Review of Gryspeerdt, et al. - The impact of sampling strategy on the cloud droplet number concentration estimated from satellite data

This paper compares MODIS droplet concentration retrievals with those from several aircraft campaigns and calculates r^2 values for several different sampling strategies for the MODIS pixels. The quality of the writing is good and a good number of field campaigns are included. The field was in need of such comparisons between in situ and satellite Nd data since previously this had only been done for a few campaigns. The overall scientific quality is good and lots of interesting aspects are explored. However, there are a few mistakes (concerning the values used for filtering relative to those suggested in the cited literature) and an over reliance on the r^2 metric, which is not always informative. Additional metrics (mean bias, RMSE, etc.) should be calculated too. The aircraft sampling method is also quite different to previous comparisons leading to more datapoints but without the ability to determine the root cause of any discrepancies (since profiles are not used). This is fine, but it would be good to explain the differences with previous work (where higher r^2 values were found) and to compare the effects of averaging larger numbers of pixels (e.g., 3x3, 5x5) rather than jumping to 100x100. I would also like to see uncertainties for the r^2 values since it is unclear whether some of the differences in r^2 are significant and there is likely a degree of randomness.

Here is a list giving more detail on the above and some additional comments :-

Grosvenor (2018) suggested that solar zenith angle was restricted to below 65 degrees not 55. And it was 55 degrees for the viewing angle, not 41.4. This will lead to quite a lot more data being discarded than would be the case if using the correct values. What effect does this have?

For the G18 sampling there are lots of factors being applied at once – it would be better to test them individually. Which has the largest influence? It would also be useful to test BR17 and Z18 separately from G18.

Can you provide uncertainties for the r^2 values? These would help determine the likelihood of the differences between the different sampling strategies being due to chance.

Is r^2 the best metric to use? And you should provide more details on the particular r^2 value that you are using – is it appropriate for the data populations? Some of the r^2 values don't seem to match with what might be expected in Figs. 3 and 5 (admittedly judging by eye). E.g., are the COPE r^2 values really so low? Can you double check? Adding the uncertainties would help here. It would be helpful to also plot the lines of best fit against which the r^2 values are calculated rather than just the 1:1 line. For Fig. 3 you should also indicate the BR17 points on there – maybe you could use crosses in the middle of the colours or something? Or maybe provide them all in separate plots in the Supplementary?

It would be useful to test how removing random data points affects the r^2 to get an idea of whether some of the results are due to chance rather than the particular sampling strategy.

Why just use r^2? Metrics other than r^2 should also be provided - e.g. mean bias, RMSE, etc. For the E-PEACE results in Fig. 3 for example the points removed by G18 sampling may make the r^2 worse (although perhaps due to the whole cluster at the top left being removed), but are likely to reduce the bias.

Table 2 – it would be useful to provide the number of samples in each case.

Compared to the in-situ comparison studies of Painemal (2011) and Kang (2021) you have many more datapoints for your comparison. Presumably this is because they selected entire profiles from the aircraft data and then found the 5x5 satellite pixels that were collocated (within an hour), but with a calculation to account for the movement of the cloud (using the wind speed and direction). It would be useful to comment on this in the paper. Those studies seemed to find much better r^2 values than you. Use of individual MODIS pixels also seems to lead to repeated sampling of the same regions of clouds (looking at Fig. 3), which may skew the statistics somewhat. Perhaps it would be better to average over a few pixels, e.g. 5x5 as in Painemal and Kang to see how this changes the results? Especially since the satellite and aircraft will not be precisely collocated.

Is there an effect from not correcting for the wind speed and direction between the aircraft and satellite observations as done in previous studies?

The results in Fig. 7 are a bit strange given the results from the G18 review where the BR17 retrievals (although without the G18 sampling applied here) tended to be lower than those from the other dataset tested (based on Grosvenor and Wood, 2014). Can these results be explained?

Line by line comments

p.1 L13 – "As the first moment of the droplet size distribution, the N_d is important for setting cloud and precipitation process rates"

Nd is the zeroth moment, not the first. Plus the fact that it is the zeroth moment doesn't seem that directly relevant for setting cloud and precipitation rates – perhaps it would clearer/relevant to say that Nd helps to determine the droplet sizes (or something similar)?

P1, L18 – you could also add https://doi.org/10.1029/2020MS002126 and https://doi.org/10.5194/acp-20-15681-2020 as examples of GCMs being evaluated using Nd.

p.2, L6 – Is Hasekamp available yet?

P3, L12 – "but the temperature dependence can produce a 50% variation in the Nd"

It would be better to more specific here – is this for temperatures typical of the range of cloud temperatures encountered? All clouds, or just shallow ones, etc.

P4, L1 – "and a degeneracy in the retrievals for a low re"

Needs some explanation of what this means in this context.

P4, L5 – "generate uncertainties, particularly in the re" – the Grosvenor paper actually showed larger optical depth effects than re at high solar zenith angles.

P7, L15 – can you provide more details on the particular coefficient of determination metric used. Is it the square of the correlation coefficient (which method)? Or is some other metric used?

Table 1 – this lists the 5km CF > 0.9 as a sampling criteria for G18, but the text suggests that this is not the case?

Fig. 3 – the green and orange dot colours don't stand out as being different enough from each other. Can you change the colour something else? Black would work well I think.

Eqn. 3 – can you provide a derivation of this please, or a reference? Using the definition of beta in Eqn.2 seems to require dividing by A when I derive it. Or using ΔlnA instead of ΔA. Does this affect the results?

Fig. 4b – it would be more useful if you tested the G18, BR17 and Z18 sampling without the SPI<30 restriction to see the variation across the full range.

Fig. 5 – why the sudden switch to using 1.6um retrievals?

P14, L4 – "In these regions, clouds are much more likely to be adiabatic (and hence satisfy the BR17 re stacking criterion)."

The 2.1 and 3.7um retrievals actually are usually quite close to each other in stratocumulus regions (e.g., see Fig. 1 of Painemal 2011), or if anything re2.1 > re3.7 and yet there is a general good match between in-situ and satellite Nd. This raises some issues with the use of the stacking sampling.

Typos

"Insitu" – should be two separate words "*in situ*" usually in italics.