Verification and Intercomparison of Precipitation Fields Modelled by LAMS in the Alpine Area: Two FORALPS Case Studies

Stefano Mariani\textsuperscript{1, 2}, Marco Casaioli\textsuperscript{1}, Alexandre Lanciani\textsuperscript{1}, Nazario Tartaglione\textsuperscript{3}
and Christophe Accadia\textsuperscript{4}

\textsuperscript{1} Agenzia per la Protezione dell’Ambiente e per i Servizi Tecnici (APAT), Rome, Italy
\textsuperscript{2} Dipartimento di Matematica, Università di Ferrara, Ferrara, Italy
\textsuperscript{3} Dipartimento di Fisica, Università di Camerino, Camerino, Italy
\textsuperscript{4} EUMETSAT, Darmstadt, Germany

10 April 2008

Number of Figures: 8

Corresponding author: Dr. Stefano Mariani
Agenzia per la Protezione dell’Ambiente e per i Servizi Tecnici (APAT)
Via Curtatone, 3
00185 Rome, Italy
E-mail: stefano.mariani@apat.it

Abstract
Sharing knowledge and experience about NWP verification among forecasting centres, research institutions and environmental agencies in the Alpine area is one of the main goals of the EU FORALPS project – INTERREG IIIB Alpine Space. In this framework, state-of-the-art verification techniques employed by project partners have been shared and applied to numerical simulations of selected case studies of rainfall events on the north-eastern Alpine area. In this work, a multi-method approach is followed. Simulations of two events with three limited area models (LAMs) are discussed and verified with traditional and more innovative techniques, including both traditional eyeball methods, categorical scores and skill scores, along with advanced object-oriented techniques such as the continuous rain area (CRA) analysis and multiscale methods like the spectral analysis. Simulations of satellite observed fields (METEOSAT-7 Water Vapour channel) are also employed. An attempt has also been made to check qualitatively the impact of initialisation on the model error growth. Most importantly, it is shown how this multiplicity of perspective allows a deeper understanding of the results and facilitates their interpretation, thus suggesting that this approach should be followed in future works.
1. Introduction

Verification is a main component of both meteorological research and operational forecasting activities. Within the EU project FORALPS – INTERREG IIIB Alpine Space (http://www.foralps.net), it has been decided to conduct a survey on the different verification methods used by the project partners. From this survey, a technical report dealing with both operational and research applications has been produced. This report includes, among others, eyeball verification, graphical summary techniques, categorical scores and skill scores, spatial verification techniques, etc. In order to share knowledge about verification techniques and best practices, and compare and evaluate the suitability of the different methods in describing the model forecast quality, a common verification activity on predefined case studies has been planned (Mariani and Casaioli, 2008).

Results presented here refer to two selected events (16–20 Nov. 2002 and 08–10 Sep. 2005) when significant rainfall over the eastern Alpine range (Friuli Venezia Giulia region, Italy) was observed. Both the events are connected with the passage of a depression over the Mediterranean region, although with a markedly different phenomenology. The complexity of Mediterranean weather – which is classically defined as “secondary” (Speranza et al., 2007) due to the deep interaction between local forcing terms (orography, surface thermal pattern etc.) and global/synoptic forcing in shaping weather systems – makes model skill in predicting rainfall patterns highly variable from case to case.

Objective (scores, CRA) and subjective verification techniques have been employed jointly since the results from objective and subjective studies can be interrelated. On one hand, objective methods can provide a quantitative basis to subjective verification results; on the other hand, the latter can suggest a physical interpretation of the quantitative verification achievements. For example, the categorical scores provide feedback on the different agreement between forecast and observed patterns for different precipitation intensities, while displacement errors are quantified by means of the CRA analysis.

However, these variables all measure point to point matching, and are sensitive to small displacement errors, due to the double penalty effect (see, e.g., Accadia et al., 2005). Higher order moments must be studied to assess if the fields being compared are defined on grids with the same effective resolution, and if they are, that they have the same amount of small-scale detail. In this regard, the power spectrum is an effective tool which has been applied on these same case studies, as reported in Lanciani et al. (2008). The results shown here are discussed in light of that work.

A further step – but a fully nontrivial task! – should be the exploitation of both results in order to identify model error sources, and eventually to define the causal chain bringing from an original error
(e.g., inadequacy in the initial analysis and/or boundary conditions) to the diagnosed displacement of forecast rainfall fields with respect to the observed ones. Due to the complex underlying meteorological phenomena, a physical explanation for the forecast error based solely on the objective verification is difficult. Still, in one of the two analysed cases, the model error can be successfully tracked through comparison with satellite water vapour imagery using the pseudo-water vapour (PWV) technique following Fehlmann et al. (2000).

The paper is organized as follows. Forecast and observation data are described in section 2. A meteorological description of the two events is presented in section 3. Subjective verification is described in section 4. Section 5 discusses the objective verification results. In section 6 error source analysis is presented by means of a satellite water vapour comparison. Conclusions are reported in section 7.
2. Meteorological models and rain gauge dataset

2.1 Models

Three LAMs were considered for this study: the Slovenian version of the Aire Limitée Adaptation dynamique Développement InterNational (ALADIN) model (http://www.cnrm.meteo.fr/aladin/) operational at the Environmental Agency of the Republic of Slovenia (EARS); the QUADRICS BOlogna Limited Area Model (QBOLAM; Speranza et al. 2004, 2007; Mariani et al. 2005) operational at the Italian National Agency for Environmental Protection and Technical Services (APAT); the Weather Research and Forecasting Model (WRF; http://wrf-model.org/) running in a research configuration at the Regional Meteorological Observatory (OSMER) of Friuli Venezia Giulia.

WRF and QBOLAM are forced by the European Centre for Medium-range Weather Forecast (ECMWF) analyses and forecasts, whereas ALADIN is forced by the Météo-France global model (ARPEGE) analyses and forecasts.

Models’ grid size is about 10-11 km. Models have a different domain size (see Fig. 1), covering from the Alpine area (ALADIN and WRF) to the Mediterranean Basin (QBOLAM). They also differ for the parameterisation schemes employed, and about initial and boundary conditions: 1200 UTC ECMWF for QBOLAM and WRF; 0000 UTC ARPEGE for ALADIN. For the precipitation comparison, forecast data have been post-processed on two common grids (with grid size of 0.1° and 0.5°, respectively) by means of a simple nearest-neighbour average method, also know as remapping (Accadia et al., 2003). The remapping procedure is to be preferred to the bilinear interpolation as post-processing procedure, since the latter tends to smooth the structure of the original field (Lanciani et al., 2008). Moreover, the bilinear interpolation does not conserve the total amount of precipitation forecast over the native grid (Accadia et al., 2003). Forecasts have been also accumulated at 24h, starting from 0000 UTC.

2.2 Observations

Two different rain gauge dataset have been considered for the two events. For the November 2002 event, precipitation measurements have been obtained from the rain-gauge network of APAT (former Italian National Hydrographic and Marigraphic Service network – SIMN; 407 gauges over the northern Italy), the Italian Regional Agency for Environmental Protection (ARPA) of Emilia-Romagna (DEXTER system; 147 gauges over the Emilia Romagna region), ARPA of Liguria (24 gauges over the Liguria region), OSMER (25 gauges over the Friuli Venezia Giulia region), EARS (18 gauges over Slovenia), and ZentralAnstalt für Meteorologie und Geodynamik (ZAMG; 145 gauges over Austria). For the September 2005 event, precipitation measurements have been collected from the rain-gauge
network of APAT (147 gauges only over the north-eastern Italy), ARPA of Emilia Romagna (240
gauges), ARPA of Liguria (119 gauges), ARPA of Lombardia (67 gauges over the Lombardia region),
OSMER (25 gauges), EARS (21 gauges), and ZAMG (162 gauges).

In order to produce an adequate 24-h observed rainfall gridded analysis (starting at 0000 UTC)
over the two common verification grids, a two-pass Barnes objective analysis scheme has been used
(Barnes, 1964, 1973), using the implementation proposed by Koch et al. (1983). Grid points that do not
have a rain gauge within a radius of 0.15° were neglected to avoid the excessive rainfall spreading
introduced by the analysis scheme on grid points far from the gauges’ locations. Put differently, the
rainfall observational analyses are available only over northern Italy, Austria and Slovenia.
3. Meteorological description of the two case studies

The November 2002 rainfall event over the eastern Alps was due to the passage of two subsequent weather systems: a synoptic trough associated with a depressions moving from the Biscay Gulf to north-east on days 15 – 17 November, and a cutoff low in the following two days. The analysis in Fig. 2a displays the complex surface pattern on day 17, associated with a large-scale trough aloft, and southerly warm advection over the Alps, with mostly orographic precipitation, as suggested by the observed distribution of lightnings (not shown) over the mountain range. During the passage of the front associated with the first depression over the eastern Alpine region another depression was forming over the Bay of Biscay (Fig. 2b). Such a depression evolved as a cut off low passing over the Tyrrhenian Sea (Fig. 2c), whose circulation was responsible of the precipitation observed over the Alpine region on days 18 and 19 November.

The second event is linked to the slow passage of a small depression over the western Mediterranean on days 7 to 10 September 2005. The analysis (Fig. 2d) displays several instability lines, seemingly connected with the observed precipitation. Lightning observations (not shown) evidence a strong convection activity, reaching its maximum over the central Italy and the eastern Alpine range on day 9 September.
4. Subjective verification

For the 2002 case study, intense rainfall is observed and predicted during days 16 and 18 November, whereas for the rest of the event both models and observations display light rainfall.

More in detail, for day 16 November, subjective verification of 24-h accumulated precipitation, remapped on the 0.1° common grid, shows the good quality of the three LAM forecasts: rainfall mostly follows the orography and the maxima and minima are well located, with small differences among them (not shown). Instead, on day 18 November, the precipitation forecast patterns (Figs. 3b,c,d) markedly differ from each other and with respect to observations (Fig. 3a). Such a decrease in the overall model skill is not so surprising. Actually, the weather system evolution on days 15-16 November is dominated by large-scale forcing (Fig. 2a), whereas the trajectory and the structure of the cut off low visible in Fig. 2c seems to be sensitive to small-scale processes, which can be either unresolved in the initial analysis, or not properly reproduced by LAMs.

Anyway, we restrict here the subjective verification to the 18 November forecasts, since this day seems to provide, for the entire event, the main contribution to the models’ error. For the same reason, later in this paper (Sect. 5), the quantitative results obtained over the whole event will be interpreted in terms of the subjective findings for day 18 November.

The rainfall pattern observed (Fig. 3a) consists in two main features: a “crescent” on the eastern Alps (with a peak over the Friuli Venezia Giulia region) and an arc of light rainfall over the north-western Apennines. This pattern is almost conserved only in the ALADIN forecast (Fig. 3b) which, nevertheless, displays a too much fragmented structure and underestimates the rainfall peak. On the contrary, QBOLAM and WRF well predict the maximum over Friuli Venezia Giulia. Concerning the secondary maximum observed over the Apennines, it is lightly shifted south-eastwards in the ALADIN forecast, and much more in the other two LAMs. Moreover, QBOLAM and WRF predict also a precipitation band going from Marche region (central Italy) to Istria peninsula (Croatia). Although the position of the band cannot be verified without observations over Istria and the Adriatic Sea, a bulk indicator of the band displacement can be found in the shifting of the predicted Apennine maximum with respect to the observed one (being the Alpine rainfall maximum correctly located by all three models). The band displacement is seemingly related with the path followed by the moist air mass responsible for the event and, on turn, with the evolution and the trajectory of the shallow low discussed in Sect. 3.

The same comparison for the 2005 event – when rainfall over the target area is concentrated on day 9 September – leads to different results. The rainfall area is reproduced fairly well by all the three
models, but ALADIN strongly underestimates the rainfall amount. Moreover, it displays again an excess of small-scale structure. QBOLAM seems to provide here a good forecast, except for some overestimation of the heavy rainfall area size, while WRF suffers for an eastward shifting of the main rainfall peak and for an excess of rainfall over the Apennines.
5. Objective verification

5.1 Quantitative precipitation forecast verification

For the quantitative precipitation forecast (QPF) verification, four categorical verification scores have been used. These scores are tallied up on 2×2 contingency tables (Wilks, 1995), which summarize in a categorical way possible combinations of paired forecast and observed events above or below a given precipitation threshold\(^1\). Four categories are then defined: hits; false alarms; misses and correct non-rain forecasts, which are usually indicated as \(a, b, c\) and \(d\), respectively.

The bias score (BIA, also referred as frequency bias; Wilks, 1995) is the ratio between the frequency of \(yes\) forecast (hits + false alarms) and the frequency of \(yes\) observed (misses + correct non-rain forecast):

\[
\text{BIA} = \frac{a + b}{a + c}. \tag{1}
\]

A forecast is said to be unbiased when forecasts and observations are above a selected threshold the same number of times, that is, when the BIA score is equal to one. A BIA value greater than one indicates that the considered model overestimate the frequency of the events above a threshold (overforecasting); whereas a BIA value less than one indicates that the model underestimate the frequency of the events (underforecasting).

The equitable threat score (ETS, also known as Gilbert skill score; Schaefer, 1990) is an accuracy measure for events, that is, it measures how well the forecast \(yes\) events correspond to the observed \(yes\) events:

\[
\text{ETS} = \frac{a - a_r}{a + b + c - a_r}, \text{ with } a_r = \frac{(a + b)(a + c)}{(a + b + c + d)}. \tag{2}
\]

The ETS skill score is a modified version of the well-known critical success index (CSI; Gilbert, 1884) that accounts for the number of hits \((a_r)\) that would be obtained purely by chance (random forecast). A score equal to one represents a perfect score; whilst a value close to zero or negative (the minimum possible value is \(-1/3\)) means that the model has forecast ability equal or worse than a random forecast.

In order to have measures of the accuracy both for events and non-events, the Hanssen and Kuipers skill score (HK; also named the true skill score; Hanssen and Kuipers, 1965) is also used:

\[
\text{HK} = \frac{(ad - bc)}{(a + c)(b + d)}. \tag{3}
\]

\(^1\) Observation thresholds used in this study are 1.0, 5.0, 10.0, 20.0, 30.0, 40.0, and 50 mm (24 h)\(^{-1}\).
This score ranges between minus one and one (perfect forecast). A positive HK value means that
the forecast outperforms the random forecast, whereas a negative HK value indicates the opposite.
Another measure, the odds ratio skill score (ORSS; Stephenson, 2000), has been used as well. The
ORSS is obtained by transforming the odds ratio ($\theta$; Stephenson, 2000), which is the ratio between the
odds to make a good forecast (i.e., to score a hit) and the odds to make a bad forecast (i.e., to score a
false alarm), into a skill score. More precisely, $\theta$ is given by:

$$\theta = \frac{H}{1 - H} \left( \frac{F}{1 - F} \right)^{-1} = \frac{ad}{bc}, \hspace{1cm} (4)$$

where $H = a/(a + b)$ and $F = b/(b + d)$ are the hit rate (or probability of detection) and the conditional
false alarm rate$^2$ (or probability of false detection), respectively. Hence:

$$ORSS = \frac{\theta - 1}{\theta + 1} = \frac{H - F}{H + F - 2HF} = \frac{(ad - bc)}{(ad + bc)}, \hspace{1cm} (5)$$

which ranges from minus one to one. A positive ORSS value denotes skill. For a perfect forecast (i.e., $a$
and $d$ different from zero; whereas $b$ and $c$ both equal to zero) the ORSS skill score is equal to one.

Skill scores’ intercomparison, including the ECMWF forecast, has been performed over the 0.5°
grid where the forecasts’ structure is comparable as evidenced by the spectral analysis (Lanciani et al.,
2008). Instead, over the 0.1° grid, the results lead to incorrect conclusions on the real skill or real
differences in skill of the compared forecasts. In particular, the skill scores on the 0.1° grid penalize the
less smooth ALADIN forecasts due to the double penalty effect.

Only grid points belonging to the model domains’ intersection (see Fig. 1) and on which there are
observations have been considered for the calculation of the categorical scores given by equations (1) –
(3) and (5).

Results from QPF objective verification are displayed on Figs. 5 and 6 for the 2002 and 2005
events, respectively. The behaviour of the categorical scores confirms the results of the subjective
analysis (see Sect. 4).

For the 2002 event, up to the 30 mm (24h)$^{-1}$ threshold, ALADIN has the highest skill scores and
the lowest bias among the models, overperforming ECMWF at all thresholds. This picture changes at
the higher thresholds: here, the best performance is provided by the global model and, among the
LAMs, by WRF (Fig. 5c), which is, nevertheless, affected by an increasingly wet BIA over 40 mm
(24h)$^{-1}$ (Fig. 5a). These results seem to agree with the subjective comparison about day 18 November

$^2$ The conditional false alarm rate must not be confused with the false alarm ratio (FAR; Mason 1989), which is the ratio
between the number of false alarms ($b$) and the total number of yes forecasts ($a + b$).
(Fig. 3), even if some caution is needed, since they refer to the whole event. At lower thresholds, ALADIN’s good skill is explained by the fact that the remapping over the coarser 0.5° grid removes the unrealistic small-scale structures. Concerning WRF, the wet BIA, especially at higher thresholds, is recognisable in Fig. 3, too; whereas the good scores at higher thresholds can be related to the good forecast of the rainfall peak on Friuli Venezia Giulia. Similarly, QBOLAM underestimation of intense rainfall area extent in the same zone, and false alarm on central Italy can explain the skill score decrease of this model at higher thresholds.

In the 2005 case, the skill score intercomparison (Fig. 6) provides further insight and detail. Overall, results confirm the subjective findings (see Fig. 4) discussed in Sect. 4. A close inspection shows the difficulty of the ECMWF model in resolving the structure of the event, evidenced by the ETS and HK drop at increasing thresholds (Figs. 6b and 6c). Concerning the three LAMs, the same trend is emphasised for ALADIN (even though the model is initialised with ARPEGE): this behaviour reflects clearly the event underestimation visible in Fig. 4b. The good score of the same model at low thresholds can be related with the good detection of the correspondent rainy area shape. On the contrary, the other two models are able to catch the event, but are affected by wet BIA at all thresholds (Fig. 6a). QBOLAM is the only model showing high skill scores on the highest thresholds, despite it has a comparatively worse performance below 30 mm (24h)\(^{-1}\). WRF seems to be also penalised, on the highest threshold, by a larger number of false alarms. The misplacement of the maxima and the false alarms of WRF, the comparatively better forecast of QBOLAM and the excess of heavy rainfall of both models are clearly visible in Fig. 4.

5.2 Displacement error quantification

The object-oriented CRA analysis (Ebert and McBride, 2000) is employed to identify and quantify the horizontal displacement of the forecast precipitation field with respect to the observed one. The method is based on a pattern matching of two contiguous areas (or entities) defined as the precipitation regions delimited by a pre-defined isohyet. For the present application, an isohyet of 0.5 mm (24h)\(^{-1}\) has been selected. This value is usually used for rain/no-rain discrimination.

The displacement error is determined by incrementally translating in the x- and y-directions the forecast rainfall entity over the observed entity until a best-fit criterion is satisfied. For this study, the maximization of the Spearman (linear) correlation has been chosen as the pattern-matching criterion. The total QPF error is then decomposed into three component sources of error – displacement, pattern and volume – by means of the decomposition proposed by Grams et al. (2006, see section 2e), which is suitable to be applied when correlation is chose as pattern-matching criterion.
However, to achieve a reasonable pattern-match between modelled and observed precipitation areas, a quality control test is also performed. A minimum correlation needs to be exceeded for each shift in the x- and y-directions. This correlation value is defined as the minimum correlation statistically different from zero at a desired significance level (in this case, 95%). The value depends also on the effective number of independent comparing samples (Xie and Arkin, 1995). The F test (Panofsky and Brier, 1958) is chosen for assessing whether the shifts in the x- and y-directions are statistically significant.

Table 1 resumes the outcomes of the CRA analysis applied to the 0.1° remapped forecast fields. For the 18 November 2002, results shown that models forecast need to be shifted mainly in the west direction in order to obtain a best match with observations. More in detail, ALADIN shows a dramatic improvement of the correlation after correcting the error location. This is also confirmed by the high displacement error of ALADIN, about 42% of the total error in terms of mean square error (MSE), although pattern error represents the main source of error. For QBOLAM and WRF, the pattern error reaches instead a magnitude of about 80%. For the 9 September 2005, it is worth to note that QBOLAM shows no location error and the pattern error is the quasi-totality of MSE. Also for ALADIN and WRF pattern error plays a major role. However, WRF forecast need to be moved eastward (two grid points) to achieve a best match with the gridded analysis, whereas ALADIN forecast needs instead to be shifted south-eastward.

These results can be interpreted with the help of the previous ones. For the 2002 event, as seen in the subjective analysis (Figs. 3a and 3b), ALADIN large-scale pattern resemble the observed one, so that a westward shift improves the forecast. This is consistent with the shifting of the rainfall pattern discussed in Sect. 4. In the same section, WRF and QBOLAM were found to forecast a rotated, or deformed forecast pattern (Figs. 3c and 3d) so that, for instance, the north-western displacement of forecast pattern needed to realign the secondary maximum over the Apennines would displace the primary maximum away. Under these conditions, the CRA analysis is less powerful. Note also the agreement between low values of volume error and little BIA at low thresholds (Fig. 5a) for all the LAMs.

For the 2005 event, the absence of displacement error in QBOLAM forecast is evident in Fig. 4. The relatively low correlation values can be due to the incorrect prediction of local minima and maxima. In the case of WRF, an eastward shifting of the forecast pattern is evident in Fig. 5d, when the 5 mm (24h)^{-1} threshold is looked at. The wet BIA of the same model (Fig. 6a) is clearly related to the
large volume error (Table 1). Instead, the underestimate of the event by ALADIN is not evidenced in the CRA volume error.
6. Verification of weather system evolution forecast against the METEOSAT-7 Water Vapour images

Results discussed in the previous sections evidence model-dependent shifting and pattern errors, indicating how challenging is these days forecast (e.g., with respect to the 16 November one), probably due to the difficulty in predicting the evolution of the related cut-off low over the Mediterranean. The same results seem to suggest a major impact of initialisation on forecast quality, since displacement errors are much less in the model (ALADIN) initialised 12h after the others.

Once that precipitation verification has been employed to characterise different models’ errors, the errors’ origin should be investigated. This will be the object of future work. Here, qualitative studies employing the FORALPS dataset have been carried out to obtain some preliminary result on the 2002 case, and in particular on day 18 November.

Firstly, ECMWF analyses and forecasts from subsequent daily runs have been compared at fixed times in order to roughly assess the impact of initial conditions on the ECMWF forecast, which drives both QBOLAM and WRF. Surface pressure and 500 hPa geopotential fields have been considered as bulk measures of the weather state. A visual inspection seems to indicate that, until 0000 UTC 17 November, strong qualitative agreement is present between subsequent forecasts but, after that time, the depression over the Tyrrhenian Sea is visible both at the surface and aloft, and its position and location varies among the different forecasts (not shown). However, the quantitative assessment given by plotting difference fields does not provide any evidence that relative differences between subsequent forecasts’ fields increase significantly after that date.

A suitable technique to verify how LAMs predict the weather systems’ evolution even in their mesoscale details is the comparison of the METEOSAT-7 Water Vapour channel imagery with the PWV (Fehlmann et al., 2000) given by the forecast temperature over the 75 mg kg\(^{-1}\) specific humidity isosurface. Dark bands in PWV, correspondent to high potential vorticity anomalies, can be used to track and verify the evolution of the forecast weather system’s key structures (Casaïoli et al., 2006). The comparison is valid only in cloud-free zones. This technique has been applied here to the QBOLAM and WRF runs starting at 1200 UTC 16 November (Figs. 7 and 8).

Looking at the METEOSAT image at 0000 UTC 18 November (Fig. 7a) and the correspondent analysis (Fig. 2c) it is worth to note, along with the main cyclonic system centred on Sardinia (cloudy area in Fig. 7a), a second vorticity centre near the Algerian coast (marked with a cross in Fig. 7a).

Following the latter structure backwards in the METEOSAT images up to 1200 UTC 17 November, its rapid development appears to be associated to convection astride the Algerian coast,
embedded in the orographic cyclogenesis over the Atlas mountain range (not shown). QBOLAM is not able to develop this second vorticity centre in the first 12h of the run. This is evidenced in Fig. 7c: only the main structure (the trough correspondent to the shallow low over Sardinia) is well reproduced, but the absence of the Algerian vortex (whose main features are reported from Fig. 7a as dashed lines) introduces some deformation in the whole structure. Such a deformation is a reasonable candidate for the eastward shifting error of precipitation forecast pattern, discussed in Sect. 4, since it may result in an eastward displacement of the cloudy mass visible over Italy in Fig. 7a. At present, this is only a hint, to be assessed quantitatively in future studies. Note that cloudiness itself prevents to compare PWV and METEOSAT images on central-northern Italy. Concerning WRF, its PWV image (Fig. 7b) is not very informative due to the small domain extent and cloud cover. The only comparable feature, the black band, does not differ from the observed one (Fig. 7a).

The models’ error seems to increase during the following 24 hours. At 0000 UTC 19 November, we can note that in the observation (Fig. 8a) the depression is more elongated than in the QBOLAM forecast (Fig. 8c, see the dashed lines over southern Italy). Again a minor vorticity centre, present in the METEOSAT image (dashed lines on the Genoa gulf) is absent in the forecast. Still, WRF domain is too small to detect these features (Fig. 8b). On the whole, the QBOLAM predicted cyclone seems to be too deep, and the sharp pressure minimum (marked with a cross in Fig. 8c) looks rather unrealistic. On the other hand, WRF seems to exaggerate the curvature of the dark band (Fig. 8c).
7. Conclusion

The combined approach adopted in this work, where quantitative verification methods provide objective findings and qualitative techniques help to interpret them from a physical point of view, allows a complex characterisation of forecast error for three different LAMs on two selected case studies.

Traditional skill-score verification is meant to provide univocal results. For the 2002 event, ALADIN gives the best forecast at lower thresholds (up to 30 mm (24h)$^{-1}$) whereas WRF is the best (even suffering for a wet BIA) at higher thresholds. For the 2005 event, WRF and QBOLAM give the best forecast at low and high thresholds respectively, whereas ALADIN strongly underestimates the event, resulting in poor scores except for the lowest thresholds.

The subjective analysis allows clarifying this picture through a more complex characterisation of the model performance. About the 16-20 November 2002 event, forecasting day 18 November rainfall seems problematic for all the models. ALADIN’s good score seems to be due to the success in prediction the overall rainfall pattern shape, despite the unrealistic small-scale structure and an insufficient prediction of the main rainfall peak. The latter is caught by the other two models, but they are penalised by the occurrence of shifting and distortion in the rainfall pattern. Shifting is present, at a lesser extent, also in ALADIN forecast, and CRA analysis is able to provide a quantitative assessment of such a shifting error (whereas, for the other two models, a simple shifting is unable to match the forecast fields to the observed ones). Subjective analysis for the 2005 case allows relating the ALADIN bad performance to rainfall underestimation, and QBOLAM and WRF pros and cons can be related to the local precipitation features (in particular, only QBOLAM predicts intense precipitation where it has been observed). Still, CRA error decomposition is able to express some of these findings through quantitative values: namely, the shifting of WRF and the good alignment of QBOLAM.

Our case-study approach is meant to demonstrate the capability of a combined verification approach on results from selected operational models (whereas, in order to assess models’ quality, or possible systematic errors, long-term, statistically significant verification is required). Differences between the forecasts cannot be easily attributed to a single feature (e.g., different initialisation, model equation, domain size, or parameterisation schemes), but can be nevertheless characterised in terms of physical meaning. Then, an analysis can be started about “what have gone wrong” in the forecast.

Here, some attempt has been done, by employing qualitative techniques such as the PWV verification. Results indicate that, in some particular cases, forecast quality may be heavily influenced by small-scale errors in the analysis fields or to the misrepresentation of small-scale processes, which
are able to influence the precipitation pattern at a larger scale and on a long distance. In order to confirm and refine these indicative results, thorough studies should be carried on. The evolution of the related weather system in the LAMs should be verified against suitable observations (which special concern with moisture distribution, instability and uplift mechanisms). The role of initial analysis should be assessed too, along with the sensitivity of results to small changes in initial and boundary conditions. The impact of model settings as the domain size, the parameterisation schemes and so on, should be studied, as well as the impact of local factors, as orographic forcing and latent heat release, on shorter and longer time scales. Such a not easy work will be the object of future studies, keeping in mind that the target is not necessarily to find a new technical advance able to improve the models’ skill, but to identify and investigate the weak points in our knowledge of the Mediterranean weather systems, given by three sources: observations, theory and numerical models.

Acknowledgements. This work was developed under the framework of the EU FORALPS INTERREG IIIB Alpine Space project (I/III/3.1/21). Our thanks go to the FORALPS partners and in particular to ARPA of Lombardia, EARS, OSMER and ZAMG for providing us the observed and forecast data used in this study. We want to thank also the ARPA of Emilia Romagna and of Liguria for precipitation data over their territory. The Met Office is acknowledged for the mean sea level pressure analysis charts. Finally we wish to thank Eng. Barbara Lastoria for her help in treating rain gauge data and Prof. Antonio Speranza (University of Camerino, Italy) and Eng. Giuseppina Monacelli (APAT) for the their support.
References


Barnes SL (1973) Mesoscale objective analysis using weighted time-series observations. NOAA Tech. Memo. ERL NSSL-62, National Severe Storms Laboratory, Norman, OK 73069, 60 pp. [NTIS COM-73-10781.]


Gilbert GK (1884) Finley’s tornado predictions. Amer Met J 1: 166–172.


FIGURE CAPTIONS

Fig. 1. Models’ domains: ALADIN (solid black line), QBOLAM (solid grey line), and WRF (dashed black line). The grey shaded indicates the verification area, obtained as intersection of all the models’ domains, that is, between 42.9° N and 48.9° N in latitude and between 8.7° W and 18.4° W in longitude.

Fig. 2. Mean sea level pressure analysis over Europe (Source: The Met Office, UK). a) At 0000 UTC 16 November 2002. b) 0000 UTC 17 November 2002. c) 0000 UTC 18 November 2002. d) 0000 UTC 9 September 2005.

Fig. 3. Contour of precipitation observed (a) and forecast by ALADIN (b), QBOLAM (c) and WRF (d) on 18 November 2002. Forecasts are remapped on the 0.1° common verification grid. In panel (a) the white areas indicate data voids.

Fig. 4. As in Fig. 3, but on 9 September 2005.

Fig. 5. Skill scores for the 2002 event reported as a function of pre-defined thresholds: (a) BIA; (b) ETS; (c) HK and (d) ORSS. Negative values are not reported.

Fig. 6. As in Fig. 5, but for the 2005 event.

Fig. 7. QBOLAM and WRF forecasts comparison with METEOSAT-7 water vapour channel images. Models’ run is initialised at 1200 UTC 17 November 2002. a) METEOSAT image at 0000 UTC 18 November. b) WRF forecast at 0000 UTC 18 November. c) QBOLAM forecast at 0000 UTC 18 November. On panels c), white dashed lines indicate the QBOLAM mean sea level pressure, whereas white solid lines indicate the QBOLAM geopotential height at 500 hPa.

Fig. 8. As in Figure 7, but at 0000 UTC 19 November.

TABLE
<table>
<thead>
<tr>
<th>Date</th>
<th>Model</th>
<th>[E, N] shift</th>
<th>No. of comparing grid points</th>
<th>Corr.</th>
<th>Shifted Corr.</th>
<th>MSE displ. [%]</th>
<th>MSE vol. [%]</th>
<th>MSE patt. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>18 November 2002</td>
<td>ALADIN</td>
<td>[−0.8, −0.1]</td>
<td>477</td>
<td>0.46</td>
<td>0.70</td>
<td>42.2</td>
<td>3.0</td>
<td>54.8</td>
</tr>
<tr>
<td></td>
<td>QBOLAM</td>
<td>[−0.9, 0.1]</td>
<td>545</td>
<td>0.31</td>
<td>0.47</td>
<td>21.4</td>
<td>0.0</td>
<td>78.6</td>
</tr>
<tr>
<td></td>
<td>WRF</td>
<td>[−0.4, −0.2]</td>
<td>545</td>
<td>0.50</td>
<td>0.57</td>
<td>12.6</td>
<td>2.2</td>
<td>85.2</td>
</tr>
<tr>
<td>9 September 2005</td>
<td>ALADIN</td>
<td>[−0.6, −0.2]</td>
<td>418</td>
<td>0.44</td>
<td>0.53</td>
<td>11.9</td>
<td>8.5</td>
<td>79.6</td>
</tr>
<tr>
<td></td>
<td>QBOLAM</td>
<td>[0.0, 0.0]</td>
<td>399</td>
<td>0.48</td>
<td>0.48</td>
<td>0.0</td>
<td>0.5</td>
<td>99.5</td>
</tr>
<tr>
<td></td>
<td>WRF</td>
<td>[0.2, 0.0]</td>
<td>409</td>
<td>0.49</td>
<td>0.56</td>
<td>12.9</td>
<td>0.9</td>
<td>86.2</td>
</tr>
</tbody>
</table>
Figure 5b
Click here to download high resolution image

(b)

ETS

THRESHOLDS (mm 24h⁻¹)

ALADIN - QBOLAM
WRF - ECMWF
Figure 6c

![High resolution image](Click here to download high resolution image)