COSMO–LEPS Verification: First Results.

C. Marsigli, F. Boccanera, A. Montani, F. Nerozzi, T. Paccagnella

ARPA-SIM, Bologna, Italy

Abstract

An objective probabilistic verification of the Limited–area ensemble system COSMO–LEPS is being carried out at ARPA–SIM. In particular, results for the 24–cumulated precipitation is considered in this work. Forecast values are compared against observations covering Northern Italy, but results for a bigger part of the COSMO–LEPS domain (covering also Germany and Switzerland) are presented. COSMO–LEPS is compared with ECMWF ensembles and different verification techniques are used. Finally, an analysis of the performances of the system at different spatial scales is shown.

1 Introduction

COSMO–LEPS is a probabilistic system for weather forecasts which combines the probabilistic information coming form the ECWMF global ensemble system with the mesoscale information introduced by Lokal Modell. Therefore the two main features of the system are the probabilistic approach and the capability of forecasting surface parameters with a greater detail with respect to global ensemble systems, leading to a better representation of mesoscale–related processes. The verification package of COSMO–LEPS is designed keeping in mind these characteristics, in order to retain and to evaluate the information coming from both of them. As regards the necessity of understanding the behaviour of a probabilistic forecast, probabilistic verification tools have been developed. Among them, the computation of Relative Operating Characteristic (ROC Curves), Brier Score and Brier Skill Score, Cost–Loss Analysis and Percentage of Outliers has been implemented in the COSMO–LEPS verification package. Though the computation of these scores is rather simple, their interpretation is not straightforward, different indices describing different features of the forecast system. In addition to this, the relationship between these scores is not a linear one. Therefore, a global evaluation of the forecast system should rely on a set of indices. In this report, for brevity reasons, only results in terms of the Brier Skill Score, ROC area and Percentage of Outliers will be presented. A description of these indices is reported in Appendix. Bearing in mind the other characteristic of COSMO–LEPS, which is the use of a mesoscale resolving model, the probabilistic evaluation is performed only in terms of surface variables, using a high–resolution dataset. In this report the focus is on the ability of the system of forecasting intense precipitation events. A verification of the 2–meters temperature is also being carried on. Different ways of comparing forecast and observed values are followed, in order to have a clean comparison and aming at the understanding of the spatial–scale properties of the forecast fields. After a brief description of the system (Section 2), verification results are presented (Section 3), organised as follows: a comparison with the ECMWF ensemble in terms of forecast precipitation over Northern Italy is shown in Section 3.1, while an evaluation over a bigger part of the COSMO area is presented in Section 3.2. Results obtained by comparing aggregated forecast and observed values over boxes of different sizes are reported in Section 3.3. Finally, the used evaluation idices are described in Section 4.

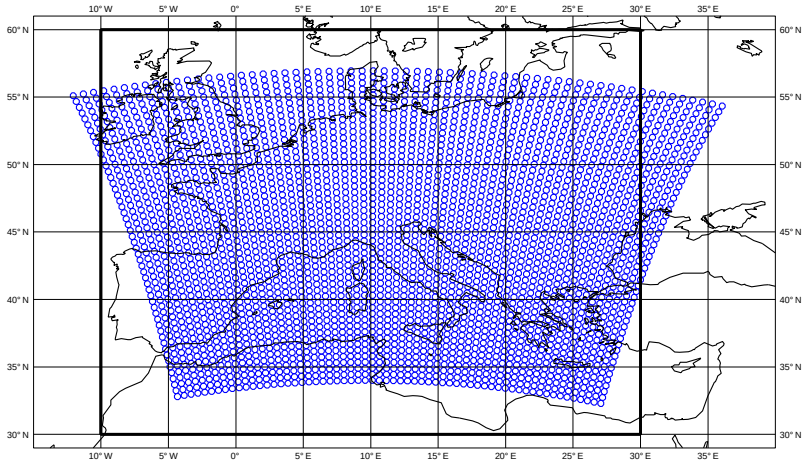

Figure 1: COSMO–LEPS operational domain (small circles) and clustering area (big rectangle).

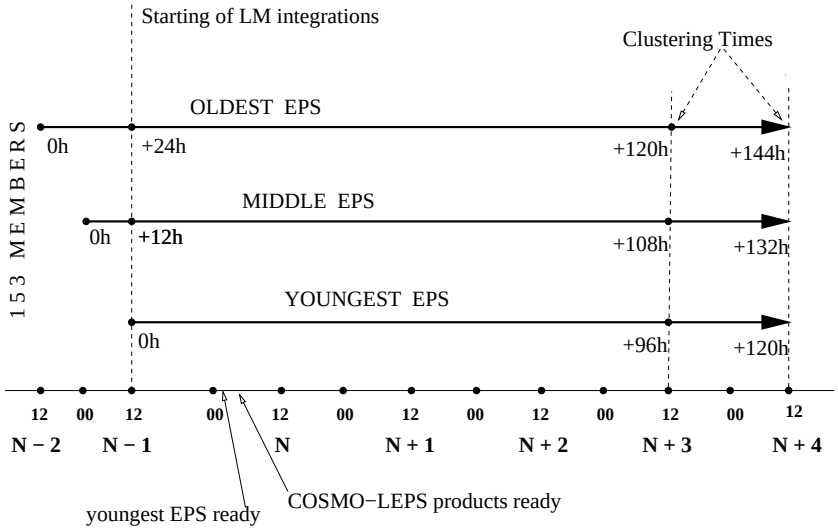

Figure 2: Details of the COSMO–LEPS suite.

2 The COSMO–LEPS operational system

The limited–area ensemble prediction system COSMO–LEPS has been running operationally at ECMWF since November 2002. The suite is run and maintained remotely by ARPA–SIM, with support given by ECMWF, and the necessary Billing Units are made avilable by the ECMWF COSMO countries. Five runs of the non–hydrostatic limited–area model Lokal Modell (LM) are available every day, nested on five selected members (the so–called Representative Members, or RMs) of three consecutive 12–hour lagged ECMWF global ensembles. The five selected members are representative of five clusters, built by grouping all the global ensemble members on the basis of their similarity in terms of upper–air fields (see Fig.2; for details on the suite implementation, the reader is referred to Montani et al., 2003b). The limited–area ensemble forecasts range up to 120 hours and are integrated over a domain covering all the countries involved in COSMO (Fig. 1). The model version is 3.3, the horizontal resolution is about 10 km and 33 vertical levels are used. LM–based probabilistic products covering a "short to medium–range" (48–120 hours) are disseminated to the weather services involved in COSMO.

Objective verification of the system has been carried out in order to quantify abilities and shortcomings of the system.

3 Verification results

The objective verification is mainly made in terms of probabilities of occurrence of selected weather events. In principle, each of the 5 COSMO–LEPS member contributes to the forecast probability of occurrence of the event by 20%. Nevertheless, due to the way the RMs driving the 5 LM integrations are selected form the super–ensemble, a weight can be assigned to each LM run, computed as the percentage of members from the super–ensemble that falls in its own RM's cluster. Scores computed by assigning this weight to each LM forecast are referred to "weighted". This "weighting procedure" is applied to the COSMO–LEPS members when probability maps are computed. Nevertheless, first verification results show that the "not–weghited" COSMO–LEPS performs slightly better than the "weighted" one or, at least, that difference between the two are not significant. In this report, results from the comparison between the two configurations are shown in Section 3.2.

Verification has been performed in two ways: by interpolating the gridded forecast values on station points where observations are available or by individuating a couple of representative observed and forecast values on boxes of fixed size. The interpolation of forecast values on station points has been performed in either of this two ways: by taking for each station the value on the nearest grid point or by averaging the values on the 4 nearest grid points. On the other hand, the unique value per each box is obtained either by computing the average value in the box or by choosing the maximum value in the box.

3.1 Comparison against ECMWF ensembles

A comparison between the COSMO–LEPS ensemble and the ECMWF available ensemble systems has been carried out over the network of stations covering Northern Italy, collected for the purpose of verifying LM. The period considered ranges from November 2002 to January 2003. Verification has been performed in terms of 24–hour cumulated precipitation.

3.1.1 Interpolation on station points

The first verification results compare COSMO–LEPS performance with that of the ECMWF ensembles. The indices have been computed for the entire super–ensemble, made up of three consecutive EPS ("epsse" model) as well as for the most recent EPS used in the construction of the super–ensemble ("eps51" model) and for the ensemble made up of the 5 ECMWF RMs chosen from the super–ensemble ("epsrm" model). Results are presented in terms of Brier Skill Score, a higher value corresponding to a better results and the zero level indicating the limit of usefulness of the forecasting system.

An interesting result is that the performance of the super–ensemble is comparable with that of the most recent EPS alone (Fig.3, black and blue lines, respectively). The 5–RM ensemble (green line) is almost always less skillful than COSMO–LEPS (red line), which benefits from the gain due to the nesting of the limited–area model. COSMO–LEPS performance is slighlty worse than that of the super–ensemble at the lower threshold (top left panel) and comparable to it at the intermediate thresholds (top right and bottom left panels). When heavy precipitation is considered (over 50mm/24h, bottom right panel), COSMO–LEPS has some skill at the 4–5 days forecast range, where the ECMWF ensembles show no skill. In order to better understand the results, it is important to remind that a bigger ensemble is favoured in terms of the Brier Skill Score (Talagrand et al., 1999). The reason of the good performance of the super–ensemble can be understood also looking at the percentage

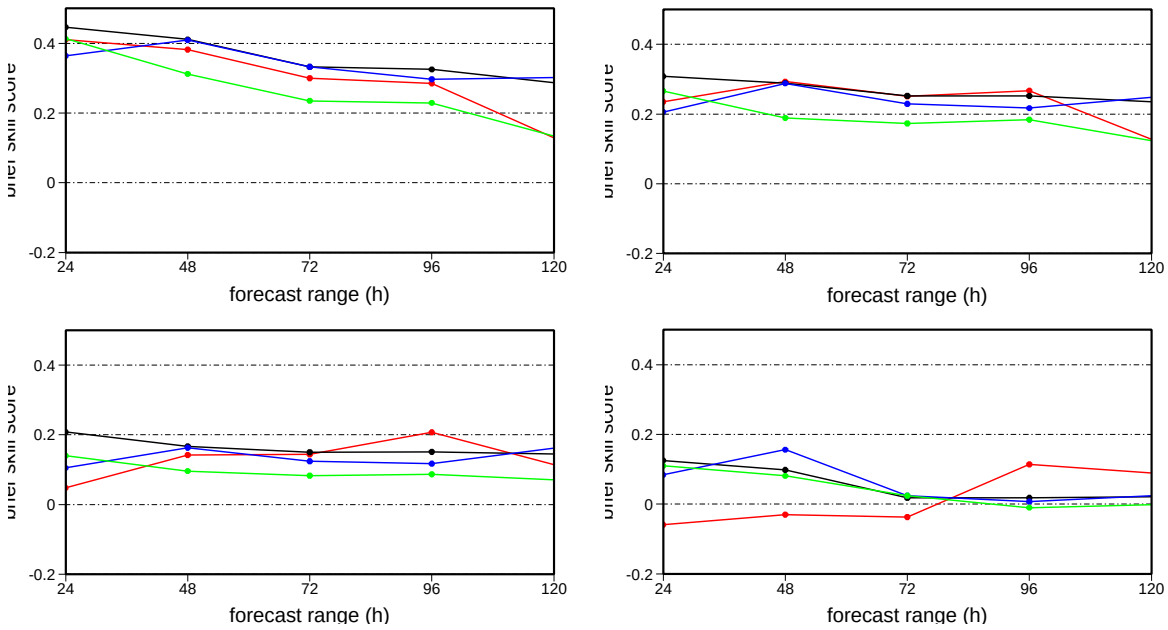

Figure 3: Brier Skill Score (the higher the better) as a function of the forecast range (in hours) relative to the 24–hour cumulated precipitation forecasts by COSMO–LEPS (cleps, red line), by the ECMWF super–ensemble (epsse, black line), by the most recent of the three EPS (eps51, blue line) and by the 5 ECWMF RMs ensemble (epsrm, green line) for different precipitation thresholds: over 10 (top left), 20 (top right), 30 (bottom left) and 50 mm/24h (bottom right). Forecast values are bilinearly interpolated over station points.

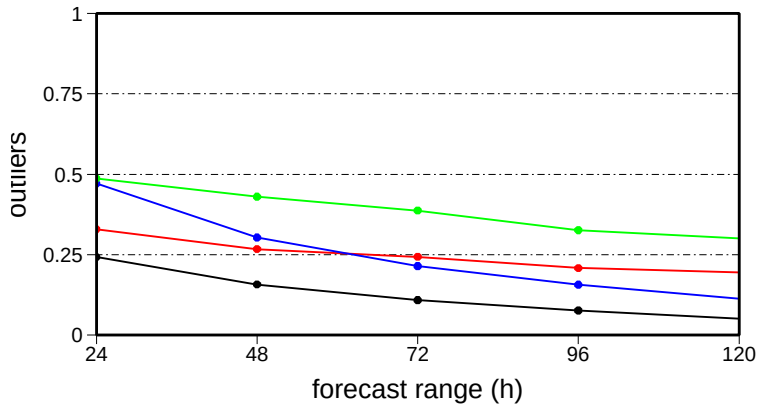

Figure 4: Percentage of outliers as a function of the forecast range (in hours) relative to the 24–hour cumulated precipitation forecasts by COSMO–LEPS (cleps, red line), by the ECMWF super–ensemble (epsse, black line), by the most recent of the three EPS (eps51, blue line) and by the 5 ECWMF RMs ensemble (epsrm, green line).

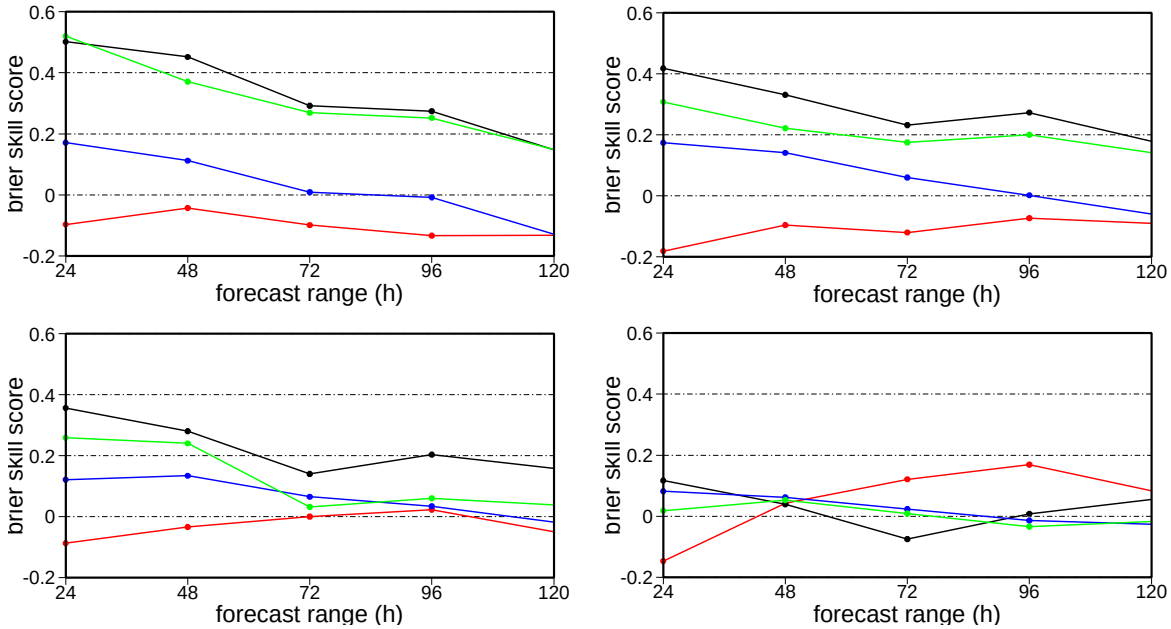

Figure 5: Brier Skill Score (the higher the better) as a function of the forecast range (in hours) relative to the 24–hour cumulated precipitation forecasts by COSMO–LEPS (red and black lines) and by the ensemble made up with the 5 RMs of ECMWF (blue and green lines) for different precipitation thresholds: over 10 (top left), 20 (top right), 30 (bottom left) and 50 mm/24h (bottom right). Black and green lines are relative to the scores computed for the average values over boxes 1.5 x 1.5, while red and blue lines are obtained by comparing the maximum values over the same boxes.

of outliers, which is the percentage of cases in which the observed value lies out of the range of the forecast values. As shown in Fig. 4, the use of the 153–member ensemble permits to have enough spread to reduce to its mimumum value the percentage of outliers for every forecast range.

3.1.2 Average and maximum values in boxes

The aggregation of both forecast and observed values over boxes seems to be very important in order to properly compare the two. This is particularly true when comparing models with different spatial resolution. In this case, COSMO–LEPS has an horizontal resolution of about 10 km, while ECWMF EPS runs have a resolution of about 80 km. In this work, boxes are built in a way that permits a partial overlapping between them, in order to avoid sharp and somewhat artificial boundaries between one box and the other. Average observed and forecast values over partially overlapping boxes with size 1.5 x 1.5 degrees are compared in this section. Due to the fact that averaging tends to smooth the precipitation field and to reduce the maximum values and having COSMO–LEPS be designed for the forecast of intense precipitation, a comparison of maximum forecast and observed values in each box has also been carried out. This analysis has been performed for COSMO–LEPS and for the 5–RM ECMWF ensemble ("epsrm") only, the verification of the super–ensemble over a great amount of boxes being highly computationally demanding.

Results are shown in Fig. 5. Black and green lines are relative to the scores computed by comparing the average observed and forecast values over boxes 1.5 x 1.5 degrees, for COSMO–LEPS and epsrm respectively. Red and blue lines are obtained by comparing the maximum observed and forecast values over the same boxes, for COSMO–LEPS and epsrm respectively. It is evident that the average values are more skillful than the maxima for moderate rainfall thresholds (10 and 20 mm/24h) and that COSMO–LEPS mean values over

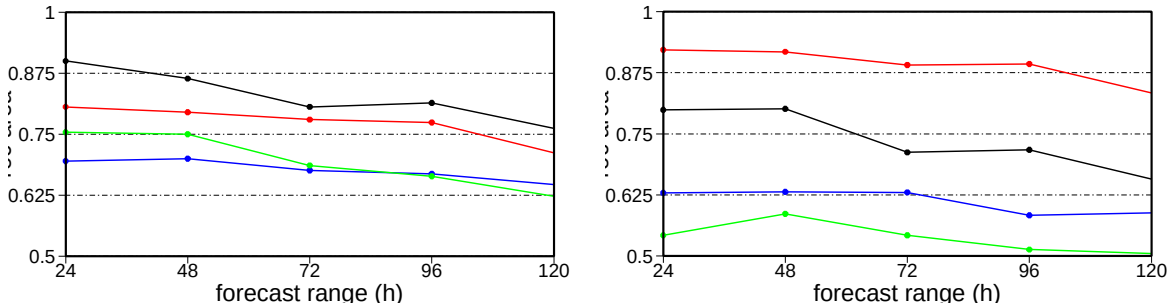

Figure 6: The same as in Fig. 5 but in terms of ROC area.

the boxes are more skillful than epsrm ones, especially for the 20 and 30 mm/24 thresholds. At the higher precipitation threshold (50mm/24h, bottom right panel), only COSMO–LEPS maxima have some skill, especially from +72 hours onward. In Fig. 6 results on terms of ROC area are also reported for the highest precipitation thresholds: 30 (left panel) and 50 mm/24h (right panel). According to this measure the better performance of COSMO–LEPS for intense precipitation is more evident, especially in terms of maximum values for the 50mm/24 threshold. The difference between the two indices can be understood referring to the Brier Score decomposition presented in Section 4.2. When an event is correctly forecast with low probability, the Brier Score increases, that is it worsens, while the ROC area increases.

3.2 Verification over Germany, Switzerland and Italy

The observational database used for COSMO–LEPS verification has recently enlarged, thanks to the efforts of the COSMO community. An agreement has been established and 24–hour cumulated precipitation data (06–06 UTC) from Germany, Switzerland and Italy are collected, put in a common format and redistributed by Ulrich Damrath. The available network is shown in Fig. 7.

This great amount of station (over 4000) has been used to verify COSMO–LEPS over a larger domain. Verification has been carried out by comparing the observed value with the forecast value on the nearest grid point. Computations is being repeated by averaging the values over boxes of different sizes and will be presented in the next future. The results obtained so far are quite unexpected: the Brier Skill Score values decrease substantially (that is, they worsen) when computed over the bigger network with respect to the values obtained by repeating verification over Northern Italy only.

This is shown in Fig. 8, where the two areas are compared: the black and the red lines are relative to the verification over Northern Italy (weighted and not–weighted, respectively), while the green and the blue line are relative to the verification made by using all the stations (weighted and not–weighted, respectively). For precipitation exceeding 10mm/24h (left panel), a decrease of the score when the whole network is used is evident, though its values are still positive. At the highest threshold (50mm/24h, right panel), the skill showed by COSMO–LEPS for the longer time ranges is missing when computations are made over the whole network. The reason for this discrepancy cannot be ascribed to the different altitude of the stations, the German stations (which constitutes the greatest part of the network) being located mainly in the plain. An analysis has been performed by subdividing the sample according to the stations' altitude (not shown) and no differences able to explain this discrepancy have been found. The different size of the two dataset can certainly play a role and a clean comparison between the two different samples is not possible. Nevertheless,

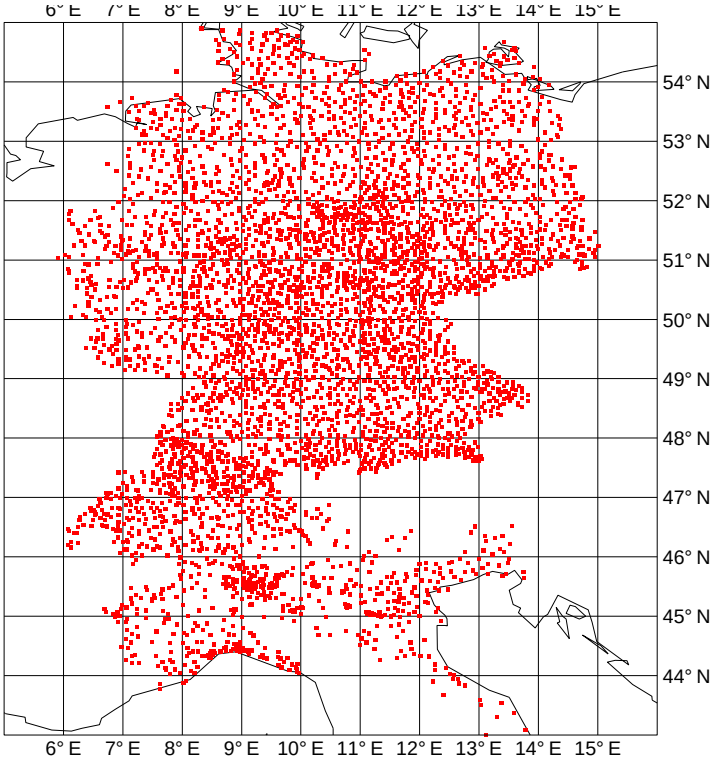

Figure 7: COSMO network of stations were observed precipitation is available. Precipitation data are cumulated over 24 hours from 06 to 06 UTC.

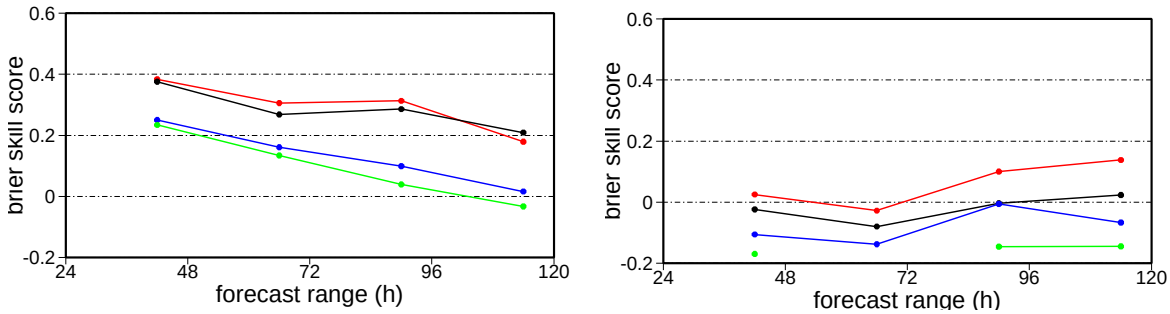

Figure 8: Brier Skill Score (the higher the better) as a function of the forecast range (in hours) relative to the 24–hour cumulated precipitation forecasts by COSMO–LEPS for precipitation thresholds 10 (left panel) and 50 mm/24h (right panel). Black and red lines are relative to the scores computed only over the italian stations (black line for the weighted configuration and red line for the not–weighted one), while green and blue lines are obtained by computing the scores over all the stations (green line for the weighted configuration and blue line for the not–weighted one).

it is possible to suggest that COSMO–LEPS, at least in Autumn, performs better over Italy than over Germany. This can be understood thinking at the mountainous terrain of the Northern Ialy, with alternating mountains and plains in a narrow region, where mountains act as a major forcing for the precipitation field, increasing the predictability associated to this parameter, especially in case of intense events. In Switzerland the situation is different with respect to Italy: the terrain is also mountainous, but we are mainly considering Autumn cases, that is cases associated with flow form the south over the Alps. The tendency of LM to underestimate the precipitation downwind could explain the rather poor performance over this area (not shown). Of course, this analysis is a very preliminar one and further investigations are needed.

As the difference between weighted and not–weighted ensemble is concerned, it is possible to notice from Fig. 8 that the weighting procedure does not imply an improvement of the scores. This is an example of a more general results: comparing the two configuration in different ways, either the not–weighted one is more skillful or no difference between the two has been found.

3.3 Comparison of different box sizes

The idea underlying the use of aggregated observed and forecast values in boxes is that the very detailed information provided by Lokal Modell contains a non–negligible stochastic component that has to be removed. This has already been expressed by Theis et al. (COSMO Newsletter No. 2). The size of the boxes, which is related to the spatial scale badly resolved by the model, is still an open problem. Every box has to be large enough to contain a number of points that permits a robust statistics, both for observations and forecasts, but the box size has to be also related to the characteristics of the model we are using, if we accept the idea that aggregating Lokal forecasts on a certain scale will lead to more robust and reliable estimate of surface parameters.

As the behaviour of the system for different spatial scales is concerned, the verification indices have been computed by comparing average and maximum values over boxes of different sizes. Forecast and observed values can have different densities in the boxes, COSMO–LEPS forecast density being about 100 points in a box of 1 x 1 degrees. In order to reduce the impact of this difference, a constraint has been imposed: only boxes where at least 10 observations are available are considered. Computation has been performed for November 2002 only, using the Northern Italy network. Precipitation has been cumulated over 24 hours, between 06 and 06 UTC.

Results (Fig. 9) indicate that the skill of the average forecast value for the moderate rainfall threshold (20mm/24h threshold, top left), is maximum when box size is 1 x 1 or 1.5 x 1.5 degrees (blue and green lines). The scores drop when the 50mm/24h threshold is considered (bottom left panel), with every box size. At this high threshold, only the maximum values are still skillful (bottom right panel), especially for bigger boxes (1 x 1 or 1.5 x 1.5 degrees, blue and green lines). At the moderate threshold, the maximum value is more skillful when little boxes are considered (top right panel). In every panel the scores obtained by comparing the observed value with the value forecast in the nearest grid point has been showed as a reference (red line).

4 Description of the probabilistic indices

The forecast produced by a probabilistic prediction system of M members can be regarded as subdivided in M+1 probability classes. The probability associated to each class is k/M, $k \in \{0, 1, \ldots, M\}$. For each probability class k and for each event, a contingency table is

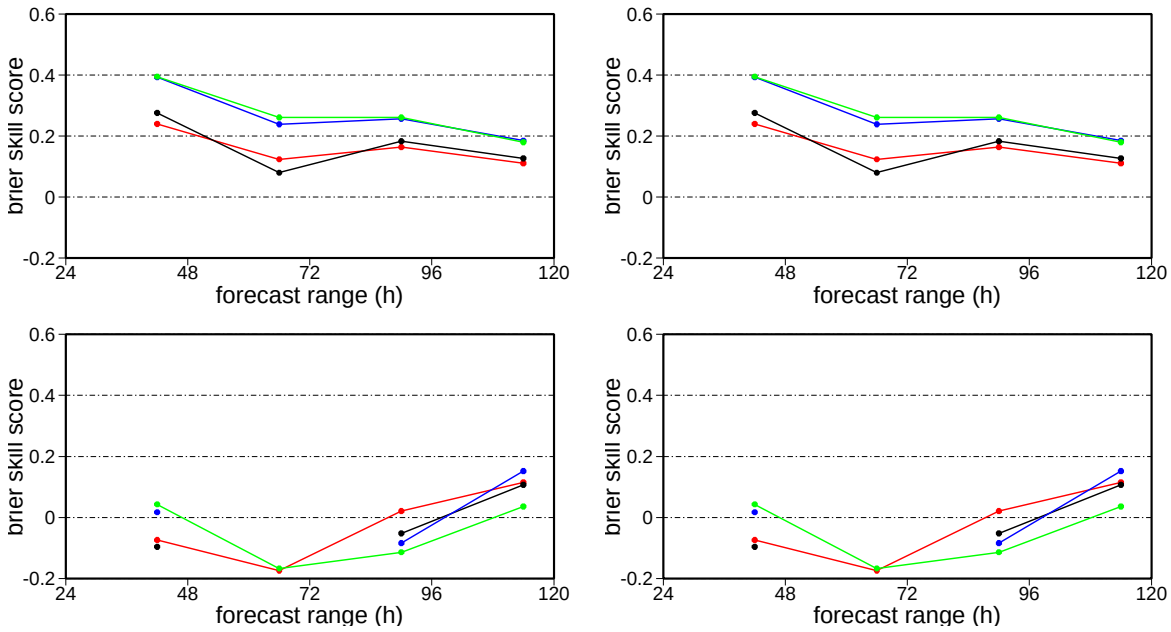

Figure 9: Brier Skill Score (the higher the better) as a function of the forecast range (in hours) relative to the 24–hour cumulated precipitation forecasts by COSMO–LEPS. Scores are relative to the average values in the left panels and maximum values in the right panels. Two precipitation thresholds have been considered: 20 mm/24h in the top panels and 50 mm/24h in the bottom panels. The red line is for the nearest point, black line for boxes of 0.5 x 0.5 degrees, blue lines for 1 x 1 degrees and green lines for 1.5 x 1.5 degrees.

built.

Contingency Table		observed	
		yes	no
forecast	yes	a_k	b_k
	no	c_k	d_k

The following relations hold: $a_k + b_k + c_k + d_k = N$, $a_k + c_k = N\bar{o}$, $b_k + d_k = N(1 - \bar{o})$, $a_k + b_k = N_k$, $a_k = N_k\bar{o}_k$ and $b_k = N_k(1 - \bar{o}_k)$. N is the dimension of the verification domain, while $\bar{o}$ is the observed frequency of the event. For each probability class, N_k is the dimension of the subspace where the event is predicted with probability k/M and $\bar{o}_k$ is the observed frequency of the event when it is predicted with probability k/M.

4.1 ROC area

The accuracy of probabilistic forecasts can be evaluated using the Relative Operating Characteristic (ROC) curves (Mason and Graham, 1999). According to the Contingency Table, the Hit Rate (H) and the False Alarm Rate (F) for each probability class are defined as:

$$H_k = \frac{a_k}{a_k + c_k} = \frac{N_k\bar{o}_k}{N\bar{o}},$$

$$F_k = \frac{b_k}{b_k + d_k} = \frac{N_k(1 - \bar{o}_k)}{N(1 - \bar{o})}.$$

The two scores indicate, respectively, the proportion of events which were predicted by k members and actually happened, and the proportion of events forecast by k members and did not occur. If several warning thresholds are used for the event, corresponding to a set of forecast probabilities, a set of cumulative hit and false alarm rates can be determined for

the same threshold. The accumulation is made for the probability classes form the k–th to the M–th. The set of cumulative H, plotted against the set of the corresponding cumulative F, generates the ROC curve. The area under the curve is commonly used as a probabilistic score, its maximum value being 1, and a value of 0.5 indicating a no–skill forecast system (Mason and Graham, 1999).

4.2 Brier Score and Brier Skill Score

The Brier Score (BS) is the mean–square error of the probability forecasts (Brier, 1950). The BS averages the squared differences between pairs of forecast probabilities and the corresponding binary observations, representing the occurrence (or non–occurrence) of the event. It is defined as follows:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2,$$

where the observation is $o_i = 1$ ($o_i = 0$) if the event occurs (does not occur), while p_i is the fraction of ensemble members which forecast the event and the index i denotes a numbering over the whole domain. BS can take on values in the range [0,1], the perfect forecast having $BS = 0$ (Stanski et al., 1989; Wilks, 1995). The Brier Score can be also expressed as a function of Hit Rate and False Alarm Rate. The relation, expressed in a different form by Talagrand et al. (1999), is:

$$BS = \bar{o}\sum_{k=0}^{M} H_k\left(1 - \frac{k}{M}\right)^2 + (1-\bar{o})\sum_{k=0}^{M} F_k\left(\frac{k}{M}\right)^2.$$

When an event is correctly forecast (high value of H_k) with a low probability (high value of $(1 - \frac{k}{M})^2$), the Brier Score increases, that is it worsens.

The Brier Skill Score (BSS) is defined as the BS percentage improvement of the forecast system with respect to climatology and it is computed as

$$BSS = \frac{BS_{cli} - BS}{BS_{cli}} = 1 - \frac{BS}{\bar{o}(1-\bar{o})}.$$

A positive BSS indicates that a system has predictive power; the perfect deterministic forecast has BSS = 1 (Stanski et al., 1989).

4.3 Percentage of Outliers

The Percentage of Outliers of a probabilistic forecast system is defined as the probability of the analysis lying outside the forecast range (Buizza, 1997). Here, it is computed as the fraction of points of the domain where the observed value lies outside the range of forecast values. The M values predicted at each grid–point by a probabilistic forecast system can be put in increasing order, thus determining M+1 intervals: v ¡ f_1, f_1 ¡ v ¡ f_2, ..., v ¿ f_M. In this way, it is possible to compute the percentage of times the observed value lies in each interval. If a sufficiently large number of cases is considered, the expected probability of the analysis being inside each of the M+1 intervals is 1/M+1 (Buizza, 1997). The percentage of times the observation lies in the first or in the last interval provides the percentage of outliers. The percentage of times the observed value is smaller than the lowest forecast value is called outliers below the minimum (first interval) and the percentage of times the observed value is greater than the highest forecast value is called outliers above the maximum (last interval).

References

Brier, G. W, 1950. Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Buizza, R., 1997. Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **125**, 99–119.

Marsigli, C., Montani, A., Nerozzi, F., Paccagnella, T., Tibaldi, S., Molteni, F., Buizza, R., A strategy for High–Resolution Ensemble Prediction. Part II: Limited–area experiments in four Alpine flood events, 2001. *Quart. J. Roy. Meteor. Soc.*, **127**, 2095–2115.

Marsigli, C., Montani, A., Nerozzi, F., Paccagnella, T., Probabilistic high-resolution forecast of heavy precipitation over Central Europe, 2003. *Natural Hazards and Earth System Sciences*, in press.

Mason, S. J. and Graham, N. E., 1999. Conditional probabilities, relative operating characteristics and relative operating levels. *Wea. and Forecasting*, **14**, 713–725.

Molteni, F., Buizza, R., Marsigli, C., Montani, A., Nerozzi, F. and Paccagnella, T., 2001. A strategy for High–Resolution Ensemble Prediction. Part I: Definition of Representative Members and Global Model Experiments. *Quart. J. Roy. Meteor. Soc.*, **127**, pp. 2069–2094.

Montani, A., Marsigli, C., Nerozzi, F., Paccagnella, T. and Buizza, R., 2001. Performance of ARPA–SMR Limited–area Ensemble Prediction System: two flood cases. *Nonlinear Processes in Geophysics*, **8**, 387–399.

Montani, A., Marsigli, C., Nerozzi, F., Paccagnella, T., Tibaldi, S. and Buizza R., 2003a. The Soverato flood in Southern Italy: performance of global and limited–area ensmeble forecasts *Nonlinear Processes in Geophysics*, **10**, 261–274.

Montani, A., Capaldo, M., Cesari, D., Marsigli, C., Modigliani, U., Nerozzi, F., Paccagnella, T., Patruno, P. and Tibaldi, S., Operational limited–area ensmeble forecasts based on the Lokal Modell, 2003b. *ECMWF Newsletter Summer 2003*, **98**, 2–7.

Stanski, H. R., Wilson, L. J. and Burrows, W. R., 1989. Survey of common verification methods in meteorology. WMO World Weather Watch Tech. Rep., **8**, pp. 144.

Talagrand, O., Vautard, R. and Strauss, B, 1999. Evaluation of probabilistic prediction systems. Proceedings of the ECMWF workshop on predictability, 20–22 October 1997, Reading, UK, pp. 372.

Wilks, D. S., 1995. Statistical methods in the atmospheric sciences. Academic Press, New York, 467 pp.