Probabilistic Verification of COSMO-LEPS Forecasts with SYNOP Stations

André Walser and Mark A. Liniger

MeteoSwiss, Krähbühlstrasse 58, 8044 Zürich, Switzerland

1 Introduction

A probabilistic verification of the limited-area ensemble prediction system COSMO-LEPS is carried out for spring and summer precipitation. COSMO-LEPS is the operational limited-area ensemble prediction system (EPS) of the Consortium for Small-Scale Modeling (COSMO) running daily at ECMWF since November 2002. Since February 2006, COSMO-LEPS is a 16-member ensemble based on version 3.17 of the non-hydrostatic limited-area COSMO model (formerly LM; see Steppeler et al., 2003) using a horizontal grid-spacing of 10 km. Previous verification efforts of ARPA-SIM, Bologna (Marsigli et al., 2005; Marsigli et al., 2006) compare upscaled high-density observations of the COSMO member states Germany, Greece, Italy, Poland and Switzerland with forecasts of COSMO-LEPS and the European Centre for Medium-Range Weather Forecasting (ECMWF) EPS on boxes of 1.5×1.5 degrees.

This study assesses the COSMO-LEPS against observation of European SYNOP stations. The skill of COSMO-LEPS precipitation forecasts at a local scale is evaluated and analyzed with regard to the overall skill, the skill in complex topography and the spatial variability. This report is organized as follows: Section 2 presents the verification strategy including a brief introduction to the verification metric used. The results are given in Section 3 including a detailed skill analysis for spring 2006 (3.1), a comparison with the skill of deterministic forecasts using a higher resolution (3.2), a comparison with other verification periods (3.3), and finally an analysis of the skill in the complex topography of the Alpine area (3.4). Conclusions are provided in Section 4.

2 Verification method

The verification is based on accumulated precipitation between 0600-1800 UTC and 1800-0600 UTC, referred to as daytime and nighttime precipitation hereafter. It is carried out on the common model domain of COSMO-LEPS and the Swiss implementation of the COSMO model using a horizontal grid-spacing of 7 km, referred to hereafter as aLMo. The model domains are illustrated in Fig. 1. The thresholds considered are 1 mm, 5 mm, 10 mm and 25 mm precipitation, respectively. For the results presented here, the predicted probabilities are derived from the average precipitation of the five grid points closest to the station location. Investigations reveal no significant differences in overall scores using an average of five, nine or just the closest grid point. COSMO-LEPS probabilities are derived by equally weighting the ensemble members, since earlier verification studies for precipitation pointed out slightly worse scores for probabilities derived by weighting the members according to the size of the cluster they represent (Marsigli 2004, personal communication). This might be related to the bias of the Brier Skill Score (see below) for weighted ensembles (Weigel et al., 2007b).

In order to evaluate the predicted probabilities for an event, we use the Brier score (BS) as proposed by Brier (1950) which measures the square of the differences between predicted

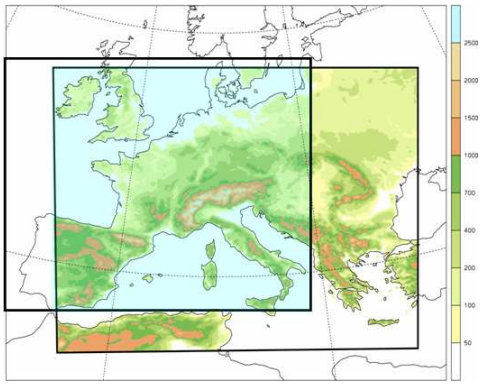


Figure 1: COSMO-LEPS domain with topography (color coded) and aLMo domain. The verification is performed on the common area (transparent light blue).

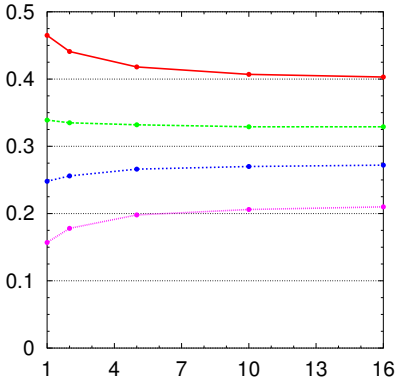


Figure 2: Debiased Brier skill score for 12h-accumulated precipitation exceeding 1 mm for spring 2006 as a function of ensemble size for lead-time +18h (red line), +42h (green), +66h (blue), and +90h (purple), respectively.

and observed probabilities. The BS can be decomposed into the scalar attributes reliability, resolution, and uncertainty (Murphy 1973):

$$\mathrm{BS} = \underbrace{\frac{1}{N}\sum_{k=1}^{J} N_k \left(y_k - \overline{o_k}\right)^2}_{reliability} - \underbrace{\frac{1}{N}\sum_{k=1}^{J} N_k \left(\overline{o_k} - \overline{o}\right)^2}_{resolution} + \underbrace{\overline{o}\left(1-\overline{o}\right)}_{uncertainty} \tag{1}$$

where N is the number of forecast-observation pairs, J the number of probability classes k, y the predicted probability, and $\overline{o}$ the relative frequency of the event in the verification period. The reliability (REL) term indicates how closely the forecast probability matches the observed event frequency as a function of forecast probability; the smaller the reliability value, the better. The resolution (RES) term measures the ability of the forecasting system to resolve the set of events into subsets with characteristically different observed frequencies. The third term, uncertainty (UNC), is independent from the forecasting system and only a function of the event frequency.

Forecast skill is measured by a comparison of forecast score with the score from a reference forecast. Therefore, the Brier Skill Score (BSS) is defined as $BSS = 1 - BS/BS_{ref}$. The BSS has a maximum of one (if BS equals zero). A BSS value below zero means no skill, i.e. the forecast is worse than the reference forecast. As reference forecast we use climatology. A climatological forecast has no resolution by definition and if the event frequency in the verification period is equal to the climatological probability, then the reliability terms becomes also zero, and hence $BSS = (RES - REL)/UNC$, i.e. the forecasts have skill if $RES > REL$. The station climatology is estimated from the observations in the correspond-

	MAM	JJA
1800-0600 UTC	735	810
0600-1800 UTC	680	806

Table 1: Number of SYNOP stations used for the verification from totally 1273 stations with a sufficient availability to build a climatology.

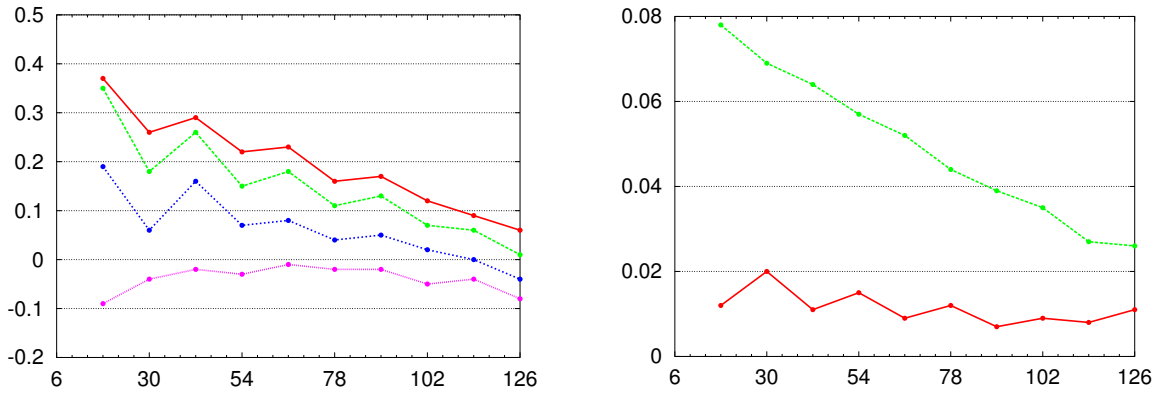


Figure 3: Verification results of COSMO-LEPS 12-h-accumulated precipitation forecasts for spring 2006 as a function of forecast range [h]. The left panel shows the Brier skill score for thresholds 1 mm (red solid line), 5 mm (green), 10 mm (blue), and 25 mm (purple), respectively. The right panel shows reliability (red solid line) and resolution (green) for the 1 mm threshold.

ing season of the past 10 years (1997-2006) for both verification times, requesting a minimal availability of 50% of the SYNOP messages. This reduces the total number of stations from 1273 inside the verification domain to about 750 depending on the season and the verification time. The exact numbers are summarized in Table 1.

From Müller et al. (2005) it is known that the BSS is negatively biased for EPSs with small ensemble sizes. Weigel et al. (2007a) gives an analytical formula to derive a biased corrected version

$$BSS_D = 1 - \frac{BS}{BS_{Clim} + D} \quad \text{with a correction term} \quad D = \frac{1}{M}\overline{o}(1-\overline{o}) \tag{2}$$

identified as the intrinsic unreliability of the forecasting system due to the limited number of ensemble members M. Figure 2 shows the BSS_D as a function of ensemble size for the 1 mm thresholds for different lead-times. The values are a result of 100 random samples. The skill decreases considerably with lead-time but is positive for all lead-times and ensemble sizes. In addition to the bias discussed above for the BSS, we note a secondary bias for small ensemble size, which is positive for short lead-times and negative for long lead-times. The positive bias for lead-time +18h indicates, that the ensemble is overconfident for that lead-time as explained in Weigel et al. (2007b).

3 Results

3.1 Verification for spring 2006

We begin the discussion of the verification results with spring 2006 (March, April and May). Fig. 3a presents the BSS as a function of forecast range for the thresholds 1 mm, 5 mm, 10 mm, and 25 mm, respectively. For the two lowest thresholds, the forecasts have skill until the end of the forecast range which is also valid for the 10 mm threshold except for the very end of the range. While the skill decreases as expected with increasing lead-time, we note a significant diurnal cycle for these lowest three thresholds reflecting a higher skill for precipitation during nighttime. This behavior is most probably related to the limited

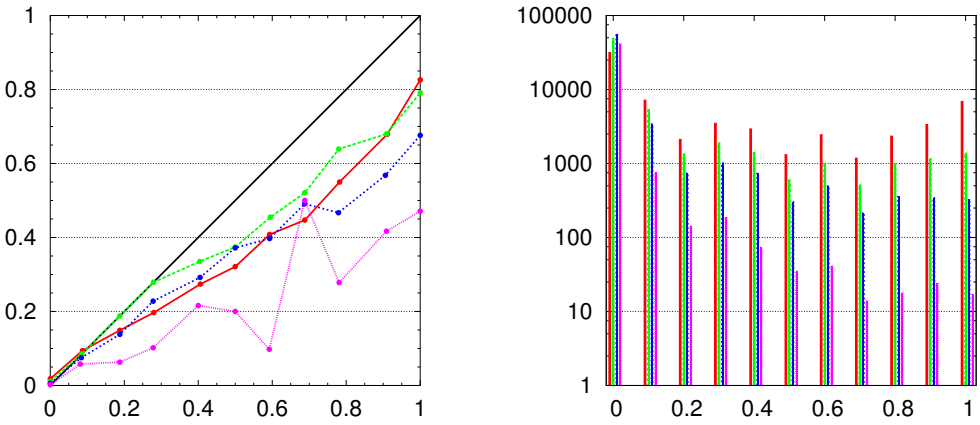


Figure 4: COSMO-LEPS 12h-accumulated precipitation in spring 2006 for the forecast range +18h for precipitation exceeding 1mm (red solid line), 5 mm (green), 10 mm (blue), and 25 mm (purple). The left panel shows the reliability diagram with forecast probability on the x-axis and observed frequency on the y-axis. The right panel indicates the number of forecast-observation pairs in each probability class (note the logarithmic scale).

predictability of convection (e.g. Walser et al., 2004) which occurs more frequently during daytime. Finally, the panel reveals a negative BSS, i.e. no skill, for precipitation exceeding 25mm/12h for all lead-times which is further discussed later on.

The reliability and the resolution term are shown in Figure 3b as a function of forecast range for the 1 mm threshold. While the reliability is quite small and rather constant with lead time, the resolution decreases almost linearly with increasing forecast range, which is the normal behavior of every weather forecasting system due to decrease in predictability with increasing forecast range.

A so-called reliability diagram for the lead-time +18h is derived presenting the reliability qualitatively for the four thresholds considered (Fig. 4a). The predicted probabilities are divided into 11 probability classes (<5%, 5-15%, 15-25%,..., >95%). The sample size for each class and threshold is given in Fig. 4b. For all thresholds, the curves lie below the diagonal (black line), in particular for classes with high predicted probabilities. Hence, COSMO-LEPS overestimates the occurrence of precipitation events when it predicts rather high probabilities, particularly for the highest threshold 25 mm. However, it should be noted that the sample sizes are small for this threshold (except for the first class of course), in particular for probabilities higher than 60% (see Fig. 4b). The reliability diagrams look similar for other lead-times except for classes with high probabilities due to the decrease of sample size for those classes with increasing forecast range.

In addition, Fig. 4 shows event resolution for all four thresholds, even for the 25 mm threshold which reveals no skill. This points to a weakness of the BSS with climatological forecasts as reference forecast as discussed in Mason (2004): if the BSS is negative, the forecasts are assumed unskillful, but there might nevertheless be some useful event resolution in the forecast that can be used after a calibration (a simple example is discussed later in section 3.4). In our case, the forecast for precipitation exceeding 25 mm would be nevertheless more valuable than a climatological forecast for a customer with a rather low cost-loss ratio (Richardson 2000) taking action already at rather low predicted probabilities.

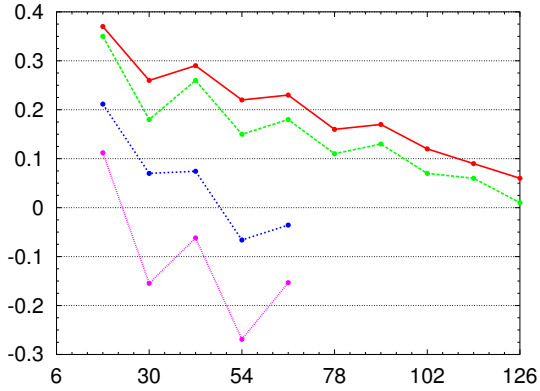


Figure 5: As Fig. 3a, but a comparison between skill for COSMO-LEPS (red and green lines) and aLMo (blue and purple) forecasts for precipitation exceeding 1 mm (red and blue) and 5 mm (green and purple), respectively.

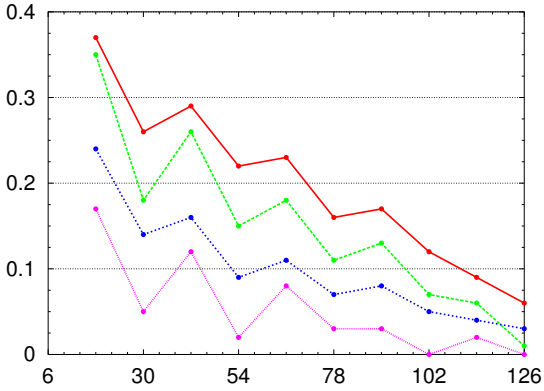


Figure 6: Comparison of skill for 12h-accumulated precipitation between spring and summer 2006 in terms of Brier skill score. Red and green lines indicate skill for threshold 1 mm and 10 mm, respectively, for spring, blue and purple lines indicate skill for the same thresholds for summer.

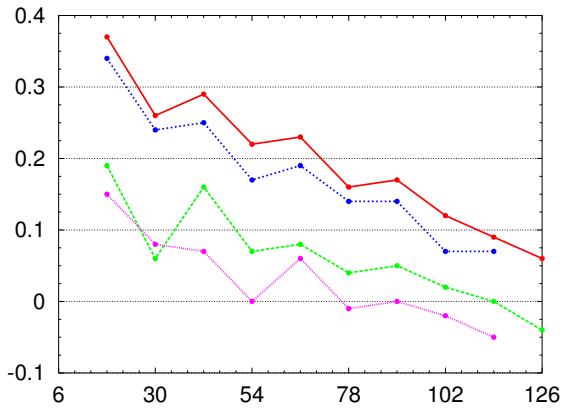
3.2 Comparison with aLMo forecasts

The deterministic model aLMo can be considered as one-member ensemble predicting the probabilities 0% or 100%, respectively, for up to +72h. The BSS of aLMo for spring 2006 and thresholds 1 and 5 mm, respectively, is compared with COSMO-LEPS in Fig. 5. The skill of the aLMo forecasts is clearly lower for all lead-times with skill only in the short-range for the 5 mm threshold and up to +42h for the 1 mm threshold. Considering BSS_D, then aLMo clearly outperforms COSMO-LEPS (not shown). For the 1 mm thresholds, BSS_D for aLMo is in the range 0.58 (+18h) and 0.44 (+54h). The largest differences in the COSMO setup between aLMo and COSMO-LEPS is the assimilation cycle of aLMo (relevant for short lead-times) and the horizontal resolution as mentioned above. The clearly higher skill of aLMo in terms of the BSS_D highlights a potential to significantly improve COSMO-LEPS forecast with a higher horizontal resolution.

3.3 Comparison with different verification periods

In this section, we compare the verification results for spring 2006 with those for summer 2006 and spring 2005. Figure 6 presents the BSS for the thresholds 1 mm and 5 mm for spring and summer 2006, respectively. For both thresholds, the skill is clearly higher for spring, which is most probably related to the larger fraction of convective precipitation events in summer than in spring.

Investigating spring 2005 and 2006, two different setups of COSMO-LEPS can be compared. Before February 2006, the COSMO-LEPS ensemble was a 10-member ensemble using COSMO version 3.15 and a coarser vertical resolution (32 levels, new setup has 40). At the same time a new cycle of the driving EPS was introduced at ECMWF with a higher horizontal (from T255 to T399) and vertical resolution (from 45 to 61 levels). Figure 7 shows a comparison in terms of BSS (left) and BSS_D (right) for the thresholds 1 mm and 10 mm, respectively. In general, the BSS is higher for spring 2006, while the BSS_D shows only marginal differences. Thus, the better skill for spring 2006 is mainly a result of the larger



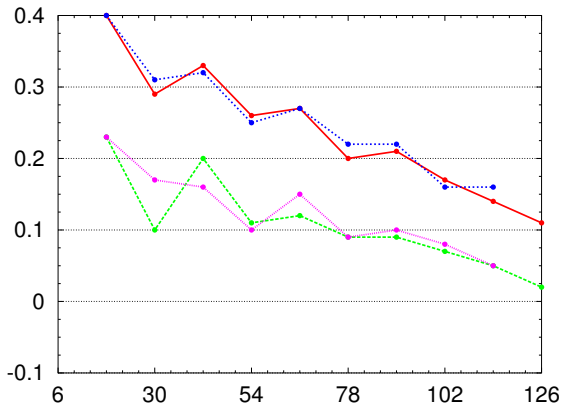


Figure 7: Comparison of skill for 12h-accumulated precipitation between spring 2006 and 2005 in terms of Brier skill score (left) and debiased Brier skill score (right). Red and green lines indicates skill for threshold 1 mm and 10 mm, respectively, for 2006, blue and purple lines indicates skill for the same thresholds for 2005.

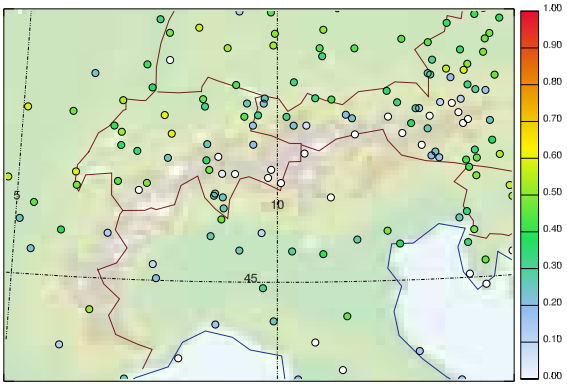


Figure 8: Brier skill score for 12h-accumulated precipitation exceeding 1 mm for spring 2006 and lead-time +42h at SYNOP stations in the Alpine region.

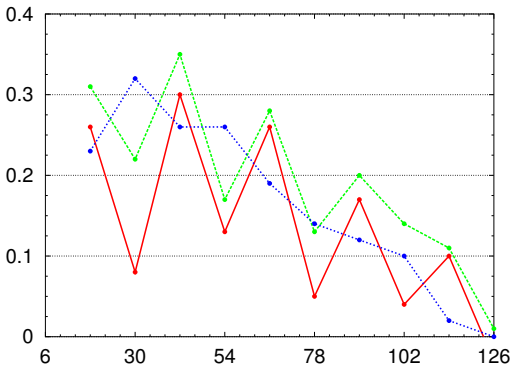


Figure 9: Same as Fig. 3a, but for Swiss stations only and without 25 mm threshold (sample size too small).

ensemble and not due to the improvements of the model setups of COSMO-LEPS and the driving ECMWF EPS.

3.4 Forecast skill in the complex topography of the Alps

In this section, we focus on COSMO-LEPS forecast skill in the complex topography of the Alpine region. The spatial variability of the skill is very large in this region (Fig. 8). Some stations show very good skill with BSS higher than 0.5, while a few stations, mainly located in inneralpine valleys, show no skill (white circles). The BSS derived with Swiss stations only (39) is presented in Fig. 9. As for the entire verification domain, we found skill until the end of the forecasting range for the thresholds 1 mm, 5mm, and 10 mm, but the skill for the different thresholds are closer together with a better skill for 5 mm than for 1 mm for all lead-times. Due to the small sample size, the 25 mm threshold is not investigated for this limited number of stations.

In the following, we analyze the skill for station Sion in the Rhone valley which reveals no

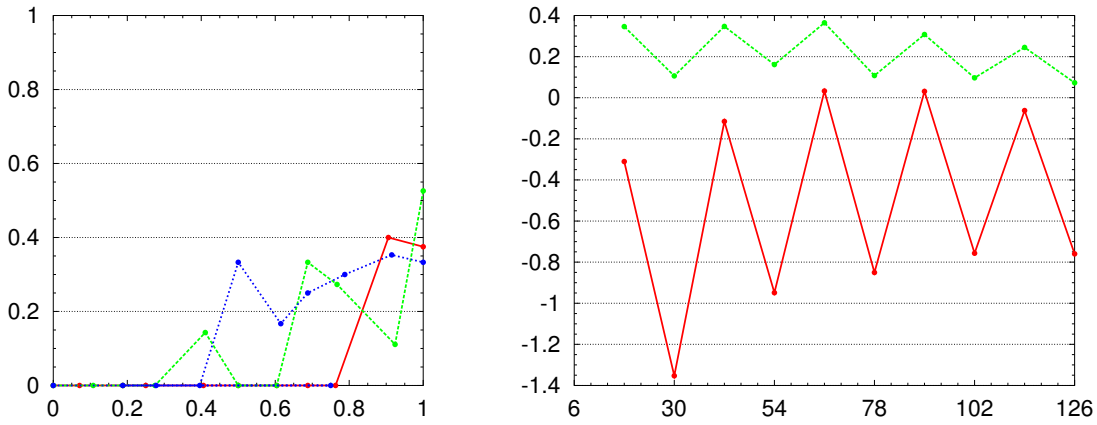


Figure 10: Verification for SYNOP station Sion in the Swiss Rhone valley for 12h-accumulated precipitation exceeding 1mm. The left panel shows the reliability diagram for lead-time +30h (red), +54h (green), and +78h (blue). Brier skill score (red) and Brier skill score of the calibrated forecast (green) is indicated in the right panel.

skill for daytime precipitation exceeding 1 mm for all lead-times and no or very low skill for nighttime precipitation, indicated as red solid line in Fig. 10b. The reliability diagram in Fig. 10a for lead-times +30h, +54h, and +78h shows that COSMO-LEPS dramatically overestimates the probability for all classes, resulting in a very low reliability. In order to investigate the impact of a postprocessing on the skill, an ad-hoc calibration is applied by multiplying the predicted probabilities: $y^* = 0.5y$. The factor 0.5 is chosen according to the reliability diagram that indicates slopes of about 0.5 for the three lead-times. The BSS using the calibrated probabilities y^* for the 1 mm threshold is indicated in Fig. 10b as green line. It is not only positive for all lead-time, it is even higher than the average skill for the Swiss stations (cf. Fig. 9).

4 Conclusion

An objective and comprehensive probabilistic verification of COSMO-LEPS 12h-accumulated precipitation has been carried out for spring and summer 2006. Overall, the results show forecast skill in terms of the Brier skill score until the end of the 5.5 days forecasting range for precipitation exceeding 1 mm, 5 mm, and 10 mm, respectively, while the BSS for 25 mm turned out to be negative for all lead-times. A comparison with results for 2005 reveals that the scores have been improved since 2005, but mainly due to the increase of ensemble size from 10 to 16 members in February 2006, while changes in the model setups do not show a significant impact on the scores. In addition, skill for spring is considerably higher than for summer precipitation, most probably related to the limited predictability of convective events.

Further investigations indicate that the ensemble is overconfident for short-lead times, i.e. has a too small spread. The skill scores exhibit a high spatial variability, in particular in complex topography where a few stations with very low or no skill are found. It is demonstrated that even an ad-hoc calibration of the predicted probabilities has a large potential to improve the skill for such stations.

This study has some notable limitations. It focuses only on weak to moderate precipitation since the sample size using seasonal verification periods is too small for larger thresholds than

25mm/12h. In addition, the verification is based only on spring and summer precipitation. In a follow-up study, the other seasons will be investigated too, including the interannual variability. In addition, the use of larger time-series will allow to examine the skill of COSMO-LEPS forecasts for extreme events which is of particular interest.

Acknowledgments

This study was supported by the Swiss NSF through the National Centre for Competence in Research Climate (NCCR-Climate).

References

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *J. Atmos. Sci.*, 78, 1–3.

Marsigli, C., F. Boccanera, A. Montani, and T. Paccagnella, 2005: The COSMO-LEPS mesoscale ensemble system: Validation of the methodology and verification. *Nonlinear Processes in Geophysics.*, 12, 527–536.

Marsigli, C., A. Montani, and T. Paccagnella, 2006: Verification of the COMO-LEPS new suite in terms of precipitation distribution. *COSMO Newsletter*, No. 6, 143–141.

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., C. Appenzeller, F. Doblas-Reyes, and M. A. Liniger, 2005: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Weather and Forecasting*, 17, 173–191.

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

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, 126, 649–667.

Steppeler, J., G. Doms, U. Schättler, H.-W. Bitzer, A. Gassmann, U. Damrath, and G. Gregoric, 2003: Meso-gamma Scale Forecasts using the Nonhydrostatic Model LM. *Meteorol. Atmos. Phys.*, 82, 75–96.

Walser, A., D. Lüthi, and C. Schär, 2004: Predictability of Precipitation in a Cloud-Resolving Model. *Mon. Wea. Rev.*, 132, 560–577.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007a: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.*, accepted.

—, 2007b: Generalization of the discrete Brier and ranked probability skill scores for weighted multi-model ensemble forecasts. *Mon. Wea. Rev.*, accepted.