Response to Reviewer #3 Comments

R3: Major Comment: What is exactly, is given in the input of the main neural network? Is it radiances, or reflectances and brightness temperature? For each radiances, is it the differences with climatology (or simulations) and the observation, or raw radiances?

My concern is that the neural network description lack about the physic that is behind such the nature of the input. Also that important information are dispatch in all the study to explain fairly some results, but they are still necessary to be mentioned in the neural network description.

AR: In total, there are 20 different inputs to the cloud masking neural network. The first 16 inputs are from the moderate resolution channels of VIIRS (reflectances for the visible channels and brightness temperatures for the infrared channels). The last four inputs are a binary land/water mask, absolute value of latitude, solar zenith angle, and sun glint zenith angle. We state this in section 2.2, but we think a table would perhaps be more useful here and would facilitate the reader referencing it later on in the paper.

Only the raw observations are used, but we believe some confusion on this point might be coming from statements we make on lines 202-203 where we state that the inputs are standardized by subtracting the mean and dividing by the standard deviation. Scaling inputs is common practice for neural networks, and other approaches are often used such as making the minimum and maximum values 0 and 1, or -1 and 1. If this standardization is not done before training a neural network, then gradients of larger parameters may tend to have larger influence during the training process. In other words, we are rescaling the inputs to the neural network so that each variable has zero mean and unit variance. Otherwise, updates to the neural network may tend to favor the use of inputs with larger values or variance. This is particularly a concern for us since our inputs have very disparate scales: typically between 0 and 1 for reflectances, and between 180 K and 340 K for brightness temperatures.

From one perspective, this might be seen as a framing the inputs as differences from climatology. However, we make no effort to ensure even representation from different seasons or times of day. Similarly, we don’t perform this separately for different locations. As a result, we don’t believe calling these inputs differences from climatology would be appropriate. Rather, we would simply call them observations rescaled with zero mean and unit variance.

R3: This is more a thought for the conclusion: How does neural network methods will react in the context of global warming and the fast modification of some surfaces? Does it mean that despite the benefits of the accuracy that provide neural network, they are countered by the fact they will need regular updates?

This is an interesting, but difficult to answer question! We can think of a few ways in which climate change could affect the performance of this approach. One obvious scenario, which we somewhat touch on in this paper, is the declining presence of arctic sea ice. Evidenced by Figure 10, the cloud masks in this paper can have large TPR differences over different geographical regions, time of day, and surface type. Based on Figure 2, Figure 10, and previous evidence (Liu et al., 2010), we suspect the MVCM underestimates cloud cover over sea ice. As a result, we expect that the decrease in Arctic sea ice, and corresponding increase in ice-free ocean would result in a change in cloud cover estimated by the MVCM even if it did not result in a real change in cloud cover. This is obviously undesirable, and lends us to believe that cloud detection ability should be invariant as possible to changes in surface type. We believe that Figure 10 illustrates that the neural network may be the least susceptible to this, although the ECM and MVCM are certainly competitive with it in certain regions/conditions.

We suspect that another way in which cloud masks could be impacted by climate change would be a change in the global or regional cloud fraction. The neural network cloud mask is ultimately a statistical model. Under uncertain conditions, it will tend to predict the majority class (usually cloudy) since this is what gives it the best results during training (as measured by binary cross-entropy in our case). We have tried to ensure our approach does not depend too heavily on the use of the background mean cloud fraction by using class-balanced metrics like BACC in our evaluation. Along this same line of thinking we have limited to the amount of geographical information to the absolute value of latitude and a land/water mask. Not including longitude, or more specific surface types was an intentional choice in an effort to reduce on the model’s reliance on how the mean cloud fraction varies with this information.

Regular updates would certainly be beneficial, but are not necessarily a specific requirement of our approach. If climate change were to change something fundamental about the decision boundary between clear-sky and cloudy scenes, then all statistical cloud detection models would be impacted. We hypothesize that drifts in sensor calibration or changes in noise levels would be the largest factor in whether a cloud detection model would need regular updates if used in climate data records.

However, it might be interesting to investigate whether machine learning models trained during periods that are dominated by a particular phase of a climate oscillation generalize well to years of the opposite phase. For example, we could imagine issues stemming from approaches trained only during a strong positive ENSO year that utilize longitude and SST as predictors. Assessing how much of an impact this would have would require a very careful experimental setup, and is perhaps specific to what a model uses as predictors and what exactly it is tasked with predicting.

This is certainly something we will be thinking about if we go forward with this approach for future cloud amount analyses. We will plan to add some of this information in our discussion.

R3: Minor modifications

It would help to provide a table of the VIIRS band.

Agreed. We will plan to add a table like this to our paper.

R3: Page 4, section 2.2: You mentioned in the discussion section (page 16, line 505) that you use ancillary data. But it is poorly described in this section 2.2 (linked to my major comment).

Yes. We could certainly do a better job of describing the inputs to the neural network. We think some confusion might be come from our use of the word “ancillary” which we intended to mean an input that is not solely dependent on the VIIRS observations. Using this definition, the land-water mask would be our only ancillary data in the final neural network model since it is a derived product. We will try to clarify this in the text, and add context to our usage of “ancillary,” or remove it entirely since its meaning could be ambiguous in the manuscript’s current form.

Section 2.2. is another section that we think could benefit from a table. We will plan to include one here that lists which inputs are used for the three neural networks: (1) The main neural network which we are proposing and evaluating the performance of, (2) the pseudolabeling model which only uses infrared channels without solar contributions and provides labels to the main neural network in sun glint scenes, and (3), a neural network that is not trained with pseudolabels and is only used in Figure 12.c to demonstrate the usefulness of pseudolabeling in this context.

R3: Page 6, lines 171-181: The second part of the section “3.1 Pseudo-Labelling Procedure” is hard to understand at some points. In this section, it is about the neural network that help to account for sun glint. What information is provided by this neural network to detect sun glint? Is this information provided to the main neural network to not perform a cloud mask, or does it simulate input that are supposed to appear in sun glint condition for the main neural network? Where comes from the information of true sun glint conditions, to be reproduced? Why the 15th day of every month in 2018?

AR: We agree that this text needs to be revised and rewritten, particularly lines 171-181. We think it could benefit from the table in the above response. We could also add a separate table of VIIRS/CrIS fusion channels rather than listing them in the text.

The reason we need the pseudolabeling model, is that we do not have any labels in regions with sun glint where the visible VIIRS channels may give the false impression of cloudiness. To fix this we train a pseudolabeling model that only uses VIIRS IR channels, VIIRS/CrIS fusion channels without solar contributions, latitude, and the land/water mask. The VIIRS/CrIS Fusion channels are used in the pseudolabeling model in an attempt to make up for some of the information lost by removing the visible channels. We removed solar zenith angle and the glint zenith angle since they would not provide useful information for the pseudolabeling model. Aside from these differences, the pseudolabeling model is trained using the same collocations as the main neural network model.

The pseudolabeling model is then used to make predictions in scenes with substantial sun glint. Our determination of substantial sun glint is somewhat subjective, and consists of images with sun glint zenith angles less than 40 degrees. Below 40 degrees, is roughly were we identified visible reflectances starting to increase due to specular reflection and thus, where we would need pseudolabels. The predicted probabilities from the pseudolabeling model are then treated as true labels (as if they were obtained from CALIOP). Essentially, the pseudolabeling model creates cloudy/cloud-free labels that incentivize the main neural network model to not misconstrue high visible reflectance for cloud cover in sun glint scenes.

Scenes from 2018 are selected since that is a year that is included in our training dataset. If 2017 or 2019 were used, our training dataset for the final neural network model would not be independent of the validation and testing datasets. The choice of using the 15th day of each month from 2018 is also somewhat arbitrary. We needed to capture the annual variability of sun angle with respect to latitude, so selecting one day from each month was preferable to selecting 12 consecutive days, for example. We could use more or fewer days, but even after substantial subsampling of these scenes, we found we had more than enough pseudolabels to work with.

R3: Page 7, line 202: what is the meaning of "binary cross-entropy"?

AR: Training a neural network requires a loss function (usually called *cost* or *error* function in meteorology and other disciplines) that is differentiable and that one is typically aiming to minimize throughout training. In binary classification problems we usually select binary cross-entropy as our cost function. We will add the equation for the binary cross entropy function to clarify this.

$$J(y, \hat{y}) = -(\, y\log(\hat{y}) + (1-y)\log(1-\hat{y}))$$

Here, J is the binary cross-entropy loss, *y* is the binary label from CALIOP, and $\hat{y}$ *is* the predicted cloud probability from the model.

R3: Page 10, lines 279- 284: Seeing the Figure 4, the difference between TPR of MVCM and the one for neural network is really small. It is most likely that their performances for low broken clouds are similar.

AR: Agreed. Our phrasing here was partially motivated by the fact that the VIIRS/CALIOP collocations often do not characterize these clouds well since they can often be smaller than the resolution of the 1 km CALIOP Cloud products. Even though the differences in TPR according to the unfiltered VIIRS/CALIOP collocations in Fig 4 are very slight, they may be indicative of a larger overall difference making this result more significant. We do mention this already in the discussion, so we will amend the text here to say the performance here for the collocations is similar, but clarify this expectation later in the discussion.

R3: Page 10, line 314-322: In Figure 6, the cloud mask with neural network is less sensitive to variation of latitudes.

AR: Indeed. This is a good point that we will plan to add here.

R3: Page 11, line 323: "All the of the previous" A word is missing!

AR: This should read "All of the previous analyses [...]." We will fix this in the final version.

R3: Page 12, lines352-354: "This is surprising … a land or water surface." This is really important information it should be mentioned in the description of the neural network input section. (major comment)

AR: We do mention this in section 2.2, but again, it could be more clear and would certainly benefit from a table describing inputs.

R3: Page 14, line 428: "is subject to a large"

AR: Thanks! Yes, this is how that line should read (the word “to” was missing in the original manuscript).

R3: Page 14, line 447-page 15, line 459: I suggest you put this section and Figure 12, with the section “3.1Pseudo-Labelling Procedure”, as it makes the understanding of the pseudo-labelling more clear. Also, because this section is quiet independent of all the analysis of the neural network performances.

AR: Agreed. In the current manuscript, there are 8 pages of text separating the description of pseudolabeling and seeing the actual impact of pseudolabeling. This is obviously not ideal. We will make this change in the final version and move Figure 12 and its accompanying text to the end of section 3.1.

R3: Page 23, Figure 1: This paper would benefits of a better scheme that describe the neural network. Better description of the input vector with geo-localisation information.

AR: Absolutely. We will plan to give a more detailed description of the neural network and its input. Figure 1 might also be more effective as a simple table describing the input size, output size, and type of each layer.

R3: In relation with section 3.2 page 6 and 7, can you say more about the meaning of dropout X% between each layer?

AR: Yes. Dropout is a very simple regularization method used in neural networks that helps prevent overfitting. In short, dropout sets a specified number of intermediate activations to a value of zero. Our first layer is a fully-connected layer with 200 units. Dropout(2.5%) indicates that 5 of these 200 units are randomly selected at each training step and set to a value of zero. This helps prevent the model from relying too heavily on any one connection. In our case, we suspect that if we set the dropout rate too high (5% or above) the model had trouble learning relevant features since it was performing worse on our validation dataset during hyperparameter tuning. When we completely removed dropout (by setting it to 0%), the model also performed worse. A dropout rate of 2.5% turned out to be ideal in preventing some amount of overfitting, while still allowing the model to learn relevant features. We will add a sentence or two in the text summarizing this.

R3: Page 34, Figure 12: There are obvious difference in the behave of the cloud mask from neural network without pseudo-label and the one with pseudo-label. The second cloud mask is more “binary” (i.e. values equal to 0 or 1) than the first one. Can you comment this result? Which neural network of figure 12, have you compared during your paper?

AR: Yes. The model that we are analyzing the performance of in the results section is Figure 12.d (the neural network with pseudo-labels). Figure 12.c is shown simply to illustrate how poorly a machine learning model can perform if we don’t account for obvious deficiencies in the training dataset (such as a dataset without any sun glint examples).

On the question of the differences between the two models – that is tough to answer for a couple of reasons. First, neural networks are notoriously difficult to interpret the predictions of, and secondly, sun glint scenes are out-of-domain predictions for the neural network without pseudolabels. Said differently, it is undefined behavior for this model so attempting to interpret its predictions here is even more difficult than usual.

Of course, as we mention in the paper, the neural network without pseudolabels makes erroneously cloudy predictions throughout the entire scene. This is likely because the model without pseudolabels has likely learned to associate high visible reflectivity over water with cloudy pixels since it has never seen sun glint.

All that being said, we can speculate on some of the reasons why Fig 12.d might be more "decisive" than the model in Fig. 12.c. For Fig 12.c there are likely two competing factors: a high visible reflectivity which usually indicates cloudy pixels, and warm infrared brightness temperatures which usually indicates clear pixels. These two factors could result in uncertain conditions since these pieces of information are somewhat contradictory. Fig 12.c is likely decisive because it is making predictions in a sun glint scene where it is exclusively trained with pseudolabels. Rather, than the probabilities being accurate assessments of uncertainty with respect to the CALIOP label, the probabilities in Fig 12.d likely portray the model's ability to accurately reproduce predictions made from a model that exploits solely IR information. To that end, the cloud probability/uncertainty estimates in sun glint regions are not especially useful (which we mention in lines 531-535), but we can verify the actual predicted labels from Fig 12.d appear reasonable compared to the other operation models.

Below we have created drafts of three tables that we will plan to include in the revised version of this manuscript. These tables are aimed at addressing many of reviewers #3 concerns on the nature of the inputs to the neural network. In addition to Table 3, we will include clarifying text that details the purpose of each model, and specifically where each model is used in this work.

Band	**Spectral Range (µm)**	**Units**
M1	0.400 – 0.421	Refl.
M2	0.436 – 0.451	Refl.
M3	0.477 – 0.496	Refl.
M4	0.541 – 0.561	Refl.
M5	0.662 – 0.680	Refl.
M6	0.738 – 0.752	Refl.
M7	0.843 – 0.881	Refl.
M8	1.225 – 1.252	Refl.
M9	1.368 – 1.383	Refl.
M10	1.571 – 1.631	Refl.
M11	2.234 – 2.280	Refl.
M12	3.598 – 3.791	BT [K]
M13	3.987 – 4.145	BT [K]
M14	8.407 – 8.748	BT [K]
M15	10.234 – 11.248	BT [K]
M16	11.405 – 12.322	BT [K]

Table 1: Shown are the names, and spectral ranges of each moderate resolution VIIRS channel. Also shown are the units and whether the channels are expressed in reflectance (Refl.) or Brightness Temperature (BT).

VIIRS/CrIS Fusion Channel	Spectral Range of MODIS Equivalent Channel (μm)
MODIS 27	6.535 – 6.895
MODIS 28	7.175 – 7.475
MODIS 29	8.400 – 8.700
MODIS 30	9.580 – 9.880
MODIS 31	10.780 – 11.280
MODIS 32	11.770 – 12.270
MODIS 33	13.185 – 13.485
MODIS 34	13.485 – 13.785
MODIS 35	13.785 – 14.085
MODIS 36	14.085 – 14.385

Table 2: Shown are the names of each infrared VIIRS/CrIS Fusion channel that is used in the pseudo-labeling model. The stated spectral ranges are those of the MODIS equivalent channels. All channels are expressed as brightness temperatures [K].

Inputs	Neural Network with Pseudo-labels	Neural Network without Pseudo-labels	Pseudo-labeling Model
M1-M11	X	X	
M12-M13	X	X	
M14-M16	X	X	X
MODIS27-MODIS36			X
| Latitude |	X	X	X
Solar Zenith Angle	X	X	
Sun Glint Zenith Angle	X		
Land/Water Mask	X	X	X

Table 3: Summary of the information used by each neural network model in this work.

Once again, we would like to express our thanks to the reviewer for volunteering their time to give us feedback on our manuscript. They have very helpfully identified several areas in which we can improve the quality, presentation, and clarity of this work.