Adv. Sci. Res., 8, 149–155, 2012
www.adv-sci-res.net/8/149/2012/
doi:10.5194/asr-8-149-2012

Advances in Science & Research
Open Access Proceedings

A quantitative evaluation of the high resolution HARMONIE model for critical weather phenomena

E. V. van der Plas, B. Wichers Schreur, and K. Kok

Royal Netherlands Meteorological Institute, De Bilt, The Netherlands

Correspondence to: E. V. van der Plas
(emiel.van.der.plas@knmi.nl)

Received: 13 January 2012 – Revised: 11 July 2012 – Accepted: 17 July 2012 – Published: 31 July 2012

Abstract. The high resolution non-hydrostatic Harmonie model (Seity et al., 2012) seems capable of delivering high quality precipitation forecasts. The quality with respect to the European radar composite is assessed using the Model Evaluation Tool, as distributed by the NCAR DTC (Developmental Testbed Center, 2012), and compared to that of the reference run of Hirlam (Unden et al., 2002), the current operational NWP model at KNMI. Both neighbourhood and object-based verification methods are compared for a week with several high intensity precipitation events in July 2010. It is found that Hirlam scores very well in most metrics, and that in spite of the higher resolution the added value of the Harmonie model is sometimes hard to quantify. However, higher precipitation intensities are better represented in the Harmonie model with its higher resolution. Object-based methods do not yet yield a sharp distinction between the different models, as it proves difficult to construct a meaningful and distinguishing metric with a solid physical basis for the many settings that can be varied.

1 Introduction

The interest in high resolution numerical weather prediction is mainly driven by the presumed ability to skillfully predict extremes in critical weather situations. Heavy precipitation and strong wind may be very local phenomena, for which even a 2.5 km grid spacing could be too coarse to resolve, but present day numerical models try to improve upon the forecasts of these events nevertheless.

The higher resolution of the models also poses a challenge for the verification of the forecasts. When the timing or location of a shower is only a few minutes or kilometers off, a pixel-per-pixel comparison will see this as a double mismatch: a shower is forecasted where it is not observed, giving rise to a false alarm, and the observed shower is not forecasted, counting as a "miss". Nevertheless, these may be very useful forecasts, giving relevant information to most of the end-users.

The IMPACT project aims to evaluate the high resolution non-hydrostatic model HARMONIE (Seity et al., 2012), being developed by the Hirlam/Aladin consortium, for a series of cases in which the weather was critical to the operations of Schiphol Airport (Amsterdam, The Netherlands).

In this paper a period with several events with convective precipitation will be studied using two verification approaches that try to circumvent the double penalty problem.

2 Methods

Various approaches to the problem of verification of high resolution precipitation forecasts have been proposed. In this paper two of those will be studied: neighbourhood methods, such as the Fractions Skill Score (FSS) (Roberts and Lean, 2008), and the object-based MODE (Davis et al., 2006). The software to apply these methods is contained in the Model Evaluation Tool (MET), developed by the developmental testbed center (DTC) at NCAR (Developmental Testbed Center, 2012).

2.1 Neighbourhood methods

The most widely used method to take into account that higher resolution forecasts may introduce localisation errors is the

fractions skill score (FSS). This method attributes merit to a forecast if the criterion (e.g. 1 h accumulated precipitation > 5 mm) is met in a *neighbourhood* of the observed event.

So, following Roberts and Lean (2008), the observed and forecasted rainfall distribution O_r and M_r are converted into binary fields I_O and I_M, set to 1 when a threshold q is exceeded and zero otherwise.

Subsequently, for every grid point in the binary fields I_O, I_M, the fraction of points with a value 1 within a square of (odd) length n are computed, O_n and M_n. This is effectively *smoothing* the respective fields.

The mean squared error (MSE) for the observed and forecast fractions is then given by

$$\mathrm{MSE}_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [O_{n,ij} - M_{n,ij}]^2.$$

This score depends highly on the frequency of the event itself, so by defining a reference MSE,

$$\mathrm{MSE}_{(n),\mathrm{ref}} = \frac{1}{N_x N_y} \left[\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} O_{n,ij}^2 + \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} M_{n,ij}^2 \right],$$

a skill score relative to a low-skill forecast can be defined: the fractions skill score (FSS):

$$\mathrm{FSS}_{(n)} = \frac{\mathrm{MSE}_{(n)} - \mathrm{MSE}_{(n),\mathrm{ref}}}{\mathrm{MSE}_{(n),\mathrm{perfect}} - \mathrm{MSE}_{(n),\mathrm{ref}}} = 1 - \frac{\mathrm{MSE}_{(n)}}{\mathrm{MSE}_{(n),\mathrm{ref}}}.$$

This score is minimal for a neighbourhood size of 1, i.e. pixel-per-pixel comparison, and tends asymptotically to a ratio of the frequencies of the observed and forecasted events, as:

$$\mathrm{FSS}_{(n)} \to 1 - \frac{(f_O - f_M)^2}{f_O^2 + f_M^2} \quad \text{for } n \to N,$$

where $N = N_x N_y$ is the total number of available grid points, and f_O and f_M are the fractions of the observed and forecasted points exceeding the threshold over the whole domain.

Using the neighbourhood aggregation over an area around the pixel, it is also possible to compute other well-known contingency table statistics (CTS, see Table 1, and e.g. Ebert (2012)). The MET suite provides most of the relevant CTS scores, but the score that will be used in this paper is the Hanssen-Kuiper discriminant (HK). Another well-known quantity in this respect is the Gilbert Skill Score (GSS), also known as the Equitable Threat Score (ETS), but the same overall behaviour emerges as for the HK, and is therefore omitted here for brevity.

The Hanssen-Kuiper discriminant is defined as

$$\mathrm{HK} = \frac{a}{a+c} - \frac{b}{b+d},$$

which gives a measure of how well the areas with precipitation are distinguished from the areas without.

Table 1. Contingency table

		Observed	
		yes	no
Forecast	yes	a	b
	no	c	d

2.2 Object-based verification: MODE

The method for object-based diagnostic evaluation (MODE, Davis et al., 2006), has a quite different approach. Again a binary field (I_O, I_M) is constructed depending on the exceedance of the threshold q. Only now, this field is convoluted by a circular kernel K of r pixels wide to make the field more contiguous, and filter out small and potentially uninteresting features,

$$I_{O,C} = \int I_O(\mathbf{x}) K(\mathbf{x}, r) \mathrm{d}^2 x,$$

thus constituting the *objects*. The combination of threshold q and convolution radius r determine the distribution of the resulting objects.

To associate a forecasted object to an observed object, an interest function F prescribes, on a scale from 0 to 1 (1 being perfect), how closely an attribute of the forecasted object matches the same attribute of the observed object. For any forecast and observed object pair, the total interest C is then defined as

$$C = \frac{\sum_{i=1}^{M} c_i w_i F_{i,j}}{\sum_{i=1}^{M} c_i w_i}.$$

Here w is the weight assigned to a certain attribute i of an object, e.g. the location of its center of mass, and c is a function of attributes that describes the confidence in a partial interest value obtained from $w_i F_{i,j}$. Total interest assumes a value between 0 and 1, and it is used to associate the objects to one another. When the total interest is larger than 0.65 objects are considered "matched". When several objects match to each other above this threshold the objects are called a "cluster".

As an example, the result of one particular time and model is shown in Fig. 1. Here, e.g. the forecasted objects 19, 20 and 21 are compared to observed object 15 (lower right corner, in green) because they are not too far apart, they have a similar intensity, the difference in the angle is not too large etc. This amounts to a total interest larger than 0.65, and hence they are matched.

3 Data

The objective of this paper is to compare the forecast produced by a non-hydrostatic high-resolution model to a current operational standard, in this case HARMONIE and the hydrostatic Hirlam RCR model, respectively. These are then

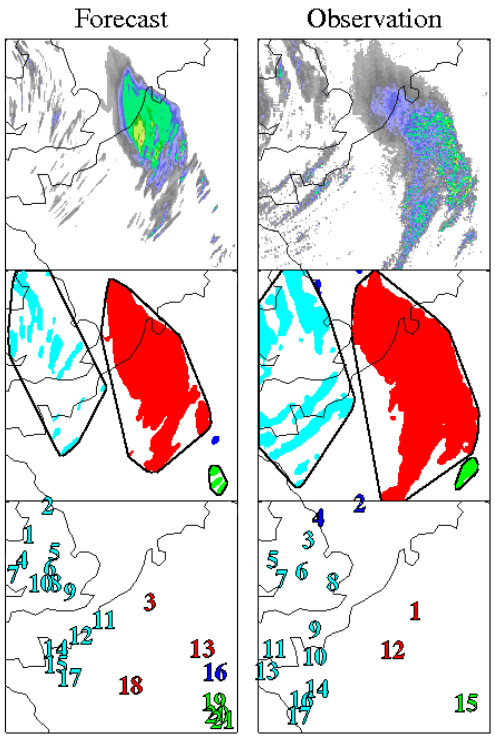

Figure 1. Typical output of the MODE algorithm, here for Harmonie with Hirlam boundaries, 14 July, 18:00 UTC: on the left the forecasted precipitation field (top: raw data, middle: thresholded and convoluted objects, bottom: objects numbered and colored by matching criteria with observations), on the right the radar data.

compared to the European radar composite at 4 km resolution.

In this study we have chosen a 10 day period from 6 to 15 July 2010. In this period weather alerts have been issued by KNMI for extreme precipitation on 10, 12 and 14 July. The latter case also developed a phenomenon which was believed to be a *micro-downburst*: the passage of the front induced a very strong yet highly localised wind field, resulting in substantial damage and the loss of life.

3.1 Model data

The HARMONIE model is the latest high resolution model that is used within the Hirlam/ALADIN community, using the AROME non-hydrostatic dynamical core (cycle 36h1.3) which was developed by the ALADIN community. It runs on a limited area, getting initial conditions and boundaries from either the Hirlam RCR run or the ECMWF operational analysis and forecast. Runs are executed on a 400^2 grid with 2.5 km resolution using both boundary strategies.

The HARMONIE runs are initialised by either the Hirlam RCR or the ECMWF analysis. However, as some of the HARMONIE prognostic variables, e.g. vertical velocity for convection and the hydrometeors, are not initialised, this will effectively be considered as a "cold start". The starting times are 00:00, 06:00, 12:00 and 18:00 UTC, with 12 h forecasts. A spin-up time of 3 h is taken into account, so the data of T+003 until T+009 are considered in this particular study.

Table 2. Model configuration

	Harmonie	Hirlam RCR	Radar
Vertical levels	40	60	n.a.
Domain	400×400	582×446	512×512
Resolution	2.5 km	0.15°	4 km
Projection	Lambert	rot. lat-lon	polar stereogr.
Assimilation	none	3DVAR	

The Hirlam RCR run is performed on a substantially larger domain (see Table 2), and performs an analysis using 3DVAR data assimilation.

3.2 Observations

The choice for the European composite radar product RADNL23 can be motivated by the assertion that in this pilot study we primarily concern ourselves with the distribution of (extreme) precipitation. It is recognised that this data is aggregated from various different radar installations, giving rise to quantitative differences for the different areas. Also, over the North Sea radar clutter can give rise to complications for the verification methods, especially the object-based methods. As the scatter has very little spatial extent but may have considerable intensity, the convolution step in MODE tends to overestimate these areas. A more reliable radar product may one day be available as a result of the OPERA project, and by cross-validating with MSG satellite data during daytime.

3.3 Grids

To compare the gridded observations and the model output it is necessary to transform the data to a uniform grid. To prevent the introduction of artificial values all data was resampled unto the highest resolution grid available, i.e. the HARMONIE Lambert-conformal grid, using a nearest-neighbour method.

4 Results

First of all, the total precipitation over the 10 day period can be compared, see Fig. 2. The radar shows the largest amount of rain, also with the largest amount of variation. The Hirlam RCR data gives a considerably smoother picture. Here we point out that the data shown here is already resampled to the Harmonie grid, giving rise to some obvious resampling artefacts. The Harmonie data resembles the radar data more closely, where the run nested in the Hirlam

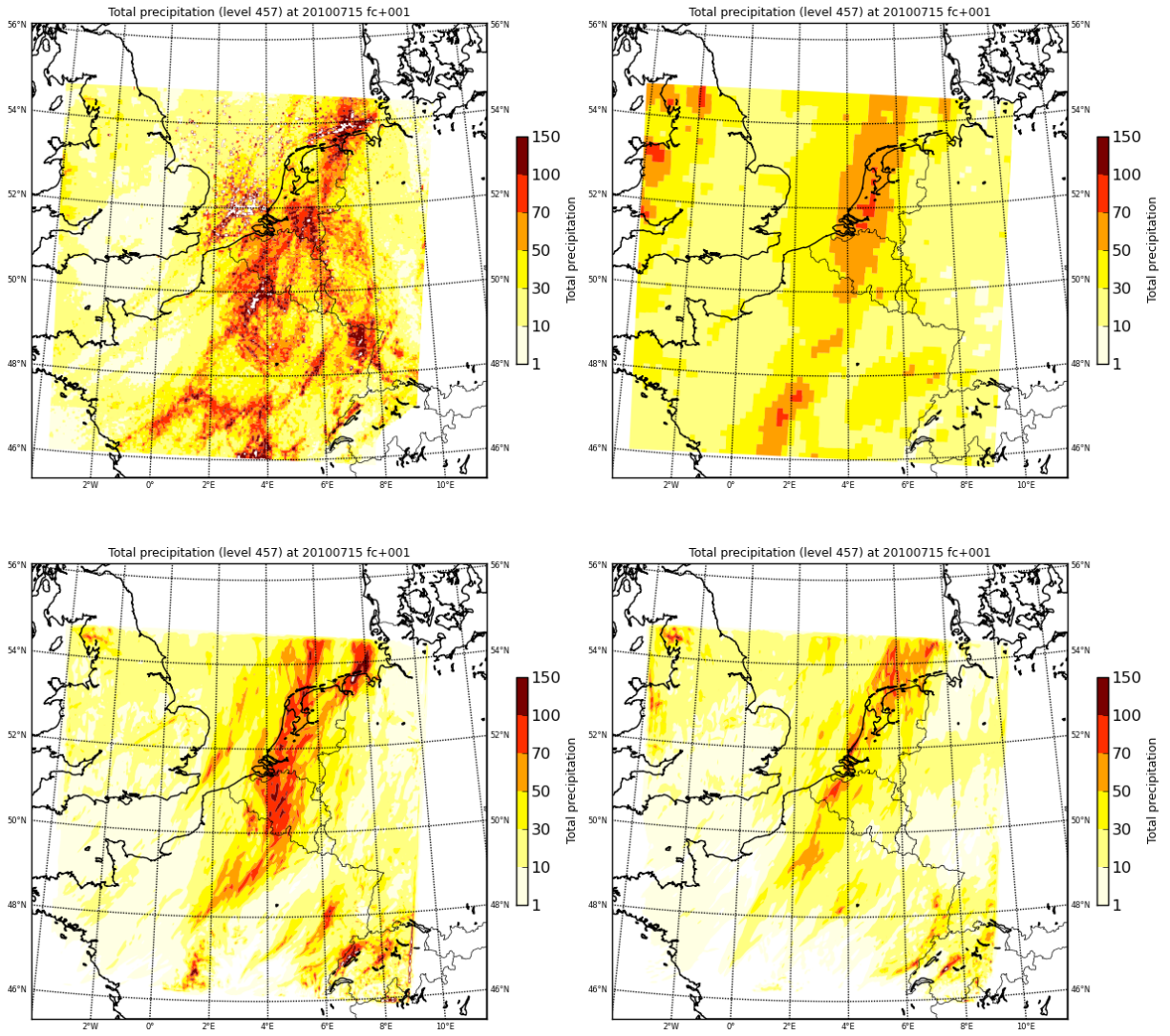

Figure 2. Total accumulated precipitation for the period 6–15 July 2010, using the resampled data. Top left: radar data, right: Hirlam RCR. Bottom left: Harmonie with Hirlam boundaries, right: Harmonie with ECMWF boundaries.

model gives higher quantities, especially in the band of intense rain that crossed Belgium and the Netherlands. These figures can be summarised in the histogram in Fig. 3 (top). The distribution of the radar data is shifted towards higher intensities than the model data, whereas the Harmonie runs have more dry pixels, reflecting the tendency of the model to underrepresent light rain. The maximum in the distribution in the Hirlam RCR precipitation around 20 mm can probably be attributed to the coarser resolution of the model set-up: the extremes (> 100 mm) are underrepresented, because generally high intensity precipitation does not extend over large areas. On the other hand the amount of "dry" pixels is reduced because of the same effect: if there is precipitation in the grid box, the whole box will give non-zero precipitation. This leads to a shift of the distribution towards its mean, in this case approximately 35 mm over the whole period.

In the bottom panel of Fig. 3 the histogram of the three hour precipitation sums in the same periodis shown. Again we see how the Hirlam RCR run tends to overestimate and smear out low intensities due to the coarser resolution. This might also lead to the overestimation of the 60–80 mm bins in the total accumulation in the top figure: many low-intensity precipitation events could lead to accumulative amounts in this range (10 days times 8 3-h accumulations times 1 mm/3 h yields 80 mm).

Harmonie however clearly underrepresents low precipitation intensities. This is in line with the experience that the forecasted showers tend to be too localised: stratiform, light rain is not allways well captured in the model.

A time series of the FSS is shown in Fig. 4. The three events with heavy precipitation are clearly distinguishable.

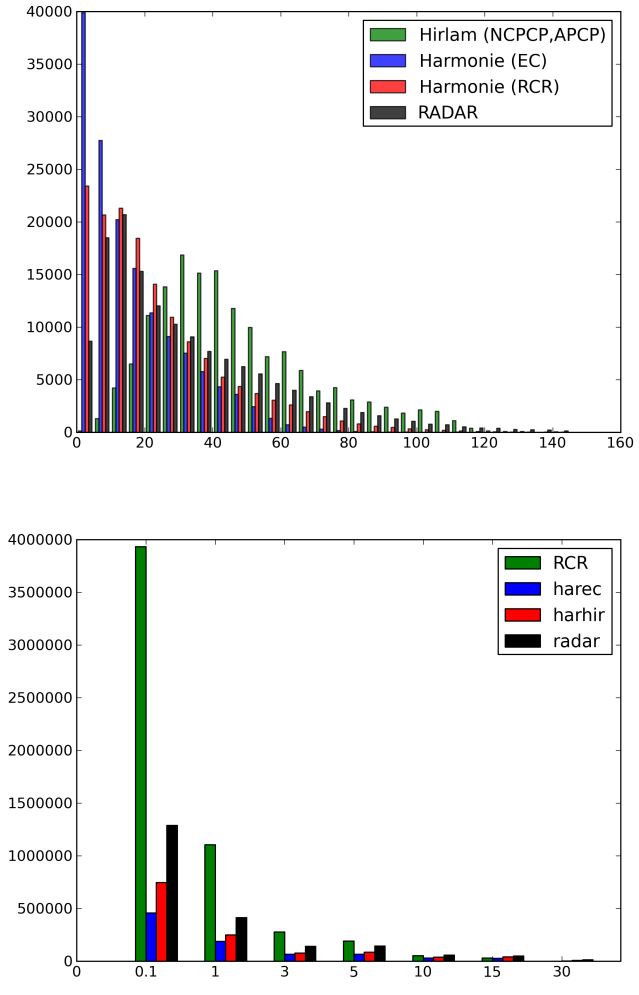

Figure 3. Top panel: histogram of the total accumulated precipitation for the period 6–15 July 2010. Here, the resampled data was binned in 5 mm bins (the interval [0,5] is depicted at 0, etc.). The bottom panel shows the accumulative histogram for each three hour period, binned between 0.1, 1, 3, 5, 10, 15, 30 and 100 mm per three hours (the interval (0.1,1] is depicted at 0.1, etc).

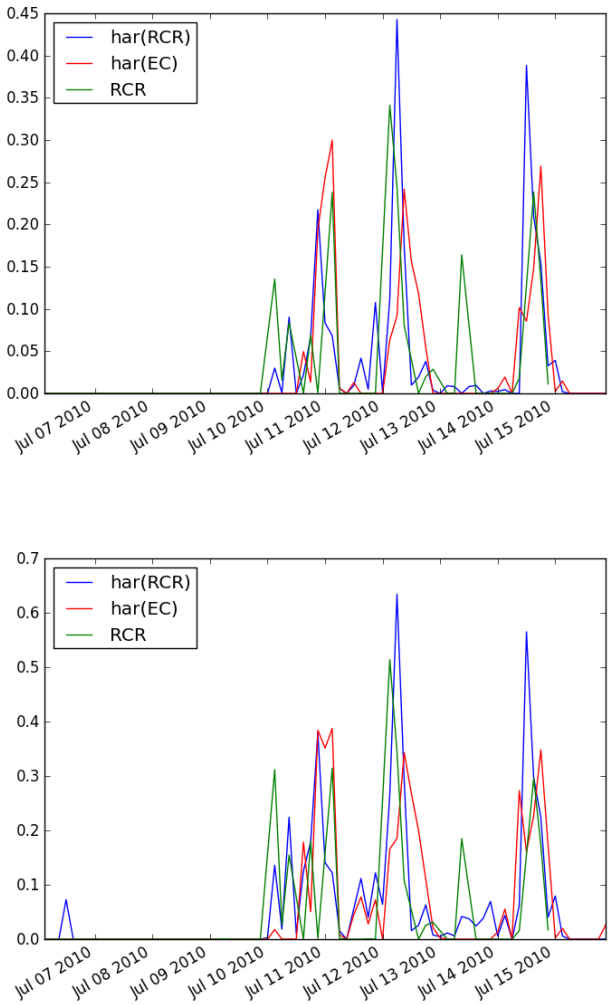

Figure 4. Time series of the FSS for precipitation over 5 mm/3 h, with a neighbourhood size of 1 pixel or 2.5×2.5 km. (top) and 15^2 pixels or 37.5×37.5 km (bottom). The green line represents the RCR run data, blue Harmonie using Hirlam boundaries and red Harmonie using ECMWF boundaries.

We compare the different runs taking either a neighbourhood size of 1 by 1 pixel, the classical CTS score, or smoothing the data over an area of between 3 by 3 (7.5 by 7.5 km) to 75 by 75 (187.5 by 187.5 km) pixels. This has been taken as the upper limit, as the amount of areas within the computed domain as well as the amount of counted events drops significantly for larger areas.

In this figure the scores for precipitation of more than 5 mm per three hours is shown. The scores are comparable for the three models. For the 1 pixel neighbourhood size, i.e. pixel-per-pixel comparison, shown on top, Hirlam scores generally slightly higher for the large scale events, and also picks up an event on 9th July (not shown here). This holds for both the FSS and the HK discriminant score. Increasing the neighbourhood size, to 15 by 15 pixels, shown in the bottom panel, results obviously in increasing scores as generally the number of counted events over larger areas will increase, but the same qualitative picture remains. The scores of the Harmonie runs increase a bit more than those of the Hirlam RCR run.

The behaviour of the FSS and HK scores as a function of the neighbourhood size is depicted in Fig. 5. The FSS for the Harmonie model runs increase monotonically and visibly faster than the FSS for the Hirlam model, where the median does not reach beyond a FSS of roughly 0.2.

The HK score is less pronounced in this respect. For smaller regions (1 by 1 to 9 by 9 gridboxes) the scores are of the same order. Harmonie with ECMWF boundaries scores slightly higher (median at HK $\approx$ 0.1 vs. HK $\approx$ 0.05). For areas of 15 and 25 gridboxes wide Hirlam scores more consistently, but from 35 gridboxes upward this advantage is reversed again.

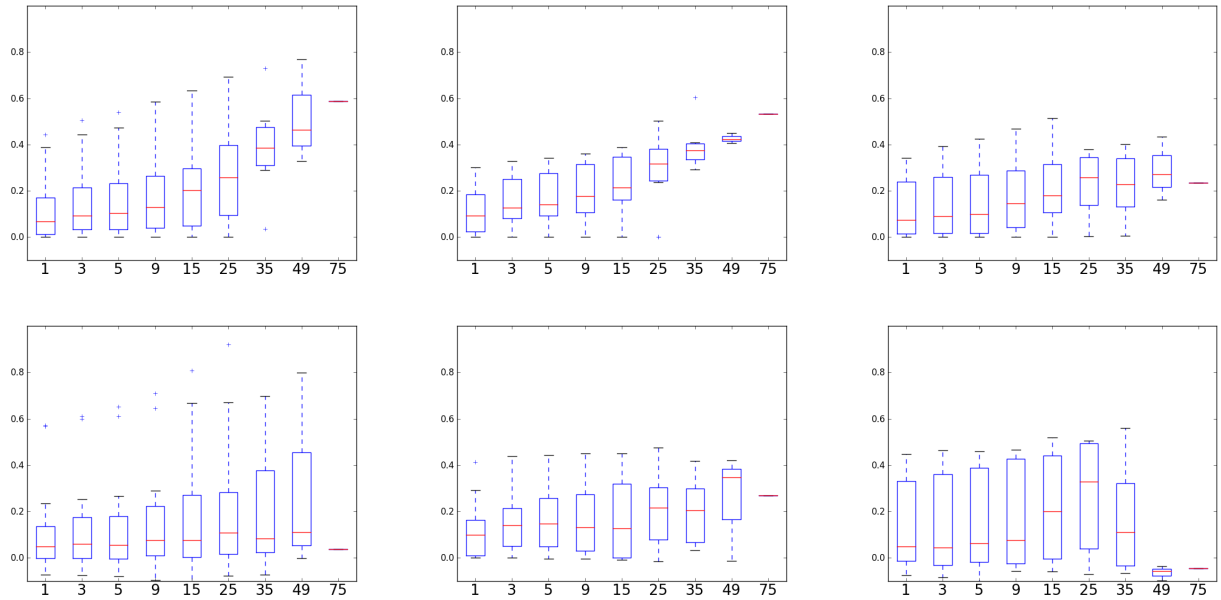

Figure 5. Box plots of the FSS (top) and HK (bottom) for rain over 5 mm per three hour as a function of the neighbourhood size for Har(RCR), Har(EC) and RCR (left, middle and right). The red line represents the median of the data, and the box extends from the 25th to the 75th percentile. The whiskers extend to 1.5 times the distance from the mean to the 25th and 75th percentile range. Points outside the whiskers are denoted with a blue +, and considered outliers from a visualisation perspective.

The outliers, corresponding to the events with the highest precipitation intensities, show a considerable edge to Harmonie with RCR boundaries. The neighbourhood method averages over a square of increasing size, and this may yield unfavourable results for the typically highly localised precipitation forecasts of Harmonie at the intermediate areas.

If we look at the MODE analysis, we first observe that the process of grouping precipitation features into objects of a certain minimum size makes it easier to do visual, subjective verification. An example of the visual output of the thresholding and convolution process as performed by MET is shown in Fig. 1. Matched events are coloured accordingly (dark blue means not matched), and one can easily perform a quick subjective assessment of the situation considered. Per case or time step one may compare certain attributes of objects one is interested in, such as total interest or centroid distance. However, to condense this information into a single score over the entire time range is less straightforward.

Furthermore, the configuration of MODE has many degrees of freedom. Some settings, such as the radius of the convolution kernel, can make a substantial difference in how the objects come out.

Statistics of the individual objects may give interesting information for model intercomparison studies (different physics, initial and boundary conditions etc.), but generally have very little meaning in the context of (longer) timeseries. One might follow e.g. the largest matched object or identified cluster, but the objects or clusters are not tracked in time, so this naive attribute does not tell whether the largest cluster in one time step is related to the largest cluster in the next. Also, the grouping of objects into a cluster is not always very consistent in the sense that storms that belong to different clusters in one time step may coalesce into the same cluster and vice versa. In Fig. 1 one sees how grouping may seem a bit arbitrary.

One more consistent method to construct a score was proposed in Davis et al. (2009), using the median of the maximum interest (MMI) of the whole domain. The matching procedure computes the interest between all features, and considers it a match when this number is above a certain, user defined threshold. By considering the median of all these interest values, we have a measure that reflects how well the forecast performed for a given moment, and that can be used to compare different models and different (dryer and wetter) periods. The results for this particular case are shown in the right panel of Fig. 6. It is remarkable that the the high-resolution Harmonie model performs comparable to the Hirlam model for this particular score, with the exception of the large lower tail for Harmonie using ECMWF boundaries.

5 Discussion

A few different verification methods have been applied to Hirlam and HARMONIE model data, using the European composite radar product to verify against.

The high level of detail in the forecast would suggest that classical methods, such as (neigbourhood) contingency table statistics, are less well equipped to attribute the forecasted

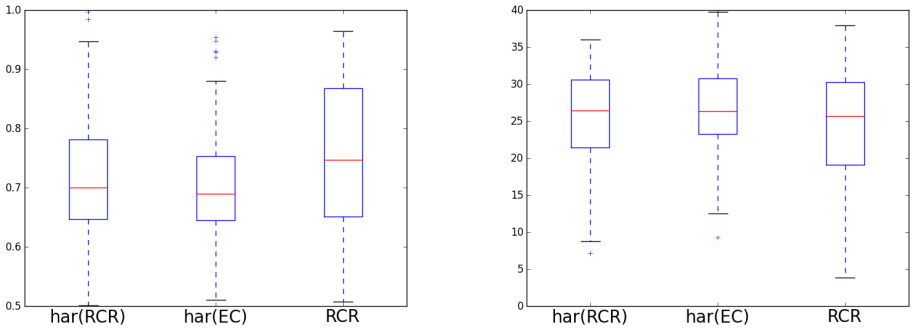

Figure 6. Box plot of the median of the maximum interest (MMI, left) and scores and the median of the centroid distance between two matched objects (right) for Harmonie using Hirlam and ECMWF boundaries and the Hirlam RCR run over the 10-day period. Box definitions as in Fig. 5.

precipitation features to its observed counterparts than the more flexible object-based methods, such as MODE. However, the qualitative picture that arises when comparing the scores of Harmonie and Hirlam shows that using the former CTS-based methods, with e.g. the fractions skill score or the HK discriminant, give relatively more credit to the higher resolution forecasts than for this specific choice of MODE results.

The higher resolution of the Harmonie model does give a better representation of the more extreme events (> 5 mm/3 h) than Hirlam. This is one of the expected advantages from a high-resolution model that properly takes into account the dynamics on the smaller scales. These higher precipitation intensities were hardly present in the Hirlam data as a result of the coarser resolution. It is noted (not shown) that the events with the highest precipitation intensities generally correspond to positive outliers in the scores, though more investigation and a larger dataset are needed. These points hardly contribute to the value of the median of the distribution, even if the median should be regarded as the score for a certain measure.

Also, the Harmonie results shown here are obtained with a model set-up that does not use data-assimilation of any kind. This means that especially quantities that are not being initialised by the analysis on which they are based, such as cloud cover and precipitation, suffer from spin-up. Experiments show that for the atmosphere to display a properly dsitributed state it takes the model between 4 to 6 h. Longer lead times show better scores for these parameters when cold starts are considered.

It is stressed that this is just a preliminary study with a very modest amount of data. Furthermore, the output of the MODE algorithm is so rich, that obviously more effort should be invested into combining the attributes in such a way as to produce a score that gives intuitive results and can be compared over a variety of cases.

Acknowledgements. Model Evaluation Tools (MET) was developed at the National Center for Atmospheric Research (NCAR) through a grant from the United States Air Force Weather Agency (AFWA). NCAR is sponsored by the United States National Science Foundation.

Edited by: H. Böttger
Reviewed by: K. Eerola, P. Unden, and P. W. Kallberg

The publication of this article is sponsored by the European Meteorological Society.

References

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V.: The AROME-France convective scale operational model, Mon. Weather Rev., 139, 976–991. 2012,

Roberts, N. M. and Lean, H. W.: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events, Mon. Weather Rev., 136, 78–97, 2008,

Davis, C. A., Brown, B. G., and Bullock, R.: Object-Based Verification of Precipitation Forecasts. Part I: Methodology and Application to Mesoscale Rain Areas, Mon. Weather Rev., 134, 1772–1784, 2006.

Davis, C. A., Brown, B. G., Bullock, R. and Halley-Gotway, J.: The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program, Weather Forecast., 24, 1252–1267, 2009.

Ebert, E.: Forecast Verification: Issues, Methods and FAQ, http://www.cawcr.gov.au/projects/verification, 2012.

Developmental Testbed Center, MET: version 3.0.1 Model Evaluation Tools Users Guide, http://www.dtcenter.org/met/users/index.php, 2010.

Unden, P., Rontu, L., Järvinen, H., Lynch, P., Calvo, J., Cats, G., Cuxart, J., Eerola, K., Fortelius, C., Garcia-Moya, J. A., Jones, C., Lenderink G., Mcdonald, A., Mcgrath, R., Navascues, B., Woetman Nielsen, N., Degaard, V., Rodriguez, E., Rummukainen, M., Sattler, K., Hansen Sass, B., Savijarvi, H., Wichers Schreur, B., Sigg, R., and The, H.: HIRLAM-5 Scientific Documentation, 2002.