Thank you for your review and comments. We have amended the manuscript to address the issues raised.

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- In line 266, "a KSS of 0.691 versus 0.589 for the JMA product and 0.472 for the BoM product" has been replaced by "a KSS of 0.632 versus 0.523 for the JMA product and 0.432 for the BoM product".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: One of the issues is the innovative contribution of this paper. As the author mentioned, using Machine Learning to facilitate satellite image recognition/categorization is a hot topic. Many studies have tried using ML/CNN to identify cloud and/or aerosols from passive sensors, for example Marais et al., 2020, Lee et al., 2021, Wang et al., 2020. Some of these studies also uses lidar as benchmark to label particle types. The new contribution from this study that is differ from the already published studies shall be clarified.

Reply 1: These studies have been added to the manuscript and a short description of the way in which this study differs from those cited has also been added to the manuscript.

- Marais et al., 2020, has been cited alongside Hughes et al., 2019, in line 72 critiquing hand-labeled datasets.
- Wang et al., 2020, and Lee et al., 2021, have been cited in line 80 and the way in which this work differs described in lines 81-86, which read, “These studies all demonstrate the success of machine learning algorithms trained on data labeled by active instruments, but do not extend the technique to geostationary instruments, which require a more sophisticated collocation technique to account for the parallax between the active and passive sensors. These studies do not focus on biomass burning plumes, which are of particular significance over Australia. In addition, in this paper we combine this technique with explainable machine learning models to better understand the influence of passive instrument channels on the outcome of the classification.”.

Comment 2: In addition, more information of the disadvantage and advantage of passive and active remote sensing techniques of clouds and aerosols are needed to justify the benefits of using active sensor to provide typing information.

Reply 2: Further discussion of the advantages and disadvantages of passive and active sensors have been added to the manuscript to reinforce our justification for using labels assigned by active sensors.

- A critique of passive instruments has been added to lines 38-40, which reads, “However, individual passive sensors can only see in 2D. For example, to classify whether a bright, cold pixel is snow/ice or a cloud top, a retrieval algorithm must be applied to the pixel to classify it. These algorithms require evaluation using ground-based instruments and active instruments to ensure that they are accurate”.
- A discussion about the advantages of using active sensors has been added to lines 42-45, which reads, “As they have their radiation source on-board, active instruments can operate independently of solar illumination and are more sensitive to thin atmospheric layers, such as thin cirrus and aerosols, than passive instruments. In addition, active sensors can retrieve the height of layers within a pixel by evaluating the strength of the return signal and time taken for the pulse to return. This makes them able to detect clouds and aerosols within their pixels much more accurately than passive instruments.”.
- Finally, we have added a summary of the advantages of using active sensors for labeling in lines 47-50, which reads, “Combining the temporal and spatial resolution of passive instruments with the more accurate classification of atmospheric layers achieved by active sensors is desirable to create an optimal algorithm for classifying clouds and aerosols. By using active instruments to label passive sensor pixels, classification algorithms for passive sensors can be developed that take advantage of the increased accuracy of active sensors.”.

Comment 3: Discussions on potential misclassification in CALIOP of identify spherical fine particles as clouds and how that is going to impact the outcome of this study needs to be discussed in the article. Related to this issue, my biggest concern is that there is little information of the uncertainty/QA procedures used when using CALIOP CAD to identify aerosols and clouds. The CAD > 50 thresholds will likely mark some of the small clouds as aerosols, which is shown in Figure 12. The upper right corner has many fine popcorn clouds, which is marked as potential cloudy in ML output and identified as clear in binary mask. In contrast, the other two cloud products marked this area as cloudy. This can cause large problem in aerosol retrieval. Due to this mislabeling is caused by how clouds are defined, it will not be marked as missing detection of clouds in validation (accuracy score). Plus, an altitude threshold of CAD will mark some elevated aerosols as clouds, such as volcano eruption/stratosphere aerosols, although the percentage of these data will be very small.

Reply 3: This point has been addressed along with comments made by reviewer 3 by discussing the CAD algorithm in section 3.2 and including a more complete description of the caveats of using CALIOP as truth.

- At the end of section 3.2, lines 117-131 now read, "While the 5km algorithm is more sensitive to optically thin clouds, after initial investigation optimum results were found using the 1km L2 cloud-layer version 4.20 product (CAL_LID_L2_01kmCLay-Standard-V4-20) (Young et al., 2018) because the higher spatial resolution leads to increased accuracy of identifying small-scale clouds in AHI pixels at 2km resolution. The version 4 product is used for this study due to improvements made in the cloud-aerosol discrimination (CAD) score algorithm for this product (Liu et al., 2019). The CAD algorithm seeks to discriminate between cloud and aerosol particles, such as fine spherical dust particles and water cloud, by using 5D probability density functions (PDF)s to assign values between -100 and 100 to each layer, with -100 being certainly aerosol and 100 being certainly cloud. The CAD algorithm improves on the previous version by applying the algorithm to all single shot retrievals, which were previously classified as cloud by default, as well as making improvements to identifying elevated aerosol layers and cloud layers under dense aerosols such as smoke (Liu et al., 2019). The version 4 algorithm is validated on the 5km product, but inspection of CAD scores between the 5km and 1km products indicate similar performance. Therefore, although the 5km product is more suitable for use with the CAD score, the 1km product is still appropriate for use in this study. However, it is important to note that extreme cases of aerosols can still lead to classification of aerosol layers as cloud from the CALIOP classification and that small scale (less then 1km across) clouds can be potentially misclassified and be a source of error in the NNs and validation, i.e. the pixel classification by CALIOP is assumed to be true throughout this study, but CALIOP misclassifying layers is a potential source of uncertainty within this study.".

Using the CAD score is, in of itself, a QA procedure that has been used in other studies (Winker et al., 2013, (https://doi.org/10.5194/acp-13-3345-2013); Watson-Parris et al., 2018 (https://doi.org/10.1002/2013JD019527)) and is used in the same way as it has been in our study. However, we have acknowledged that the CAD score may lead to some small clouds being misclassified due to the CAD score.

With regards to the altitude limit, please see the statement given at the start of this document.

Comment 4: It is also not clear to me how the NN model is set up. Is small batch of horizontal pixel from AHI used as input. If so, what is the size of batch? How is CALIOP labeling work for each batch?

Reply 4: The model uses information from a single pixel at 2km resolution and uses this to classify the pixel. This is done for every pixel in an AHI scene. The phrase "to analyse a scene pixel-by-pixel" has been added to line 179 to clarify this point.

Comment 5: In terms of validation, due to the ambiguity in determine CALIOP cloud and thick aerosols, external data, such as ground lidar can be used to validate the cloud/aerosol mask as well as more cases of intense smoke from wildfire and pollution are needed.

Reply 5: This is a great idea. Unfortunately, we could not identify any good quality LIDAR data for the case studies over Australia. However, there are proposals to use weather radar for tracking biomass burning plumes, although this is still in the very early stages of development and may be available for future studies.

Comment 6: Another suggestion is that if the main purpose of the model is to separate thick aerosols from clouds while maintain reliable cloud mask, instead of comparing the cloud/aerosol mask to other cloud mask products, comparisons between ML cloud mask to cloud mask within other aerosol products is more appropriate. Because cloud mask, which is made to remove “unclear” sky, is known to have “clear sky bias”; while aerosol products try their best to preserve these aerosol scenes.

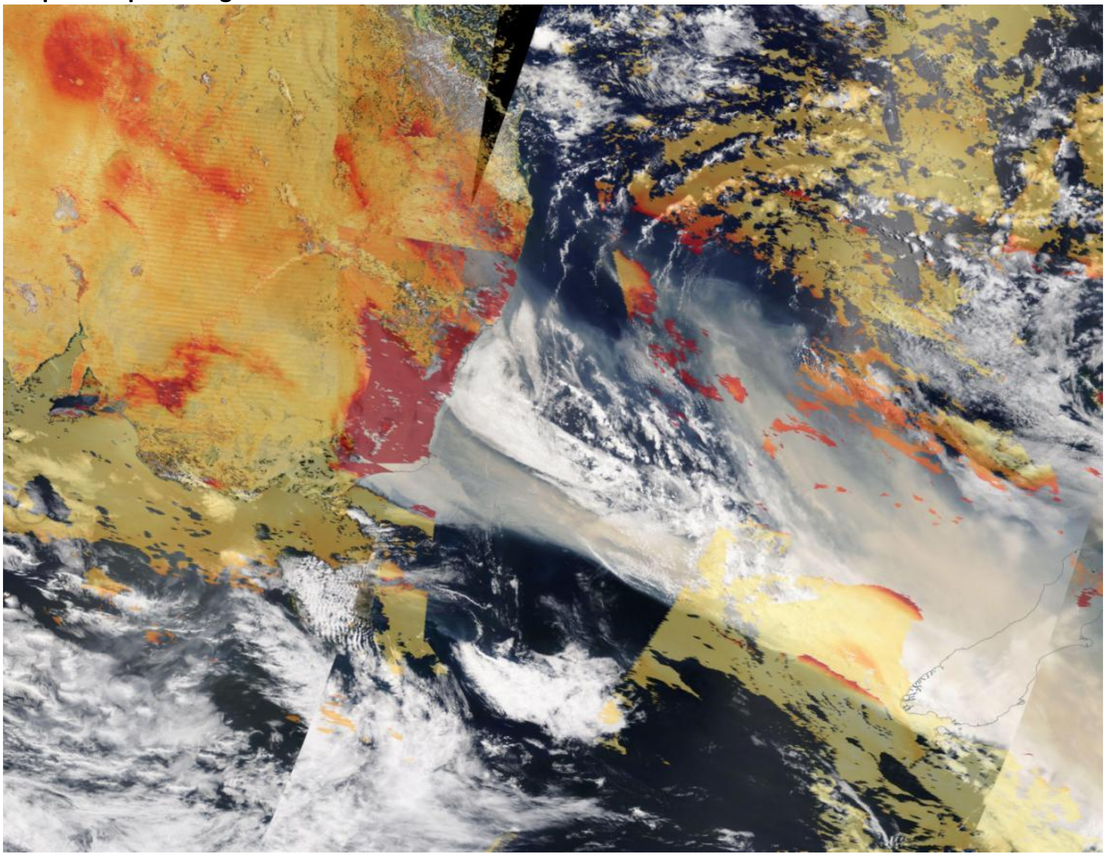
Reply 6: This is an interesting idea. However, from visual inspection of aerosol product masks, it can be seen that these masks also have issues with removing cloud (see MAIAC mask for MODIS taken during the 2019/2020 bushfires in response-specific figures at the end of this document). For example, the JAXA AOD product is based on the deep-blue method (She et al., 2020 (http://dx.doi.org/10.3390/rs12244125)) over ocean and operates in a similar fashion to the JAXA standard cloud mask. It is not obvious that comparing against these masks will make a difference to the statistical analysis, but including these in future studies may be considered.

Comment 7: For reader’s benefits, reword the description of the parallax correction. From my understand, the pseudo-CALIOP vertical profile is generated using layer information from different CALIOP lidar pulse along the AHI airmass pathway. However, the description of the parallax correction is very confusing mentioning the angle from CALIOP needs to match angles from AHI.

Reply 7: This description has been reworded to clarify that the AHI observation angle for each layer is what must be matched to account for parallax.

- Lines 154-161 now read, “The parallax correction for each layer was performed by:
  – Calculating the observation angles of the CALIOP layer as it would be seen by AHI at the position and altitude specified in the CALIOP data, i.e. the angle that corresponds to the dashed line beneath the cloud layer in Fig. 2.
  – The observation angles of the CALIOP layer as seen by AHI were then matched with the observation angles for AHI corresponding to the Earth’s surface.
  – Where the AHI observation angles matched, the layer was assigned to the collocated AHI pixel, i.e. the cloud layer in Fig. 2 would be assigned to the pixel that corresponds with the red star. As the match is to the closest pixel, this leads to a spatial uncertainty of approximately ±1km at nadir for AHI.
  – This was repeated for every layer and a pseudo-CALIOP profile was generated for each AHI pixel. This includes thin layers that AHI may struggle to observe and are accepted as a potential source of error in the final cloud mask.”

Response-specific figures



The MAIAC product from 1st January 2020, showing the retrieval mask misclassifying cloud.

Updated Manuscript Figures and Tables

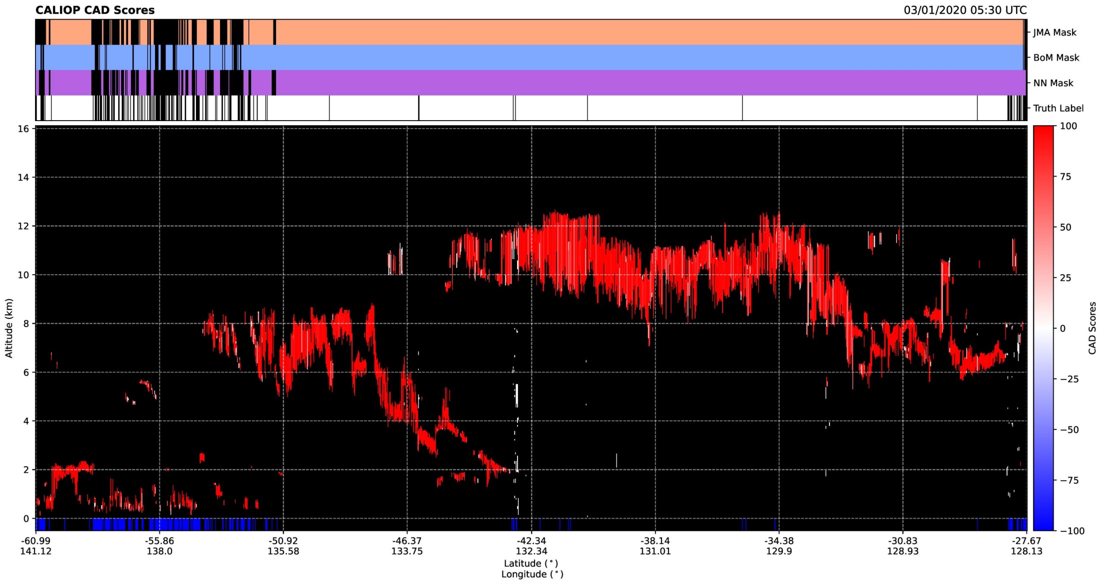


Figure 4 - Old

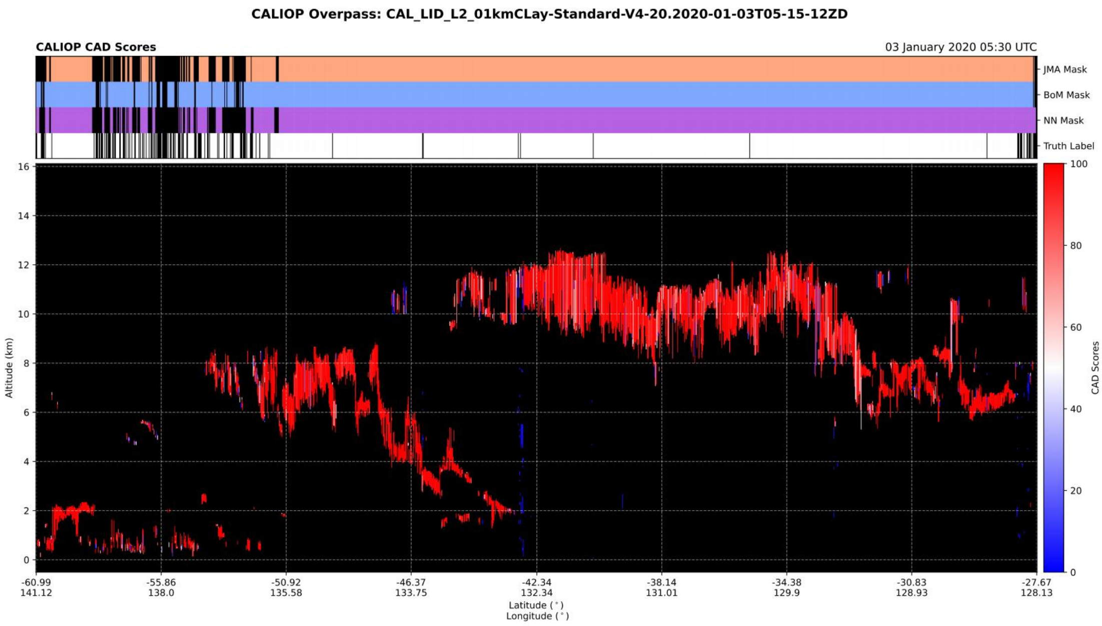


Figure 4 - Updated

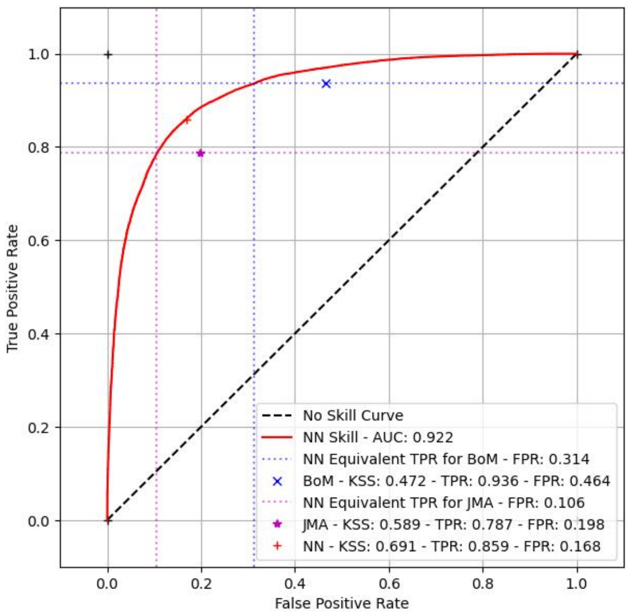


Figure 5 - Old

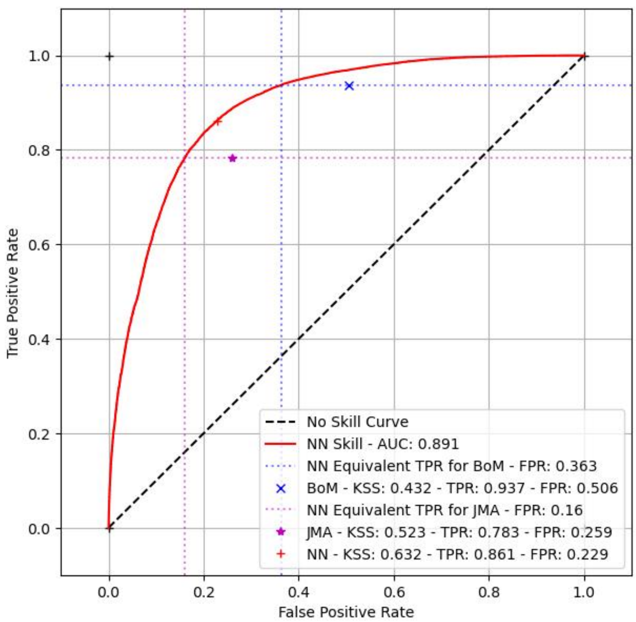


Figure 5 - Updated

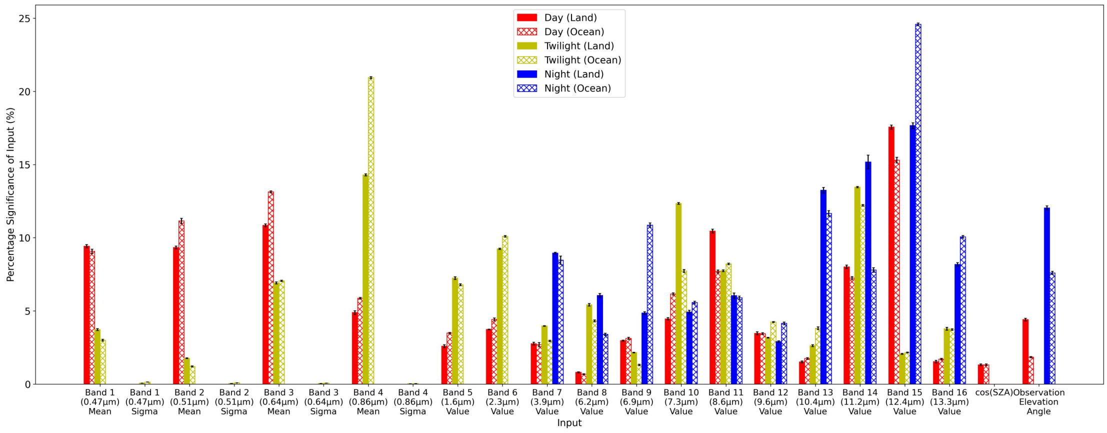


Figure 6 - Old

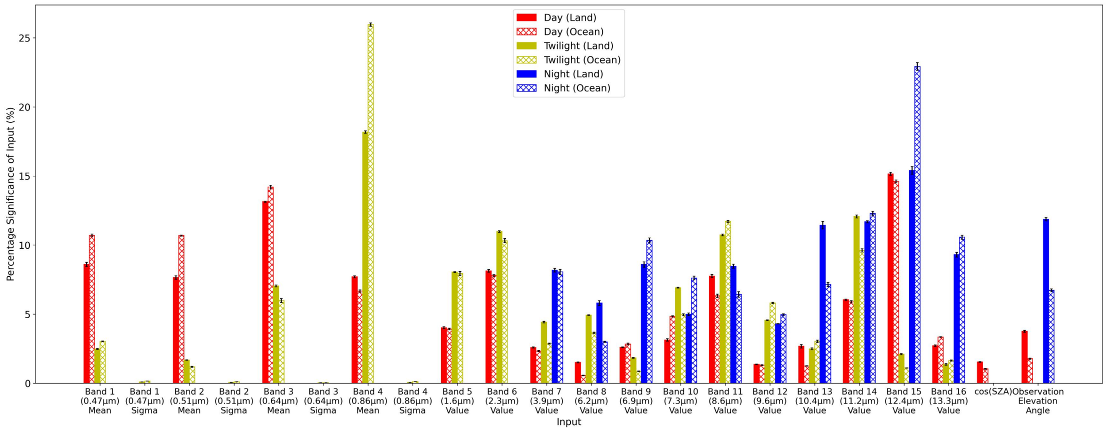


Figure 6 - Updated

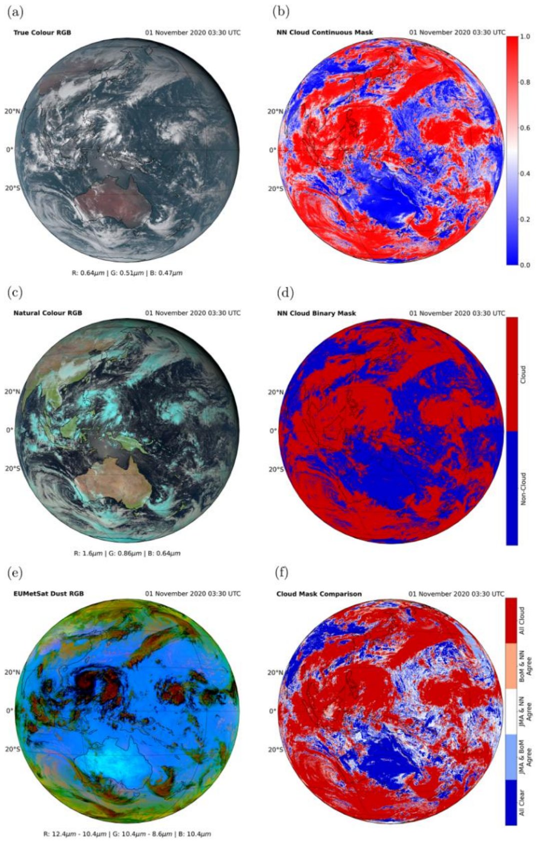


Figure 7 - Old

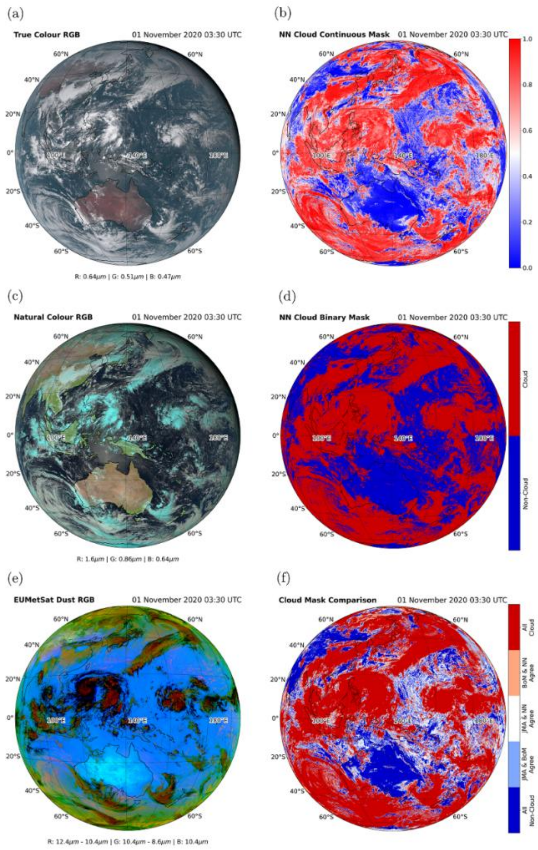


Figure 7 - Updated

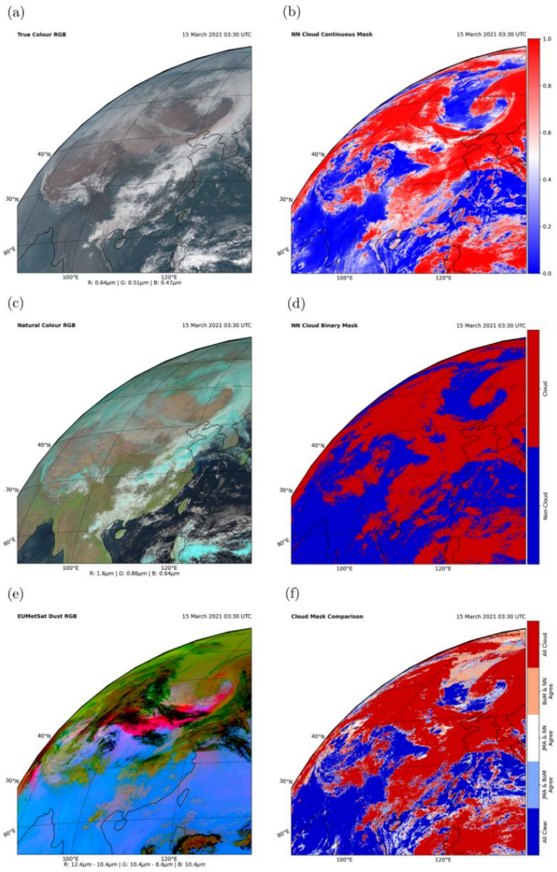


Figure 8 - Old

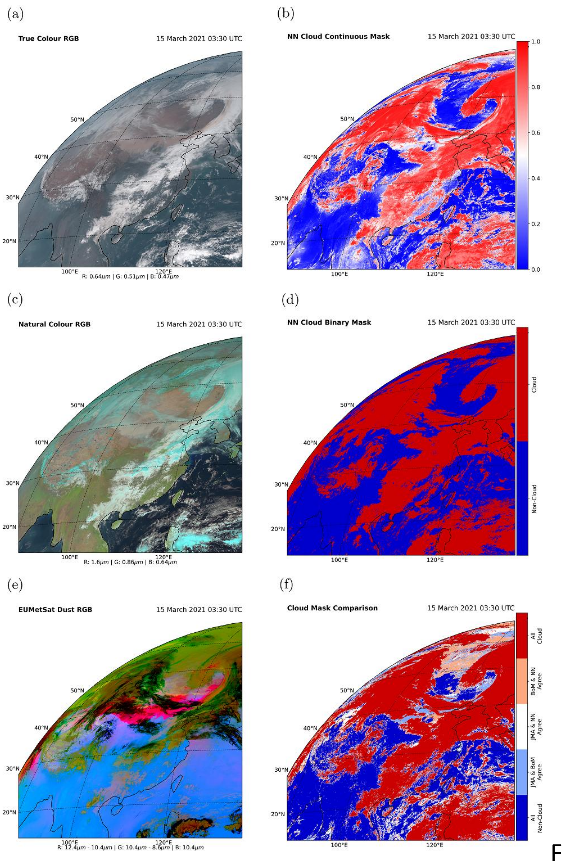


Figure 8 - Updated

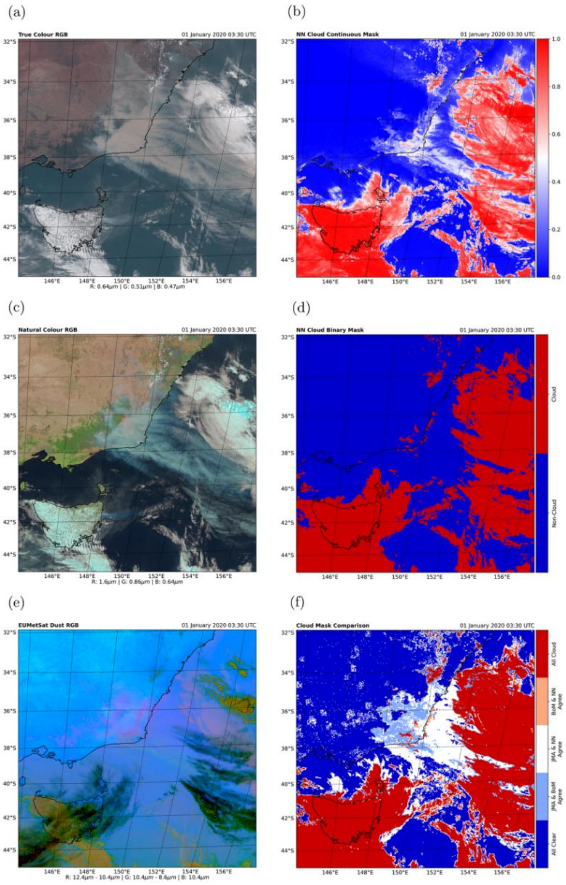


Figure 10 - Old

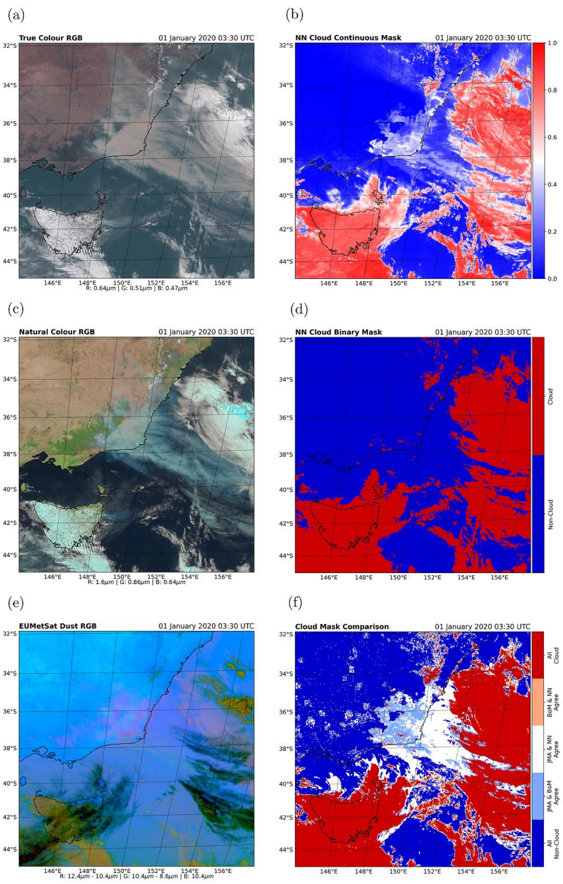


Figure 10 - Updated

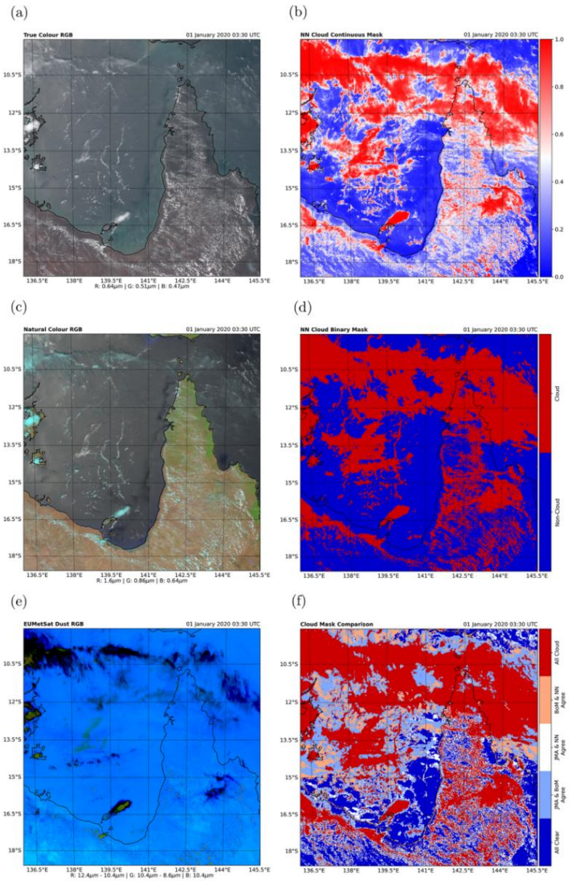


Figure 11 - Old

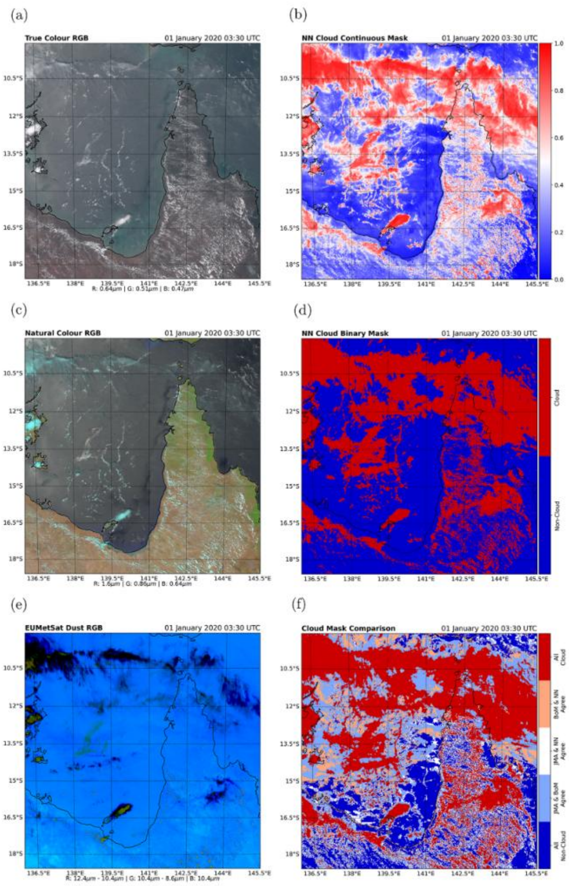


Figure 11 - Updated

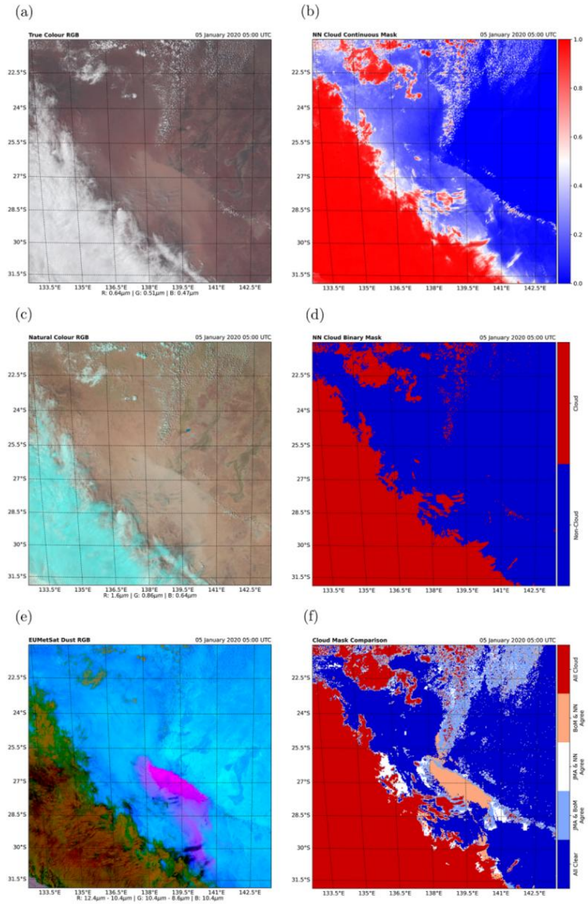


Figure 12 - Old

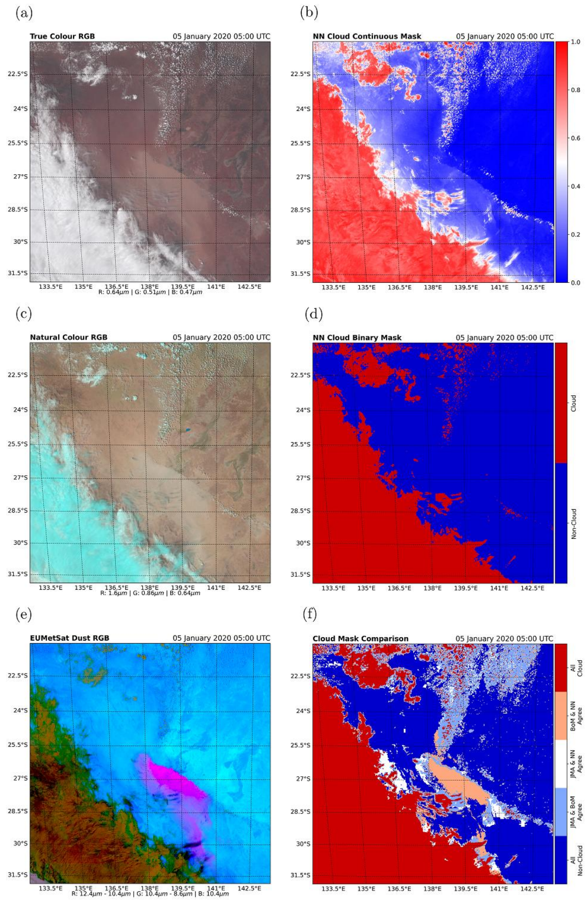


Figure 12 - Updated