Comparison of NWP based nowcasting (AROME) with classical system

II. Results

RC - LACE stay report

by Mirela Pietrisi
Supervisors: Christoph Wittmann, Florian Meier and Yong Wang
15^{th} May - 7^{th} July 2017

1 Assessing the hourly precipitation forecast skill for July 2016

The evaluation of hourly precipitation forecast at different spatial scales was investigated using the fractions skill score (FSS) by varying the width of the verification neighborhood in grid squares. Neighborhood widths of n = 3, 5, 7, 9 and 11 were applied to the AROME and INCA nowcasting systems. For each raw threshold and $n \times n$ size, a fractional coverage value is computed at each forecast and observation grid point. Therefore, instead of comparing forecast/observation values at individual grid points, areas of forecast values were compared to areas of observation values.

To give meaning to the forecast verification results, the statistics should be associated with uncertainty information (E. Gilleland, 2010). Therefore, the 95 % confidence intervals (CI) were applied to the computed statistics for the estimation of the sampling uncertainty. There are two main types of CI's available within MET package: parametric (normal approximation) and non-parametric (bootstrap method). For practical considerations, the percentile method was used for the computation of confidence intervals.

Fractions Skill Score was computed to obtain an objective measure of how forecast skill of both nowcasting systems varied across different spatial scale and thresholds. FSS is a variation on the Brier Skill Score (Roberts, 2008) and given by:

$$FSS = 1 - \frac{FBS}{FBS_{worst}} \tag{1}$$

$$FBS = \frac{1}{N}\sum_{j=1}^{N}(O_j - F_j)^2 \tag{2}$$

O_j and F_j are the observed and forecast fractions

$$FBS_{worst} = \frac{1}{N}\left[\sum_{j=1}^{N} O_j^2 + \sum_{j=1}^{N} F_j^2\right] \tag{3}$$

FBS_{worst} - gives the largest possible FBS that could be obtained from the observed and forecast fractions.
Roberts and Lean (2008) estimated the forecasts useful skill that would be obtained at the grid scale (width = 1) as $FSS_{uniform}$:

$$FSS_{uniform} = 0.5 + \frac{f_0}{2} \tag{4}$$

where f_0 is the base rate. This value is considered to be a suitable "target skill", because is halfway between random skill ($FSS_{random} = f_0$) and perfect skill ($FSS_{perfect} = 1$). FSS ranges between 0 and 1, with 0 representing no overlap and 1 representing complete overlap between forecast and observed events. The score is sensitive to rare events (or for small precipitation areas)

(https://dtcenter.org/met/users/docs/users_guide/MET_Users_Guide_v5.2.pdf).

Figures 1 and 2 show the FSS computed against different neighborhood widths (n = 1, 3, 5, 7, 9 and 11) for AROME nowcasting system using as rainfall thresholds $\geq$ 0.25 mm/h (Figure 1) and $\geq$ 1.0 mm/h (Figure 2). The x-axis corresponds to the forecast time (hours) and the y-axis corresponds to the starting hours (24 runs per day).

It can be noticed that the FSS values increased as the neighborhood widths increased and it show a slight decrease when the rainfall thresholds were increased. As it was expected, the lowest skill is obtained at the grid scale (the neighborhood is only one grid point) and the highest skill from the largest neighborhood. Likewise, the best scores are obtained in the case of light rainfall.

All simulations exhibit an increase of the FSS values starting at 10 UTC, when usually the convective activity is triggered. From 10 UTC to 16 UTC, the FSS values continue to be high in regards to the rest of the simulations, emphasizing the period of convective activity throughout the day.

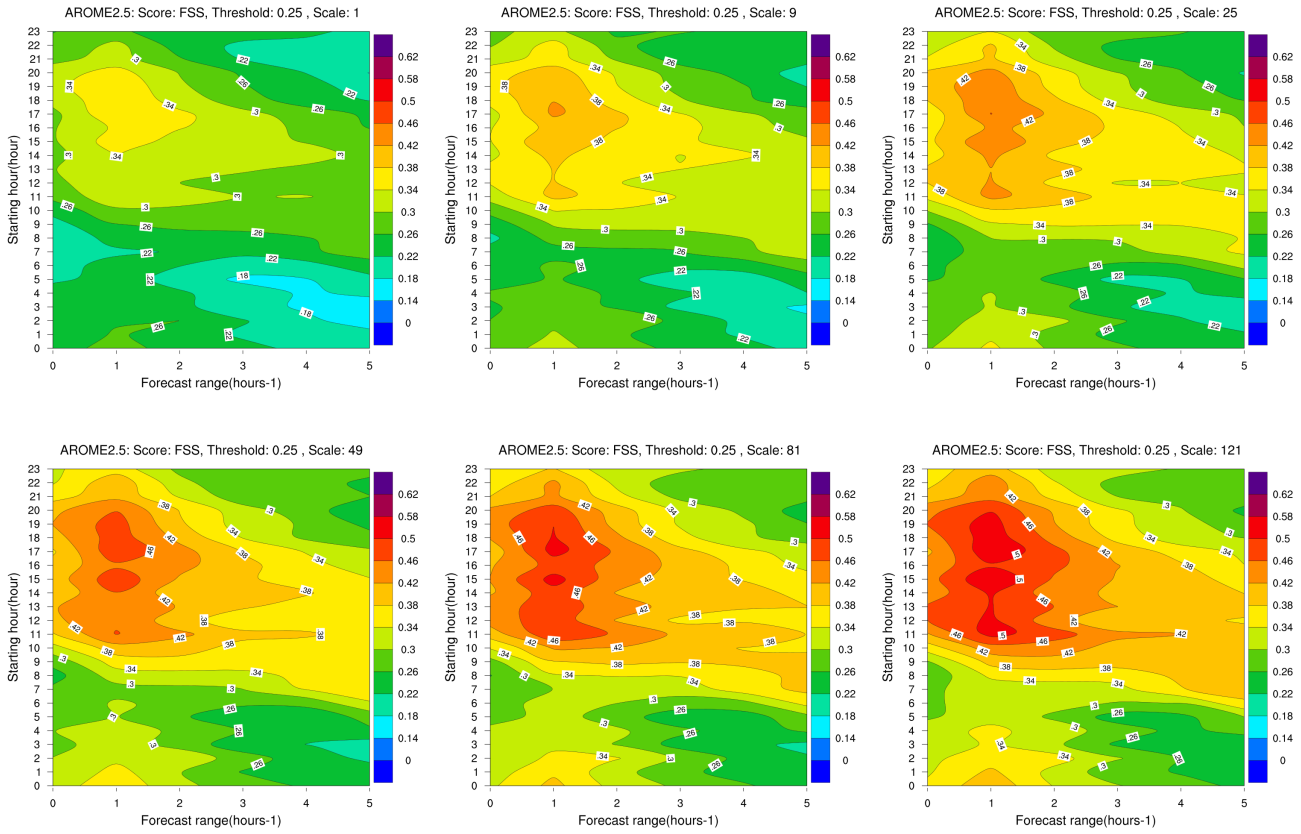

Figure 1: The evolution of FSS over different neighborhood widths for AROME nowcasting system; threshold $\geq$ 0.25 mm/h (up: n = 1, 3, 5 and bottom: n = 7, 9, 11)

The hourly cycling 3D-Var system assimilates radar reflectivity and Doppler winds from 4 radars, as well as other observations from local database and the OPLACE system, with the observation cuttoff time of 25 minutes. Due to its narrow assimilation window, the system assimilates mostly surface observations and radar data.

It was found that the hourly accumulated precipitation forecast skill, as measured by the FSS, for higher thresholds starts to decrease. This fact does not imply that AROME nowcasting has poor skill for heavy rain forecasting because the higher thresholds are associated with very small scale features, like intense localised convective storms (Mittermaier and Roberts, 2010). The 2.5 km convective-permitting AROME model has the capacity to resolve the convective storms, but the exact location of any individual storm has limited predictability. It is a known that for high spatio-temporal forecast precipitation resolution it is difficult to match the forecast and observation perfectly, particularly for hourly accumulated precipitation.

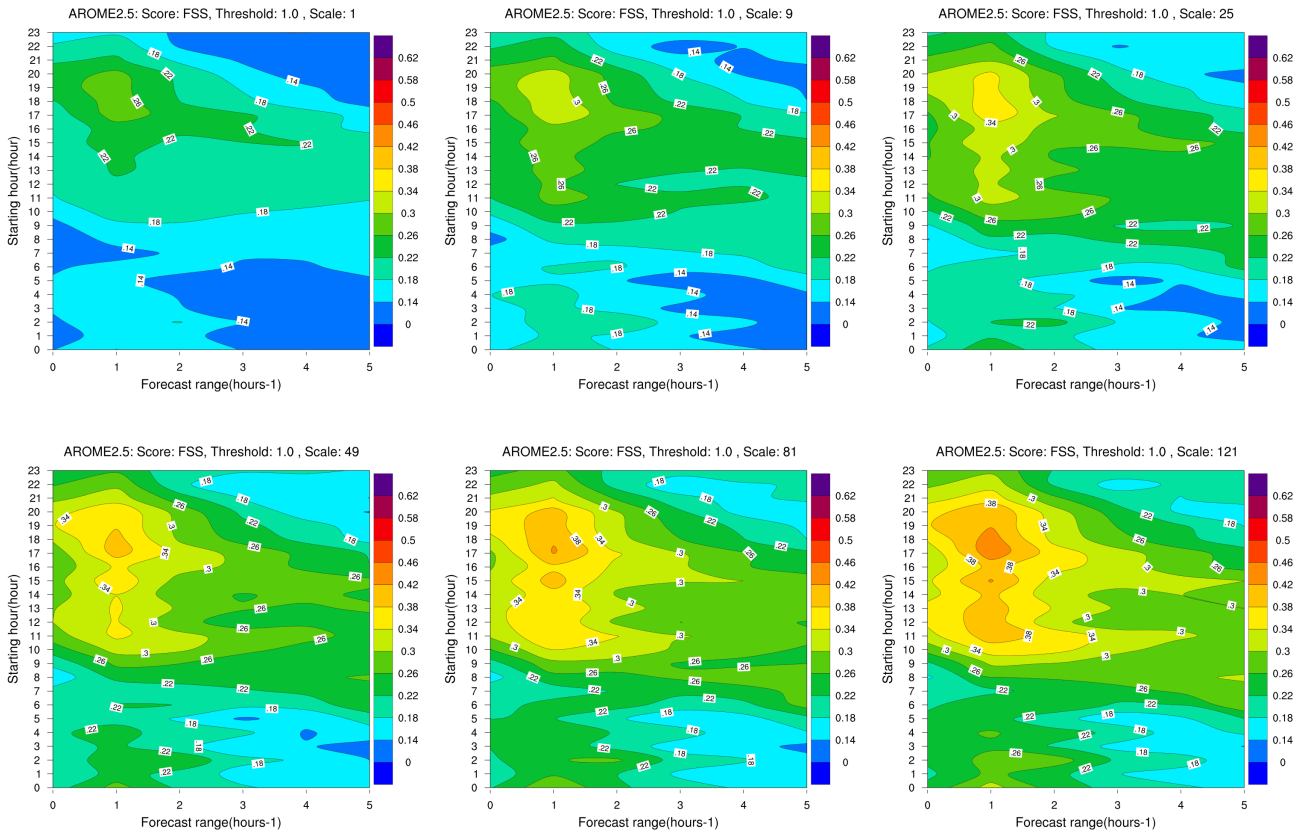

Figure 2: The evolution of FSS over different neighborhood widths for AROME nowcasting system; threshold $\geq$ 1.0 mm/h (up: n = 1, 3, 5 and bottom: n = 7, 9, 11)

The same procedure was applied for the computation of FSS for the INCA nowcasting systems (observation-based extrapolation and NWP based). The following notation will be used: **INCA_N** for the INCA observation-based extrapolation system and **INCA_O** for the INCA NWP based system.

Figures 3 an 4 show the FSS computed against different neighborhood widths (n = 1, 3, 5, 7, 9 and 11) for INCA_N and INCA_O nowcasting versions with rainfall thresholds $\geq$ 0.25 mm/h.

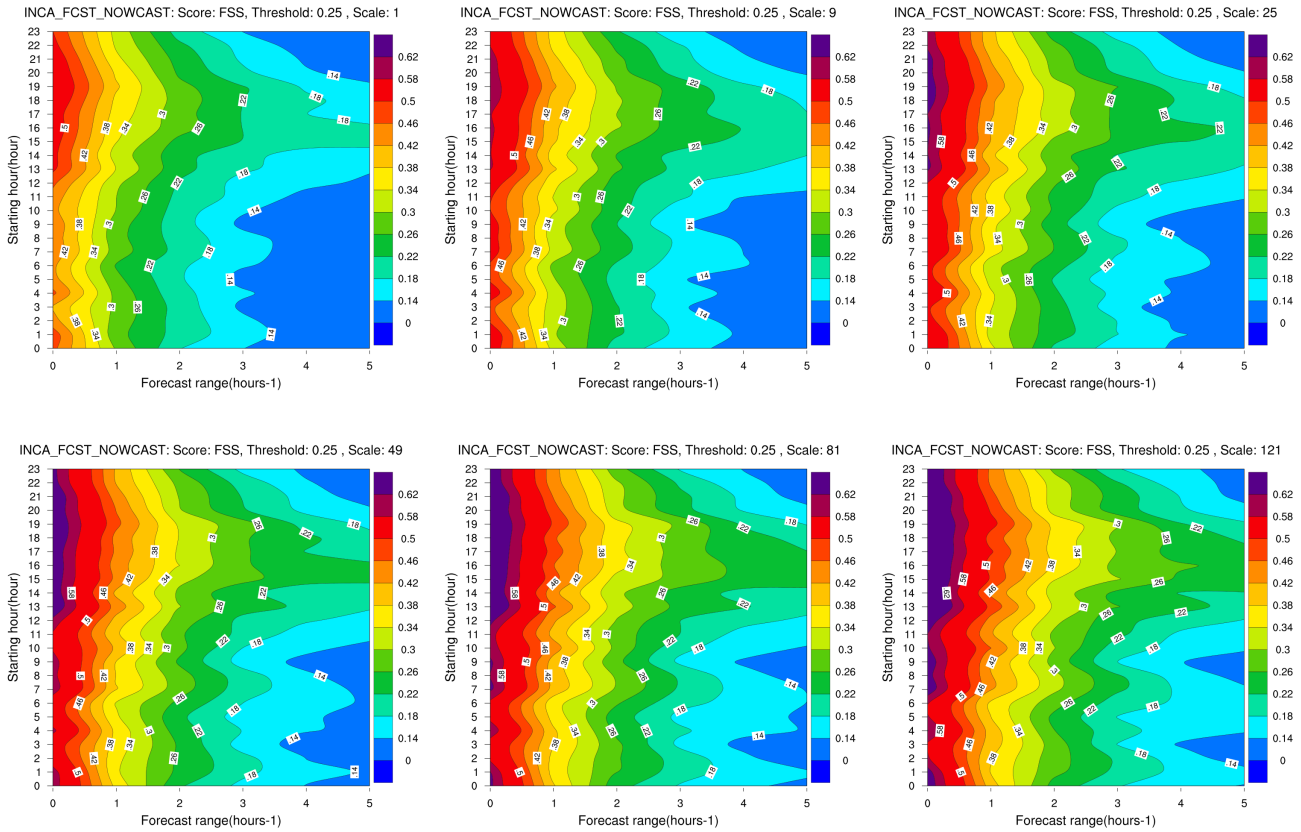

Figure 3: The evolution of FSS over different neighborhood widths for INCA_N system (up: n = 1, 3, 5 and bottom: n = 7, 9, 11), threshold $\geq$ 0.25 mm/h

The skill of the INCA_N is the highest at 0 - 2 forecast lead time, decreasing after the first two hours. This is due to the fact that extrapolation techniques can't initiate new convective storms or change the intensity or motion of the convective structures.

The decrease in the forecast skill with lead time can also be seen for the INCA_O system. Whilst the INCA_N skill drops rapidly after T+2 forecast range, it is not entirely the case for INCA_O. It can be noticed from the FSS's shape that from 10 UTC to 16 UTC, INCA_O retains some forecast skill. This fact is due to the benefit of blending the extrapolated forecasts with the latest available operational 2.5 km AROME forecast, which improved the extrapolated nowcast precipitation.

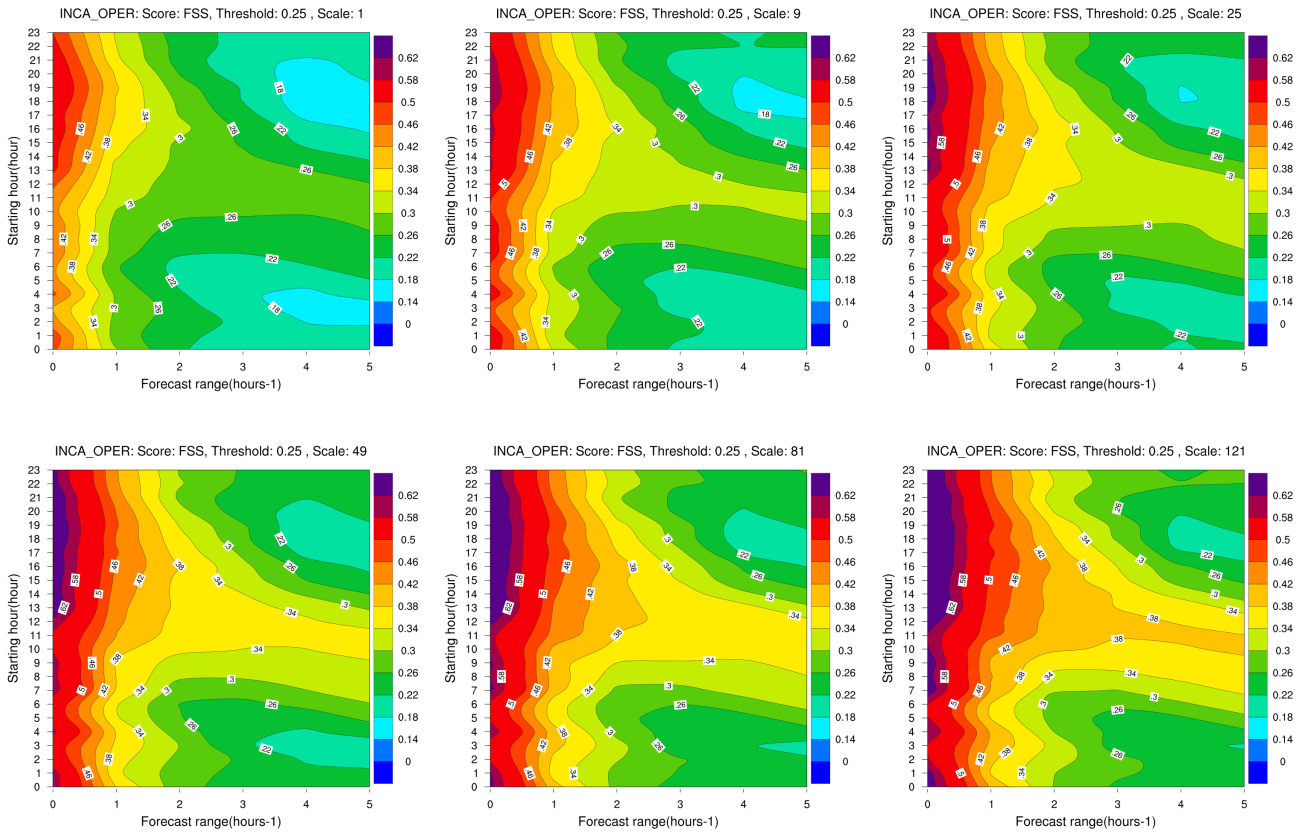

Figure 4: The evolution of FSS over different neighborhood widths for INCA_O system (up: n = 1, 3, 5 and bottom: n = 7, 9, 11), threshold $\geq$ 0.25 mm/h

Figures 5 and 6 show the hourly accumulation precipitation FSS difference between INCA_N, INCA_O and AROME nowcasting systems for thresholds $\geq$ 0.25 mm/h (top panel) and $\geq$ 1.0 mm/h (bottom panel).

The graphs exhibit a positive difference in the skill of INCA in regards with AROME for the first two forecast lead time. Yet, after T+2 hours of simulation AROME outperforms INCA for both precipitation thresholds.

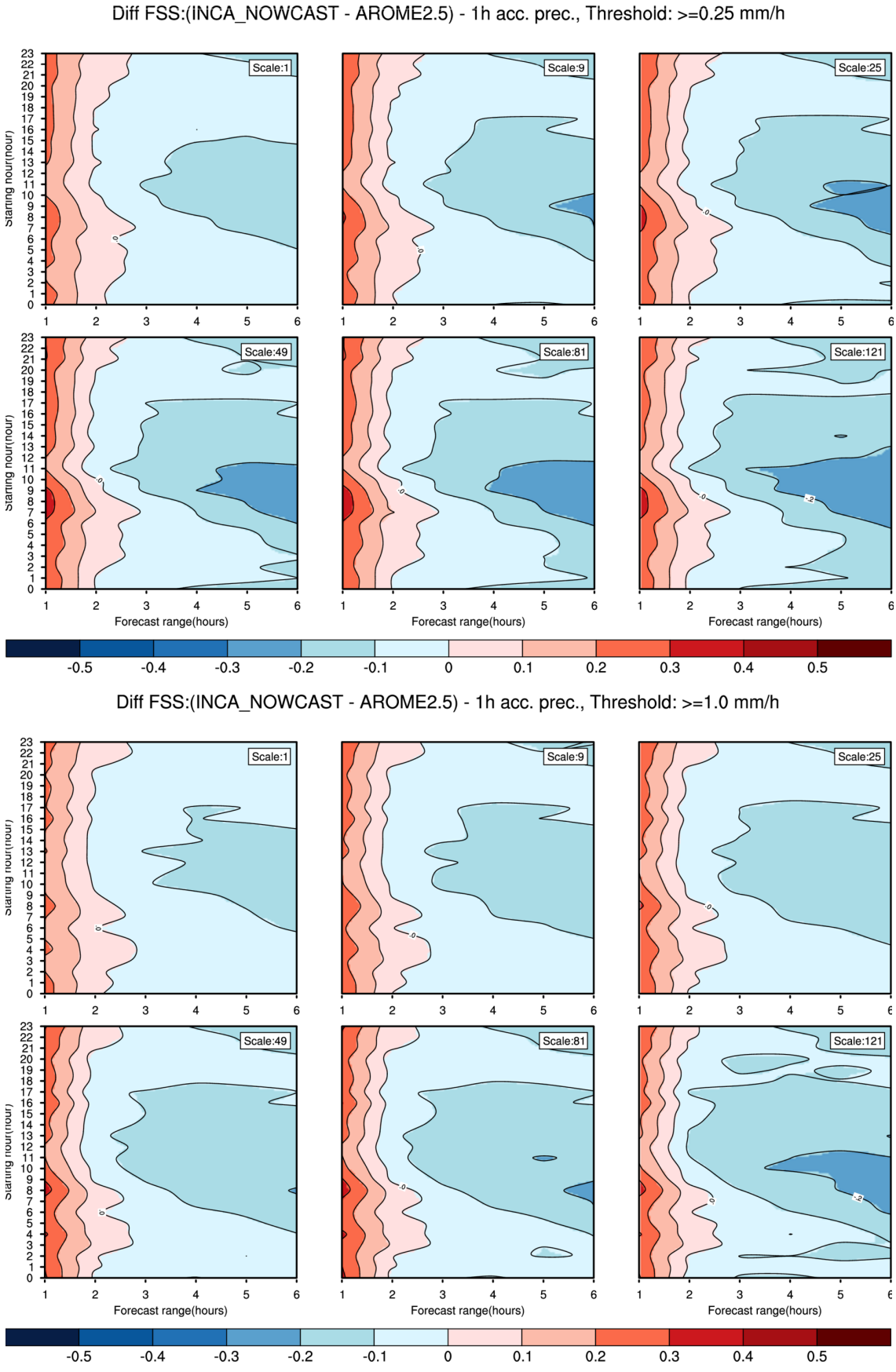

Figure 5: hourly accumulation precipitation FSS difference between INCA_N and AROME nowcasting systems: up - threshold $\geq$ 0.25 mm/h and bottom - threshold $\geq$ 1.0 mm/h

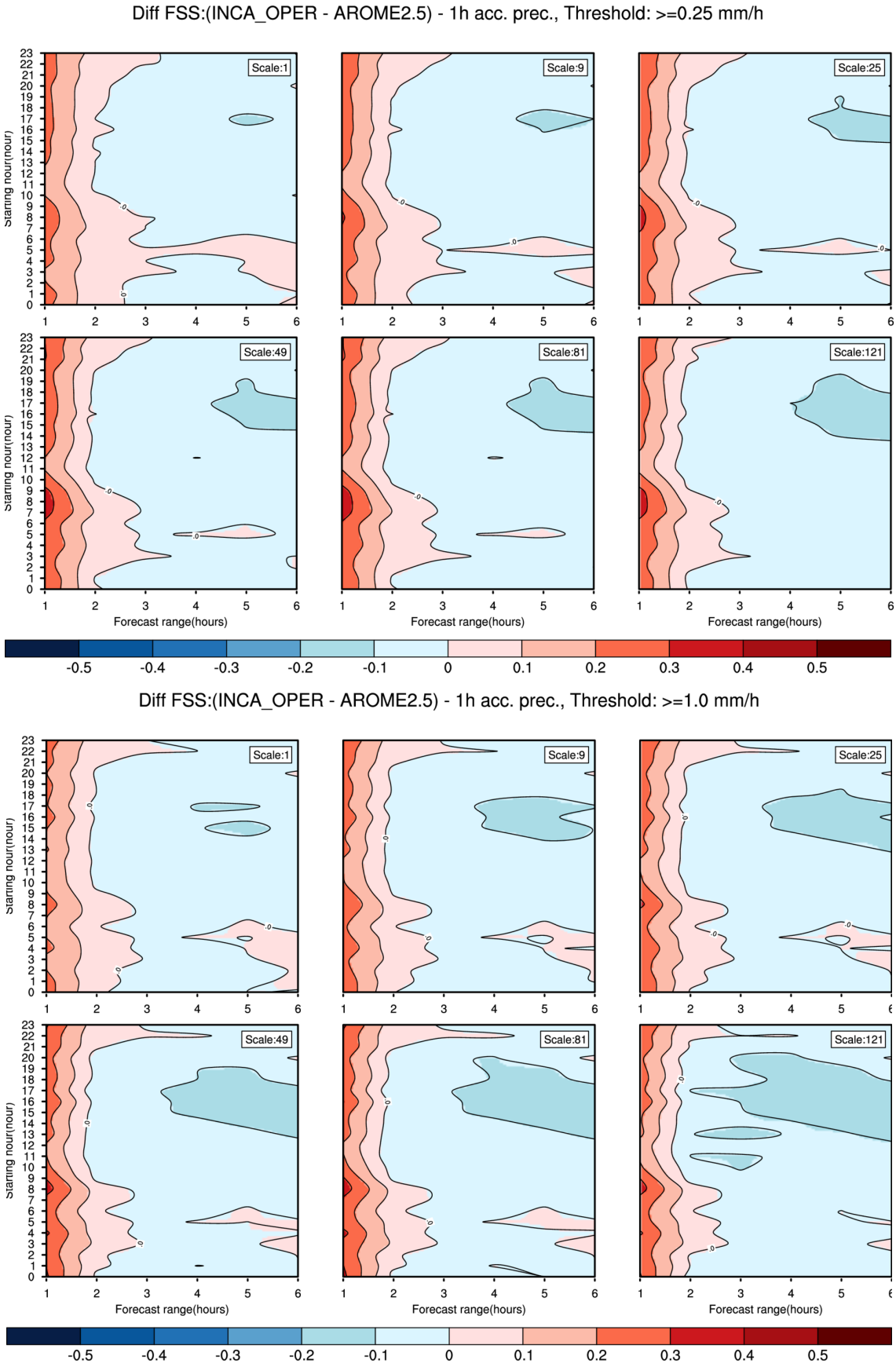

Figure 6: hourly accumulation precipitation FSS difference between INCA_O and AROME nowcasting systems: up - threshold ≥ 0.25 mm/h and bottom - threshold ≥ 1.0 mm/h

1.1 Aggregated results

Aggregated values were computed for the traditional (not shown) and FSS methods using the Stat_Analysis tool. This tool ties togheter results from Grid_Stat tool by providing summary statistical information.

First, to obtain a general view about how fast the nowcasting systems are drifting from the truth, the statistics were stratified across all model initialization by the lead time. Figures 7 - 10 show the FSS values aggregated for all model initialization by the lead time for AROME and INCA nowcasting versions for the thresholds: $\geq$ 0.25 mm/h, $\geq$ 1.0 mm/h, $\geq$ 2.0 mm/h and $\geq$ 4.0 mm/h. It seems that for higher thresholds, AROME ($\geq$ 2.0 mm/h and $\geq$ 4.0 mm/h) outperforms INCA before T+2 lead time, but the FSS values are smaller than $\text{FSS}_{uniform}$ values.

Second, to obtain the overall performance for both nowcasting systems, the statistics were aggregated across their neighborhood results over 4464 precipitation forecasts sample. Figure 11 shows a verification quilt of FSS as a function of spatial scale and threshold for the 6 hours lead time. The y-axis corresponds to the neighborhood size which increases toward the top of the plot and the x-axis corresponds to the threshold which increases toward the right. It can be seen from the FSS quilt plots, the higher thresholds are always associated with FSS values less that $\text{FSS}_{uniform}$ values. The concept of FSS_{useful} is sometimes difficult to quantify, it depends on the forecast application.

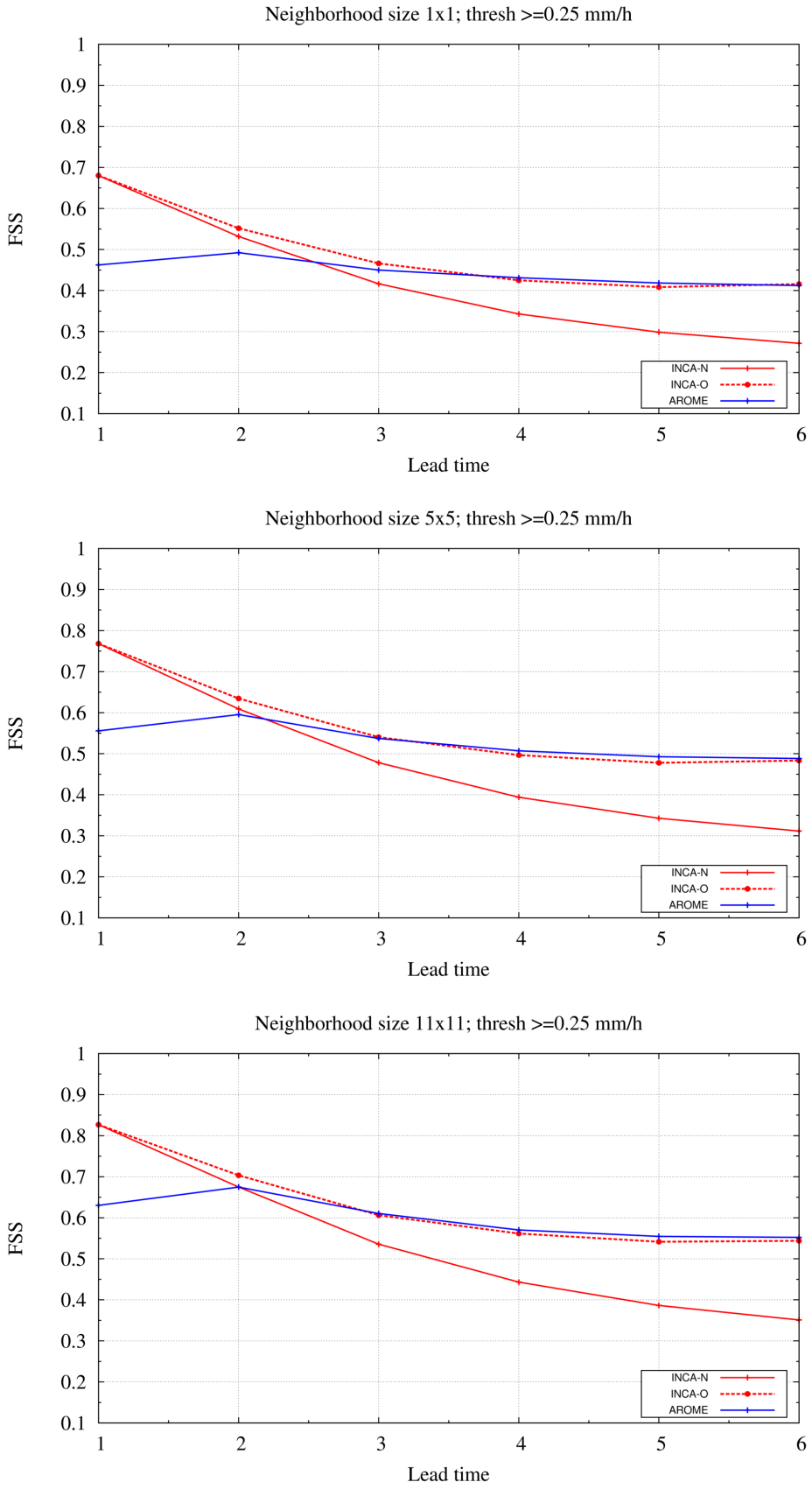

Figure 7: The FSS values aggregated all model initialization by the lead time for AROME (blue color) and INCA nowcasting versions (red colors) , threshold $\geq$ 0.25 mm/h

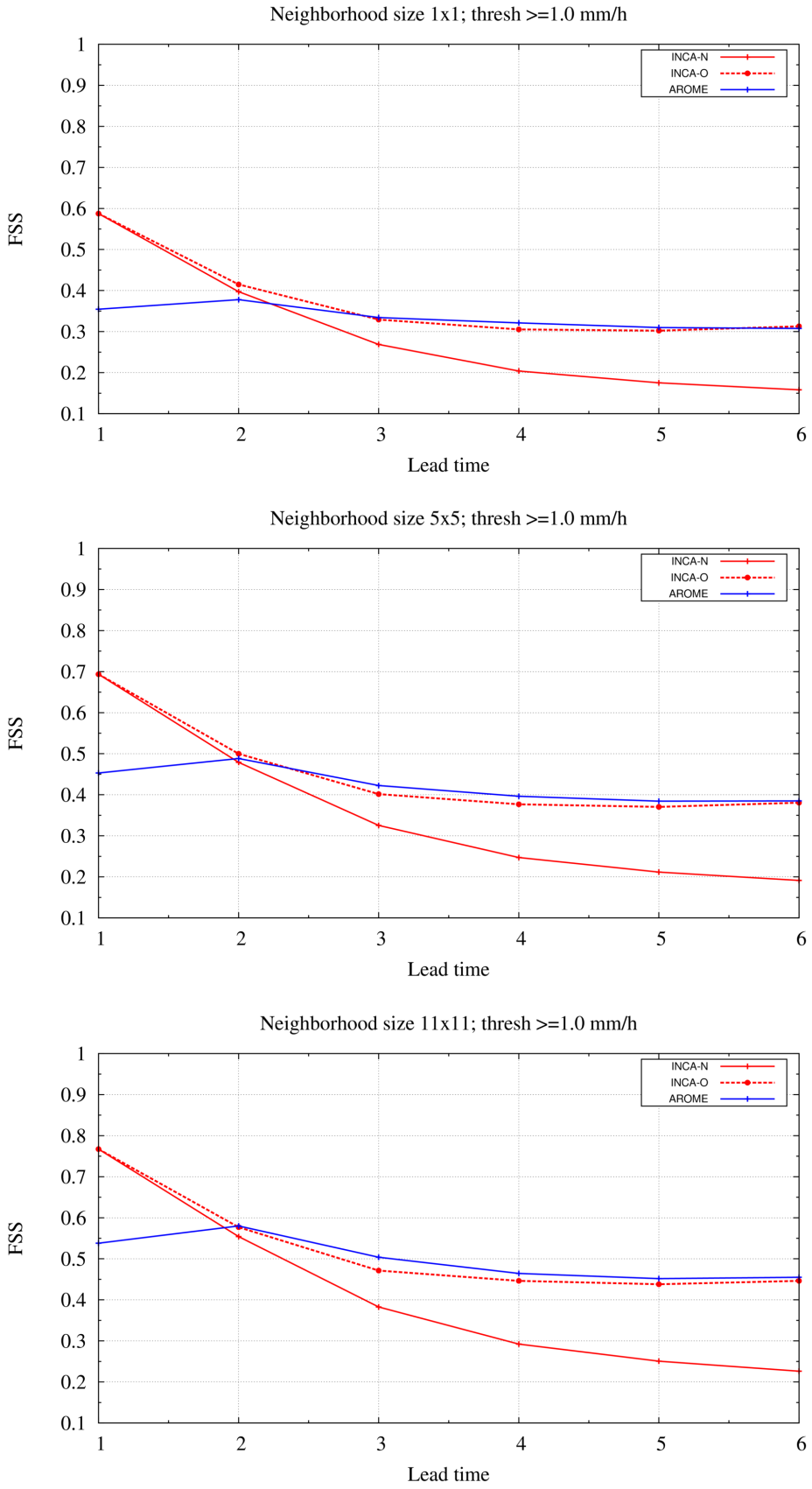

Figure 8: The FSS values aggregated all model initialization by the lead time for AROME (blue color) and INCA nowcasting versions (red colors) , threshold $\geq$ 1.0 mm/h

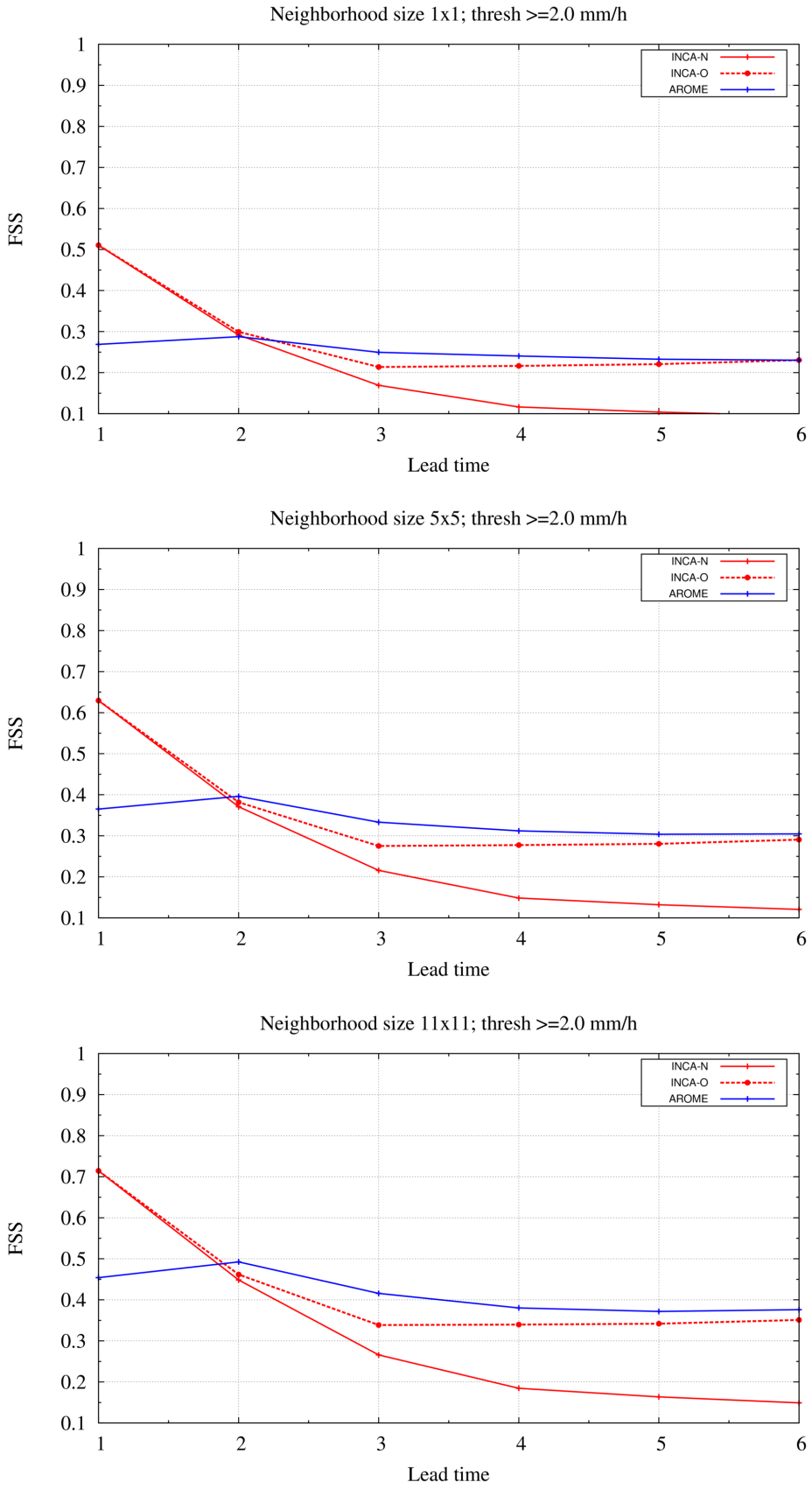

Figure 9: The FSS values aggregated all model initialization by the lead time for AROME (blue color) and INCA nowcasting versions (red colors) , threshold $\geq$ 2.0 mm/h

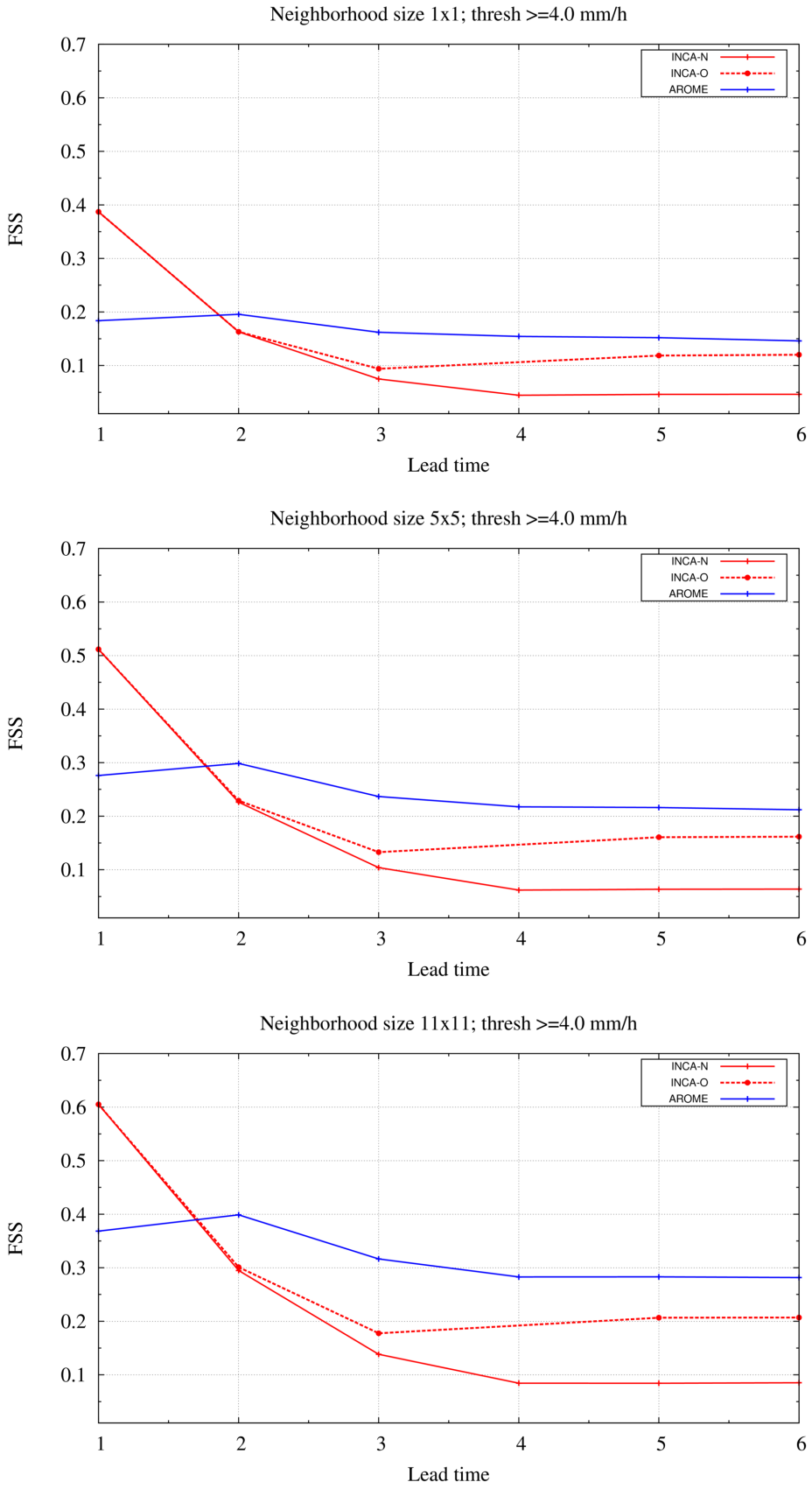

Figure 10: The FSS values aggregated all model initialization by the lead time for AROME (blue color) and INCA nowcasting versions (red colors) , threshold $\geq$ 4.0 mm/h

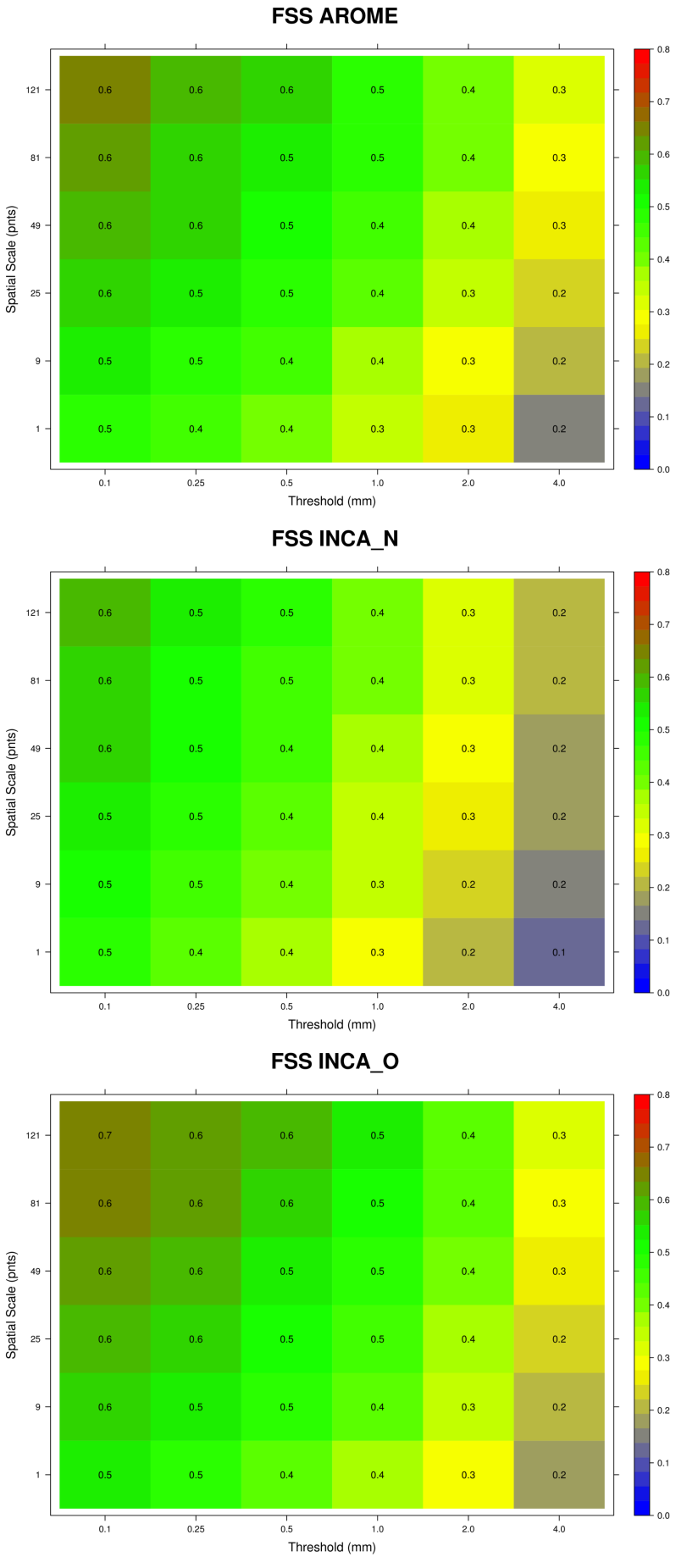

Figure 11: Quilt plots of FSS as a function of spatial scale and threshold across all model initialization for the 6 hours lead time: AROME (top), INCA_N (middle) and INCA_O (bottom)

1.2 Summary

Briefly, the findings regarding the FSS are:

- Firstly, taking into account the "acceptable forecast skill" estimated by Roberts and Lean (2008) it was found that AROME maintains skill at a scale of 9 grid squares throughout the 6 hours forecast lead time, whereas skill decreases for high thresholds. Regarding the INCA systems, the smallest neighborhood size which provides sufficiently skillful forecast is about 5 grid squares for lighter precipitation, however increasing the threshold, the skill decrease;

- Secondly, from the FSS differences graphs, it can be seen that INCA forecasts are more skillful than AROME forecasts for the first 2 hours, but they lose skill with the lead time. Yet, after T+2 forecast lead time, AROME outperforms INCA.

  The real crossing point is not simple to estimate, mainly because the INCA is basically verified against its own analysis for the first forecast hour. Ideally, an independent variable measure of skill is needed. This fact is sustained also by Ballard et al. (2015) and Simonin et al. (2017). Another fact could be the AROME spin-up period which can affect deeply the magnitude of FSS, especially in the first forecast hour.

- Thirdly, from the aggregated results it seems that for higher thresholds, AROME outperforms INCA before T+2 forecast range, although the values of FSS fall below the uniform value.

For the analyzed period, the overall conclusion is that the AROME provided better forecasts after T+2 lead time. This is due partially from the 3DVaR data assimilation scheme, latent heat nudging approach and from cycling frequency (hourly cycling implies the use of more observations).

2 Assessing the hourly precipitation forecast skill for January 2017

During winter time, the synoptic scale systems tend to produce large area of more stratiform precipitation events. Therefore, different accumulation thresholds were chosen for the FSS's computation (rainfall thresholds: ≥ 0.1 mm/h ≥ 0.25 mm/h ≥ 0.5 mm/h ≥ 1.0 mm/h ≥ 1.5 mm/h).

Figure 12 shows the evolution of FSS over different neighborhood widths for AROME nowcasting system, threshold ≥ 0.1 mm/h. According to Roberts and Lean work (2009), the FSS "useful skill" was not achieved at any shown scale for this winter period. One possible explanation for this behaviour might be the small sample of precipitation forecasts for January 2017.

Same graphs were obtained for the INCA nowcasting systems. It is shown only the evolution of FSS over different neighborhood widths for INCA_O system (Figure 13) In this case, the FSS "useful skill" is achieved at a scale of 7 grid squares.

Figure 14 show the hourly accumulation precipitation FSS difference between the INCA_O and AROME nowcasting systems for thresholds ≥ 0.1 mm/h. It is clearly that for this winter month, INCA_O outperforms AROME system after T+2 lead time.

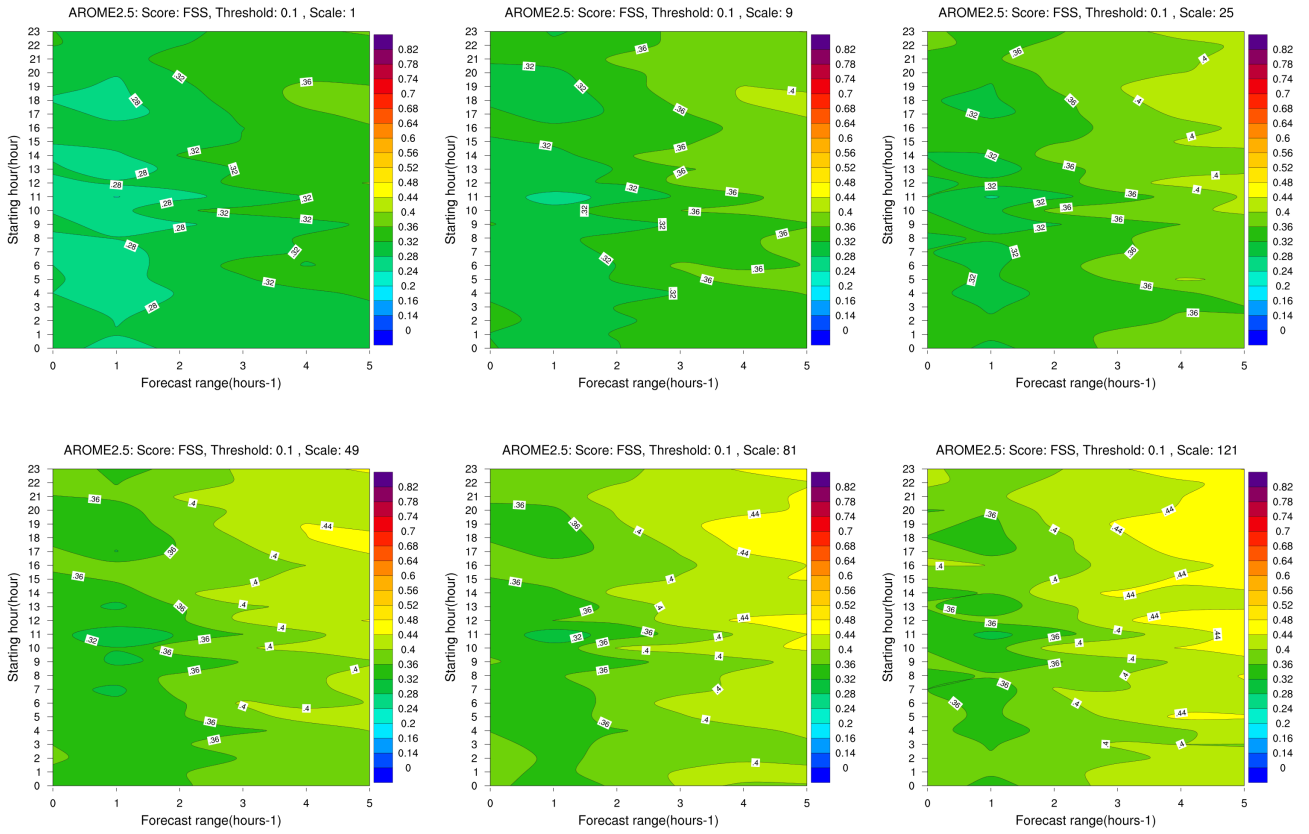

Figure 12: The evolution of FSS over different neighborhood widths for AROME nowcasting system; threshold ≥ 0.1 mm/h (up: n = 1, 3, 5 and bottom: n = 7, 9, 11)

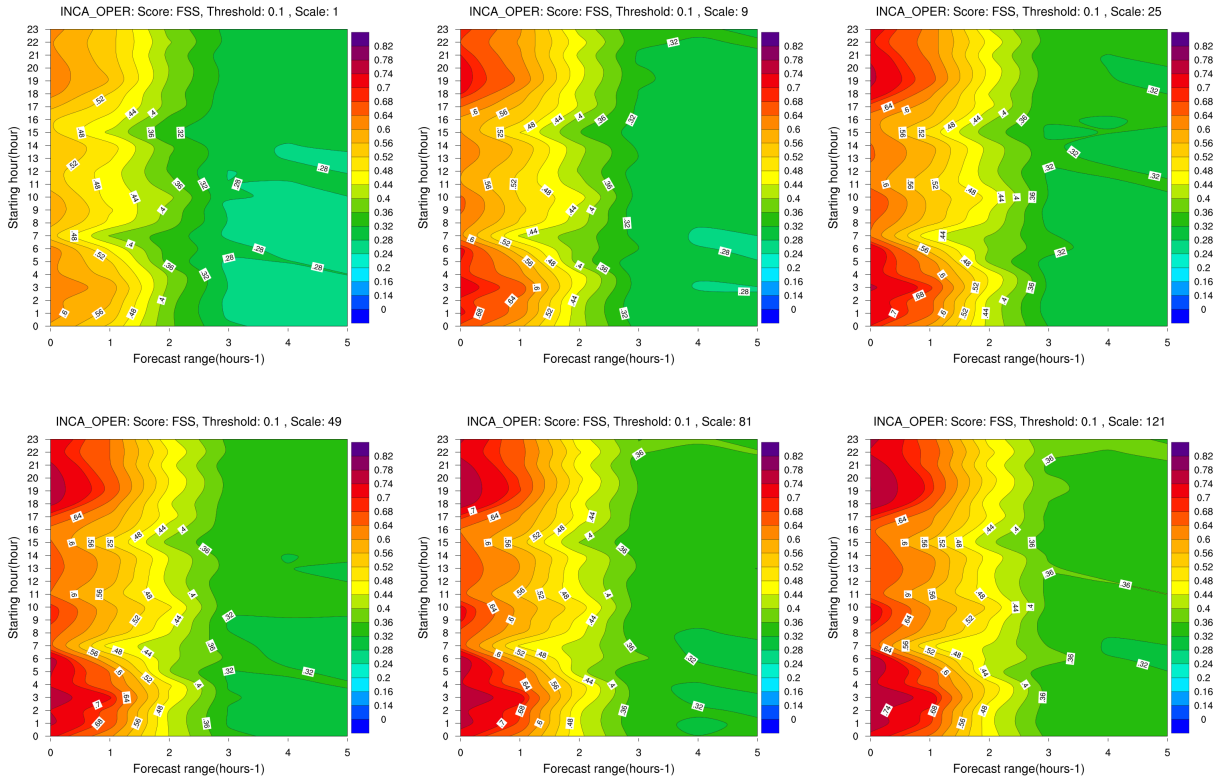

Figure 13: The evolution of FSS over different neighborhood widths for INCA_O system (up: n = 1, 3, 5 and bottom: n = 7, 9, 11), threshold $\geq$ 0.1 mm/h

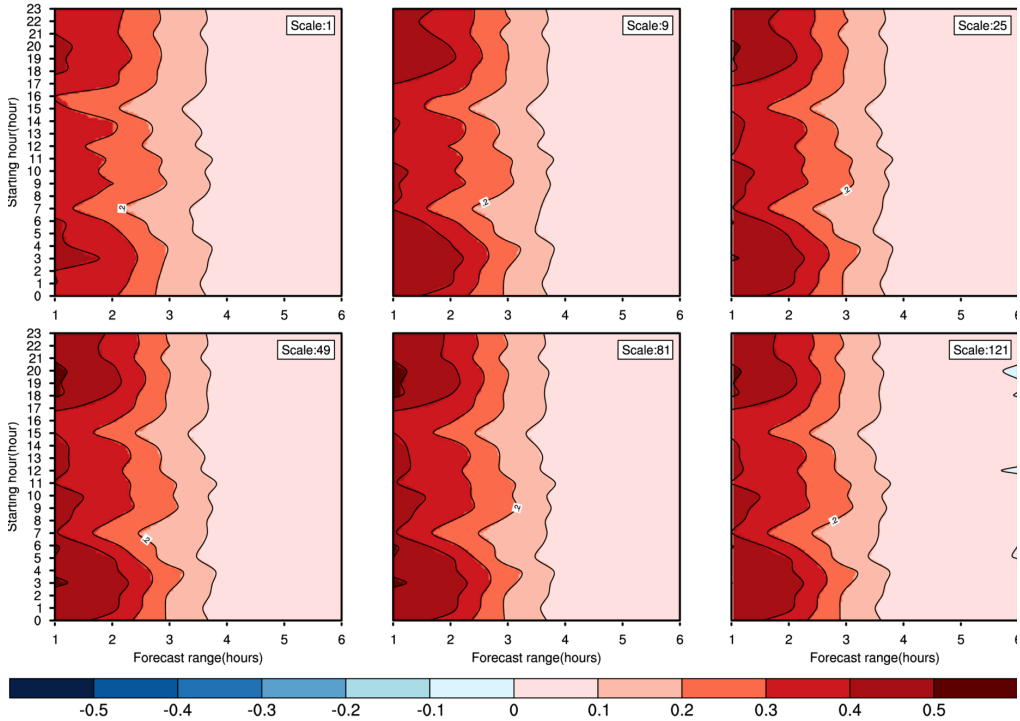

Figure 14: hourly accumulation precipitation FSS difference between INCA_O and AROME nowcasting systems, threshold $\geq$ 0.1 mm/h

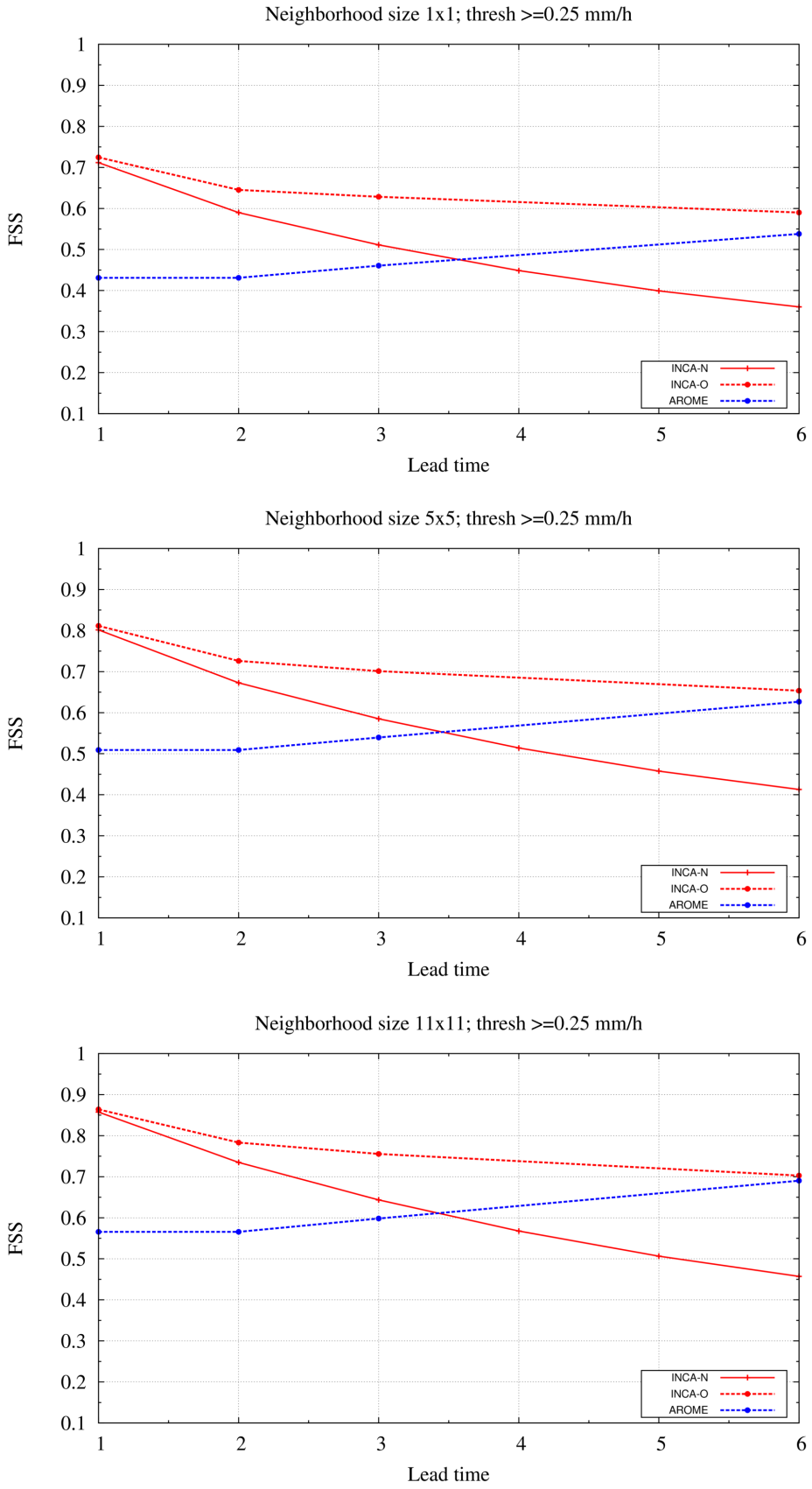

Figure 15: The FSS values aggregated all model initialization by the lead time for AROME (blue color) and INCA nowcasting versions (red colors) , threshold $\geq$ 0.25 mm/h

- The high-resolution AROME nowcasting system was run over the Austrian domain for one summer and winter month. The system uses an hourly cycling which allows the 3DVAR data assimilation to use recent observations.
- In general, the AROME nowcasting achieves a level of skill comparable to INCA systems around the 2h forecast lead time. Beyond this crossing point, AROME outperforms INCA for all precipitation thresholds.
- Taking into account the "useful scale" defined by Roberts and Lean, for the summer period the scale for AROME was found at 81 pnt and for INCA at 25 pnt. For the winter period, the "uniform value" of the FSS was hard to find it at shown scales, mainly due to the small sampling of precipitation forecast.

Model Evaluation Tools (MET) was developed by the National Center for Atmospheric Research (NCAR) Developmental Testbed Center (DTC) through the grants from the National Science Foundation (NSF), The National Oceanic and Atmospheric Administration (NOAA), the United States Air Force (USAF), and the United States Department of Energy (DOE). NCAR is sponsored by the United States National Science Foundation.

Acknowledgements

I would like to thank to Christoph Wittmann, Florian Meier and Yong Wang for their constant help, advice and the entire support during my stay.

References

[1] Roberts NM, Lean HW. 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Monthly Weather Review 136: 78 97;

[2] Roberts, N. M., 2008: Assessing the spatial and temporal variation in skill of precipitation forecasts from an NWP model. Meteor. Appl., 15, 163169;

[3] Ebert, 2009: Neighborhood verification: A strategy for rewarding close forecasts. Wea. Forecasting, 24, 14981510, doi:10.1175/ 2009WAF2222251.1.

[4] Marion Mittermaier, Nigel Roberts and Simon A. Thompson, 2013. A long-term assessment of precipitation forecast skill using the Fractions Skill Score. Meteorol. Appl. 20: 176-186, DOI: 10.1002/met.296

[5] Wolff, J., M. Harrold, T. Fowler, J. Halley-Gotway, L. Nance, and B. Brown, 2014: Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. Wea. Forecasting, 29, 14511472, doi:https://doi.org/10.1175/WAF-D-13-00135.1

[6] Mittermaier MP, Bullock R. 2013. Using MODE to explore the spatial and temporal characteristics of cloud cover forecasts from high-resolution NWP models. Meteorol. Appl. 20: 187196 DOI: 10.1002/met.1393

[7] Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. Wea. Forecasting, 25, 343354. DOI:10.1175/2009WAF2222260.1

[8] Gilleland E, 2008. Confidence intervals for forecast verification. Submitted as an NCAR Technical Note. Available at: http://www.ral.ucar.edu/ericg/Gilleland2008.pdf