Reviewer #2

General comments:

The Introduction is now more clear, more focused, and much improved. The simple example added to the discussion of bootstrap sampling is helpful and will make it clear to the community why bootstrapping was used to compute the performance metrics. Overall, the manuscript is much improved but a few additional changes are necessary to be ready for publication.

Specific comments:

Line 168 – the criteria for classification as Category 6 (pressure at cloud top less than 440 mb and non-opaque) should be mentioned here so the reader understands how this class is selected.

Lines 174-175 – When the CAD algorithm gives a CAD score near zero, the algorithm finds the probability of aerosol and the probability of cloud are nearly equal. This is most often because the detected 'layer' did not match the characteristics of either an aerosol or a cloud, often because the detection algorithm triggered on a noise spike or other signal artifact and not on an actual aerosol or cloud layer.

Line 288 – I found the use of "pixel" here to be confusing. I think this refers to a 5-degree lat-lon grid cell.

Lines 301-305. Sassen (2009) used an earlier version of the cloud product and also used different criteria for selecting and screening cloud layer data. Both of these likely contributed to differences when compared with the later results. Higher cirrus occurrence at night is primarily due to better sensitivity due to a lack of solar background (see Winker et al. 2024). The true diurnal difference in cirrus occurrence is complicated, as convective clouds have different diurnal cycles depending on geographic region. The day-night difference in background noise likely produces an artificial diurnal difference which outweighs the true diurnal differences.

Lines 317-319 – Yes, lidar systems are more sensitive to optically thicker clouds, but they also have much greater sensitivity at night due to a lack of solar background and higher signal-to-noise ratio. Whether higher frequency of cirrus detection at night is (partly) due to increased nighttime optical depth is open to debate.

Line 325 – I am still confused by "parameters that precluded the use of the test", used in line 325 and the caption of Table 2. What does 'parameters' refer to? To me, a parameter is something like reflectance or radiance. The authors provided a clear response to my previous comment on this: By the phrase "precluded the use of the test," we meant that the specific indicators in question reach values that, in our judgment, make it impossible to use these tests directly for identifying Cirrus cloud masks. Given that the tests indicated by numbers in bold do not help in identifying cirrus for the cloud mask, do the bolded numbers in Table 3 give us a threshold value of the metric (ROP, POD, FAR, etc) where the test is not useful below (or above) that threshold? A little more explanation is necessary.

Line 471 – Figure 10 shows results as a function of cloud optical depth, up to an optical depth of 10. Winker et al. (2024) points out that CALIOP retrievals of cirrus with optical depth become very uncertain

when the optical depth is greater than 2 or 3. Optical depth uncertainty can grow to much larger than 100%. The authors should consider whether this large uncertainty at large optical depths might impact the results shown.

Technical corrections:

There are several instances of 'p.p.' which I think should be '%'

Line 181 – polar orbits with 16-day revisit cycle

Line 182 – CALIPSO followed the Aqua spacecraft

Line 188 – only the 5 km product

Line 258 – The balancing of the sample …

Line 261 – 'more accurate results', 'more insightful results', rather than 'more reliable'?

Line 417 – indicates a very high level …

Line 419 – high values of POD are observed ?

Reference:

Winker, D., X. Cai, M. Vaughan, A. Garnier, B. McGill, M. Avery and B. Getzewich, 2024: "A Level 3 monthly gridded ice cloud dataset derived from 12 years of CALIOP measurements", *Earth Syst. Sci. Data*, **16**, 2831–2855, https://doi.org/10.5194/essd-16-2831-2024.