General comments. This paper uses polarimetric radar observations to test simulations using various microphysics schemes for summer convective cases over Germany. A major strength of the paper is the statistical evaluation over many cases in contrast to the approach often employed of focusing on a single case study. This topic is important as the representation of microphysics is a major source of uncertainty in NWP models. The paper is reasonably well written and the conclusions seem to be robust. My main concerns on the science center on 1) consistency between the radar forward simulator and internal assumptions and characteristics within the microphysics schemes, and 2) inconsistencies in the analysis of P3 (primarily for graupel/hail). These concerns are detailed in major comments below. I also include several minor comments, mainly to clarify certain issues or improve the presentation. Finally, a handful of technical/editorial comments are included at the end. Overall, I think the paper could be acceptable if the main comments are addressed.

Recommendation: *Major revision*

Major comments.

1. One of my main comments concerns the consistency of assumptions and ice properties internal to the microphysics schemes with assumptions in the radar forward operator calculations using CR-SIM. It's mentioned a few times that these calculations are consistent, but in my view more detail is needed on this. For instance, what ice particle properties are needed or assumed by the forward operator calculations? I guess this includes particle density, size distribution information, and particle habit? Are there other inputs or assumptions made about ice particles in CR-SIM? I think it should be straightforward to couple the traditional bulk schemes (Thompson, Morrison) with CR-SIM. However, this seems less straightforward for coupling with SBM and particularly P3. In P3, the particle density (e.g., particle mass-size relation) varies across different regions of the size distribution. How was this accounted for? Particularly relevant for this paper, in all the schemes what is assumed for the distribution of liquid water on melting ice particles? This is particularly important in this paper given the focus on radar signatures of hail and particularly the reflectivity bias in all schemes in conditions of large hail (i.e., high reflectivities).

Perhaps these issues are discussed in some of the previous papers on CR-SIM, but more discussion is needed in this paper.

2. Figure 5 and ~p. 16. I think the analysis here is a misinterpretation of graupel and rimed ice as simulated by P3. Graupel-like or even hail-like ice particles in P3 don't necessarily consist wholly of rimed ice mass, -- they may contain substantial mass grown by vapor deposition as well (but still have high particle densities expected for graupel/hail). Thus, I think it's incorrect to only include rimed mass for the mixing ratio threshold in Fig. 5 for P3. If the hydrometeor ID identifies a particular time/location as hail/graupel from the P3 output, I would include all ice (rime plus vapor grown) in the analysis. This is particularly relevant to the analysis of graupel/hail since there is no separate rimed ice category in P3. It seems quite likely that ice

present at the 1 km level in P3 for summer convective cases will be fast falling and with a high rime fraction (and therefore hail-like) anyway.

3. I didn't see any discussion on the scale of the radar observations versus the model, other than discussion of 400 m grid spacing possibly explaining the low bias for the instances of high reflectivity associated with graupel/hail. Is there a general scale mismatch between the radar and the observations? Any scale mismatch is likely to be quite important for instances of heavy rain and hail. If the radar data is higher resolution, can it be appropriately averaged to give a comparable scale to the model data? Keep in mind that the effective resolution in models like WRF is about 5-7 times the grid spacing (Skamarock 2004, MWR).

4. The SBM results from Figure 4 look rather strange, in particular, the sharp drop at diameters just above 4 mm. What is the explanation for this? Does it have something to do with how breakup is treated in this scheme? Or is this the maximum size of bins in the scheme?

Minor comments.

Abstract, line 7. "heavy rain events" is imprecise. Can you state here the specific rain rates used to define such events?

Line 26. I disagree that increasing resolution eliminates problems caused by inaccurate parameterizations, it addresses *some* of these problems. Thus, I suggest adding "some" before "inaccurate parameterizations".

Line 32. Add "usually" before "predict", since not all bulk schemes assume a predefined statistical function for the particle size or mass distributions (a few schemes predict processes directly from moments without assuming any functional form of the size distribution, e.g. Kogan and Belochitski 2012, JAS).

Line 41. Not sure I agree that it is not known exactly which processes are poorly represented and I don't think the paper cited here (Morrison et al. 2020) makes this argument either – we have some idea of which processes are most uncertain or most poorly represented. Perhaps reword this sentence to "It is known that many processes, especially those involving ice microphysics, are poorly represented in numerical weather prediction models (Morrison et al., 2020)."

Line 151. I'd add "single" before "case study", as I think this states the point here better.

Line 184. Not clear what you mean by "The missing information about the area of rain events is presented in the top right of Fig. 3." What is missing? I think what you mean is simply "The area of rain events is presented in the top right of Fig. 3."?

Line 188. Confusing as written. Suggest removing "number of".

Line 240. Are the differences really “astounding”? Maybe “substantial” ort “major” would be better?

Lines 244-246. You might note that these relations of mass and reflectivity to diameter are true for liquid drops (or more generally, isometric particles).

Line 247. Technically, all of the bulk schemes here use complete size distributions, meaning they extend mathematically from 0 to infinity. Thus, it’s better to just say “..few large particles may contribute significantly…” rather than “few or no large particles…”.

Line 268. I think this could be reworded more clearly – I suggest replacing “due to the missing large raindrops” with “due to the lack of large raindrops”.

Line 290. Technically droplet fall velocity and droplet size distribution are not processes. Suggest rewording this to “is affected by processes such as evaporation and drop sedimentation, …”

Line 300. They don’t evaporate faster because of high surface tension, it’s because evaporation in the schemes depends mainly on the number concentration times the mean radius (referred to as the integral radius), with some additional modification to account for ventilation.

Lines 315-316. This is not correct – the P3 scheme simulates not just number and mass mixing ratios of ice. By predicting additional ice attributes, it can distinguish between graupel-like and hail-like ice (for example, by differences in mean density, size, and fallspeed).

Lines 331-332. This is confusing. I’d reword to “a grid spacing less than about 250 m is required…”.

Line 333. I’d add “when” before “further”.

Line 339. Note that a multiple “free” category version of P3 exists (see Milbrandt and Morrison 2016, JAS). Thus, I’d reword this to “…this version of the P3 scheme uses only one ice class…”.

Line 378. “likely” seems too strong of a word to use here. Perhaps reword to “which might be a resolution problem”. Same comment on line 415.

Technical/editorial comments.

Line 4, abstract. Suggest adding “the” before “observation dataset”.

Line 44. Perhaps replace “just this” with “such”.

Line 78. Replace “are” with “were”.

Line 84. Add “schemes” after “microphysics”?

Line 99. The first “is” should be “are” (data here is plural).

Line 179. I feel “then” could be removed.

Line 226. Replace “it is relying on” with “it relies on”.

Line 227. “mixing ratio” should be “mixing ratios”.

Line 260. I think “github” should be “GitHub”?

Line 260. Reword to “Because the bulk schemes do not actually have fixed size bins…”

Line 272. “simulate” should be “simulates”. Alsop “produce” should be “produces”. Same comment on the next line (line 273) as well.

Line 278. “is” should be “was” and “note” should be “noted”.

Line 279. “attribute” should be “attributed”.

Lines 277-281. This sentence is very long, perhaps break it up into 2 sentences.

Lines 281-286. I’d suggest using past tense in the writing here, since you’re describing what previous studies found.

Line 297. Add a comma after “schemes”. Also, replace “is getting smaller” with “becomes smaller”.

Line 298. “schemes” should be “scheme”.

Line 301. I think you can remove “also”.

Line 312. Add “are” before “of interest”.

Line 331. There’s an extra right parenthesis after “(2015)”.

Line 332. Space is missing between “area” and “converge”.

Line 333. Remove “with”.

Line 335. I think “reflection” should be “reflectivity”?

Line 338. I'd replace "from" with "with".

Line 343. Replace "distribution" with "distributions".

Line 374. Typo: "is" is repeated twice.

Figure 3 caption. "Oue et al. (2020)" should be "(Oue et al., 2020)". Same comment with the Figure 5 caption as well.

References.

Kogan, Y. L., and A. Belochitski, 2012: Parameterization of cloud microphysics based on full integral moments. *J. Atmos. Sci.*, 69, 2229-2242.

Milbrandt, J. A., and H. Morrison, 2016: Parameterization of cloud microphysics based on the prediction of ice particle properties. Part 3: Introduction of multiple free categories. *J. Atmos. Sci*., 73, 975-995.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, 132, 3019-3032.