Thank you very much for your very kind review and comments. We have addressed the issues raised and amended the manuscript.

Please note, upon reflecting on some of the comments made about this study, we have rerun the models with labels set without the altitude limit to the CAD score filter. This has led to some minor improvements in the models' ability to classify cloud and aerosol when compared to the JMA and BoM masks. We have updated figures 4-8 and 10-12, as well as table 2, with the results from the new models (please see updated manuscript figures and tables section at the end of this document). With the exception of the case study of the dust storm over China, where the NN mask no longer misclassifies the dust plume, our analysis remains the same.

- In line 11, "0.106 and 0.198" and "0.314 and 0.464" have been replaced with "0.160 and 0.259" and "0.363 and 0.506" respectively.
- In line 12, "1.11 and 1.28 times" has been replaced with "1.13 and 1.29 times".
- We have removed the reference to the Labonne et al., 2009, study in line 196.
- In line 266, "a KSS of 0.691 versus 0.589 for the JMA product and 0.472 for the BoM product" has been replaced by "a KSS of 0.632 versus 0.523 for the JMA product and 0.432 for the BoM product".
- In lines 278-279, "the associated FPRs would be 0.314 versus 0.464 for NN and BoM algorithms and 0.106 versus 0.198 for the NN and JMA algorithms respectively" has been replaced with "the associated FPRs would be 0.363 versus 0.506 for NN and BoM algorithms and 0.160 versus 0.259 for the NN and JMA algorithms respectively".
- In line 279, "This implies that the NN accurately identifies 1.11 and 1.28 times" has been replaced with "This implies that the NN accurately identifies 1.13 and 1.29 times".
- In lines 336-340, "Over land, bands 4, 10 and 14 have approximately equivalent significance in the NN. Bands 4 and 14 serve the same role over land as they do over ocean. However, unlike over ocean, some land surface types can be bright in band 4 at twilight. This causes the NN to require a water vapour absorption band to effectively identify cloud over land during twilight and the NN has found band 10 to be most useful for this purpose" has been replaced with, "Over land, bands 11 and 14 have approximately equivalent significance in the NN. Bands 4 and 14 serve the same role over land as they do over ocean. However, unlike over ocean, some land surface types can be bright in band 4 at twilight. This causes the NN to require an additional cloud-detection band to effectively identify cloud over land during twilight and the NN has found band 11 to be most useful for this purpose".
- In lines 383-394, "However, all the masks fail to effectively classify the dust plume, with the exception of the NN mask accurately classifying a small section of the dust storm to the north of the Korean peninsula. Given that this event was a historically significant event with an unusually high plume (Filonchyk, 2022), the failure of the cloud masks might be expected. In particular, it shows that the NN cloud mask is only as effective as its training data and extreme events that it is not trained for will cause the mask to fail, although under more extreme scenarios than the JMA and BoM masks. In panel b of Fig. 8, pleasingly it can be seen that the section of the dust plume that is towards the centre of the scene is assigned scores significantly below values given to clouds - the plume has values of approximately 0.5, whereas clouds have values close to 1 - indicating that the NN mask is not confident the plume is cloud. A future algorithm could use this information within a convolutional NNs to improve the performance for large plumes or to develop uncertainty metrics" has been replaced with, "The JMA and BoM masks fail to effectively classify the dust plume, which the NN mask accurately identifies as non-cloud. Given that this event was a historically significant event with an unusually high plume (Filonchyk, 2022), the failure of the cloud masks might be expected. However, large areas of the dust plume are assigned relatively high values by the NN mask. In panel b of Fig. 8, it can be seen that the section of the dust plume that is towards the centre of the scene is assigned scores slightly below those assigned to cloud - the plume has values of approximately 0.5, whereas clouds have values close to 1 - indicating that, although the NN mask is not confident the plume is cloud, the dust storm poses a challenge to the NN masks classification algorithm. A future algorithm could use this information within convolutional NNs to improve the performance further for large plumes or to develop uncertainty metrics".

- In line 456, “the NN accurately detects 1.11 and 1.28 times” has been replaced with “the NN accurately detects 1.13 and 1.29 times”.

Comment 1: l.152 The final inputs to the neural networks are listed. Line 91, however, states: "The auxiliary information from AHI is also included in the collocated data, such as the latitudes, longitudes, solar and observation angles." Are the latitudes and longitudes included as inputs for the neural networks?

Reply 1: Only the final inputs described in line 152 are used in the final models. All other auxiliary information was used for further analysis of the results. We have added a clarification of this point to line 184, which reads, “Only these inputs are used for the models. Auxiliary data, such as satellite zenith angle, latitude and longitude is used only for further analysis of results.”.

Comment 2: l.172 The neural network training section lacks some minor details for reproducing the results. For example, how many epochs were used to train the neural networks? Was there an early stopping criterion to stop the optimization? How do the authors ensure the convergence of the trained neural networks?

Reply 2: The NNs are trained over 200 epochs. This number ensures convergence as the cost can be seen to asymptote before this value in all the models that were trained. Line 213 has been amended to include “Each NN is trained over 200 epochs to ensure convergence”.

Comment 3: l.227 Please describe what is meant by a collocated dataset ("30 collocated datasets"). For example, do a single dataset correspond to some specific time instant, or are the pixels selected randomly, or something else?

Reply 3: We have added a short definition of a collocated dataset to line 256, which reads “where each dataset is a CALIOP overpass that has been collocated with AHI data”.

Comment 4: l.252 Please clarify that by the surface type you mean land or ocean.

Reply 4: We have added the clarification of surface type to line 282, which reads “surface type (over ocean or over land)”.

Comment 5: l.267 Typo "Shapely". It should be "Shapley".

Reply 5: The typo “Shapely” in line 296 has been corrected to “Shapley”.

Updated Manuscript Figures and Tables

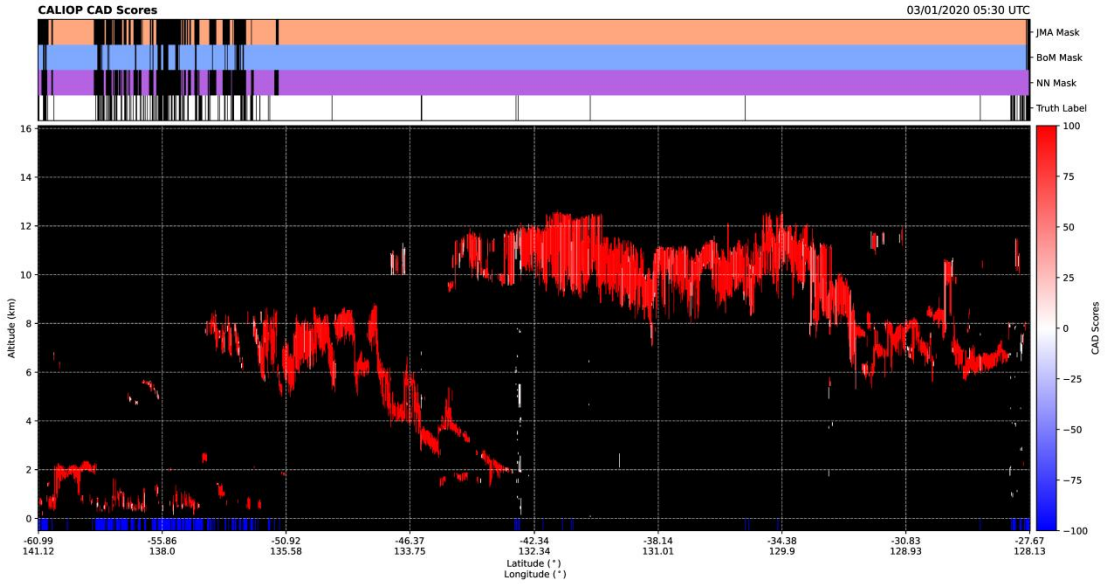

Figure 4 - Old

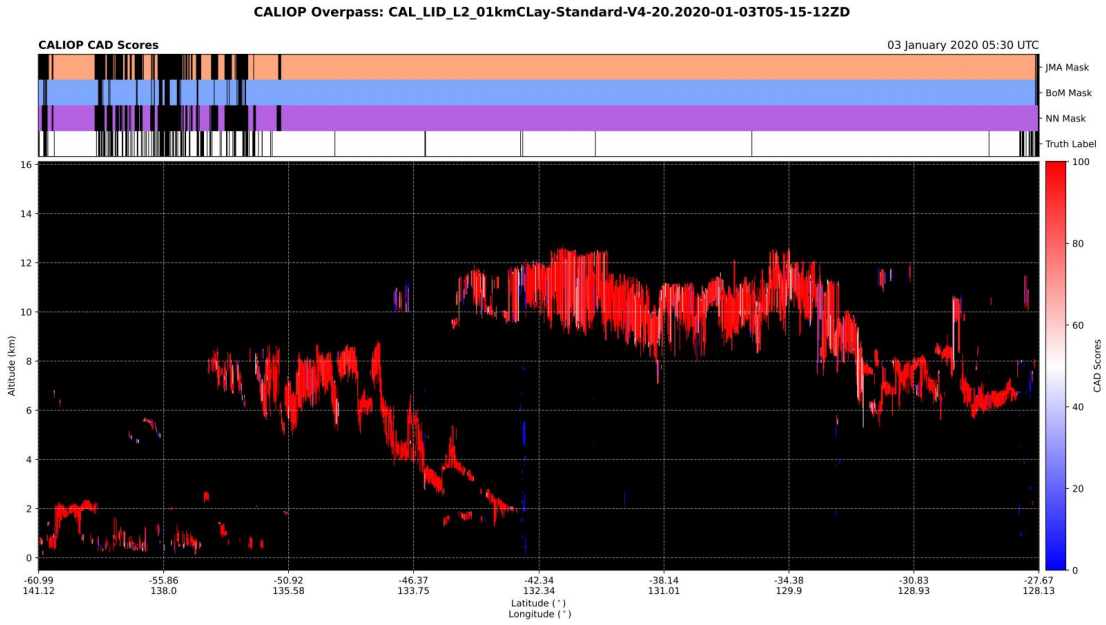

Figure 4 - Updated

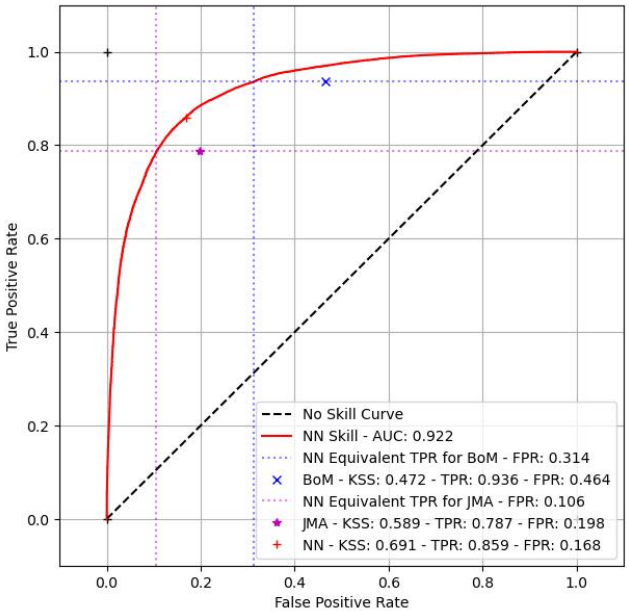

Figure 5 - Old

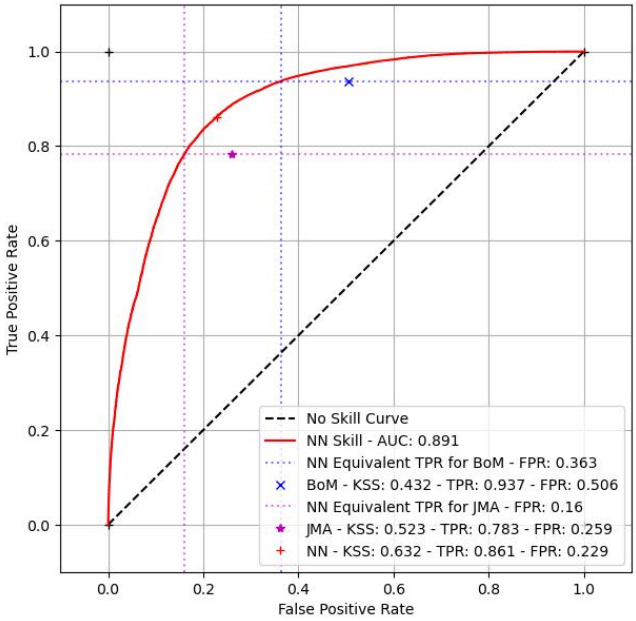

Figure 5 - Updated

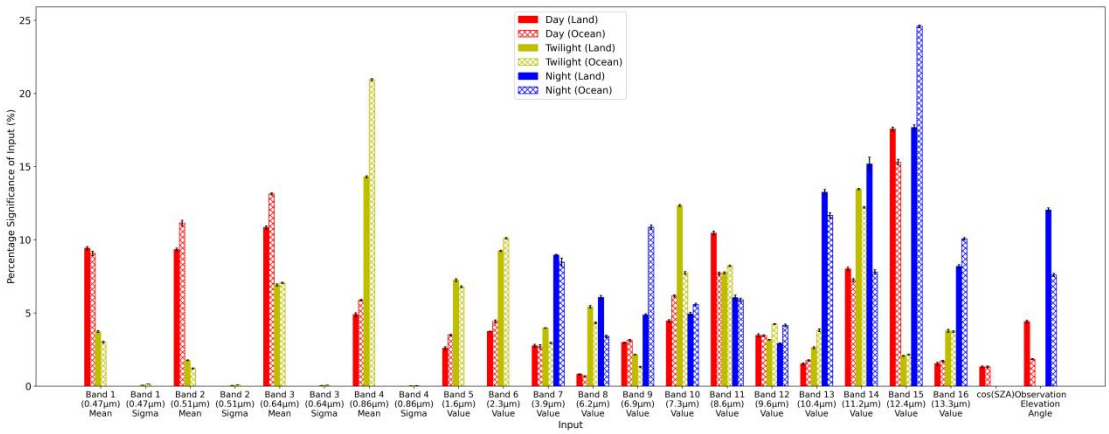

Figure 6 - Old

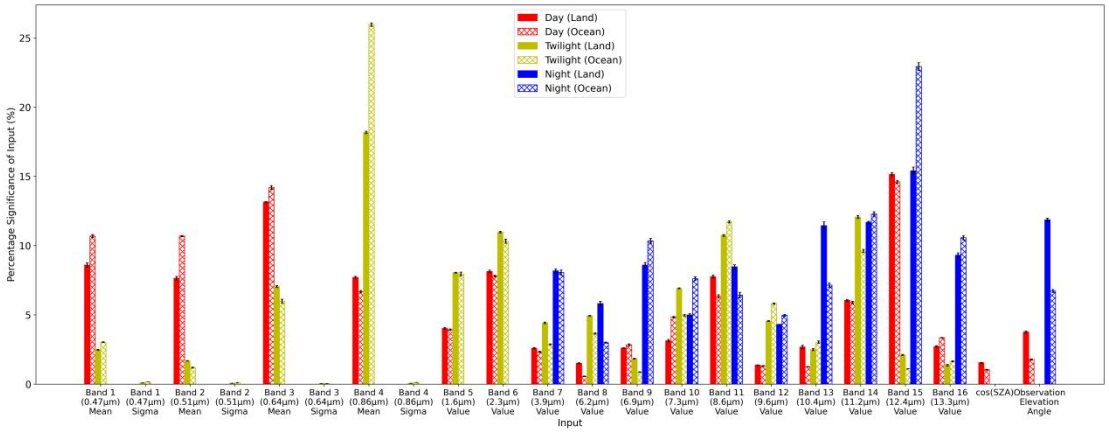

Figure 6 - Updated

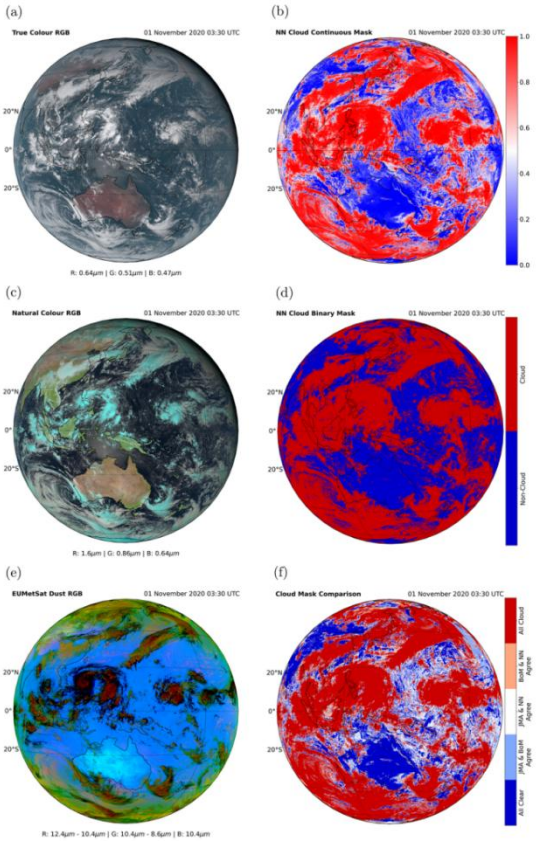

Figure 7 - Old

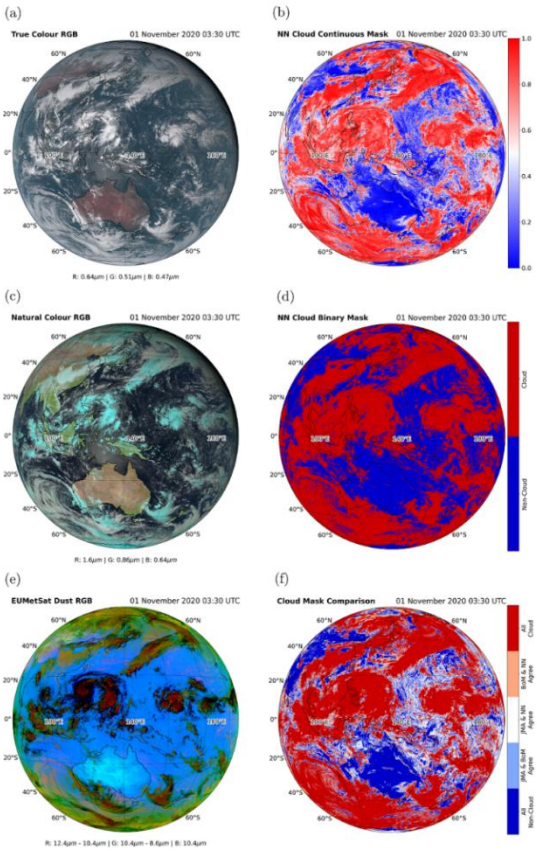

Figure 7 - Updated

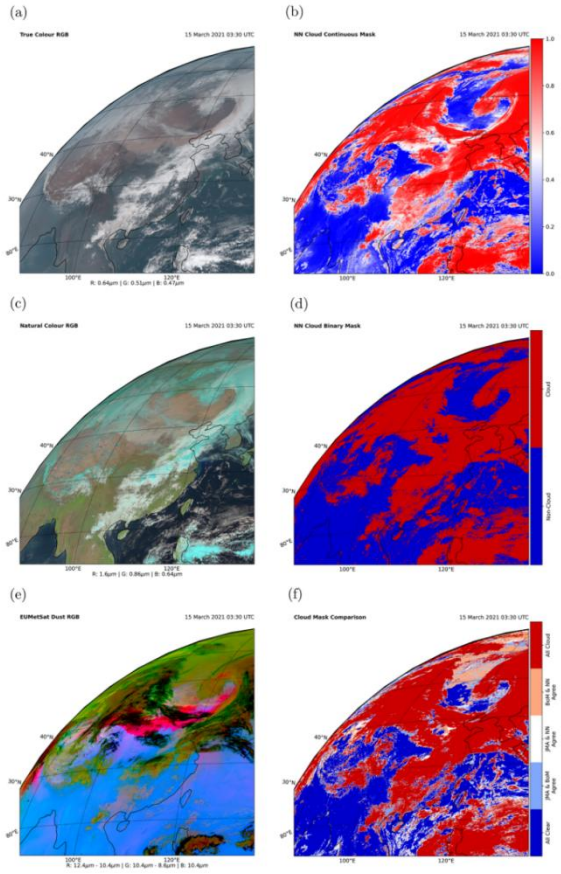

Figure 8 - Old

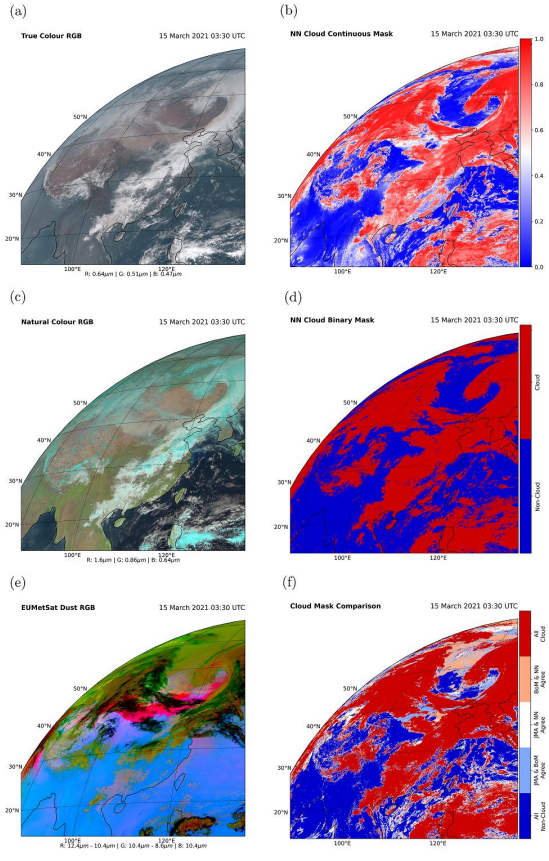

Figure 8 - Updated

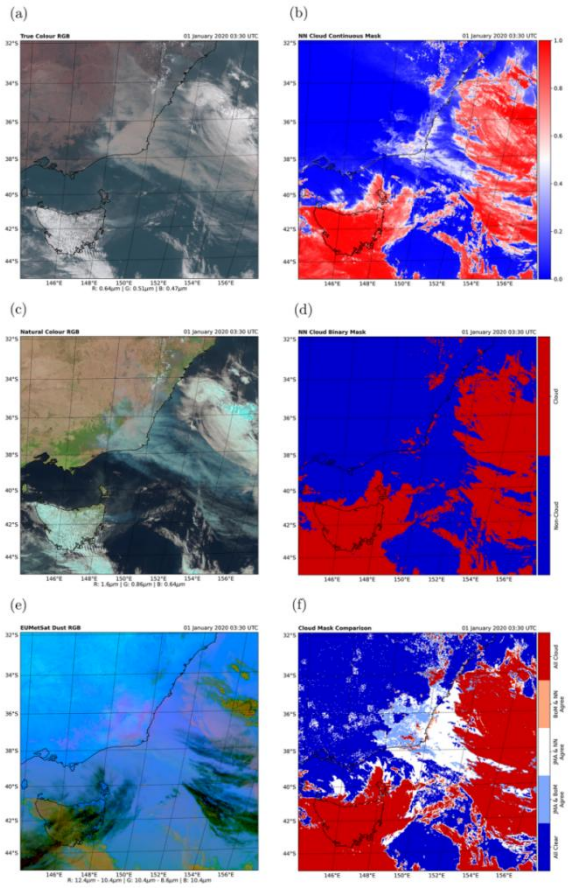

Figure 10 - Old

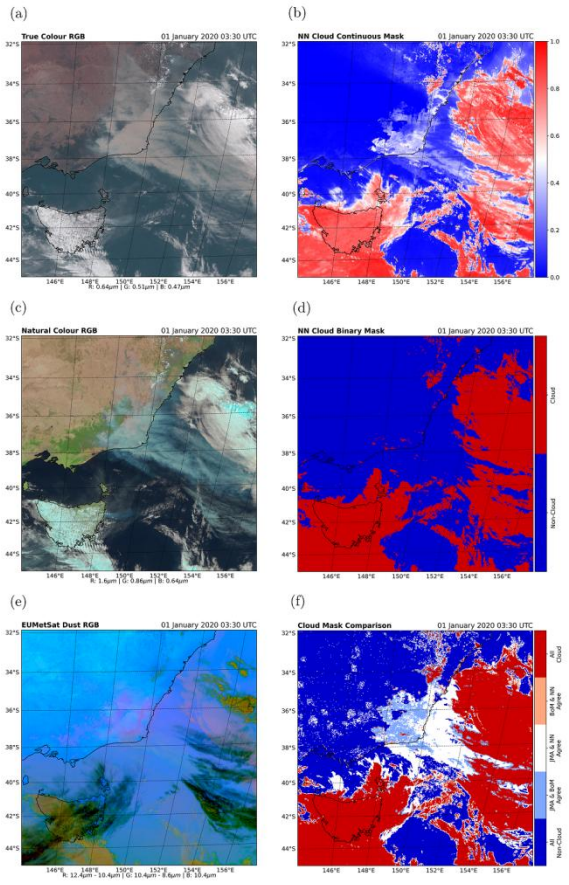

Figure 10 - Updated

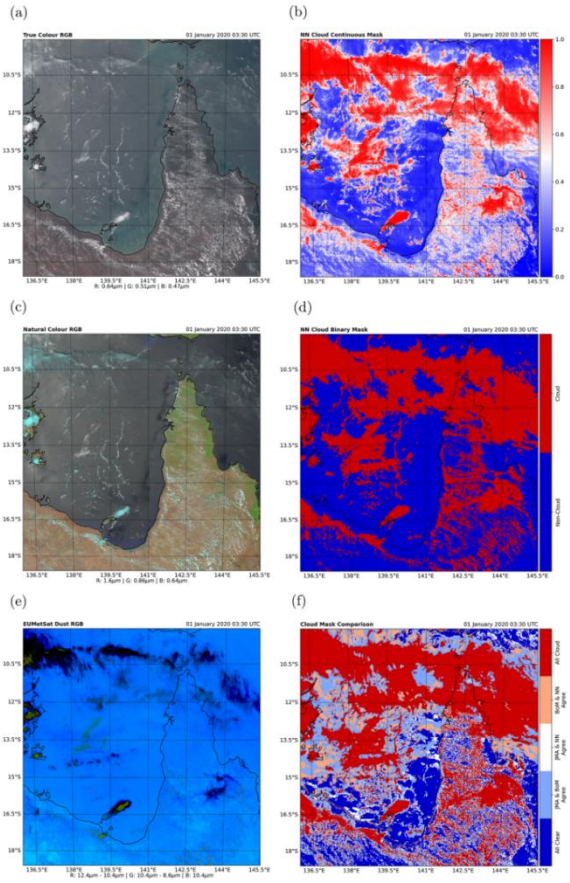

Figure 11 - Old

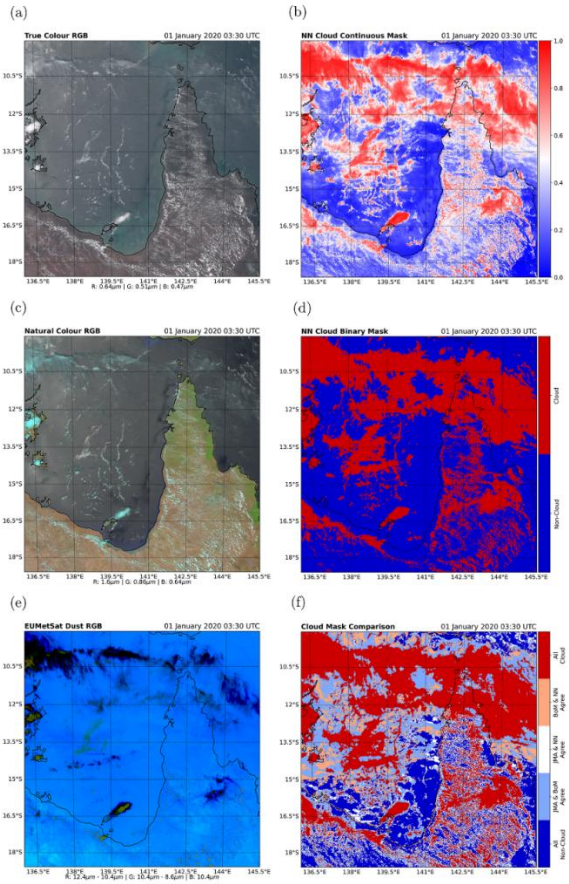

Figure 11 - Updated

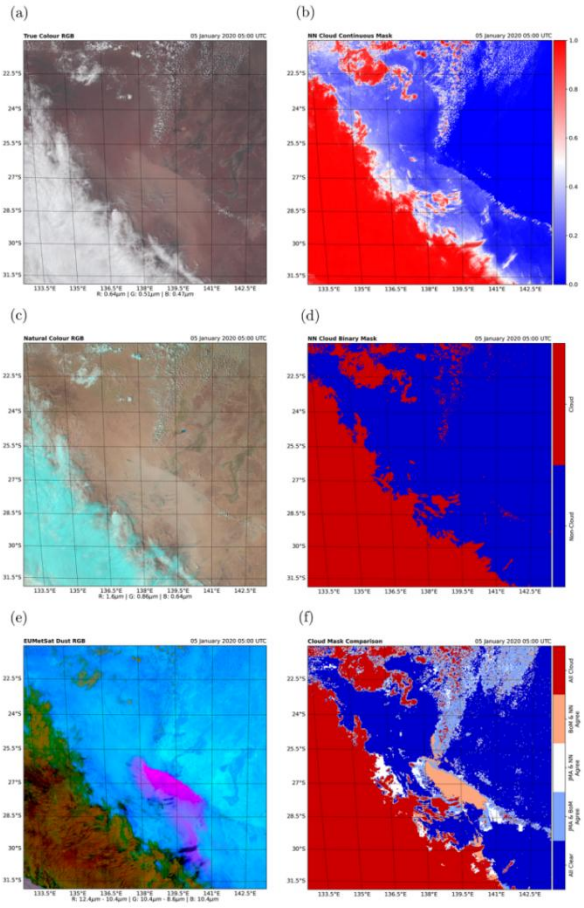

Figure 12 - Old

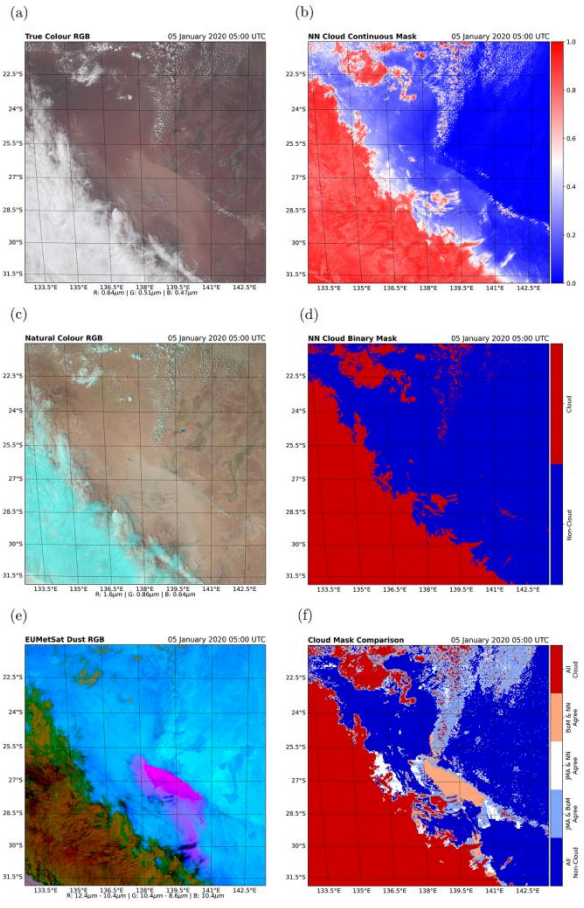

Figure 12 - Updated