

Response to: Reviewer #1

Overall evaluation: In this manuscript, the authors carried out a series of validations and experiments on three different DCC classification algorithms against observations from CloudSat-CALIPSO lidar-radar observations. This research would be helpful for gaining insights on how the popular algorithms perform with different thresholds and over different regions. The major drawback is on the paper writing, especially in the introduction part which consists numerous unclear descriptions, including but not limited to my specific comments #1~12 below. Therefore, I would recommend a minor revision from the authors to make this manuscript better reader-friendly to the general peers in and outside the relevant field before it gets published.

Specific comments:

- Line 24, ‘Although the least-frequent cloud type’ should be ‘Although/Despite being the least frequent cloud type’.

Corrected, as suggested.

- Line 28, there should be a comma between ‘Agency’ and ‘economic’.

Corrected, as suggested.

- Line 30, ‘As global climate warms’ should be ‘With global warming’.

Corrected, as suggested.

- Line 38, ‘limit spatial coverage’ should be ‘the spatial coverage is limited’.

Corrected, as suggested.

- Line 38-39, I believe that ‘only useful when operated by a human’ can be part of ‘limited spatial coverage’, since it is aforementioned as ‘manual observation’.

Rephrased, as suggested.

- Line 41, it should be like ‘imagers that provide frequent observations with global coverage’.

Corrected, as suggested.

- Line 52, it should be ‘Here, Cirrus pose a threat as well’ to be more formal in writing.

Corrected, as suggested.

- Line 54, for ‘spectral radiances’, do you mean ‘multispectral radiances’?

Yes, ‘multispectral’ sounds more appropriate.

- Line 57, ‘measures’ should be ‘measurements’.

Corrected, as suggested.

- Line 75-76. This is a relatively big statement which needs some reference citing to support it, like how even observations from CALIPSO cannot be regarded as direct observation or ‘ground truth’.

Unfortunately, this is true in cloud studies. There is no 100% reliable source of reference data for cloud detection and classification. The only solution is to use a reference dataset that is believed to be more accurate than others. That said, ‘believe’ is not a matter of guessing, or a random choice, but results from a deep understanding of the physics of (remote) sensing. For instance, active profiling with lidar is much more accurate for cloud detection than passive imaging in the infrared range, thus lidars (like CALIOP) have been widely used for validating imagers. A true ‘ground truth’ that reports clouds at every scale, at any time, and at every wavelength is the Holy Grail for cloud remote sensing. And we wish we could have it!

- Line 92, what do you mean by ‘the latter authors’, as there is only one group of authors mentioned in this newly-beginning paragraph.

Clarified, as suggested.

- Line 95, the full names of As and Ns should be presented.

Clarified, as suggested.

- Line 173, by mentioning T_b used by Zou et al., (2021), do you mean they are using observations from $8.1 \mu\text{m}$ as a symbol for T_{tropo} ? It seems like the observation or derivation of T_{tropo} is never mentioned in this paragraph.

Corrected. T_b stands for T_{bIR} (the ‘IR’ in the subscript was missing).

- Line 180, which scheme is the ‘latter’ one?

Clarified, as suggested.

- Line 293, by ‘every full hour’ do you mean data collected hourly?

Clarified, as suggested (hourly and every full hour)

- Line 296~297, I find this sentence extremely difficult to understand (or guess) the meaning.

Rephrased, as suggested.

- Line 299, what numbers are selected as the ‘fixed thresholds’ for each method?

Clarified, as requested. Threshold values have been added.

- Line 366~369. For the discussion, since the reference dataset can also have misclassified DCCs, would the relative differences between different methods be affected as well, such as one method happens to misclassify some DCCs in the same way as the reference when the other doesn’t but give the correct results instead?

There is no reason or evidence to suspect that any of the evaluated methods miss some clouds more or less frequently than others, or the reference. These methods are ‘genetically’ similar,

all three rely on passive sensing in the thermal infrared range. Perhaps the use of another type of detection method (e.g. machine learning or object-based), would introduce the suggested inhomogeneity. But at this stage, and with the available data, we are not in a position to draw any conclusions on the matter.