Improving Cloud-Phase Detection over Snow and Ice using EarthCARE and Machine Learning

Colin Andres, Roy Grainger, Rui Song

Earth Observation Data Group, Department of Physics, University of Oxford, UK

1 Introduction

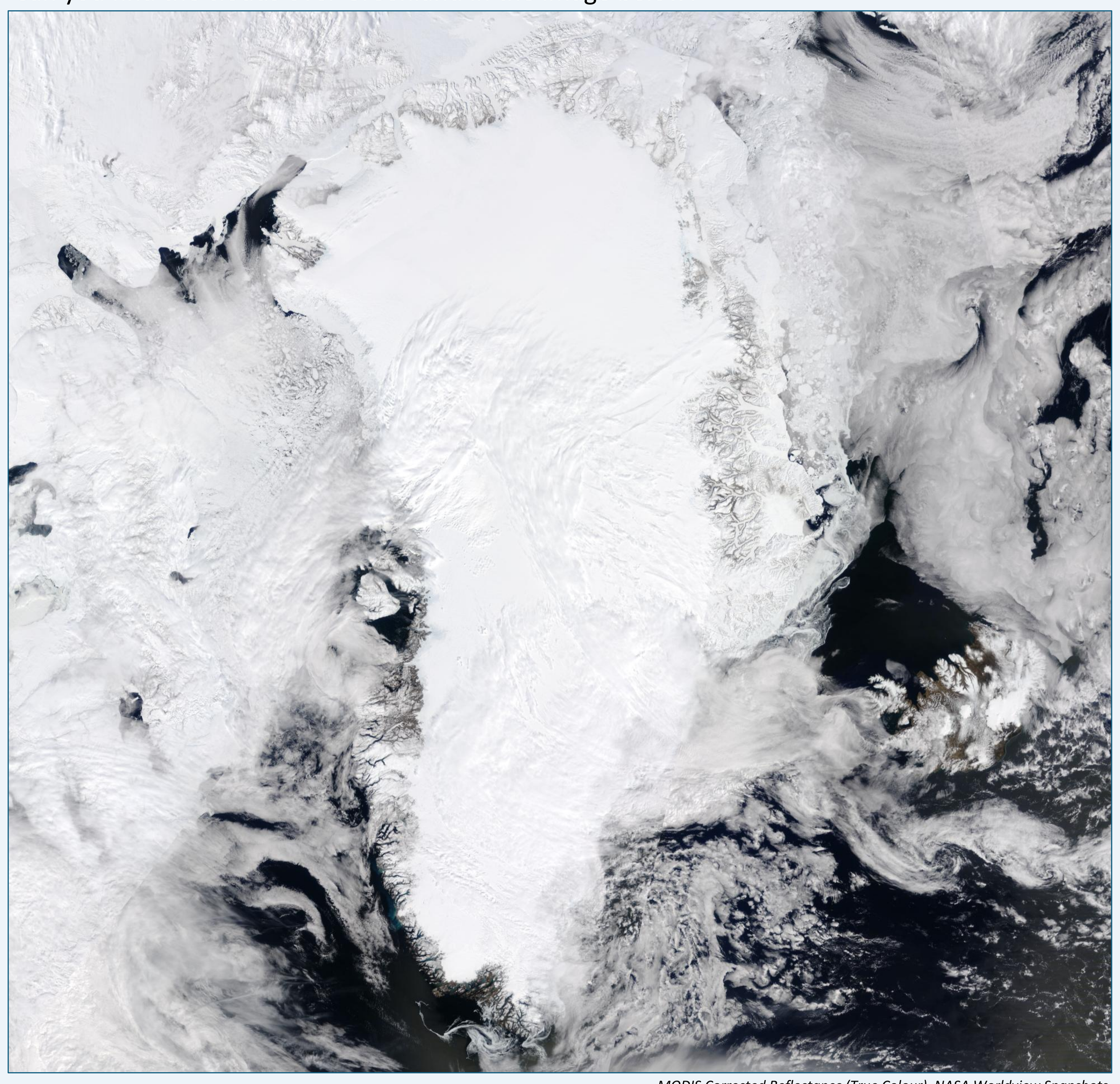
- Identifying clouds over snow- and ice-covered surfaces continues to be a challenge for imaging satellites because their visible reflectances closely resemble those of clouds
- As an active sensor, the ATLID on EarthCARE mitigates this limitation by reducing the impact of surface reflectance and thus greatly improving aerosol and cloud retrievals
- Using machine learning, raw MSI data are trained to match ATLID-informed cloud-phase classifications over snow and ice surfaces
- ATLID products are used as a high confidence reference to supervise model training
- The model has the potential to be modified for use with other satellite data which lack associated active lidar measurements

OBJECTIVES

1. Enhance cloud-phase identification in MSI-only observations
2. Develop a model that can be transferred to use on satellite imager data where there is no associated lidar
3. Gain experience working with EarthCARE data

2 Problem Overview

Can you find the clouds over Greenland in the image below?

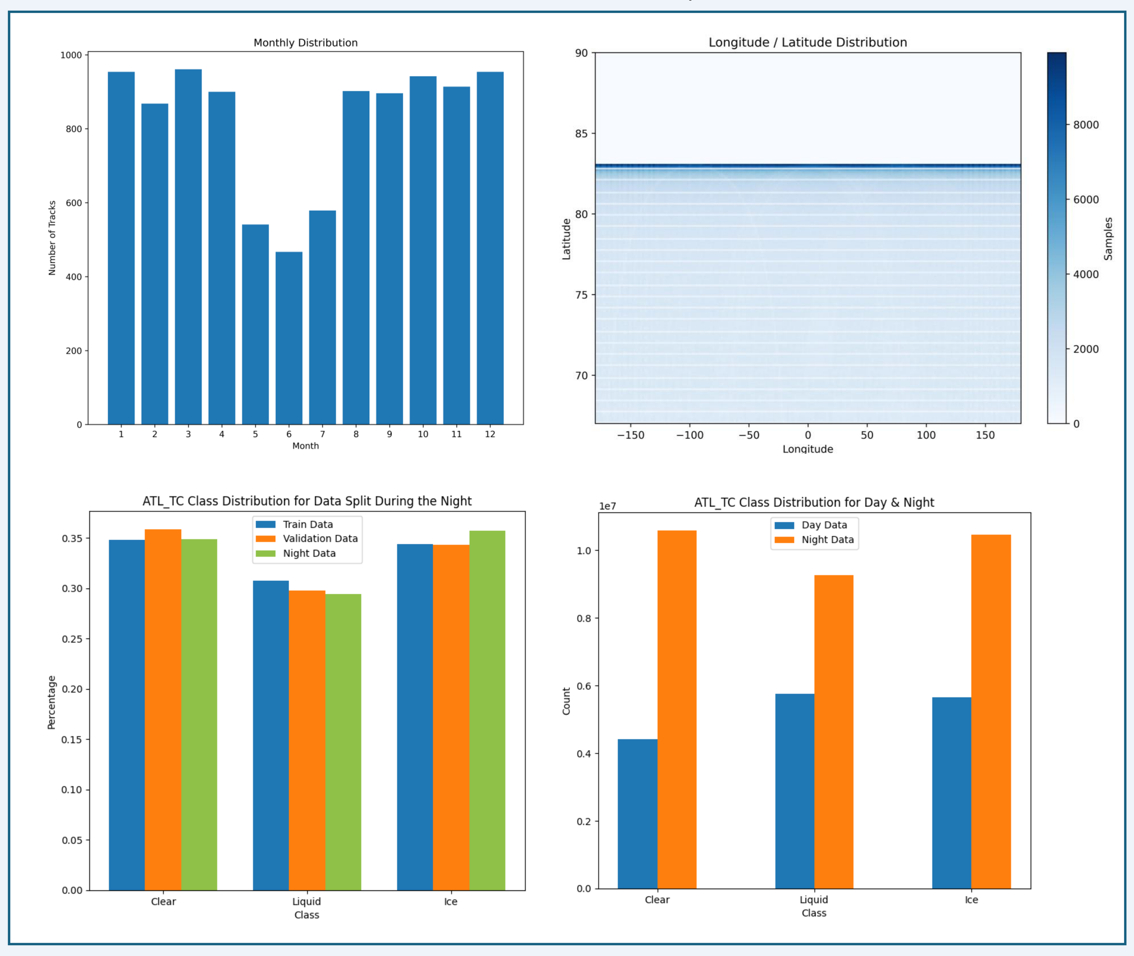


MODIS Corrected Reflectance (True Colour), NASA Worldview Snapshots

- On a pixel-by-pixel basis, it would be hard to distinguish clouds from Greenland's surface
- However, when viewed on larger scales, we – as humans – can tell clouds apart from ice or snow below
- An ML model may be able to identify these cloud structures like us and go beyond by identifying their phase

3 Data

- MSI_RGR_1C was used as the training data and ATL_TC__2A as the target (versions BA+)
- The model used all the data in frame C, the Arctic Circle, from July 25th 2024 to May 4th 2026
- MSI has more along track values than ATLID, so to make the dimensions match, only the nearest MSI points were selected
- For input into the model, only the following parameters were used:
  pixel values & uncertainty, solar azimuth & elevation angles, surface elevation, land flag
- Time, longitude, and latitude were not included in the input to avoid temporal and spatial autocorrelation (Ploton et al., 2020)
- Below are the distributions of all the data across time, space, and class



- In the top left figure, the data is approximately evenly distributed between months except for May, June, and July because the data is from July 25th 2024 to May 4th 2026
- The data are also distributed across longitude and latitude in the Arctic Circle evenly, except near its peak latitude
- In the bottom figures, we see approximately even representation of the target classes – discussed further in the thresholds section – across day and night and the train-validation-test split of 80-10-10

4 ML Models

MLP
A Multi-Layer Perceptron is an ML model of fully connected hidden layers that treats inputs as independent
→ If this model is successful, the problem is easy and more advanced models may not be needed

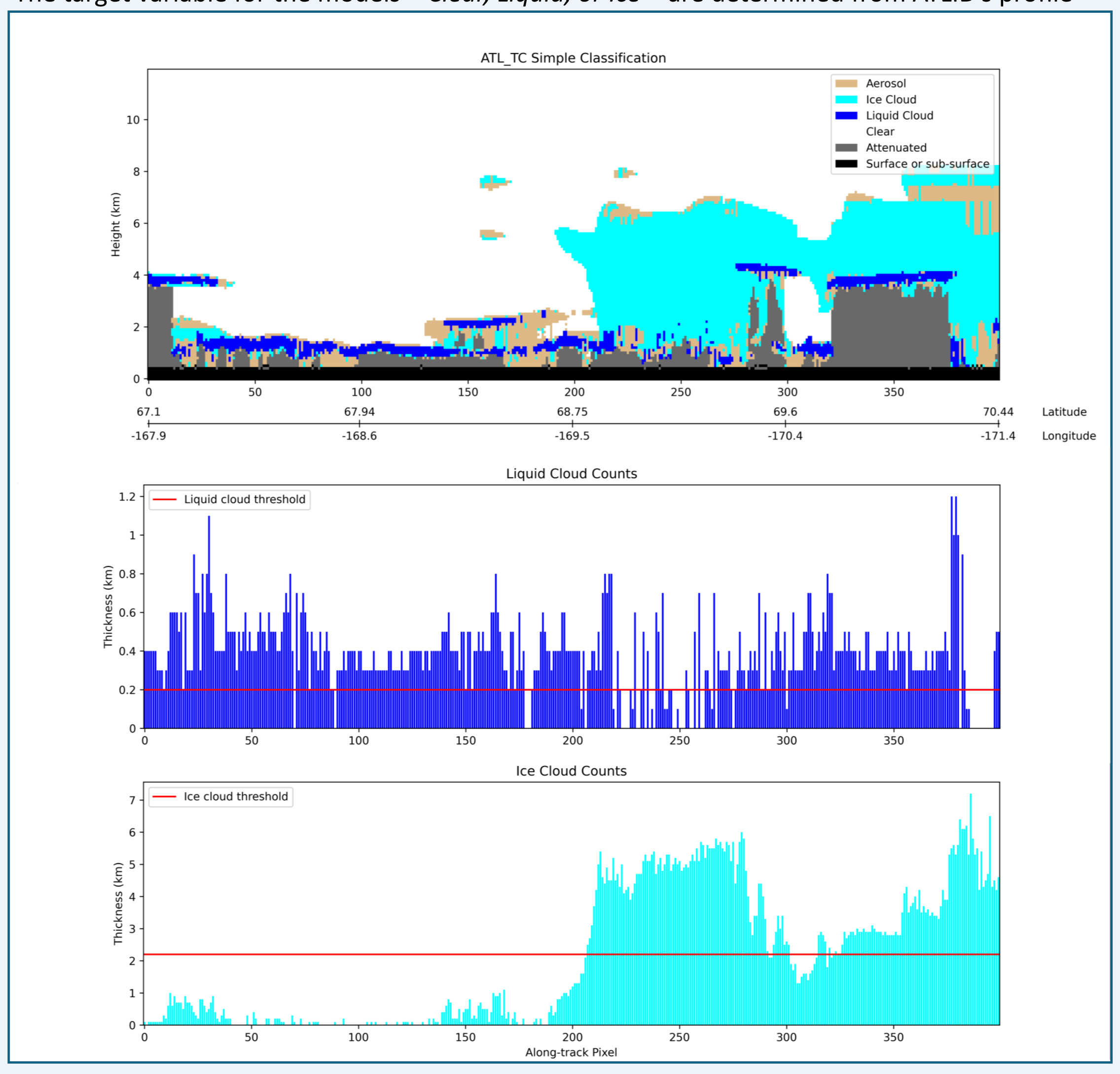
CNN
A Convolutional Neural Network uses convolutions to identify patterns and features in an "image"
→ If this model is successful, local patterns contain significant information

Attention
An attention model investigates the similarity between every point with every other point
→ If this model is successful, it tells us that there is contextual information between points

U-Net
Like a CNN but uses skip connections to maintain high-resolution features
→ If this is successful, local and global patterns are important

5 Thresholding

The target variable for the models – *Clear, Liquid, or Ice* – are determined from ATLID's profile



- MSI can only see the top of clouds, but at times the clouds may be too thin to register
- Cloud thickness thresholds were used to create a target variable based on what clouds MSI can see
- Thresholds were determined with the MSI_CM product and adjusted to minimise the models' losses during training
- If the ice cloud threshold was passed, the target was set to ice cloud
- If the liquid threshold was passed, and the ice threshold was *not* passed, it was set to liquid
- The target was set to clear if no thresholds were passed
- Initially, an overlap category was used when both ice and liquid clouds were present, but it severely reduced performance. At present, it seems there is not enough information in the MSI data to identify thin ice over thick liquid

6 Training

- The MLP, CNN, and Attention models were tested on the nearest MSI_RGR track to ATLID

The MLP performed admirably with an average F1 score across classes of 0.54

→ The problem is not so simple, but a significant amount of information is in a single point

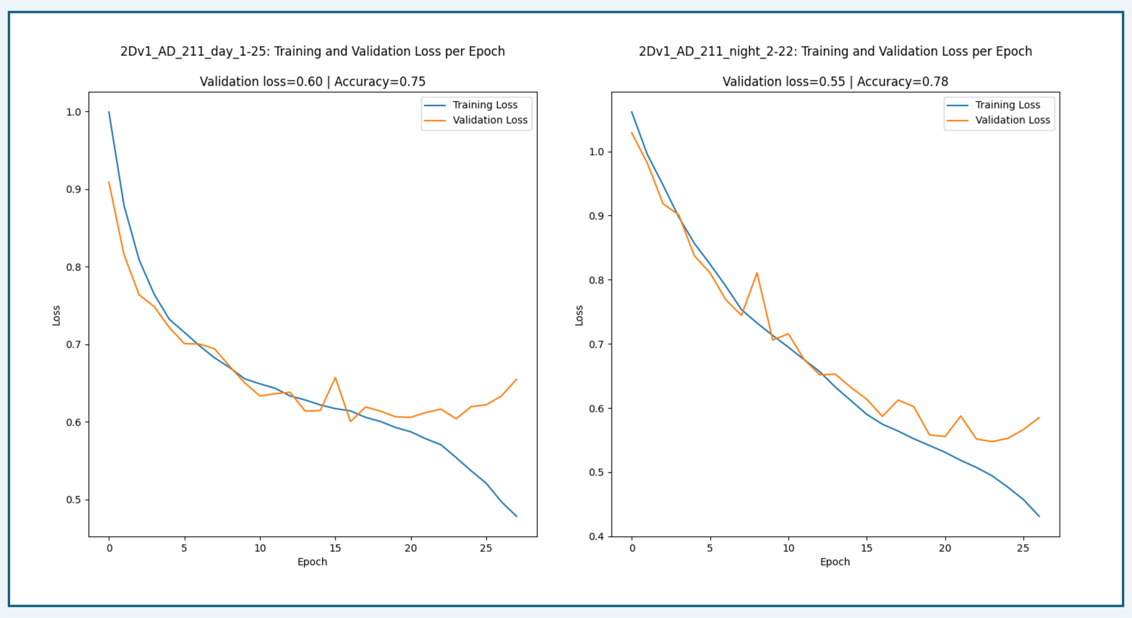
The CNN often predicted everything as a single class

→ The CNN failed and it is difficult to make conclusions

The Attention-based model performed best of the three initial 1D models but was unstable

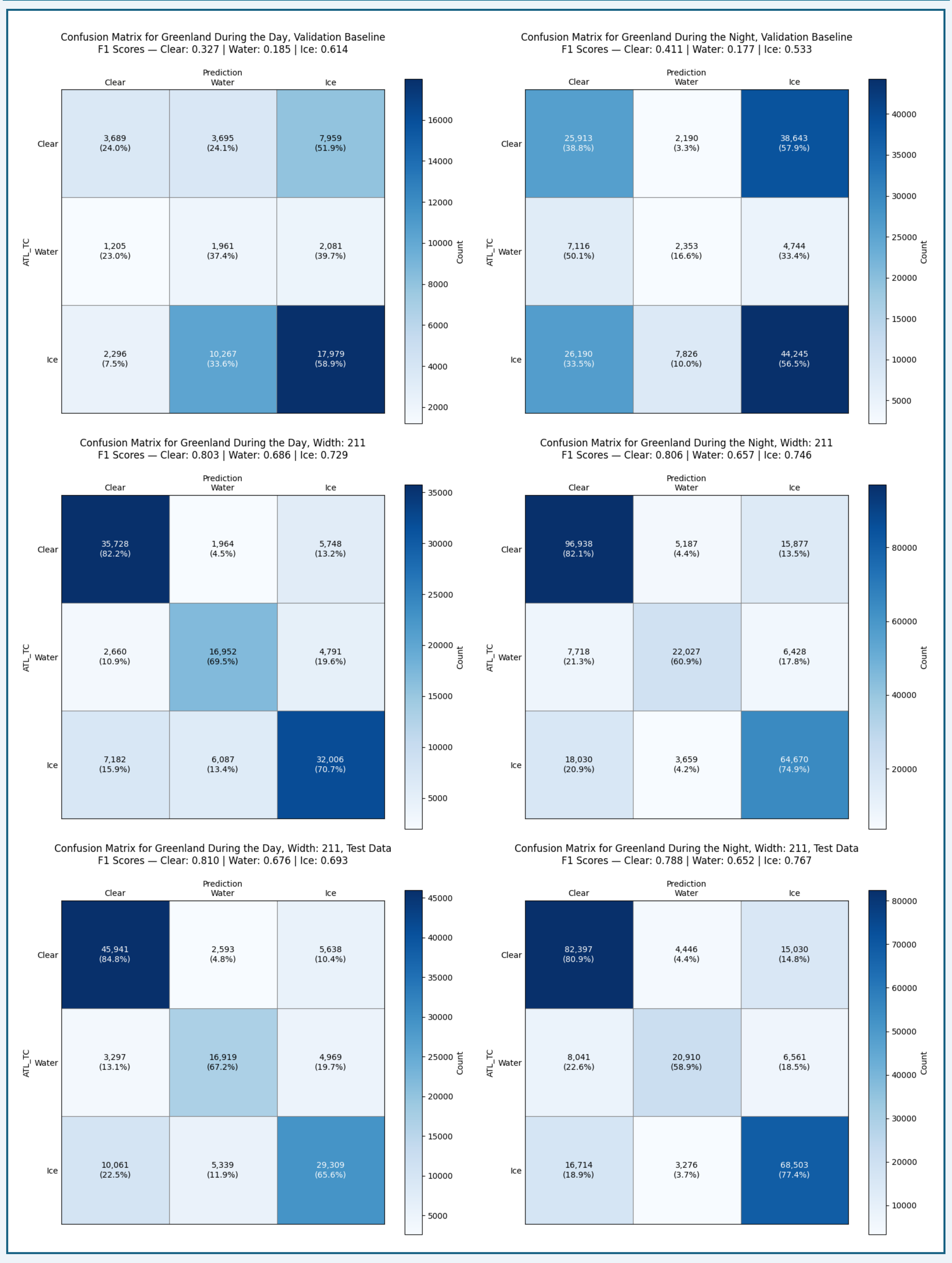
→ There is important contextual information between points

- The 1D models show information is held in local MSI observations and spatial context
- MSI's 2D data may provide additional spatial context to improve model performance
- The U-Net model was selected as the next step because it may recognise small- and large-scale cloud structures and preserve finer details with skip connections
- The U-Net's performance is shown below for the day and night models



- It was found that the best MSI_RGR width for input into the model was the highest value at 211 across track pixels, about 100 km
- In the plots above, it can be seen that the model starts overfitting to the training data and the models' performance on the evaluation data gets worse
- The optimal liquid cloud thickness threshold, i.e. the point where MSI is able to detect liquid clouds, was 200 m for night and 100 m during the day
- The optimal ice cloud thickness threshold was 2.2 km at night and 2.5 km at day
- Separate models were trained for regions over the North Atlantic and Greenland, but they performed worse than the model trained on all data, even when evaluated in their respective regions

7 Results



- The images above show the results of the baseline MSI_CM product (top row) and models using confusion matrices
- A confusion matrix compares how well the predicted classes (x-axis) matches with the true classes (y-axis)
- Ideally, a confusion matrix should look like a diagonal matrix, with perfect agreement between truth and prediction
- In each cell, the number of data points and the row percentage is shown
- To quantify how well the predictions match the true data, we can use the F1 score:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Where precision is the ratio of true positives to all predicted positives and recall is the ratio of true positives to all ground truth positive values.

- Compared to the baseline in the top row, the U-Net models have an average improvement on the validation data set in the F1 score of 0.37 for day and 0.39 for night
- The bottom row shows the test data, which was not touched at all in the training of the model

→ The model's performance on the test data shows average improvements in F1 score over the baseline of 0.36 and 0.38 for day and night, respectively.

- However, it appears that even with the model, there is difficulty in identifying liquid clouds over Greenland at night
- The models' performances at night is better than at day likely because there were more data samples at night

8 Key Findings

- The MSI cloud phase identification is significantly better using an ML model trained on ATLID data
- The model trained on the whole Arctic circle performed better in Greenland and the North Atlantic than the models trained on specifically those locations
- Using the majority of the MSI_RGR swath as input resulted in better model performance than the single nearest track to ATLID, though not as large as expected (improvement in F1 score of 0.01 for the day model and 0.13 for night)
- The models perform significantly better when considering the whole track as opposed to single points
- The U-Net model was successful at identifying the small and large cloud patterns that we can detect with our eyes

9 Conclusions & Outlook

- ML models improve significantly upon the base product
- U-Net seemed to learn at least some sort of cloud structure, ice vs liquid
- Retrieving the cloud phase of every observation in the MSI swath is an improvement of 0.35 in F1 score on average compared to the baseline MSI_CM cloud phase product, reaching accuracies over 70%

Next steps

- Optimise model further by investigating more widths, thresholds, and model parameters
- Test the models on Antarctica
- Test the models on other instruments

Acknowledgements

This work was conducted as part of the Intelligent Earth CDT supported by funding from the UK Research and Innovation Council

We acknowledge the use of data from the EarthCARE mission, courtesy of the European Space Agency (ESA) and the Japan Aerospace Exploration Agency (JAXA).

We acknowledge the use of imagery from the Worldview Snapshots application (https://wvs.earthdata.nasa.gov), part of the Earth Science Data and Information System (ESDIS).

References

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., & Pélissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications, 11(1), 4540. https://doi.org/10.1038/s41467-020-18321-y

Contact

Colin Andres

DPhil student in the Intelligent Earth CDT

colin.andres@jesus.ox.ac.uk

EarthCARE 2026 Science and Validation workshop

8–12 June 2026 | Rhodes House | Oxford, UK