Performances of COSMO-based ensemble systems for cases of High-Impact Weather over Italy

G. Pincini (1), A. Montani (2), T. Paccagnella (1), M.S. Tesini (1), C. Marsigli (3)

(1) Arpae-SIMC, (2) ECMWF, (3) DWD

Introduction

The prediction of weather events related to strong winds, heavy rain and snowfall is still nowadays a serious challenge, especially when high spatio-temporal details are required. Despite Numerical Weather Prediction (NWP) modelling has made great progress in recent decades, thanks to the increases in model resolution, better understanding of atmospheric dynamical processes and advantages in data assimilation techniques, the above-mentioned atmospheric events, usually referred to as "High-Impact Weather" (HIW), can have horizontal dimension too small to be explicitly resolved. HIWs provide the most dramatic examples of how the atmospheric affects people daily lives, since they may cause both human and economic costs. Therefore, there is a need of better ways to predict this type of phenomena, also accounting for their inherent degree of non-predictability.

The ensemble forecasting provide a representation of model uncertainty, due to the imperfect knowledge of atmospheric initial conditions and the approximate model formulation. Instead of running just one forecast with an unknown error, an ensemble of slightly different forecasts are run, in order to integrate the deterministic forecast with an estimate of the "forecast of forecast skill". Probabilistic forecasts provide a more complete, reliable and accurate view of what might happen in the future, ideally providing information on the relative frequency of an event occuring. Therefore, they bring definite benefits for decision-makers. The estimation of uncertainty is even more crucial when local effects come into play and a high spatio-temporal detail is required as in the case of precipitation, where NWP limitations become more evident.

The aim of this work is to assess the added value of the enhanced horizontal resolution in the probabilistic prediction of surface fields. In particular, the performance of three different ensemble prediction systems were compared: ECMWF ENS (51 members, 18 km horizontal resolution), COSMO-LEPS (16 members in 2016, 20 members now; 7 km horizontal resolution) and COSMO-2I-EPS (10 members in 2016, 20 members now; 2.2 km horizontal resolution). While the first two ensemble systems are operational, COSMO-2I-EPS is still in a pre-operational phase. The intercomparison window covers two limited periods, which range from 20 to 27 June 2016 and from 15 October to 15 November 2018. As for the surface variables, 2-metre temperature and precipitation are verified against the non-conventional station network provided by the National Civil Protection Department.

The ensemble spread and the root mean square error of 2-metre temperature are computed, while Ranked Probability Score and Percentage of Outliers are considered for precipitation. The best scores are mainly obtained by the COSMO-based ensemble systems with higher horizontal resolution and lower ensemble size; in particular COSMO-2I-EPS often achieves the most satisfactory performances. Although the results are based over two relative short periods due to limited data availability and further investigations is needed, the added value of high resolution in mesoscale ensemble systems seems to play a crucial role in the probabilistic prediction of atmospheric fieds at all levels. In particular, the more detailed description of mesoscale and orographic-related processes in COSMO-ensembles provides an added value for the prediction of localised High-Impact Weather events.

Global and limited-area ensemble prediction systems and description of the experiments

A summary of the technical characteristics of the three ensembles used in the verification is shown in the table (Fig. 1).

doi:10.5676/dwd_pub/nwv/cosmo-nl_19_09

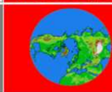
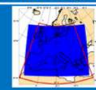
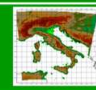
ENSEMBLE SYSTEM / MAIN TECHNICAL FEATURE	ECMWF ENS (European Centre for Medium-range Weather Forecast ENSemble prediction system)	COSMO-LEPS (Consortium for Small-scale Modelling Limited-area Ensemble Prediction Systems)	COSMO-2I-EPS
Integration domain			
Horizontal resolution (km)	18	7	2,2
Vertical resolution (Model level)	91	40	65
Forecast range (hours)	240	132	48
Type of model	Hydrostatic model	Non-Hydrostatic model	Non-Hydrostatic model
Type of convection	Parameterized convection	Parameterized convection	Explicit convection
Ensemble size	51	16	10
Starting times (UTC)	00, 06, 12, 18	00, 12	00

Figure 1: The table shows the technical characteristics of ECMWF ENS, COSMO-LEPS and COSMO-2I-EPS in 2016; now COSMO-LEPS has 20 members as well as COSMO-2I-EPS

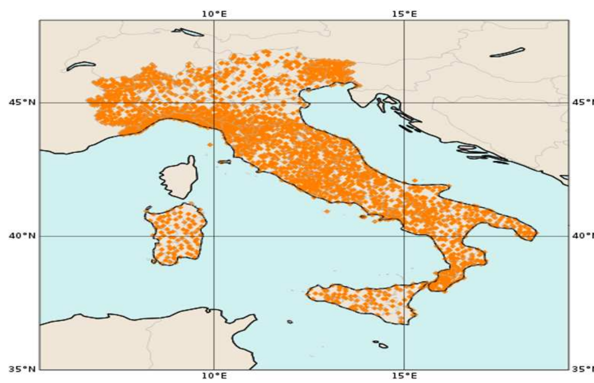


Figure 2: The domain, centered over Italy, considered for the verification of the three ensemble systems. The points are the 5524 stations of National Civil Protecion Department used for the verification of precipitation.

The intercomparison between the three ensemble systems is performed starting at 00 UTC and with a forecast range of 48 hours, because COSMO-2I-EPS runs once a day at 00 UTC and the forecast stops on the second day. The verification domain was selected in such a way as to include the entire Italian territory, more precisely the domain having the following geographic coordinate as borders (Fig. 2)

- latitude: 35^{o}N - 48^{o}N
- longitude: 6^{o}E - 19^{o}E

The station networks, used in the evaluation procedure, are:

- *the Northern-Italy non-GTS* [3] *(local) network*: it refers to about 1000 stations, over most Northern Italy and shared by the regional weather services operating in the area. These stations provide hourly data;
- *network from National Civil Protection Department (DPCN-Dipartimento Protezione Civile Nazionale)*: this network is composed of about 5524 stations over the national territory. Also these stations provide hourly data.

These station networks were used for the verification of 2-metre temperature and precipitation respectively. DPCN stations have been subdivided, in three groups depending on the location altitude. For the subdivision it was decided to adopt the WMO (World Meteorologiacal Organization) directives on the subject, as follows:

- *lowland station* (under 200 m of altitude) 2311 DPCN observatories belong to this category;
- *hill stations* (between 200 m and 599 m of altitude) 1690 observatories belong to this category;
- *mountain stations* (above 600 m of altitude) 1523 observatories belong to this category.

[3] Global Telecommunications Systems

The evaluation of the performance of the model consists in the comparison of gridded model output against point observations. A number of statistical scores evaluate different aspects of model performance while the forecast "error" is simply defined as the difference between the forecast value and the observation. In a "standard" deterministic approach, the uncertainty associated with the forecast value is not estimated. An EPS aims at quantifing this uncertainty using a set of perturbed Initial Conditions (ICs) and/or perturbed model formulations. Verification methods applied to ensemble forecasts have two main objectives:

- to assess the characteristics of the ensemble distribution;
- to verify the probability forecast.

Since all perturbed ICs should be equally possible be true and all perturbed physics or varying physics schemes or alternative models be equally plausible, the performance of any ensemble member should, in principle, be equivalent to that of another member on average. If this is not the case, that is indicative of problems with the choice of ensembling the technique employed. For example, either the IC perturbations are too large or alternative models, physics schemes or perturbations are not equally plausible. In the verification the evaluation method of the **nearest grid point** will be used: since observations seldom occur at the precise locations represented by the grid points of one particular model, it is necessary to compare the forecast values in the grid points with those of the nearest observations (ECMWF Forecast User Guide). In the experimental verification of the three ensemble systems will be used the following scores:

The Root Mean Square Error (RMSE) provides the square root of the average square error of the forecasts, which has the same units as the forecasts and observations. Here, the forecast corresponds to the ensemble mean value and an 'error' represents the difference between the ensemble mean $\overline{Y}$ and the observation x. The equation for the RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{Y_i})^2}$$

RMSE of the ensemble mean measure the distance beetween forecasts and analyses (or observations). **The ensemble spread (SPRD)** is calculated by measuring the deviation of ensemble forecasts from their mean [11]. Usually, SPRD is defined as:

$$SPRD = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(\overline{f} - f(n))^2}$$

Where $\overline{f} = \frac{1}{N}\sum_{n=1}^{N} f(n)$ is for the ensemble mean and f is for the ensemble forecast. In general, an ideal ensemble forecast will be expected to have the same size of ensemble spread as their RMSE at the same lead time in order to represent full forecast uncertainty [11] [2]; but most of the ensemble systems are underdispersed (lower spread) for longer lead times due to an imperfect model system (or physical parameterizations) and other factors. Anyway over a large number of ensemble forecasts, the statistical properties of the true value X_{TRUE} of any quantity X are identical to the statistical properties of a member X_j of the ensemble; in particular:

$$\overbrace{\overline{|X_j - X_{MEAN}|^2}}^{ensemblevariance} = \overbrace{\overline{|X_{TRUE} - X_{MEAN}|^2}}^{meansquarederror}$$

where X_{MEAN} is the ensemble mean. The time-mean ensemble spread around the mean equals the time-mean RMSE of the ensemble mean [3].

The Ranked Probability Score (RPS) is an extension of the RMSE to the probabilistic world and to the multi-category events; it ranges between 0 and 1.

$$RPS = \frac{1}{J-1}\sum_{m=1}^{J}[(\sum_{j=1}^{m} f_i) - (\sum_{j=1}^{m} o_j)]^2$$

where

- J is the number of forecast categories
- $o_j = 1$ if the event occurs in category j, $o_j = 0$ if the event does not occur in category j
- f_j is the probability of occurrence in category j

This score is used to assess multi-category forecast, where J is the number of forecast categories (for example, rainfall bins). The RPS penalizes forecasts less severely when their probabilities are close to the true outcome and more severely when their probabilities are further from the actual outcome. The lower the RPS, the better the ensemble system.

The Percentage of Outliers of a probabilistic forecast system is defined as the probability of the analysis (or observation) lying outside the forecast range [1]. Therefore this can be seen as the percentage of times the "truth" falls out of the range spanned by the forecast values. Here, it is computed as the fraction of points of the domain where the observed value lies outside the minimum or maximum forecast value.

Performance of the ensemble systems

First verification period: from 20^t to 29^{th} June 2016

To begin the performance of the three ensemble systems is verified against the two-metre temperature. As already mentioned before, for this verification it was decided to consider the observational dataset coming from the regional networks of the weather services on Central-Northern Italy. In this way, data coming from only one part of the Peninsula were considered. Infact, the temperature data from the national civil protection network could have been used, but these data are from time to time of low-quality in Central and Southern Italy and their use would have provided wrong evaluation on the model skill. The period under investigation is from 20^{th} June 2016 at 00 UTC to 29^{th} June 2016 at 00 UTC, infact, although the last runs examined are those at 00 UTC on 27^{th} June 2016, a 48-hour forecast range must always be considered. The performance of the three ensemble systems is evaluated by calculating the spread and the RMSE of the ensemble, the verification method used is the nearest grid point. The table 1 summarizes the characteristics of the verification.

Verification features	
variable:	2-metre temperature;
period:	from 20/06/2016 00UTC to 29/06/2016 00UTC (9 days);
region:	Central-Northern Italy;
method:	nearest grid point;
obs:	non-GTS local fiduciary network, no obs error;
fcst ranges:	0-48h (verification every 6h);
systems:	ECMWF EPS, COSMO-LEPS, COSMO-2I-EPS;
scores:	spread, RMSE;

Table 1: 2-metre temperature verification features

The results are reported in (Fig. 3) and can be summarised as follows:

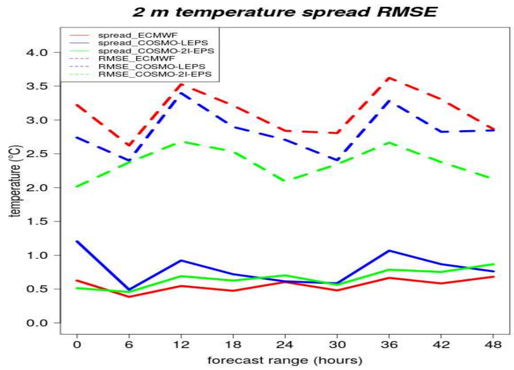


Figure 3: The figure shows the spreads (continuous lines) and the RMSE (dotted lines) values obtained for the 48 hours of the forecast range every 6 hours. The ECMWF EPS scores appear in red, COSMO-LEPS in blue and COSMO-2I-EPS in green. The foreacst range (in hours) is shown in the abscissa, in the ordinate the value of spread and RMSE (in ^{o}C). All details are indicated in the legend at the top left.

- the spread values are similar for all the three ensemble systems;
- the spread values are smaller with respect to the RMSE ones, showing a tendency of all ensembles to be underdispersive;
- with the exception of the shortest time range, COSMO-based models always show slightly higher (and therefore better) spread values than ECMWF EPS;
- RMSE values show a marked diurnal cycle, with maxima during the central hours of the day and the minimums in the night. This daytime cycle is very pronounced for ECMWF EPS and for COSMO-LEPS, less for COSMO-2I-EPS;
- the RMSE of COSMO-2I-EPS is the lowest of the three ensemble systems on the entire forecast range.

Therefore, from this 2-metre temperature verification, COSMO-based models get excellent results, especially COSMO-2I-EPS.

The performance of ECMWF ENS, COSMO-LEPS, COSMO-2I-EPS is verified also against the 6-hourly precipitation. For this verification work it was decided to use the precipitation data recorded by the rain gauges of National Civil Protection Department network. In this way, the results obtained are representative of what happened on the entire national territory between the 20^{th} June 2016 at 00 UTC and the 29^{th} June 2016 at 00 UTC. The method of the nearest grid point was used for the calculation of Ranked Probability Score and percentage of outliers. In table 2 are reported all the details of the verification.

Verification features	
variable:	6-hourly total precipitation ;
period:	from 20/06/2016 00UTC to 29/06/2016 00UTC (9 days);
region:	Italy;
method:	nearest grid point;
obs:	DPCN network, no obs error;
fcst ranges:	0-48h (verification every 6h);
systems:	ECMWF EPS, COSMO-LEPS, COSMO-2I-EPS;
scores:	RPS, outliers;
thresholds:	1mm, 5mm, 10mm, 15mm, 25mm, 50mm in 6 hours

Table 2: 6-hourly total precipitation verification features

In the Fig.4, the results obtained for the RPS can be consulted.

Considering all DPCN staions, regardless of the altitude (top left graph), it is worth pointing out:

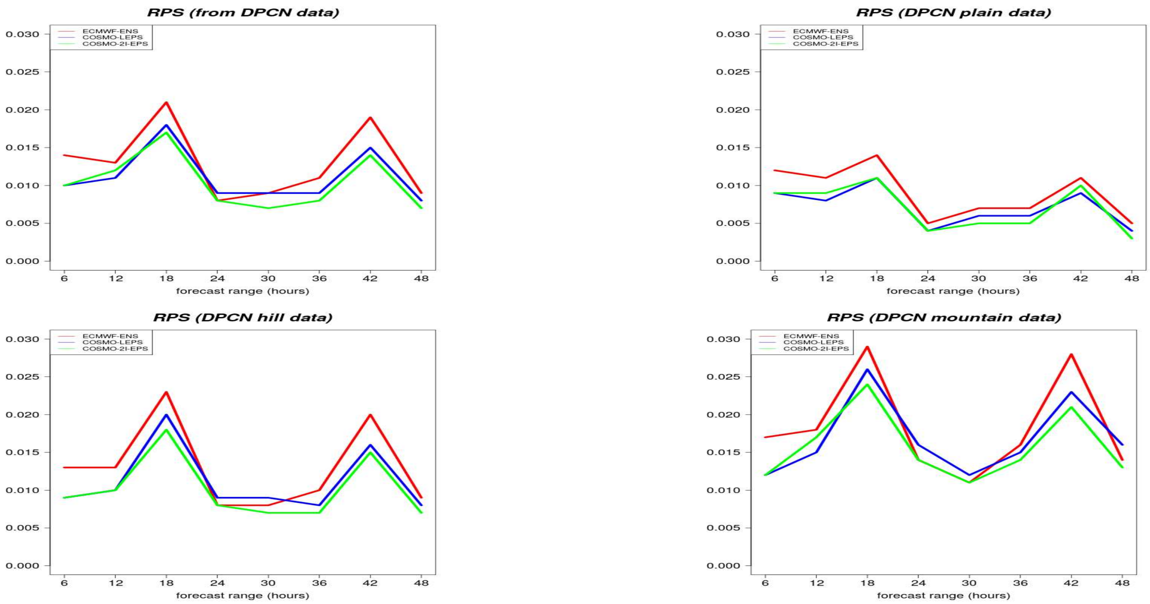


Figure 4: The figure shows the RPS for four different observational dataset, indicated in the caption under each image. The ensemble systems are ECMWF ENS in red, COSMO-LEPS in blue, COSMO-2I-EPS in green. The forecast range of 48 hours, in 6-hour steps, is shown in the abscissa; the dimensionless values of the RPS are marked in the ordinate.

- the diurnal cycle of precipitation is very marked. Since it is almost exclusively afternoon convective precipitation, the highest, and therefore the worst, RPS are just in the afternoon time slots: forecast range 12-18 h, 36-42 h;
- however, the RPS of COSMO-2I-EPS, and generally the RPS of COSMO ensemble systems, is better than ECMWF ENS one over the whole forecast range.

In this case the RPS points out to the added value of COSMO-2I-EPS.

The station of DPCN has been subdivided, according to their altitude, in three groups: plain, hill and mountain. Therefore, the purpose of this further study is to evaluate RPS variations according to the station altitude and see how this affects the results. Looking at the plots it can be concluded that:

- the RPS values obtained for the lowland stations are lower (therefore better) than those obtained for hill and mountain ones, in particular the results of mountain stations are the highest;
- in most cases, regardless of altitude, the RPS obtained for COSMO-2I-EPS is always lower (therefore better) than for COSMO-LEPS and ECMWF ENS;
- in the plain stations (top-right panel), there is a good gap beetween COSMO-2I-EPS and ECMWF ENS in the first day of forecast range. For the other stations this gap extends no longer than the first 18 hours, then the RPS tend to be similar for the three ensembles, except for the precipitation cumulated beetween the 36^{th} and the 42^{nd} hour of the forecast range;
- in the graph for hill and mountain stations (bottom left and bottom right panel respectively), the RPS follows a very strong daytime cycle, that is definitely less visible on the plain: this is could be due to the pluviometric regime of those days, with rainfall concentrated almost always in the afternoon hours and on the internal areas of hills and mountains.

So all the observational networks, built on altitude, confirm that the RPS of COSMO-based ensembles, but in particular COSMO-2I-EPS, are better than the global ensemble of Reading.

The percentages of outliers for the ensemble system considered as a function of the forecast range are shown in Fig. 5.

Considering all DPCN staions (top left graph), it is possible to see that despite the lower ensemble size, COSMO-2I-EPS has often the lowest values, compared to the other two ensemble systems with a lower horizontal resolution. So, it can be stated that in this case too, the results obtained by COSMO-2I-EPS are satisfactory. Looking at the other three panels of the Fig. 5 it can be stated that:

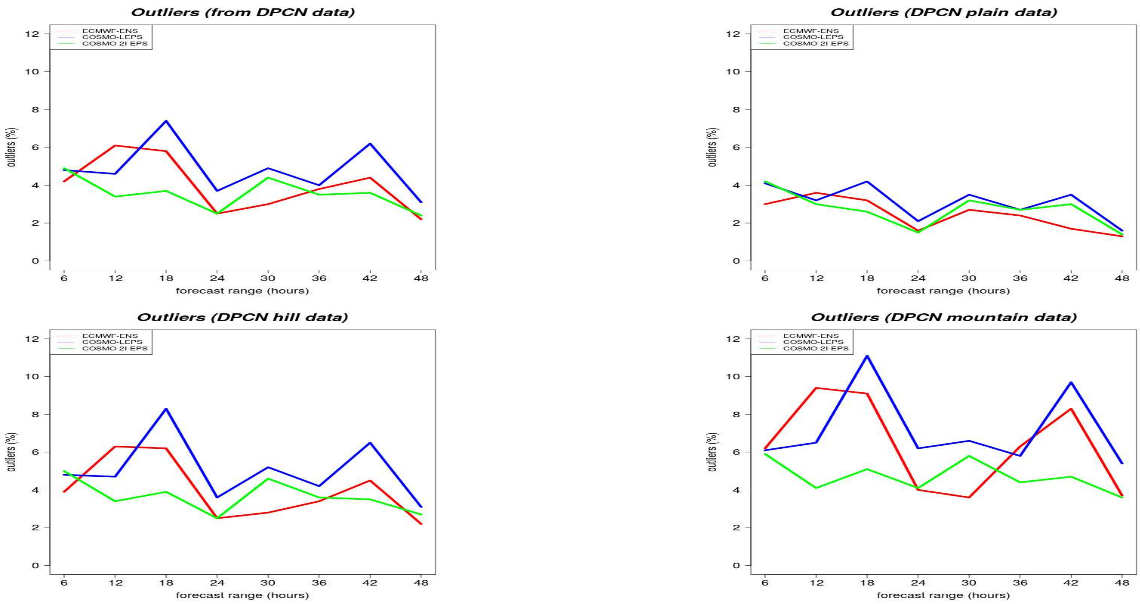


Figure 5: The figure shows the percentage of outliers for four different observational dataset, indicated in the caption under each image. The ensemble systems are ECMWF ENS in red, COSMO-LEPS in blue, COSMO-2I-EPS in green. The forecast range of 48 hours, in 6-hour steps, is shown in the abscissa; the percentage of outliers is marked in the ordinate.

- the percentage of outliers increases according to the station altitude: there are less outliers in the plains than in the mountains;
- in the plain there is little difference between the three ensemble systems; these differences increase with the altitude, infact the percentage of outliers obtained with the only mountain stations shows considerable dissimilarity beetween the ensembles;
- for hill and mountain observation datasets, a diurnal cycle is visible only in systems with parametrized convection (ECMWF ENS, COSMO-LEPS); instead, the diurnal cycle is hardly identifiable for the lowland stations;
- for almost all forecast ranges COSMO-2I-EPS has the lowest percentage of outliers.

Therefore also the percentages of outliers, studied according to the altitude of DPCN stations, indicate the good skill of COSMO-2I-EPS.

Second verification period: from 15^{th} October to 15^{th} November 2018

In this second period it has considered only the 24-hour total precipitation, the verification is performed with the rank historam.

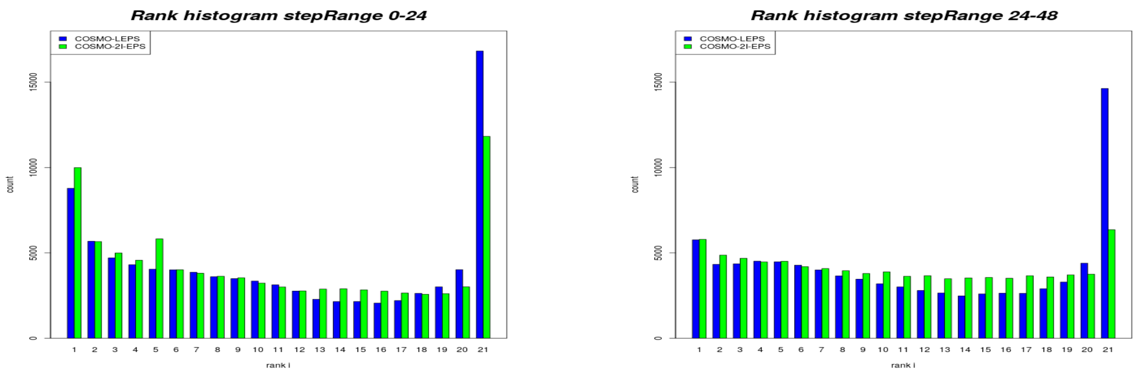


Figure 6: The rank histograms for COSMO-LEPS in blue and COSMO-2I-EPS in green, on the left for the first 24 hours of the forecast range, on the right for the second 24 hours.

The rank histogram is not a verification method per se, but rather a diagnostic tool to evaluate the spread of an ensemble. The underlying assumption is that the ensemble member forecasts are distributed so as to delineate ranges or "bins" of the predicted variable such that the probability of occurrence of the observation within each bin is equal. For each specific forecast, the bins are determined by ranking the ensemble member forecasts from lowest to highest. The interval between each pair of ranked values forms a bin. If there are N ensemble members, then there will be N+1 bins. The outer bins, lowest and highest-valued, are open-ended. Rank histograms are prepared by determining which of the ranked bins the observation falls into for each case, and plotting a histrogram of the total occurrences in each bin, for the full verification sample. The assumption underlying the rank is that the probability that the observation will fall in each bin is equal. If this is true, then over a large enough sample, the histogram should be flat or roughly so. Then one can conclude that on the average, the ensemble spread correctly represents the uncertainty in the forecast. Also in this case the verification method is the nearest grid point and the comparison is only between COSMO-LEPS and COSMO-2I-EPS, because these systems have the same number of member (20 in 2018).

The U-shape of the rank histograms (see Fig. 6) indicates the subdispersion of both ensemble systems, in particular COSMO-LEPS. This subdispersion is stronger in the last bin of the most intense precipitation and in particular for COSMO-LEPS.

Summary and Outlook

The present work aims to establish the performance of three ensemble systems with different characteristics, but in particular with a different horizontal resolution. While ECMWF ENS and COSMO-LEPS run on an daily basis, COSMO-2I-EPS is still on a pre-operational phase, with a full operational implementation planned towards the next months. Therefore, particular attention has been paid to this new ensemble, especially because it provides new types of numerical modeling products which needs to be assessed and because the best performances were expected from it. A systematic comparison between the three ensemble systems was undertaken during a "pilot period" from 20^{th} to 27^{th} June 2016. During this period, characterised by particularly unstable weather situation over the Italian Peninsula, the performances of the three systems were compared in terms of 2-metre temperature and precipitation. The forecasts in terms of 2-metre temperature and 6-hourly cumulated precipitation were verified against the Northern-Italy non-GTS network and the National Civil Protection Department network respectively. The results for 2-metre temperature indicate the under-dispersion issue for the different ensemble systems, although it can be noticed that the performance obtained by COSMO-2I-EPS (and in general by the COSMO-based ensembles) is quite satisfactory.
Rank Probability Score and percentage of outliers were considered to evaluate the skill of the three ensemble systems in terms of precipitation. In most cases, the scores indicate COSMO-2I-EPS having the best performance. In order to provide more insight on the obtained results and to assess the dependence of the scores on the altitude, it was decided to divide the stations of the National Civil Protection Department into three groups: plain, hill and mountain stations. With this division, it turns out that the performance of the systems tends to worsen with the altitude, also accentuating the diurnal cycle. This happens because it has rained more over mountain areas and during the afternoon. Anyway the scores obtained by COSMO-2I-EPS remain the best in most cases. COSMO-2I-EPS achive good results also in the verification with rank histograms, for the period from from 15^{th} October to 15^{th} November 2018. This work can be seen as a pilot study, there is no claim to consider it complete and exhaustive, but rather a starting point for further developments and investigations or a "modus operandi" for similar studies. In fact, the periods examined are too short to have solid results from a statistical point of view. This would take a longer evaluation time, comparing the three

ensembles for different atmospheric phenomena and weather types. All the results shown in this work have been obtained with the verification method of the nearest grid point. So a further idea for future studies may be to use the method of boxes to calculate the probabilistic scores in other cases; it will be interesting to see if the results will be better or worse than those obtained with the nearest grid point.

References

[1] Buizza, 1997. Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF Ensemble Prediction Systems. *Mon. Wea. Rev.*, **125**, 99-119.

[2] Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, Y. Zhu, 2005: A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems. *Monthly Weather Review*, **133**, 1076-1097.

[3] Gofa, F., Tzaferi, D. and Charantonis, T., 2010. Application and verification of ECMWF products. *Hellenica National Meteorological Service (HNMS)*

[4] Ghelli, A., 2009. Observations and their importance in the verification process: View of the Joint Working Group on Forecast Verification Method Research. *ECMWF*, *Twelfth Workshop on Meteorological Operational Systems, 2-6 November 2009*

[5] Montani, A., Capaldo, M., Cesari, D., Marsigli, C., Modigliani, U. and co-authors, 2003a. Operational limited-area ensemble forecast based on the Lokal Modell. ECMWF Newsletter 98, 2-7. Available at: http://www.ecmwf.int/publications/.

[6] Montani, A., Marsigli, C., Nerozzi, F., Paccagnella, T., Tibaldi S. and Buizza, R., 2003b. The Soverato flood in Southern Italy: performance of global and limited-area ensemble forecasts. *Nonlin. Proc. Geophys.*, **10**, 261-274.

[7] Montani, A., Cesari, D., Marsigli, C. and Paccagnella, T., 2011. Seven years of activity in field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges. *Tellus*, **63**, 605-624.

[8] Murphy, 1969. On the "Ranked Probability Score". *J. Appl. Meteor.*, **8**, 988-989.

[9] Schättler, Doms and Schraff, 2016. A Description of the Nonhydrostatic Regional COSMO-Model. Part VII: User Guide. (www.cosmo-model.org).

[10] Tibaldi, S., Paccagnella, T., Marsigli, C., Montani, A. and Nerozzi, F., 2006. Limited area ensemble forecasting: the COSMO model. Predictability of Weather and Climate. *Cambridge University Press*, 489-513.

[11] Zhu, Y., 2005: Ensemble forecast: A new approach to uncertainty and predictability, *Advance in Atmospheric Sciences*, **22**, No. 6, 781-788.