A Comparative Verification of High-Resolution Precipitation Forecasts Using Model Output Statistics

EMIEL VAN DER PLAS, MAURICE SCHMEITS, NICOLIEN HOOIJMAN, AND Kees Kok
Royal Netherlands Meteorological Institute, De Bilt, Netherlands

(Manuscript received 6 July 2016, in final form 30 June 2017)

ABSTRACT

Verification of localized events such as precipitation has become even more challenging with the advent of high-resolution mesoscale numerical weather prediction (NWP). The realism of a forecast suggests that it should compare well against precipitation radar imagery with similar resolution, both spatially and temporally. Spatial verification methods solve some of the representativity issues that point verification gives rise to. In this paper, a verification strategy based on model output statistics (MOS) is applied that aims to address both double-penalty and resolution effects that are inherent to comparisons of NWP models with different resolutions. Using predictors based on spatial precipitation patterns around a set of stations, an extended logistic regression (ELR) equation is deduced, leading to a probability forecast distribution of precipitation for each NWP model, analysis, and lead time. The ELR equations are derived for predictands based on areal-calibrated radar precipitation and SYNOP observations. The aim is to extract maximum information from a series of precipitation forecasts, like a trained forecaster would. The method is applied to the nonhydrostatic model Harmonie-AROME (2.5-km resolution), HIRLAM (11-km resolution), and the ECMWF model (16-km resolution), overall yielding similar Brier skill scores for the three postprocessed models, but somewhat larger differences for individual lead times. In addition, the fractions skill score is computed using the three deterministic forecasts, showing slightly higher skill for the Harmonie-AROME model. In other words, despite the realism of Harmonie-AROME precipitation forecasts, they only perform similarly or somewhat better than precipitation forecasts from the two lower-resolution models, at least in the Netherlands.

1. Introduction

The direct model output of current high-resolution precipitation forecasts has a degree of realism that is very appealing for most end users of meteorological data. However, it is not at all obvious to show that the high-resolution precipitation forecast really contains more skillful information than that of a lower-resolution counterpart [see the review by Clark et al. (2016)].

From a user’s perspective it seems straightforward to compare a rain forecast to a nearby rain gauge measurement. However, its localized nature gives rise to representativity issues: the rain gauge observation data are very limited in the spatial sense, possibly missing an event entirely (e.g., Witter 1984; Ren and Ren 2012). Considering both the observations and the forecast data only at the locations of the SYNOP stations is a projection operation in which potentially much valuable data are neglected.

The availability of high-resolution radar or satellite imagery seems to solve one side of this issue. The increased accuracy in the location of rain events comes at the cost of a decreased accuracy in estimating the amount of rain that has actually reached the ground. More importantly, precise observation reveals that at the (sub)kilometer scale rain is an essentially stochastic process, dependent on small-scale processes that are not within the scope of numerical weather prediction. This is best illustrated by the double-penalty problem (e.g., Nurmi 2003). If a thunderstorm is forecast at a certain time and location, and the timing is slightly off, or the storm occurs only several tens of kilometers farther, it will rain where it is not predicted (a missed event), and the rain forecast will be a false alarm. Standard verification based on pixelwise comparison to obtain a mean (squared) error will yield poor results, even though it may be a very valuable forecast.

Verification methods using traditional (“point to point”) accuracy measures like the root-mean-square error (RMSE) and the mean absolute error (MAE) tend
to favor smoother forecasts (e.g., Baldwin et al. 2001; Gallus 2002; Mass et al. 2002; Kok et al. 2008; Ebert 2008), among others due to the double-penalty problem, especially for localized phenomena such as convective precipitation. Because of their smoothing effect lower-resolution models tend to have an advantage in the comparison between models of different resolution.

There are several approaches that aim to overcome this problem (Ebert et al. 2013) by considering a region around the actual data point in the case of neighborhood methods such as the fractions skill score (FSS) (Roberts and Lean 2008; Mittermaier and Roberts 2010; Mittermaier et al. 2013; Mittermaier 2014) and the “pragmatic” method of Theis et al. (2005). In Azorin-Molina et al. (2014) a comparison for sea-breeze thunderstorms was made between the HIRLAM and HIRLAM Aladin Regional Mesoscale Operational NWP In Europe-Application of Research to Operations at the Mesoscale (Harmonie-AROME) numerical weather prediction (NWP) models that are also considered in the present paper, indicating improved skill for the nonhydrostatic Harmonie-AROME model. Neighborhood methods have also been applied to other weather parameters by Mittermaier and Bullock (2013) and Mittermaier (2014) to compare skill between two models of different resolution. These papers show that the higher resolution does add skill when the forecasts are considered as probabilities on a lower resolution.

Techniques that make use of fuzzy logic to associate patches of forecast and observed rain, such as the structure, amplitude, and location method (SAL; Wernli et al. 2008) and the Method for Object-based Diagnostic Evaluation (MODE; Davis et al. 2006a, b, 2009), can give extra diagnostic information on how well the distribution of storms is represented in a model (Wolff et al. 2014). These methods consider the spatial structure of rain areas, so one can compare quantities such as the area, center of mass, and orientation of a precipitation feature between the forecast and (radar) observation. In van der Plas et al. (2012) MODE and FSS were applied to extreme weather cases in 2010 using the Harmonie-AROME model and HIRLAM.

Furthermore, as pointed out in, for example, Kok et al. (2008), Mittermaier (2014), and Lewis et al. (2015), because both the forecast and the observation at the current spatial scales should not be interpreted as the result of a deterministic process, it is clear that probabilistic tools are needed to assess them. Historically, model output statistics (MOS; Glahn and Lowry 1972) has been used to give quantitative guidance using a NWP (ensemble) forecast (e.g., Applequist et al. 2002; Hamill et al. 2004; Bentzien and Friederichs 2012), removing model biases for specific locations, or giving the probability of the occurrence of phenomena that are not available from the direct model output like (severe) thunderstorms (e.g., Schmeits et al. 2005, 2008).

The goal of the current study is to make a comparative verification of high-resolution precipitation forecasts from NWP models with different resolution using a verification method based on MOS. This method, extensively described in Kok et al. (2008), attempts to extract all relevant characteristics (predictors) from the available precipitation forecasts that have predictive potential for the probability of rain (exceeding some threshold) at a location or area during a time interval (predictand). Here not only the percentage of exceedances in a neighborhood is taken into account, as in most of the fuzzy methods (Ebert 2008), but also other potentially skillful information from the forecast precipitation like intensity and proximity to the predictand location (or its combination). This approach selects the most important objective combination from these forecast attributes and as such can be seen as an extension of the previously mentioned fuzzy verification methods in which only one of the attributes is included. In other words, we concentrate on the, hereafter called, information content, assessed using statistical postprocessing, including testing its skill on an independent dataset. The information content is therefore expressed in terms of the combined set of selected predictors, including their relative importance, and is in general different for different predictands and lead times.

With a suitable choice of predictors the resulting probabilities will suffer less from systematic errors or double-penalty effects. In our comparative verification we compare the differences in information content between models with different grid resolution and both hydrostatic and nonhydrostatic model formulations using probabilistic verification metrics. These models are the hydrostatic HIRLAM, which is still an operational NWP model at KNMI (11-km resolution, but archived at 22-km resolution; Unden et al. 2002), the nonhydrostatic operational Harmonie-AROME model (2.5-km resolution; Seity et al. 2011), and the hydrostatic ECMWF Integrated Forecast System (IFS; T1279 corresponding to about 16-km resolution), while in Kok et al. (2008) only two hydrostatic and lower-resolution versions of the ECMWF IFS were used in the comparative verification; namely, the operational and control runs with resolutions of about 25 and 50 km, respectively. In addition, in the current study also three types of predictands are investigated, defined as 3-h accumulated observed rain exceeding several intensity thresholds, based on rain gauge measurements at 7 SYNOP stations and areal mean and
maximum precipitation in circular areas of 5 and 25 km around these SYNOP stations (Fig. 1) using calibrated radar data (Holleman 2007), while Kok et al. (2008) only used 3-h accumulated observed precipitation at the SYNOP station in De Bilt, Netherlands. So both the user-oriented viewpoint of “single observation–neighborhood forecast” as well as the model-oriented viewpoint of “neighborhood observation–neighborhood forecast” (Ebert 2008) are included in the current paper.

The paper is organized as follows. In sections 2 and 3 the NWP model data and observation data that have been used are described, respectively. Also the procedures that have been used to arrive at the three different predictand types are discussed. In section 4 the statistical post-processing method is discussed, and the procedure of the predictor selection is elucidated. The choice of the potential predictors that have been constructed from the forecast data is explained in section 5 and in the appendix. In section 6 the results are presented for the different predictands. A qualitative comparison is made with the well-known and more established FSS in section 7. Finally, in section 8 the results are discussed.

2. NWP models

Statistical postprocessing using MOS is applied to the output from each of three NWP models with different resolutions: Harmonie-AROME (Seity et al. 2011), HIRLAM (Unden et al. 2002), and the ECMWF IFS model.

The model with the highest resolution is the Harmonie-AROME model, which is based on AROME physics, as used within the HIRLAM community. A model version (36h1.4) with a 2.5-km resolution grid of 789 by 789 points, covering the Netherlands and large parts of western Europe, with 3DVAR assimilation of conventional data (SYNOP, radiosonde) has been operational since the summer of 2012.

This run is nested in the KNMI operational run of HIRLAM on an 11-km grid with 3DVAR data assimilation, which in turn is nested in the ECMWF IFS model, which is using 4DVAR data assimilation. The HIRLAM model output used in this paper was archived at a 22-km resolution. This means that the variance of the output can be considered as coming from an 11-km resolution model, but that the thinning of the data reduces the information content for spatial scales below 25 km. The ECMWF operational T1279 data have been used, with a horizontal resolution of approximately 16 km. In this study the operational model output datasets run from August 2012 to July 2014.

We have used 3-h accumulated precipitation data from the 0000 and 1200 UTC ECMWF model runs for lead times ending at +12 up to +30 h. In this study we take into account the actual availability of the model output at any given time, so the 0600 and 1800 UTC Harmonie-AROME and HIRLAM model output is compared to the 0000 and 1200 UTC ECMWF model output for the corresponding valid times, respectively. For HIRLAM and Harmonie-AROME we have used 3-h accumulated precipitation data for lead times ending at +6 up to +24 h.

3. Definition of the predictands

For the present experiment we investigate three types of 3-hourly accumulated precipitation predictands at or in areas around seven SYNOP stations in the Netherlands. At these stations, depicted in Fig. 1, automated observation data are available.

For the first predictand we use rain gauge data, because it is a fairly reliable observation source. For the other two predictands we have produced datasets on circular areas of 5 and 25 km around the seven stations, constructed from a calibrated composite of the two Doppler precipitation radars operated by KNMI, with a pixel resolution of approximately 1 km (Holleman 2007). For the second type of predictand we calculate the areal mean precipitation within these circular areas by simply averaging the radar rainfall amounts over all
pixels within these areas over a period of 3 h. In a similar fashion we took the areal maximum precipitation within these areas as the third type of predictand. All three types of predictands are defined as precipitation exceeding thresholds of 0.3, 1, and 2 mm (3 h)$^{-1}$ at least times every 3 h.

Despite a calibration by Holleman (2007) the radar dataset still suffers from “clutter”: data points with unrealistically high reflectivity that could be caused by reflections of structures such as buildings or windmills, and is especially pronounced in atmospherically very stable conditions due to anomalous reflections at the sea surface. This clutter was largely removed by applying a land–sea mask to the radar data, and disregarding maxima over 50 mm (3 h)$^{-1}$. Even though some anomalous values may still be present in the observation dataset, in a comparison of the performance of different NWP models all are equally affected.

As mentioned in the introduction, both predictand and predictor may suffer from poor representativeness. Considering the issue of representativeness, by using a disk of 25-km radius around the stations it is possible to capture considerably more events than at the SYNOP stations alone. Note that for stations near the coast the radar observations only cover the part above land due to the application of the land–sea mask.

We have also stratified the data between the “convective” (warm) season, which stretches from April to October, and the nonconvective (cold) season from November to March. The high-resolution Harmonie-AROME model is able to partly resolve convective scales, so some extra skill for higher thresholds during the warm season can be expected. In this paper only the results for the warm season are presented.

4. Statistical method

A historic dataset is constructed consisting of a binary predictand $y_i$ that indicates if the 3-h accumulated precipitation is exceeding a threshold $q$ at a given time $i$ and a given place or area (not indexed), and a set of predictors $s_{1,i}$, $s_{2,i}$, $s_{3,i}$, $s_{4,i}, \ldots$ for a given time $i$, location (not indexed), model and lead time, based on NWP model output. This set is split into training data and independent data, so that the derived relation between forecasts and observations can be verified. We have split the data by taking 20 subsequent days per month for the 14 warm months between August 2012 and July 2014 as training data (i.e., $\sim 280$ days), and the remaining days as test data (i.e., $\sim 140$ days), so that the weather in the two periods is mostly independent, but still having training and test cases in each month.

With a limited amount of data available, it was deemed necessary to pool the data over the seven different locations in order to be able to have enough training and test cases. The data will no longer be independent in the statistical sense. A balance between the amount of input data and the accuracy of the statistical model has to be sought. In the present paper the overall model skill is assessed, and we do not attempt to distinguish between the individual stations. Note that all three models are verified using the same test cases.

As our statistical postprocessing method we have used extended logistic regression (ELR; Wilks 2009), where the threshold is also a predictor so that the regression equation is a function of the threshold, effectively yielding a complete probability distribution. This has several advantages compared to standard logistic regression; namely, monotonicity of the probability distribution with respect to the threshold, possibly skillful probabilistic forecasts for higher thresholds, and much more robustness of the resulting regression equations. It does, however, make the outcome dependent on the choice of the thresholds on which the method is trained. In this study we have used 0.3, 1, and 2 mm in 3 h, and for precipitation maxima within a 5- and 25-km circle 0.3, 1, 2, and 3 mm in 3 h.

For a binary predictand $y_i$, here for an event with precipitation exceeding a threshold $q$, we try to find the probability $p_i$ as a function of the threshold $q$ and the other predictors $s = (s_{1,i}, s_{2,i}, \ldots)^T$, according to the nonlinear logistic equation:

$$p_i(s,q) = \frac{\exp[b_0 + b_1 s_{1,i} + b_2 s_{2,i} + \cdots + f(q)]}{1 + \exp[b_0 + b_1 s_{1,i} + b_2 s_{2,i} + \cdots + f(q)]}.$$  

In this study, we have chosen a linear function of the threshold in the extended logistic regression equation: $f(q) = a q$. Assuming only one predictor with a positive correlation to the predictand leads to an S-shaped curve as a function of $s_1$, where $\lim_{s_1\to-\infty} p_i(s_{1,i}) = 0$, and $\lim_{s_1\to\infty} p_i(s_{1,i}) = 1$. The parameters $\mathbf{b} = (a, b_0, b_1, b_2, \ldots)^T$ are fitted by maximizing the likelihood: if the $i$th event occurs ($y_i = 1$), the likelihood is $L_i = p_i$, and if the event does not occur ($y_i = 0$):

$$L_i = 1 - p_i = 1/[1 + \exp(b_0 + b_1 s_{1,i} + \cdots + a q)].$$

The joint likelihood is then

$$L(\mathbf{b}) = \prod_{i=1}^{n} \frac{y_i \exp(b_0 + b_1 s_{1,i} + \cdots + a q) + (1 - y_i)}{1 + \exp(b_0 + b_1 s_{1,i} + \cdots + a q)}.$$

The selection of the predictors is done in a forward selection procedure (Wilks 2011), using maximum likelihood.
First, the predictor with the best correlation is selected, where maximization of the log-likelihood yields the corresponding intercept and initial coefficients. Then, from the set of remaining predictors the predictor is selected which minimizes the Akaike information criterion (AIC; Akaike 1974).

\[ \text{AIC} = -2 \ln L + 2k, \]

with \( k \) the dimension of the free parameters in the statistical model. When the decrease of the AIC is less than 1% of the AIC of the regression equation using only a constant, AIC0, the procedure is stopped. The threshold \( q \) is treated in a similar fashion as the other predictors in the selection process, but it is always selected, as should be the case for ELR. Each step thus yields coefficients \( a, b_0, b_1, \ldots \) for the selected predictors.

To study the sampling uncertainty, the model selection and testing window on the training and test data: starting the training period at day 1 of each of the 14 months gives one set of regression equations, starting at day 2 gives another, etc. This yields a distribution of 30 different outcomes.

5. Potential predictors

To estimate the probability of precipitation exceeding a threshold \( q \) at a point or area for different analysis and lead times, a set of potential predictors is devised using the model precipitation output. If the objective of this exercise would be to construct an optimal probabilistic forecasting model for precipitation, predictors based on convective available potential energy (CAPE) and cloud water and ice content would make interesting candidates in addition to precipitation characteristics. In this paper we restrict ourselves to precipitation output to make a comparative verification of very different NWP models. Still, it could be argued that in order to compare the information content of a model, all information that can be derived from model output should be considered.

The idea behind the construction of these predictors is that a trained forecaster may use these attributes of a forecast to estimate the probability of precipitation in a point (area). Knowing the characteristics of a model, a forecast with many small-scale showers may be interpreted as a high probability of rainfall in a point (area) nearby that was forecast to remain dry.

While the predictands are defined on points and on circles with a maximum radius of 25 km, the potential predictors are constructed on circular neighborhoods around the central SYNOP stations of radii up to 100 km. They can be subdivided in the following groups:

- \( q \): the exceedance threshold \( q \);
- \( \text{dmax} \): precipitation as given by the model at the closest grid point to the station (direct model output);
- \( \text{dist} \): predictors based on the (weighted) distance from the station to the closest wet or dry grid point;

and within a disk of radius \( R \) around a station, with \( R = 25, 50, 75, \) and 100 km:

- \( \text{max} \): the (square root of the) maximum areal precipitation;
- \( \text{coverage} \): predictors based on the coverage of precipitation (i.e., the percentage of the area covered by precipitation);
- \( \text{total} \): the (square root of the) areal mean precipitation;
- \( \text{weightmax} \): predictors based on the maximum precipitation weighted by the distance to the station; and
- \( \text{weightint} \): predictors based on the precipitation integrated over the area, weighted by an exponentially decaying function from the station.

In the appendix the definition of the predictors is given. Note that similar predictors were defined in Kok et al. (2008) and Scheuerer (2014). Figure 2 shows an example of a radar image (left) and a Harmonie-AROME forecast (right) of 3-h accumulated precipitation between 1200 and 1500 UTC 5 December 2013. Also depicted are the circles of radii 25–100 km, and the locations of the maximum precipitation and nearest dry point.

The total SYNOP dataset of the warm seasons consists of 14 months of data for 7 stations and 7 lead times, which is in the order of 20,580 cases, of which approximately 570 cases have precipitation exceeding 3 mm (3 h)−1. Because of highly correlated potential predictors special care should be taken to prevent overfitting. This was the main reason for sorting the potential predictors into groups. All model predictors (except direct model output and distance to nearest dry/wet point) are defined as a function of the radius of the disk around the point of interest. Furthermore, several predictors are clustered in the sense that they represent a quantity that is weighed with different monotonically decreasing functions of the distance to the station (using linear or exponential decay). Only one predictor per group is allowed in the final regression equation to mitigate overfitting problems.

6. Results

In Fig. 3 the observation and model climatologies are shown for the three types of predictands that are considered in this paper. The histograms of the precipitation
distribution for the seven stations combined show that especially HIRLAM and ECMWF have a tendency to forecast too much precipitation up to 2 mm (3 h$^{-1}$). The panel in the inset shows the same data [including the fraction of the cases with precipitation less than 0.1 mm (3 h$^{-1}$)] on a logarithmic scale. For SYNOP observations we see that the Harmonie-AROME model is capable of forecasting larger amounts of precipitation for 3-h intervals than the HIRLAM and ECMWF models, in agreement with the rain gauge climatology. The same picture is visible in the bottom-right panel of Fig. 3 for the higher precipitation maxima within a 25-km radius disk, whereas the HIRLAM and ECMWF model climatologies are closer to the observed climatology for the lower precipitation maxima. The three models show mean precipitation values >5 mm (3 h$^{-1}$) on the 25-km radius disk similar to the radar precipitation observations (bottom-left panel of Fig. 3), while the Harmonie-AROME climatology is closest to the observed climatology for the lower mean precipitation values.

The procedure described in section 4 yields a regression equation for each NWP model, analysis time and lead time. The ranked probability skill score (RPSS; e.g., Wilks 2011) with respect to the sample climatology is computed using the thresholds that were used to derive the ELR equation.

When we look at the RPSS as a function of the number of predictors used (e.g., see Fig. 4 for SYNOP observations as predictand, and one combination of analysis and lead time), we generally see that for both the training set and the verification set the skill saturates after selecting three to four predictors (including the threshold).

The mean RPSS of the models on the verification set is lower, as expected, and for this lead time postprocessed Harmonie-AROME shows the highest RPSS. In this example the distribution of the RPSS is broader for the independent set, and postprocessed Harmonie-AROME and the ECMWF IFS have the narrowest distribution.

a. Selected predictors

In principle, each model may use very different predictors to achieve the best correlation or highest skill given the restrictions discussed in section 5. The results for most of the predictands introduced in section 3 are very similar, except for the maximum precipitation in a circle of 25 km.

When we look at which predictor was selected first for the different models (see Fig. 5 for precipitation measured by SYNOP stations as predictand), we see that in general the large-scale predictors, such as the square root of the mean areal precipitation from the total group (see the appendix) and coverage [the fraction of the circular area where the precipitation exceeds 0.3 mm (3 h$^{-1}$)] are preferred. The threshold is always selected as the second predictor. A predictor of the dist group is often picked third, indicating the (relative) importance of the distance of the predicted precipitation to the SYNOP station. The longest lead time that was studied here shows a slightly different pattern, where instead of the total predictor (i.e., the mean amount of precipitation that is forecast in a certain area) the
coverage and the dist predictors [i.e., areal coverage and location of precipitation >0.3 mm (3h)^{-1}, respectively] are chosen more often. This may be related to the forecast mean precipitation amount being less informative than the precipitation coverage for longer lead times. However, for the maximum within a 25-km radius disk as a predictand the total predictor remains the most important predictor for all lead times, except for HIRLAM (not shown).

Also, it is noteworthy that the predictor based on the direct model output (dmo) is never selected before the stopping criterion is met. Apparently, it contains no skillful information in addition to the selected predictors. This is in complete agreement with the fact that high-resolution forecasts should not be taken at face value but should be interpreted probabilistically.

In Fig. 6 the distribution of the scales of the selected predictors is given for the same four lead times and predictand as in Fig. 5. As expected, there is a general tendency to select predictors on larger spatial scales with increasing lead time. It is remarkable that even for SYNOP observations as a predictand source and especially for Harmonie-AROME, predictors defined on disks with a radius of 25 and 50 km are hardly selected.
b. Verification results

The exceedance probabilities calculated for the independent data are verified with the Brier skill score (BSS) (Brier 1950) with respect to the sample climatology for the following distinct thresholds: 0.1, 0.3, 1, 2, 3, and 5 mm (3 h)$^{-1}$. These results are shown in the following subsections.

1) RAIN GAUGE OBSERVATIONS

The most localized predictand source discussed here is the SYNOP precipitation observation by the rain gauge. Especially short-lived or localized convective events are likely to be missed by both the rain gauge and the direct model output alike, making it difficult to interpret the associated skill scores. When we look at the BSS (with respect to the sample climatology) as a function of the exceedance threshold (Fig. 7a), we see that there is hardly any difference in skill between the postprocessed HIRLAM, Harmonie-AROME, and ECMWF model. In the boxplots every box represents the distribution of BSS values from each of the 30 different test sets for all analysis and lead times combined. There is no skill for thresholds $\geq$5 mm (3 h)$^{-1}$.

To test the statistical significance of the difference between the BSS values of the three models, we have used only three nonoverlapping verification sets per analysis time and lead time to minimize the dependence of these datasets. In this way we have a set that uses days 1–20 of each month within the 14-month dataset as the training set and days 21–30 as the verification set, another set that uses days 11–30 as training set and days 1–10 as verification set, and a set that uses days 21–30 and days 1–10 as training set and days 11–20 as the verification set. Subsequently, we have performed paired $t$ tests on the 42 BSS values for two of the models at a time, following Hamill (1999). We can conclude that the BSS differences between the models are generally not statistically significant at the 0.01 level.

From Fig. 8, which shows the BSS as a function of the exceedance threshold for the same four lead times as in Figs. 5 and 6, it seems that postprocessed Harmonie-AROME performs better than the two other postprocessed models for the 1800 UTC + 6 h and 1800 UTC + 12 h lead times, especially for the higher thresholds, whereas postprocessed ECMWF seems to perform best for the last lead time (1200 UTC + 30 h). However, by determining the statistical significance in
the same way as described above, most BSS differences are not statistically significant at the 0.01 level.

It is remarkable that the higher BSS of postprocessed Harmonie-AROME in the top panel of Fig. 8 is a result of predictors on only larger scales, whereas in the ECMWF and HIRLAM ELR equations predictors on smaller scales are selected as well (Fig. 6, top panel).

To illustrate how the postprocessing adds skill with respect to the direct model output, the direct model output was verified in the same way but merely using the binary DMO exceedances of the thresholds (top panel of Fig. 9). For low precipitation intensities \([<1 \text{ mm (3 h)}^{-1}]\) HIRLAM performs clearly worse, which can be partly explained by its bias for these thresholds as can be deduced from Fig. 3.

If we apply the postprocessing method using ELR as before, but only include the direct model output \((\text{dmo})\) and the threshold as predictors, we get substantially higher BSS values, as can be seen in the bottom panel of Fig. 9. However, these BSS values are still significantly lower than those from the fully postprocessed version (Fig. 7a), indicating the importance of neighborhood information.

2) AREAL MEAN PRECIPITATION

Forecasting the mean over an area is easier than for a single point, as local effects are evened out more with increasing radius of the measurement area. For the mean 3-h accumulated precipitation over a circular area with a radius of 5 and 25 km around the stations, there are only small differences in Brier skill scores between the three postprocessed models (Figs. 7b and 7c, respectively), when the analysis and lead times are combined. When we look at the BSS as a function of the exceedance threshold for the same four lead times as in Fig. 8, we see a very similar pattern (not shown) to the one in that figure. Also, the choice and scale of the predictors is very similar to those selected for the SYNOP measurements (cf. Figs. 5 and 6), with the same tendency to switch from quantity-based predictors \((\text{total})\) to area- and location-based predictors (i.e., \text{coverage} and \text{dist}) for longer lead times.

3) AREAL MAXIMUM PRECIPITATION

Because of the higher resolution and nonhydrostatic dynamics of Harmonie-AROME, one might expect...
that the model is better capable of forecasting more extreme events, yielding large amounts of rain in smaller areas. However, for 3-h accumulations this rain will generally still be dispersed over an area as the storm passes over it.

When we look at the maximum 3-h accumulated precipitation within the circular areas of 5 and 25 km as a predictand (Figs. 7d and 7e), we see that the maxima for the larger areas (with a radius of 25 km) show skill up to higher thresholds than for the smaller areas (with a radius of 5 km). This can be expected, because it is harder to forecast probabilities of larger precipitation amounts within a smaller area. The maximum within an area as forecast by the NWP models is rarely selected as a predictor, even for the nonhydrostatic Harmonie-AROME model.

Regarding the relative skill between the models, we notice that the postprocessed Harmonie-AROME precipitation is better than the postprocessed HIRLAM precipitation for the higher thresholds \( \geq 3 \, \text{mm/hr} \), being statistically significant at the 0.01 level in the paired t test. For the BSS as a function of the exceedance threshold for the same four lead times as in Fig. 8, very similar results to those in that figure are obtained again (not shown).

7. Qualitative comparison with fractions skill score

Another method to circumvent the double-penalty problem is by using the FSS (Roberts and Lean 2008; Skok 2015). The same 14-month period was processed using the 3-hourly accumulated precipitation output from the 3 models and the 3-hourly accumulated calibrated radar data. The method compares a forecast and an observation field on a common grid by considering a region or neighborhood around each grid point and computing fractions of grid points inside this region exceeding a threshold, giving a metric for the error. The SpatialVx R package of Gilleland (2015) has been used, where FSS defaults to 0 when no rain is observed. To be able to compare the FSS for the different models and resolutions, all forecasts were regridded to the highest-resolution (2.5 km) grid by taking the nearest neighbor value.

The procedure outlined in this paper can be qualitatively compared to the FSS when only taking the coverage into account, and looking at only a few grid points instead of all. For instance, a neighborhood size of 3 grid points \((7.5 \times 7.5 \, \text{km}^2)\) would be similar to a circular area with a radius of 5 km \((\approx 79 \, \text{km}^2)\), and in this way one might compare the FSS for a neighborhood size.
FIG. 7. Brier skill scores for the three postprocessed NWP models for three predictand types: (a) SYNOP observations; (b), (c) radar mean precipitation in circles with a radius of 5 and 25 km; and (d), (e) radar maximum precipitation within circles with a radius of 5 and 25 km combined for all different test periods, analysis and lead times, as a function of the threshold, and tested on independent data. Boxes represent the 25th–75th percentile range or IQR with a line indicating the median. The whiskers (dashed lines) represent the 25th percentile minus 1.5 times the IQR and the 75th percentile plus 1.5 times the IQR, extending to the most extreme data point in this data range. The plus signs indicate data outside these ranges.
of 3 with the BSS for the mean precipitation within that circular area, using only the coverage as a predictor. An important difference, however, is that the FSS usually is a monotonically increasing function of the neighborhood size, asymptoting to a value that is a function of the bias. A lower limit of a useful spatial scale is defined (Roberts and Lean 2008) as the scale where $\text{FSS} = 0.5 + f_o/2$, with $f_o$ the fraction of observed points in the domain exceeding the threshold. The predictor method on the other hand may select a scale for which a model offers an optimal amount of information pertaining to the probability of an event occurring or not.
The results presented in Fig. 6 suggest that this scale is at least 75–100 km (note that larger scales have not been investigated), but especially for short lead times features at smaller scales are also important.

A difficulty in computing the FSS for an extended period of time is how to deal with dry days. When both forecast and observation do not have rain inside the area covered by both the radar and the model, the score will default to zero. One way to overcome this is to apply a threshold to the base rate, and only consider the cases when the base rate exceeds this threshold. The choice of this threshold has a very clear impact on the absolute values of the FSS, as can be seen in Fig. 10.

For all analysis and lead times combined, the Harmonie-AROME model appears to perform somewhat better than HIRLAM and the ECMWF model.

**Fig. 9.** As in Fig. 7, but (top) using the direct model output verified against SYNOP observations and (bottom) postprocessed direct model output verified against SYNOP observations. Box-and-whisker plots as in Fig. 7.
with respect to this metric, both for areas larger than 15 by 15 grid points and thresholds above 0.3 mm (3 h)$^{-1}$. The average useful scale that may be inferred from Fig. 10 (using the 0.05 base rate) is at best in the order of 100 km for Harmonie-AROME, whereas for the other two models the useful scale seems much larger.

**8. Discussion**

Even though precipitation forecasts of a high-resolution nonhydrostatic model produce very realistic precipitation patterns, it remains challenging to show better skill in a quantitative manner. In this comparative verification study we have not focused on the direct output of the respective models valid for point or area locations of interest, but instead we have tried to extract, using model output statistics, as much skillful information as possible from the precipitation forecasts from a large neighborhood around the observation location, using a dataset of 14 warm months between August 2012 and July 2014. The most important forecast features appeared to be aggregated, area-averaged, quantities like the mean precipitation and the so-called coverage (i.e., the percentage of the area covered by precipitation) over circular areas that are increasing with lead time. Predictors with a higher variance like the

![Fig. 10](image-url)
proximity of the forecast precipitation or the maximum within the circular areas offered only slight additional information depending on the observation type. The fact that the direct model output is never selected as a predictor is another confirmation that the use of probabilistic tools to interpret (deterministic) precipitation forecasts is essential.

When looking at different aspects of a precipitation forecast, such as the capability to estimate the probability of (the mean or maximum) precipitation exceeding some threshold in a point (an area), overall there appears to be only relatively small differences in the Brier skill scores between postprocessed Harmonie-AROME, HIRLAM, and the ECMWF model precipitation output using ELR. The differences for individual lead times are larger, but not statistically significant at the 0.01 level in a paired t test. A different choice of thresholds to train the ELR equations has also been used, but the thresholds cannot be chosen over a much larger range, as for such a relatively short dataset the higher thresholds are so rarely exceeded. This resulted in very similar results to those presented in this paper.

Using the fractions skill score (FSS) as a metric, the Harmonie-AROME model appears to perform somewhat better than HIRLAM and the ECMWF model using the same 14-month dataset. However, there are some notable differences between the FSS and our method. Our method essentially looks at all skillful properties that can be obtained or derived from (in our case) the forecast precipitation field that together yield the highest predictive potential for the predictand at hand, whereas the FSS considers only one property, the coverage, which in many cases is not the best indicator of skill, as our results show. In other words, the method tries to verify the total information content rather than verifying the degree in which the observed and predicted fractions agree. The method can easily be extended by including other forecast attributes (e.g., derived from the higher model levels) (like instability indices such as CAPE). It does not suffer from bias [as is also the case for the FSS when frequency thresholds are used; Roberts and Lean (2008)], all wet and dry cases are included in the verification and our method can be applied to predictands of all scales. However, a drawback of our method is the requirement of a rather long training period in order to assess which features at which scale are the most relevant ones. Furthermore, the assessment can only be made on the features and scales that are included in the analysis.

Although the FSS yields the smallest scale over which the model has useful (or acceptable) skill, an upper limit is not easy to define (Mittermaier et al. 2013). Our method provides an estimate of the “optimal” scale that contains all relevant information for a specific predictand. Skillful features with smaller scale than this optimal scale are explicitly accounted for in the verification, while larger scales do not provide a meaningful addition. The procedure does not involve a qualification about the usefulness of the resulting scales. The relation between usefulness or value and the accuracy (skill) is a complicated one (e.g., Murphy and Ehrendorfer 1987) and highly depends on the forecast application.

The advantage that high-resolution models may have in representing very intense rainfall events is not clear from the data presented in this paper. For thresholds above 5 mm (3 h)\(^{-1}\) none of the models show any skill, apart from the maximum precipitation in a circle with 25-km radius. This may be due to the limited amount of cases. When we would verify the model data for the whole Harmonie-AROME domain (789 × 789 points on a 2.5 km × 2.5 km grid, covering a considerable part of western Europe including the Alps), using SYNOP observations, more extreme events exceeding 5 mm (3 h)\(^{-1}\) could be included in the derivation of the ELR equations. It is expected that Harmonie-AROME would show higher skill in mountainous areas like the Alps, as case studies for other high-resolution models suggest (e.g., Colle et al. 2005).

Messner et al. (2014) introduced heteroscedastic extended logistic regression, where the spread of an ensemble was used to improve the calibration of the ensemble, using that information to predict the width of the calibrated distribution. We applied this method in a preliminary experiment, using the variance of the precipitation forecast within different circles around the stations as a measure of the forecast uncertainty. In principle, a forecast employing a higher resolution could give a better indication of the uncertainty, especially for convective precipitation, even if it misses the storm at the observation location. However, this did not change the conclusions: the skill using heteroscedastic extended logistic regression hardly improved, and the relative skill between the models did not change as well.

A more detailed study is needed to assess in which circumstances the benefit of the more realistically looking precipitation forecast of the high-resolution Harmonie-AROME model is obscured by forecasts that have less skill. One possible way of doing that is to stratify the data between convective and large-scale precipitation cases on the basis of the predicted synoptic situation instead of stratifying between a warm and cold season set. In this way the focus is much more on the convective cases, leading to conditional exceedance probabilities, conditional on the predicted convection through some criterion. An additional advantage is that
the relative frequency of severe precipitation events is much higher than in the corresponding larger set, which means that MOS is presumably capable of generating skillful equations for higher precipitation thresholds than the ones we have presented.

While the double-penalty problem for higher-resolution forecasts has been greatly reduced in the current method, there are also other factors that could result in similar skill of the higher-resolution models in comparison with the ECMWF model. These factors include the type and settings of the data assimilation scheme, the choice of background/lateral boundary conditions for the regional models, and the parameterization schemes. Although it is outside the scope of this paper, it would be interesting to investigate how much of an impact these other factors have.

Acknowledgments. The authors thank the anonymous reviewers for their comments, which have helped to improve earlier versions of the manuscript to a large extent. Seijo Kruizinga and Adri Buishand are acknowledged for their advice on statistical significance testing. MET and SpatialVx were developed at the National Center for Atmospheric Research (NCAR) through a grant from the U.S. Air Force Weather Agency (AFWA).

APPENDIX

Predictors

To predict the probability of precipitation at a point $x_i$ above threshold $q$ for different analysis and lead times, a set of predictors is devised. For the computation of the predictors such as coverage, a threshold $q_p = 0.3 \text{ mm (3 h)}^{-1}$ is used. Note that this is a different threshold than the one for which the regression equation is constructed, because a lower precipitation intensity may have a certain predictive value for the occurrence of precipitation of higher intensity. In this experiment a constant $q_p$ is considered, but one might easily conceive the threshold to be a function of the predictand threshold $q$ (e.g., $q_p = \sqrt{q}$). The following predictors are defined, where applicable in a disk $D$ around $x$, with radius $d$ of 25, 50, 75, or 100 km:

1) Threshold $q$, treated as a predictor in extended logistic regression.
2) Direct model output ($dmo$) at $x_i$, or $pcp(x_i)$ (3-h accumulated precipitation). This is retrieved using bilinear interpolation.
3) The group $dist$, which is either the closest dry grid point [if $pcp(x_i) > q_p$], {i.e., $\min[|x| \text{where } pcp < q_p] - x_i |]$}, or the closest wet grid point [if $pcp(x_i) < q_p$], {i.e., $\min[|x| \text{where } pcp > q_p] - x_i |]$}.
4) The (square root of the) maximum precipitation within the disk (max),
$$\max_{x \in D} pcp(x), \max_{x \in D} \sqrt{pcp(x)}.$$ 5) Coverage (coverage), defined as
$$\int_D H[pcp(x) > q_p] d^2x,$$
with $H(x)$ the Heaviside function: $H(x < 0) = 0$, $H(x > 0) = 1$.
6) Average precipitation in the area (total),
$$\int_D pcp(x) d^2x \int_D d^2x,$$
and the square root of this quantity (sqrttotal).
7) A group of predictors using (the square root of) the maximum precipitation $pcp(x)$ within the disk $D$ with different weighting expressions of the distance of the maximum to the station (weightmax):
   • Linear weighting function ($linmaxnorm$ and $linmaxsqrt$), using
$$1 - \frac{|x - x_i|}{d},$$
   • Exponential weighting function ($expmaxnorm$ and $expmaxsqrt$), using
$$\exp\left(\frac{|x - x_i|}{d}\right).$$
   • Exponential weighting function with slower decay ($exp2maxnorm$ and $exp2maxsqrt$), using
$$\exp\left(-\frac{|x - x_i|}{2d}\right).$$
8) A group of predictors that weigh the integrated amount of precipitation within the disk $D$ around the point (weightint)
   • Linear integral weight of precipitation within disk ($lin_int$),
$$\int_D pcp(x) \left(1 - \frac{|x - x_i|}{d}\right) d^2x.$$
   • Exponential integral weight of precipitation within disk ($exp_int$),
\[ \int_D p(x) \exp \left( -\frac{\|x - x_i\|^2}{d^2} \right) \, d^2x. \]

- Exponential integral weight of precipitation within disk (with slower decay as function of distance) 
  \( \exp_{2\text{int}} \),

\[ \int_D p(x) \exp \left( -\frac{\|x - x_i\|^2}{2d^2} \right) \, d^2x. \]

REFERENCES


