R4: *A new neural network cloud mask for VIIRS measurement is presented. The neural network is trained with collocated CALIOP observations. Using a global testing dataset of one year, the performance of the neural network is evaluated, using several metrics, for different categories like land/water, day/night, latitude range and varying COT threshold. Results show general good agreement with mean cloud fraction from CALIOP, including consistency between different categories, though larger differences are found for small-scale, low-level clouds. Comparison to two operational VIIRS cloud mask show that the neural network outperforms them for almost all conditions, however also here the struggle with small-scale, low-level clouds is evident. The largest improvements are found for collocations at higher latitudes.*

In general the manuscript is well structured. The method is clearly presented, including many corresponding references , and considerations made during the set-up of the neural network well explained. The manuscript could benefit from some additional information on the data used as well as from more details on the two operational cloud masks. The assessment of the performance and comparisons are done in multiple ways and accompanying figures clearly presented and explained. Issues and differences are analyzed and extensive discussion provided.

Minor comments/questions: Line 85: Before going straight to the Collocation Methodology I would recommend to add a small subsection on the VIIRS instrument/observations as well as for the two operational cloud mask swith which a lot of comparisons are done.

AR: Agreed. We will plan to add more details about the VIIRS instrument, and explain the general approaches of the two operational cloud masks.

R4: *Line 86: Also some more information on the CALIOP data could be provided, like what is the width of one cloud layer, time of overpass etc.?*

AR: Agreed. We will plan to add more information about the CALIOP data.

R4: *Line 105: Would be nice to see a global map of sampling frequency of valid collocations for the training dataset, maybe even per season, like is presented for the testing dataset (Fig. 6 b).*

AR: Yes, the distribution of collocations based on time of year and lat/lon is very similar between the testing dataset, but it certainly wouldn't hurt to show it explicitly in a figure.

R4: *Line 108: Observations in form of radiances/brightness temperatures? Please provide more detail on the input for the neural network.*

AR: Agreed. Other reviewers also raised this concern as well so it is clear we need to do a better job of explaining the inputs to the neural network. We will make sure this is improved in the final revised version.

R4: *Line 109: How are the eight categories combined?*

AR: The categories are 0=shallow ocean, 1= land, 2=coastline, 3=shallow inland water, 4= ephemeral water, 5 = deep inland water, 6 = continental/moderate ocean, 7=deep ocean.

We combine these categories to more generally describe land/water classification. In our binary land/water mask (land=1, water=0) the land category is made up of "land" and "coastline" categories of the original mask. All other water/ocean categories of the original mask are combined into a simply a water category. We will give more details about this in the final version

---

R4: *Line 175: It is not clear to me how the sun glint scenes are labeled, on a pixel-basis? There is a reference, but some more information would be nice.*

AR: Yes. The pseudolabeling model, which is invariant to sun glint, is used to make predictions where sun glint is present. The only differences between our main cloud detection model and the pseudolabeling model are that we include the infrared VIIRS/CrIS fusion channels in the pseudolabeling model, and remove all channels with solar contributions. It similarly accepts a a 3x3 pixel patch from each channel as input, and makes a cloudy/cloud-free prediction for the center pixel. After running this model for every pixel in several scenes with sun glint, we subsample these predictions (since we had more than enough data to work with).

Other reviewers have also suggested revising this section, so it is clear that we need to make some changes to the text in order to more clearly present this information. We will plan to revise this section in the final version of this paper.

R4: *Line 202: All inputs are standardized.. meaning for the 3 x 3 pixels?*

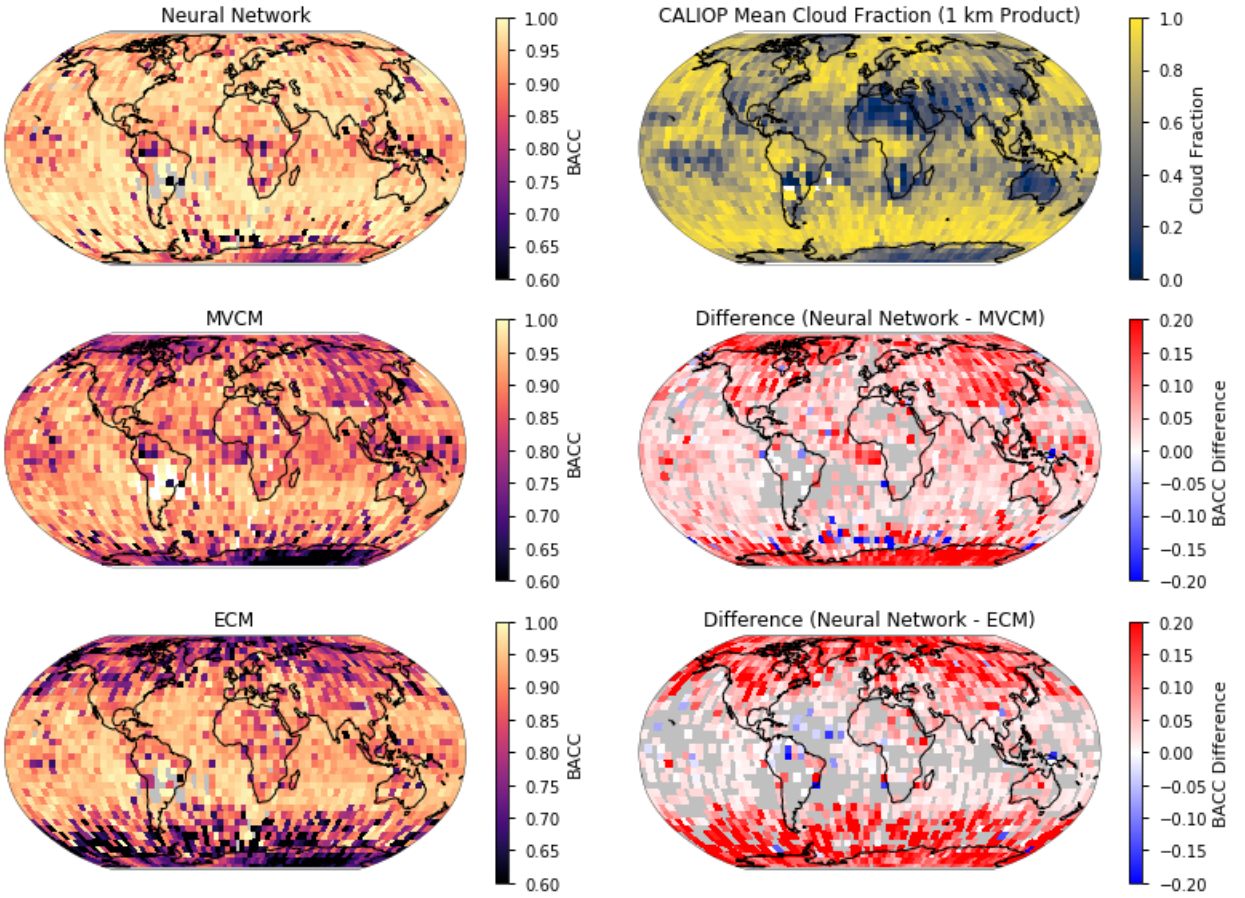
AR: The datasets are standardized using means and standard deviations computed for the entire training dataset, not just for each 3x3 patch. So for M15 (the ~10.8 μm brightness temperature), the one value for the mean is computed from all M15 observations in our training dataset (and the same for the standard deviation). The reason we do this is merely a technical consideration and is common practice for neural networks. We want the inputs to have approximately similar ranges. Inputs with very disparate scales can sometimes cause issues during training for neural networks.

R4: *Line 217: Already refer to corresponding equation numbers.*

AR: Yes. This is the first reference of those abbreviations/metrics so we will fix this accordingly.

R4: *Line 314: Why not continue with BACC?*

AR: Reviewer 2 asked a similar question so we are quoting our response here "We initially did not include a figure with BACC since there were grid cells where the mean CALIOP cloud fraction was particularly high, and the BACC was mostly dependent on a relatively small amount of cloud-free CALIOP collocations (the Southern Ocean, for example.) Below is the same figure with BACC, and panel (b) is replaced with the CALIOP mean cloud fraction. Similar to the ACC figure in the manuscript, these maps are calculated using the filtered dataset." We will plan to add this in the either the supplement or the main text.



R4: *Line 351: How are the surface temperature from the model matched, spatially and temporally, with the measurements? Some more detail should be provided*.

AR: First, we obtain the surface temperature from the 6-hourly 0.5 degree GFS 12-hour forecast (we incorrectly said this was the analysis in the main text – we will fix this) for the file occurring before and the file occurring after the VIIRS scene. Then, linear interpolation is used in time and space to approximate the resolution and time of the VIIRS scene.

One consideration that we did not discuss was the disparity in spatial resolution and differences in time between the VIIRS observations and the GFS 12-hour forecasts. Some of the differences in Figure 8 are perhaps due to discrepancies between the actual surface temperature and what was estimated by GFS due to these issues. Maybe the relatively poor performance the operational models in Figure 8 of the models is not only due to low thermal contrast, but mischaracterization of the surface temperature? We will think about these differences a bit more and add them to the main text in the final version.

R4: *Line 354: are smaller than*

AR: I think we should reword this statement altogether. Originally we were meaning the signed value was greater (in the sense that 4 > -6) but we now realize this is confusing. Maybe we should instead say something along the lines of "The performance of all models decreases as the

VIIRS 10.8μm brightness temperatures become more similar to or larger than the surface temperature.

R4: *Line 360/Fig 6.: The large negative difference for the grid cell in front of the coast of Namibia, could that be related to biomass burning aerosol layers?*

AR: This was what we originally suspected but didn't look into it much further.

We tracked down the poor performance in this grid cell to an individual nighttime scene over the where the neural network achieved an accuracy only a 25% over a stretch of roughly 550 CALIOP collocations over the ocean near the coastline that were 97% cloud-free. To compare, the ECM had 81.5% accuracy and the MVCM has 97.8% accuracy over this same stretch.

The very poor performance of this specific scene appears to be from a number of factors. One was a processing error where the land/water mask was not reduced from the 8 categories to the binary mask (we suspect a job on our cluster was preempted or terminated early).

We checked the closest daytime overpass and adjacent regions had aerosol optical depth of roughly 0.15 to 0.3 estimated by a VIIRS data. Moderate aerosol loading was somewhat apparent in the true color images.

We think the most influential factor was what appeared to be particularly cold SSTs along the coastline indicative of upwelling. Without the visible channels in this nighttime scene, the neural network misconstrued these cold surface temperatures as cloud cover. We suspect this to be a scenario where a rough estimate of surface temperature could improve the accuracy of our approach. After fixing the error in the land/water mask the accuracy over this small stretch of CALIOP collocations improves from 25% to 63% -- still trailing behind the ECM and MVCM significantly.

R4: *Line 472: Could some (pseudo) labeling technique be useful here? Or using a larger pixel matrix than 3 x 3? Maybe combined with taking information from not only 1 CALIOP profile but from adjacent profiles as well?*

AR: Sure! As another reviewer suggested we might be able to pseudolabel using the MVCM or ECM to address specific deficiencies in the neural network. For the purposes of this paper we wanted to demonstrate how the neural network could be implemented in a stand-alone capacity (assuming that neither of the operational masks were available).

We briefly experimented with a convolutional neural network that used 5x5 and 7x7 patches. These models actually performed slightly worse for broken clouds and only improved performance in more homogenous scenes (represented by the filtered dataset).

Using adjacent CALIOP profiles could be interesting, but we would have to think carefully about how this might work. We noticed in previous experiments when we trained to the 5 km CALIOP product that resulting cloud mask was extremely "smooth" and did not at all capture the fine-

scale variability of scenes with broken clouds. We worry that using information from adjacent profiles might lead to a similar effect.

This might be a situation where manual labeling could be a good option. One way to do this efficiently could be to find regions of broken clouds in VIIRS images, draw a bounding box around such regions, and choose a threshold on the 11µm channel that reliably separates cloud-free from cloudy pixels (using a threshold specific to each manually selected region). This is, of course, very subjective.

R4: *Technical corrections Line 29: ..large amounts of training data.. Line 47: ..how a very simple.. Line 298: .. compared to the MVCM.. Line 323: All of the previous.. Line 325: ..depend on the particular.. Line 363: the distribution of the.. Line 423: .. may be a result of sea ice cover. Line 428:.. is subject to a large amount..*

AR: Thanks for pointing out these errors! We will make sure they are fixed in the final version.