Comment on essd-2021-427
Leif Denby (Referee)

Schulz presents an in-depth and well-written description of a very useful dataset of manual classifications of shallow cumulus clouds covering the EUREC4A campaign period. The comparison with more traditional techniques for measuring organisation provides a valuable reference for interpreting the manual classifications. And contrasting the results of using different datasources is very valuable. The detail with which the dataset is described and the openness by which the tools used have been shared is commendable and inspiration for our community as a whole.

There are two aspects I would find very valuable to giving a little more depth. First, is the definition of "truth" in this manual classifications. This would focus on answering questions such as "which of workflow dataset should we trust as the most truthful and why?" "is it possible to produce a kind of consensus among the four workflows?" This could draw in prior studies using manual classifications referenced in this publication. Second, currently there is little analysis of the manual classifications created for the simulation data (as compared to the observation-based workflows), why is this and was this analysis done?

Detailed comments

- [p 1 l 21] "studies concentrated on the classification of meso-scale patterns...":
  
  - You could also mention Denby 2020 here as an example of using a neural network for classification. It would also be good to mention Wood & Hartmann 2006 here already. You should also include reference to Janssens et al 2021
as it quickly turned out that the identification of the patterns in the model simulation was too demanding. The features had too little similarity with those found in nature.

I think this needs reformulating. “Accumulated intentionally” isn’t so clear, you mean that people tended to label the observations and not the simulation output? You could refer to the totals in Figure 2 to make this point. I would also emphasise that you are continuing to talk about classifications made in the ICON workflow in the following sentences. Because the you lead the paragraph with the total number of classifications I initially read this as if “sugar” was hard for people to identify across all workflows (but that isn’t the case I think, cf Figure 2)

This process eliminates overlaps of same-user classifications for each pattern and turns the data into masks, rather than coordinates (see Fig. 3):

What happens if the user draws two bounding-boxes with different types of classification that spatially overlap? (It shouldn’t really happen, but maybe you could just state that it didn’t)

Figure 4:

There is a bit of aliasing for the text in figure 4. Maybe storing as a pdf/svg would be better?

Comparison with other classifications:

There doesn’t seem to be any analysis of the ICON simulation classifications. It would be good to mention why this was left out.

“org, Tompkins Adrian M. and Semie Addisu G. (2017)) with the mean cluster size (S)"
- In choice of these metrics maybe you could mention how this fits with the findings for Janssens et al 2021?

- [p 8 l 136] "we focus on a domain size of 10×10 degrees to do the comparison":

  - I would rephrase to be clearer to something like "we calculate this metrics over a 10x10 degree sub-domain". I would say “Specifically” rather than “Precisely” next. Also 10N to 15N is only 5 degrees, should it be 5N to 15N?

- [p 8 l 134] "detect the patterns in geostationary infrared images of 135 GOES-16 ABI (Schulz et al., 2021)“:

  - Does this mean that the network was trained on the dataset from the EUREC4A IR (channel 13) workflow to predict the manual classifications (masks) created by participants in the same workflow? It would be helpful to emphasise this, and specifically in say the caption of Figure 7, to so that the neural network was trained on the IR data (which would also explain why the agreement is better with the IR rather than the visible manual classification)

- [p 8 l 144] "we expect Gravel and Flowers to be rather regularly distributed and therefore to 145 have a lower Iorg compared to Fish and Sugar”:

  - I don’t quite understand why “flowers” should be more “regularly distributed” than “sugar”? My understanding is that “sugar” is scatter small cumuli which I would expect to be very regularly distributed.

- [p 8, Figure 6] "mean cluster size (S)“:

  - What are the units of "mean cluster size"? Are they pixels?

- [p 9 l 150] "we applied a threshold of 0.1 on the frequencies“:

  - This is a bit unclear. Above you talk about the “percentage of agreement” but here of “frequencies of the level 3 dataset”. How do you go from
“percentage of agreement” to “frequency”? Are they the same? Does a threshold of 0.1 mean that only 10% of participants needed to say that they label an area as a given pattern? Is that a reasonable number? It seems to quite low to me, wouldn’t that lead to very large masks? What happens if two users classify a given pixel with two different labels (I don’t think that is taken into account with the current calculation, but the word “frequency” suggests to me that it should)?

- Also, having a cumulative area fraction larger that 1.0 is quite confusing. It would be good to discuss what that means and why it is reasonable. How would I read from Figure 7 what area fraction is unclassified?

- [p 9 l 162] "While the Iorg/S metric is computationally cheap and can be easily applied to different regions, the manual classifications are naturally more accurate":

- This doesn’t quite follow for me. Figure 7 isn’t attempting to produce classification into the four organisation patterns using only Iorg/S, and so I don’t think this analysis shows that it isn’t possible with just Iorg/S.

- [p 9 l 164] "manual classifications are most 165 accurate":

- I would like to understand a little better how you draw this conclusion. What is your measure of accuracy? Aren’t the manual classifications being used as “truth” here? If so, and assuming that any other method will classify differently in some way, how can any other method predict something better than what is being used as the reference?

- [p 10 l 179] "demands further investigations on how":

- should be "demands further investments into how..."