Review of egusphere-2025-685 #1

Indications of authors : The authors would like to express their gratitude to the anonymous referee for his/her meticulous examination of the paper, which will undoubtedly enhance its quality. We hope that readers will find the answers useful.

The response to each comment is indicated in blue. Citations to parts of the paper will be quoted, and, if necessary, modified text according to the reviewer's comments will be underlined.

General comments:

The paper analyzes the ability of the AROME NVP convection model to represent convective thunderstorms. The model is run with two different microphysical schemes, which are validated on the basis of the French polarimetric radar network for 34 convective days in 2022.

In general, the paper is well written, the work is comprehensive but also quite lengthy as a result. However, I would not have a direct suggestion on how it should be shortened: The introduction is well researched, chapters 2 and 3 are mandatory for data and methodology, and the results section with comparisons of a) precipitation b) polarimetric moments c) ZDR columns is adequate. I also liked the appendix with the more technical insights into the forward operators melting scheme. Nevertheless, I would have a few minor points regarding clarity. Otherwise, I am pleasantly surprised that I found hardly any technical corrections.

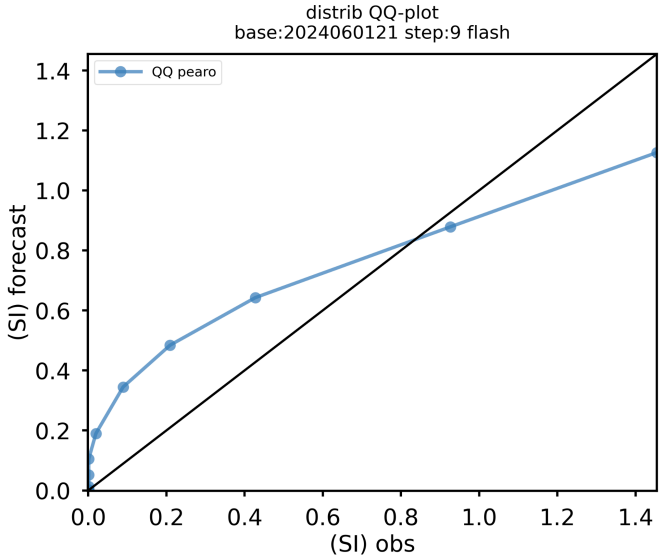
Specific comments:

L117: Why were not all 51 days selected for this study? Is it because only 34 days of these 51 days have QC1 or QC2? That is not entirely clear here.
We understand that this is confusing. In fact, it took a lot of time to get from the archives the radar data, and to replay them. Furthermore, we also had to reforecast each studied day two times (one per microphysics). Since we worked with a temporal resolution of five minutes (so our model data was consistent with our radar data), it generated a huge amount of data to analyze. Therefore, we decided to focus on events that occurred during spring and early summer (i.e. from April to July). However, 2022 had been a year rich in strong convection events, and because we were aiming to have a wide diversity of study cases, we finally decided to include 3 more study cases in our analysis sample : a major thunderstorm event all across France (16 Aug), the Corsica bow echo (18 Aug), and the Bihucourt tornado (23 Oct).
Referee #2 also noted that there was a lack of clarity surrounding this decision. Thus, we will rephrase as follows : "To target severe convective events, we restricted the selection to days where hailstones of size equal or greater to 2cm were reported in France in 2022, while keeping the number of cases reasonable in terms of computation and processing time. Thus, we included

all convective events from April to July, a period well characterized by severe convective storms and in particular supercells. In order to incorporate a variety of storm types, three high-stakes days were added to the study sample (a major thunderstorm event all across France on August 16, the Corsica bow echo on August 18, and the Bihucourt tornado on October 23), for a total of 34 convective days analyzed."

L124: I wonder whether AROME is perhaps predicting too many convective events in general because the events were selected for days on which thunderstorms actually occurred. It would be interesting to know whether AROME perhaps has a too high false alarm rate, i.e. whether too many thunderstorms were predicted even on calm days.

On days where there is no potential for strong thunderstorms, AROME does not predict high intensity storms. Similarly, on calm days, AROME does not forecast thunderstorms. In fact, the intensity dependence is quite robust. Overall, AROME has a frequency bias and predicts too many weak thunderstorms (in particular at night), but this does not apply to strong thunderstorms. As a result, it does not affect the choice of cases studied. This is illustrated by the enclosed plot.



⇦ Quantile-quantile diagram of AROME lightning density forecasts vs observed lightning density (from Météorage data), both evaluated on the same 10-km resolution grid, over June 2024.

SI = strike index (flashes per hour per square km)

Note : the figure is similar to other Spring months and years.

It shows that, on average, AROME predicts too many weak thunderstorms (QQ curve above the diagonal) where none is observed (a caveat being that some weak flashes may go undetected by the Météorage system), whereas AROME slightly underpredicts the frequency of the strongest thunderstorms (QQ curve under the diagonal for observed lightning densities larger than 0.8).

Since the paper is already quite long, we do not propose to include such a plot, but we propose to insert a sentence to summarize this information : "On average, AROME tends to slightly underpredict the frequency of the strongest thunderstorms (not shown)."

L126: So the newly calculated polarimetric data come from all three bands that were mixed together? Depending on the bands, the distance and the hydrometers present, the polarimetric signal could be different. What is selected then? I assume that ANTILOPE QPE takes this into account, but how is this taken into account in section 4.3?

Each polarimetric radar is processed through the polarimetric chain, our input radar data are hence already corrected accordingly to their frequencies, in particular for the attenuation correction (see Figueras i Ventura et al., 2012). However, we avoided using X-bands if they could have been included. So, for each event, we only used either exclusively C-bands or S-bands, or a mix of the two. We added a precision : "For the 3D evaluation, we only used C- and S-bands radar data, as X-band signal is quickly attenuated during rainy events."

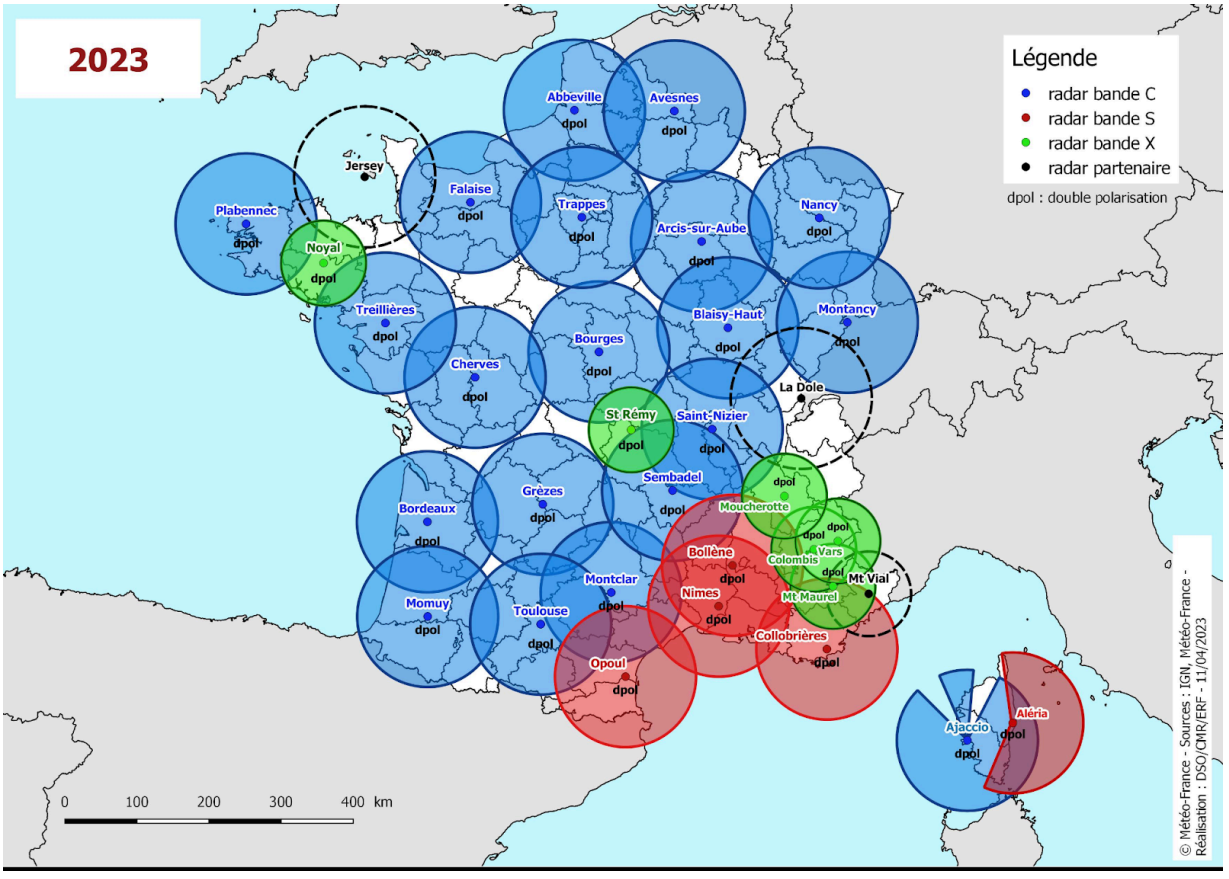
Among the studied events (see Table A1 in the Appendix), S- and C-band may have been mixed for those that occurred in the southeast or south of France (where the S-band radars are located). In this case :

1. during the interpolation step of the observations, contribution from each radar at a given grid point is weighted by its distance to the grid point. This means that, in the vicinity of an S-band radar, the resulting polarimetric field will be more influenced by the S-band radar than by a more distant C-band radar.
2. when we applied the forward operator on model outputs, we have chosen the frequency band according to the majority radar band (so for example S band if there were a majority of S-band radars that contributed to the observational grid). This has been clarified at the end of section 3.1 : "Finally, the radar forward operator is applied to model outputs to simulate polarimetric fields, with the frequency band chosen in accordance to the majority radar band in the corresponding observational dataset."

However, we did not apply a correction factor to the output of the simulated S-band datasets to bring us back to the values of a C-band radar (or inversely). For example, as KDP is inversely proportional to the wavelength, a simple multiplication by two would allow us to homogenize KDP values of S-band ($\lambda \approx 10$cm) and C-band ($\lambda \approx 5$cm). But it is not that straightforward with ZH and ZDR. For subsequent work, we will keep in mind this idea, and apply correction if the number of S-band datasets is significant compared to C-band datasets.

L255: Is coverage by 3 different radar stations per grid box required? Or do you mean at least 3 radar observations per grid box? If 3 radar stations are really required, you would have to exclude a lot of data, or how dense is the radar network? And again, how are the S, C and X bands combined?

Here is an illustration of the French radar network, with radii of 100 km for C- and S-bands, and 50km for X-bands (for illustration purpose). The maximum range of our C- and S-bands radar is 255 km.



However, the higher the elevation angle, the lower the maximum range. To ensure enough coverage and avoid holes in the gridded data (especially on top at 10–13 km high), we determined that we needed at least the contribution of 3 radars on the considered domain (one domain per event), so that at least one elevation angle among all elevations available (from all radars) contributes to each grid box. This number was found to be a good compromise between data loss in the upper part of the grid and computational cost. Hence, we carefully selected the radars that were to contribute to the grid (in most cases we used more than 3). Since our events are centered on the domain, only the edges of the domain can suffer from gaps in the gridded data. We added "In other words, at least one elevation angle among all elevations available (from all radars) contributes to each grid point." in section 3.1 to make it clearer.

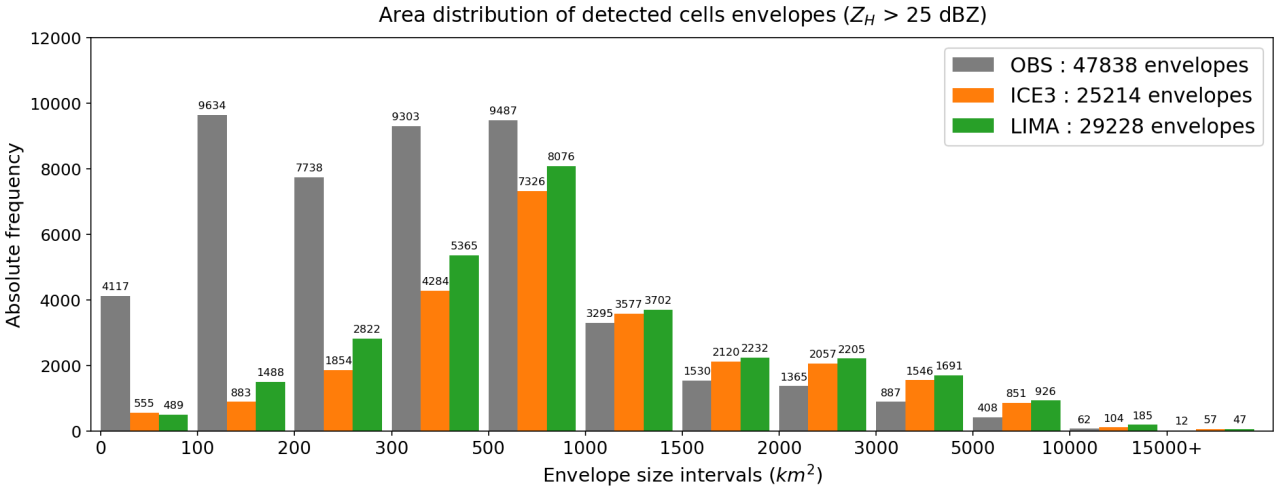
L259: Have you analyzed the impact of the lead time of your prediction? In other words, does it make a difference how long after the NWP start the event took place? I would imagine that your results could be different depending on the lead time, as the model tends to produce fuzzier predictions with longer lead times.
This is an excellent suggestion, as NWP forecasts are indeed usually evaluated in terms of their lead time (see for example Figure 10 of Brousseau et al., 2016 which show the HSS compared to persistence forecast averaged over different thresholds of 6h cumulative rainfall against rain-gauge measurements, as a function of the forecast range, for forecasts initialized at 0000 and 1200 UTC over a four-month convective period). However, we did not analyze the data given their lead time in this study, as the main objective was to focus on the ability of the model microphysics, combined with the forward operator, to accurately simulate the storm characteristics in terms of polarimetric radar data (and not in terms of location and temporality). Different configurations of AROME, with LIMA microphysics, initialized at 00 UTC are currently running in real time for a long-term evaluation over the whole France domain. With these new generated data, it will be possible to take lead time into account, but this is beyond the scope of the present study.

L357: 50 x 50 km box means that with the regrided radar resolution of 500m per lon/lat in each box there are 10000 values, right? And from that the 99 percentile is taken? And what is true for the model data with 1.3km resolution? Something like 1600 values per 50 x 50 km box? Why are you not using the same grid?
The horizontal resolution of our observation grid is 1 km (500m is on the vertical). So in a 50x50 km box, the 99th percentile is computed from 2500 obs values versus 1480 model values. However, what we are comparing in this section is AROME with ICE3 microphysics against AROME with LIMA microphysics. All contingency tables have been constructed identically, which means that the use of 1x1km observation QPE affects AROME-ICE3 and AROME-LIMA scores the same way.

L365: How is the bootstrapping done? Out of the 50x50 km boxes? How often?
We also calculated, for each rainfall threshold, the difference between the ICE3 and LIMA scores (i.e. $\Delta POD = POD_{LIMA} - POD_{ICE3}$, idem for FAR and HSS). The bootstrapping was performed on these ΔPOD/FAR/HSS values (for each threshold, one Δ value per case study = max 38 values). Note that for the highest rainfall thresholds, it was sometimes not possible to calculate a difference (e.g. no occurrences), which reduced the number of values used for the bootstrap (this number is shown in bold in our Figure 2). We used the `scipy.stats.bootstrap` function (see the description here) with `statistic=np.mean`, `confidence_level=0.95` and `method='basic'`. A precision has been added : "Additionally, a bootstrap test is performed using the scores' differences from all 38 cases to determine if the differences between ICE3 and LIMA are significant."

L386/ Figure 3: As far as I understand, there is a discrepancy between the resolution of the model and the observation grid. So does it make sense to compare the very small observed cells with the coarser model? I don't understand that here. And the discussion about the small cells being 'simulated too rarely' is then misleading. How should a model with a resolution of 1.3 km be able to simulate these very small cells (i.e. with a size of about 0 km^2)? I think there should be a discussion about model resolution in this section. The same goes for the discussion about lifetimes: I don't think AROME is able to predict the very small lifetimes. That is related to the model resolution.

We agree that our paper lack of a model resolution discussion. In our Figure 3, we displayed the convective cores size distribution, but you may be interested in the same plot for the cell envelope :



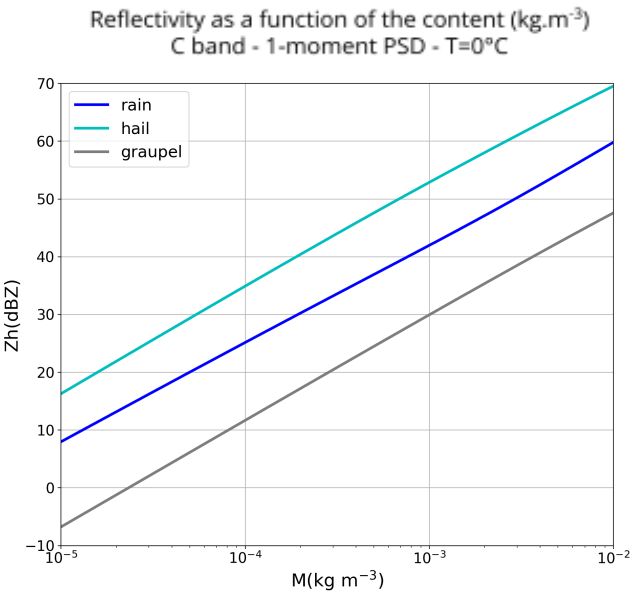
The difference in resolution between the observations and the model is unavoidable, as the best possible observations are usually preferred, especially in an assimilation context (which is our long-term goal). This paper is precisely about the possibility of an obs/model comparisons despite this inconsistency. However, Ricard et al. (2013) diagnosed the effective resolution of AROME to be of the order of 9-10Δx, where Δx=1.3km is the grid resolution. It implies that, in the above figure, objects of size smaller than 13x13 = 169km^2 cannot be fully resolved (neither in terms of size nor lifetime). A brief discussion will be added in section 4.2 : "This behavior is expected, as Ricard et al. (2013) diagnosed the effective resolution of AROME to be of the order of 9 − 10Δx, where Δx = 1.3 km is the grid resolution. This implies that objects of size smaller than 136 – 169 km2 cannot be fully resolved by AROME."

L416: Why should LIMA be compared to OBS_no_hail? LIMA is not directly including hail, but it is included in the graupel class.

This is right, in this study hail is not considered as a separate class but is instead included in the graupel class, which implies it has the parametrization of the graupel (same 1-moment PSD, fall speed, mass-size relationship, etc.).

Please find on the right an illustration of the theoretical output from the forward operator for ZH values as a function of the content (at C-band, for a 1-moment PSD and T=0°C).

Furthermore, the forward operator inherits the microphysics properties of the graupel to compute the resultant polarimetric values. How could we fairly compare observed reflectivities (due to rain and sometimes hail) with simulated reflectivity



values due to rain only ? Thus, we have chosen to distinguish between rain-only precipitation and total precipitation. We also added a precision regarding the “OBS_no_hail” distinction : “[Fig5] has been complemented by another curve (in gray) that only includes observed cores not associated with medium hail (5 − 20 mm) or large hail (> 20 mm) as detected by the A13 radar hydrometeor classification algorithm (see Appendix C of Forcadell et al., 2024).”

L425/Figure 6: Why CFADs instead of CFTDs (Contoured Frequency by Temperature Diagrams)? Maybe then the melting layer would be more sharp and the whole discussion in 4.3 more meaningful. I could imagine that the 44 events vary a lot in surface temperature and by that the distributions of observed pol. Moments in Fig. 6 are broader.
Thanks for this relevant question. We started the work with a grid in km altitude levels because of the radar data. The interpolation was easy from radar gates altitudes to regular gridded altitude data. Hence, we kept the altitude resolution for the model grid. As you rightly suggested, we are planning to switch to CFTDs for our current work on the forward operator, that more specifically focuses on stratiform events and bright band simulation.

Technical corrections:

L366/Figure 2: “The HSS score”: leave out “score” as it is already in HSS. Done.

Figure 9: “Altitudes are given in km AGL”: Is this a remark for Figure 8? Yes, now corrected.