Dear Reviewer,

Thank you for your constructive comments and helpful suggestions, which have allowed us to improve the manuscript and dataset more effectively. Below the comments are responded point-by-point, with the original comments in **black** and our response in **blue**. The revised sentences in the manuscript are indicated in ***italics***.

Overview:

Wu et al. present a global daytime and nighttime low cloud morphology dataset classified based on deep learning methods. The work builds on the algorithms of Yuan et al (2020; for deep learning) and Wang et al (2022; for nighttime COT retrieval) to expand the range of cloud classification, for the first time, to nighttime retrievals. The dataset is novel, unique, and of high quality, with significant potential for use in cloud and climate studies. However, I find that the manuscript lacks some important information, particularly regarding their choice of testing data, sensitivity studies, and data screening techniques. Moreover, the data could be presented in a better way for a wider scientific community. Therefore, I recommend a major revision of the manuscript/dataset by addressing the following comments before it can be considered for potential acceptance in ESSD.

We appreciate your recognition of the novelty and quality of our dataset. Regarding to your concerns, we realize the need for further clarification and improvement. In the revised manuscript, we have provided additional details on the selection criteria for the testing data, conducted sensitivity studies to evaluate the robustness of our model, and further clarified our data processing and screening methods. Furthermore, we have improved the presentation of our dataset by providing additional descriptions and modifying its storage format, which significantly enhances its accessibility. We believe these revisions will address your concerns and improve the quality of our manuscript. Please see the response below for further details.

Comments:

1. While the TIR-CNN-based retrieval of cloud properties in Wang et al (2022) could be a better one compared to the TIR-based algorithm, it cannot replace the standard daytime retrieval algorithm in MODIS. Therefore, to justify using the TIR-CNN-based COT used in training the MCC-classification, I suggest the authors include an additional validation in this study, which includes comparing their MCC classification with that from the outputs of a CNN trained on MODIS daytime COT. Even though I agree that the choice of TIR-CNN-based COT is methodologically justified to stay consistent in their application to both daytime and nighttime retrievals, the inclusion of this additional validation/sensitivity study will strengthen their results and the MCC classification dataset.

Thank you for your valuable suggestion. We agree that an additional validation comparing the algorithms based on TIR-CNN-based COT and MODIS daytime COT would be beneficial. In response, we have performed a sensitivity experiment by training a CNN model using MODIS COT and the original three infrared channels. The results indicate that the accuracy of the CNN model based on MODIS COT is 91.3%, which is almost identical to the 91.5% accuracy achieved by the model using the COT retrieved from Wang et al. (2022). The new training results based on MODIS COT are shown in Figure R1. This demonstrates that the TIR-CNN-based COT can effectively serve as a reliable alternative to MODIS COT for cloud classification, which further reinforce the reliability of our model and the MCC classification dataset. Therefore, we have added some supplementary statements regarding this sensitivity experiment in the '2.2 Data' Section, as follows: "*To validate the reliability of using TIR-CNN-based COT as a replacement for MODIS COT, we conducted a sensitivity experiment: comparing our classification with the outputs of a CNN trained on MODIS daytime COT. The results (Fig. S3) showed that the accuracy of both models is nearly identical, indicating that TIR-CNN-based COT is a reliable alternative to MODIS COT.*" (Lines 142-145).

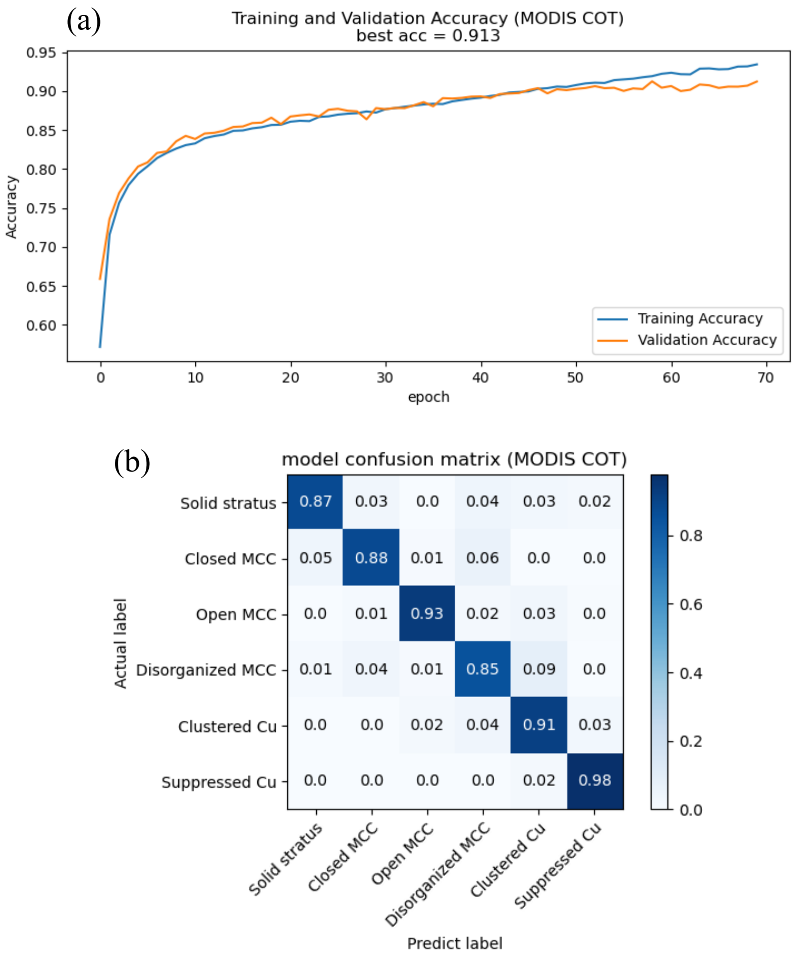

Figure R1. Model training results based on MODIS COT. (a) the model's accuracy on the training and validation datasets. (b) the confusion matrix of the model.

2. The authors use different numbers of samples in each MCC category to train the model.

For example, (0.6 times) 9,900 labeled suppressed Cu are used compared to just (0.6 times) 3,548 solid stratus samples. Shouldn't this disparity impact the performance of the classification for different MCC categories? Can the authors comment on this?

We agree that the disparity in the number of samples across different cloud types can potentially affect the model's performance. When there is an imbalance in the training data, the model might become biased towards the categories with larger sample sizes and perform less accurately on the underrepresented categories.

However, since each category in our training set contains a sufficient number of samples and the sample size ratio between categories does not exceed 4:1, the impact of the sample imbalance on our training should be relatively small.

To approve this, we conducted an experiment in which 2,000 samples were randomly selected from each category to train a new CNN model. The training result (Figure R2) shows a prediction accuracy of 88.8%, which is slightly lower than our previous result. This may be due to the relatively small sample size. Despite this, it indicates that sample imbalance has little impact on our training process. Therefore, we have added some explanations in Section '2.2 Data': *"Despite the disparity in sample sizes within our training dataset, it is capable of yielding superior model performance compared to a balanced dataset (Fig. S2)."* (Lines 132-133)

In the future, we plan to add more manually labeled samples to the category with the fewest samples to explore whether a larger balanced training dataset can lead to a better performance.

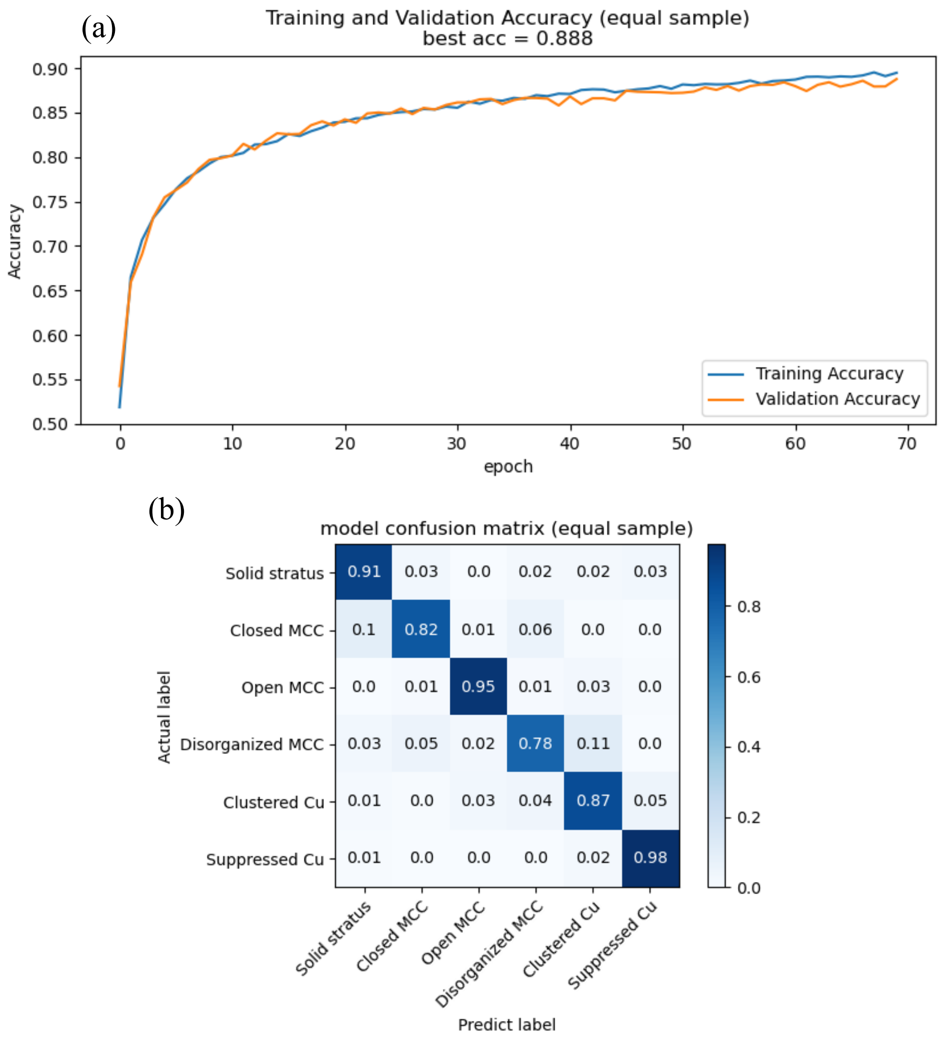

Figure R2. Model training results with 2,000 samples for each category. (a) the model's accuracy on the training and validation datasets. (b) the confusion matrix of the model.

3. Why do the authors interpolate the data within a 1° × 1° scene to 128 × 128 pixels? Even if I consider the finest resolution of 1km, the number of pixels within a 1° × 1° scene would be less than 128 × 128 pixels, leading to extrapolation-related truncation error. Also, did the authors perform any sensitivity test regarding the size of the scene considered in training the model except from 1° × 1°? Wood and Hartmann (2006) use native MODIS 256 × 256 pixels in their classification. Increasing the grid size may reduce the probability of misclassifications (Fig. 3). For instance, considering a smaller domain may result in misclassification of edges of open cells into clustered Cu. In case you achieve a better classification, the resulting dataset can be resampled to a finer grid easily for future use in conjunction with other climate and weather-related datasets.

Thank you for pointing out these issues. For the first question, we did not perform extrapolation; rather, we performed refinement, where the 128 × 128 pixels images obtained through interpolation are included within a 1-degree grid. We have made the following revisions in the manuscript to clarify it: "*To align with conventional climate datasets, we developed a standard 1° gridded datasets by applying the trained model to 1°-resolution images, where the 1°×1° satellite images were interpolated and refined to 128 × 128 pixels.*" (Lines 175-176)

Although this may still lead to issues near the cloud edges, it avoids interference from other grid pixels. Particularly at higher latitudes, a 128 × 128 pixels scene will cover multiple grids, thus the classification of center grid may be interfered by the cloud features in neighboring grids. By directly classifying the cloud pixels within a 1° grid, this type of error can be minimized.

For the secondary question regarding to the sensitivity test of the scene size, first, the size of the scene can influence the definition of cloud types. A larger scene may necessitate a redefinition of cloud types, such as the Sugar, Gravel, Fish, and Flowers categories defined by Stevens et al. (2020) (10 ° ×10 °). Secondly, Yuan et al. (2020) also noted that larger scene sizes increase the probability of multiple different cloud types appearing within a single scene, while smaller scenes may lack sufficient contextual information for effective classification. Beyond the consideration of Yuan et al. (2020), we took additional factors into account and ultimately decided to adopt a standardized grid scene to better address the issue of pixel stretching at high latitudes. Moreover, standardized cloud classification datasets are more convenient for the community to use, as most current studies utilize 1-degree grid datasets for analysis.

Thirdly, we agree that using a larger scene size helps to constrain and reduce the misclassifications caused by smaller domains. Therefore, a more reasonable approach would be classifying the integration of scenes of different sizes using the automatic unsupervised learning, which is an area we plan to explore in our future work.

4. Information is missing regarding why channels 29, 31, and 32 were particularly used in training and classification when multiple other cloud-top-related channels (33-36) are available in MODIS.

Channels 29, 31, and 32 were chosen because they effectively represent cloud properties and cloud-top temperature, which are critical for cloud classification. Specifically, channels 29 (8.7μm) is sensitive to water-vapour absorption, while Channels 31 (10.8μm) and 32 (12.0μm) provide valuable information on cloud-top temperature. In contrast, Channels 33-36 are more focused on cloud-top altitude and other related properties. We previously conducted an experiment using all six channels (29, 31, 32, 33, 34, and 35) for model training, the results were very similar to those obtained using channels 29, 31, 32 only (unfortunately, the experimental data from that trial was not properly archived). In order to reduce the amount of data, we ultimately chose the three channels 29, 31, 32 as inputs.

Therefore, we added explanations in the manuscript to justify our selection of these three channels: *"Thermal infrared (TIR) channels 29 (8.7μm), 31 (10.8μm) and 32*

(12.0μm) were specifically chosen as they most effectively represent the cloud properties and cloud-top temperature." (Lines 135-136)

5. What are the parallel yellow and red lines in the panels of Figure 4? Are these physical and being used in classification or graphics-related artefacts?

The striped patterns (yellow and red lines) visible in Figure 4 are graphical artifacts generated by satellite sensors. Although these artifacts may create some visual noise, they have minimal effect on our pattern identification process since our model relies on qualitative pattern recognition rather than quantitative analysis. Furthermore, the CNN model can filter the noise out, so we opted not to eliminate the striped noise during the training and classification processes.

Nevertheless, we also tried several methods to eliminate this noise and improve the visual quality, such as mean filtering, Fourier transform, and directional filtering. However, the stripe noise in our data is not traditionally horizontal or vertical, and there are no significant numerical characteristics, so none of these methods were effective in removing it. When the noise was removed through these methods, the image became very blurry, as shown in the examples below. Perhaps AI-based methods could help eliminate it, and we will continue to try that.

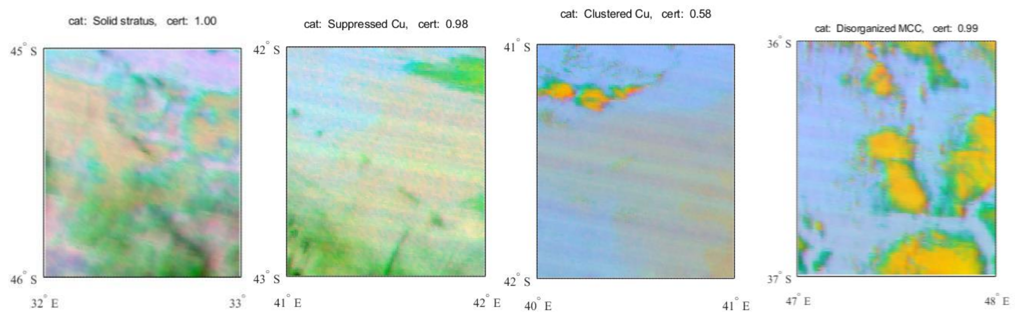

Figure R3. The image processed with directional filtering.

6. Regarding the dataset, I highly recommend using standard data formats used in atmospheric sciences like netCDF and HDF for easy cross-platform and cross-software accessibility. Not all users will be accustomed to the Python-specific NumPy format.

We fully agree that using standard data formats would enhance cross-platform and cross-software accessibility. In response, we have converted all the files previously in NumPy format to HDF files.

7. Since this is a data-descriptor paper, some important information on the contents (variables and how they are calculated) and the file nomenclature should be included

in the manuscript. It may be presented as a separate sub-section within Section 2 and summarized using an additional table. This information is currently missing from the manuscript.

Indeed, our manuscript lacks a description of the dataset-related content. We have added the following sub-section in Section 2 to explain the contents of our dataset:

"2.3 Marine Low-cloud Mesoscale Morphology Dataset
Our cloud dataset provides global classifications of daytime and nighttime marine low-cloud mesoscale morphology for the years 2018-2022, with a spatial resolution of 1° × 1° and a temporal resolution of 5 minutes. The dataset is provided in two kinds of files: those prefixed with "day" store the daytime classification results for each year, while files with the prefix "night" contain the nighttime classification results for each year. Both sets of files include the same variables. Table 3 provides an overview of the variables and their associated information. The key variables in the dataset include 'date' (representing the time of the 1° × 1° scene, formatted as the MODIS granule date), 'lon' and 'lat' (indicating the central longitude and latitude), and 'cat' (assigned cloud category, the values from 0 to 5 correspond to 'Solid Stratus', 'Closed MCC', 'Open MCC', 'Disorganized MCC', 'Clustered Cu', and 'Suppressed Cu', respectively). Additionally, 'cert' represents the model certainty, quantifying the probability that the cloud morphology belongs to the assigned category. 'low_cf' denotes the low cloud fraction, and 'COT_CNN', 'CER_CNN', and 'LWP_CNN' provide the in-cloud average cloud optical thickness, effective radius, and liquid water path respectively, as derived from the TIR-CNN model from Wang et al. (2022). The 'Sensor_zenith' variable indicates the scene average sensor zenith angle." (Lines 191-203)

Table 3 Variables of the Daytime and Nighttime Global Marine Low-cloud Mesoscale Morphology Dataset

Variable Name	*Description*	*Source*	*Spatial Resolution*	*Temporal Resolution*	*Units*
date	*Time of the 1°×1° grid point, formatted as 'YYYYDDD.HHHH'*	*MODIS MYD021*	*1° ×1°*	*5 minutes*	-
lon	*Central longitude (-180,180)*	*MODIS MYD021*	*1° ×1°*	*5 minutes*	*degree (°)*
lat	*Central latitude (-60,60)*	*MODIS MYD021*	*1° ×1°*	*5 minutes*	*degree (°)*
cat	*Category of the cloud morphology: 0-Solid stratus, 1-Closed MCC, 2-Open MCC, 3-Disorganized*	*Cloud Classification Model*	*1° ×1°*	*5 minutes*	-

	MCC, 4-Clustered Cu, 5-Suppressed Cu				
cert	*Model certainty*	*Cloud Classification Model*	*1 ° ×1 °*	*5 minutes*	-
low_cf	*Cloud fraction of low clouds*	*MODIS MYD06*	*1 ° ×1 °*	*5 minutes*	-
COT_CNN	*In-cloud average cloud optical thickness (COT)*	*TIR-CNN model of Wang et al. (2022)*	*1 ° ×1 °*	*5 minutes*	-
CER_CNN	*In-cloud average cloud effective radius (CER)*	*TIR-CNN model of Wang et al. (2022)*	*1 ° ×1 °*	*5 minutes*	*μm*
LWP_CNN	*In-cloud average cloud liquid water path (LWP)*	*Calculated from COT_CNN and CER_CNN*	*1 ° ×1 °*	*5 minutes*	*g/m^2*
Sensor_zenith	*Scene average sensor zenith angle*	*MODIS MYD021*	*1 ° ×1 °*	*5 minutes*	*degree (°)*

Minor comments:

1. Line 49-51: More recently Goren et al (2019) showed a similar delay in closed-to-open transition using LES.

Yes, references have been added.

2. Line 60-61: The cloud morphology dataset by Wood and Hartmann (2006) has been expanded to more than a decade of MODIS observations, the Morphology Identification Data Aggregated over the Satellite-era (MIDAS), by McCoy et al (2023).

Thank you for reminding us! We have revised the sentence to: *"Their work was pioneering and has since been extended to more than a decade of MODIS observations by McCoy et al. (2023)"* (Line 62)

3. Line 64: Abbreviation VGG not defined!

The abbreviation "VGG" has already been defined as "Visual Geometry Group" in the manuscript.

4. Line 64: "… for daytime scenes …". All the morphology datasets discussed prior to this point correspond to daytime observations, don't they?

Yes, they did. We have removed the original sentence *"Their dataset has higher spatial resolution, at 128×128 pixel, but also only includes classifications for daytime scenes."*

5. Line 102-103: "Disorganized MCC … larger droplets and lower optical thickness." Can the authors cite studies that have demonstrated this fact?

The citations for this fact have been added, and the original sentence has been checked and revised to: *"Disorganized MCC are a mix of convective elements and extensive stratiform clouds, marked by smaller droplets and lower optical thickness (Yuan et al., 2020; Liu et al., 2024)."* (Lines 104-105)

6. Line 106: Citation missing!

References have been added.

7. Line 116: "spatial resolution of 1 km" This resolution is for nadir pixels. It changes with sensor zenith angle.

Thank you for pointing that! The original sentence has been further clarified as: *"The primary observation data utilized in this study were derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA's Aqua satellite, including the Level-1B radiance product MYD021KM and the Level-2 cloud product MYD06 (Platnick et al., 2017), both with a spatial resolution of 1 km at nadir point."* (Lines 115-117)

8. Line 121: The authors state that they filter out scenes with more than 10% high clouds or ice clouds. How do the authors deal with ice/high cloud pixels in scenes where they are less than 10%? Are they set to missing values and not used in either training or classification steps?

Thank you for pointing out this oversight! We initially overlooked this detail and directly used the remaining ice/high cloud pixels in both training and classification steps, which may introduce noise into the model. In future model iterations, we will exclude these pixels by setting them to missing values, ensuring they do not interfere with our training and classification process.

9. Line 172: How is the reanalysis data co-located? Do you select the nearest timestamp or interpolated the data to MODIS observations?

We interpolated the ERA5 data temporally to match the MODIS observation time. We added a detailed explanation in the manuscript to clarify our co-location method: "*For the purpose of investigating the influence of meteorological conditions on low-cloud morphologies, we conducted some statistical analyses utilizing the co-located hourly ERA5 reanalysis data (1° × 1°) from European Centre for Medium-Range Weather*

Forecasts (ECMWF). The co-location is achieved by spatially selecting the nearest ERA5 grid point to each MODIS observation and temporally interpolating the ERA5 data to match the exact time of the MODIS observations. This ensures accurate alignment between the two datasets in both space and time." (Lines 178-182)

10. A link to the classification dataset is missing in the "Data Availability" section.

The classification dataset (training, validation, and test sets) has been added to the same link as our cloud classification product dataset, along with descriptions of the variables included. We informed in the manuscript that: *"Daytime and nighttime cloud classification datasets as well as our training dataset are accessible on the https://doi.org/10.5281/zenodo.13801408 (Wu et al., 2024)."* (Lines 426-427)

11. No information on the file "example.xlsx" in the data repository.

Thank you for noting that. The description for the file "example.xlsx" has been added into the data repository, which is: "*File 'example.xlsx': A sample of the variable data from our cloud classification dataset, showcasing the classification results of a MODIS granule captured on January 1, 2018, at 00:25 UTC. This sample is provided to help users better understand the content of our dataset."*

Language-related suggestions:

Line 21: Abbreviation RFO defined in abstract is not used.

Fixed.

Line 84: dependent?

Yes, fixed.

Line 91: Prior to "Section 2 intro…", perhaps insert an introductory sentence like "The manuscript is organized as follows."

Done.

Line 184: Abbreviation ML is not defined

Fixed.

Line 210: Remove underscore after Yuan et al (2020)

Done.

Line 383: Consider changing the word "worse"

The sentence has been changed to: *"which largely explains the performance gap between our nighttime model and the daytime model proposed by Yuan et al. (2020)."* (Lines 409-410)

Line 409: "… nightly …"Do you mean nighttime?

Yes, fixed.

References

Eytan, E., Koren, I., Altaratz, O., Kostinski, A. B., and Ronen, A.: Longwave radiative effect of the cloud twilight zone, Nature Geoscience, 13, 669-673, https://doi.org/10.1038/s41561-020-0636-8, 2020.

Klein, S. A. and Hartmann, D. L.: The Seasonal Cycle of Low Stratiform Clouds, Journal of Climate, 6, 1587-1606, https://doi.org/10.1175/1520-0442(1993)006<1587:TSCOLS>2.0.CO;2, 1993.

Lang, F., Ackermann, L., Huang, Y., Truong, S. C. H., Siems, S. T., and Manton, M. J.: A climatology of open and closed mesoscale cellular convection over the Southern Ocean derived from Himawari-8 observations, Atmos. Chem. Phys., 22, 2135-2152, https://doi.org/10.5194/acp-22-2135-2022, 2022.

Liu, J., Zhu, Y., Wang, M., and Rosenfeld, D.: Cloud Susceptibility to Aerosols: Comparing Cloud-Appearance vs. Cloud-Controlling Factors Regimes, EGU General Assembly 2024, Vienna, Austria, 14–19 Apr 2024, EGU24-4059, https://doi.org/10.5194/egusphere-egu24-4059, 2024.

McCoy, I. L., McCoy, D. T., Wood, R., Zuidema, P., and Bender, F. A.-M.: The Role of Mesoscale Cloud Morphology in the Shortwave Cloud Feedback, Geophysical Research Letters, 50, e2022GL101042, https://doi.org/10.1029/2022GL101042, 2023.

Mohrmann, J., Wood, R., Yuan, T., Song, H., Eastman, R., and Oreopoulos, L.: Identifying meteorological influences on marine low-cloud mesoscale morphology using satellite classifications, Atmos. Chem. Phys., 21, 9629-9642, https://doi.org/10.5194/acp-21-9629-2021, 2021.

Platnick, S., Meyer, K. G., King, M. D., Wind, G., Amarasinghe, N., Marchant, B., Arnold, G. T., Zhang, Z., Hubanks, P. A., Holz, R. E., Yang, P., Ridgway, W. L., and Riedi, J.: The MODIS Cloud Optical and Microphysical Products: Collection 6 Updates and Examples From Terra and Aqua, IEEE Transactions on Geoscience and Remote Sensing, 55, 502-525, https://doi.org/10.1109/TGRS.2016.2610522, 2017.

Schulz, H., Eastman, R., and Stevens, B.: Characterization and Evolution of Organized Shallow Convection in the Downstream North Atlantic Trades, Journal of Geophysical Research: Atmospheres, 126, e2021JD034575, https://doi.org/10.1029/2021JD034575, 2021.

Stevens, B., Bony, S., Brogniez, H., Hentgen, L., Hohenegger, C., Kiemle, C., L'Ecuyer, T. S., Naumann, A. K., Schulz, H., Siebesma, P. A., Vial, J., Winker, D. M., and Zuidema, P.: Sugar, gravel, fish and flowers: Mesoscale cloud patterns in the trade winds, Quarterly Journal of the Royal Meteorological Society, 146, 141-152, https://doi.org/10.1002/qj.3662, 2020.

Wang, Q., Zhou, C., Zhuge, X., Liu, C., Weng, F., and Wang, M.: Retrieval of cloud properties from thermal infrared radiometry using convolutional neural network, Remote Sensing of Environment, 278, 113079, https://doi.org/10.1016/j.rse.2022.113079, 2022.

Wu, Y., Liu, J., Zhu, Y., Zhang, Y., Cao, Y., Huang, K.-E., Zheng, B., Wang, Y., Wang, Q., Zhou, C., Liang, Y., Wang, M., and Rosenfeld, D.: Global Classification Dataset of Daytime and Nighttime Marine Low-cloud Mesoscale Morphology [dataset], https://doi.org/10.5281/zenodo.13801408, 2024.

Yuan, T., Song, H., Wood, R., Mohrmann, J., Meyer, K., Oreopoulos, L., and Platnick, S.: Applying deep learning to NASA MODIS data to create a community record of marine low-cloud mesoscale morphology, Atmos. Meas. Tech., 13, 6989-6997, https://doi.org/10.5194/amt-13-6989-2020, 2020.