Nat. Hazards Earth Syst. Sci., 12, 591–606, 2012
www.nat-hazards-earth-syst-sci.net/12/591/2012/
doi:10.5194/nhess-12-591-2012

Natural Hazards and Earth System Sciences

Intercomparison of two meteorological limited area models for quantitative precipitation forecast verification

E. Oberto[1], **M. Milelli**[1], **F. Pasi**[2], **and B. Gozzini**[3]

[1]ARPA Piemonte, Torino, Italy
[2]Consorzio LAMMA, Sesto Fiorentino (FI), Italy
[3]IBIMET-CNR, Firenze, Italy

Correspondence to: E. Oberto (elena.oberto@arpa.piemonte.it)

Received: 15 April 2011 – Revised: 5 October 2011 – Accepted: 31 October 2011 – Published: 8 March 2012

Abstract. The demand for verification of numerical models is still very high, especially for what concerns the operational Quantitative Precipitation Forecast (QPF) used, among others, for evaluating the issuing of warnings to the population. In this study, a comparative verification of the QPF, predicted by two operational Limited Area Models (LAMs) for the Italian territory is presented: COSMO-I7 (developed in the framework of the COSMO Consortium) and WRF-NMM (developed at NOAA-NCEP). The observational dataset is the precipitation recorded by the high-resolution non-GTS rain gauges network of the National Civil Protection Department (NCPD) over two years (2007–2008). Observed and forecasted precipitation have been treated as areal quantity (areal average of the values accumulated in 6 and 24 h periods) over the 102 "warning areas", defined by the NCPD both for administrative and hydrological purposes. Statistics are presented through a series of conventional indices (BIAS, POD and POFD) and, in addition, the Extreme Dependency Score (EDS) and the Base Rate (BS or 1-BS) have been used for keeping into account the vanishing of the indices as the events become rare. Results for long-period verification (the whole 2 yr) with increasing thresholds, seasonal trend (3 months period), diurnal error cycle and error maps, are presented. Results indicate that WRF has a general tendency of QPF overestimation for low thresholds and underestimation for higher ones, while COSMO-I7 tends to overestimate for all thresholds. Both models show a seasonal trend, with a bigger overestimation during summer and spring, while during autumn and winter the models tend to be more accurate.

1 Introduction

Many works have been devoted to the verification of numerical models, and special efforts have been dedicated to the operational QPF verification: in the present work the QPF of two operational LAMs, the COSMO-I7 model (COnsortium for Small-scale MOdelling) and the WRF model (Weather Research and Forecasting) is investigated. Although, to the authors' knowledge, this is the first time that COSMO and WRF QPF performances are compared, other studies have been carried out on the models: for what concerns COSMO, for instance, we could mention Dierer et al. (2009) or Wernli et al. (2009), while for WRF (or his progenitor MM5), we could quote Colle et al. (2000), Mass et al. (2002) or Clark et al. (2007). Moreover the QPF of other mesoscale models such as BOLAM (Bologna Limited-Area Model) or ETA model (NCEP) has been studied as an example in Lagouvardos et al. (2003) and in Papadopoulos and Katsafados (2009), respectively. All these studies confirm that the advantage of LAMs to be able to resolve meso-alpha and meso-beta features might be penalized by the use of the wrong verification method and by the wrong choice of indices, which becomes crucial.

Recently, after the Third International Workshop on Verification Methods hosted by ECMWF in February 2007, a special issue of Meteorological Applications (2008) has been published with an extensive collection of papers on verification. In their comprehensive review, Casati et al. (2008) define the verification process as an indispensable part of the meteorological research and operational activities, warning on the fact that the methodology should be properly designed to meet the needs of different groups, including modellers, forecaster and end-users. In our case the work is addressed

to the Italian NDCP mainly, the most important (and most demanding) end-user, since we give great emphasis to the verification of heavy rain events over predefined areas (warning areas) that might cause floods and/or landslides in very densely populated zones. These events are, fortunately, rare in a statistical sense; consequently, particular care has to be addressed to the use of the most appropriated statistical indices. In fact, standard statistical indices are problematic as they generally vanish as events become rare (Stephenson et al., 2008). As pointed out by Ghelli and Primo (2009), the definition of forecast accuracy for rare events has been a subject for 2 yr and authors suggest the use of Extreme Dependency Score (EDS) in conjunction with at least another skill score (e.g., False Alarm Rate, or F) in order to avoid hedging.

As a consequence of these considerations, a set of conventional indices (BIAS, POD, POFD) has been used together with EDS and BR (1-BR). Models' skill scores have been calculated for long-period verification (the whole 2 yr), seasonal trend (3 months period) and diurnal error cycle. As suggested by Jolliffe and Stephenson (2003) and applied by Accadia et al. (2004) for similar purposes, a testing hypothesis has been adopted in which a confidence interval has been built for the performance differences in order to evaluate if the results of the two models are statistically different (Hamill, 1999). In addition, error maps for the entire domain are presented with the aim of providing a general outlook of QPF errors with respect to the warning areas properties, season and rainfall climatology.

The observational dataset is the Italian high-resolution non-GTS network, which has more than 1300 rain gauges distributed over a large part of the national territory. The verification period is 2 yr (2007 and 2008) and precipitation (on 6 h and 24 h accumulation time) is treated as an areal quantity, being averaged over "warning areas", which have a national value as the administrative sectors for the issuing of weather warnings by the NCPD. This procedure is similar to a simple fuzzy verification method, as explained extensively by Ebert (2008). In fact, the basic assumption of the fuzzy methods is that it is still acceptable for the forecast to be slightly displaced in space and/or in time.

For what concerns numerical models, we are reminded that they predict meteorological conditions on spatial scales different from the observed ones, therefore, verification against irregularly distributed data (e.g., high resolution observation network) might be liable to misinterpretation. Usually the observations are in some way upscaled to pre-defined areas together with the model values, or the model values are interpolated to the station point, but any process of interpolation has to be made with caution as it normally assumes that the starting field is continuous. For precipitation, especially for the convective part of it, this is not the case. However, recent sensitivity studies (e.g., Cherubini et al., 2002) have shown little influence of the averaging technique on verification results, especially when treating precipitation as an areal quantity (Pappenberger et al., 2009).

The paper is structured in the following way: Sect. 2 explains the main features of the numerical models, the domain of integration, the chosen configuration and the data assimilation; Sect. 3 describes the dataset characteristics and the averaging methods used to build the pairs (observations and model forecasts) for the objective verification; Sect. 4 is dedicated to the methodology of verification; in Sect. 5 the most representative results are illustrated and eventually in Sect. 6 the main conclusions are drawn.

2 Models description

2.1 COSMO Model

The model is based on non-hydrostatic, fully compressible hydro-thermodynamical equations in advection form (Steppeler et al., 2003). Generalized terrain-following coordinates with rotated geographical coordinates are used. The model equations are solved on an ARAKAWA C-grid with user-defined vertical grid staggering. They are spatially discretised with second-order finite differences. Time integration uses a 2nd order leapfrog (horizontally explicit, vertically implicit) time-split integration scheme including extensions proposed by Skamarock and Klemp (1992). A 4th order linear horizontal diffusion is calculated. 2-dimension divergence damping and vertical off-centring are applied in split-time steps. Rayleigh damping is calculated in the upper layers (Doms and Schaettler, 2002). Data at the lateral boundaries are prescribed using a Davies-type one-way nesting (Davies, 1976, 1983). Subgrid-scale turbulence is parameterised by a prognostic turbulent kinetic energy closure at level 2.5 including effects from subgrid-scale condensation and from thermal circulations (Mellor and Yamada, 1974). The surface layer parameterisation is based on turbulent kinetic energy and includes a laminar-turbulent roughness layer. The formation of precipitation is described by a bulk microphysics parameterisation including water vapour, cloud water and ice, rain and snow with a fully prognostic treatment of precipitation, i.e., three-dimensional transport of rain, and snow is calculated. Condensation and evaporation are parameterised by saturation adjustment while depositional growth/sublimation of cloud ice is calculated using an explicit non-equilibrium growth equation. Subgrid-scale cloudiness used for radiation calculations is parameterised by an empirical function depending on relative humidity, ice content and height. Moist convection is parameterised using a mass-flux scheme with an equilibrium closure based on moisture convergence following Tiedtke (1989). Radiation is calculated using a two-stream scheme for short- and long-wave fluxes (eight spectral intervals) including a full cloud-radiation feedback (Ritter and Geleyn, 1992). A two-layer soil model (TERRA, see Doms and Schaettler, 2002) is applied. As far as the data assimilation is concerned, a nudging or Newtonian relaxation scheme is implemented. It

consists of relaxing the model's prognostic variables towards prescribed values within a given time window. In the present scheme, nudging is performed using observations, which is more appropriate for high-resolution applications than nudging towards 3-dimensional analyses (Stauffer and Seaman, 1994). A relaxation term is introduced into the model equations and the nudging term usually remains smaller than the largest term of the dynamics, so that the dynamic balance of the model is not disturbed strongly. Moreover, an explicit balancing by a hydrostatic temperature correction for surface pressure updates, a geostrophic wind correction and a hydrostatic upper-air pressure correction are applied. COSMO performs a nudging data assimilation cycle before the main run using upper level parameters (wind, temperature, humidity) and surface parameters (pressure, wind, humidity).

2.2 WRF Model and configuration

The WRF system supports two dynamical cores: the Advanced Research (ARW) and the Non-hydrostatic Mesoscale Model (NMM), object of this work and used operationally at LaMMA Consortium for the regional weather service. The model dynamics is fully described in Janjic (2003) and the model physics, including the different options available is, for example, described in Chen and Dudhia (2000).

The operational configuration used at LaMMA for WRF-NMM version 3.0 is the NCEP suggested standard for mid-latitudes. More in detail:

- Long and short wave radiation: Geophysical Fluid Dynamics Laboratory (GFDL, Lacis and Hansen, 1974).
- Surface layer: Janjic Similarity (Janjic, 1996, 2002).
- Planetary Boundary Layer: Mellor-Yamada-Janjic Turbulent Kinetic Energy scheme (TKE-MYJ, Janjic, 1996, 2002).
- Land-surface: NOAH (Chen and Dudhia, 2001).

The microphysics scheme used is the Eta Grid-scale Cloud and Precipitation also known as Eta Ferrier scheme (Ferrier et al., 2002). This is a 4-class scheme that predicts changes in water vapour and condensate in the form of cloud water, rain, cloud ice and precipitation ice (snow/graupel/sleet). The individual hydrometeor fields are combined into total condensate, and this quantity plus the water vapour is advected into the model. A more detailed description of this scheme can be found in Ryan (1996).

The cumulus parameterisation scheme used is a modified version of the Kain and Fritsch (1993) scheme. In the original scheme convection is triggered by lifting a lower-level slab layer with an impetus heating as a function of grid-scale vertical motion at the lifting condensation level. Convection adjustment is based on convective available potential energy (CAPE) removal in a grid column within a convective timescale. The used scheme has been updated by imposing a minimum entrainment rate to suppress widespread convection in marginally unstable, relatively dry environments. Shallow (non-precipitating) convection is allowed for any updraft that does not reach minimum cloud depth for precipitating clouds and this minimum depth varies as a function of cloud-base temperature. For more details see also Kain (2004).

2.3 Model initialisation and domain

2.3.1 COSMO

The operational version of COSMO runs over the area shown in Fig. 1, using 297×313 grid points horizontally (0.0625° resolution) and 40 vertical levels. Initial and 3h-boundary conditions come from ECMWF-IFS (GCM, T_L799 about 0.25° resolution). It runs twice a day (00:00 and 12:00 UTC) for 72 h, with an assimilation cycle of 18 h. Additionally, assimilation of the same kind of observation described in Sect. 2.1 is performed in the first 4 h of the runs. It has to be underlined that while the Deutscher Wetterdienst (DWD) uses a variational Soil Moisture Analysis for the COSMO-EU model, the Italian version of the model (COSMO-I7) has no external analysis (neither Soil Moisture nor Sea Surface Temperature, therefore, the model gets this information only from the ECMWF-IFS initial conditions). A brief description of the currently used datasets to determine the external parameters is given here:

- GLOBE dataset from the National Geophysical Data Center contains the orographical height of the land surface in a resolution of 30 arcseconds.
- GLC2000 is provided by the Joint Research Center of the European commission. Except for Antarctica, data for plant characteristics are given with a resolution of 1 km. It is used to determine parameters such as Leaf Area Index (LAI), plant cover, root depth, land cover, surface roughness.
- Digital Soil Map of the World (FAO) for soil type. The resolution of the dataset is 5 arcminutes.

2.3.2 WRF

WRF Preprocessing System (WPS, version 3.0) is used to initialise WRF-NMM. WPS acts as an interpolator to provide the needed analysis and boundary conditions that must be specified at one lateral grid-point, plus 4 more points of relaxation.

Initial and boundary conditions data come from the operational analysis and 6 h forecasts of ECMWF-IFS, respectively, like COSMO model. A cold start is used operationally, that is no observation data is assimilated into the model. Sea surface temperature (SST) is taken from the global daily

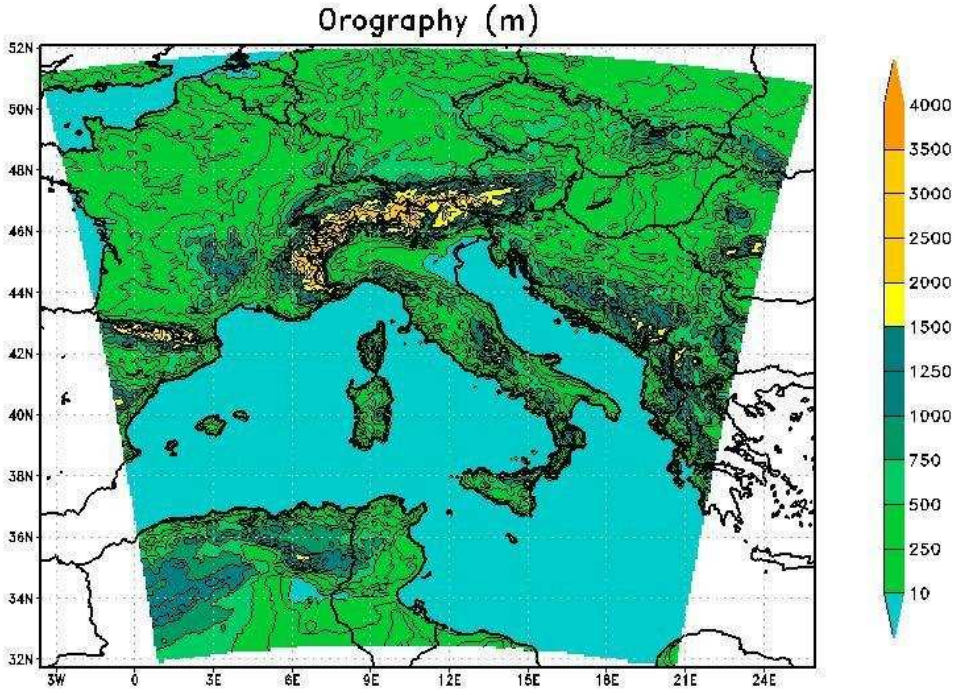

Fig. 1. COSMO-I7 domain and orography description.

NCEP real-time SST product at 0.083° resolution and is kept constant throughout the whole simulation. Soil moisture is taken from the GCM data. The domain of integration consists of 128×232 points, covering the whole Italian peninsula and partly central Europe as shown in Fig. 2, at a horizontal resolution of 0.07° (about 7.5 km). The eastern domain boundary is very close to the target area (south-eastern tip of the Italian peninsula, Puglia region) and this might introduce spurious effects from the boundary. The potentially undesired effects are mitigated by the exclusion of the Puglia region from the verifying dataset for missing data (see next paragraph). On the vertical, 35 vertical levels unequally spaced from ground to 100 hPa and with the first 10 levels being concentrated in the boundary layer (around 1.0 km above ground level), are present. Land-use and soil category come from the standard United States Geological Survey (USGS) categories (24 for land use and 16 for soil). Topography is derived from the global 30-s USGS topography data with a 4-point average.

3 Dataset description

The precipitation measurements for this study have been obtained from the high-resolution non-GTS network that counts for more than 1300 rain gauges distributed all over the Italian peninsula (Fig. 3). These data come from different sources:

- Data from Piemonte region network.
- Data gathered among some of the northern Italy regions available inside the regional data exchange in the COSMO framework.
- Data from NCPD network.

Data quality is ensured by a two-step procedure: first a simple automatic check is applied, and then a manual check is performed.

As a second step, in the frame of the NCPD warning system, the Italian territory has been subdivided into 102 "warning areas" mainly based on hydrological criteria, but keeping into account also meteorological, orographical and administrative criteria. As a further consequence of the Italian complex orography, these areas have been classified following also elevation criteria, ranging from plain-hill to mountain (the limit being fixed at around 600 m a.s.l.). At the end of this process, these areas, ranging between 1200 and 7400 km^2, are homogeneous from a meteorological and hydrological point-of-view. On average, with the actual LAM resolution, 60–70 models' grid points belong to each warning area.

From Fig. 3, one can infer that data coverage is satisfactory in the northern part of Italy (except from the two central plain areas), over the central and southern Apennine chain and on the central Adriatic coast (Abruzzo and Molise region). Data

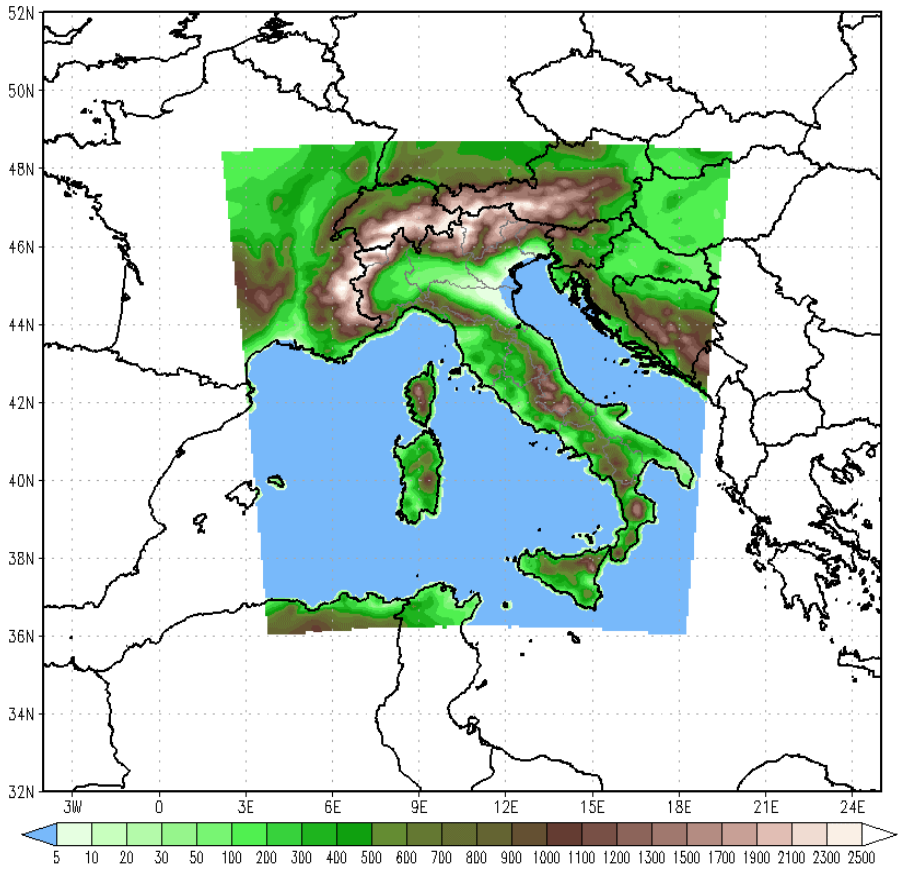

Fig. 2. WRF domain and orography description (m).

coverage is less dense over the central Thyrrenian coast, central Adriatic coast (Marche region) and Sardinia Island, and is missing over the southern Adriatic coast (Puglia region) and Sicilia Island. Warning areas with less then 3 stations are not taken into account in order to obtain consistent and significant estimates of mean areal rainfall. This leads to a reduction of the sample to 90 warning areas from the original 102. It is worth mentioning that the limit of 3 stations per warning area is reached in only 3 areas out of 90.

The verification dataset is based on the period January 2007 – December 2008 with 6 and 24 h accumulated rainfall (e.g., 00:00–24:00 UTC for 24 h period and 00:00–06:00, 06:00–12:00, 12:00–18:00, 18:00–24:00 UTC for 6 h periods). All model data come from runs initialised at 00:00 UTC.

4 Methodology description

Ghelli and Lalaurette (2000) propose to build up a gridded analysis (upscaled gridded observation), allocating each station to a grid box, averaging all the values within each grid box and assigning the averaged value to the appropriate grid point, because precipitation is as an areal quantity. For a mesoscale model, Cherubini et al. (2002) set the horizontal resolution limit to 9 km to be compared to a single observation avoiding representativity errors. Dealing with hydrological catchments, Pappenberger et al. (2009) has widely described the uncertainties related to the upscaling procedure to get areal precipitation averages. Several interpolation techniques (e.g., quasi-kriging, linear, cubic, nearest neighbour) are compared to finding out the best resolving upscaling. Pappenberger's study is performed over hydrological catchments (from small to large) very similar to the investigated 90 warning areas. The results show that, in the 43 catchments studied, there are small differences between the various interpolation methods with a slight preference for the nearest neighbour.

In this study the authors are dealing with warning areas that are very similar (if not the same) to small catchments. The horizontal resolution of the models is just below the limit proposed by Cherubini et al. (2002), but models treat convection with parameterisations. As a consequence the chosen approach is based on considering precipitation as an average areal quantity and upscaling over the warning areas with the technique proposed by Ghelli and Lalaurette (2000).

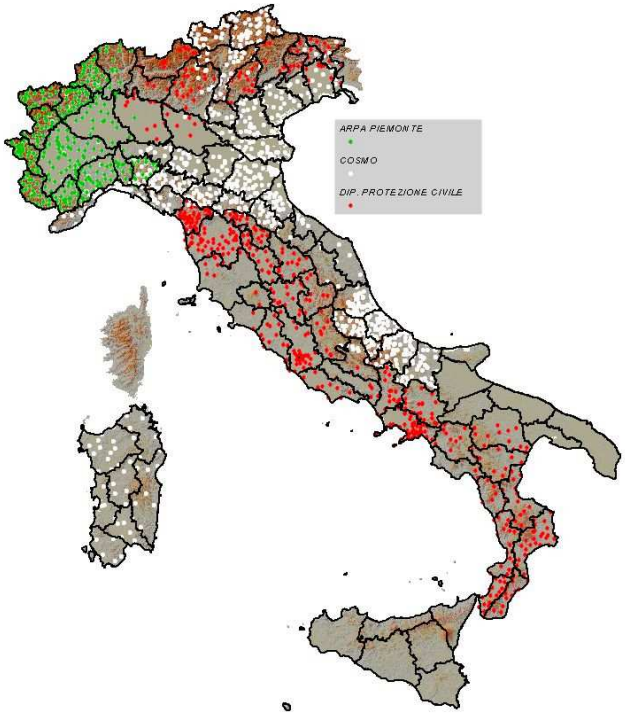

Fig. 3. National Civil Protection warning areas (black contour) and high-resolution non-GTS network (green dots: Piemonte data source, white ones: COSMO consortium, red ones: National Civil Protection Department).

Table 1. The contingency table, where A is the number of forecasted and observed events (hits), B is the number of forecasted but not-observed events (false alarms), C is the number of observed but not-forecasted events (misses), D is the number of not-forecasted and not-observed events (correct negatives).

		OBS	
FOR		YES	NO
	YES	A	B
	NO	C	D

Each station, of the observing network, belongs to a specific warning area and the average of all the observations within each area is calculated and assigned to the corresponding warning area for the chosen accumulation time (e.g., 24 h and 6 h). The same is made for the model: the average of all model grid-point (scalar quantity, centre of the grid) belonging to the chosen warning area is calculated and assigned to the corresponding warning area for verification. The pair of two upscaled quantities (observation and model precipitation forecast) is used for the objective verification.

The observed rainfall is composed by an on/off and a quantitative component: for our needs, it is necessary to evaluate the models' ability in predicting both components, the event occurrence and the estimated QPF. In this context the most useful verification approach is the rainfall changeover from continuous amounts to "exceedance" categories (yes-no statements indicating whether rainfall equals or exceeds a chosen threshold). The thresholds have to be chosen properly, keeping into account the verification's goals and must be supported by reliable statistics in the considered period of time (Efron and Tibshirani, 1986). Finally the forecasted rainfall field's quality can be inferred with classical statistical indices based on contingency tables (Wilks, 1995; Jolliffe and Stephenson, 2003). It is important to underline that an incorrect set of indices may lead to conflicting skill indications (Harvey et al., 1992). Table 1 shows a classical contingency table that is needed to define the indices cited or used in the following, and summarized in Table 2. In order to fully describe the system, a set of three suitably chosen indices is needed (Stephenson, 2000).

Table 2. Verification indices based on 2 × 2 contingency table. See Table 1 for the definition of A, B, C and D.

Index	Formulation
n (total number of cases)	A+B+C+D
BIAS	(A+B)/(A+C)
POD	A/(A+C)
POFD	B/(B+D)
EDS	2[ln((A+C)/n)/ln(A/n)] – 1
BR	(A+C)/n

The selected indices, as suggested by Wilks (1995), are BIAS (namely frequency BIAS, i.e., ratio of the forecasted and observed rain frequency), POD (Probability Of Detection, i.e., the relative number of times an event has been forecasted and it actually occurred) and F or POFD (False Alarm Rate or Probability Of False Detection, i.e., the fraction of observed non-events that were forecast to be events). In addition, as suggested by Ghelli and Primo (2009), in order to analyse the models' capability to predict the rare events correctly, we have chosen the Extreme Dependency Score (EDS) in conjunction with Base Rate (BR).

Generally, model skill is a function of space, time, thresholds and observation network. A long period is necessary to have a sufficient number of data to describe the rainfall statistics significantly (in this case 731 days are taken into account since January 2007 to December 2008), but sometimes, when the data are not homogeneous in space-time, the smoothing of the skill can mask the performance differences. It is suggested to filter the sample into subsets (for example season, morphologically similar areas, etc.) to highlight the model behaviour during a certain weather regime or orographic area.

In this study, Hamill (1999) test hypothesis have been applied, in which a confidence interval is calculated with a bootstrap method (Accadia et al., 2003, 2004) in order to establish the real difference between the skill scores of two competitive forecasts. The test hypothesis assumes that time series have negligible autocorrelations and a 5 % significance level. Doing so, the 95 % confidence intervals for the difference itself are added to the metric of the selected forecast system (the error bars in the graphs are the 2.5 and 97.5 percentile of re-sampling distributions). This means that if the other forecasting system metric is outside this interval, the differences may be considered statistically significant with a confidence of 95 %. As a consequence of the symmetry of this test, it is possible to add the confidence intervals to the other forecast system.

5 Results

In the following sections, a series of objective verification results are presented:

- Long-period verification with increasing thresholds.
- Seasonal trend.
- Diurnal error cycle.
- Error maps.

5.1 Long period verification

This general-purpose verification has been performed over the whole two-year period (January 2007 – December 2008) with the aim of addressing the overall models' behaviour over the whole territory and for a long period of time. BIAS, POD, POFD and EDS are calculated together with 1-BR for the first and second day of forecast, with increasing thresholds ranging from the lowest value of 0.2 mm 24 h^{-1} to the highest of 75 mm 24 h^{-1}. Statistical indices represent the average models' skill in predicting the 24 h accumulated rainfall averaged over each warning area. First and second day results do not show substantial differences except a slight worsening for the second day, so in Figs. 4 and 5 only the results for the first day of forecast are shown. In detail, we can state that:

- By definition, 1-BR plotted vs. increasing thresholds represents the probability that precipitation amount does not exceed a certain threshold. So, in the considered statistical sample, in the 60 % of the cases the average areal precipitation does not exceed 0.2 mm 24 h^{-1}, on the other hand, in the 99 % of the cases the average areal precipitation does not exceed 35 mm 24 h^{-1} and, therefore, the average areal precipitation higher than 35 mm 24 h^{-1} can be considered a "rare event" in our sample.

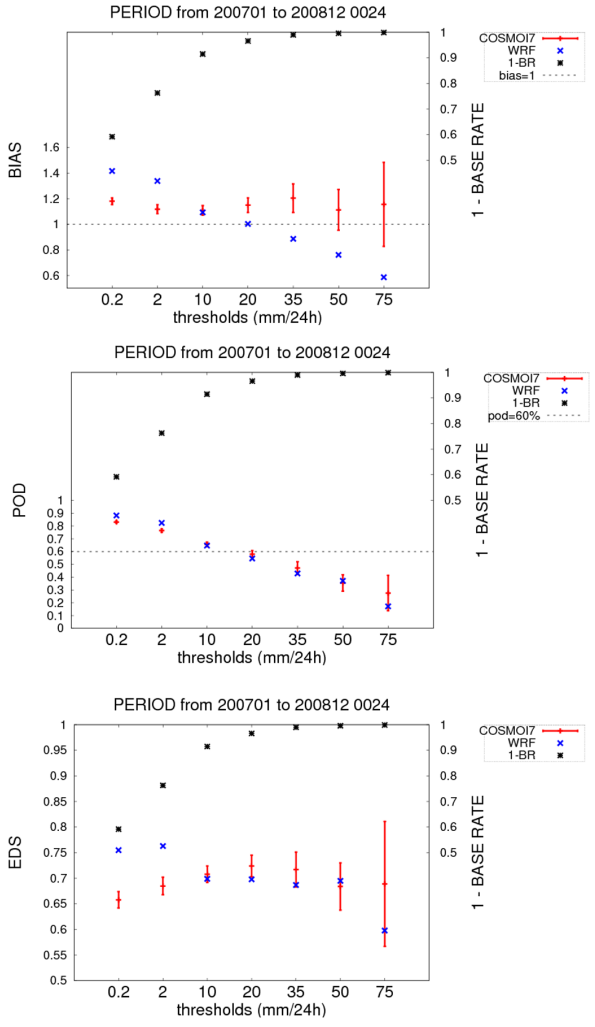

Fig. 4. Long-period verification: BIAS, POD, EDS vs. increasing thresholds for the first day of forecast. The dashed red lines and the blue daggers represent COSMO-I7 and WRF, respectively. The Base Rate (stars) is plotted for each threshold.

- COSMO-I7 shows a quite constant overestimation for all the thresholds (BIAS>1), while WRF shows an overestimation below 20 mm 24 h^{-1} and an underestimation above that threshold (Fig. 4 top panel).
- The differences are statistically significant for all thresholds with the exception of 10 mm 24 h^{-1} (Fig. 4 top panel).
- Concerning POD and POFD (Fig. 4 middle panel and Fig. 5, respectively), both models have decreasing scores with respect to increasing thresholds: even if it is clear a worsening of POD, it is also noticeable that POFD is as low, as rare is the event.
- For very low thresholds (0.2 and 2 mm 24 h^{-1}) WRF has a POFD significantly larger than COSMO-I7 (Fig. 5).

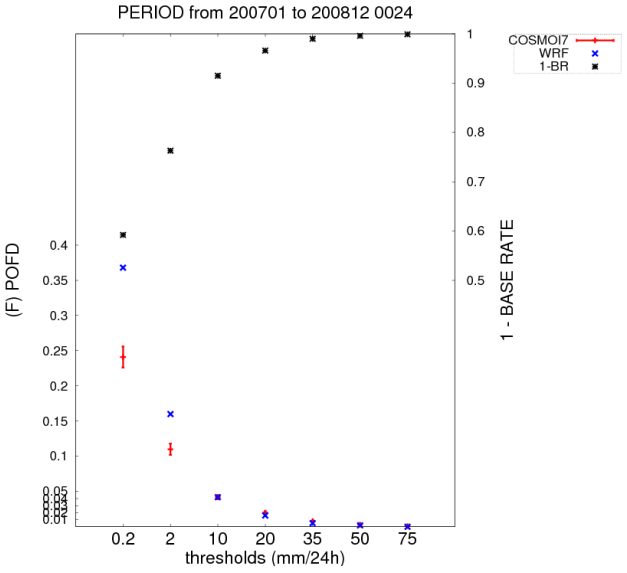

Fig. 5. Long-period verification: POFD vs. increasing thresholds for the first day of forecast. The dashed red lines and the blue daggers represent COSMO-I7 and WRF, respectively. The Base Rate (stars) is plotted for each threshold.

5.2 Seasonal trend

A seasonal (3 months period) aggregation is used to investigate models' pluviometric regime skill dependence. In Italy, autumn (September-October-November, SON) and winter (December-January-February, DJF) seasons are generally wetter with widespread stratiform rainfall. Spring (March-April-May, MAM) and summer (June-July-August, JJA) are generally drier with mainly convective scattered rainfall.

The calculated statistical indices represent the seasonal warning areas averaged forecasted and observed 24 h accumulated rainfall exceedance for three chosen thresholds ($0.2\,\mathrm{mm\,24\,h^{-1}}$, $10\,\mathrm{mm\,24\,h^{-1}}$, $20\,\mathrm{mm\,24\,h^{-1}}$). In addition, the 1-BR is plotted for each season to check the Base Rate variability during the time.

In Fig. 6, the BIAS for the first day of forecast starting from winter 2007 (only January and February) to autumn 2008 is shown:

- For the lowest threshold there is a little Base Rate variability with a higher 1-BR value during summertime when we expect a large number of rare events. Moreover, there is an evident cycle with minimum winter BIAS and maximum summer BIAS. The amplitude of this cycle is more pronounced in WRF than in COSMO-I7 and the differences are always statistically significant (Fig. 6 top panel). Besides, there is a general overestimation of rainfall for both models.

- For the medium threshold (Fig. 6 middle panel), it can be noticed that there is a downshift of models BIAS, but again in summer the WRF overestimation is definitely present. Spring and autumn 2008 belong statistically to the same population. For this thresholds there are not significant variations of Base Rate.

- For the highest threshold (Fig. 6 lower panel), while WRF maintains the same features shown for the other ones, COSMO-I7 worsens its performance in spring and summer 2007 drastically. It has to be pointed out that for this threshold the statistics is poorer, with respect to, the previous cases, even though the number of cases is high enough to permit a statistical description, as shown in Table 3. Also for this threshold the Base Rate does not vary significantly.

- Focusing on spring 2007, we notice a large BIAS for COSMO-I7. Although a deeper and dedicated investigation is needed, it has to be said that that period was abnormally dry, as reported in periodic reports written by the Italian National Met Service (VV. AA., 2007). The presence of a persistent high-pressure structure determined (especially in March and April 2007) very stable conditions, with some weak midday convection close to the mountains. In these conditions, with almost no synoptic forcing, the Tiedtke scheme seems to overestimate the convection intensity, and it is also confirmed in the statistics given by the Swiss Met Service regarding the Swiss version of COSMO (Hug, 2007). Moreover, south of the Alps they observed a positive (hence overestimation) Relative Humidity BIAS between 600 and 150 hPa.

Table 3. Number of cases with a $20\,\mathrm{mm\,24\,h^{-1}}$ threshold in each three-months period (first column): observations (second column), COSMO-I7 forecast (third column) and WRF forecast (fourth column).

	OBS.	COSMO-I7	WFR
JF'07	109	140	87
MAM'07	201	361	190
JJA'07	136	214	225
SON'07	254	319	257
DJF'08	159	163	138
MAM'08	353	325	288
JJA'08	149	168	224
SON'08	480	450	445

In Figs. 7 and 8, the seasonal behaviour of POD and POFD, respectively, for the three thresholds is shown. The large overestimation in BIAS for the lower threshold (Fig. 6 upper panel) is reflected in the very high POD (Fig. 7 upper panel) and, with a minor impact during the summer periods, also in POFD (Fig. 8 upper panel). For the medium and highest thresholds POD lowers to more typical values,

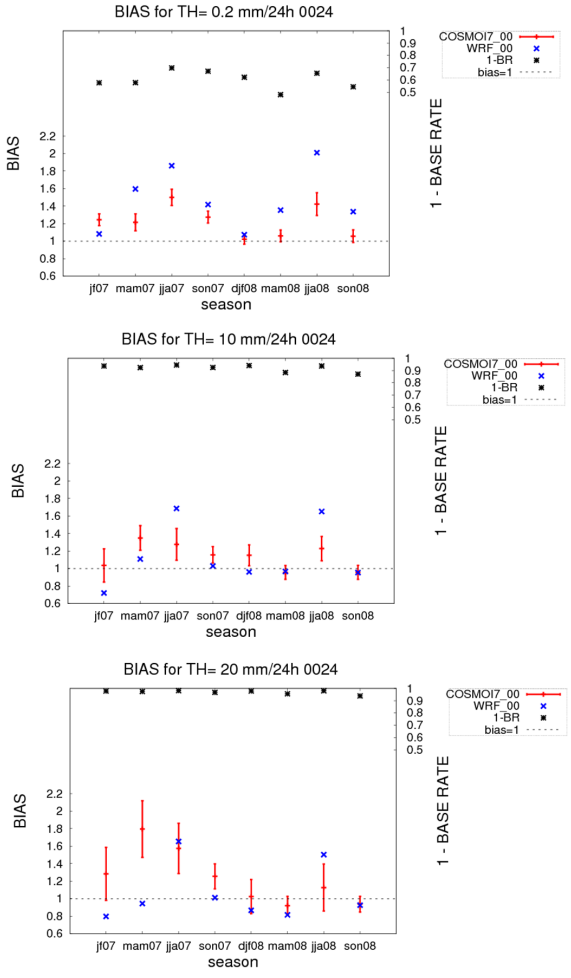

Fig. 6. Seasonal trend: BIAS vs. increasing thresholds ($0.2\,\mathrm{mm\,24\,h^{-1}}$, $10\,\mathrm{mm\,24\,h^{-1}}$, $20\,\mathrm{mm\,24\,h^{-1}}$) for the first day of forecast. The dashed red lines and the blue daggers represent COSMO-I7 and WRF, respectively. The Base Rate (stars) is plotted for each threshold.

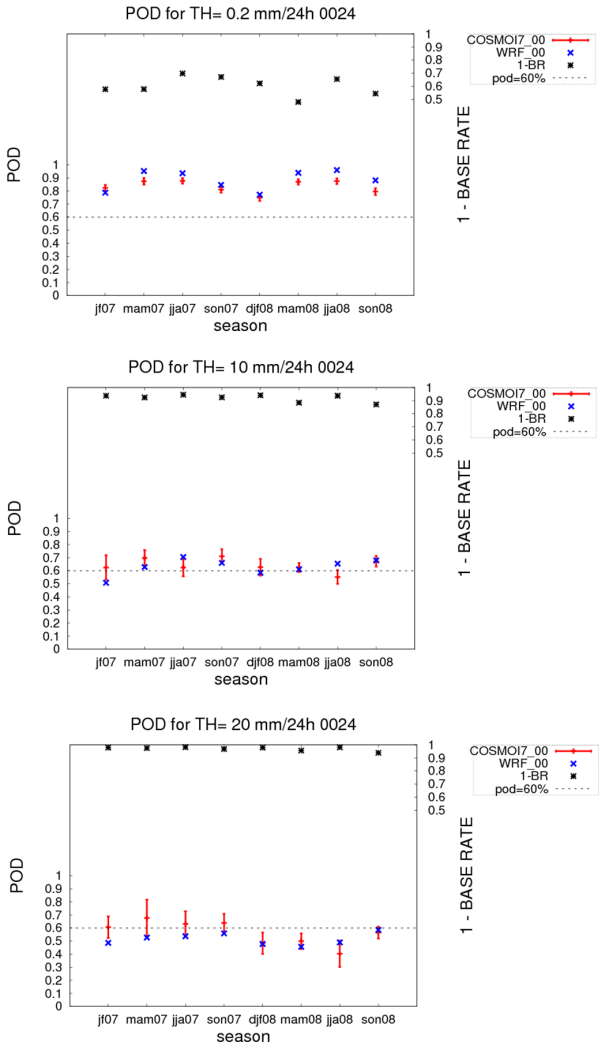

Fig. 7. Seasonal trend: POD vs. increasing threshold ($0.2\,\mathrm{mm\,24\,h^{-1}}$, $10\,\mathrm{mm\,24\,h^{-1}}$, $20\,\mathrm{mm\,24\,h^{-1}}$) for the first day of forecast. The dashed red lines and the blue daggers represent COSMO-I7 and WRF, respectively. The Base Rate (stars) is plotted for each threshold.

and POFD tends to decrease, but presents higher values during spring and summer. Concerning EDS (Fig. 9), for the lowest threshold WRF presents in general values higher than COSMO-I7, but for higher thresholds (larger number of rare events), we obtain a different behaviour during almost the whole period: in 2007 the association between forecasted and observed events is better for COSMO-I7, while in 2008 both the models worsen, in particular COSMO-I7.

The worsening of WRF during summer (BIAS) is probably due to the onset and development of convective rainfall, which is clearly enhanced in an unrealistic way. The reason for this behaviour can be found in the characteristics of the Kain-Fritsch scheme (see Kain and Fritsch, 1990, 1993 and Bechtold et al., 2001): here the closure assumption is based on CAPE and the onset of convection depends on the large-scale vertical velocity (in the Tiedtke scheme the closure assumption is based on moisture convergence and convection is triggered if the parcel's temperature exceeds the environment temperature by a fixed temperature threshold of 0.5 K, see Tiedtke, 1989). This determines an overestimation of the average rainfall over the plains and a reduction of the maxima over the mountains, with respect to the Tiedtke scheme. Consequently, having a large domain, more and more flat regions are included and this counterbalances the under prediction over the Alps. This result confirms what has been found in Milelli et al. (2008).

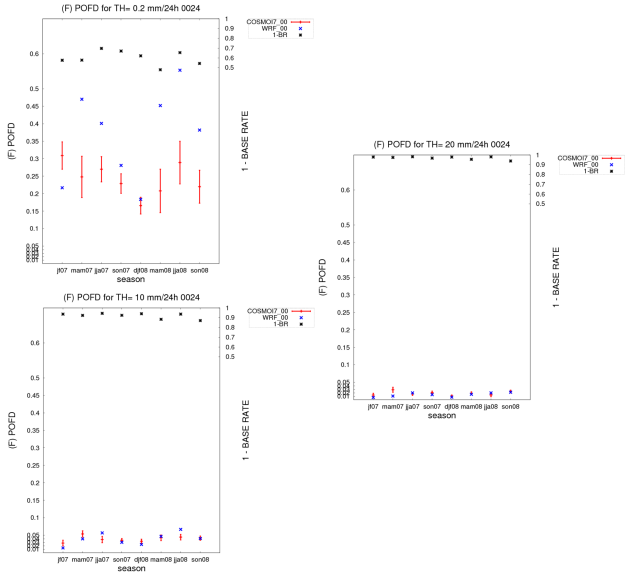

Fig. 8. Seasonal trend: POFD vs. increasing threshold ($0.2\,\mathrm{mm\,24\,h^{-1}}$, $10\,\mathrm{mm\,24\,h^{-1}}$, $20\,\mathrm{mm\,24\,h^{-1}}$) for the first day of forecast. The dashed red lines and the blue daggers represent COSMO-I7 and WRF, respectively. The Base Rate (stars) is plotted for each threshold.

A consideration has to be done for very high thresholds such as $50\,\mathrm{mm\,24\,h^{-1}}$: the authors agree that it would be useful to include also this information in order to draw an overall conclusion on the models' capability to reproduce intense precipitation events, but we count a very limited number of cases for the three-months aggregation. For this reason we believe that this threshold might be misleading, being the statistics very poor.

5.3 Diurnal error cycle

The error in the diurnal cycle considering the 6 h-accumulated rainfall (averaged over each warning area) is investigated in this section. This is generally a more difficult task for LAMs because they are requested to forecast the event in a shorter period of time.

Two thresholds have been chosen as a reference: $2\,\mathrm{mm\,6\,h^{-1}}$ (Fig. 10) and $10\,\mathrm{mm\,6\,h^{-1}}$ (Fig. 11). In addition, the 1-BR is plotted at each time step to check the Base Rate variability, which remains quite constant for both thresholds.

For the lowest threshold there is a clear diurnal cycle, with maxima at noon and minima at 06:00 UTC. The main difference between the models appears in the first 24 h of forecast, where WRF has higher values of indices at noon and lower at night with respect to COSMO-I7. After the first day, the forecasts are much similar and the differences do not belong to different populations from a statistical point of view (Fig. 10). It has to be underlined a general worsening with the forecast time, in particular for POD and EDS and

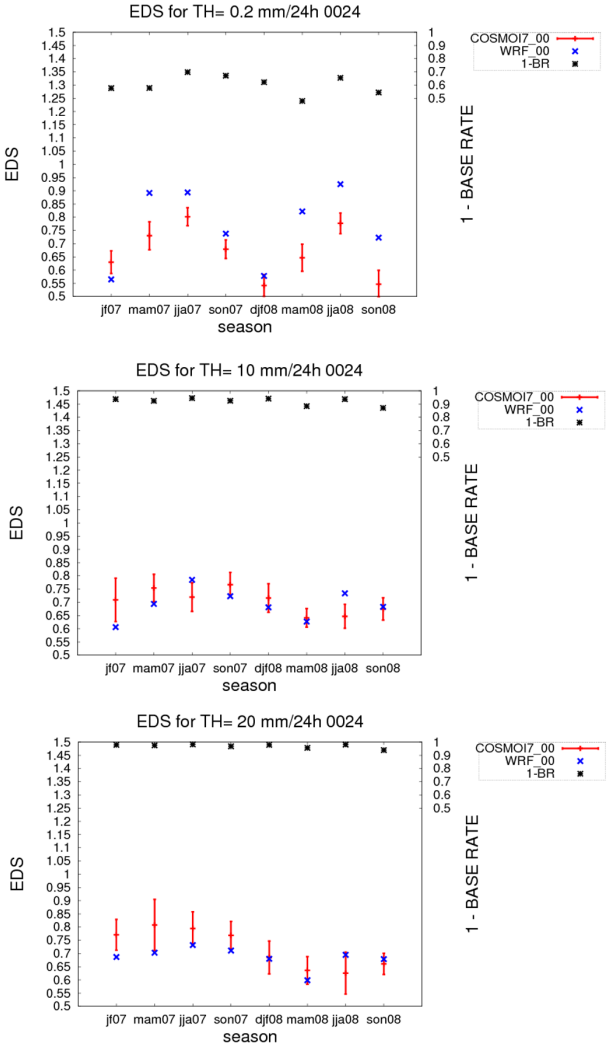

Fig. 9. Seasonal trend: EDS vs. increasing threshold ($0.2\,\mathrm{mm\,24\,h^{-1}}$, $10\,\mathrm{mm\,24\,h^{-1}}$, $20\,\mathrm{mm\,24\,h^{-1}}$) for the first day of forecast. The dashe red lines and the blue daggers represent COSMO-I7 and WRF respectively. The Base Rate (stars) is plotted for each threshold.

an overestimation of rainfall in both models (Fig. 10 upper panel). On the contrary, for $10\,\mathrm{mm\,6\,h^{-1}}$ threshold, WRF presents a constant underestimation (Fig. 11 upper panel). Both models have a worsening with the forecast time and after the first day of forecast they tend to be statistically independent: COSMO-I7 seems to perform better at least in terms of POD and EDS, but with a greater overestimation and False Alarm Rate. Having a BIAS smaller than COSMO-I7, WRF has also a lower POD and a lower POFD (Fig. 11 middle and lower panel, respectively).

In correspondence with the $10\,\mathrm{mm\,6\,h^{-1}}$ threshold, WRF has a much lower BIAS with respect to COSMO-I7 and this might be explained with the lack of a data assimilation cycle,

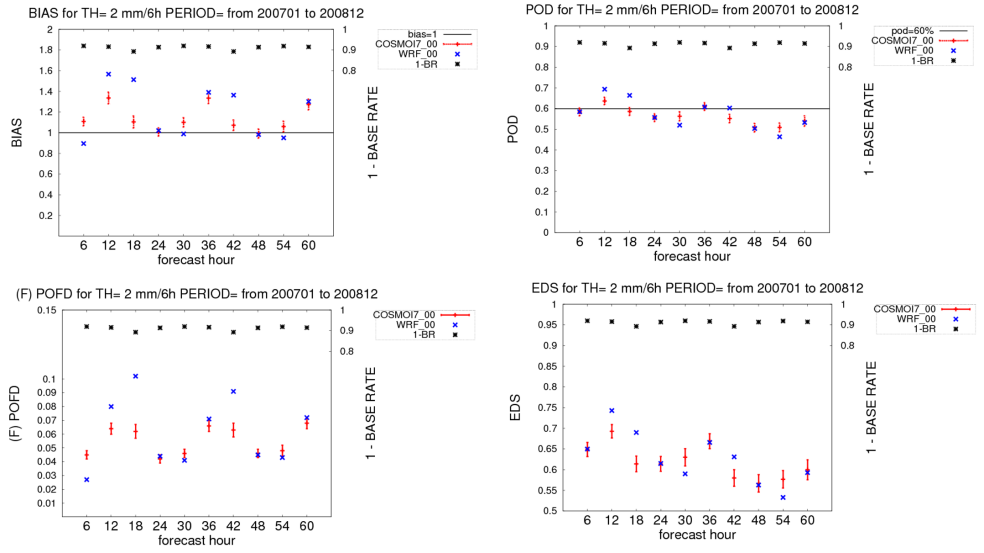

Fig. 10. Diurnal cycle: BIAS, POD, POFD and EDS for $2\,\mathrm{mm}\,6\,\mathrm{h}^{-1}$ threshold. The dashed red lines and the blue daggers represent COSMO-I7 and WRF respectively. The Base Rate (stars) is plotted for each threshold.

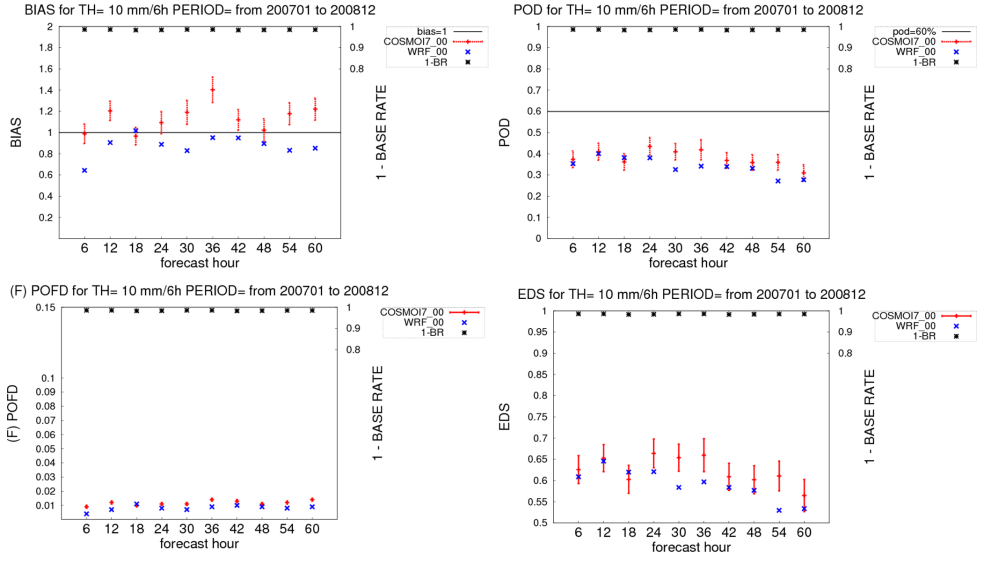

Fig. 11. Diurnal cycle: BIAS, POD, POFD and EDS for $10\,\mathrm{mm}\,6\,\mathrm{h}^{-1}$ threshold. The dashed red lines and the blue daggers represent COSMO-I7 and WRF respectively. The Base Rate (stars) is plotted for each threshold.

present in COSMO-I7. Moreover, COSMO model has an extra assimilation window of 4 h as explained in Sect. 2.1. As a consequence WRF spin up is more pronounced with respect to COSMO-I7.

5.4 Error maps

In the last paragraph, a selection of relative error maps is presented. These maps are useful because the main model characteristics (QPF error) with respect to the warning areas properties (e.g., topography, location), the season and the rainfall climatology are highlighted. Results for the autumn period are presented because is the wettest period of the year and most of the severe events are registered. Besides, we show also the results for the summertime to highlight the big models' overestimation, in particular, for WRF: it is a crucial point for the NCPD to predict correctly the precipitation amount during the most convective periods because a very intense and localized flash flood can cause damages.

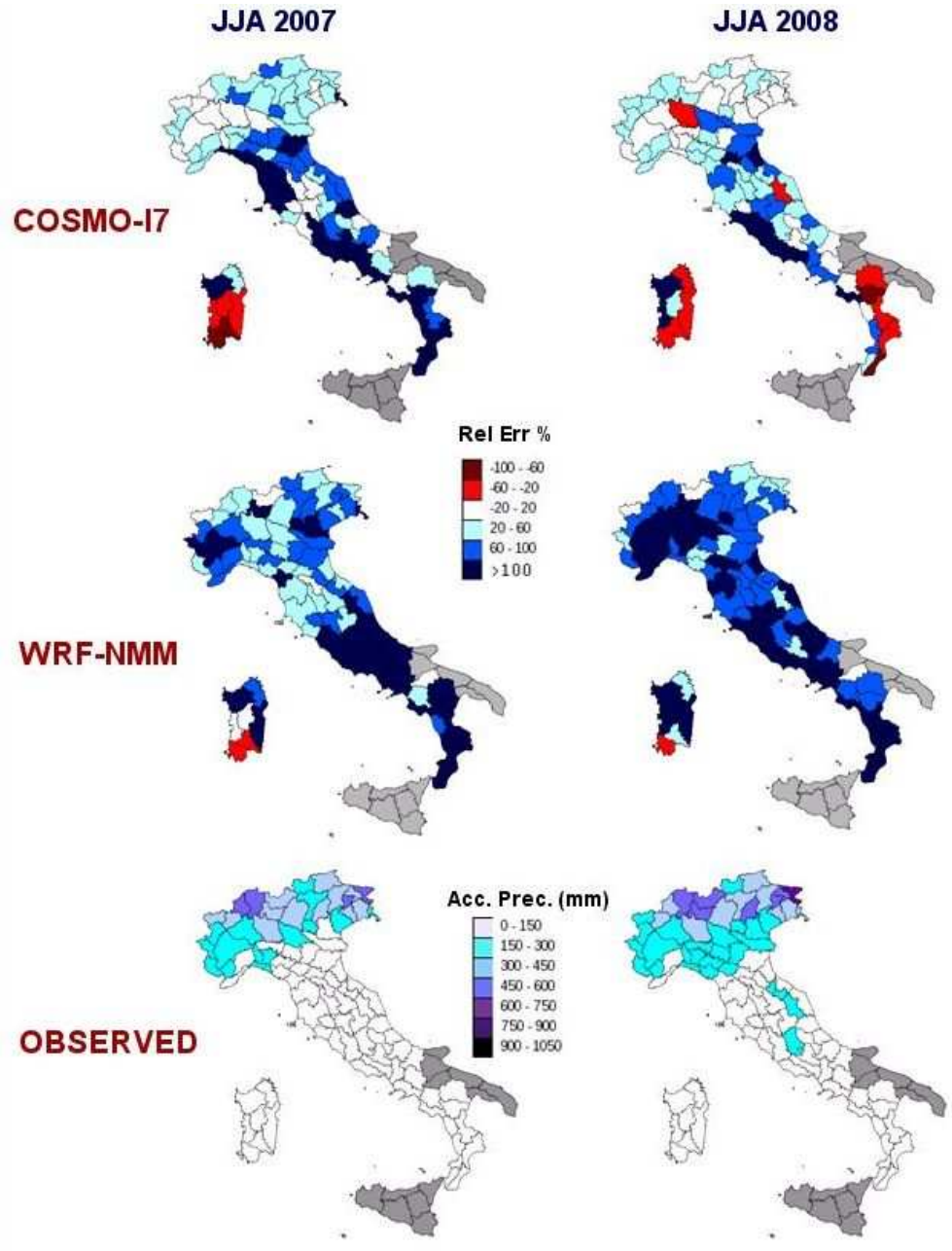

Fig. 12. Model error maps (first two rows) and corresponding observed rainfall (bottom row) for summer 2007 (left column) and 2008 (right column). First row: COSMO-I7 relative errors. Second row: WRF relative errors. For relative error maps red colours means model underestimation, blue colour model overestimation.

The first column of Fig. 12 shows results for the 3 months period of JJA 2007, the second column for the JJA 2008 period. The first row (COSMO-I7) and the second one (WRF) of the same figure show model relative error maps. The third row shows the observed total averaged rainfall over a 3 months period. A similar structure has been built up in Fig. 13 for SON 2007 and 2008.

Models maps are calculated by taking the relative difference between forecast and observation of 24 h accumulated precipitation for the 3 months period, averaged over each warning area (Mullen and Buizza, 2001). More specifically:

$$\text{error} = \frac{\text{For} - \text{Obs}}{\text{Obs}}\%. \quad (1)$$

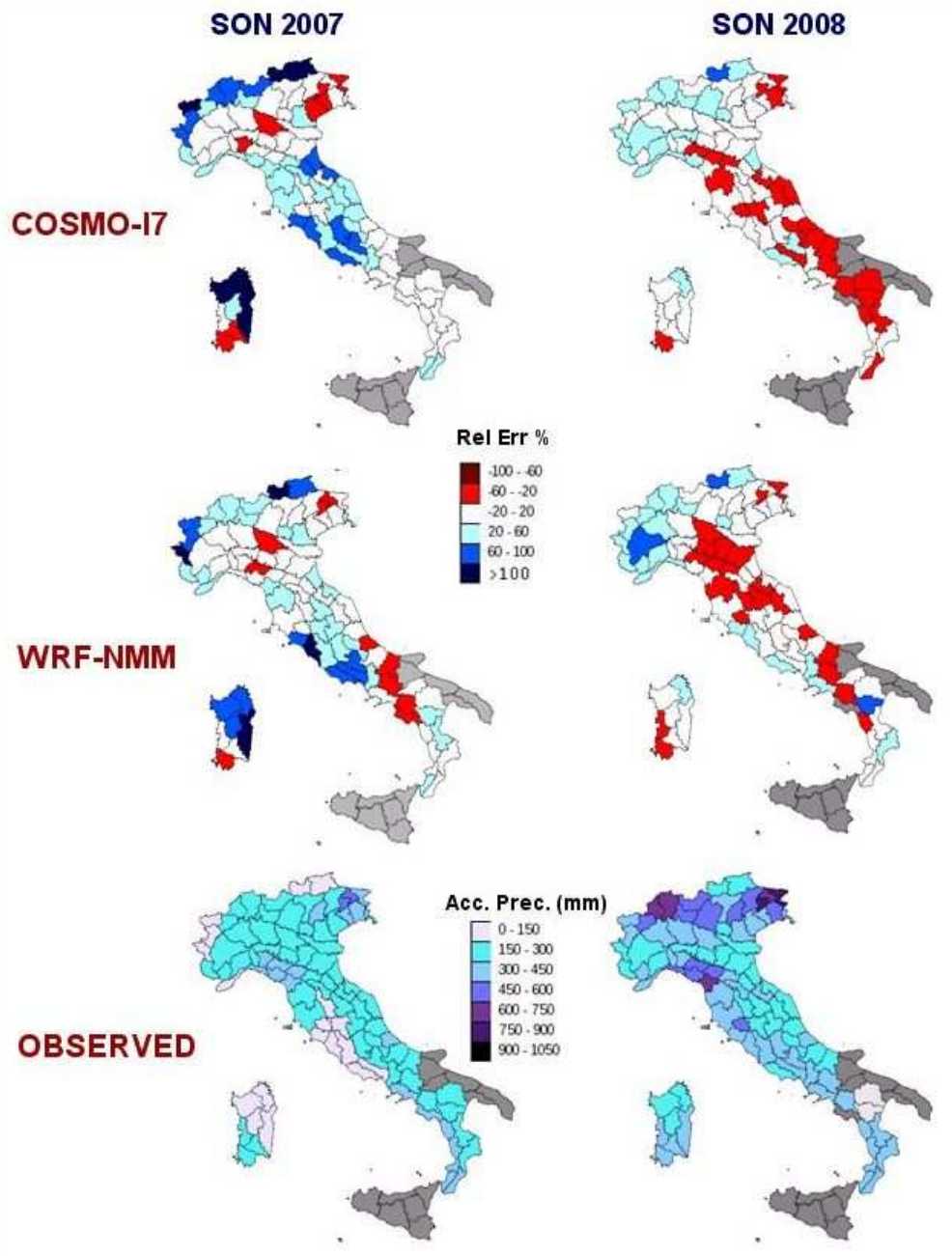

Fig. 13. Model error maps (first two rows) and corresponding observed rainfall (bottom row) for autumn 2007 (left column) and 2008 (right column). First row: COSMO-I7 relative errors. Second row: WRF relative errors. For relative error maps red colours means model underestimation, blue colour model overestimation.

Where "For" in the Eq. (1) represents the sum of the daily cumulated precipitation averaged over all the grid points falling inside a certain warning area, from the first day of June (September) to the last day of August (November). Similarly, "Obs" in the Eq. (1) represents the sum of the daily cumulated precipitation averaged over all the station points falling inside a certain warning area, from the first day of June (September) to the last day of August (November). Only the first day of forecast has been considered here. Red values indicate that model underestimates, blue ones that model overestimates. The observation map is the term "Obs" in Eq. (1). Colours indicate the rainfall amounts with 150 mm increments.

The main considerations are the following:

- A general models overestimation during summertime (Fig. 12), more pronounced for WRF especially over plain areas. That confirms the hypothesis explained in Sect. 5.2 regarding the Kain-Fritsch scheme.
- The pluviometric regime of the models during the two summer seasons seems quite similar (Fig. 12 bottom panels).
- In northern Italy, especially over the Alps, SON 2007 was drier than 2008. In central and southern Italy, the amount of rainfall does not change drastically from one year to the other (Fig. 13 bottom panels).
- The behaviour of the model changes according the area, being more accurate over northern Italy (and the Alps), and less good over central-south Italy where there is a general overestimation in 2007 and an underestimation in 2008.

As a general remark it seems that models behaviour depends critically on the circulation pattern. Nonetheless, the models seem to forecast on medium term (seasonally) amounts of rainfall around some kind of "climatological" value. This aspect should be investigated more, taking into account a larger number of seasons clustered together with the aid of weather types definition.

6 Conclusions

In this study we have performed a two-years QPF verification of two operational LAMs producing forecasts for the Italian territory. The verification has been realized to give a better understanding about models behaviour and utility in producing reliable forecasts for the Italian NCPD. In order to achieve this task, the averaged rainfall values over 90 warning areas, which are more useful for hydrological purposes, have been used. Results from long period verification with increasing thresholds, seasonal trend, diurnal error cycle and error maps, are presented.

In general, the two models have good performances in estimating this kind of areal averaged QPF. WRF has a general tendency of QPF overestimation for low thresholds and underestimation for higher ones, while COSMO-I7 tends to overestimate for all thresholds. Quite obviously the results are generally better in the first 24 h than in the second 24 h, POD and POFD decreases with increasing thresholds and COSMO-I7 has a lower percentage of non events forecasted "yes" for low thresholds. For very high thresholds COSMO-I7 presents a better hit-rate (POD) and the association between forecasted and observed events is definitely better, a valuable feature for Civil Protection purposes.

The results have shown that a seasonal trend for both models exists: they overestimate during summer (more WRF, because the Kain-Fritsch scheme produces an enhancement of convective rainfall especially over flat areas) and spring (more WRF for lower thresholds, more COSMO-I7 for higher ones), while during autumn and winter the models tend to be more accurate. Models present also an evident diurnal cycle with a general overestimation during daytime (both for lower thresholds, more COSMO-I7 for higher ones). The presence of spin up in WRF (very low BIAS in the first hours of the run, evident for high thresholds) is an indication of the lack of data assimilation. Moreover, error maps have allowed us to better understand the spatial distribution of the models' error, also in relation to pluviometric and weather regime, showing that there is a defined bias towards overestimation during summertime. Besides, model results depend critically on the general circulation affecting the territory, an aspect that should be further investigated with the aid of weather types or re-analysis runs.

As a general remark about the methodology, authors are aware of the limits caused by using the same thresholds for all the areas, especially in very complex orography. By doing so, the threshold that defines a "rare event" might be sensibly different from an area to the other. As a first step, the work focuses on the general behaviour of the two models over the whole Italian territory, paying more attention to the differences between the two systems rather than on the performances over specific areas. A concrete improvement to this aspect would be the use of variable thresholds calculated using the percentiles for each verifying area. In future work, this aspect will be addressed and investigated more carefully.

Finally, it is the authors' belief that these indications might be useful for the operational use of the models under study. In fact, the two models are operationally used by the NCPD (COSMO) and by some Italian regional weather service (WRF) for the evaluation and the issue of warnings in case of potentially severe events leading to floods and flash floods, which sadly are quite common for the particular topography of the Italian Peninsula.

Acknowledgements. The authors wish to thank their colleagues at ARPA Piemonte, LAMMA and CNR for the contribution in the discussions during these months. In particular, Marco Turco (now at University of Barcelona) for the implementation of the bootstrap technique. We thank the National Department of Civil Protection as the main end-user and supporter of our work and the Italian regional institutions for the data sharing, in particular: ARPA Emilia Romagna (Maria Stefania Tesini), ARPA Liguria (Nicola Arena), ARPA Sardegna (Paolo Boi), ARPA Friuli Venezia Giulia (Dario Giaiotti), Provincia autonoma di Trento (Paolo Cestari), Provincia Autonoma di Bolzano (Martin Pernter), Centro Operativo di Agrometeorologia delle Marche (Michela Busilacchi), Dipartimento di Fisica dell'Università dell'Aquila (Marco Verdecchia). We are also grateful to two anonymous referees for their valuable comments and to the editor for having organized the production of this paper.

Edited by: K. Lagouvardos
Reviewed by: two anonymous referees

References

Accadia, C., Casaioli, M., Mariani, S., Lavagnini, A., Speranza, A., De Venere, A., Inghilesi, R., Ferretti, R., Paolucci, T., Cesari, D., Patruno, P., Boni, G., Bovo, S., and Cremonini, R.: Application of a statistical methodology for limited area model intercomparison using a bootstrap technique, Il Nuovo Cimento C, 26, 1, 61–77, 2003.

Accadia, C., Mariani, S., Casaioli, M., Lavagnini, A., Speranza, A.: Verification of Precipitation Forecasts from Two Limited-Area Models over Italy and Comparison with ECMWF Forecasts Using a Resampling Technique, Wea. Forecast., 20, 276–300, 2004.

Bechtold, P., Bazile, E., Guichard, F., Mascart, P., and Richard, E.: A mass flux convection scheme for regional and global models, Q. J. R. Meteorol. Soc., 127, 869–886, 2001.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown B. G., and Mason, S.: Forecast verification: current status and future directions, Meteorol. Appl. 15, 3–18, 2008.

Chen, S. H. and Dudhia, J.: Annual report: WRF physics, Air Force Weather Agency, 38pp, (available online at http://www.mmm.ucar.edu/wrf/users/docs/wrf-phy.html), 2000.

Chen, F. and Dudhia, J.: Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity, Mon. Wea. Rev., 129, 569–585, 2001.

Cherubini, T., Ghelli, A., and Lalaurette, F.: Verification of precipitation forecasts over the alpine region using a high-density observing network, Weather Forecast., 17, 238–249, 2002.

Clark, A. J., Gallus Jr., W. A., and Chen, T. C.: Comparison of the Diurnal Precipitation Cycle in Convection-Resolving and Non-Convection-Resolving Mesoscale Models, Mon. Wea. Rev., 135, 3456–3473, 2007.

Colle, B. A., Mass, C. F., and Westrick, K. J.: MM5 Precipitation Verification over the Pacific Northwest during the 1997–99 Cool Seasons, Weather Forecast., 15, 730–744, 2000.

Davies, H. C.: A lateral boundary formulation for multi-level prediction models, Q. J. Roy. Meteor. Soc., 102, 405–418, 1976.

Davies, H. C.: Limitations of some common lateral boundary schemes used in regional NWP models, Mon. Wea. Rev., 111, 1002–1012, 1983.

Dierer, S., Arpagaus, M., Seifert, A., Avgoustoglou, E., Dumitrache, R., Grazzini, F., Mercogliano, Milelli, M., and Starosta, K.: Deficiencies in quantitative precipitation forecast: sensitivity studies using the COSMO model, Meteorol. Z., 18, 6, 631–645, 2009.

Doms, G. and Schaettler, U.: A description of the non-hydrostatic regional model LM. Part I: Dynamics and Numerics, 2002.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorol. Appl. 15, 51–64, 2008.

Efron, B. and Tibshirani, R.: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, Stat. Sci., 1, 54–77, 1986.

Ferrier, B. S., Lin, Y., Black, T., Rogers, E., and Di Mego, G.: Implementation of a new grid-scale cloud and precipitation scheme in the NCEP Eta model, Preprints, 15^{th} Conference on Numerical Weather Prediction, San Antonio, TX, Amer. Meteor. Soc., 280–283, 2002.

Ghelli, A. and Lalaurette, F.: Verifying precipitation forecasts using up-scaled observations, ECMWF Newsletter, No. 87, ECMWF, Reading, UK, 9–17, 2000.

Ghelli, A. and Primo, C.: On the use of the extreme dependency score to investigate the performance of an NWP model for rare events, Meteorol. Appl. 16, 537–544, 2009.

Hamill, T. M.: Hypothesis tests for evaluating numerical precipitation forecasts, Weather Forecast., 14, 155–167, 1999.

Harvey Jr., L. O., Hammond, K. R., Lusk, C. M., and Mross, E. F.: The application of signal detection theory to weather forecasting behavior, Mon. Wea. Rev., 120, 863–883, 1992.

Hug, C.: Verification for March/April/May 2007, available athttp://www.cosmo-model.org/content/tasks/verification.priv/switzerland/default.htm, 2007.

Kain, J. S.: The Kain–Fritsch Convective Parameterization: An Update, J. Appl. Meteor., 43, 170–181, 2004.

Kain, J. S. and Fritsch, J. M.: A one-dimensional entraining / detraining plume model and its application in convective parameterization, J. Atmos. Sci., 47, 2784–2802, 1990.

Kain, J. S. and Fritsch, J. M.: Convective parameterization for mesoscale models: The Kain-Fritsch scheme. The representation of cumulus convection in numerical models, Meteor. Monogr., 27, 165–170, 1993.

Janjic, Z. I.: The surface layer in the NCEP Eta Model, 11th Conference on Numerical Weather Prediction, Norfolk, VA, 19-23 August 1996; Am. Meteor. Soc., Boston, MA, 354–355, 1996.

Janjic, Z. I.: Nonsingular Implementation of the Mellor-Yamada Level 2.5 Scheme in the NCEP Meso model, NCEP Office Note No. 437, 61 pp., 2002.

Janjic, Z. I.: A Nonhydrostatic Model Based on a New Approach, Meteorol. Atmos. Phys., 82, 271–285, http://dx.doi.org/10.1007/s00703-001-0587-6, 2003.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: A Practitioner's Guide in atmospheric Science, John Wiley and Sons, 2003.

Lacis, A. A. and Hansen, J. E.: A parameterisation for the absorption of solar radiation in the earth's atmosphere, J. Atmos. Sci., 31, 118–133, 1974.

Lagouvardos, K., Kotroni, V., Koussis, A., Feidas, H., Buzzi, A., and Malguzzi, P.: The Meteorological Model BOLAM at the National Observatory of Athens: Assessment of Two-Year Operational Use, J. Appl. Meteor., 42, 1667–1678, 2003.

Mass, C. F., Ovens, D., Westrick, K., and Colle, B.: Does Increasing Horizontal Resolution Produce More Skillful Forecasts?, BAMS, 83, 3, 407–430, 2002.

Mellor, G. L. and Yamada, T.: A hierarchy of turbulence closure models for planetary boundary layers, J. Atmos. Sci., 31, 1791–1806, 1974.

Milelli, M., Oberto, E., and Parodi, A.: Sensitivity experiments of a severe rainfall event in North-Western Italy: 17 August 2006, Adv. Sci. Res., 2, 133–138, 2008.

Mullen, S. and Buizza, R.: Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System, Mon. Wea. Rev., 129, 638–663, 2001.

Papadopoulos, A. and Katsafados, P.: Verification of operational weather forecasts from the POSEIDON system across the Eastern Mediterranean, Nat. Hazards Earth Syst. Sci., 9, 1299–1306, doi:10.5194/nhess-9-1299-2009, 2009.

Pappenberger, F., Ghelli, A., Buizza, R., and Bodis, K.: The skill of

probabilistic forecasts under observational uncertainties within the generalized likehood uncertainty estimation framework for hydrological applications, J. Hydrometerol., 10, 807–819, 2009.

Ritter, B. and Geleyn, J. F.: A Comprehensive Radiation Scheme for Numerical Weather Prediction Models with Potential Applications in Climate Simulations, Mon. Wea. Rev., 120, 303–325, 1992.

Ryan, B. F.: On the global variation of precipitating layer clouds, B. Am. Meteor. Soc., 77, 53–70, 1996.

Skamarock, W. C. and Klemp, J. B.: The Stability of Time-Split Numerical Methods for the Hydrostatic and the Nonhydrostatic Elastic Equations, Mon. Wea. Rev., 120, 2109–2127, 1992.

Stauffer, D. R. and Seaman, N. L.: Multiscale four-dimensional data assimilation, J. Appl. Meteor., 33, 416–434, 1994.

Stephenson, D. B., Casati, B., Ferro, C. A. T., and Wilson, C. A.: The extreme dependency score: a non-vanishing measure for forecasts of rare events, Meteorol. Appl., 15, 41–50, 2008.

Stephenson, D. B.: Use of the odds ratio for diagnosing forecast skill, Weather Forecast., 15, 221–232, 2000.

Steppeler, J., Doms, G., Schaettler, U., Bitzer, H. W., Gassmann, A., Damrath, U., and Gregoric, G.: Meso-gamma scale forecasts using the non-hydrostatic model LM, Meteorol. Atmos. Phys., 82, 75–96, 2003.

Tiedtke, M.: A comprehensive mass flux scheme for cumulus parameterization in large-scale models, Mon. Wea. Rev., 117, 1779–1800, 1989.

Various Authors, Aeronautica Militare Italiana, available online at http://clima.meteoam.it/bollettinoMensile.php, 2007.

Wernli, H., Hofmann, C., and Zimmer, M.: Spatial forecast verification methods intercomparison project: application of the SAL technique, Weather Forecast., 24, 1427–1484, 2009.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences. An Introduction, Academic Press, San Diego, xvii + 627 pp., 1995.