Consortium

for

Small-Scale Modelling

Technical Report No. 22

The CONSENS Priority Project

July 2013

DOI: 10.5676/DWD_pub/nwv/cosmo-tr_22

Deutscher Wetterdienst
MeteoSwiss
Ufficio Generale Spazio Aereo e Meteorologia
ΕΘΝΙΚΗ ΜΕΤΕΩΡΟΛΟΓΙΚΗ ΥΠΗΡΕΣΙΑ
Instytucie Meteorogii i Gospodarki Wodnej
Administratia Nationala de Meteorologie
ROSHYDROMET
Agenzia Regionale per la Protezione Ambientale del Piemonte
Agenzia Regionale per la Protezione Ambientale dell'Emilia-Romagna
Centro Italiano Ricerche Aerospaziali
Amt für GeoInformationswesen der Bundeswehr

www.cosmo-model.org

Editor: Massimo Milelli, ARPA Piemonte

The CONSENS Priority Project

C. Marsigli [*], **T. Diomede** [*], **A. Montani** [*] **and T. Paccagnella** [*],

P. Louka [**] **and F. Gofa** [**],

A. Corigliano [***]

[*] ARPA-SIMC, Bologna, Italy
[**] HNMS, Athens, Greece
[***] University of Bologna, Bologna, Italy

Contents

1 Abstract

This project was aimed at consolidating the COSMO ensemble forecasting systems for the mesoscale. The operational ensemble COSMO-LEPS and the experimental COSMO-SREPS system, both running at 7 km horizontal mesh-size, have been designed for different forecast ranges (day 3-5 and 1-2, respectively) and with different perturbation strategies. The aim of the project was to propose a solution for a unique "multi-perturbation-strategy" COSMO ensemble system, benefiting of perturbations which can produce appropriate spread for the entire forecast range (i.e., day 1-5), including a set of model perturbations which should guarantee a good description of the COSMO model error. Both perturbations of the mode physics parameters and perturbation of the soil moisture field had been considered. Furthermore, a calibration strategy was developed and applied to the QPF issued by the COSMO-LEPS ensemble. The project lasted for 3 years.

2 Introduction

Since November 2002, COSMO is running a Consortium ensemble, COSMO-LEPS (Montani et al., 2003). In the present configuration, COSMO-LEPS is a 16 member ensemble based on the COSMO model with a 7 km horizontal mesh-size and 40 vertical levels. The system is running at ECMWF using the Billing Units provided by the COSMO Countries which are also ECMWF member states (Germany, Greece, Italy, and Switzerland) and it is developed an maintained by ARPA-SIMC. COSMO-LEPS is a dynamical downscaling of the ECMWF EPS, taking initial and boundary conditions from 16 selected members of the EPS. COSMO-LEPS is mainly designed for the early medium range (day 3-5) and the forecast range is 5.5 days. Some perturbations to the model are included, following the outcomes of the SREPS and CONSENS PPs. The soil moisture analysis provided by DWD for COSMO-EU is also used (COSMO-LEPS is currently using the same grid as COSMO-EU, but for a smaller domain). For more detail, the reader is referred to Montani et al., 2011. In September 2006, COSMO started the SREPS Priority Project, aiming at the development of an ensemble system targeted for the short-range, COSMO-SREPS (Marsigli et al., 2009). Two different kind of perturbations are applied: initial and boundary conditions benefit of a multi-model approach, being provided by 4 different operational global models and the COSMO model itself is perturbed by changing the values of a set of physics parameters. The SREPS project ended in September 2008, delivering an ensemble system for the short-range based on the COSMO model with a 10 km horizontal mesh-size and 40 levels in the vertical. Initial and boundary conditions to the 10 km COSMO-SREPS runs were provided by 4 COSMO run at 25km hor. res. nested on IFS, GME, GFS, UM, performed by AEMET as part of their SREPS system. Then, a new COSMO-SREPS suite was implemented in 2010, with direct nesting of COSMO at 7 km on the 4 global model. The system was running quasi-regularly at ECMWF, using the Billing Units provided by some COSMO Countries. COSMO-SREPS provided initial and boundary conditions to the convection-resolving ensemble for Germany under development at DWD (COSMO-DE-EPS) and it was used to compute a flow-dependent B-matrix for a test version of the 1d-Var assimilation of satellite data under development at ARPA-SIMC. In order to avoid duplications, COSMO aimed at a confluence of the two 7 km ensemble systems into a unique COSMO ensemble, covering both the short and the medium range and based on the most appropriate perturbation strategy for the entire forecast range. This implies the development of a "multi-perturbation-strategy" ensemble system, with perturbations more appropriate for the short range in the beginning (day 1-2), and with perturbations more appropriate for the medium range thereafter, or a combination of both. It should be reminded that the importance of such an ensemble system resides both in its inherent value as assistant to the 7-km operational runs and in its capability of providing boundary conditions for the convection-permitting ensemble system under development in the COSMO countries. The issue of ensemble merging and of multi-perturbation strategy design has been dealt in the Project and results are presented in Section 4. A satisfactory representation of the uncertainty affecting the mesoscale was still lacking. The quantification of the errors which are made by the model in the description of the meteorological phenomena at this scale needs to be improved. The development of mesoscale ensemble should then include 1) the perturbation of the lower boundary forcing and 2) the perturbation of the parametrised physical processes. The importance of further perturbing the physics parameters has been underlined by the outcome of the SREPS PP. Within COSMO it has also been suggested that adding perturbations in the soil parameters can be crucial in order to get a better representation of the model error in terms of surface variables. The issue of model perturbation was also addressed in the Project, and results are presented in Section 3. Within this work on the consolidation of the COSMO ensemble systems, attention has also been paid to the post-processing. Calibration of ensemble forecasts in terms of precipitation and surface temperature has been studied and applied widely in recent years (e.g. Hamill and Whitaker, 2006; Hagedorn et al., 2008; Hamill et al., 2008; Santos-Muñoz et al., 2007). As for the COSMO-LEPS ensemble, it has been recognised that a calibration for 24 h precipitation would be desirable to improve the forecast skill (Marsigli et al., 2008). Last year, work on COSMO-LEPS calibration has been carried out at MeteoSwiss (Felix Fundel et al., 2008). Thirty year of re-forecast of one member of COSMO-LEPS have been computed and stored at ECMWF, and they have been used for calibrating the COSMO-LEPS output over Switzerland. This work has shown the potential of using re-forecast to improve the forecast skill. Therefore, calibration of QPF, especially aimed at hydrological applications, was also addressed in the Project, and results are shown in Section 5.

3 Model perturbations

3.1 Parameter perturbations

On the basis of the work carried out in the SREPS Priority Project (COSMO Technical Report n. 13, 2009), further investigations on the perturbation of model parameters had been carried out. In CONSENS, it was decided to start by evaluating a set of parameters, already tested in SREPS, over a different season, JJA (summer 2008, 92 days). For this purpose, an ad hoc test suite have been set-up at ECMWF, CSPERT, where 16 runs of COSMO are performed, with the same configuration of the COSMOSREPS/LEPS runs, but with IC and BCs provided by the IFS deterministic forecast. The model domain is shown in Figure 1. The 16 runs have 16 different set-up of the physics parameters, as described in Figure 2. An analysis has been carried out in terms of 2m temperature and dew-point temperature, for both northern Italy and Greece (see Figure 3 and Figure 4 respectively). Results confirm what has been found for the autumn 2007 season:

- qc0: no detectable impact
- tur_len: small impact for tur_len=150
- c_soil, c_lnd and pat_len parameters: marked impact, even suspicious for c_soil and c_land

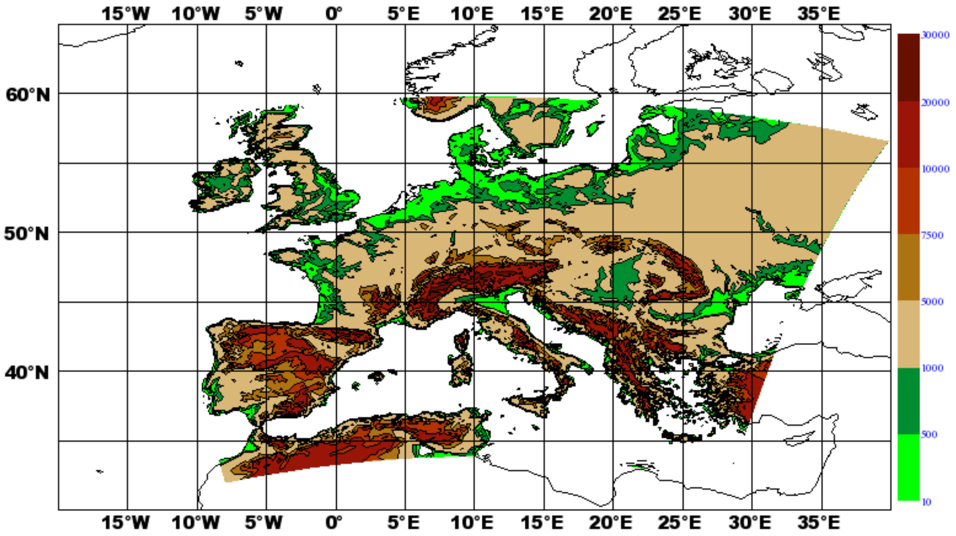


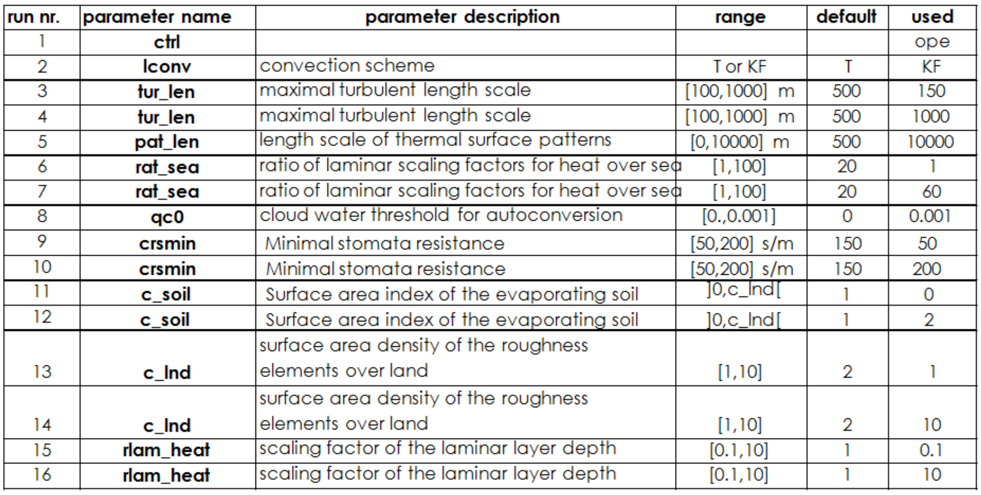
Figure 1: Orography of the model in the ensemble systems, showing the model domain.

- rlam_heat is an important parameter to perturb and it is sensitive to small changes, especially over Greece

On the basis of these results, and on the previous ones, it was decided to test other parameters or combinations of parameters in the CSPERT test suite (Figure 5). In this set up the new parameters considered are: cloud droplet number concentration (cloud), exponent of the raindrop size distribution (mu_rain) and switch on/off of the graupel scheme (gscp). These tests were run over MAM, JJA and SON 2009 and evaluation was performed in terms of precipitation over Northern Italy and Greece, then over Greece also for T, Td, and 10m wind. Some results of the evaluation are listed below:

- 2m temperature: patlen=10000 has strong detrimental effect over Greece (⟶ decision to switch to patlen=2000, also based on previous results); crsmin=200 has little positive impact
- 10 m wind: no strong effects (not expected, due to the perturbed parameters, at least for the horizontal wind)
- precipitation (very little precipitation in spring, though, especially over Greece):
  - crsmin=200 determines an increase the precipitation amount. Detrimental effect over Greece in terms of BSS
  - itype_gscp=3 determines a decrease of precipitation over Northern Italy
  - there is an impact of both perturbing cloud_num and mu_rain

In Figure 6 the scores in terms of 6h precipitation over Northern Italy are shown, for some parameter set-up for the spring 2009 season. The perturbation of the mu_rain parameter has an impact, increasing POD for intense precipitation and decreasing it for light precipitation,



run nr.	parameter name	parameter description	range	default	used
1	**ctrl**				ope
2	**lconv**	convection scheme	T or KF	T	KF
3	**tur_len**	maximal turbulent length scale	[100,1000] m	500	150
4	**tur_len**	maximal turbulent length scale	[100,1000] m	500	1000
5	**pat_len**	length scale of thermal surface patterns	[0,10000] m	500	10000
6	**rat_sea**	ratio of laminar scaling factors for heat over sea	[1,100]	20	1
7	**rat_sea**	ratio of laminar scaling factors for heat over sea	[1,100]	20	60
8	**qc0**	cloud water threshold for autoconversion	[0.,0.001]	0	0.001
9	**crsmin**	Minimal stomata resistance	[50,200] s/m	150	50
10	**crsmin**	Minimal stomata resistance	[50,200] s/m	150	200
11	**c_soil**	Surface area index of the evaporating soil	]0,c_lnd[	1	0
12	**c_soil**	Surface area index of the evaporating soil	]0,c_lnd[	1	2
13	**c_lnd**	surface area density of the roughness elements over land	[1,10]	2	1
14	**c_lnd**	surface area density of the roughness elements over land	[1,10]	2	10
15	**rlam_heat**	scaling factor of the laminar layer depth	[0.1,10]	1	0.1
16	**rlam_heat**	scaling factor of the laminar layer depth	[0.1,10]	1	10

Figure 2: Set up to the CSPERT suite for the 2007 and 2008 runs.

with almost no effect on the false alarm rate. Its association with the cloud_num perturbation gives detectable, though small, impact. The same scores are computed also for the autumn season (Figure 7), confirm that the impact of these 2 perturbations is detectable, especially for intense precipitation, and that they do not lead to a worsening of the forecast. On the basis of these results, a final set-up for the COSMO-SREPS perturbations (shown in Figure 8) has been implemented in 2011. The same perturbations have been adopted also in the COSMO-LEPS suite, but they can be combined differently.

3.2 Soil moisture perturbation

The interaction between the surface and the lower troposphere determines the development of fluxes close to the ground. Soil moisture is of primary importance in determining the partition of energy between surface heat fluxes, thus affecting surface temperature forecasts. The ensemble forecasts usually suffer of a lack of variability among the members, which is typically worse near the surface rather than higher in the troposphere. Therefore, the aim of this work is to ameliorate this deficiency by implementing a technique for perturbing soil moisture conditions and explore its impacts on the variability of the members for the different forecasted surface parameters (e.g. 2m air temperature, accumulative precipitation). The proposed technique is based on the method by Sutton and Hamill, 2004. The steps followed for its implementation are:

- use daily soil moisture data for a period with soil moisture variability for continuous years, in order to have a “climatology”
- calculate a 30-day running mean from this set of dataset
- subtract the running mean from the daily soil moisture data to calculate daily deviations
- apply an Empirical Orthogonal Function (EOF) analysis to calculate the perturbations in the variability categories appearing in the data

For the retrieval/archiving of soil moisture data, daily soil water content data were provided by DWD COSMO-EU surface analysis. The period selected for the dataset, is three months

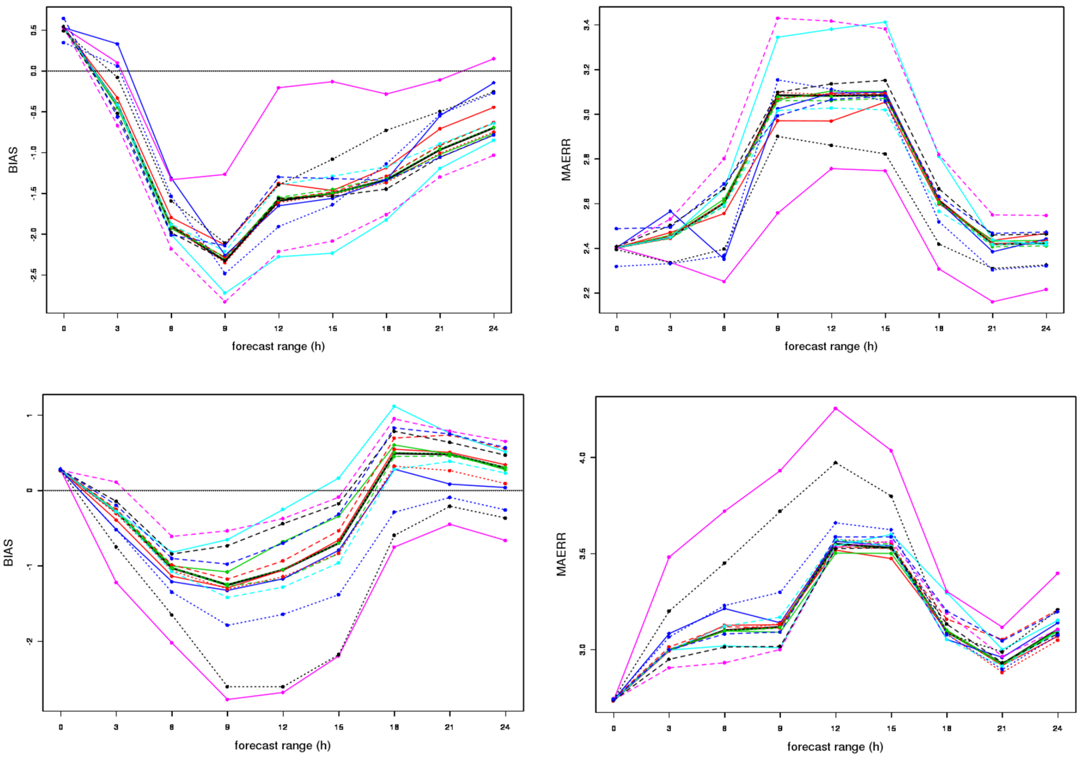


Figure 3: Bias and Mean Absolute Error of 2m T and Td for the CSPERT run, computed over Northern Italy against SYNOP station data, for JJA 2008.

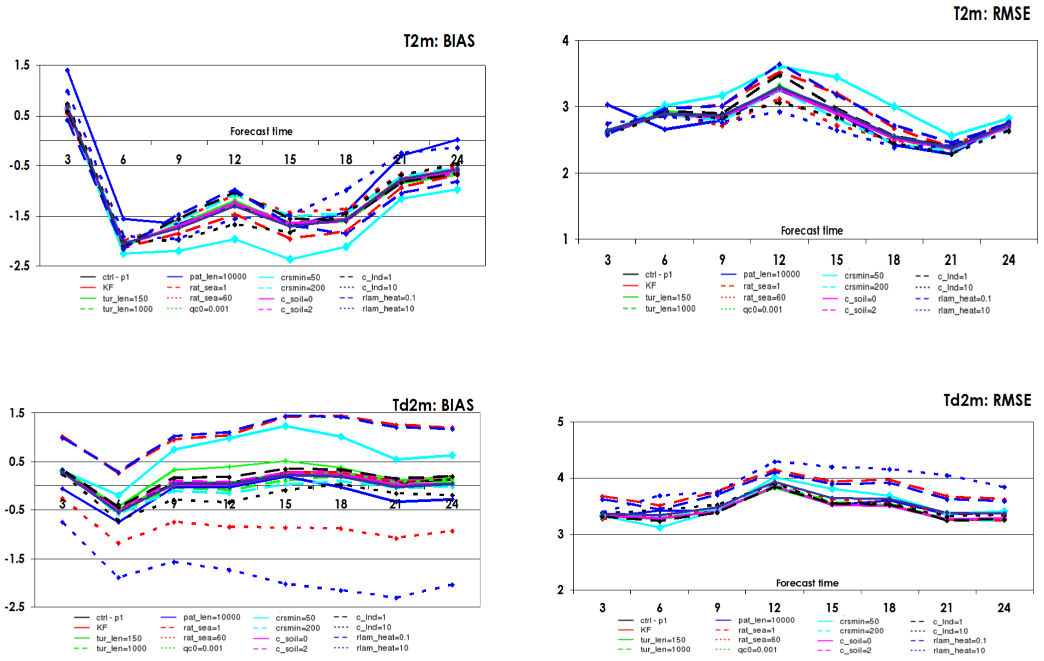
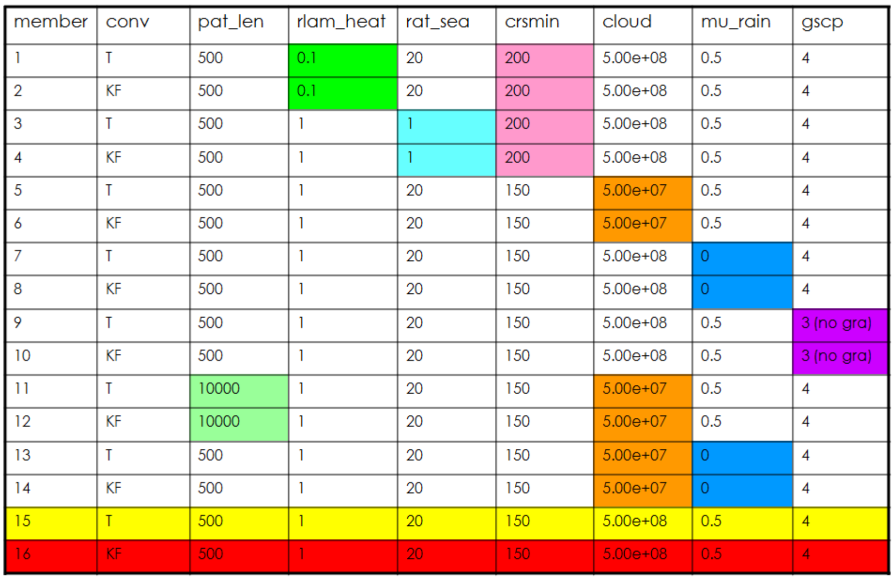


Figure 4: Bias and Mean Absolute Error of 2m T and Td for the CSPERT run, computed over Greece against SYNOP station data, for JJA 2008.



member	conv	pat_len	rlam_heat	rat_sea	crsmin	cloud	mu_rain	gscp
1	T	500	0.1	20	200	5.00e+08	0.5	4
2	KF	500	0.1	20	200	5.00e+08	0.5	4
3	T	500	1	1	200	5.00e+08	0.5	4
4	KF	500	1	1	200	5.00e+08	0.5	4
5	T	500	1	20	150	5.00e+07	0.5	4
6	KF	500	1	20	150	5.00e+07	0.5	4
7	T	500	1	20	150	5.00e+08	0	4
8	KF	500	1	20	150	5.00e+08	0	4
9	T	500	1	20	150	5.00e+08	0.5	3 (no gra)
10	KF	500	1	20	150	5.00e+08	0.5	3 (no gra)
11	T	10000	1	20	150	5.00e+07	0.5	4
12	KF	10000	1	20	150	5.00e+07	0.5	4
13	T	500	1	20	150	5.00e+07	0	4
14	KF	500	1	20	150	5.00e+07	0	4
15	T	500	1	20	150	5.00e+08	0.5	4
16	KF	500	1	20	150	5.00e+08	0.5	4

Figure 5: Set up to the CSPERT suite for the 2009 runs.

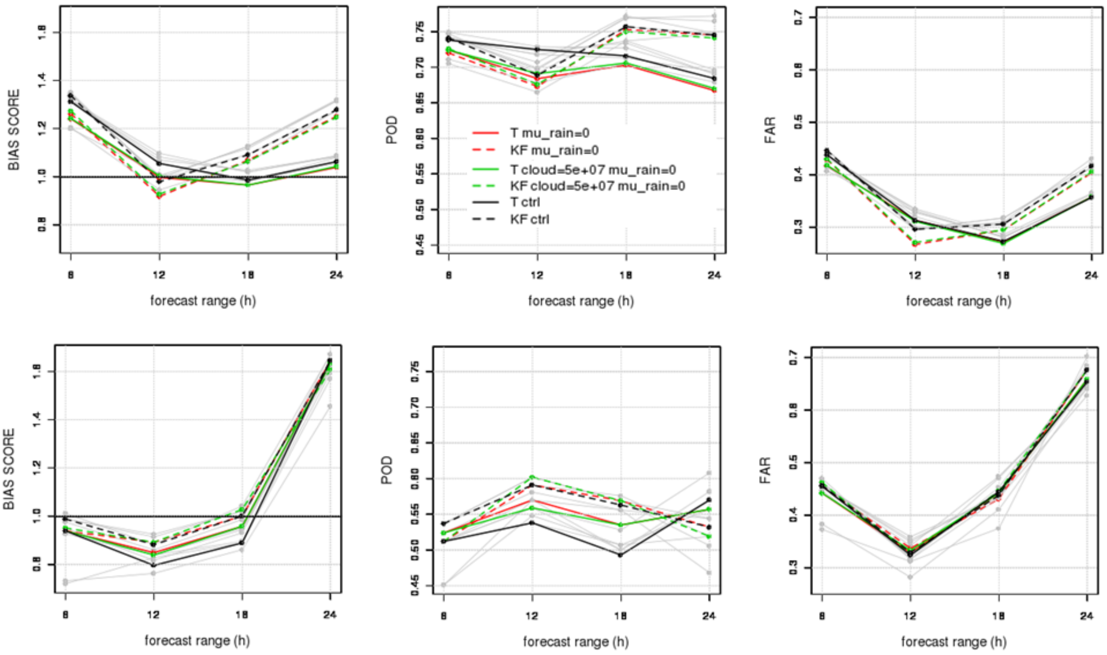


Figure 6: Bias Score, POD and FAR for the 6h-precipitation forecasted by some of the CSPERT runs, computed against high-resolution precipitation observations over Northern Italy, upscaled as average precipitation over boxes of 0.5 x 0.5 degrees exceeding 1 (top row) and 10 (bottom row) mm/6h. Scores are for spring 2009.

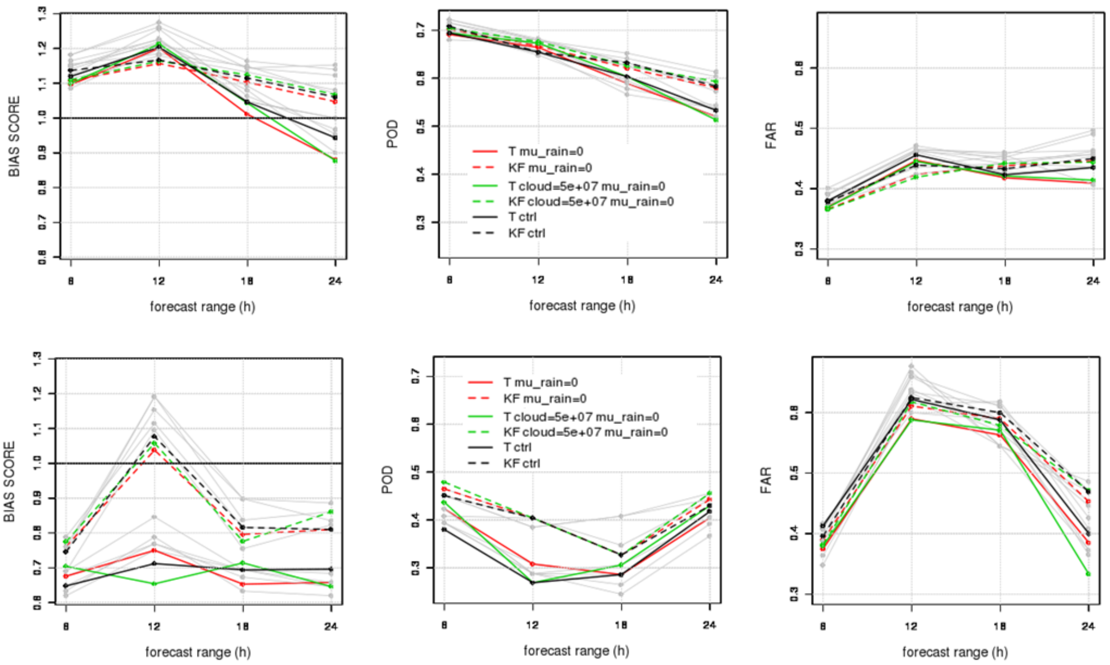
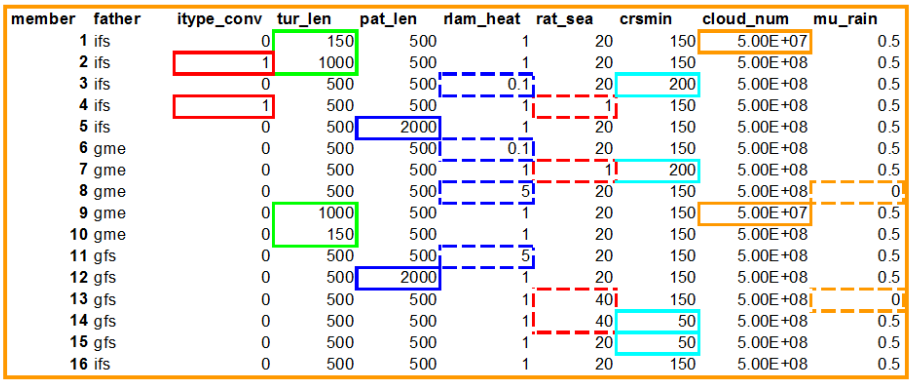


Figure 7: Bias Score, POD and FAR for the 6h-precipitation forecasted by some of the CSPERT runs, computed against high-resolution precipitation observations over Northern Italy, upscaled as average precipitation over boxes of 0.5 x 0.5 degrees exceeding 1 (top row) and 10 (bottom row) mm/6h. Scores are for autumn 2009.



member	father	itype_conv	tur_len	pat_len	rlam_heat	rat_sea	crsmin	cloud_num	mu_rain
1	ifs	0	150	500	1	20	150	5.00E+07	0.5
2	ifs	1	1000	500	1	20	150	5.00E+08	0.5
3	ifs	0	500	500	0.1	20	200	5.00E+08	0.5
4	ifs	1	500	500	1	1	150	5.00E+08	0.5
5	ifs	0	500	2000	1	20	150	5.00E+08	0.5
6	gme	0	500	500	0.1	20	150	5.00E+08	0.5
7	gme	0	500	500	1	1	200	5.00E+08	0.5
8	gme	0	500	500	5	20	150	5.00E+08	0
9	gme	0	1000	500	1	20	150	5.00E+07	0.5
10	gme	0	150	500	1	20	150	5.00E+08	0.5
11	gfs	0	500	500	5	20	150	5.00E+08	0.5
12	gfs	0	500	2000	1	20	150	5.00E+08	0.5
13	gfs	0	500	500	1	40	150	5.00E+08	0
14	gfs	0	500	500	1	40	50	5.00E+08	0.5
15	gfs	0	500	500	1	20	50	5.00E+08	0.5
16	ifs	0	500	500	1	20	150	5.00E+08	0.5

Figure 8: Final set up to the COSMO-SREPS suite.

(April to June) for three years (2007-2009). This spring-to-summer period can provide the necessary soil moisture variability in the data. These data (water content for each soil layer kg/m^2) are extracted at 8 different levels in the soil, namely 1, 2, 6, 18, 54, 162, 486, and 1458 cm. In this way, 241 deviation data files (=3years*3months*30days-29days=241 days) have been created, each of them containing 436,905 lines (grid points). Then an EOF analysis was developed and implemented to the daily soil water content data. The EOF analysis was implemented to the first soil layer from the surface. The perturbations are created through the following equation suggested by Houtekamer, 1993:

$$\varepsilon_j = \sum_{i=1}^{N=244} d_i \lambda_i \sigma_i \tag{1}$$

where, ε_j is the j-th perturbation, d_i a standard normally distributed random number, λ_i the square root of the eigenvalues and σ_i the corresponding eigenvectors. In order to solve the above equation a method for creating random numbers was used based on Press et al., 1992. Routines calculating EOFs have been found and selected the one by Ziemke J.R. (http://acdb-ext.gsfc.nasa.gov/People/Ziemke/) who has based his code on Kutzbach, 1967. This routine has been built to work for matrices with moderate size. Therefore, the main problem encountered with the available data is that the routine could not handle the extremely large data matrices (436,905 lines x 436,905 rows) as they require a very large stack memory. For this reason it was necessary to find solutions to overcome this problem. An efficient method was proposed by von Storch and Hannoschock, 1984 and Legler, 1984. They proposed to inverse the matrices, i.e. if initially there are M lines (grid points) and N days, with N $\ll$ M, it is possible to invert the problem and instead of creating MxM matrices to end up with NxN matrices that would lead to much less computationally intensive problem. Then, the input to the EOF analysis has been changed to an NxM data matrix in which there are N rows (corresponding to the available days) and M columns (corresponding to the grid points). After this change, the EOF analysis can run without stack memory problems. Then work has been performed on:

- adding the calculated "perturbations" to the initial soil water content file
- calculating the resulting fractional soil moisture
- constraining it to between 3% and 100% and then
- get the final perturbed soil moisture fields

The perturbations are calculated from:

$$P = d\Lambda c = \begin{pmatrix} \sum_{i=1}^{M} d_i c_{1,i} \\ \sum_{i=1}^{M} d_i c_{2,i} \\ \vdots \\ \sum_{i=1}^{M} d_i c_{M,i} \end{pmatrix} \tag{2}$$

where "d" are the random numbers , "c" are the space eigenvectors and "lambdas" are the space eigenvalues. Each line corresponds to each point in space and the sum is over the 244 EOF categories (if all are kept). In order to solve the equation a method for creating random numbers was used, based on Box-Muller method for generating random deviates with normal distribution (Press et al., 1992).

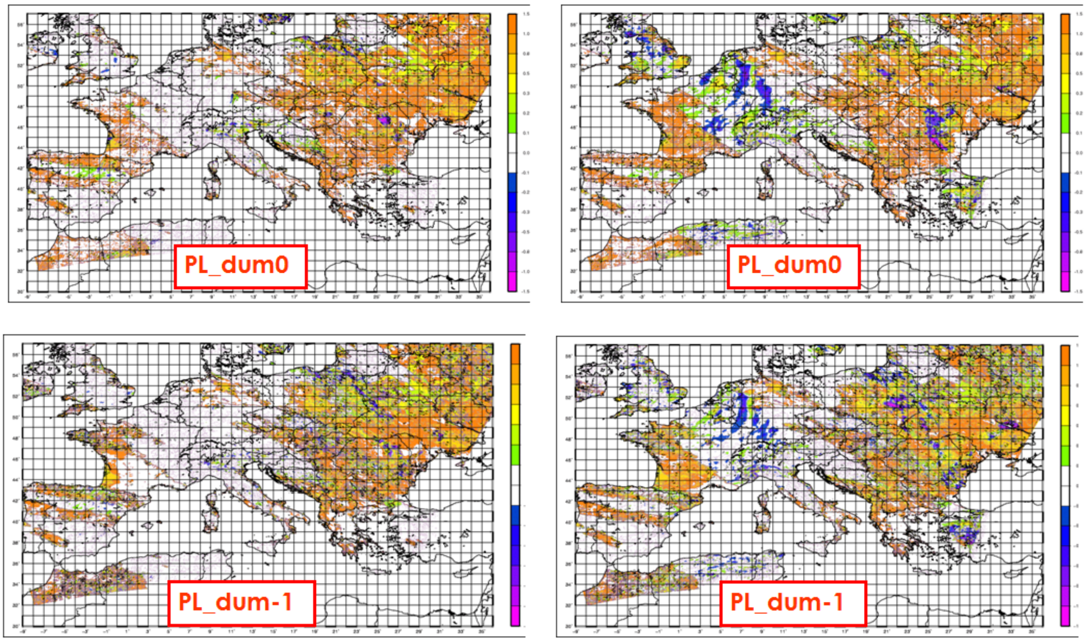


Figure 9: Differences of the soil moisture fields at the first level between perturbed runs (for different seeds, as indicated) and the control, at +24h (left) and +48h (right) forecast range.

Once the perturbed fields of soil moisture (on 7 levels) has been obtained, they have been tested in COSMO runs (Figure 9 and Figure 10). Different random seeds were used. The perturbations obtained (one set per each soil level) are added (the perturbations can be positive or negative) to the initial (original) soil moisture fields. A constraint is imposed: the perturbed value should not be more than 50% of the original value. Perturbations of the soil moisture is performed at all the soil levels (except the climatological level). COSMO was run on one case (9th June 2011) in the COSMO-SREPS configuration with perturbed soil moisture fields, for each perturbation set (no random number, seed=-1, seed=-5), at +48-hour. COSMO was also run with unperturbed soil moisture, for comparison (control run). Results have been presented at the COSMO GM 2011 (WG7 parallel session). As for the soil moisture difference, it is highlighted that positive values correspond to wetter soil layer for the CTRL run. The areas with positive values seem to be more than those with negative values at t+24h. For the particular date it seems that the method leads mostly to differences over East and some parts of West Europe although there are also differences in smaller areas elsewhere. At t+48h, the differences with the CTRL are greater and appear generally over the whole domain. There are also some changes among the results from the tests with various random numbers used (which could lead to different members of an ensemble). The control run is more wet in eastern Europe. Then, the control has more LHF than the perturbed, which is coherent with the fact that the control is wetter. For the SHF, this depends very much from the hour of the day. During the night (as it is at +24h, since the run starts at 00) it should be more from the ground towards the atmosphere, which means negative. Hence, since the difference it is positive, this may indicate that the SHF of the perturbed is stronger than that of the control, which is in agreement with the fact that in the perturbed the LHF is reduced, so the two are complementary. As for 2mT, the control is colder than the perturbed (and wetter in the soil and with more LHF and less SHF). The differences (positive or negative) between the CTRL run and the new runs in precipitation are larger at +48h, and they interest larger areas of the domain (Figure 11). This technique has not been implemented yet in an ensemble configuration for testing. Its applicability will be further investigated in the future.

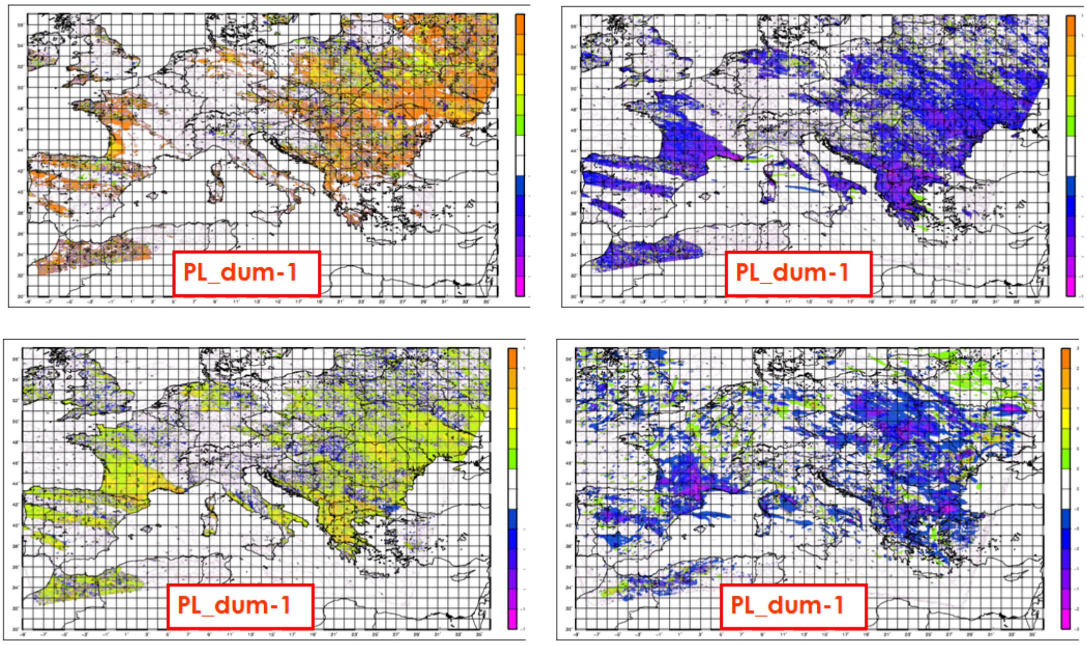


Figure 10: Difference between control and perturbed at +24h for: soil moisture (top left), latent heat flux (top right), sensible heat flux (bottom left) and 2m temperature (bottom right).

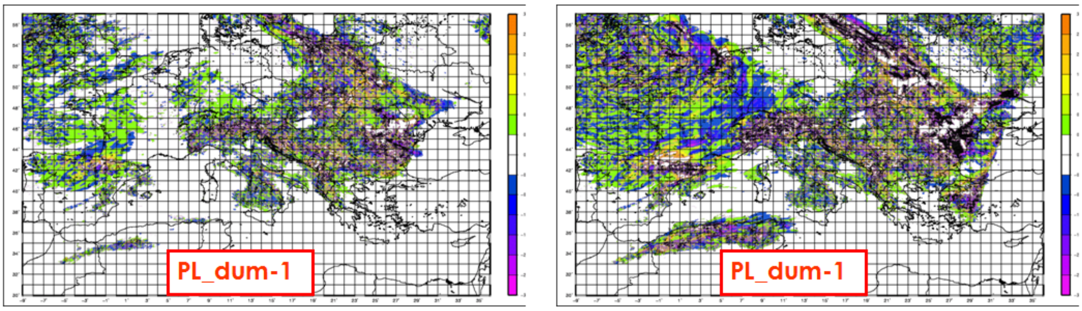


Figure 11: Difference in accumulated precipitation between the perturbed run and the control at 24h (left) and 48h (right) forecast range.

4 Ensemble merging

4.1 Clustering applied to a multi-ensemble

In the framework of the TIGGE-LAM coordination, it is proposed to analyse the possible benefit of a multi- ensemble for driving the limited area ensemble. More specifically, it is studied if a downscaling of a combination of more than one global ensemble can bring benefit with respect to downscaling a unique global ensemble. In this study, the member selection technique which is used to select the EPS members driving COSMO-LEPS (Molteni et al., 2001, Marsigli et al., 2001) is used to select members in the multi-ensemble made ECMWF EPS and the global MOGREPS ensemble of UKMO. The following global ensemble are considered:

- EPS (50+1): 51 members
- MOGREPS (23+1): 24 members
- MINI-MIX (EPS24 +MOGREPS24): 48 members
- MEGA-MIX (EPS51 +MOGREPS24): 75 members

For each of the above-mentioned ensembles, a 16-member cluster analysis and representative-member identification is performed. Then, 10-member global ensembles (EPS_REDU, MOGREPS_REDU, MINI_REDU, MEGA_REDU) are thus generated and the properties of the "REDU" ensembles are studied. It is worth pointing out that ECMWF EPS control run was always included in ECMWF24. More specifically, the following questions will be addressed:

- how many times do the 2 ensembles mix ?
- where do the best (and the worst) elements of "REDU" ensembles come from? How to they score depending on their "origin" ?
- how do "REDU" ensembles rank with respect to EPS, MOGEREPS, MINI-MIX and MEGA-MIX ?

The investigation was carried out over the period March to May 2009 (MAM09) for a total of 184 cases, considering both 00 and 12UTC runs of the two systems, as archived in TIGGE database. Z500 at fc+96h was used as clustering variable. As verifying analysis (at 00 and 12 UTC), a "consensus analysis" (average of UKMO and ECMWF high-resolution analyses) was taken. The performance of the MINIMIX48 ensemble was assessed in terms of root-mean-square error (RMSE) and anomaly correlation coefficient (ACC) of the geopotential height at 4-day forecast time (fc+96h), averaged over a domain covering Central and Southern Europe (30-60N, 10W-30E). The skill of MINIMIX24 is compared to that of UKMO24 and ECMWF24, using the common "consensus analysis". Figure 12 shows the time-series of the RMSE for the 3 ensembles means: a marked day-to-day variability is evident, with peaks of high (low) values of the score, indicative of particularly inaccurate (accurate) forecasts. It can also be noticed that ECMWF24 tends to outperform UKMO24, with lower values of RMSE for most of the days under investigation. As expected, the RMSE of the MINIMIX ensemble mean tends to stay somewhat in between the RMSE of the two other ensembles. It is worth pointing out that the use of MINIMIX turns out to be particularly useful in the days of failure of the more skilful ECMWF24. If we consider all 184 cases, the average RMSE of

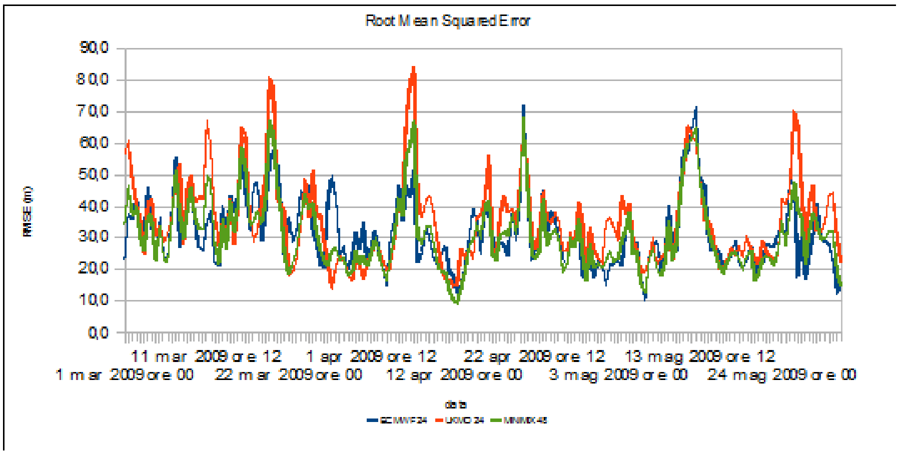


Figure 12: Time-series of the rmse (in metres) of the ensemble means for UKMO24 (red), ECMWF24 (blue) and MINIMIX48 (green) in terms of geopotential height at 500 hPa at fc+96h, averaged over the domain 30-60N, 10E-30W.

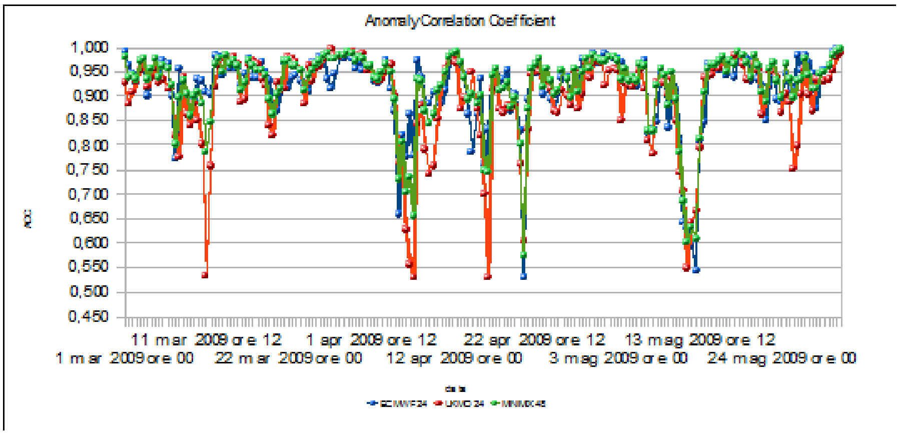


Figure 13: The same as Fig. 9, but for the anomaly correlation coefficient.

MINIMIX is slightly lower than ECMWF24, indicating the added value of multi-model even a low-population version of a very skilful global ensemble is used to generate a multi-model system.

Similar conclusions can be drawn if the ACC is considered (Figure 13). This score is sensitive to how the climate is computed. In fact, the ACC measures the correlation between the forecast and analysed deviations from climate. As an estimate of the climate, we used the three years of analyses present in TIGGE database and only the MAM season was considered. Also the ACC coefficient varies a lot from day to day and, as before, ECMWF24 tends to have better scores than UKMO 24, as shown by the higher values in ACC. It is again shown that the ACC of MINIMIX ensemble mean is, on average, slightly higher than that of ECMWF24, while UKMO24 has generally lower correlation coefficient. As a measure of the ensemble properties, we also considered the ensemble standard deviation (usually referred to as “spread”) around the ensemble mean for the 3 ensembles. Figure 14 shows the time-series evolution of the spread for ECMWF24, UKMO24 and MINIMIX48. This quantity, rather than a pure verification score like RMSE and ACC, reflects the properties of the ensembles, more precisely the extent to which the different ensemble members get more and more differentiated as the forecast step increases. It is immediately evident that, at least for this

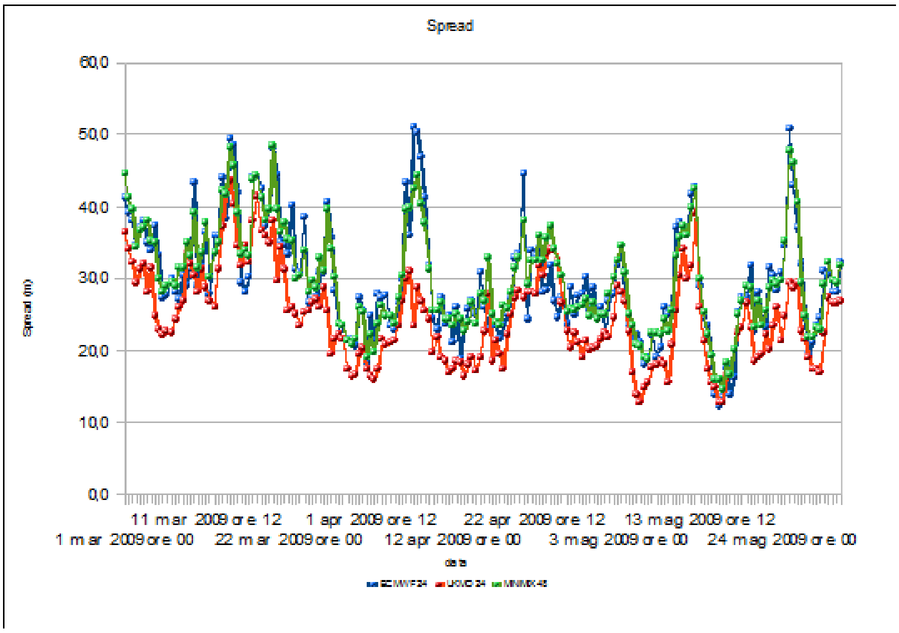
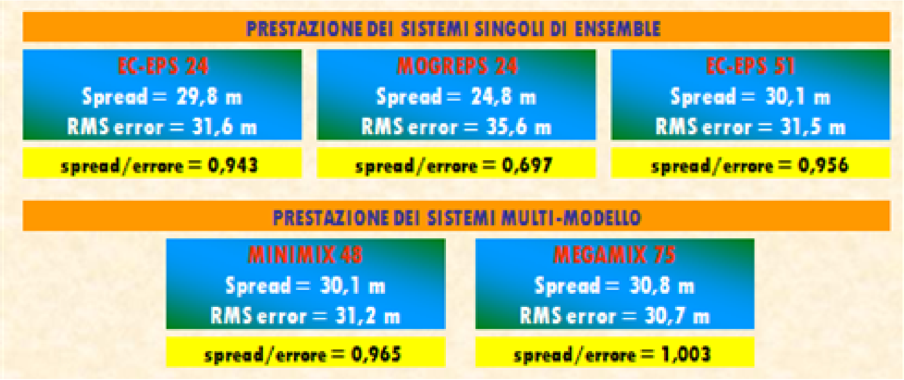


Figure 14: The same as Fig.9, but for the ensemble standard deviation (spread) among the members of the ensembles.

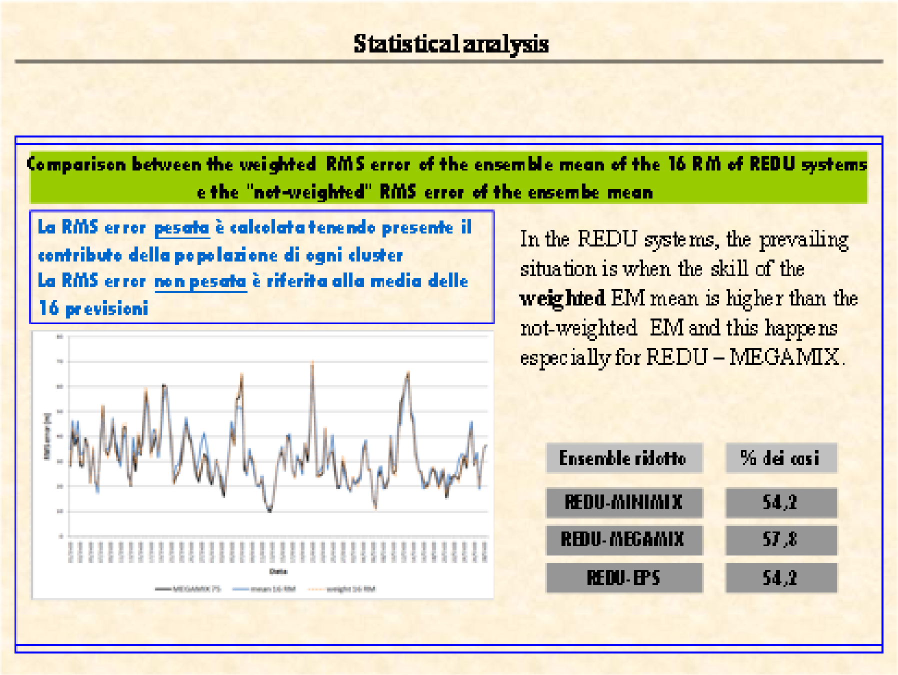

PRESTAZIONE DEI SISTEMI SINGOLI DI ENSEMBLE

EC-EPS 24	MOGREPS 24	EC-EPS 51
Spread = 29,8 m RMS error = 31,6 m	Spread = 24,8 m RMS error = 35,6 m	Spread = 30,1 m RMS error = 31,5 m
spread/errore = 0,943	spread/errore = 0,697	spread/errore = 0,956

PRESTAZIONE DEI SISTEMI MULTI-MODELLO

MINIMIX 48	MEGAMIX 75
Spread = 30,1 m RMS error = 31,2 m	Spread = 30,8 m RMS error = 30,7 m
spread/errore = 0,965	spread/errore = 1,003

Figure 15: Synthesis of the performance for single-model and multi-model systems in terms of Z500 rmse and spread.



Statistical analysis

Comparison between the weighted RMS error of the ensemble mean of the 16 RM of REDU systems e the "not-weighted" RMS error of the ensembe mean

La RMS error pesata è calcolata tenendo presente il contributo della popolazione di ogni cluster
La RMS error non pesata è riferita alla media delle 16 previsioni

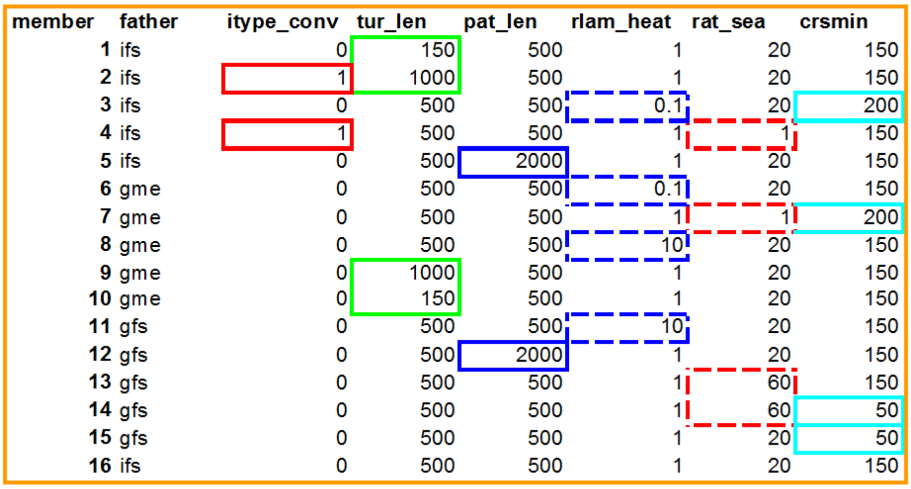
In the REDU systems, the prevailing situation is when the skill of the **weighted** EM mean is higher than the not-weighted EM and this happens especially for REDU – MEGAMIX.

Ensemble ridotto	% dei casi
REDU-MINIMIX	54,2
REDU-MEGAMIX	57,8
REDU-EPS	54,2

Figure 16: Synthesis of the performance for the REDU-ensembles.

forecast range, ECMWF24 has systematically larger spread than UKMO24. This is possibly related to the different strategies followed to generate the initial perturbations in the two ensembles (Singular Vectors for ECMWF24 versus Ensemble Transform Kalman Filter for UKMO24), although other interpretations are being put forward. As a further comment, it can be noticed that the MINIMIX48 ensemble has a spread very close to that of ECMWF24. In some cases, the MINIMIX48 spread is the largest one, suggesting that the use of a multi-model system allows the exploration of a wider region of the phase space of the atmosphere. If also the performances of EC-EPS51 (the full ECMWF-EPS system), MOGREPS24 (the full UKMO system) and MEGAMIX75 are considered, then the results indicate that the latter system outperforms all the others and that MINIMIX48 performs better than EC-EPS51 in terms of RMSE of the ensemble mean (see also Figure 15). As for the evaluation of the skill of the 16-member REDU ensembles, the attention is focus on the performance of the ensemble mean of these systems, calculated in the different ways: EM_w is the ensemble mean obtained weighting each member with the population of the cluster, while EM_now is the ensemble mean computed in the "classical way", that is without any weight. The attention is focussed of REDU-MINIMIX, REDU-MEGAMIX and REDU-EPS and results are summarised in Figure 16, which indicates a higher skill for the REDU systems where the ensemble mean is weighted according to the cluster population. This happens especially for the "best" system: REDU-MEGAMIX. Results can be summarized as follows:

- MINIMIX (48 members) performs better than EPS51 in terms of z500 RMSE of the ensemble means
- REDU-MEGAMIX outperforms the others ensembles in terms of "RMSE of the best element"
- REDU-MINIMIX performs slightly better than EPS51 in terms of "RMSE of the best element"
- for each of the 16-member ensembles, the weighted ensemble mean gives better score



member	father	itype_conv	tur_len	pat_len	rlam_heat	rat_sea	crsmin
1	ifs	0	150	500	1	20	150
2	ifs	1	1000	500	1	20	150
3	ifs	0	500	500	0.1	20	200
4	ifs	1	500	500	1	1	150
5	ifs	0	500	2000	1	20	150
6	gme	0	500	500	0.1	20	150
7	gme	0	500	500	1	1	200
8	gme	0	500	500	10	20	150
9	gme	0	1000	500	1	20	150
10	gme	0	150	500	1	20	150
11	gfs	0	500	500	10	20	150
12	gfs	0	500	2000	1	20	150
13	gfs	0	500	500	1	60	150
14	gfs	0	500	500	1	60	50
15	gfs	0	500	500	1	20	50
16	ifs	0	500	500	1	20	150

Figure 17: Set up of the COSMO-SREPS suite adopted for the comparison with COSMO-LEPS (winter 2010/2011 and spring 2011).

than the non-weighted one

4.2 Multi-model vs single-model approach for IC and BC to ensemble

The Consortium was running two mesoscale ensemble system, COSMO-LEPS, up to 6 days, mainly targeted for the early medium range, and COSMO-SREPS, which was developed for the aim of short-range forecasting (up to 48 h) in the last years (SREPS PP). One task of CONSENS was dedicated to decide what to do out of these two systems: - how is performing COSMO-SREPS for short-range forecasting, with respect to the already available COSMO-LEPS system? This also considering that recently the ECMWF EPS, on which COSMO-LEPS is based, has improved its spread-skill relation in the short-range, due to the introduction of the EDA.

- is it worth to run two separate systems at the same spatial scale, one for short-range only ?
- is it feasible to merge the two systems ? And if yes, how ?

The set-up of COSMO-SREPS is presented in Figure 17. An intercomparison of COSMO-LEPS and COSMO-SREPS in terms of short-range precipitation forecasting has been carried out, including also two combinations of the systems:

- a 20-member ensemble made up of the 16 COSMO-LEPS runs + 4 runs selected from COSMO-SREPS (mix20)
- a 16-member ensemble made up by the first 12 COSMO-LEPS runs + 4 runs selected from COSMO-SREPS (mix16)

The 4 runs selected from COSMO-SREPS are members 1 (nested on IFS), 6 (nested on GME), 11 (nested on GFS), 16 (control run). Verification is performed in terms of Probabilistic Quantitative Precipitation Forecast (PQPF), 6-hourly accumulated, over Northern Italy (more than 400 stations) for the winter 2010-2011 (20 Nov 2010 - 28 Feb 2011). The verification method is DIST (Marsigli et al., 2008), a distributional method. Some parameters

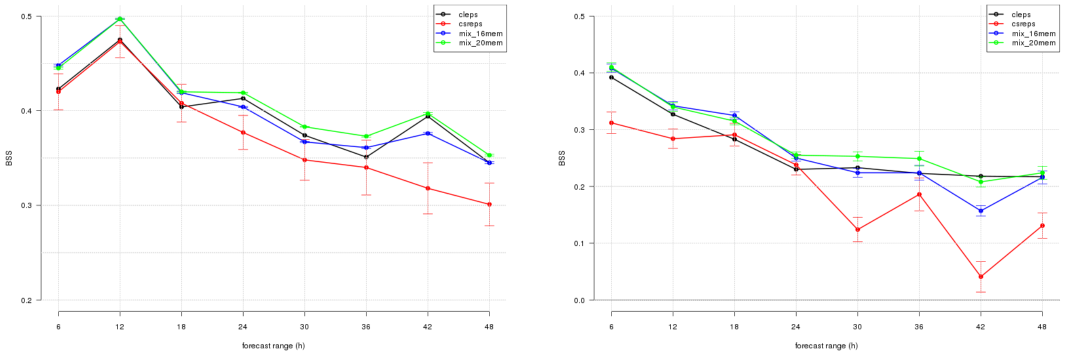
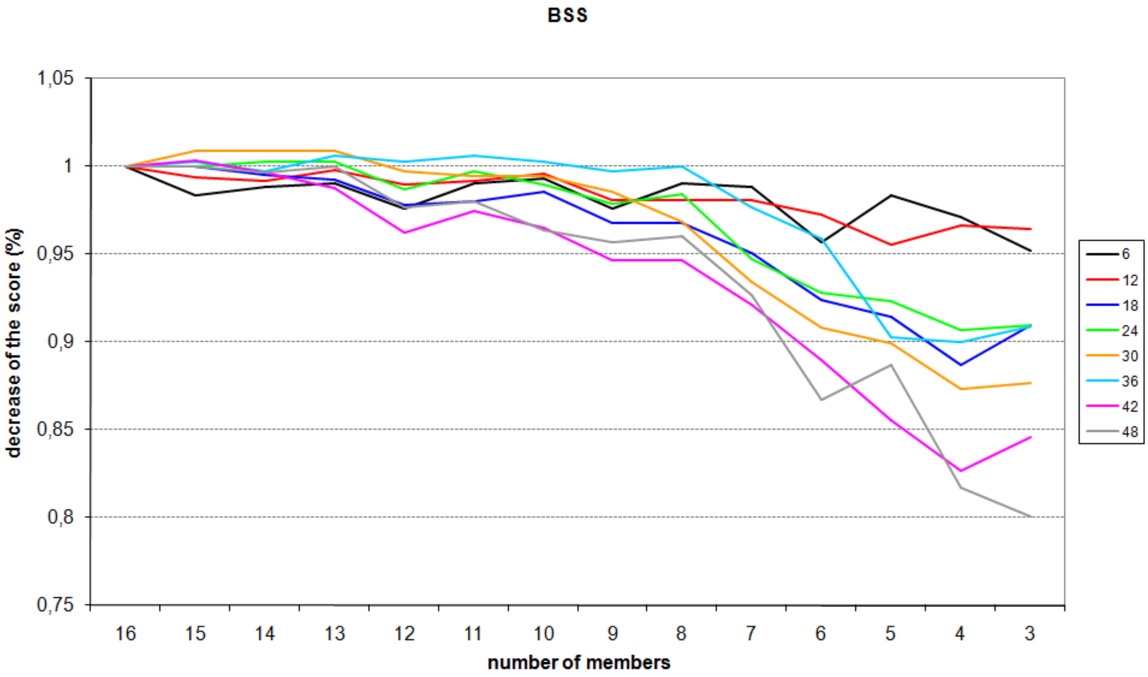


Figure 18: BSS as a function of the forecast range relative to the 4 ensemble systems for the event average precipitation exceeding 1 (left) and 5 (right) mm/6h.

of the distribution of both observed and forecasted precipitation are computed over boxes of 0.5 x 0.5 degrees size, then are compared. Results show that COSMO-LEPS performs better than COSMO-SREPS for almost all the forecast ranges, with some overlapping in the forecast range 18-24h (Figure 18). The mix16 ensemble outperforms COSMO-LEPS for the first 18 h in terms of BSS, while is comparable in terms of ROC area (not shown). For the higher threshold, mix16 performs better than COSMO-LEPS for the first 24. The mix20 ensemble performs better than COSMO-LEPS for almost the whole forecast range, especially for the higher precipitation threshold. Results are confirmed for different parameters of the precipitation distribution (not shown) and for a different season (spring 2011). It is shown that, for COSMO-SREPS, with only 3 global models providing initial and boundary conditions, the scores increase after 8 members is very limited (Figure 19). Therefore, it is not suggested to run an ensemble with more than 5-8 members with only 3 different perturbed initial and boundary conditions. Furthermore, the scores of both ensembles saturates around ensemble size 13-14. This implies that with a 16-member downscaling of the EPS we are already at the maximum skill attainable in the short-range, while there is some indication of room for improvement in the medium range. In summary, results (analysis of 6h precipitation over northern Italy for winter 2010/2011) indicates that:

- generally COSMO-LEPS outperforms COSMO-SREPS in terms of probabilistic indices
- the multi-model approach for i.c. and b.c. proves valuable even if model with different qualities are used
- comparing a 5-member COSMO-LEPS with a 5-member COSMO-SREPS (where all the 3 different i.c. and b.c. sets are included + 2 extra IFS-driven members), the two performs similarly
- comparing a 3-member COSMO-LEPS with a 3-member COSMO-SREPS (where the 3 members have the 3 different i.c. and b.c. sets), the reverse happens: COSMO-SREPS outperforms COSMO-LEPS
- hence, for the multi-model approach to be effective in providing i.c. and b.c., several models are needed to get a performance similar (or better) to a downscaling from a well constructed ensemble (like EPS)
- combining the 16 COSMO-LEPS members with 4 COSMO-SREPS members (taking 3 members with the 3 different i.c. and b.c. sets + the control), this 20-member ensemble outperforms both systems



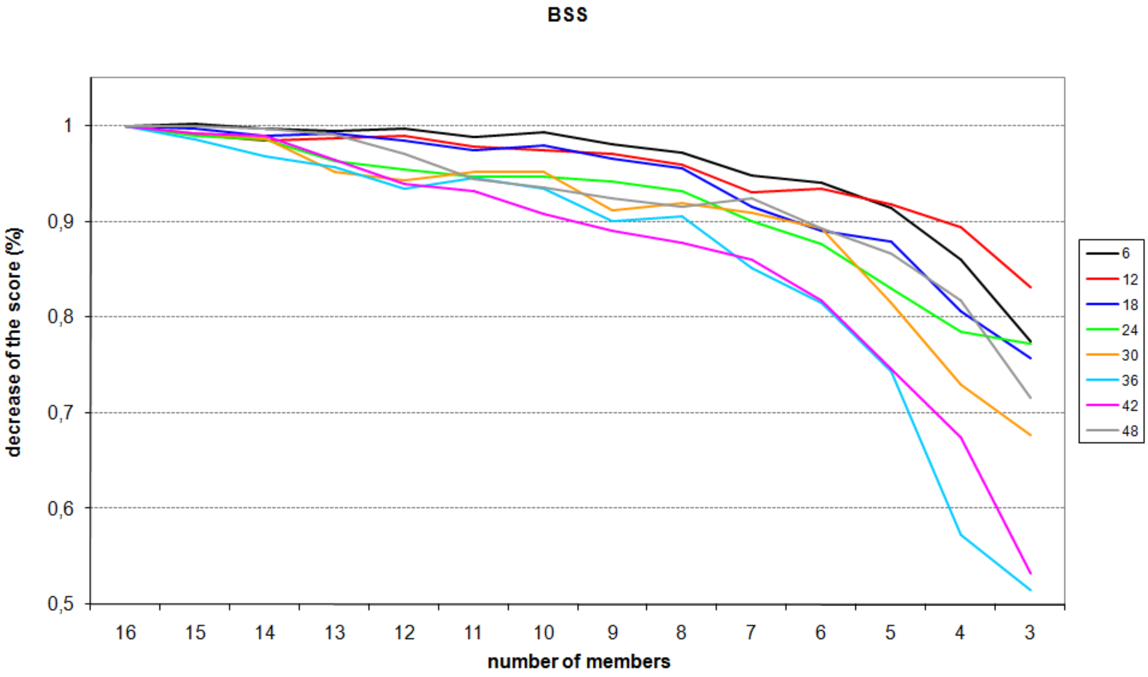


Figure 19: Dependence of the ensemble size of the BSS for COSMO-SREPS (top) and COSMO-LEPS (bottom).

- combining 12 COSMO-LEPS members with 4 COSMO-SREPS members (taking 3 members with the 3 different i.c. and b.c. sets + the control), this 16-member ensemble outperforms COSMO-LEPS for the first day (from 18 to 30 h)

A more comprehensive evaluation is described in Marsigli et al., 2013. On the basis of these results, it was decided to stop running the extra COSMO-SREPS members (nested on the same sets of IC and BC but with different physics), and to Keep only the COSMO-SREPS members which receive IC and BC by different global models, but prolonging the runs up to +132h, as COSMO-LEPS. Then, a new ensemble system has been created, COSMO-HYBEPS, by merging the 16 COSMO-LEPS runs with the 4 BC-EPS runs, creating an additional product which is currently under evaluation in a test suite.

5 Calibration

Three different calibration methods have been implemented and tested on 24h accumulated precipitation forecast issued by COSMO-LEPS. These techniques provide corrections based on: the cumulative distribution function (quantile-to-quantile mapping, hereafter, CDF), the linear regression (hereafter, LR) and analogs. The basic implementations of the CDF and LR methods were developed by generating a seasonal correction function for each model grid point. Each function was defined by using the historical data which were forecasted and observed for the considered grid point during each season. This approach generates correction functions that are location (i.e. grid point) specific. The analogs search was performed in terms of the similarity of the forecasted 24-h precipitation fields over the area under investigation (i.e. ER). The calibration strategy was based on historical forecast and observed rainfall data which were available over the area under investigation. Thirty years of reforecast of one member of COSMO-LEPS were run by MeteoSwiss. One reforecast run with a 90-h lead time was available every three days from 1971 to 2000. This model climatology is used to calibrate forecasts of all lead times, although forecasts of longer lead times might require a longer lead time reforecast data set. According to the model climatology, the climatological observed precipitation data were collected over the period 1971-2000 for ER and Switzerland; unfortunately, the observed data over Germany were available only for the period 1989-2000. Finally, the impact of the calibration process was also verified by performing the coupling of the ensemble QPFs with an hydrological model. This experiment was carried out only for a medium-sized catchment located in Emilia-Romagna, as other implementations of the meteo-hydrological coupled system were not available over the study areas under investigation.

5.1 Comparison of the calibration methodologies in terms of precipitation forecast

The impact of the calibration process on the 24-h QPFs operationally provided by COSMO-LEPS was verified for the years 2003-2007. Three study areas (Figure 20) have been analysed: the total area of Germany (about 357000 km^2), Switzerland (about 41000 km^2) and Emilia-Romagna (about 22000 km^2). Four different lead times were considered, ranging from day-2 to day-5.

According to the 24-h time window over which the observed rainfall amounts were available for each study area, these lead times correspond to, respectively, the 18-42, 42-66, 66-90 and 90-114 forecast hours for Germany and Switzerland and the 20-44, 44-68, 68-92 and 92-116

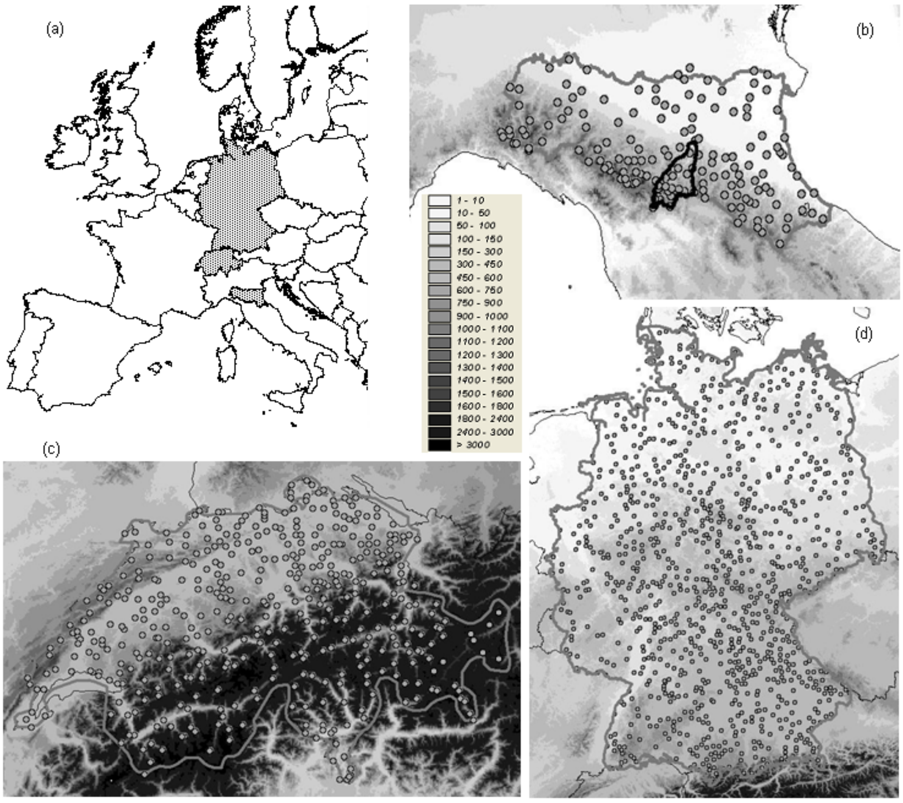


Figure 20: The three study areas localised in western Europe (a). Rain gauge locations (dots) of observed climatology and verification datasets for (b) Emilia-Romagna, northern Italy (in evidence the Reno river basin); (c) Switzerland and (d) Germany. The legend of the digital elevation model (in meters above sea) is common for the latter three panels.

forecast hours for Emilia-Romagna. The verification methods used here were the attributes diagram and the Brier Skill Score (BSS). The thresholds for the verified events were defined as percentiles of the climatological distribution of observed 24-h precipitation, different for each grid point. The use of percentiles of climatological distribution as thresholds guarantees that the sample points share the same climatological frequency of the event. This approach enables to avoid that chosen metrics might report unexpectedly large skill in some points due to the variation of the climatological event frequencies within the verification area. In the following of this Section, results are analysed for each study area separately. The forecast verification measures were first computed independently for each model grid point and were then averaged over the area under investigation. Attributes diagrams are shown for each season only at the day-2 lead time, considering the 80-th and the 95-th percentiles of 24-h climatological rainfall amount as verification thresholds. For brevity, we have chosen not to show attributes diagrams at other thresholds and lead times, as results do not depend strongly on the forecast range (forecasts show a slightly better reliability for the shorter lead times than for the longer lead times) and approximately shows a blend of the hereafter discussed diagrams, giving a little additional information only. Inset histograms for the frequency of usage were plotted on a log-10 scale, in order to provide a better visualization of the distribution in the tails. Results in terms of BSS are shown for each season, considering different forecast ranges (from day-2 up to day-5) and verification thresholds (the 80-th and 95-th percentiles of 24-h climatological rainfall amount). The observed climatology was used as the reference forecast for the computation of the BSS.

a Germany

  1 ATTRIBUTES DIAGRAM
  The raw ensemble provides poor performance at forecast probabilities lower than 60% in all the seasons; only forecasts of higher probabilities show reliability lines which lie above the no skill line, except for summer (Figure 21). The un-calibrated forecasts provided in autumn are better than forecasts provided in the remaining seasons. Generally, the calibration increases the forecast reliability at all the probabilities; the LR method is particularly effective in spring and summer. The best performances resulting from the forecast calibration are obtained in spring and winter. The improvement of the calibrated forecasts is somewhat less salient in summer, especially for probabilities greater than 60%. The calibrated forecasts show more reliability at lower probabilities for all the seasons. Generally, the raw and calibrated forecasts are overconfident; except for the forecast calibrated by the LR method which results underconfident at the lower probabilities of the higher threshold, especially in spring

  2 BSS
  The BSS values are always improved by calibrating the COSMO-LEPS system with the CDF method, for all the seasons and both the thresholds (Figure 22). The ANL and LR methods provide a less significant improvement, especially for the spring and summer seasons. The greatest amount of skill improvement is obtained in winter. Forecasts of lower precipitation events are slightly skilful than forecasts of higher precipitation events. Both the raw and calibrated forecasts show a remarkable decay of performance with lead time

b Switzerland

  1 ATTRIBUTES DIAGRAM
  The raw ensemble shows low reliability; un-calibrated forecasts for winter are particularly worse than forecasts for the other seasons (Figure 23), due to the

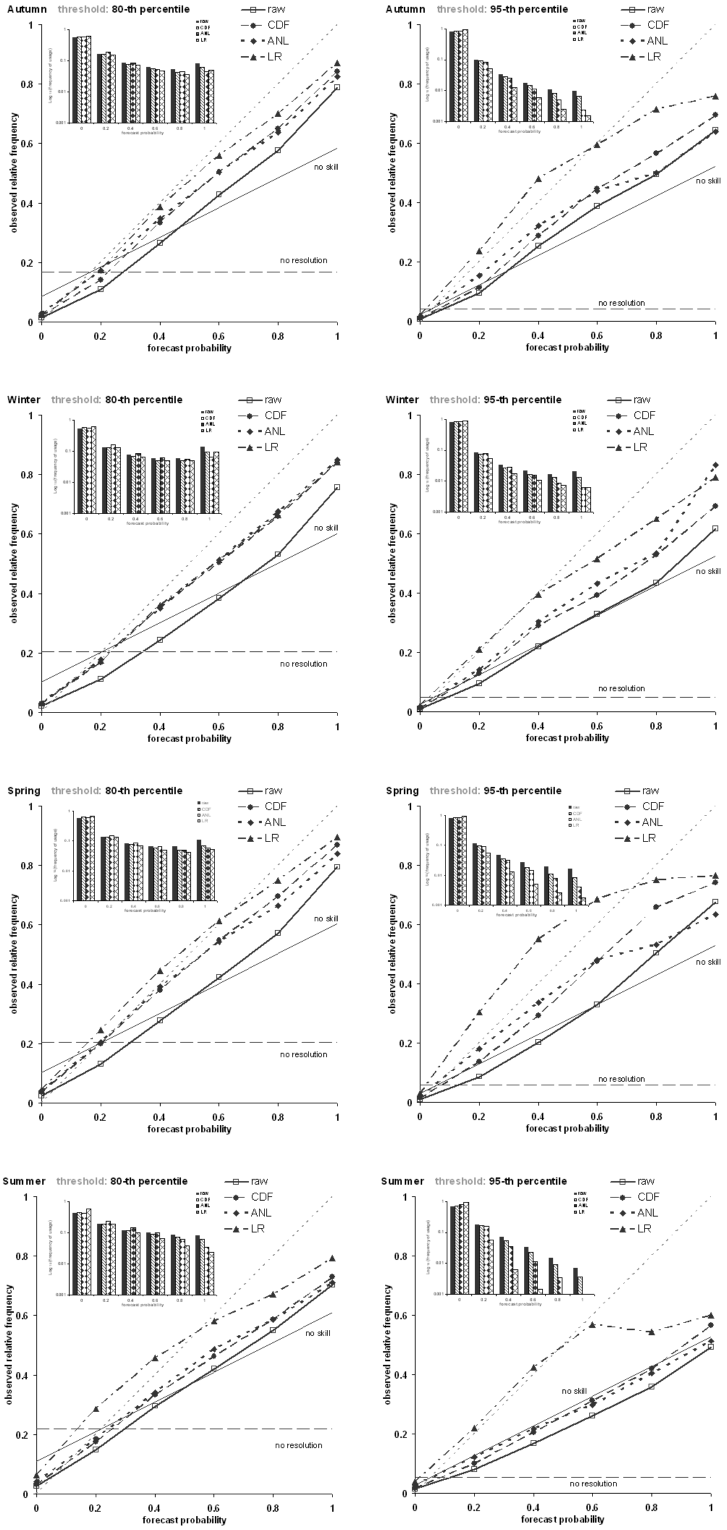


Figure 21: Attributes diagrams of the calibrated and raw forecasts over Germany at day2 lead time for the (left panels) 80-th and (right panels) 95-th percentile thresholds. The diagrams refer to (first row) autumn, (second row) winter, (third row) spring, (fourth row) summer. The inset histograms denote frequency of forecast usage of each probability bin.

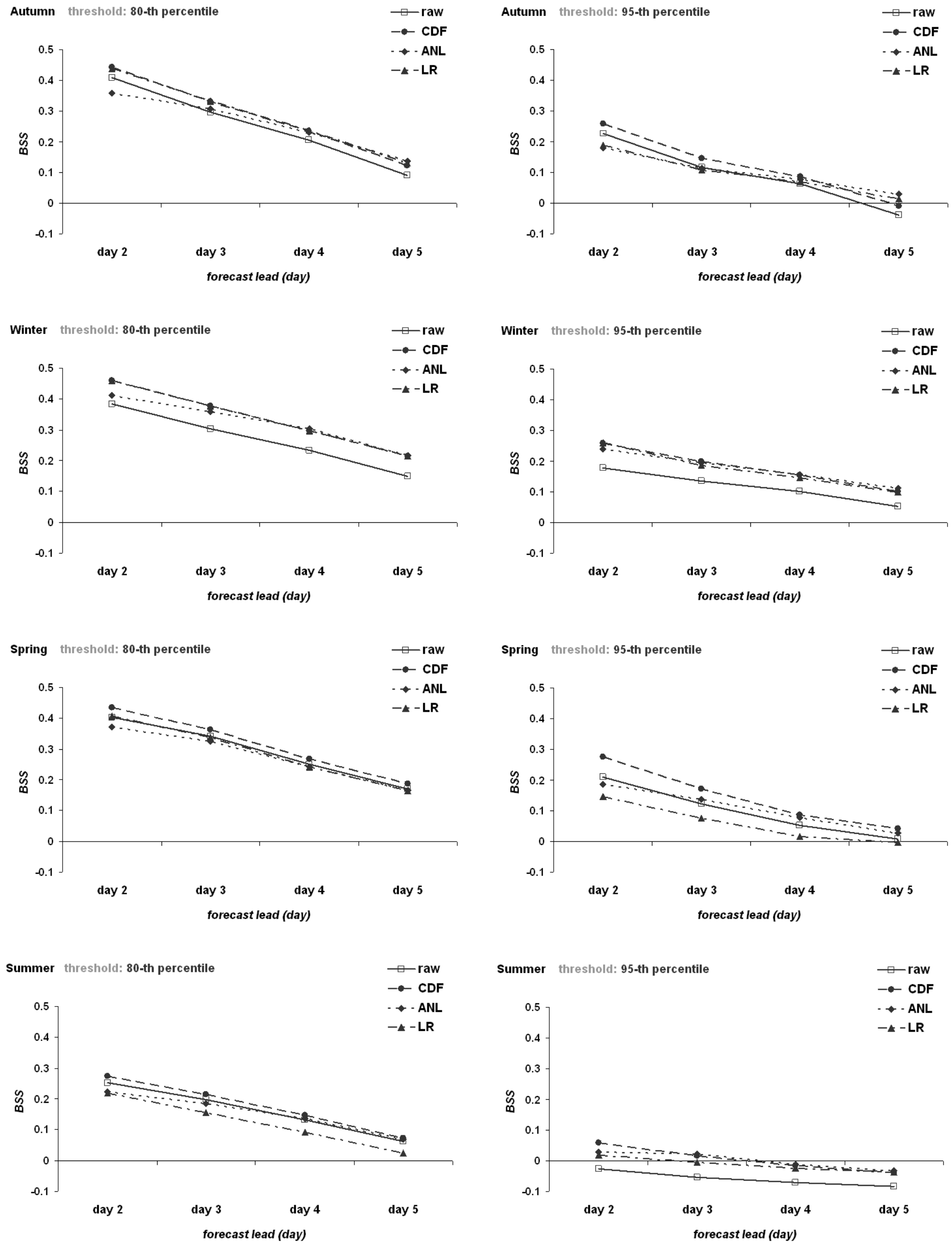


Figure 22: BSS at (left panels) 80-th and (right panels) 95-th percentile thresholds for the raw and calibrated ensembles over Germany in (first row) autumn, (second row) winter, (third row) spring, (fourth row) summer. The results refer to different forecast ranges, up to day5 lead time.

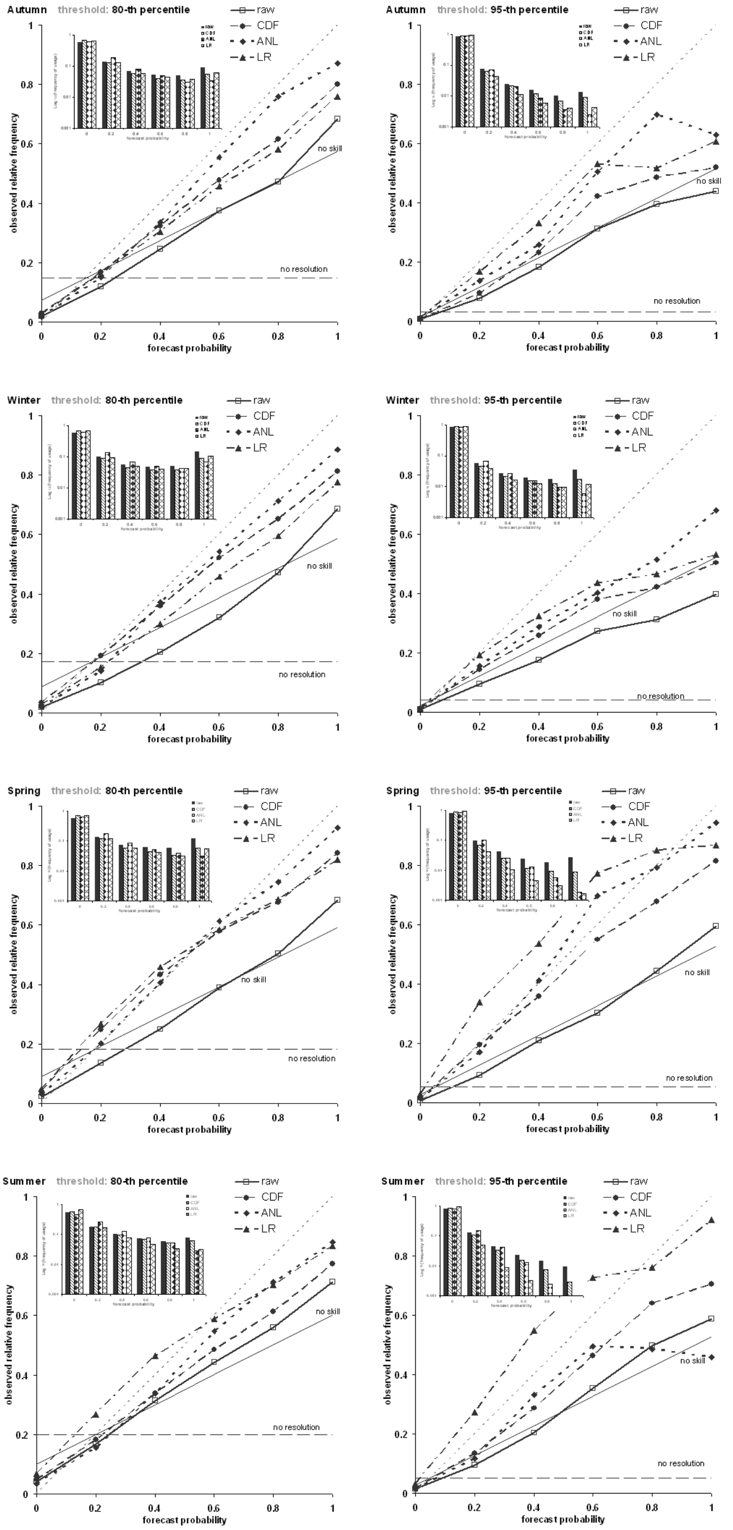


Figure 23: Same as Figure 21, but results refer to Switzerland.

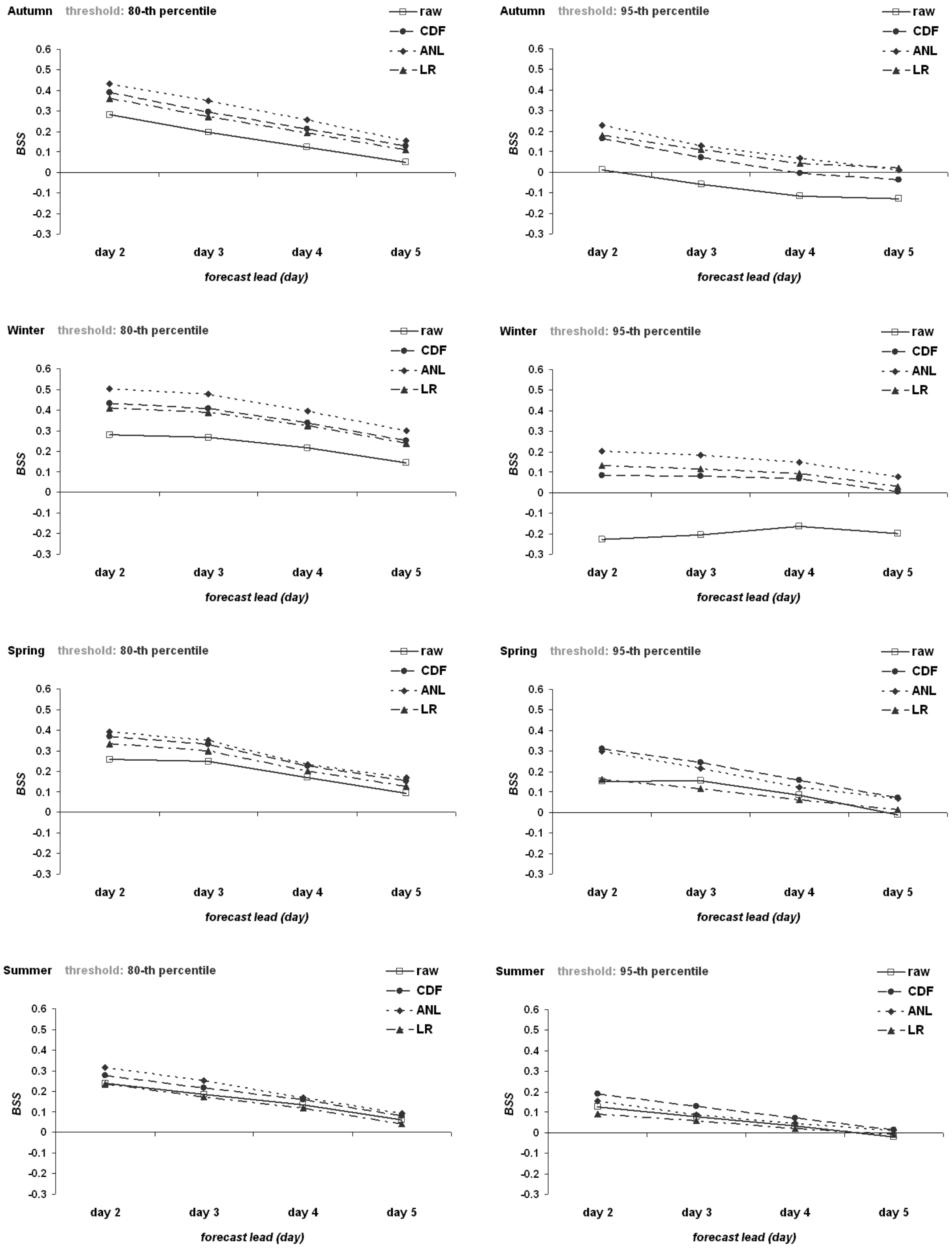


Figure 24: Same as Figure 22, but results refer to Switzerland.

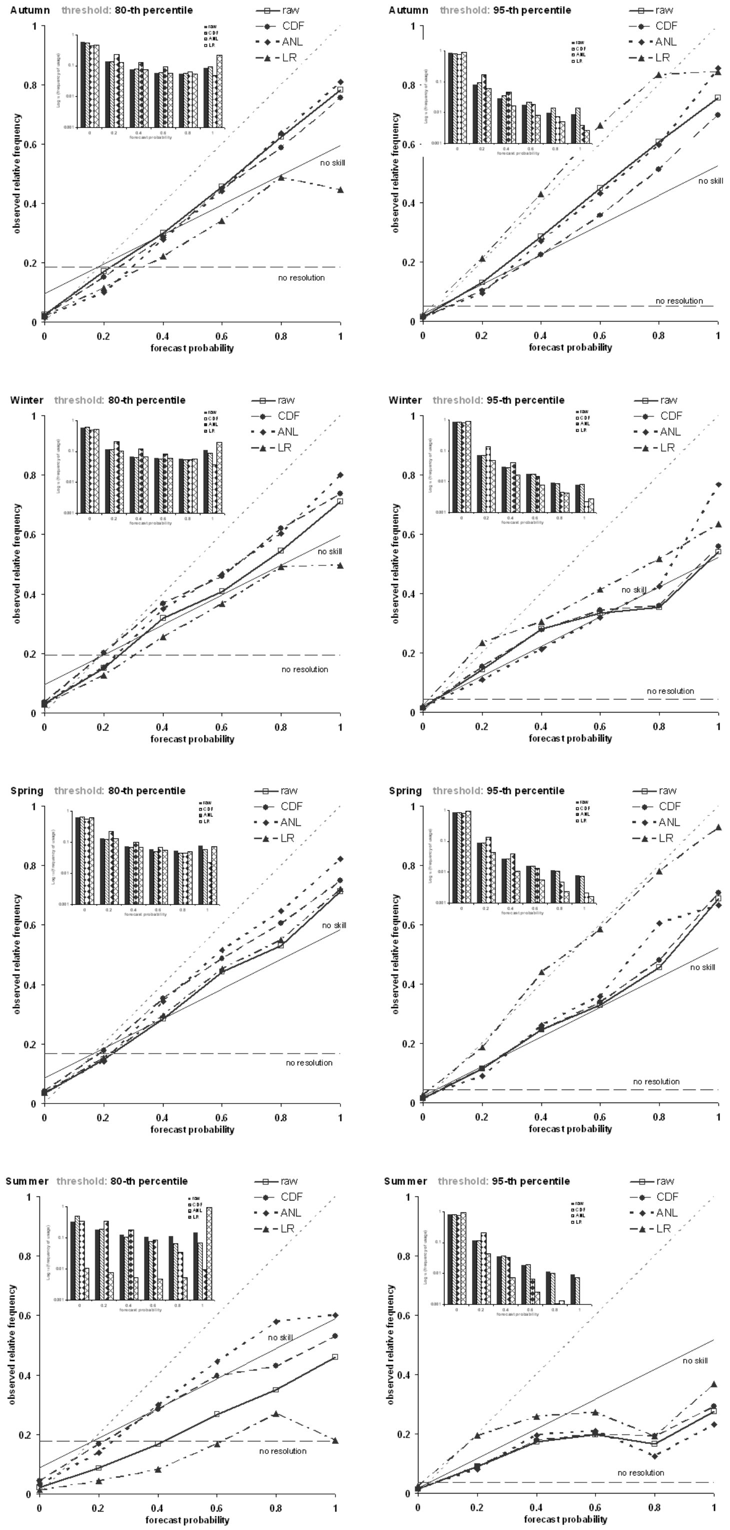


Figure 25: Same as Figure 21, but results refer to Emilia-Romagna.

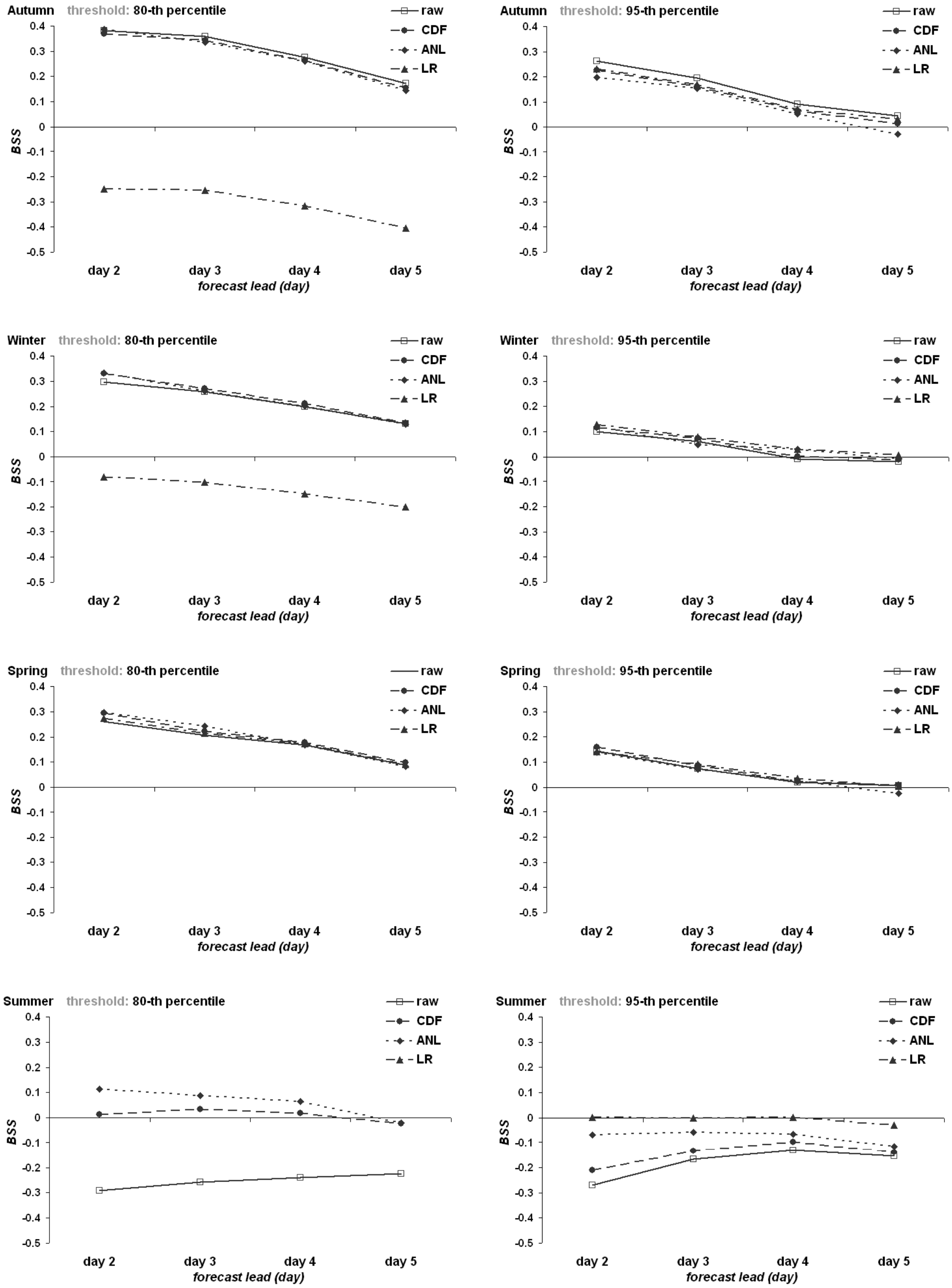


Figure 26: Same as Figure 22, but results refer to Emilia-Romagna.

larger systematic error which results in this season (Fundel et al. 2010). The calibration process enables to increase the reliability in all the seasons; generally, the ANL method performs better than the other methods (except for summer). Calibration is particularly effective in spring. The improvement of the calibrated forecasts is somewhat less salient in summer, as the raw forecasts have a relatively higher reliability. For all the seasons, the calibrated forecasts show more reliability at probabilities lower than 60%. Generally, both the raw and calibrated forecasts are overconfident in all the seasons, except for the forecasts calibrated by the LR method which result as underconfident in spring and summer at the lower probabilities

2 BSS
Generally, for all the seasons and both the thresholds, the BSS values are always improved by calibrating the COSMO-LEPS system with the CDF and ANL methods (Figure 24). The LR method proved to produce a less significant improvement, especially in spring and summer. A large amount of skill improvement is obtained in autumn and winter. Even, with respect to the higher threshold, unskilful raw forecasts for autumn and winter can be turned into skilful forecasts. The amount of skill improvement is smaller in spring and the smallest in summer. As explained in Fundel et al., 2010, the latter result can be due to the more convective and localized nature of precipitation events which occur in summer over Switzerland. During this season, systematic errors in the forecast system might be less important than spatial errors that cannot be corrected for with the proposed calibration methods. Forecasts of lower precipitation events are more skilful than forecasts of higher precipitation events. The decay of performance with lead time is quite evident for the raw and calibrated forecasts, especially for the lower threshold, and with a greater extent in autumn and spring

c Emilia-Romagna

1 ATTRIBUTES DIAGRAM
Generally, the raw and calibrated forecasts are overconfident in all the seasons (Figure 25). The raw ensemble has no good reliability; the un-calibrated forecasts provided in autumn are better than forecasts provided in the remaining seasons. This outcome can be ascribed to the higher predictability of precipitation in autumn, when precipitation events are more frequent and driven by large scale forcing. Raw forecasts for spring and winter show reliability lines which lie quite close to or under the no skill line; the higher threshold worsening this outcome. Even, raw forecasts for summer show reliability lines which lie quite under the no skill line. The results obtained by the forecast calibration process show high variability with respect to the season and the methods. The calibration does not substantially improve reliability in autumn, except for the LR technique at the higher threshold. However, a slight improvement is provided by the ANL method at the lower threshold for probabilities greater than 50%. Improvements are detectable in the winter and spring seasons at the lower threshold when the CDF or ANL calibration techniques are applied. At the higher threshold, only LR enables an increase of reliability, especially for spring. A little gain in reliability is obtained at the lower threshold for summer with the CDF and ANL methods, but results are still quite poor. As the performance of the calibration methods provide only slight improvements in such cases with respect to the raw forecast, a procedure based on a 1000-member block bootstrap sample was used, following Hamill, 1999, to test whether the possible benefits were statistically significant. Generally, the error bars showed a small magnitude, not overlapping to the raw

forecast line (not shown), revealing that the improvements can be evaluated as statistically significant, at a 95% level

2 BSS
The calibration process does not provide a clear beneficial impact on the ensemble QPFs (Figure 26), except for summer. For the autumn season, the values of BSS associated to the calibrated ensembles are lower than the BSS of the raw ensemble for all the lead times. For the winter and spring seasons, the BSSs associated to the calibrated ensembles are quite similar, sometimes slightly higher, with respect to the BSS of the raw ensemble. A clear improvement is obtained in summer, but the raw ensemble is quite unskilful and the calibrated ensembles perform similar to climatology. Generally, at the higher threshold, the calibration methods provide quite similar performances. Rather, at the lower threshold, the LR method provides very bad corrections with respect to the CDF and ANL methods (except for spring); even, very low negative BSS values (out of scale in the graph) result in summer. This outcome can be ascribed to the correction typical of the LR function. In particular, the value of the intercept strongly influences the calibrated value for lower QPF, providing a systematic increase of the raw forecast. Therefore, this aspect strongly penalizes the BSS when computed with respect to low QPF values as verification threshold, such as the 80-th percentile of observed climatology of Emilia-Romagna. Actually, this value is lower than 2.5 mm over the large part of Emilia-Romagna, especially in the plain areas. The scores for forecasts of lower precipitation events are slightly better than the scores of forecasts of higher precipitation events. The decay of performance with lead time is significant, except in summer

5.2 Verification of the calibration process by coupling QPFS with an hydrological model

The hydrological model used to generate simulated discharges is TOPKAPI (TOPographic Kine-matic Approximation and Integration), a physically-based distributed rainfall-runoff model. A detailed description of the model can be found in Liu and Todini, 2002. The performance of the meteo-hydrological coupled system has been evaluated over the Reno river basin (for a more detailed description of the basin refer to Diomede et al. 2008); in particular, the stream flow forecasts were evaluated at Casalecchio Chiusa, the closure section of its mountainous part which dimension is about 1000 km^2 (Figure 20 panel b). The stream flow forecast experiments covered the autumn and spring seasons in the period 2003-2008. Each hydrological run driven by COSMO-LEPS QPFs is 120-h long, starting at 1200 UTC on each day. Every hydrological forecast was run starting with basin initial conditions which were generated by running TOPKAPI in a continuous mode for the whole period 2003-2008 driven by observed precipitation and temperature data. Both QPF and rain gauge observations were downscaled to the hydrological model grid resolution (i.e., 500 x 500 m) by assigning to each cell, respectively, the QPF and the observed value provided on the nearest COSMO-LEPS grid point.

As the hydrological model runs with an hourly time step, hourly calibrated QPFs were used. For each grid point and 24-h lead time, the hourly calibrated QPFs were obtained from the hourly raw QPFs by using a scaling factor defined as the ratio of the calibrated and raw 24-h precipitation amounts. The hydrological simulations generated using rainfall observations were used as the reference measure, in order to assess the skill of the meteorological forecasts. Results of the meteo-hydrological coupling were investigated by statistical analyses. The analysis has firstly been carried out in terms of root-mean-square error (hereafter, RMSE). A

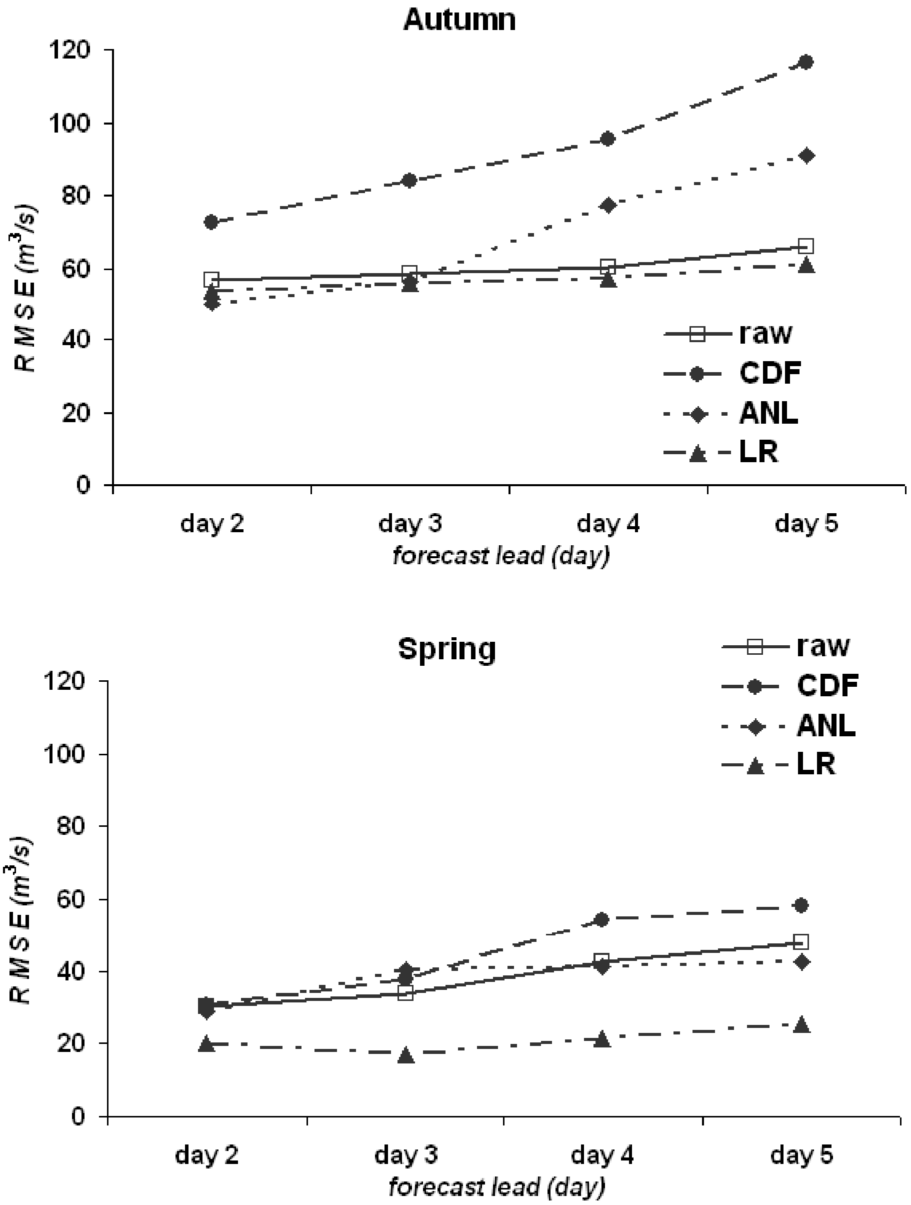


Figure 27: Root-mean-square error for the 95-th percentile of the ensemble of discharge forecasts provided at Casalecchio Chiusa in the autumn (panel a) and spring (panel b) seasons of the years 2003-2008. The discharge forecasts were driven by the QPFs provided by the raw ensemble (black line) and the ensemble calibrated by the CDF (dashed line with circles), ANL (dotted line with diamonds) and LR (dotted-dashed line with triangles) methods. The statistics are displayed as a function of the lead time of the forecast.

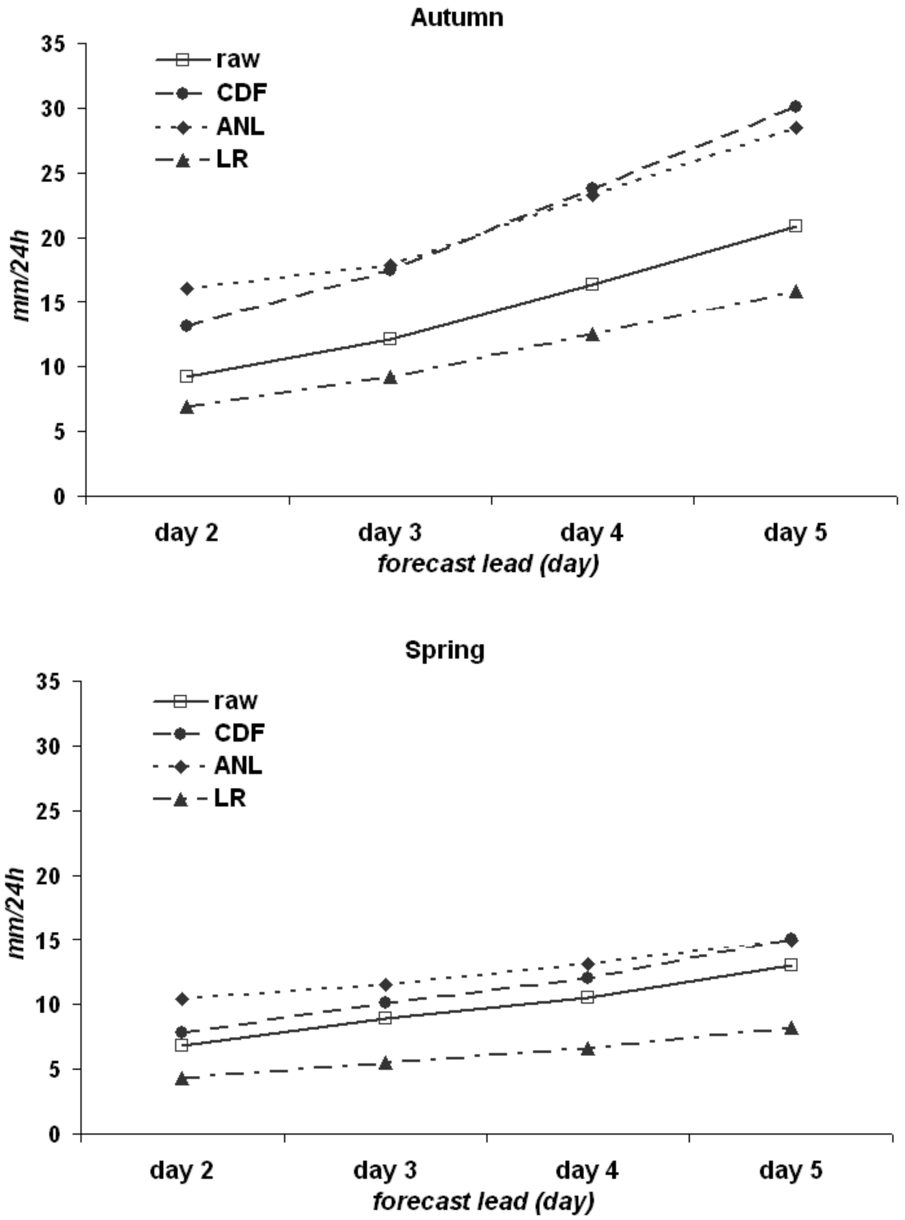


Figure 28: Spread of 24-h rainfall amount for the raw and calibrated ensembles, as a function of the lead time of the forecast. Results are shown for the autumn (panel a) and spring (panel b) seasons of the years 2003-2008.

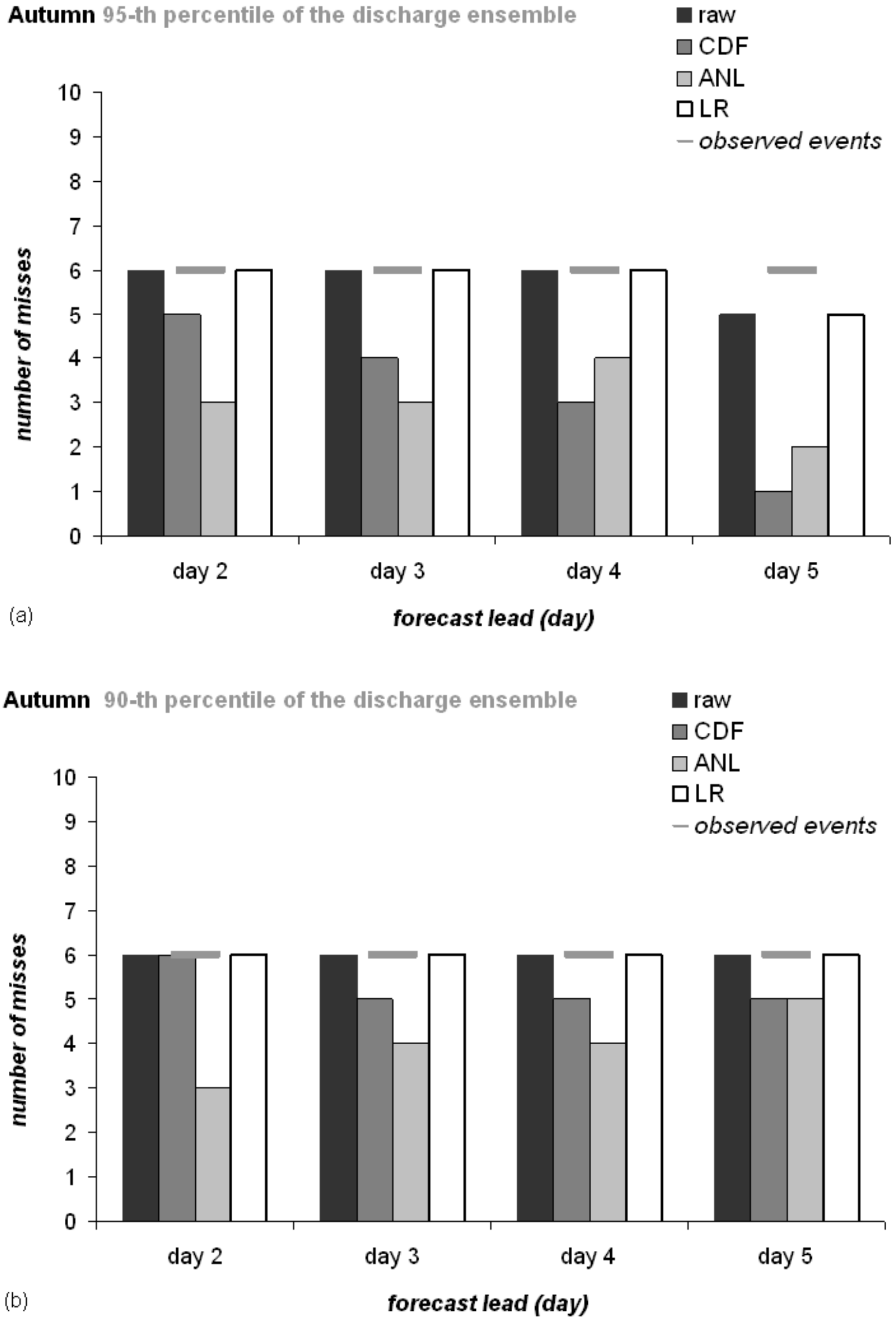


Figure 29: Number of missed events provided by the discharge forecasts driven by the raw and calibrated QPF ensembles for the autumn seasons in the years 2003-2008, as a function of the lead time. The statistics refer to the (panel a) 95-th and (panel b) 90-th percentiles of each discharge ensemble, with respect to the exceeding of the warning threshold at the Casalecchio Chiusa river section. The number of observed events (six) is displayed by the horizontal thick grey line.

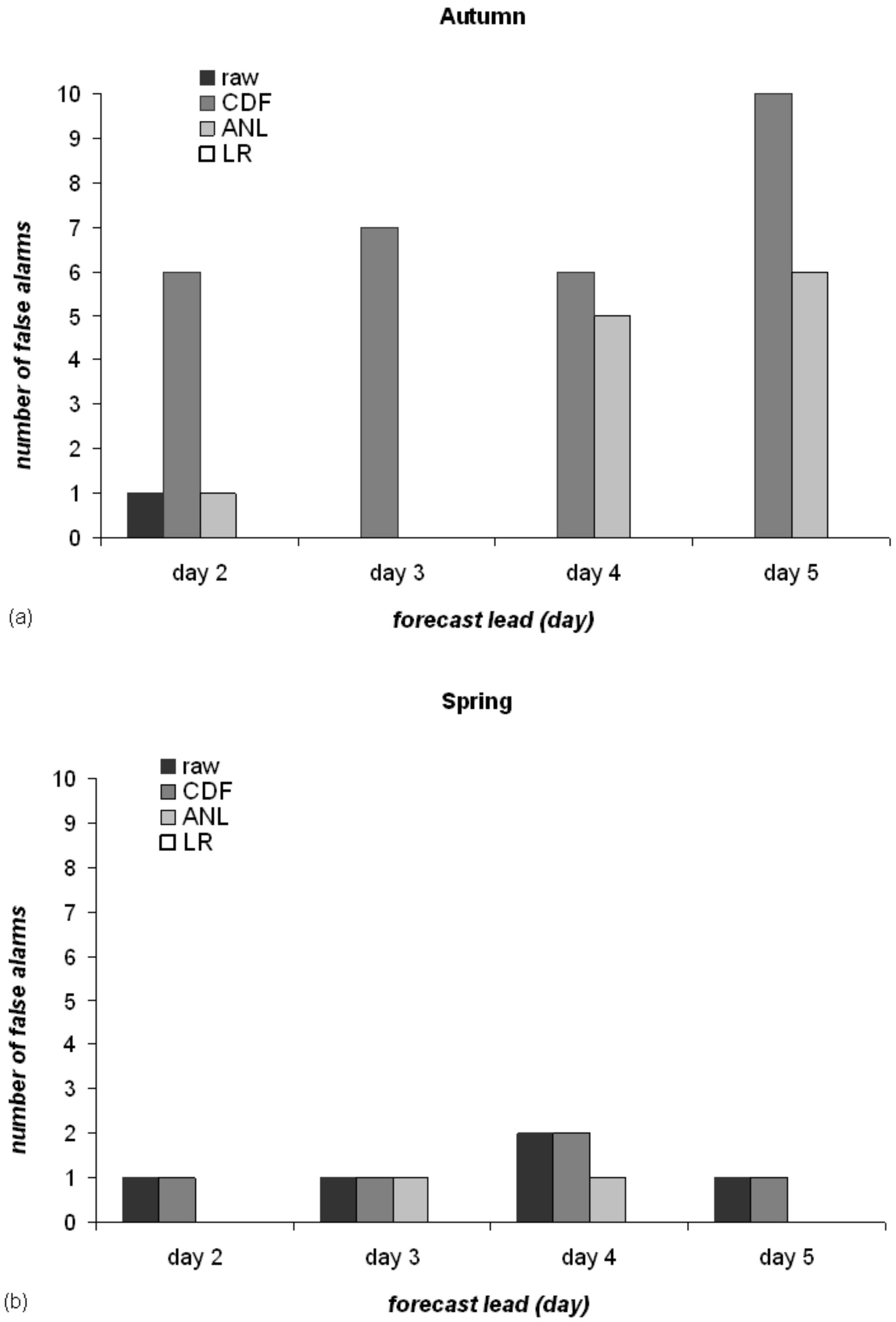


Figure 30: Number of false alarms provided by the discharge forecasts driven by the raw and calibrated QPF ensembles for the autumn (panel a) and spring (panel b) seasons in the years 2003-2008, as a function of the lead time. The statistics refer to the 95-th percentile of each discharge ensemble, with respect to the exceeding of the warning threshold at the Casalecchio Chiusa river section.

noteworthy outcome stands out by the calibration when the 95-th percentile of the discharge ensemble is considered. With respect to the raw ensemble, the CDF and ANL methods provided an increase of RMSE, especially in autumn for longer lead times; instead, the LR method provided a decrease of RMSE, more evident in spring (Figure 27). At the same time, for both autumn and spring, the RMSE values of the discharge ensemble mean driven by the raw and calibrated QPFs are quite similar (not shown). By the light of these outcomes, it seems that the CDF and ANL methods tend to increase the spread of the discharge ensemble. An additional investigation for the period 2003-2008 was then performed on the 24-h rainfall fields forecasted over the sub-area of Emilia-Romagna which includes the Reno river basin, in order to deeper investigate the impact of the calibration methodologies on the spread of the ensemble. This analysis confirms that the ANL and CDF methods tend to increase the spread of the ensemble (evaluated as the difference between the minimum and the maximum values of QPF provided by the ensemble members), especially in autumn (Figure 28). The increased spread associated to the ANL method can be due to the lower quality of analogs, namely very different situations which are not related to the current forecast are selected as analogs. Such a tendency to degrade the calibrated ensemble towards the climatology can also characterise the CDF method, which by construction does not correct for spread deficiencies (Hamill and Whitaker, 2006) and neglects the conditional relationship between observations and forecasts (Hopson and Webster, 2010). Instead, the reduced spread of the ensemble calibrated by the LR method could be explained by considering that in case of forecasts and observations are uncorrelated, this method is not able to remove the error systematically, providing a correction which could adjust all member forecasts to very close values (probably the climatological mean), regardless of their initial value (Hamill and Whitaker, 2006). Then, the impact of the calibration approach focussed on the verification of warning messages which would have been issued on the basis of the discharge scenarios driven by COSMO-LEPS. In particular, missed events and false alarms for the autumn and spring seasons in the years 2003-2008 have been investigated with respect to the exceeding of the warning threshold (i.e., the second level out of three levels of alert defined for the aims of civil protection) at Casalecchio Chiusa within each 24-h forecast period, up to day-5. Noteworthy outcomes were provided by the higher percentiles of the discharge ensemble. Considering the 95-th percentile, the impact of the calibration process in terms of reduction of misses for the autumn season is remarkable for the CDF and ANL methods (Figure 29, panel a). Actually, the discharge forecasts driven by the raw QPFs missed the six observed events totally for the lead times up to day-4; at day-5, five out of six events were missed. Rather, the discharge forecasts driven by the QPFs calibrated by the CDF and ANL methods reduced the missed events. In particular, the ANL method performs better up to day-3 (the number of misses is reduced by half), whereas the CDF method provides the best performance for longer lead times (one missed event only out of six cases at day-5). The LR method does not enable a reduction of misses with respect to the raw forecast. The positive impact provided by the CDF and ANL methods in terms of decrease of misses is reduced when the 90-th percentile of the discharge ensemble is considered (Figure 29, panel b). Nonetheless, the performance of the ANL method is quite similar up to day-4 for both the percentiles of the ensemble. No events occurred in the spring season in the period 2003-2008, thus no missed events result.

In terms of false alarms, when the 95-th percentile of the ensemble is considered, the discharge forecasts driven by QPFs calibrated by the CDF and ANL methods provide an increase of false alarms in the autumn season, with respect to the raw forecast (Figure 30, panel a). However, the worsening for the ANL method is evident only from day-4. The LR method does not provide any false alarm. When the 90-th percentile of the discharge ensemble is considered (not shown), only the CDF method still provides some (e.g., no more than four) false alarms. The number of false alarms provided by the 95-th percentile of the discharge ensemble is quite reduced (no more than two false alarms) in the spring season (Figure

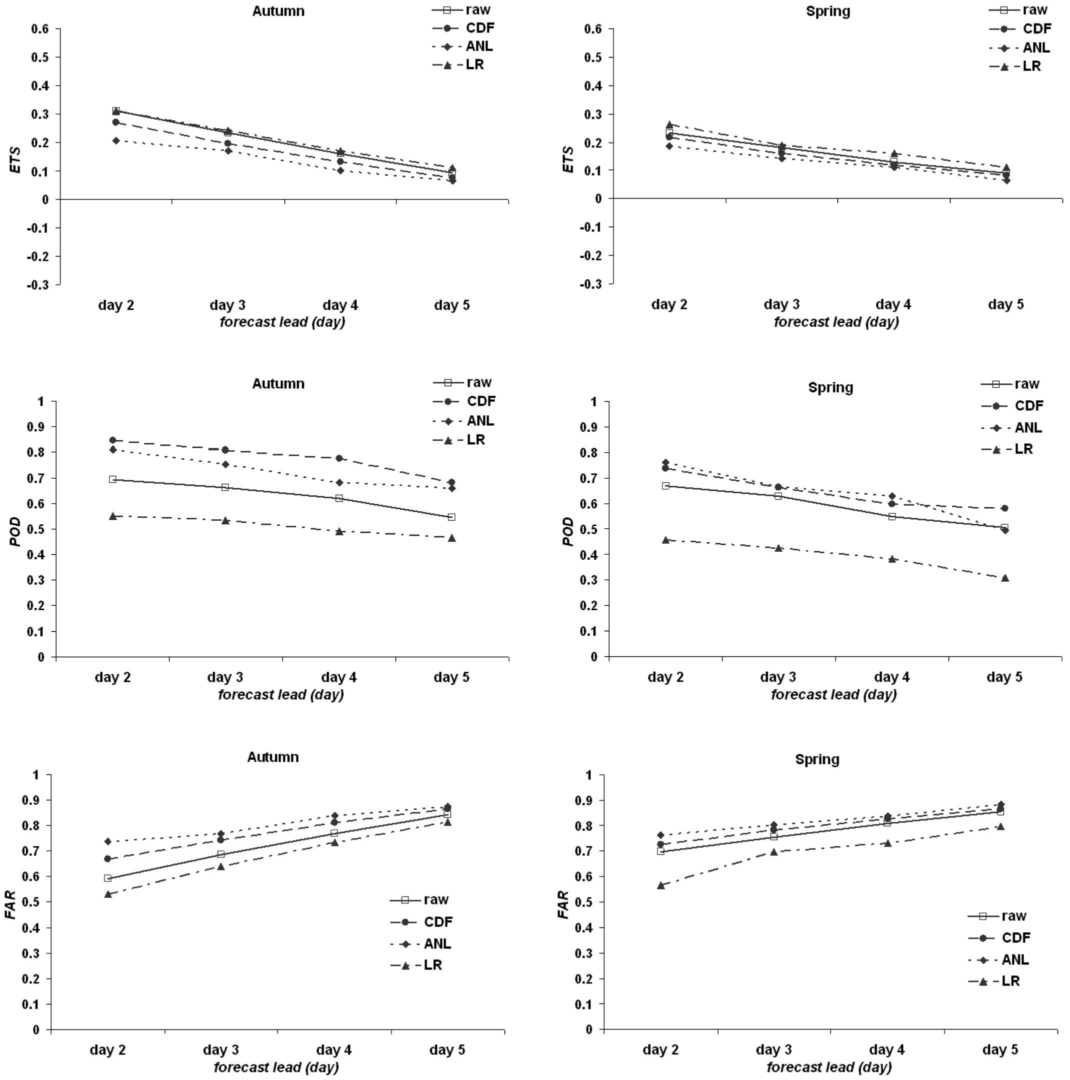


Figure 31: Equitable Threat Score, Probability Of Detection and False Alarm Rate at the 95-th percentile threshold of rainfall climatology for the 95-th percentile of QPFs of the raw and calibrated ensembles over the sub-area of Emilia-Romagna which includes the Reno river basin. The results refer to different forecast ranges, up to day 5 lead time, for the autumn (left panels) and spring (right panels) seasons in the years 2003-2008.

30, panel b). In particular, the discharge ensembles driven by QPFs calibrated with the CDF and ANL methods does not provide an increase of false alarms with respect to stream flow forecasts driven by the raw QPFs; even, the ANL method enables a reduction of false alarms. The LR method does not provide any false alarm. Similar results are obtained when the 90-th percentile of the discharge ensemble is considered (not shown). Summarising, the investigation over the Reno river basin shows that the calibration performed with the ANL and CDF methods enables to decrease the number of missed events, at the expenses of an increase of false alarms. Decision-makers have to evaluate (for instance, on the strength of a cost-loss analysis) which percentile of the discharge ensemble should be more suitable for supporting their activities. The noteworthy outcomes provided by the meteo-hydrological coupling suggested to perform an additional statistical analysis in terms of QPFs for the sub-area which includes the Reno river catchment, in order to investigate the ability of the calibrated ensembles to discriminate precipitation events. Actually, the calibration did not show a remarkable beneficial impact in terms of reliability and BSS on 24-h QPFs, especially in autumn . The attributes diagrams and the BSSs relative to the sub-area of the Reno river basin (not shown) provide results which are similar to those of the total area of Emilia-Romagna (shown in Figure 25 and Figure 26 respectively). The Equitable Threat Score (ETS), Probability Of Detection (POD) and False Alarm Rate (FAR) were thus computed for the autumn and spring seasons in the years 2003-2008. Figure 31 shows the forecast skills in terms of ETS, POD and FAR for the 95-th percentile of raw and calibrated ensembles, with respect to the 95-th percentile of rainfall climatology as verification threshold. The increase of POD and FAR for the ensembles calibrated by CDF and ANL with respect to the raw ensemble confirms the outcomes of the meteo-hydrological coupling. On the one hand, the improved QPF performance in terms of POD for the ensembles calibrated by CDF and ANL constitutes a real issue for the beneficial impact of calibration, as end-users are usually more concerned about missed events than by false alarms. On the other hand, these calibrated forecasts are characterised by an increase of FAR, even though not at a preoccupying rate. The ETS does not change significantly, due to the balance between the increases of both hits (influencing positively the ETS) and false alarms (influencing negatively the ETS).

5.3 Comments and conclusions

The direct model output of COSMO-LEPS is biased and shows deficiencies concerning both reliability and sharpness. Consequently, there is a need for calibration and post-processing techniques that address these issues and provide to the forecasters more reliable products for operational use. Here, the calibration approach utilized a 30-yr reforecast dataset produced by MeteoSwiss. In the recent past, several studies showed that reforecasts turn out to be useful for additional forecast improvement and diagnostic capability, making worth the extra computational resources required to produce them. In this work, the impact of the calibration approach was not so beneficial with respect to it had been reported in some comparable studies, in particular in terms of increase of reliability and skill. This can be due to the fact that in this study high resolution precipitation forecasts is dealt with. The paper investigated three methods for postprocessing the ensemble precipitation forecasts: quantile-to-quantile mapping, linear regression and analogs. The accuracy of calibrated QPFs was verified, with a particular focus on the use of such fields in hydrological applications. There appeared to be no single best forecast method for all applications and study areas, among the three tested. This study demonstrated that it is possible to improve raw ensemble forecasts of precipitation, but the improvements vary strongly from to site to site. It is reasonable to assume that this was more related to the non systematic model error than to the lack of capability of these methods to extract predictive information. The statistical analyses over Switzerland and

Germany revealed a positive impact of the calibration process. The calibrated ensembles increased the forecast reliability in all the seasons. The best performance over Switzerland was generally obtained with the ANL method, except for summer. Rather, the LR and CDF methods provided the best performances over Germany. No significant improvements resulted over Emilia-Romagna by the statistical analysis on the calibrated QPFs. But, the quality of these forecasts proved to be as improved for hydrological applications. Actually, the output of the hydrological model TOPKAPI driven by the calibrated QPFs revealed a beneficial impact of calibration for the discharge predictions over a medium-size catchment (i.e., the Reno river basin) selected as case study. In particular, the investigation on the hydrological runs covering the autumn and spring seasons of the years 2003-2008 showed that the calibration performed with the ANL and CDF methods enabled to obtain a decrease of missed events. But, an increase of false alarms resulted by the application of these two calibration methods, even though this trend was evident for the ANL method for longer lead times only. Decision-makers, who are usually more concerned by missed events than false alarms, have to evaluate which percentile of the discharge ensemble should be more suitable for supporting their activities. It seems clear that objective post-processing of ensemble forecasts will remain a critical component of the forecast process. The difficulty of accurately forecasting the intensity, location and timing of intense precipitation events at the spatial scale typical of small/medium-size catchments limits the ability of COSMO-LEPS to confidently and reliably capture observed intense peak discharges. Therefore, although calibrated forecasts do convey added-value in comparison to raw ones, precipitation forecasts need to be further improved to guarantee sufficiently reliable flood predictions over small/ medium-size catchments. The promising results of this study indicate that further research is merited. Whatever the case, there is definitely room for improvement in the calibration of ensemble systems. Reduction of errors that vary by weather element and flow configuration from ensemble precipitation forecasts is considered a necessary step to improving forecast quality and benefiting end users (Gneiting and Raftery, 2005; Yuan et al., 2007). There is a multitude of techniques that could be applied to regime-based calibration approach. Considering the techniques which were applied in the present study, a future improvement for the analog search could be obtained by a multi-variable approach, in order to find a better matching of the space-time evolution of the synoptic pattern. With respect to the CDF and LR methods, a solution could be the division of the training sample size according to some criteria which would allow to pool data which have similar model errors with respect to a certain meteorological situation.

6 Conclusions

The CONSENS PP has ended in 2011 and the work carried out within the Project has provided useful indications about the update and/or modification of the ensemble suites, together with suggestions about future works. The main outcomes are:

- the work on parameter perturbation has lead to the choice of a definite set of parameters and values which have been implemented in both the COSMO-SREPS and COSMO-LEPS ensembles. In the latter, though, perturbations can have different combinations with respect to those here analysed, since the combination is randomly selected day by days

- the soil moisture analysis perturbation technique has been developed and implemented, and it is available for further testing. This technique has not been implemented yet in an ensemble configuration for testing, but its applicability will be further investigated in the future

- the work on multi-clustering has confirmed that the ensemble reduction methodology is able to retain a good fraction of the information contained in the original ensemble and that there is a relationship between cluster size and skill of the corresponding representative member of the large scale ensemble (for upper air variables). Therefore, the clustering methodology will be further developed and updated, aiming at improving also the short-range performances

- on the basis of the results of the COSMO-LEPS/COSMO-SREPS comparison, it was decided to stop running the extra COSMO-SREPS members (nested on the same sets of IC and BC but with different physics), and to keep only the COSMO-SREPS members which receive IC and BC by different global models, but prolonging the runs up to +132h, as COSMO-LEPS. These four extra runs (one nested on IFS with Tiedkte convection, one nested on IFS with Kain-Fritsch convection, one nested on GME and one nested on GFS) are now run as part of an experimental suite, which in combination with COSMO-LEPS constitutes the COSMO-HYBEPS 20 member ensemble, currently under evaluation

- an additional product of COSMO-LEPS has been made available, the 24h QPF calibrated with the CDF methodology over the 3 domains which have been considered in the study here presented

These results have lead to an effective consolidation of the mesoscale ensemble system of the Consortium and have provided useful indications about possible further improvements. Furthermore, the experience gained in the field of model perturbation techniques is useful for the development of the convection-permitting ensembles which is on-going in several COSMO Countries and which will be addressed in a forthcoming PP.

7 Acknowledgments

This project was carried out in the framework of the COSMO Consortium, therefore the authors would like to thank the COSMO colleagues for fruitful discussions and useful suggestions. We are also grateful for their help to Davide Cesari, Paolo Patruno and Fabrizio Nerozzi (ARPA-SIMC). The COSMO-LEPS and COSMO-SREPS ensembles have been run at ECMWF thanks to the Billing Units provided by the COSMO Countries which are also ECMWF members (Germany, Italy, Switzerland, Greece). The Italian institutions which provided observed data for the verification are also acknowledged.

8 References

Diomede, T., Nerozzi, F., Paccagnella, T. and E., Todini, 2008. The use of meteorological analogues to account for LAM QPF uncertainty. *Hydrol. Earth Syst. Sci.* **12**, 141-157.

Fundel, F., Walser, A., Liniger, M.A., Frei, C. and C., Appenzeller, 2010. Calibrated Precipitation Forecasts for a Limited Area Ensemble Forecast System Using Reforecasts. *Mon. Wea. Rev.* **138**, 176-189.

Gneiting, T. and A.E., Raftery, 2005. Weather forecasting with ensemble methods. *Science* **310**, 248-249.

Hamill, T.M. and J.S., Whitaker, 2006. Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev.* **134**, 3209-3229.

Hopson, T.M. and P.J., Webster, 2010. A 1-10-Day Ensemble Forecasting Scheme for the Major River Basins of Bangladesh: Forecasting Severe Floods of 2003-07. *J. Hydrometeor.* **11**, 618-641.

Houtekamer, P.L., 1993. Global and local skill forecasts. *Mon. Wea. Rev.* **121**, 1834-1846.

Kutzbach, J.E., 1967. Empirical eigenvectors of sea-level pressure, surface temperature and precipitation complexes over North America. *J. App. Meteorol.* **6**, 791-802.

Liu, Z. and E., Todini, 2002. Towards a comprehensive physically-based rainfall-runoff model. *Hydrol. Earth Syst. Sci.* **6**, 859-881.

Marsigli C., Montani A., Nerozzi F., Paccagnella T., Tibaldi S., Molteni F. and R., Buizza, 2001. A strategy for high-resolution ensemble prediction. Part II: Limited-area experiments in four Alpine flood events. *Q. J. R. Meteorol. Soc.* **127**, 2095-2115.

Marsigli, C., Montani A. and T., Paccagnella, 2008. A spatial verification method applied to the evaluation of high-resolution ensemble forecasts. *Meteorol. Appl.* **15**, 125-143.

Marsigli, C., 2009. Final report on priority project SREPS (Short Range Ensemble Prediction System). COSMO Technical Report N. 13, available at
http://www.cosmo-model.org/content/model/documentation/techReports/default.htm.

Marsigli C., Montani A. and T., Paccagnella, 2013. Perturbation of initial and boundary conditions for a limited-area ensemble: multi-model versus single-model approach. *Q. J. R. Meteorol. Soc.*, in press.

Molteni F., Buizza R., Marsigli C., Montani A., Nerozzi F. and T., Paccagnella, 2001. A strategy for high-resolution ensemble prediction. Part I: Definition of representative members and global model experiments. *Q. J. R. Meteorol. Soc.* **127**, 2069-2094.

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

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

Press, W.H., Teulosky S.A., Vetterling, W.T. and B.P., Flannery, 1992. Numerical Recipes in Fortran 77. 2nd ed. Cambridge Univ. Press, 280 pp.

Sutton, C. J. and T. M., Hamill, 2004. Impacts of perturbed soil moisture conditions on short-range ensemble variability. 16th NWP/20th W&F Conference, Seattle, American Meteorological Society.

von Storch, H. and G., Hannoschock, 1984. Comments on "Empirical Orthogonal Function Analysis of Wind Vectors over the Tropical Pacific Ocean". *Bulletin of the Meteorological Society of America* **65**, 162. (Appeared as a letter to the editor concerning: Legier, D.M.,

1983. Empirical Orthogonal Function Analysis of Wind Vectors over the Tropical Pacific Region. *Bulletin of the Meteorological Society of America* **64**, 234-241.)

Yuan, H., Gao, X., Mullen, S.L., Sorooshian, S., Du, J. and H-M. H., Juang, 2007. Calibration of Probabilistic Quantitative Precipitation Forecasts with an Artificial Neural Network. *Wea. Forecasting* **22**, 1287-1303.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. 3rd edn. Academic Press, New York, 676 pp.

List of COSMO Newsletters and Technical Reports

(available for download from the COSMO Website: www.cosmo-model.org)

COSMO Newsletters

No. 1: February 2001.

No. 2: February 2002.

No. 3: February 2003.

No. 4: February 2004.

No. 5: April 2005.

No. 6: July 2006.

No. 7: April 2008; Proceedings from the 8th COSMO General Meeting in Bucharest, 2006.

No. 8: September 2008; Proceedings from the 9th COSMO General Meeting in Athens, 2007.

No. 9: December 2008.

No. 10: March 2010.

No. 11: April 2011.

No. 12: April 2012.

No. 13: April 2013.

COSMO Technical Reports

No. 1: Dmitrii Mironov and Matthias Raschendorfer (2001):
Evaluation of Empirical Parameters of the New LM Surface-Layer Parameterization Scheme. Results from Numerical Experiments Including the Soil Moisture Analysis.

No. 2: Reinhold Schrodin and Erdmann Heise (2001):
The Multi-Layer Version of the DWD Soil Model TERRA_LM.

No. 3: Günther Doms (2001):
A Scheme for Monotonic Numerical Diffusion in the LM.

No. 4: Hans-Joachim Herzog, Ursula Schubert, Gerd Vogel, Adelheid Fiedler and Roswitha Kirchner (2002):
LLM ¯ the High-Resolving Nonhydrostatic Simulation Model in the DWD-Project LITFASS.
Part I: Modelling Technique and Simulation Method.

No. 5: Jean-Marie Bettems (2002):
EUCOS Impact Study Using the Limited-Area Non-Hydrostatic NWP Model in Operational Use at MeteoSwiss.

No. 6: Heinz-Werner Bitzer and Jürgen Steppeler (2004):
Documentation of the Z-Coordinate Dynamical Core of LM.

No. 7: Hans-Joachim Herzog, Almut Gassmann (2005):
Lorenz- and Charney-Phillips vertical grid experimentation using a compressible non-hydrostatic toy-model relevant to the fast-mode part of the 'Lokal-Modell'.

No. 8: Chiara Marsigli, Andrea Montani, Tiziana Paccagnella, Davide Sacchetti, André Walser, Marco Arpagaus, Thomas Schumann (2005):
Evaluation of the Performance of the COSMO-LEPS System.

No. 9: Erdmann Heise, Bodo Ritter, Reinhold Schrodin (2006):
Operational Implementation of the Multilayer Soil Model.

No. 10: M.D. Tsyrulnikov (2007):
Is the particle filtering approach appropriate for meso-scale data assimilation ?

No. 11: Dmitrii V. Mironov (2008):
Parameterization of Lakes in Numerical Weather Prediction. Description of a Lake Model.

No. 12: Adriano Raspanti (2009):
COSMO Priority Project "VERification System Unified Survey" (VERSUS): Final Report.

No. 13: Chiara Marsigli (2009):
COSMO Priority Project "Short Range Ensemble Prediction System" (SREPS): Final Report.

No. 14: Michael Baldauf (2009):
COSMO Priority Project "Further Developments of the Runge-Kutta Time Integration Scheme" (RK): Final Report.

No. 15: Silke Dierer (2009):
COSMO Priority Project "Tackle deficiencies in quantitative precipitation forecast" (QPF): Final Report.

No. 16: Pierre Eckert (2009):
COSMO Priority Project "INTERP": Final Report.

No. 17: D. Leuenberger, M. Stoll and A. Roches (2010):
Description of some convective indices implemented in the COSMO model.

No. 18: Daniel Leuenberger (2010):
Statistical analysis of high-resolution COSMO Ensemble forecasts in view of Data Assimilation.

No. 19: A. Montani, D. Cesari, C. Marsigli, T. Paccagnella (2010):
Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO–LEPS system: main achievements and open challenges.

No. 20: A. Roches, O. Fuhrer (2012):
Tracer module in the COSMO model.

No. 21: Michael Baldauf (2013):
A new fast-waves solver for the Runge-Kutta dynamical core.

COSMO Technical Reports

Issues of the COSMO Technical Reports series are published by the *COnsortium for Small-scale MOdelling* at non-regular intervals. COSMO is a European group for numerical weather prediction with participating meteorological services from Germany (DWD, AWGeophys), Greece (HNMS), Italy (USAM, ARPA-SIMC, ARPA Piemonte), Switzerland (MeteoSwiss), Poland (IMGW), Romania (NMA) and Russia (RHM). The general goal is to develop, improve and maintain a non-hydrostatic limited area modelling system to be used for both operational and research applications by the members of COSMO. This system is initially based on the COSMO-Model (previously known as LM) of DWD with its corresponding data assimilation system.

The Technical Reports are intended

- for scientific contributions and a documentation of research activities,
- to present and discuss results obtained from the model system,
- to present and discuss verification results and interpretation methods,
- for a documentation of technical changes to the model system,
- to give an overview of new components of the model system.

The purpose of these reports is to communicate results, changes and progress related to the LM model system relatively fast within the COSMO consortium, and also to inform other NWP groups on our current research activities. In this way the discussion on a specific topic can be stimulated at an early stage. In order to publish a report very soon after the completion of the manuscript, we have decided to omit a thorough reviewing procedure and only a rough check is done by the editors and a third reviewer. We apologize for typographical and other errors or inconsistencies which may still be present.

At present, the Technical Reports are available for download from the COSMO web site (www.cosmo-model.org). If required, the member meteorological centres can produce hard-copies by their own for distribution within their service. All members of the consortium will be informed about new issues by email.

For any comments and questions, please contact the editor:

Massimo Milelli
Massimo.Milelli@arpa.piemonte.it