 2008. A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.* accepted

Müller, W., 2004: Analysis and Prediction of the European Winter Climate, Dissertation, ETH Zürich Nr. 15540, in *Veröffentlichung der MeteoSchweiz*, **69**, 101 pp.

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, *Journal of Climate*, **18**, 1513-1523.

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

Weigel A.P., Liniger M.A. and **C. Appenzeller**, 2007b. Generalization of the discrete Brier and ranked probability skill scores for weighted multi-model ensemble forecasts. *Mon. Wea. Rev*. **135**, 2778-2785

Weigel A.P., **Baggenstos D**., **Liniger M.A., Vitart F.** and **C. Appenzeller.** 2008a. Probabilistic verification of monthly temperature forecasts. *Mon. Wea. Rev*., accepted

Weigel A.P., Liniger M.A. and **C. Appenzeller.** 2008b. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Met. Soc*., **134**, 241-260

Weigel A.P., Liniger M.A. and **C. Appenzeller.** 2008c. Seasonal ensemble forecasts: Are recalibrated single models better than multi-models? *to be submitted*