Adv. Sci. Res., 4, 89–98, 2010
www.adv-sci-res.net/4/89/2010/
doi:10.5194/asr-4-89-2010

Advances in
Science & Research
Open Access Proceedings

9th EMS Annual Meeting and 9th European Conference on Applications of Meteorology 2009

Evaluating multi-scale precipitation forecasts using high resolution analysis

C. Wittmann, T. Haiden, and A. Kann

Central Institute for Meteorology and Geodynamics, Vienna, Austria

Received: 21 December 2009 – Revised: 25 June 2010 – Accepted: 29 June 2010 – Published: 15 July 2010

Abstract. The SAL (Structure, Amplitude, Location) method is used for verification of precipitation forecasts at horizontal grid spacings ranging from 2.5 km to 25 km, using a high-resolution 1 km precipitation analysis as a reference. The verification focuses on a summertime period with predominantly convective precipitation. The verification domain contains lowland as well as alpine areas. Evaluation of the individual SAL components shows that with regard to area mean values (A) the benefit of high resolutions models becomes apparent only in high impact weather situations. For the summertime period studied, the subjective impression of better structured precipitation fields (S) in higher resolution models can generally be confirmed. The most significant improvement appears to be associated with explicit simulation of deep convection.

1 Introduction

An increasing number of operational numerical weather prediction (NWP) models are run with horizontal grid spacings in the 1–5 km range, generating highly structured forecast fields. Proper evaluation of the actual benefit of such forecasts compared to those in the more classical 10–20 km range is not straightforward, especially for precipitation. The use of conventional scores, based on point-to-point comparison of gridded data, or for station locations, penalizes the model that generates higher field variances. Spatial up-scaling to a common reference grid solves the variance problem but masks potential structural skill of the high-resolution model. Similarly, temporal up-scaling such as verification of 24-h totals masks potential timing skill concerning, for example, the diurnal cycle of convection, or frontal passages.

Prompted by these issues, several pattern-oriented verification methods have recently been developed (Casati et al., 2008). Gilleland et al. (2009) provide a comprehensive overview and intercomparison of these methods which they categorize into “filtering” and “displacement” types. The primary purpose of this study is to use the displacement-type method SAL (which stands for structure, amplitude, and location) developed by Wernli et al. (2008), to evaluate precipitation forecasts during the convective season over a wide range of model resolutions against a high-resolution precipitation analysis. Previously, raingauge data interpolated to a 5 km grid (Früh et al., 2007) and a 10 km grid (Wernli et al., 2008), or radar rainfall data were used, the latter however only for a 9-day period (Wernli et al., 2009). Also, we compare NWP models over a broader range of resolutions than has been done in the past, with grid spacings ranging from 2.5 to 25 km. Forecasts of 3-hourly accumulated precipitation are considered over a 2 month period in the summer of 2009, for different domains covering lowland and mountain areas.

Section 2 describes the high resolution precipitation analysis of the INCA (Integrated Nowcasting through Comprehensive Analysis) system, and is followed by an overview of the different NWP models and a summary of observed precipitation characteristics during the chosen period in Sect. 3. The verification method SAL is briefly discussed in Sect. 4, and results and conclusions are presented in Sects. 5 and 6, respectively.

2 INCA precipitation analysis

The INCA precipitation analysis is a combination of station data interpolation, including elevation effects, and radar data. It is designed to combine the strengths of both observation types, the accuracy of the point measurements and the spatial structure of the radar field. In the following, the analysis method for 15-min precipitation amounts is briefly described. A detailed description of the entire analysis and nowcasting system is given in Haiden et al. (2009).

Correspondence to: C. Wittmann
(christoph.wittmann@zamg.ac.at)

Published by Copernicus Publications.

2.1 Interpolation of station data

The irregularly distributed point precipitation values measured at the station locations are interpolated onto the regular 1×1 km INCA grid using inverse-distance-squared weighting (IDW). To reduce the occurrence of bulls-eyes, only the nearest 8 stations are taken into account in the interpolation. Furthermore, a modification to the classical IDW method has been introduced. It takes into account the inhomogeneous azimuthal distribution of stations around the grid point in question and reduces the relative weight of stations located “behind” (i.e. at a similar azimuth to) nearer stations. The resulting field is denoted by $P_{\mathrm{STAT}}(i,j)$.

2.2 Climatological scaling of radar data

The radar data, which are available at 5 min intervals, are aggregated in time and bilinearly interpolated onto the INCA grid. Since the radar field is strongly range-dependent and contains biases due to topographic shielding it must be scaled before use in a precipitation analysis. In a first step, a “climatological” scaling is performed. Operationally, a 3-month temporal average of scaling factors RFC (i,j) centered at the actual month is used. This climatological scaling only partially corrects the reduced radar return in the case of snowfall as compared to rain.

In order to compensate for some of the artifacts in the radar field caused by topographic shielding of the radar beam, the interpolated scaling factor is replaced by a local scaling factor in regions where the radar beam is strongly shielded (indicated by beamlike structures with high local scaling factors). The local scaling factor is the ratio of the monthly accumulated precipitation gained from the interpolated station observations to the accumulated radar precipitation at the respective gridpoint.

2.3 Re-scaling of radar data using the latest observations

In a next step the climatologically scaled radar field is rescaled on the basis of a comparison at analysis time of station observations and radar values at the stations. In this comparison, a spatial shift of a maximum of 4 km in either direction between the station and the corresponding radar pixel is allowed to take into account effects due to the finite settling time of hydrometeors, effects of wind-drift, etc. The scaling for a gridpoint is a weighted average of the ratio between station and radar precipitation at the nearest stations, where the weight decreases with increasing distance, with increasing difference in climatological scaling, and with decreasing precipitation at the station (relative to the precipitation at the gridpoint). The resulting field is denoted by $P^{**}_{\mathrm{RADAR}}(i,j)$ (cf. Haiden et al., 2009).

2.4 Final combination

The two precipitation fields $P_{\mathrm{STAT}}(i,j)$ and $P^{**}_{\mathrm{RADAR}}(i,j)$ are finally combined to a field $P_{\mathrm{INCA}}(i,j)$ that gives a better estimate of the precipitation distribution than each individual field. The combination is obtained through a weighting relationship

$$P_{\mathrm{INCA}}(i,j) = P_{\mathrm{STAT}}(i,j) + v\left[P^{**}_{\mathrm{RADAR}}(i,j) - P^{**}_{\mathrm{RADSTAT}}(i,j)\right], \quad (1)$$

where the weight v is given by

$$v(i,j) = \begin{cases} 1 & \mathrm{RFC} < \mathrm{RFC}_0 \\ \exp\left[-\ln(2)\left(\frac{\mathrm{RFC}-\mathrm{RFC}_0}{\mathrm{RFC}_\mathrm{H}-\mathrm{RFC}_0}\right)^2\right] & \mathrm{RFC} \geq \mathrm{RFC}_0 \end{cases} .(2)$$

In Eq. (2), RFC is the spatially and seasonally varying climatological scaling factor described in Sect. 2.2. The auxiliary field $P^{**}_{\mathrm{RADSTAT}}(i,j)$ is created by interpolating onto the grid, analogous to the station observations, the scaled radar values at the station locations. The threshold value RFC_0, above which the weight of the radar begins to decrease, is 3. The value of RFC_H, at which the radar weight has decreased to one-half, is 5.

Figure 1 shows an example of the stepwise procedure and final analysis from the June 2009 flood event in Austria. Note the large difference between un-scaled radar and station interpolation (top panels), and the importance of the final combination (lower right panel) as a means to smoothly connect areas seen by radar with those only covered by stations. If the final combination step is not made, “edges” remain in the analysis (lower left panel).

The INCA precipitation analysis reproduces (within limits imposed by the grid spacing) the observed values at the raingauge locations. The quality of the analysis as determined by cross-validation based on single-station denial experiments, is superior to raingauge-only IDW interpolation. As expected, the benefit of incorporating radar data is largest for convective precipitation, where the reduction of mean absolute error (MAE) compared to IDW is 20–50%. For stratiform precipitation, the reduction of MAE is typically 5–10%. A paper describing crossvalidation results of the INCA analysis scheme has been submitted. A parameterization of the elevation dependence of precipitation, which was introduced to the system in 2007, is described in Haiden and Pistotnik (2009).

Any precipitation analysis contains measurement errors, sampling and representativeness errors, and uncertainties due to assumptions made in the analysis scheme. Following Pappenberger et al. (2009) we roughly estimate the raingauge measurement errors (undercatch) to be of the order of 10% for lowland and valley stations, and 20% for some of the more exposed mountain stations. Note that we are dealing with cases of rainfall only. According to Skok and Vrhovec (2006), errors of up to 50% for point values can be introduced by the analysis itself. However, the SAL method is equivalent to using spatially aggregated values, which reduces the

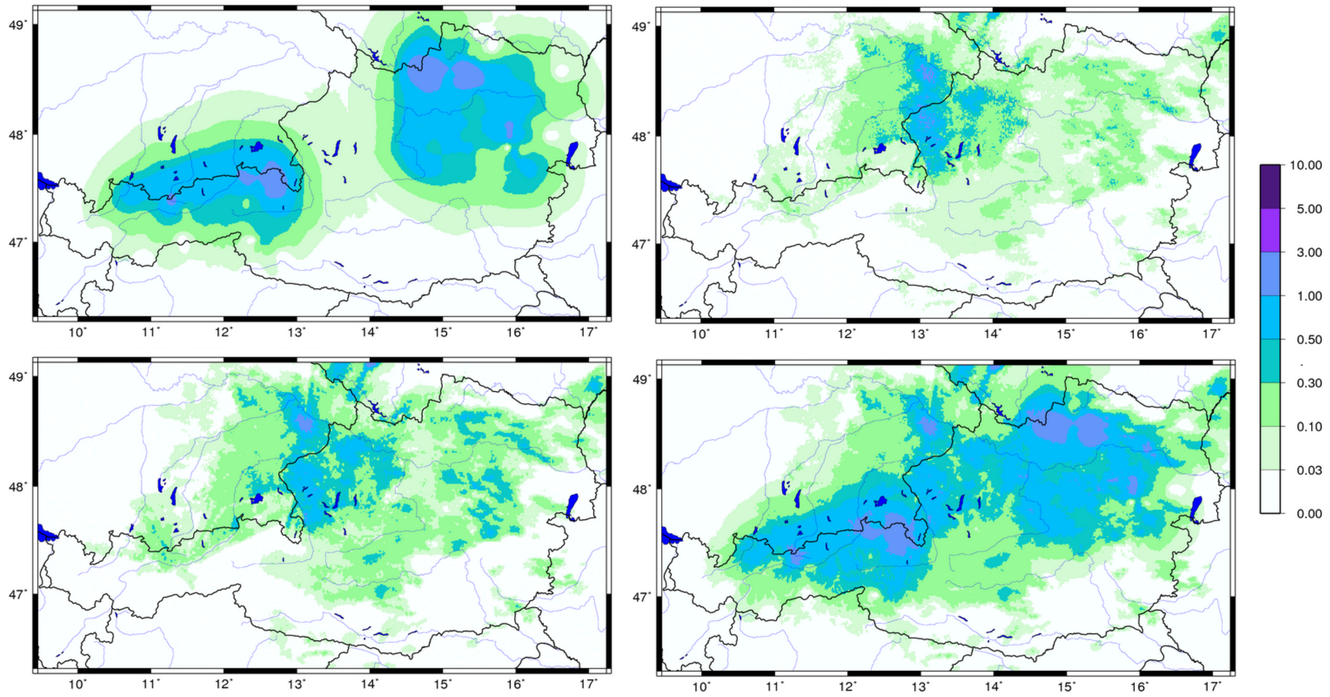

Figure 1. Example of a 15-min INCA precipitation analysis (23 Jun 2009, 00:15–00:30 Z) during a flooding event in Central Europe. Upper left panel: pure station interpolation, upper right panel: uncorrected radar field, bottom left panel: corrected radar field, bottom right panel: final INCA precipitation analysis. Units are mm, with color scale on the right.

error. For a comprehensive analysis and discussion of the uncertainty issue we refer to Pappenberger et al. (2009) and references therein.

3 Model data

The operational forecast model used at ZAMG is ALADIN (Aire Limitée Adaption Dynamique Developpement International), a spectral Limited Area Model (LAM) which is being developed within the ALADIN consortium, a cooperation of several mainly European national weather services. ALADIN is part of a software library consisting of several models, which allows to choose among various physical schemes and parameterizations. The operational ALADIN version implemented at ZAMG uses the ALARO physics package. It has been developed to address problems in the difficult grid spacing range between 3 and 7 km, where deep convection starts to be at least partly resolved. The ALARO package comprises a prognostic large-scale cloud and precipitation scheme and prognostic convection scheme named 3MT, which stands for Modular Multi-scale Microphysics and Transport. Detailed information about ALARO and 3MT can be found in Gerard et al. (2009).

Apart from the operational ALADIN model (hereafter referred to as ALA-AUT), several other model configurations are used in this verification study. Table 1 gives an overview of the main characteristics of the different models: a 5 km hydrostatic ALADIN version (ALA5) using ALARO physics; a 5 km non-hydrostatic ALADIN version (ALA5-NH); a 6.5 km ALADIN version (ALA-EUR) using ALARO physics; and the 2.5 km AROME (Applications of Research to Operations at Mesoscale) model. These models are used at ZAMG in a pre-operational environment. AROME is a new model which more or less represents the merger of the physical package from the research model Méso-NH (Lafore et al., 1998) and the non-hydrostatic version of ALADIN. The microphysical scheme used in AROME is ICE3 (Pinty and Jabouille, 1998).

In addition to the models listed in Table 1, two more models are considered: The deterministic global model of the European Center for Medium-Range Weather Forecasts (ECMWF, 25 km horizontal resolution) and INCA precipitation forecasts. INCA forecasts are computed on a two dimensional grid with 1 km resolution and start with kinematic extrapolation of analyzed precipitation fields (as described in the previous section) in the nowcasting range (up to +2 h). Outside the nowcasting range the extrapolated fields smoothly merge into a combination of ALA-AUT and ECMWF precipitation forecasts. The combination of these two models uses weights computed on the basis of several years of archived precipitation forecasts. More details about the INCA precipitation forecast are given by Haiden et al. (2009).

Table 1. Main characteristics of the different limited area models used at ZAMG.

	ALA-AUT	AROME	ALA5-NH	ALA5	ALA-EUR
Timestep [s]	415	60	207	207	285
Coupling model	ARPEGE	ALA-AUT	ARPEGE	ARPEGE	ECMWF
Initialization	ARPEGE	ALA-AUT	ARPEGE	ARPEGE	ECMWF
Resolution [km]	9.6	2.5	4.9	4.9	6.5
Grid size	300×270	432×320	540×512	540×512	720×810
Levels	60	60	59	59	45
Forecast Range [h]	72	30	48	48	72
Convection parameterization	Yes (3MT)	No	Yes (3MT)	Yes (3MT)	Yes (3MT)
Microphysics	ALARO	ICE3	ALARO	ALARO	ALARO
Kernel H/NH	hydrostatic	non-hydrostatic	non-hydrostatic	hydrostatic	hydrostatic

The main objective of this study is to evaluate model performance with respect to convective precipitation, so June and July 2009 were chosen as verification period. The number of observed days with precipitation was 76–91% out of the total of 61 days, with higher values found in the more mountainous domains. The corresponding numbers for the forecasts are 87–100%. Only 00:00 UTC runs are considered. This is true for all models except for INCA where the 06:00 UTC runs are taken into account. This is done to allow a comparison of all models available to the forecaster around or shortly after 06:00 UTC in the morning. The common forecast range is defined by the model with the shortest integration time, which in our case is AROME, so all models are compared up to forecast time +30 h, using 3 h accumulated precipitation intervals.

In order to quantify the proportion of days with convective precipitation with respect to the total number of days with precipitation for the chosen period, lightning data was used. Assuming that a day can be classified as "convective" when more than two (cloud-to-ground) lightening strokes are detected within the considered area, the percentage for convective precipitation days ranges from 76–82% with higher values again found in the mountainous domains. The reason for choosing two lightning strokes as a threshold is to reduce erroneous detections. This type of characterization does not distinguish between days with convective cells growing and decaying within the considered area and convective systems moving into and/or over the domain (often related to large scale systems). A subjective, semi-quantitative estimation of this partitioning is based on visual inspection of INCA analyses of 24-h accumulated precipitation. Out of the total number of days with precipitation, the percentage of days with local convection is about 40%, the percentage of days with large scale systems moving over the considered domains is about 20%. The remaining days can not be clearly assigned, showing a mixture of both types. This is valid for the mountainous areas. For the lowlands the proportion of days with local convection is smaller (about 30%), whereas the days with advective convection occur with similar frequency.

4 Verification method

In order to allow a fair comparison of the precipitation forecasts coming from models running on different horizontal (and vertical) resolutions it was decided to use the verification package SAL, which stands for Structure, Amplitude and Location. In the following, a short description of the main components of SAL is given. A comprehensive description can be found in Wernli et al. (2008).

SAL is an object-based method which allows to evaluate QPFs on a given geographic domain according to three criteria: structure (S), amplitude (A) and location (L). A is a measure of the deviation of the areal mean QPF relative to the observed value. S gives information on whether the precipitation objects created by the model correspond to the observed objects in terms of size and shape. Finally L yields information about the displacement of the precipitation objects with respect to the observed objects.

To calculate the S and L components it is necessary to identify individual precipitation objects in the forecast and the observation fields. In SAL this is done by using a threshold $R^* = fR_{\max}$, where $R_{\max}$ denotes the maximum precipitation amount found in the domain. Starting from a local precipitation maximum (provided it is larger than R^*), all surrounding grid points with values $R_{ij} > R^*$ are considered to be part of the actual precipitation object. Following Wernli et al. (2008) the empirical factor f is set to 1/15. According to their sensitivity analysis, the S and L results are not particularly sensitive to the exact value of this factor. In cases where the threshold R^* happens to be close to the minimum value at the "saddle" between two maxima, the object definition becomes sensitive to f, but as these cases comprise only a small fraction of the sample their effect on the overall statistics will be small.

4.1 Amplitude *A*

The amplitude component A is defined as the normalized difference of the forecasted and observed domain average precipitation ($D(R_{\text{mod}})$ and $D(R_{\text{obs}})$). Thus, A measures the quality of the total precipitation amount simulated by the model. This normalized difference is computed according to:

$$A = \frac{D(R_{\text{mod}}) - D(R_{\text{obs}})}{0.5(D(R_{\text{mod}}) + D(R_{\text{obs}}))}. \qquad (3)$$

Values of A range from −2 to +2. Positive values indicate an overestimation of the domain-averaged precipitation coming from the model, negative values an underestimation, and 0 stands for a perfect forecast.

4.2 Location *L*

The location component L is the sum of two components, named L_1 and L_2. The first part L_1 measures the normalized distance between the centers of mass of the modeled and the observed precipitation field ($\boldsymbol{x}(R_{\text{mod}})$ and $\boldsymbol{x}(R_{\text{obs}})$)

$$L_1 = \frac{|\boldsymbol{x}(R_{\text{mod}}) - \boldsymbol{x}(R_{\text{obs}})|}{d}, \qquad (4)$$

where d stands for the maximum distance found in the given domain between two boundary points. Values for L_1 range from 0 to 1, where for $L_1 = 0$ the centers of mass for model and observed precipitation fields are equal.

The second part L_2, ranging from 0 to 1, can be described as the weighted mean distance between the centers of mass and the individual precipitation objects ($r(R_{\text{mod}})$ and $r(R_{\text{obs}})$ and is computed according to

$$L_2 = 2\frac{|r(R_{\text{mod}}) - r(R_{\text{obs}})|}{d}. \qquad (5)$$

The total location component $L = L_1 + L_2$ can assume values between 0 to 2, where 0 would only be obtained when the total mass centers *and* the averaged distance between individual objects and the total mass center are identical in the observed and in the model field.

4.3 Structure *S*

The structure component S is defined in such a way as to give information about the quality of the forecast with respect to size and shape of the precipitation objects. This is done by calculating a scaled volume V_n for every object R_n. Finally a weighted average Volume V is calculated separately for the forecast and observational fields ($V(R_{\text{mod}})$ and $V(R_{\text{obs}})$). The final score for structure is then calculated through

$$S = 2\frac{V(R_{\text{mod}}) - V(R_{\text{obs}})}{V(R_{\text{mod}}) + V(R_{\text{obs}})} \quad . \qquad (6)$$

S again takes a continuum of values from −2 to +2. Positive values mean that the predicted forecast objects are too large or widespread with respect to the observed objects. Negative values for S stand for predicted objects being too small or peaked.

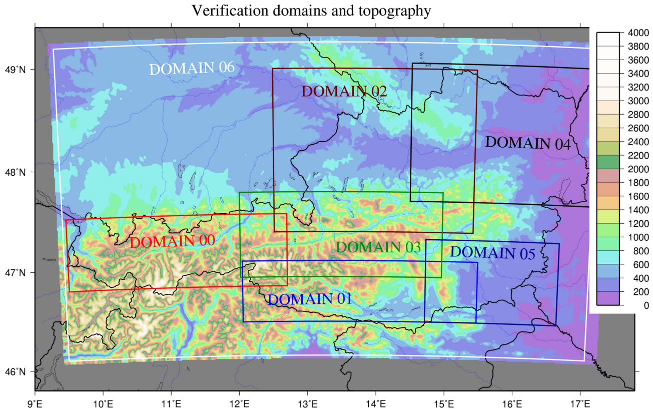

Figure 2. Verification domains numbered 00 through 06, and topography. Height in meters with color scale on the right.

4.4 Verification domains

The SAL verification is performed for several rectangular domains over Austria covering either Alpine areas or flatland areas. Figure 2 shows the different domains together with the topography. In order to make the various forecasts comparable and usable for the SAL software package, the precipitation fields are interpolated to the same grid, which is the 1 km INCA grid in this case. To discuss the results in Sect. 5 two representative verification domains have been chosen: domain 00, representing an Alpine area in the western part of Austria, and domain 04, an area consisting of mainly rather flat terrain.

5 Results

The SAL verification has been performed for the period 20090601–20090731 based on 00:00 UTC model runs (and 06:00 UTC runs for INCA) up to a forecast time of +30 h, considering 3-h accumulations. In the following subsections the results are discussed with regard to each of the individual SAL components. In Sect. 5.4 some verification results using more classical grid-point based scores instead of the SAL components are briefly discussed.

Following the WMO recommendations (WMO, 2009) regarding statistical significance tests for verification results, a bootstrap method is used because the presented scores (and the underlying precipitation fields) do not allow any assumptions concerning theoretical distributions. Further, since we are dealing with fields which are autocorrelated in time, a blocked bootstrap method was chosen. Whenever "significant" or "not significant" is mentioned in the following subsections it is meant in its statistical sense based on a blocked bootstrap and hypothesis test as described in Wilks (1997) using the following test parameters: significance level 0.95; block length $l = 7$; bootstrap sample size $n_b = 5000$. In addition to the bootstrap test a Wilcoxon-Mann-Whitney rank-sum test was performed (Wilks, 2006).

5.1 Area mean precipitation

The top panel in Fig. 3 shows the mean amplitude score A for each model for the entire period for an Alpine area in the western part of Austria as a function of lead time, so for e.g. lead time +6 h the values in the plot represent the mean amplitude score averaged over 61 forecasts of the 3-hourly precipitation between +3 and +6 h. It is worth noting that all models, including the one with explicit deep convection (AROME), tend to overestimate convective activity in the afternoon. The amplitude score reaches maximum values (the strongest overestimation of area mean precipitation) between 12:00–15:00 UTC. Thus, even if the higher resolution models turn out to have better structural scores, this does not by itself solve the problem of overestimation of alpine afternoon convection. This corresponds to the subjective impression gained from visual inspection of the precipitation forecast fields. All models tend to produce an "envelope" of precipitation over the entire alpine area, whereas in the real atmosphere the precipitation is more selective and clustered into specific regions. The higher resolution models generate more fine-scale structure but still exhibit the envelope characteristics.

Comparing the different models we must keep in mind that the INCA forecast has the advantage of starting by design with values near 0, since in the nowcasting range (up to 4 h) it is an extrapolation of the analysis which is used for verification. For higher lead times the INCA curve is sometimes further from the zero-line than one of its constituent models ECMWF and ALA-AUT. As INCA should represent an optimal combination of ECMWF and ALA-AUT it is concluded that the weights for the combination of these two models are "out-of-date". Several changes have been made to the operational ALA-AUT model and also in the ECMWF model, changing the characteristics of the precipitation forecasts significantly. At the time of writing the weights for the combination have already been updated. It is interesting to see the ECMWF model in Fig. 3 (top panel, yellow curve) as the one showing the fastest decrease of overestimation during late afternoon and evening, whereas the values even become slightly negative (underestimation) in the evening and during night. In contrast to ECMWF, AROME is the model showing the strongest overestimation during the afternoon and evening.

Figure 4 shows the scores for an area in the north eastern part of Austria, representing a predominately flat terrain. The scores exhibit a similar diurnal cycle to the case of the Alpine area, but there are noticeable differences. Most of the models start with an underestimation of the area mean precipitation, which is turning into an overestimation before midday. It is obvious that compared to the results for the Alpine area the phase of the diurnal cycle in terms of the amplitude score is shifted towards shorter lead times. Furthermore it seems that this shift depends on the resolution of the model, that is to say the ECMWF error characteristic is the first to change

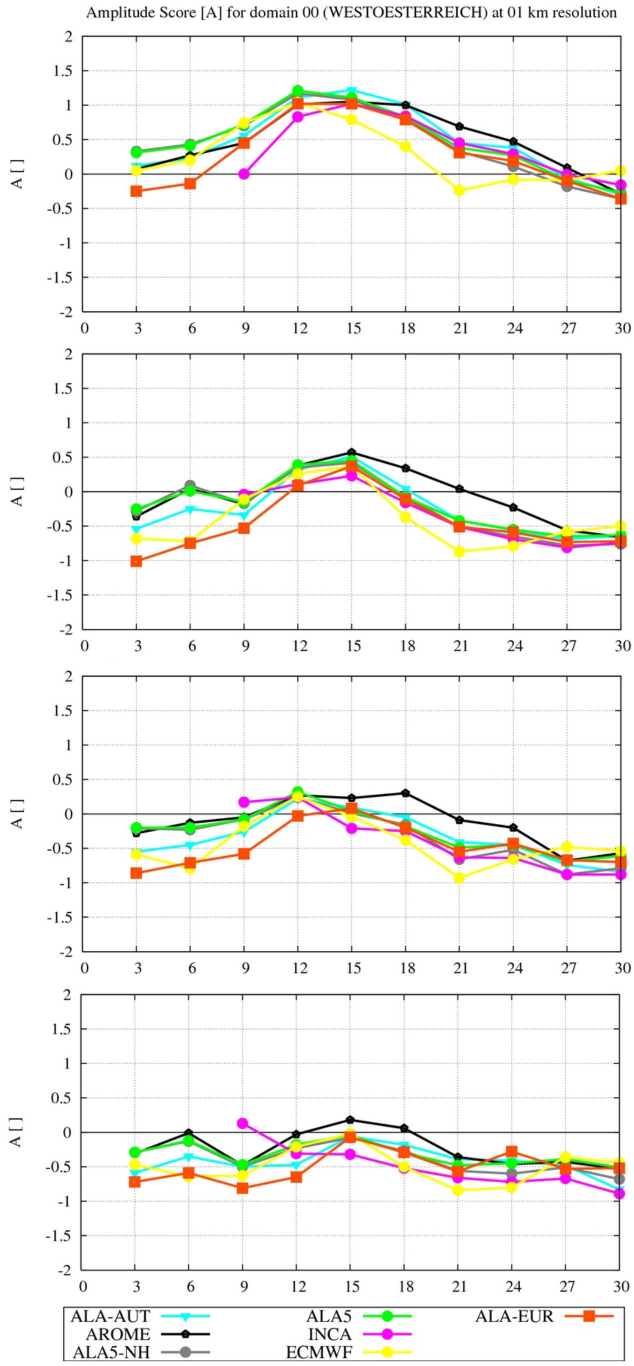

Figure 3. Mean values for SAL amplitude component A as a function of lead time for various models for different domain average thresholds: (from top to bottom) all events with an observed domain average greater than 0 mm, 0.5 mm, 1.0 mm, 2.5 mm (verification period 20090601–20090731).

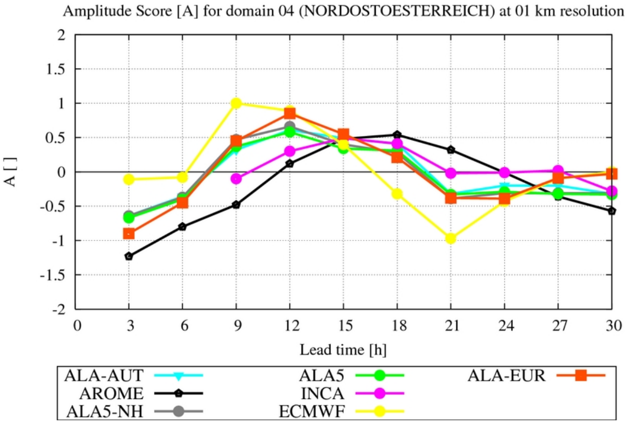

Figure 4. Mean values for SAL amplitude component A as a function of lead time for different models (verification period 20090601–20090731).

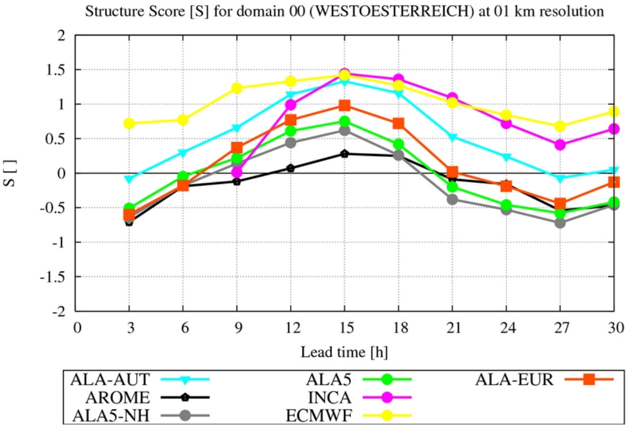

Figure 5. Mean values for SAL structure component S as a function of lead time for different models (verification period 20090601–20090731).

from underestimation to overestimation and to reach a maximum before midday. For AROME this is occurring later in the day. As we are dealing mainly with convective precipitation events, the interpretation can be as follows. As the models with higher resolution generally are better capable of simulating the evolution of convective cells, they show higher skills in predicting the spread of convective activity from the mountainous areas to the flatland areas in Austria. But this positive aspect is reduced by the significant negative amplitude scores in the morning, which should be a topic for further investigation.

Two points should be further mentioned. First, from comparing the hydrostatic ALA5 with its non-hydrostatic counterpart ALA5-NH it is evident that in terms of mean scores over a two month period the differences between the two versions are not significant. It is hardly possible to distinguish the curves for ALA5 and ALA5-NH in Figs. 3 and 4. This does not necessarily mean that two versions always produce the same forecasts, but it would need some case studies to point out the more significant differences. The second point is that the INCA precipitation forecast shows consistently higher skill than each of its constituents ALA-AUT and ECMWF from +18 h onwards, and partly also at shorter lead times. Hence, in contrast to the Alpine area the weights for the combination of ALA-AUT and ECMWF over the lowlands still seem to work reasonably well. Forecasting experience has shown that the change of precipitation characteristics of ALA-AUT due to the use of new physics packages was most pronounced in mountainous areas.

Summarizing the results seen from Figs. 3 and 4 one may not see great benefit in running a high resolution model like AROME or ALA5/ALA5-NH. But it is important to point out that the scores shown in the top panel of Fig. 3 represent mean values over a 2 month period without taking into account the intensity of the events. According to Eq. (2) the amplitude score reaches values of 2 for all cases with $D(R_{\mathrm{mod}}) > D(R_{\mathrm{obs}}) = 0$, i.e. when there is precipitation in the model but not in reality. So even when the model domain average is very small, the score is 2 if the observed domain average is 0.

In order to take this important fact into account the scores were recomputed for different domain average thresholds, i.e. different precipitation intensities. Figure 3 also shows the scores separated for events with observed domain averages greater than 0.0 mm, 0.5 mm, 1.0 mm and 2.5 mm respectively (top to bottom). Comparing the curves it can be seen that the higher the intensity of the event gets, the closer the curves for high resolution models (like AROME, ALA5 and ALA5-NH) get to 0 (closer to the observed domain average). This indicates that the benefit of the high resolution models reveals itself for strong precipitation events, i.e. for high impact weather. For events with domain averages greater than 2.5 mm (bottom panel in Fig. 3) it can be seen that the models generally underestimate the amplitude, with AROME showing the best overall A score. During the convectively most active period (12–18 Z) AROME differs from the other models in that it slightly overestimates the amplitude. One needs to keep in mind, however, that the number of considered cases decreases with increasing intensities. With respect to the total number of 61 days the proportions of days with domain averages greater than 0.0 mm, 0.5 mm, 1.0 mm and 2.5 mm are 88%, 37%, 32% and 15%, respectively. This naturally has an effect on the statistical significance of the results, but we expect that the main characteristics of the results do not change when extending the verification period.

5.2 Structure and location of precipitation fields

The interpretation of the results addressing the structure of the precipitation forecasts is more straightforward. Figure 5 shows the scores for the structure component for the different models, again as a function of lead time. It can be clearly seen that the higher the resolution of the model, the better the structure of the precipitation forecasts gets in terms of the S component. The structure component is largest for the model with the lowest resolution (ECMWF), implying that predicted precipitation patterns are too big or flat compared to the observed fields. The higher-resolution models ALA-EUR, ALA5, and ALA5-NH show generally significant smaller S values during daytime but apparently underestimate the up-scale development of convective systems in the evening and night hours, as S values become significantly negative there. Averaged over the day, AROME has the smallest S values. Comparing the non-hydrostatic and the hydrostatic 5 km ALADIN version it turns out that the non-hydrostatic version tends to produce smaller or more intense features, as the mean structure scores are noticeable smaller. This is valid for Alpine areas, while the difference is much smaller for areas located in flat terrain (not shown). So in general the benefit of using a high resolution model can be clearly shown with the structure component of SAL.

The results for the location component L do not yield much information when computing mean values for a 2 month verification period, as the values for the various models do not differ significantly. The mean L values (over all lead times) for the different models range from 0.29 to 0.36 for domain 00 and from 0.29 to 0.33 for domain 04. An evolution of L with respect to the lead time is hardly noticeable. Two aspects are worth mentioning. First, the advantage of INCA in the nowcasting range is also visible in the L component. Second, if one had to choose, without considering statistical significance, one model showing the overall lowest L scores it would be ECMWF, in particular for the mountainous domain 00. Taking a look at some case studies while keeping in mind that the L component is constructed as the sum of two components, one gets the impression that the mass centers for the different models are often rather similar. This means that the differences in terms of the L value arise mainly because of the second component of L.

5.3 Sensitivity of results to grid resolution

Up to now the results were discussed for verification domains with 1 km horizontal resolution. To obtain information about the sensitivity of the SAL results to the resolution of the underlying verification grid, the scores were re-computed for horizontal resolutions of 5 km and 10 km. The up-scaling of the analysis and model fields was performed by thinning the initially created 1 km fields using median values for the final 5 km and 10 km resolution grid, respectively. The median was chosen as it is less sensitive to extreme values than the mean.

Comparing the A scores for the different resolutions for domain 00 (not shown) indicates that the curves for the high-resolution models (AROME, ALA5-NH and ALA5) are slightly shifted towards lower A values for all lead times (0.1 in terms of the mean A value over all lead times) when going to 5 km, and further to 10 km resolution, respectively. For ALA-EUR, ALA-AUT, ECMWF and INCA the differences are smaller, in general even a slight shift towards higher A values may be observed. So it seems that the high resolution model fields lose at least a part of their overestimation characteristics when upscaled to lower resolutions. For domain 04 there is a noticeable shift towards higher values of A, this time for all models. The magnitude of this shift is larger (up to 0.3 for ECMWF) for the lowlands than for the mountain areas. The sign of the change is opposite for the high resolution models for the mountainous and the lowland domain. So for the lowlands the upscaling appears to be less beneficial.

The results for the structure component S change in the way that the curves for the different models tend to be located closer together when moving to lower grid resolution. In general the S values decrease for ALA-AUT and ECMWF and increase for the other models. This implies that ALA-AUT and ECMWF precipitation fields gain structural skill with respect to the verifying analysis, while the high resolution model forecasts indeed loose structure in their fields but still retain a better correspondence to the observed field in terms of volume and shape of the precipitation objects. For example, the difference of the mean S value between AROME and ECMWF over all lead times reduces by 0.35 when moving to 10 km grid resolution. But even at 10 km resolution, the S values are significantly lower (and closer to 0) for the high resolution models with respect to ALA-AUT and ECMWF. It may be interesting to extend the verification to even lower resolutions (e.g. the one of ECMWF), but for that it would be necessary to increase the size of the verification domains as the remaining number of grid points within domain 00 and 04 gets too low so that the object identification algorithm within SAL might not work properly anymore.

Finally, when comparing the L component for the different verification grid resolutions, it can be seen that again a shift of the curves is noticeable. The slight but noticeable shift happens towards lower L values. This in turn implies that there is a tendency of the location of the center of mass and the mean distance of individual objects with respect to the center of mass to coincide better when upscaling the analysis and model fields. But as mentioned in the previous section, the differences in terms of L values do not differ greatly between the models.

5.4 Grid point verification

In order to evaluate the precipitation forecasts in a more classical way, various grid point scores were computed in addition to the SAL scores: Probability of Detection (POD), False Alarm Ratio (FAR) and Equitable Threat Score (ETS). Definitions and properties of these scores are given in Wilks (2006). For the computation of ETS random chance was chosen as reference forecasts.

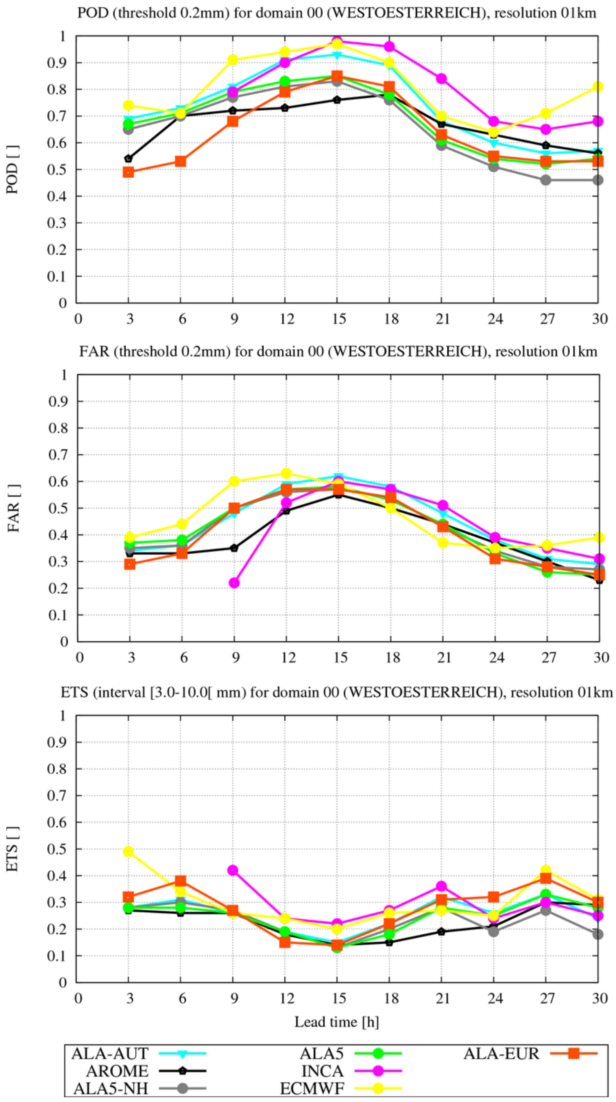

Figure 6. FAR, POD and ETS for different models (verification period 20090601–2009731).

Figure 6 shows POD (top), FAR (middle) and ETS (bottom) for the different models as a function of lead time for the mountainous domain 00, again for verification period 20090601-20090731 on a common grid with 1 km horizontal resolution. POD and FAR presented in Fig. 6 just include grid points with precipitation intensities over 0.2 mm, the ETS scores are shown for the intensity interval 3.0–10.0 mm. The diurnal cycle which could be observed in the SAL scores in Figs. 3 and 4 is also visible for the three scores presented in Fig. 6, whereas in contrast to FAR and ETS, POD shows higher values during the period with strongest convective activity. This is consistent with the result for the *A* component in Fig. 3, as a stronger overestimation (in terms of areal means) implies an increase of POD but at the same time also an increase of FAR. Taking a closer look on the curve for POD, one can notice that there is a tendency towards lower POD values for the high resolution models. The difference in terms of POD is around 0.2 during afternoon. This is even better visible in the scores for the lowland domain 04 (not shown), where the difference in terms of POD increases to 0.3. For both domains it is the model running on highest resolution (AROME) showing the lowest POD values for the period 06:00–18:00 UTC, but at the same time the lowest FAR values. This statement is true when taking ALA-EUR and INCA aside for a moment. As the model physics used inside ALA-EUR is the same as it is the case for ALA-AUT and ALA5, the significant difference in the first 12 forecast hours should be related to the initial conditions (ECMWF vs. ARPEGE). This will be a topic for further investigation, as in general the creation of initial conditions for ALADIN/ALARO/AROME models from ECMWF model output is not straightforward, in particular for surface fields like temperature and humidity. Difficulties arise from the differences in the surfaces schemes used in the ECMWF and the ALADIN/ALARO/AROME/ARPEGE models.

In general INCA shows the best scores for ETS and FAR during the first 3 h, as would be expected from a nowcasting system. It must be kept in mind, however, that INCA has the advantage of starting from the analysis which is used for verification. For FAR and ETS the difference between lowlands and mountainous areas is less obvious. In terms of ETS it is ECMWF and INCA showing the highest values during the period with strongest convective activity. When using different intensity intervals to compute the ETS score (e.g. 10–20 mm, > 20 mm) no additional information is obtained, except that in terms of ETS the models show low skill for the forecast period with strongest convective activity.

6 Summary

Various models used at ZAMG running on different horizontal grid spacings ranging from 2.5 to 25 km are compared according to their precipitation forecasts for a 2-month period in summer 2009 which is dominated by convective activity. The method used for verification is SAL, a package evaluating the domain average precipitation and also the structure and the location of the precipitation features in the model. On observational side the INCA high resolution precipitation analysis is used, which is a part of the operational analysis and nowcasting system at ZAMG.

It turns out that SAL is a valuable tool to address more aspects than just the quantitative quality of precipitation forecasts. In addition to the *A* component (evaluating the area mean values), the structure component *S* seems to be able to confirm “human” subjective impressions when there is a need to choose the model creating the most realistic precipitation forecast patterns. The location component seems to yield less valuable information due to its rather small variation from model to model (in terms of mean values). *L* should be more valuable when considering single case studies.

The evaluation of the different models shows that the benefit of high resolution models running on grid spacings of 2.5 km or 5 km can be seen when the structure of the model precipitation fields are considered. This is true even when verification grids with lower horizontal resolution are used. In terms of domain average values the benefit is not apparent, unless one concentrates on strong precipitation events. For high impact weather the advantages of models with grid spacing finer than 5 km are visible. Using conventional grid point scores the benefit of high resolution models is not visible.

Among the models used in the present verification study two 5 km model versions are compared, one running with a hydrostatic kernel and one with its non-hydrostatic counterpart. The results show that noticeable differences between the two versions are just visible in the structure component and primarily in mountain areas. In terms of areal mean values the differences are small, but one can expect to reveal more details by performing case studies.

Edited by: D. Giaiotti
Reviewed by: S. Niemela, D. Ahijevych, and
three other anonymous referees

References

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.

Früh, B., Bendix, J., Nauss, T., Paulat, M., Pfeiffer, A., Schipper, J. W., Thies, B., and Wernli, H.: Verification of precipitation from regional climate simulations and remote-sensing observations with respect to ground-based observations in the upper Danube catchment, Meteorol. Z., 16, 275–293, 2007.

Gerard, L., Piriou J.-M., Brožkova, R., Geleyn, J.-F., and Banciu, D.: Cloud and precipitation parameterization in a meso-gamma scale operational weather rediction model, Mon. Weather Rev., 137, 3960–3977, 2009.

Gilleland, E., Ahijevych, D., Brown, B.G., Casati, B., and Ebert, E. E.: Intercomparison of spatial forecast verification methods, Wea. Forecasting, 24, 1416–1430, 2009.

Haiden, T. and Pistotnik, G.: Intensity-dependent parameterization of elevation effects in precipitation analysis, Adv. Geosci., 20, 33–38, doi:10.5194/adgeo-20-33-2009, 2009.

Haiden, T., Kann A., Pistotnik, G., Stadlbacher, K., and Wittmann, C.: Integrated Nowcasting through Comprehensive Analysis (INCA), System description, available at: www.zamg.ac.at/fix/INCA_system.pdf, 2009.

Lafore, J. P., Stein, J., Asencio, N., Bougeault, P., Ducrocq, V., Duron, J., Fischer, C., Héreil, P., Mascart, P., Masson, V., Pinty, J. P., Redelsperger, J. L., Richard, E., and Vilà-Guerau de Arellano, J.: The Meso-NH Atmospheric Simulation System. Part I: adiabatic formulation and control simulations, Ann. Geophys., 16, 90–109, doi:10.1007/s00585-997-0090-6, 1998.

Pappenberger, F., Ghelli, A., Buizza, R., and Bodis, K.: The skill of probabilistic forecasts under observational uncertainties within the generalized likelihood uncertainty estimation framework for hydrological applications, J. Hydromet., 807–819, 2009.

Pinty, J. P. and Jabouille, P.: A mixed phase cloud parameterization for use in a mesoscale nonhydrostatic model: Simulations of a squall line and of orographic precipitation. Preprints, Conf. on Cloud Physics, Everett, WA, Amer. Meteor. Soc., 217–220, 1998.

Skok, G. and Vrhovec, T.: Considerations for interpolating rain gauge precipitation onto a regular grid, Meteorol. Z., 15, 545–550, 2006.

Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL – A novel quality measure for the verification of quantitative precipitation forecasts, Mon. Weather Rev., 136, 4470–4487, 2008.

Wernli, H., Hofmann, C., and Zimmer, M.: Spatial forecast verification methods intercomparison project: application of the SAL technique, Wea. Forecasting, 24, 1472–1484, 2009.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 2nd Edn., Academic Press, 627 pp., 2006.

Wilks, D. S.: Resampling hypothesis tests for autocorrelated fields, J. Climate, 10, 65–82, 1997.

WMO: Recommendations for the verification and intercomparison of QPFs and PQPFs from operational NWP models, World Meteorological Organization, Revision 2, 2009.