Review comments for “characterizing clouds with the CCClim dataset, a machine learning cloud class climatology” by Kaps et al.

This manuscript describes a cloud class climatology dataset that merges passive and active observations together in a fixed coarsened grid that is suitable for model comparisons. This dataset employs the CloudSat+CALIPSO 8 WMO cloud class as the “truth”, and Aqua-MODIS Level 2 physical cloud property retrievals as the inputs for the training in consideration of obtaining large amount of collocated training samples and to make the model interpretable (hence not using MODIS Level 1 measurements). Then the labels predicted from native resolution MODIS data are then coarsened to GCM grid size and training coarse-grained MODIS data to generate another ML model that is finally applied to the ESACCI-AVHRR dataset to generate a 34-year long cloud class dataset. The authors then gave several examples illustrating how to use this dataset for GCM comparison to help identifying issues associated with certain cloud types.

Overall, I think there’s some scientific merits of this current research. I particularly like the example regarding comparison with ICON-A outputs, and appreciate the “MODIS-equivalent filter” is applied to make the effort for a true apple-to-apple comparison. However, there are some fundamental issues with the design of the ML architecture that cause inevitable flaws to make this dataset really useful. These issues need to be addressed before publication of this work.

Major issues:

(1) The training “truth” CC-L coming from 2B-CLDCLASS only has CloudSat and some MODIS information. If you read the ATBD on CloudSat website, you’ll find CALIPSO data is in their plan to use jointly, but was not implemented so far. The 2B-CLDCLASS-LIDAR product has cloud mask from joint CloudSat-CALIPSO observations, however, that dataset doesn’t have cloud type classification. Because only CloudSat and limited MODIS information was fed into the “truth”, it inherently underestimates cirrus clouds and then propagates this bias into your product.

(2) Aqua-MODIS always collocates with CloudSat at “nadir” view. That means the off-nadir correction must be made in order to not “overpredicting” cloud masks because of the slantwise integration length making the off-nadir view easier to detect cloud. This correction factor was never discussed in the current manuscript. The overprediction of the overall cloudiness in your CCClim dataset might likely attributes largely to this factor.

(3) For ancillary data from ERA5 reanalysis, I don’t understand why temperature and water vapor profiles are not included. Aren’t they the closest atmospheric variables to determine whether to form a cloud or not?

(4) It is not explained why you can apply two-step ML models trained by MODIS data directly to AVHRR data. Admittedly you use similar L2 cloud retrieval products for training and prediction, but AVHRR has so few bands (literally only one visible band), so the products are not quite comparable. Even the Cloud_CCI project had published a paper illustrating their discrepancies (https://doi.org/10.5194/essd-9-881-2017).

(5) By training on multi-year collocated MODIS-CloudSat data, I don't quite understand why the inter-annual variability is not learnt by the ML model, resulting in no-interannual variation (e.g., ENSO) in your timeseries shown in Fig. 4. Although it is clarified later on the manuscript that this dataset is not suitable for trend study, it is never claimed that it is not suitable for interannual variability study either. Ultimately, if a 34-yr long dataset is not intended for studying inter-annual variability, why product that? Why not just stop at Step-2 model and produce a MODIS cloud class that suits every application presented in this manuscript.

(6) As also notified in this manuscript, MODIS (all passive sensors) have issues distinguishing clouds against snow-cover surface in polar regions. However, the statistics (e.g., Fig. 2, Fig. 5) were summarized globally. I'd strongly suggest you exclude polar areas in computing your statistics.

Minor issues:

Fig. 3 – recommend adding a map from CCL for straightforward visual comparison. For example, I don't see the Gill model distribution in the Western Pacific cirrus clouds (might be obscured by the annual cycle or your coarse colorbar). Same recommendation for Fig. 5 (i.e., if using CCL annual cycle, do you still see the same bias suggested by your CCClim product?)

Fig. 10 – It's not understandable for cirrus clouds, how can ICON model produces a bunch of thin cirrus (i.e., IWP low) with large size ice particles? Please double check your graphing codes.

Line 277 – Eliasson et al., citation year missing.