Responses to the reviewers

Significance of microphysical processes for uncertainties in ensemble forecasts of summertime convection over central Europe

by Christian Barthlott, Beata Czajka, Christoph Gebhardt, and Corinna Hoose March 17, 2026

We thank both reviewers for reading the manuscript and providing detailed comments. We have carefully considered all comments and changed the manuscript accordingly. Please find below our responses in blue.

Reviewer 2

General Comments

The manuscript addresses a highly relevant and topical issue in convective weather prediction: the role of cloud microphysical uncertainties in ensemble forecast spread. Using a 108-member ensemble generated by perturbing key microphysical parameters, such as CCN and INP concentrations, graupel sedimentation velocity, and droplet size distribution, the authors systematically explore how these perturbations affect convective precipitation intensity, spatial distribution, and ensemble spread in convection-permitting simulations with the ICON model.

The study makes several important contributions:

- It demonstrates substantial variability in precipitation outcomes attributable solely to microphysics perturbations, despite identical initial and boundary conditions, highlighting microphysics as a major source of forecast uncertainty.
- Results linking cold-rain vs warm-rain process dominance to aerosol conditions offer valuable physical insight and suggest new diagnostics for regime classification.
- The comparison between large ensembles and operational-scale subsets meaningfully illustrates how ensemble size influences forecast spread representation.

Overall, the methodology is innovative, and the results are of broad interest to both weather prediction and microphysical parameterization communities. However, key areas require clarification and expansion before the manuscript can be considered for publication.

Major Comments

1. The paper convincingly shows that perturbations in CCN, INP, and sedimentation rates significantly alter precipitation totals. However, the mechanistic explanations linking these perturbations to changes in convective evolution (e.g. why cold-rain dominance increases with higher CCN/INP) could be expanded to provide a more complete physical narrative.
   Thanks for this comment. We expanded the ICON model code to write out microphysical process rates with the intention of quantifying the physical processes and their impact on precipitation totals. Already in the description of the precipitation deviations in section 3.2, we referred to changes in process rates affecting precipitation totals, e.g. the reduction of the warm-rain processes or the reduction of riming with increasing CCN concentrations. We made similar explanations in section 3.5 about the relative role of warm- and cold-rain processes. For example: *"Physically, higher CCN concentrations delay warm-rain initiation via suppressed autoconversion and reduced collision-coalescence efficiency. As a result, more cloud water is available to be lofted into the mixed-phase region, where it can participate in cold-rain processes."* We added

some more statements in the text to better explain why individual processes change with our perturbations and what that could mean for precipitation totals.

2. The analysis uses four convective cases over central Europe. It would help the reader if the authors justify why these specific cases were chosen and whether they capture a wide range of synoptic environments. Are the conclusions generalizable beyond these particular events?
These cases were chosen as they took place during two field campaigns conducted in southwestern Germany. These campaigns were led and conducted by KIT, which means that we know about the occurrence and evolution of clouds and precipitation. As already indicated in our manuscript, these cases include single cells, multicells, supercells, and larger convective clusters in weather regimes with different strengths of synoptic forcing. The synoptic weather charts presented in Fig. 3 document the different strengths of synoptic forcing. We therefore believe that the case selection is representative of summertime convective situations in that area. Even if only 4 cases have been studied, we believe that by analysing four different typical cases, the conclusions about the importance of microphysical uncertainties are most likely generalizable.

3. The comparison between the 108-member microphysics ensemble and operational ensemble spread is valuable. Yet, the operational ensemble design (physics schemes, perturbation strategies) should be described in more detail to support a fair qualitative comparison. Differences in model configuration, resolution, or perturbation strategy could confound interpretations.
Section 2.3 describes the ensemble setup used at DWD and already gives information about their perturbation strategy. We added some statements about the physics schemes and further model settings to better demonstrate the similarities/differences between our ensembles.

4. In several cases presented (e.g., as noted by existing RC1 comments), peak precipitation occurs near the end of the simulation window (e.g. Figure 7). This raises the question of whether the full lifecycle of convective events is captured for all ensemble members. Analyses truncated before event completion could bias statistical measures.
As reviewer 1 raised this point as well, we can repeat our reply here: We agree with the reviewer that in three cases the rainfall event did not come to an end. In two of these cases (cases 1 and 4), however, there is a morning/midday precipitation peak, which is covered in our 24-hour runs. In each case, the heavier precipitation occurs in the evening. An important point is that the maximum of the event is always captured in our 24-hour runs, and the precipitation rates are already declining at the end of the integration time. One reason why we only simulated 24 hours is the fact that we did not want to have a very long lead time. In operational forecasts, new (ensemble) runs are started every one to three hours to keep the integration times relatively short. With 24 hours, we tried to find a compromise between not deviating too much from reality and taking into account at least the maximum convective activity.

   To be able to access the impacts of not covering the entire convective event throughout the night, we redid all 108 simulations with a longer integration time of 30 h for case 1 (23 June 2021, Fig. R.1). Unfortunately, the compilers used for these simulations are not available anymore on our HPC system. So the re-compilation of the model code with new compilers could also lead to slightly different outcomes. It turns out that although the same compilers are not available anymore, the sensitivities to the applied perturbations are not changed. In Fig. R.2, the general precipitation characteristics with respect to our microphysical perturbations remain the same even if individual points are not identical.

   We also recalculated the ratio of the cold to warm rain processes with the 30-h simulation data and found an almost identical characteristic behaviour (Fig. R.3). As the sensitivities to our microphysical perturbations are robust and do not change for the longer integration time for case

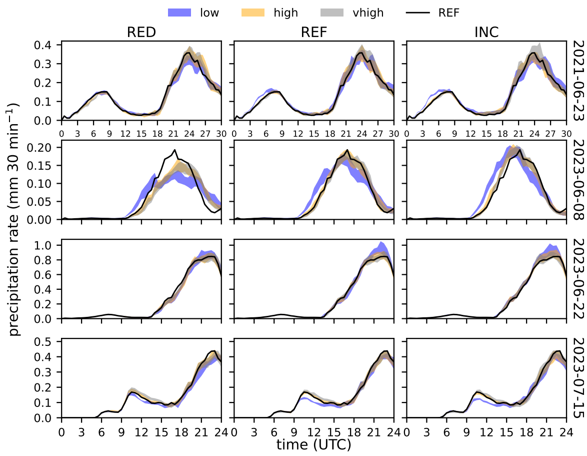


Figure R.1: Domain-averaged precipitation rates for reduced (left), reference (middle), and increased graupel sedimentation (right). Top row is for 30-h integration time.

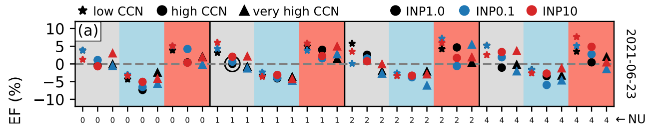
1, we decided not to redo the other three cases, which would require another 130k CPU h for the full ensemble. In section 3.2 of the revised manuscript, we now mention that a longer simulation time, including the decaying stage of the convective events, did not change the sensitivity to our perturbations.

5. Introducing additional objective ensemble verification metrics (e.g., continuous ranked probability score, rank histograms, reliability diagrams) would contextualize the spread relative to forecast skill and probabilistic performance, beyond just spread magnitude.
   We fully agree that, for a mature forecasting system, such scores are indispensable to quantify the relative quality of two ensembles. Unfortunately, the present study cannot provide a meaningful answer to the question of which of the two ensembles performs better. With only four cases, the sampling distribution of any verification score is extremely broad. Even a perfectly calibrated ensemble may obtain a very low (or very high) score on a single day by chance. Consequently, a direct comparison of the two ensembles would be dominated by sampling noise rather than by systematic differences in model physics.

   To demonstrate the large day-to-day variability, we have computed the Fractions-Skill-Score (FSS) of 24-h total precipitation amounts compared to radar-derived precipitation for the two ensembles (Fig. R.4). Across all four days, the spatial distribution of the FSS is broadly similar between the two ensembles, indicating that MPHYS-ENS is capable of generating a level of skill comparable to that of the operational EPS. Nevertheless, marked day-to-day variability is evident. For the 8 June 2023 case, the scores are uniformly low, reflecting poor agreement with observations, whereas the 22 June 2023 case exhibits consistently higher FSS values, signalling substantially better performance. The 23 June 2021 and 15 July 2023 cases fall in between, with mixed regions of high and low skill. These plots illustrate that, while the ensembles can achieve comparable FSS magnitudes, the skill is highly sensitive to the specific weather situation. Consequently, any quantitative assessment of relative performance would require a much larger sample of cases; the present four-day set primarily serves to demonstrate that the spread produced by MPHYS-ENS is of the same order as that of ICON-D2-EPS.

   Because of the points listed above, we deliberately refrain from presenting CRPS, rank histograms, reliability diagrams, or FSS as definitive performance indicators for MPHYS-ENS in the manuscript. Doing so would give the impression that the four case-studies are sufficient to draw statistically sound conclusions about the relative quality of the two ensembles – which they are not. We have included this text after the ensemble spread analysis:

(a) 30-h integration



(b) 24-h integration

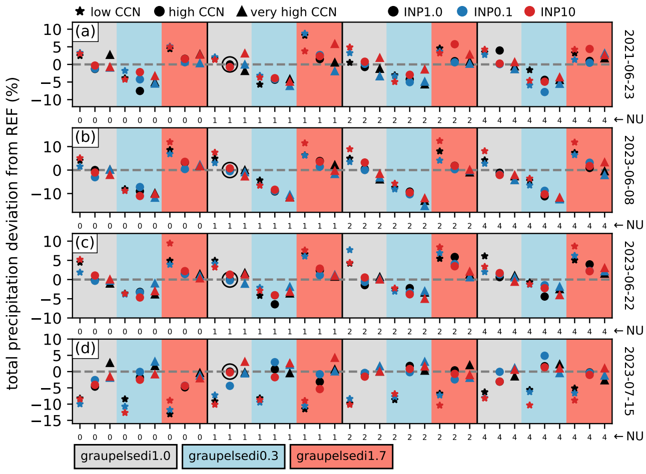
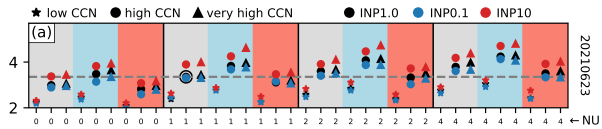


Figure R.2: (a) Precipitation deviation from the respective reference run (marked with a black circle) for case 1 with 30-h integration time and (b) all cases with 24-h integration time.

(a) 30-h integration



(b) 24-h integration

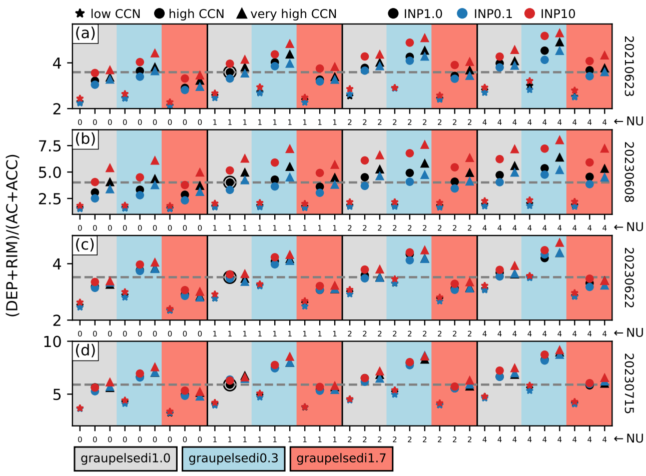


Figure R.3: Ratio of cold-rain formation (deposition DEP and riming RIM) to warm-rain formation (autoconversion AC and accretion ACC).

> *"The present study is based on four 24-h precipitation events. For such a small sample, probabilistic scores (e.g., continuous ranked probability score, rank histograms, reliability diagrams) do not allow a statistically robust comparison of MPHYS-ENS and the operational ICON-D2-EPS. Nevertheless, we have computed the Fractions-Skill-Score (FSS) of 24-h total precipitation amounts compared to radar-derived precipitation (not shown). Across all four days, the mean values for the selected precipitation thresholds and scales of the FSS is broadly similar between the two ensembles, indicating that MPHYS-ENS is capable of generating a level of skill comparable to that*

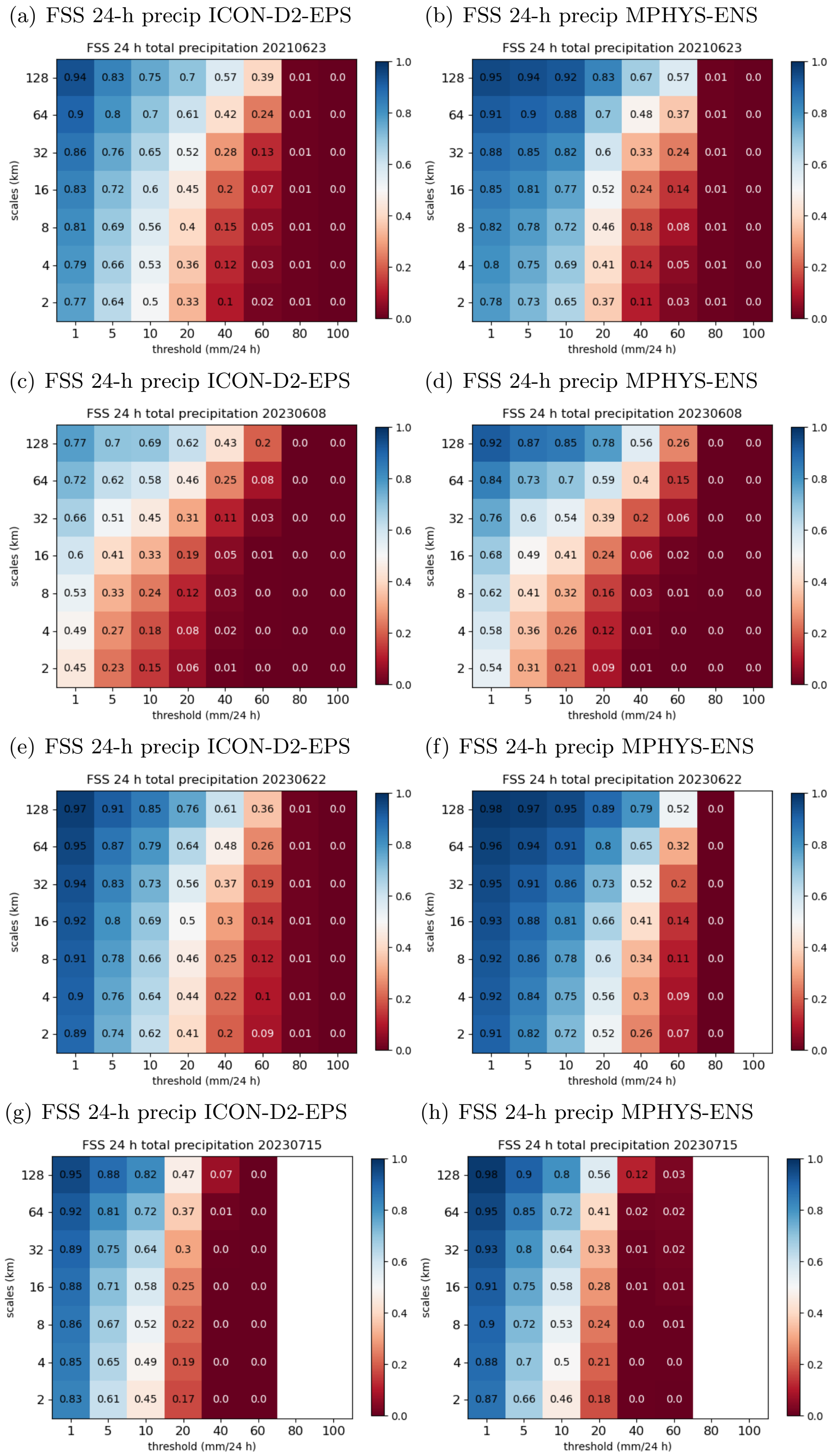


Figure R.4: (a) and (b) 2021-06-23, (c) and (d) 2023-06-08, (e) and (f) 2023-06-22, (g) and (h) 2023-07-15.

of the operational EPS. Nevertheless, marked day-to-day variability is evident. For the 8 June 2023 case, the scores are uniformly low, reflecting poor agreement with observations, whereas the 22 June 2023 case exhibits consistently higher FSS values, signalling substantially better performance. The 23 June 2021 and 15 July 2023 cases fall in between, with mixed regions of high *and low skill. While the ensembles can achieve comparable FSS magnitudes, the skill is highly sensitive to the specific weather situation. Consequently, any quantitative assessment of relative performance would require a much larger sample of cases; the present four-day set primarily serves to demonstrate that the spread produced by MPHYS-ENS is of the same order as that of ICON-D2-EPS. "*

We trust that this explanation clarifies why a conventional verification analysis would be misleading for the dataset currently at hand, and why we focus on spread as a qualitative indicator of the ensemble's ability to represent forecast uncertainty. We hope that the added illustrative figure and the revised manuscript text address the reviewer's concern while preserving the scientific integrity of the study.

Minor Comments and Suggestions

1. The authors should situate their approach relative to prior work perturbing similar microphysical parameters (e.g., Thompson et al. 2021-style approaches), to better frame the novelty of the perturbation strategy.
   This was also suggested by reviewer 1, and we should have mentioned that paper already in our first manuscript. We have now inserted some sentences in the introduction about their perturbation strategy and findings.

2. Ensure consistent use of terminology (e.g., "cold-rain pathways," "warm-rain processes") and define early how these categories are diagnosed.
   Thanks for pointing that out. We now consistently use either "cold/warm-phase pathways" or "cold/warm-rain processes" and also include explanations of those earlier in the text.

3. The finding of rapid domain-wide error growth even in cloud-free regions is important. A discussion linking this to broader convective-scale predictability limits and non-local coupling would enhance the manuscript's implications for ensemble design.
   We also consider this an important finding and added a discussion about that point in the conclusions.

Thompson, G., Berner, J., Frediani, M., Otkin, J. A., & Griffin, S. M. (2021). A stochastic parameter perturbation method to represent uncertainty in a microphysics scheme. Monthly Weather Review, 149(5), 1481-1497. https://doi.org/10.1175/MWR-D-20-0077.1