Comparative verification of nowcasting and numerical weather prediction using spatial verification methods as part of the SINFONY project at DWD

Gregor Pante and Michael Hoff

German Weather Service - Research & Development

Report on sub-task 3.6 of COSMO Priority Project AWARE (Appraisal of Challenging WeAther FoREcasts)

Contents

1 Introduction

Germany is exposed to various kinds of high impact weather phenomena. Strong impacts are expected from convective events during summer which happen to be especially hard to predict. The Seamless Integrated Forecasting System (SINFONY) project at DWD focuses on such events, which mostly take place on the kilometer scale. One aim of the project is therefore the development, adaptation, and operationalization of innovative, spatially based verification methods of the entire process chain of the integrated forecasting system consisting of data assimilation, nowcasting and numerical short-term prediction. The advantage of spatially based verification methods is that exact matching of forecasts and observation no longer needs to prevail to obtain good scores because these methods circumvent the "double penalty" problem, i.e. a miss due to a displaced observation event and a false alarm due to a displaced forecast event.

Following Gilleland et al. (2009) there exist mainly four categories of spatial verification – neighborhood (or fuzzy) and scale-separation basically applying filtering methods, as well as feature (or object) based and field deformation basically yielding information about displacements. In the SINFONY project, we decided to apply neighborhood as well as object-based verification methods. Both methods are well established and cover a huge amount of information which is helpful for model development, user interpretation and many more.

The neighborhood (also known as fuzzy) approaches compare values of forecasts and observations in space–time neighborhoods relative to a point in the observation field. Properties of the fields within neighborhoods (e.g., mean, maximum, existence of one or more points exceeding a certain threshold) are then compared using various statistical summaries, which are often simply the traditional verification statistics. Such comparisons are typically done for incrementally larger neighborhoods so that it is possible to determine the scale at which a desired level of skill is attained by the forecast (Gilleland et al., 2009). The neighborhood methods apply a smoothed filter on the original field(s). Summary statistics, such as traditional verification statistics, can be applied to the smoothed field. The process is typically repeated using increasingly larger neighborhoods. The most established neighborhood method is called Fractions-Skill-Score developed by Roberts and Lean (2008).

Of particular interest, especially in SINFONY, are object-based methods which require a threshold-linked object identification algorithm. It is applied to pixel-based forecast and observation fields of radar reflectivity. The resulting objects contain certain attributes regarding their geometry (e.g., centroid, area), intensity (e.g., min, max), and forecast information (e.g., trajectory). In SINFONY, we focus on the object-based evaluation metric called median of maximum interest (MMI) after Davis et al. (2009) to assess the quality of the predicted precipitation objects. The object-based evaluation is extended to cope with ensemble forecasts. Besides basic single member verification a new technique to define a so called "pseudomember" (Johnson et al., 2020, J20 hereafter) is analyzed. The pseudomember comprises a reasonable and representative selection of objects from all ensemble members that have locally the highest probability of occurrence.

2 Data and methodology

2.1 Data sets

In SINFONY will be a variety of data available for verification. The case study period of the underlying data will be mentioned in the respective section.

2.1.1 Grid-based data

For numerical weather prediction, NWP, the underlying model is the regional ICON-D2-EPS in a quasi-operational setup since 2019. Before 2019, we were using COSMO-DE-EPS in a quasi-operational setup. The EMVORADO operator (Zeng et al., 2016) simulates synthetic radar reflectivities for each of the 17 polarimetric Doppler-C-Band radar systems in the DWD radar composite. Subsequently, the model volume scans will be processed by POLARA and mapped onto a Radolan grid with a horizontal resolution of 1 km. We are using 40 members for data assimilation and 20 members for the forecast of up to 8 hours.

For nowcasting, we are using STEPS DWD with a localization filtering approach (planned to submit in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing) to generate an ensemble with 30 member (20 member used for verification) with a nowcasting time of two (or four) hours. All nowcasting data are on the 1 km horizontal resolution Radolan grid.

2.1.2 object-based data

For object identification, we are using our in-house product KONRAD3D. With the help of adaptive thresholding schemes and other filtering methods, which will not be specified in this report, KONRAD3D identifies cell objects in each radar volume scan. The default basic threshold for object identification is 35 dBz whereas a subcell within such regions must fulfill a minimum-maximum difference to the basic threshold of at least 7 dBz. This means that cells mostly obtain a minimum value of 42 dBz. By optimized combination of objects in each radar volume scan, three-dimensional objects will be built taking into account the entire DWD radar network.

KONRAD3D is used for object nowcasting as well. Currently in development is an ensemble based object nowcasting which will unfortunately not be available for the current study. However, since EMVORADO simulates reflectivity for all radars and respective volume scans, KONRAD3D can be applied to the NWP forecasts, described in the previous section. Therefore, we can fall back on an 20 member ensemble object-based NWP forecast of 8 hours with temporal resolution of five minutes to test our object-based verification methods. Further, a comparison of 1-moment vs. 2-moment microphysics scheme will be done. As the latter is able to produce higher reflectivities, it is expected to better capture extreme events.

Combined product Seamless combination of nowcast and model forecast, grid-based and object-based.

2.2 Spatial verification methods

2.2.1 Neighborhood-based methods

We apply mostly well-known neighborhood-based verification methods to our data. The most established method is the fractions skill score, FSS, (NO-NF) by Roberts and Lean (2008). Further, we implemented the minimum coverage method (NO-NF), Fuzzy-logic (NO-NF), fuzzy-logic with joint probabilities (NO-NF), multi-event contingency table (SO-NF) and pragmatic approach (SO-NF) for which the reader is referred to as Ebert (2008). All necessary information about the underlying methods can be found in this publication. Since the above mentioned methods for building a contingency table from neighborhood probabilities have weaknesses in their bias behavior, we implemented the neighborhood-based contingency table including errors compensation by Stein and Stoop (2018). This method uses a practical approach in which it the same number of misses and false alarms in a certain neighborhood compensate each other to hits and correct negatives, i.e. it is a correct forecast in the respective neighborhood. A positive side effect of this method is that the frequency bias is independent of the neighborhood size (small deviations on domain edges are possible), which makes it quite practical using it for verification.

Another useful method we implemented is the displacement estimation of precipitation fields based on fractions skill score by Skok and Roberts (2018). The authors used the FSS = 0.5 threshold for a useful forecast to estimate a global distance metric. The results are quite promising even though the method is not applicable for frequency biases larger than two and lower then 0.5. Also for frequency biases larger than 1.5 (< 0.75) the method exhibits shortcomings. However, for the remaining data, the displacement metric is a useful information apart from the classical categorical verification metrics. To go one step further, G. Skok presented a new metric called displacement from NSS (neighborhood skill score) at 2020 International Verification Methods Workshop. Further, Skok showed that the results are closer to the real displacement and also in more realistic cases the score showed more reliable results. The Displacement-NSS is no more limited to small biases which makes it quite useful for application in our SINFONY project. Therefore, with the help of G. Skok, we implemented this metric as well. However, the deviation of the NSS displacement from real displacement becomes larger the closer precipitation objects are to the domain edges. Up to now, we did not correct this fact in our verification analyses but postpone it to future work.

Another useful method, we implemented, is Neighborhood-Ensemble-Probabilities (NEP) proposed by Schwartz et al. (2010). Here, the thresholding, neighborhood-smoothing (for different box lengths) will be done for all M ensemble member separately. Finally, the resulting M neighborhood probabilities will be averaged to obtain NEP. On the NEP field, all above described methods can be applied, however, not all methods will give benefits for using NEP. The most reliable method in combination with NEP is FSS. The NEP is most beneficial for smaller neighborhood sizes around a certain point of interest. For larger neighborhoods, the effect will be smoothed out or the areas of precipitation probabilities become to large in comparison with the observation.

Finally, we implemented reliability and ROC diagrams for analyzing our grid-based deterministic and ensemble data. As reference, however, we made a compromise and took only binary observation into account, since otherwise the huge quantity of verification data is not manageable in an operational framework.

Further implementations are planned for the future.

2.2.2 Object-based methods

Total Interest and Median of Maximum Interest

The TI (Davis et al., 2009) is a measure for the similarity of two objects with respect to the objects' attributes. For each selected attribute i of an object pair j a "fuzzy logic function" (F) is defined in order to transform the value of i into the interval [0,1]. For example, the function of the centroid distance (F_{CD}) – one attribute of an object pair – is defined to be equal 1 if CD is less than 10 km, then linearly decreases with increasing CD and equals 0 for CD larger than 100 km. The F-values of different attributes result in the "interest" (I) by multiplying with weights (w) and confidence factors (c). The TI of an object pair j, finally, is the weighted sum of all I-values of all considered attributes i (Davis et al., 2009):

$$TI_j = \frac{\sum_i I_{ij}}{\sum_i w_i c_i} \quad (1)$$

$$I_{ij} = F_{ij} w_i c_i \quad (2)$$

Attribute	w, %	c	f_{min}	f_{max}
Centroid distance	28	Area ratio	10 km	100 km
Minimum boundary distance	40	1	5 km	50 km
Area ratio	19	1	0.0	0.8
Intersection area ratio	13	1	0.0	0.25

Table 1: Attributes and parameters used to calculate the total interest TI. f_{min} and f_{max} are the lower and upper limits below and above which the fuzzy logic function of the respective attribute takes its minimum, respectively maximum value.

In the presented analysis we employ the settings as described in Davis et al. (2009) and listed in Table 1 to calculate the TI. Having one set of observed and one set of predicted objects, the TI-values of all possible object pairs are calculated. They fill the so called TI matrix which contains all observed objects as columns and all predicted objects as rows. The next step selects the maximum values along each row (column) and adds them as a new column (row) at the right (bottom) of the TI matrix. The median over all these maximum values builds the final score for the object-based ensemble verification, i.e., the median of maximum interest MMI.

Ensemble forecasts

The object-based evaluation of ensemble forecasts is one major challenge in the verification for two reasons. First, the amount of objects to be processed can be very large depending on the weather situation and number of ensemble members. And second, new methods must be developed to reveal a fair score. Two rather simple ideas are the verification of the objects from each single ensemble member separately or of the merged set of all objects from all members. The first one yields simply the quality of each member and can additionally provide information about the spread of the ensemble. The second one is very likely to generate so called "over-forecasting", i.e., the combined set of objects comprises much more objects than the observation which may generate many false alarms. Therefore, a third method is analyzed in which a reasonable selection of objects is chosen to build the so called "pseudomember" which comprises the objects from all ensemble members that are locally the most representative ones of the ensemble distribution (J20).

The selection of objects for the pseudomember follows five steps as described in J20:

1. *"Make a list of all objects in the forecast ensemble, together with the objects' probabilities, calculated from the percentage of ensemble members with a matching (i.e., total interest > 0.2) object.*

2. *Sort all of the objects by probability, breaking ties according to the average total interest with all the objects from other ensemble members that it matched to.*

3. *Add the highest probability object to the object list of the pseudomember.*

4. *Remove from consideration the added object, as well as all matching objects in other members that contributed to the probability of the added object, leaving a new, smaller list of objects.*

5. *Repeat from step 2 until no objects remain in the list of ensemble forecast objects."*

Here these steps are performed for constructing the pseudomember but another matching criterion (first step of J20) was used. For the comparison of one specific object from one member with all objects from all other members, the TI of this specific object with all other objects is calculated as

$$TI = \frac{2 \cdot F_{CD} + F_{AR}}{3} \tag{3}$$

where F_{CD} and F_{AR} are the interest functions of centroid distance (CD) and area ratio (AR). These functions are defined as

$$F_{CD} = \begin{cases} 1 & CD < CD_1 \\ \frac{1}{2} \cdot \left[\cos\left(\frac{CD - CD_1}{CD_2 - CD_1} \cdot \pi\right) + 1\right] & CD_1 \leq CD \leq CD_2 \\ 0 & CD > CD_2 \end{cases} \tag{4}$$

$$F_{AR} = \begin{cases} 0 & AR < AR_1 \\ \frac{1}{2} \cdot \left[\sin\left(\frac{AR - AR_1}{AR_2 - AR_1} \cdot \pi - \frac{\pi}{2}\right) + 1\right] & AR_1 \leq AR \leq AR_2 \\ 1 & AR > AR_2 \end{cases} \tag{5}$$

Below CD_1 and AR_1 and above CD_2 and AR_2, which are set to $CD_1 = 10$ km, $CD_2 = 70$ km, $AR_1 = 0$, and $AR_2 = 0.8$, the interest functions take their minimum (0), respectively, maximum (1) values. For object pairs to be defined a match the TI must exceed a value of 0.7. This criterion limits the ranges of CD and AR within which matches are possible. Hence, if CD is larger than 38 km no match can occur even if AR was perfect while no matches occur for AR below 0.16 even if CD was perfect.

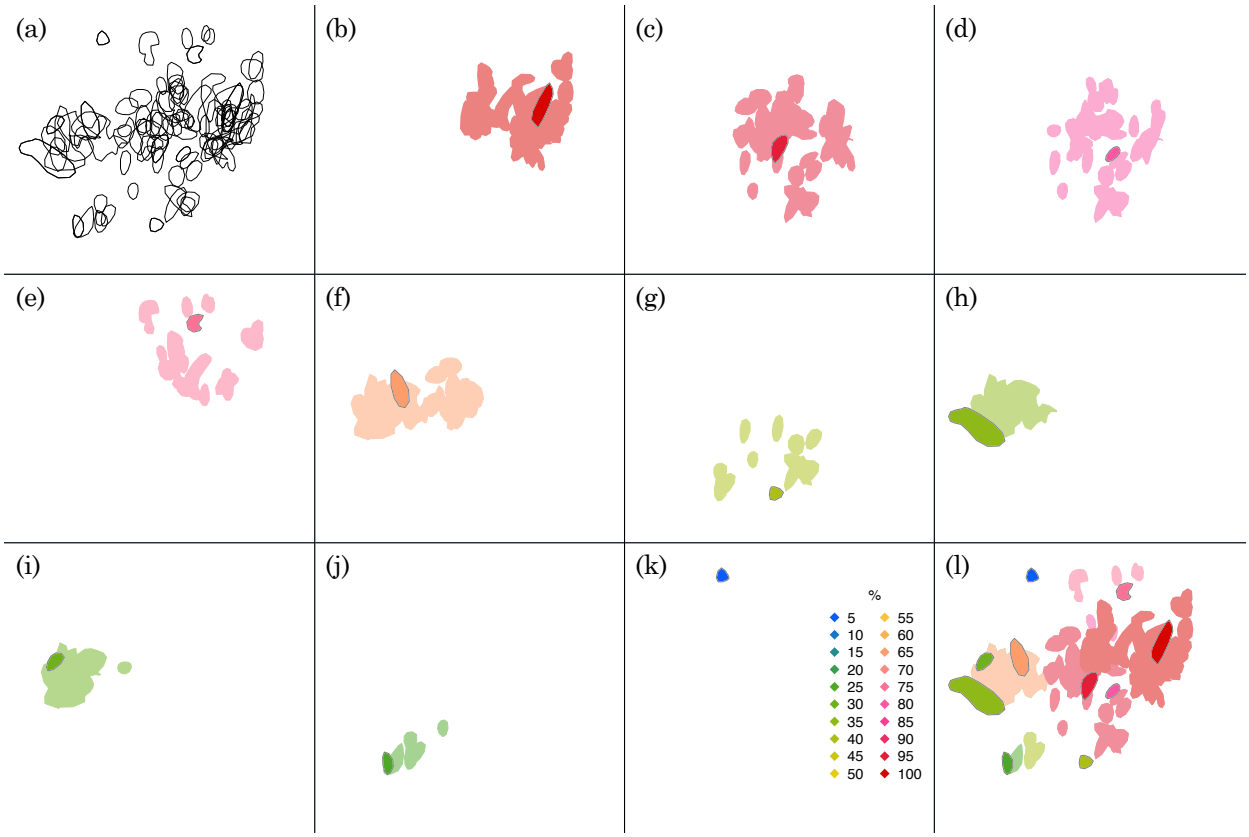

Figure 1: Procedure of selecting the objects of the pseudomember for a forecast initialized on 30 May 2016 12 UTC with a lead time of 3 hours. All objects of all ensemble members in a given region are shown in panel (a). Panels (b)–(k) depict the single pseudomember objects (gray bordered polygons) according to their probability of occurrence (colors). The lighter colors around these objects mark the uncertainty regions, i.e., the unified area of all objects from other members that were defined a "match" with the respective pseudomember object. The combined result with all pseudomember objects is given in panel (l). Colored areas of the uncertainty regions are stacked on top of each other with increasing probability, hence, regions with low probability can be covered by those with higher probabilities.

Figure 1 illustrates the procedure of selecting the pseudomember objects following the steps described above. Technically, the pseudomember is a list of polygons, i.e., the selected objects, together with their probabilities of occurrence and uncertainty regions. The probability of occurrence p (color scale in Fig. 1) is the percentage of ensemble members with at least one matching object. The member of the object in consideration itself is counted as well, hence, for a 20 member ensemble, as used in this study, p varies in 5% steps between 5% and 100%, where 5% means no other member has a matching object and 100% all other members have at least one matching object. If a member has more than one matching object all these objects are removed from further consideration (step 4 in the description above) but this member still counts as only one member with regard to the probability. The uncertainty region of a pseudomember object is the unified area covered by all the matching objects from other members (light colors in Fig. 1). In the example one object has matching objects in all other ensemble members and gets a value of $p = 100\%$ (Fig. 1b). The probability of the subsequently selected objects decreases until only one object remains which has no matching objects in other members and therefore $p = 5\%$ is assigned to this object (Fig. 1k).

3 Results

The different spatially based verification methods described before are applied to predictions from the SINFONY reference period between 27 May – 25 June 2016. This early summer period is characterized by almost daily strong convective activity over Germany. Unfortunately, only COSMO-DE-EPS runs are available for this time period.

3.1 Neighborhood-based methods

In this section, we show some representative results from the SINFONY reference period in May/ June 2016.

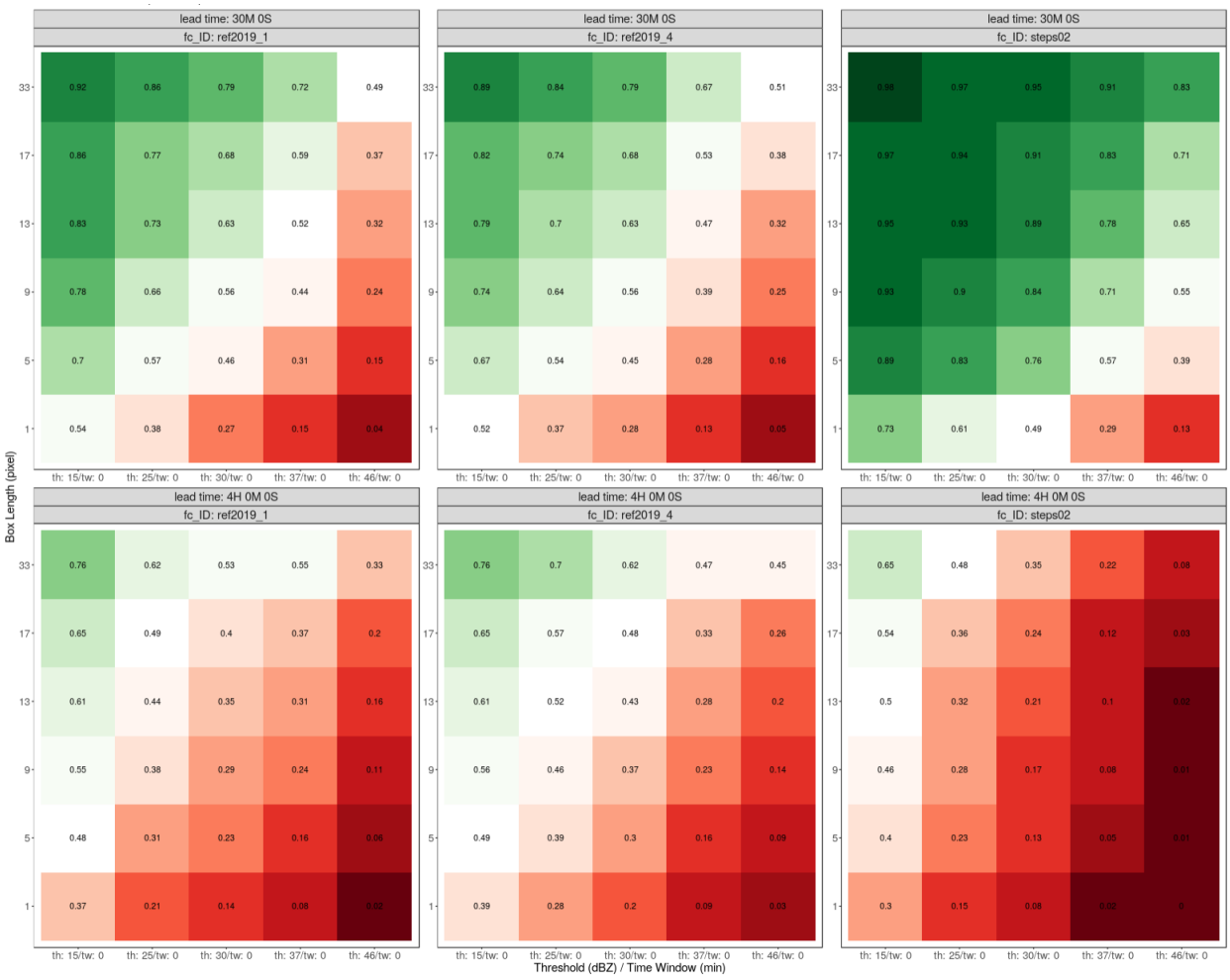

Figure 2: FSS tiles plots for reflectivity (dBz) averaged over the SINFONY reference period (27 May – 25 June 2016) and over all initial times (11 – 15 UTC) and all ensemble members (1 – 20, incl. NEP). The top row shows results for a lead time of 30 minutes, the bottom row for 4 hours. COSMO-DE-EPS 1-moment microphysics scheme (left panels), 2-moment microphysics scheme (middle panels) and STEPS nowcasting (right panels). Greenish colors represented a skillful FSS (≥ 0.5), reddish colors represent non-skillful FSS (< 0.5).

A first overview of the quality of the forecasts is given by Fig. 2. It shows FSS tiles plots for 30 minutes lead time (upper row) and 4 hours lead time (bottom row), as well as three different model setups, COSMO-DE-EPS 1-moment microphysics scheme (left panels), 2-moment microphysics scheme (middle panels) and STEPS nowcasting (right panels).

Aggregated over all parameters, the FSS shows normal behavior, i.e. increasing values with increasing box length (neighborhood size) and decreasing values with increasing thresholds. The STEPS nowcasting (right panels) is, as expected, of better quality after 30 minutes in comparison to the NWP setups. Especially the higher thresholds show better scores in the nowcasting, mostly because the NWP is not able to produce such high reflectivities. However, after 4 hours lead time (lower panels), the NWP quality is superior to nowcasting quality, which is not surprising since the nowcasting does not include dynamical information.

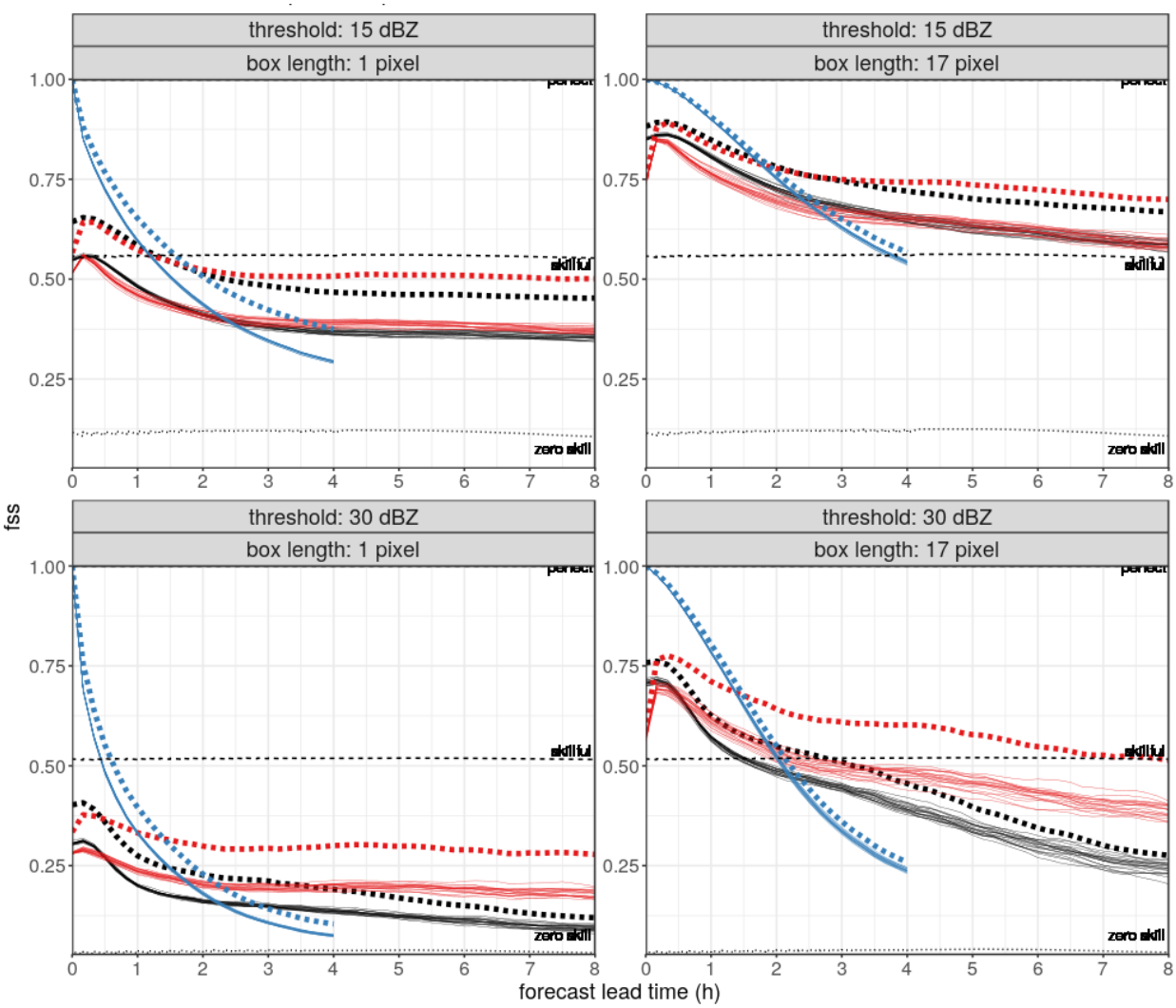

Figure 3: FSS as a function of lead time for reflectivity (dBz) averaged over the SINFONY reference period (27 May – 25 June 2016) and over all initial times (11 – 15 UTC). The top row shows results a threshold of 15 dBz, the bottom row for 30 dBz. The left column shows results for a box length of 1 pixel (1 km), i.e. no neighborhood, the right columns for 17 pixel (17 km). Thin solid lines show the FSS of all ensemble members, the thick dashed line shows the FSS of the NEP field. In black 1-moment microphysics scheme NWP, in red 2-moment microphysics scheme NWP and in blue STEPS-DWD nowcasting.

Fig. 3 shows the FSS results for reflectivity (dBz) as a function of lead time, aggregated of the SINFONY reference period and all initial times (11 – 15 UTC). The top row shows results a threshold of 15 dBz, the bottom row for 30 dBz. The left column shows results for a box length of 1 pixel (1 km), i.e. no neighborhood, the right columns for 17 pixel (17 km). Thin solid lines show the FSS of all ensemble members, the thick dashed line shows the FSS of the NEP field. In black 1-moment microphysics scheme NWP, in red 2-moment microphysics scheme NWP and in blue STEPS-DWD nowcasting.

It can be seen that the NWP (red, black) exhibits a short spin-up phase, whereas the spin-up effect is much stronger for the 2-moment microphysics scheme (red). The reason for this was that the model produced way to many reflectivity features in the early lead times. This effect is correct for ICON-D2-EPS in 2020 and 2021 (not shown). It is obvious that the NEP (thick dashed lines) has a quite positive impact on the score, especially for smaller box lengths. This fact answers the question whether we need an ensemble for our forecasting systems.

Another powerful tool in neighborhood verification is a respective reliability and ROC diagram. First, it must be clarified which type of observation should be taken into account. Since neighborhood verification methods potentially produce a huge amount of data, we decided for a compromise and used the binary

observation as reference for the diagrams. Otherwise, the user has to decide which observation neighborhood probability threshold he is interested in.

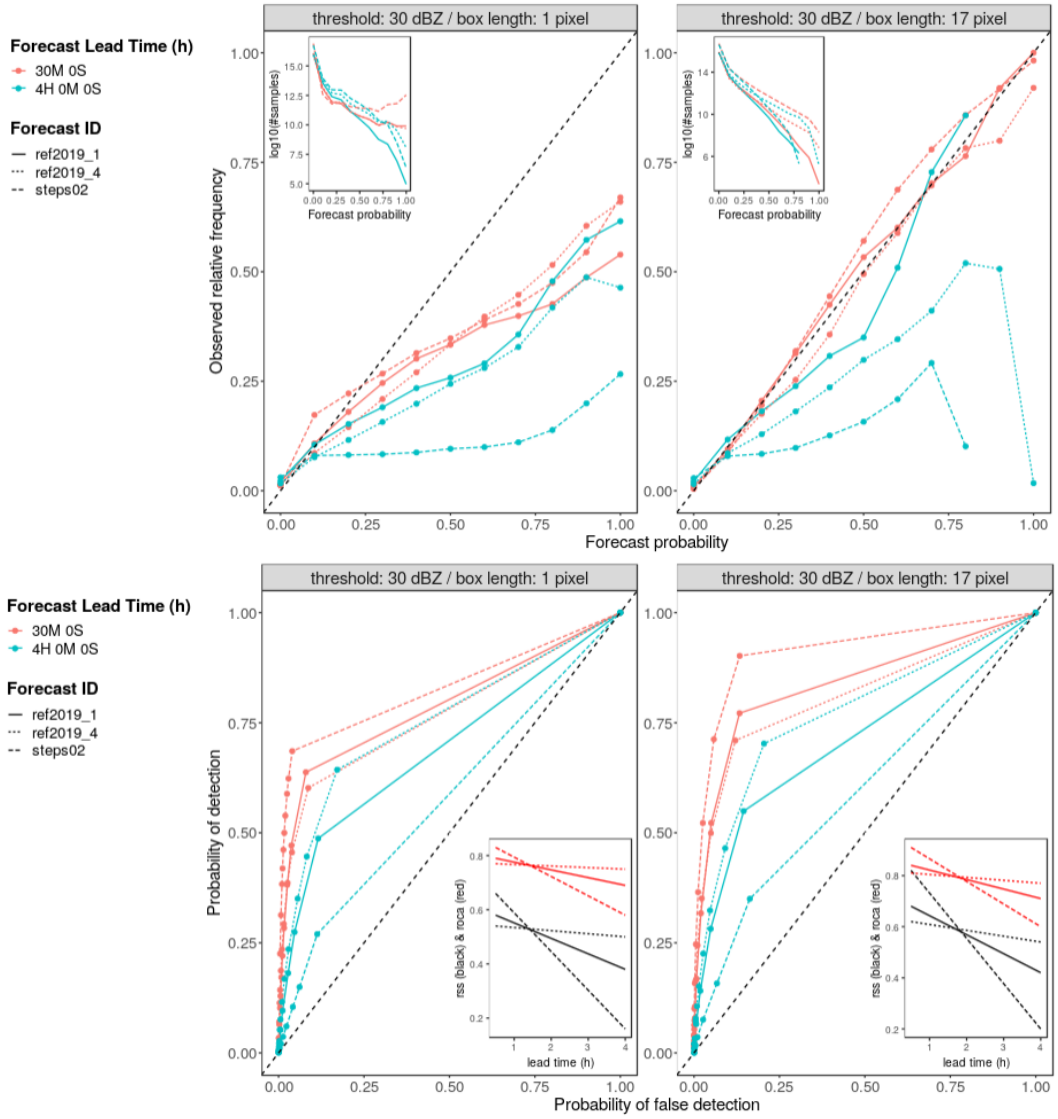

Figure 4: Reliability diagram (upper panels) and ROC diagram (lower panels) of NEP member for 30 dBz and two different box lengths, pixel-based (left) and 17 pixel (right). The model setups are coded as different line types. Red lines represent the lead time of 30 minutes and turquoise of 4 hours. The reference observation is of type binary.

Fig. 4 shows reliability diagrams (upper panels) of NEP member for 30 dBz and two different box lengths, pixel-based (left) and 17 pixel (right). The left panel shows the classical reliability diagram based on ensemble probabilities. It can be seen that there is over-forecasting for almost all cases, which increases for greater lead times (4 hours, turquoise). However, when we include a neighborhood box length of 17 pixel (17 km, right panel), there is almost perfect reliability of all model setups after 30 minutes lead time (red) and for some setups even after 4 hours lead time (turquoise). This confirms the fact that including a neighborhood can exhibit a massively increased forecast quality. A similar picture is given by the ROC diagrams in the lower panels of Fig. 4. The discrimination of events and non-events is much better when including a neighborhood box length of 17 pixel.

Another advantage of the neighborhood-based reliability diagrams is that they can be computed even for deterministic forecasts, i.e. based on neighborhood probabilities. This gives another great added value to forecast verification.

Finally, we want to show results for Displacement FSS and Displacement NSS developed by Skok and

Roberts (2018) and Skok (2021, not yet published). In contrast to the previously described results, we have now chosen STEPS DWD nowcasting data from May/June 2021 period.

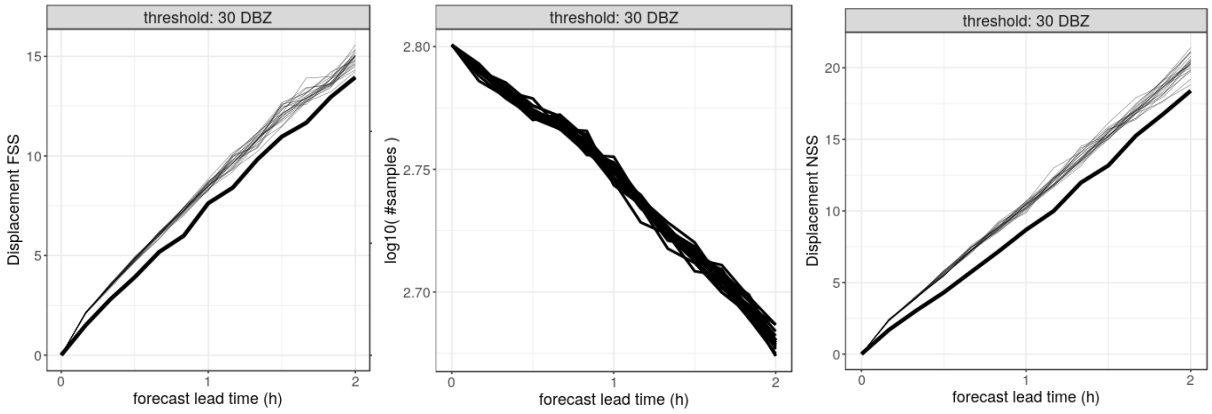

Figure 5: Displacement FSS (left), number of samples for D-FSS with $0.5 \leq FBI \leq 2$ (middle) and Displacment NSS (right) for STEPS DWD nowcasting in May/ June 2021 with 20 members. Data are aggregated over initial times from 6 – 18 UTC, 1-hourly.

The left and middle panels of Fig. 5 show the Displacement FSS (D-FSS) and respective number of samples for D-FSS with $0.5 \leq FBI \leq 2$, which are taken into account. It can be seen that the displacement is increasing almost linearly, which is in correspondence with the mechanism of nowcasting. After 2 hours of lead time, the global displacement ended up with about 14 km ensemble and 13 km deterministic. However, the number of samples with low bias decreased with increasing lead time.

In contrast, the Displacement NSS (D-NSS) score in the right panel of Fig. 5 has no limitation to the bias. Biased fields could simply be bias-corrected via constant factor. The displacement from D-NSS ended up at around 20 km ensemble and 16 km deterministic. This is slightly more than for D-FSS, however, the D-NSS score should be more confident than D-FSS. Not only because there is no bias limitation, also because some shortcomings of D-FSS are corrected in D-NSS score (see presentation of G. Skok at 8th IVMW 2020).

All in all, we found that D-FSS and D-NSS are very useful scores for interpreting other neighborhood scores, since most of them give no information about deviations in physical parameters. Even if the absolute values are not that exact as the reality, the relative values when comparing two experiments give added value to the verification. However, there is a problem of not negligible deviations from real displacement at domain edges. Up to now, we found no solution for this but this will be done in future work.

3.2 Object-based methods

3.2.1 Deterministic predictions

The MMI is calculated for the nowcasting and two sets of deterministic COSMO-DE forecasts, the first one employing the one-moment-, the second one the two-moment micropysics scheme. Nowcasts are initialized hourly between 12 UTC and 16 UTC and run for seven hours. The model is initialized hourly between 11 UTC and 15 UTC and evaluated for the first 8 forecast hours. The shift of one hour in the initialization explains with the fact that it takes about one hour from the model start until the predictions are available. For a fair comparison the 11 UTC model forecast is therefore compared to the 12 UTC nowcast and so on.

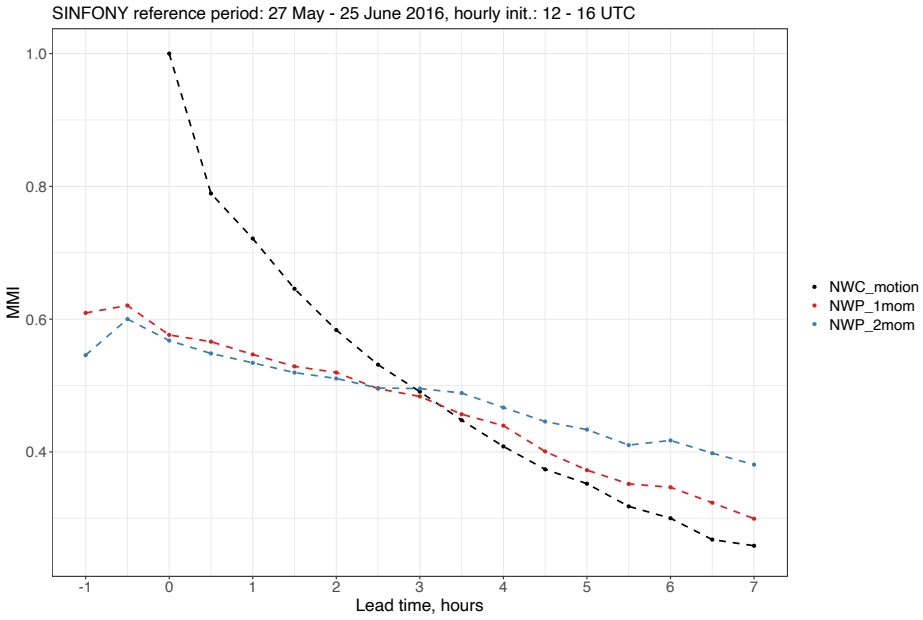

Figure 6: MMI vs lead time averaged over the SINFONY reference period (27 May – 25 June 2016) and over all initial times (12 – 16 UTC). Predictions are shown in black for the nowcasting and in red and blue for the deterministic model forecasts employing the one- and two-moment-microphysics scheme, respectively. The lead time of the model starts at -1 hour (i.e., 11 – 15 UTC), since about one hour is required for forecasts started at that time to become available.

The nowcasting starts at forecast time 0 with the perfect value of 1 (Fig. 6) because the observations serve as initialization for the nowcast and the fields are identical. The MMI decreases rapidly and is below the model forecasts after about 3 hours. The one-moment model forecasts start with higher MMI-values than the two-moment model data. At initialization, i.e., lead time -1 hour, this difference is most distinct. The artificial initialization of too many objects in the two-moment model causes the bad performance (see also discussion of Fig. 8). The MMI of the model forecasts approach after 30 minutes and the two-moment model is superior to the one-moment model after 4 hours of forecast time, i.e., 3 hours lead time in Fig. 6. From that lead time on the model forecasts perform better than the nowcasting with the clear trend that the two-moment model is superior to the one moment model.

3.2.2 Ensemble predictions

The analysis of ensemble forecasts is restricted to the two-moment model because its advantages at the longer lead times compared to the one-moment model.

Example of pseudomember characteristics

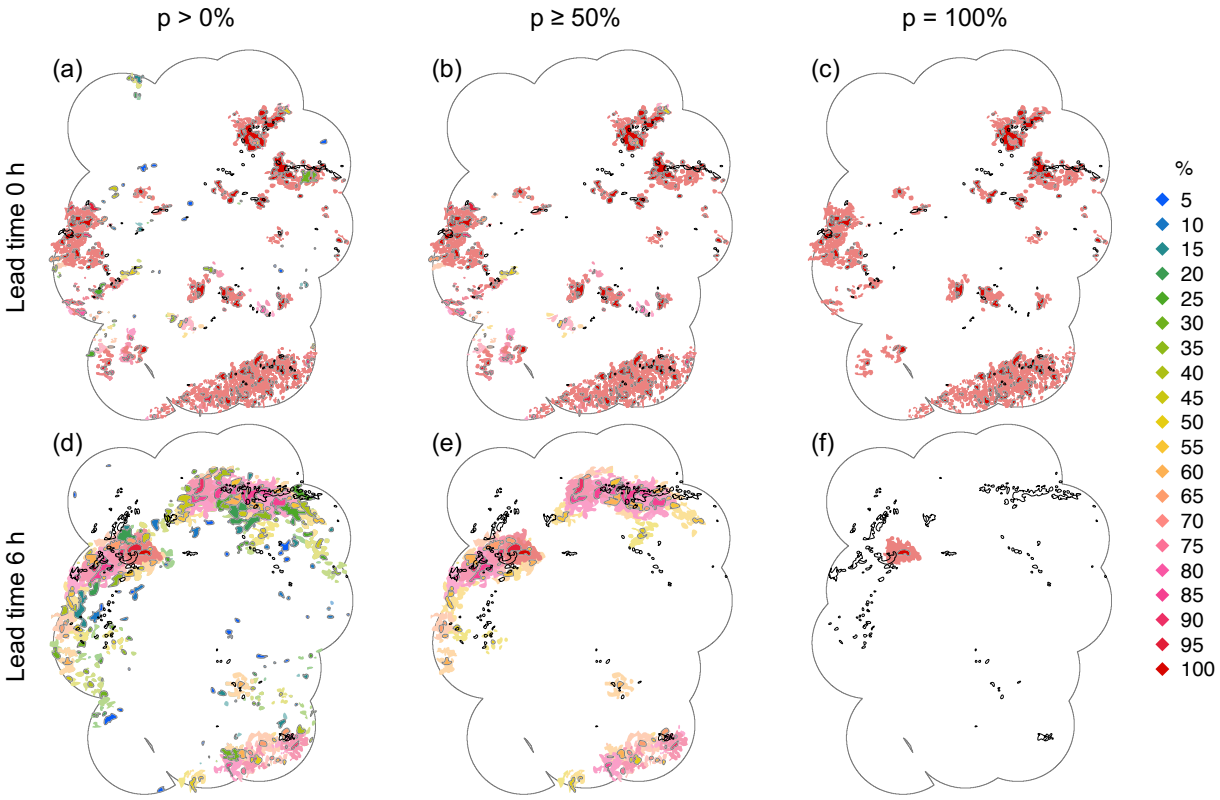

Figure 7: Observed objects (black) and objects of the pseudomember (gray-bordered) for forecasts initialized on 30 May 2016 12 UTC at lead times of 0 hours (top) and 6 hours (bottom). Pseudomember objects are colored according to their probabilities and areas in the respective lighter colors around these objects mark their uncertainty regions (see text for further details). The effect of considering only pseudomember objects exceeding a certain probability of occurrence p is illustrated by plotting all pseudomember objects (a, d: $p > 0\%$) and only objects with $p \geq 50\%$ (b, e) and $p = 100\%$ (c, f), respectively.

panel	lead time, h	p, %	# obs	# prd	MMI
a	0	0	82	260	0.55
b	0	50	82	217	0.56
c	0	100	82	136	0.58
d	6	0	137	210	0.56
e	6	50	137	70	0.49
f	6	100	137	1	0.03

Table 2: Number of observed (# obs) and predicted (# prd) objects and MMI for the examples shown in Fig. 7.

Figure 7 illustrates the objects of the pseudomember, their probabilities and uncertainty regions, and the observed objects for 30 May 2016. The forecast was initialized at 12 UTC. The corresponding numbers of observed and predicted objects and the resulting MMI are listed in Table 2. The ensemble shows little spread for a lead time of 0 hours as evidenced by the fact that most of the pseudomember objects have a probability of 100% (Fig. 7 top). For a lead time of 6 hours this has massively changed and only one object with $p = 100\%$ remains (Fig. 7f).

In comparison with the observed objects the pseudomember contains objects that represent the observations over large parts of the domain well. For lead time 0 all objects, i.e., $p > 0\%$, contain several false alarms, e.g., in the north-western and south-western part of the domain (Fig. 7a). In the Southeast the pseudomember has many objects with $p = 100\%$ where several but much less objects are observed (Fig. 7c). Removing objects with low probability from consideration generally reduces the number of false alarms

while introducing only few missed events over the central to western areas (Fig. 7c). This leads to a slight increase in the MMI from 0.55 ($p > 0\%$) to 0.58 ($p = 100\%$). For all p-values the number of predicted objects is clearly overestimated by a factor of 3.2 for $p > 0\%$ and still 1.7 if only objects with $p = 100\%$ are considered (Table 2).

After 6 hours all objects ($p > 0\%$) still contain false alarms over the south-western and south-eastern parts of the domain (Fig. 7d) and the total number of objects is overestimated by a factor of 1.5 (Table 2). Considering only objects with $p \geq 50\%$ again removes many false alarms on the one hand but the number of missed events increases on the other hand, over the central-western areas, for example (Fig. 7e). This leads to an underestimation in the number of predicted objects, 70, compared to 137 observed objects (Table 2). In comparison with lead time 0, the behavior of the MMI is reversed. Considering all objects yields the highest MMI (0.58) although about 50% more objects are predicted than observed. Constraining the pseudomember to objects with $p > 50\%$ causes a strong reduction in the number of predicted objects leading to a lower MMI of 0.49 (Table 2). Constraining the objects to $p = 100\%$ is not useful for this forecast range because all but one objects have lower probabilities (Fig. 7f) yielding a MMI of 0.03.

Number of objects

The number of objects can be used as a first criterion for the quality of a forecast and it can give a rough overview about false alarms and missed events in the prediction. The mean numbers for the SINFONY reference period at all initial times between 11 and 15 UTC are shown in Fig. 8. The observations have maximum 85–100 objects at early lead times between 0 and 3 hours, i.e., 11–18 UTC depending on the initial time. The number decreases with lead time to 13 objects at +8 hours lead time, i.e., 19–23 UTC. This reflects the diurnal cycle of convective activity with most objects occurring in the afternoon that become less during the evening and early night-time hours.

The number of pseudomember objects obviously decreases with increasing values of p (Fig. 8). At lead time 0 the model has too many objects which is a well known issue in the initialization of simulations employing the 2-moment microphysics scheme. These artificially initialized, unphysical objects have vanished after 30 minutes and from that lead time onward the numbers of the pseudomember objects scatter around the observed number of objects depending on p. The number of objects with $p > 30\%$ (light green in Fig. 8) represents the average number of observed objects best in both the temporal evolution with lead time and in the mean number (65 observed and 70 predicted objects).

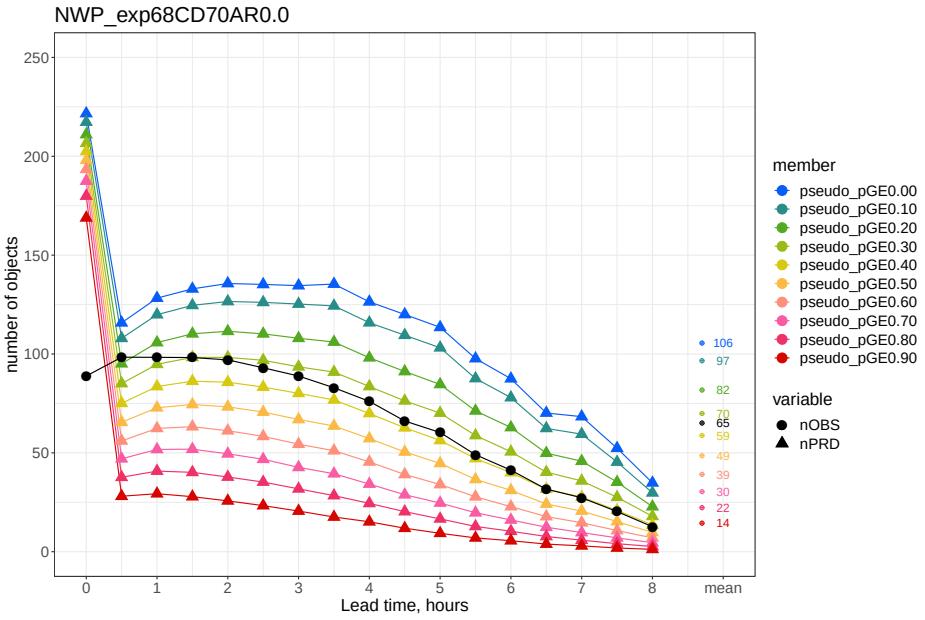

Figure 8: Number of observed (black) and predicted (colors) objects depending on the lead time, averaged over the SINFONY reference period (27 May – 25 June 2016) and over all initial times (11–15 UTC). Different colors distinguish which objects of the pseudomember are considered depending on their probability of occurrence from blue (all objects, $p > 0\%$) to red ($p > 90\%$). Mean values at the right are averaged over all lead times but 0 hours in order to remove the impact of artificial objects at initialization time.

MMI vs lead time

The forecast quality of different prediction types is again quantified in terms of the MMI. The following analysis comprises the MMI of the nowcasting, the deterministic forecast, all the single ensemble members, the pseudomember with $p > 30\%$, and two "best member" selections. For the latter the MMI is calculated for each forecast and each single ensemble member separately. The best member then is selected for the evaluation. We distinguish between the best member at each forecast time step ("best member at each step") and the best member on average over forecast lead time ("best member over lead time"). For these selections the observations for all lead times are required, hence, they can not be used as forecasts. Compared to the other real forecasts this method globally (over the entire domain) selects the best ensemble member as if one knew a priori which member will be the best for each forecast. The best member selections help to classify the quality of the other members.

The MMI of all these prediction types is illustrated in Fig. 9. The nowcasting (black) and the deterministic forecast (dashed blue) are the same as in Fig. 6. The MMI of the nowcasting is below the different model forecasts (blue) after about 2–4 hours. The deterministic forecast is slightly better than any individual ensemble member (dotted). The quality of the pseudomember is persistently the best, except lead times -1 and 7 hours, surpassing the quality of the nowcasting after only 2 hours. The pseudomember even outperforms the "best member" selections showing that it is better to do a localized selection of representative objects from the ensemble distribution than to choose the member that is globally the best. Again, the pseudomember is purely based on the ensemble forecasts while the "best member" selections need observational data for all lead times. This shows the enormous potential of the pseudomember for the object-based forecasting of precipitation.

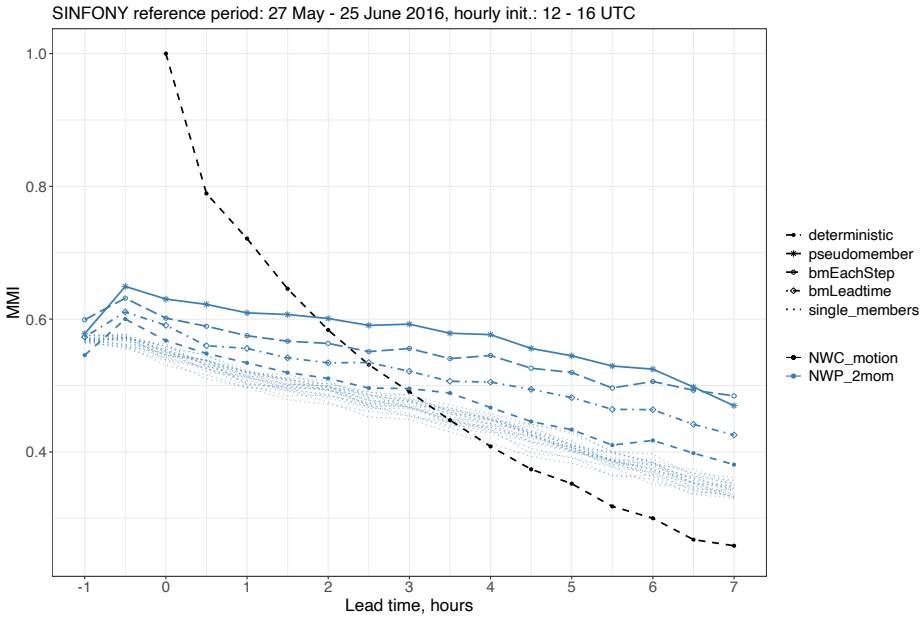

Figure 9: MMI vs lead time averaged over the SINFONY reference period (27 May – 25 June 2016) and over all initial times (12 – 16 UTC). Predictions are shown in black for the nowcasting and in blue for the model. The lead time of the model starts at -1 hour (i.e., 11 – 15 UTC), since about one hour is required for forecasts started at that time to become available. Different forecast types are distinguished by line types and symbols. The pseudomember is restricted to objects with $p > 30\%$. "bmEachStep" and "bmLeadtime" stand for "best member at each step" and "best member over lead time", respectively. See text for further details.

4 Conclusions

In the running PP-AWARE period, we have applied a lot of verification metrics which are already established (neighborhood verification) and tested also new verification metrics based on MMI (pseudomember by Johnson et al. (2020)). Especially the latter is quite useful in the SINFONY project. When using a 40 member object ensemble from NWP, nowcasting and combined products, the number of existing objects could become massively huge and not manageable without applying filter methods like pseudomembers.

All above described methods, and some more, are implemented in R-packages predominantly for DWD-internal usage. However, if the packages are well developed, they could be provided to the community. The R-packages are applicable by namelist control but also interactively. We will provide a flexible reading capability. The packages will have a flexible aggregation functionality over different parameters. A visualization via R-Shiny app will give the possibility to interactively visualize and aggregate scores in a way the user desired. Up to now, we do not plan to integrate an extensive pre-processing like regridding or restructering. We focus only on the computation of the scores and the user has the responsibility to unify the data in advance.

References

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 and Forecasting, 24, 1252 – 1267, https://doi.org/10.1175/2009WAF2222241.1, 2009.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorological Applications, 15, 51–64, https://doi.org/10.1002/met.25, 2008.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of Spatial Forecast Verification Methods, Weather and Forecasting, 24, 1416 – 1430, https://doi.org/10.1175/2009WAF2222269.1, 2009.

Johnson, A., Wang, X., Wang, Y., Reinhart, A., Clark, A. J., and Jirak, I. L.: Neighborhood- and Object-Based Probabilistic Verification of the OU MAP Ensemble Forecasts during 2017 and 2018 Hazardous Weather Testbeds, Weather and Forecasting, 35, 169 – 191, https://doi.org/10.1175/WAF-D-19-0060.1, 2020.

Roberts, N. M. and Lean, H. W.: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events, Monthly Weather Review, 136, 78 – 97, https://doi.org/10.1175/2007MWR2123.1, 2008.

Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R., Kong, F., Thomas, K. W., Levit, J. J., Coniglio, M. C., and Wandishin, M. S.: Toward Improved Convection-Allowing Ensembles: Model Physics Sensitivities and Optimizing Probabilistic Guidance with Small Ensemble Membership, Weather and Forecasting, 25, 263 – 280, https://doi.org/10.1175/2009WAF2222267.1, 2010.

Skok, G. and Roberts, N.: Estimating the displacement in precipitation forecasts using the Fractions Skill Score, Quarterly Journal of the Royal Meteorological Society, 144, 414–425, https://doi.org/https://doi.org/10.1002/qj.3212, 2018.

Stein, J. and Stoop, F.: Neighborhood-Based Contingency Tables Including Errors Compensation, Monthly Weather Review, 147, 329–344, https://doi.org/10.1175/MWR-D-17-0288.1, 2018.

Zeng, Y., Blahak, U., and Jerger, D.: An efficient modular volume-scanning radar forward operator for NWP models: description and coupling to the COSMO model, Quarterly Journal of the Royal Meteorological Society, 142, 3234–3256, https://doi.org/https://doi.org/10.1002/qj.2904, 2016.