Manuscript: Lenhardt, J., Quaas, J., Sejdinovic, D., and Klocke, D.: CloudViT: classifying cloud types in global satellite data and in kilometre-resolution simulations using vision transformers, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2024-2724, 2024.

Julien Lenhardt, Johannes Quaas, Dino Sejdinovic, and Daniel Klocke
14.07.2025

Response to Referee #1 on the manuscript "CloudViT: classifying cloud types in global satellite data and in kilometre-resolution simulations using vision transformers" submitted on 25 Mar 2025.

Dear anonymous referee,

We would like to thank you for the comments and the constructive discussion. Please find our response below, in which the review comments are in bold and followed by our response.

The complete edits can be viewed in the revised version of the manuscript which includes tracked changes. The line numbers referenced below correspond to the manuscript's lines before the revisions.

Best regards,
Julien Lenhardt on behalf of the authors

Suggestions for revision or reasons for rejection:

The authors have addressed all of my comments, and the paper might be accepted for publication.
Thank you for your review of our answers and revised manuscript.

Response to Referee #2 on the manuscript "CloudViT: classifying cloud types in global satellite data and in kilometre-resolution simulations using vision transformers" submitted on 10 Apr 2025.

Dear anonymous referee,

We would like to thank you for the comments and the constructive discussion. Please find our response below, in which the review comments are in bold and followed by our response.

The complete edits can be viewed in the revised version of the manuscript which includes tracked changes. The line numbers referenced below correspond to the manuscript's lines before the revisions.

Best regards,
Julien Lenhardt on behalf of the authors

Suggestions for revision or reasons for rejection:

Again, this is ACP and the results should focus much more on the atmospheric physics side of things here. A modelling tool is just a means to an end. Right now, the paper focuses too heavily on the modeling details, but lacks discussion on physics. More critically, my main criticism remains: the performance is not good. It does not seem to be useful if the model constantly confuses one class with another. With so much noise, one cannot obtain meaningful insights using this tool, which defeats the purpose. The revision did not address this concern. I am giving an honest assessment of the paper in its current form. One shall not sugar coat the performance and try to argue that somehow the results are acceptable. I submit that this kind of performance would not pass any conference screening in an average AI/ML meeting. However, if the editor thinks this is acceptable, I am willing to reconsider my decision/recommendation.

1. Cloud type classification has a large body of literature. The authors added a few that are talking about different schemes of cloud types discussed in this paper. More relevant papers such as Wood and Hartman, 2006; Muhlbauer et al., 2014; Stevens et al 2020; Yuan et al., 2020; Geiss et al., 2023; McCoy et al., 2023. They are dealing with cloud types more relevant to what is being studied here, it seems. Yet, none of them are mentioned.
To be more complete in the overview of cloud classification studies, sentences are now added in the relevant section to mention the studies recommended here which are focusing on mesoscale cloud organization and subsequent consequences on cloud radiative effects and feedback as this is indeed relevant to the context of the manuscript.
(l69) "[...] allow to then better constrain radiative effects of mesoscale convection (Wood and Hartmann, 2006; Bony et al., 2020; Stevens et al., 2020) which would prove to be too cumbersome manually. The application of deep learning to the classification of mesoscale cloud patterns in particular (Muhlbauer et al., 2014; Yuan et al. 2020; McCoy et al., 2023) additionally demonstrates how specific cloud organization patterns, observable by experts, can be learned by machine learning

models, and allows a deeper analysis of their radiative effects and characteristics on longer time periods and larger spatial scales."

(l 308) "This could further propagate to the classification performance on the related classes, e.g. mesoscale convection clouds or cumulonimbus, whose intricate patterns are better assessed on their own directly (Bony et al., 2020; Rasp et al., 2020; Stevens et al., 2020; Yuan et al. 2020; McCoy et al., 2023)."

2. The relative occurrence of the cloud types plot is interesting and should be kept, but the absolute number of samples are also very important. It should not be put into the appendix.

The absolute number of samples was added to the Figure 2 on top of each bar of the histogram (cf. revised manuscript) to make the information clearer.

3. The performance is still poor. It does not matter what method one uses and what the limitations are. If the performance is so poor, what is the point of publishing the results? Certainly, the authors do not consider the results here as negative ones, i.e., failed experiments for the community to learn from. They want to push its application for wider usage. The model should not be applied because of its lack of skill. The authors need to either improve the model skill so that it is useful for usage. Alternatively, results can be presented as an interesting but ultimately not successful attempt at using an advanced tool to address a challenging issue. Reporting negative results is meaningful since it shares valuable experience with the community. For this purpose, I suggest the authors include a hypothesis on why the model ultimately failed to produce outstanding performance.

We agree with the reviewer's point that the results should be put into perspective, and discussion points should be added to reflect that throughout the manuscript. Nevertheless, the developed method brings interesting and relevant developments, notably with the straightforward application of cloud classification to both satellite and model data. Such seamless application would be greatly useful with the development of km-scale global simulations, and is worth noting in this manuscript. However, a clear and straightforward message addressing the limitations of the method and its applications is indeed needed in the manuscript. To follow the suggestion of the reviewer, we have adapted the abstract and some paragraphs throughout the manuscript to clarify certain points that we think are important and relevant:

- To clarify the intended message of the study regarding the achieved results, the method's shortcomings, and the improvement potential, the revised manuscript was adapted with the following sentences:

  (l26) "While the application of the method in its current state comes with apparent uncertainties due to limited performance, improvements to mitigate that emerge in the reduction in mismatches between data sources, the extension of the colocated dataset, and the refinement of the classification model. To foster CloudViT's advancements, the dataset and model are available from Zenodo (Lenhardt et al., 2024b)."

  (l111) "Eventually, we discuss the benefits of the presented method, the potential improvements, and the remaining challenges to make CloudViT's usage reliable and capable of achieving notable performance on the classification task."

  (l 468) "Using the classification model developed here thus comes with apparent uncertainties across the different cloud types. Efforts were made with the aim to classify all cloud types consistently from the limited training dataset available but to limited outcomes. The extension of the training dataset appears as an obvious way to purposefully improve the classification performance of the model. An extended colocated dataset would allow stricter

filtering, mainly with respect to the colocation time-window, which would help improve the representativeness of the samples. The analysis of the classification performance shows here the limitations of a reduced-size dataset with potential underlying discrepancies between data sources during colocation."

(l 612) "Cloud type diagnostics such as CloudViT could be a resourceful addition to the panel of assessment methods for model data (Kuma et al., 2023; Kaps et al., 2023) given improvements to achieve remarkable performance in its classification ability as previously described."

(l 617) "Improving the representativeness of the training samples could solve the performance issues faced by the model presented here, and potentially achieve performance aligned with a wider usage of the method for cloud type analysis."

(l 647) "In conclusion, the method presented here showcases and highlights a wide array of potential applications in the study of cloud types, their characteristics and evolution, and their past, current, and future effects on the Earth's climate: from the extension of sparse surface observations to global yearly predictions, and the proof of concept of transferring the method to a global model simulation. Despite the relatively imbalanced performance assessment of the method which shows both great promise in capturing large scale characteristics of cloud types distributions but struggles to capture precisely the features in the training dataset, the design of CloudViT and its straightforward application to model data outputs potentially makes it a useful tool for cloud type diagnostics modulo its performance reliability being improved. We recommend further advancements of the method being firstly focused on data curation and followingly on model tuning once the performance has been raised to desirable levels. To this extent, the necessary datasets and model architecture code are made available on Zenodo (Lenhardt et al., 2024b)."

- A limitation of the study lies, from our point of view, in the design of the input data where the performance is hindered by the colocation between surface and satellite data sources. While the parameters for the colocation process have been explored with the aim to provide a satisfying dataset, it is apparent that it constitutes a remaining point of improvement here. To reflect this, several additions and modifications were made to the revised manuscript:

  (l26) "[...] improvement directions emerge in the extension of the training dataset [...]"

  (l334) "A main caveat arising from colocating these two data sources is the potential mismatch between the actual clouds jointly depicted. Contrarily to methods like Zantedeschi et al. (2019) which relies on joint retrievals of cloud properties and cloud type or Kuma et al. (2023) which aggregates observations at daily time scales, the presented colocated dataset leaves room for misaligned surface observations and satellite retrievals. As it will be also highlighted later on, this potential misalignment between data sources constitutes a hurdle in the development of the cloud classification method. Indeed, if the model needs to learn from satellite data that actually does not visibly fit the surface observation, then the learning process is hindered. Attempts to reduce this risk have not yielded satisfying results. For example, decreasing the time-window described in section 2.2 did not ultimately yield improvements in the classification performance, especially due to generalisation limitations from a lower number of samples. Furthermore, these attempts are mainly limited by the amount of satellite data that would be necessary to build a substantial and consistent colocated dataset which would span a larger timeframe than the two years used in this study."

(l 516) “Even though the limitations of ground-based observations are evident, they still provide quality observations on which a classification model can be trained. The colocation between these surface observations and the satellite retrievals is thus of crucial importance and guides the performance of the later trained model. It partly contributes in the case of CloudViT to a hurdle to achieve notable classification performance. The model, however, shows its ability to generalise from limited samples to consistent and physically-relevant distributions of cloud properties among the predicted cloud types. By refining the training dataset, the improvements can be expected to reflect directly on the classification performance. The characteristics observed in the histograms across cloud types contribute to an increase in confidence in the ability of CloudViT to discern various cloud types in large remote sensing datasets despite the method’s limited ability described in the previous section.”

(l 614) “The hypothesis as to why the model fails to achieve great performance in this study rests heavily on the colocation process between surface observations and satellite data. The method would benefit from including further ground-based observations through the colocation process but then much larger storage and computational facilities would be needed as global MODIS data represents thousands of granules each day. More training samples could simultaneously solve performance issues by providing a clearer vision of the different cloud types for the classification model to learn from. The improvements through a larger training dataset will yield relevant benefits only if the potential mismatches occurring during the colocation process are tackled.”

- The vision transformer model showcases some limitations in its ability to reconstruct samples across all areas with the same accuracy (see following reviewer comment about Figure 4). The spatial patterns of the reconstruction error were not addressed for now in the manuscript, but they hold importance in the assessment of the whole method as it can propagate uncertainties to the subsequent classification. To reflect on this, the following additions or modifications were made in the revised manuscript:

  (l 308) “The results across the training, validation and test datasets are shown in Figure 3 for the training process and some examples of reconstructed samples belonging to all three splits, while Figure 4 highlights the spatial distribution of the reconstruction error per channel and across splits. On the left panel of Figure 3, the losses show a consistent decreasing trend even at the end of the training epochs. The training process was halted after 100 epochs due to computational limitations, but would gain to be extended as the vision transformer’s performance seems to still be improvable. On the right panel of Figure 3, the reconstructions presented for some random samples reveal where the model would benefit from an improved performance: the reconstructions appear realistic, but fail to reproduce the exact sharpness that is visible in the satellite retrievals. While this aspect would not guarantee a decisive improvement in the downstream task which only relies on the encodings, it would greatly help build more trust in the model. A related case that can lead to observed patterns of reconstruction errors in Figure 4 lies in the reconstruction of cloud scenes with convective cells. The invigorated core of the convective cell stands much higher and holds more water compared to its surroundings which can lead to steep gradients in the cloud quantities when observed from space. As the reconstructions are not able to reproduce these features, larger errors can arise from such cloud scenes. This could further propagate to the classification performance on the related classes, e.g. mesoscale convection

clouds or cumulonimbus, whose intricate patterns are better assessed on their own directly (Bony et al., 2020; Rasp et al., 2020; Stevens et al., 2020; Yuan et al. 2020; McCoy et al., 2023). The additional patterns in the reconstruction error of Figure 4, in particular for COT, are visible in some consistent areas over land. A deeper analysis of the spatial generalisation skill of the model than the one presented in section 3.3.2 covering only the colocated dataset might help constrain the spatial generalisation performance of the vision transformer and infer potential performance caveats still remaining."

Figure 4: strong patterns of reconstruction error is concerning since it suggests the model has strong performance variations, depending on the location.

This comment is indeed important to make when assessing the skill of the model. We noticed these patterns during the training, and a visual analysis of some random samples showing larger errors in the reconstruction metric pointed towards different reasons. Generally, as seen on the right panel of Figure 3, the reconstructions appear realistic but lack the sharpness of the input data which leads to a consistent error at the sharp edges of the cloud structures. We acknowledge that this could be improved with a longer training, supported by the decreasing gradient observable in the reconstruction loss. Nevertheless, the model was trained for only 100 epochs due to computing limitations. Some areas of larger errors can be attributed to areas with higher variability of cloud properties in the cloud scenes due to the ITCZ and convection in the Tropics. This contrast is clearly visible for cloud optical thickness where oceanic areas with observed large and uniform cloud structures (e.g. stratocumulus or stratiform structures) show smaller errors. The spatial generalisation of the model from Appendix B.4 was limited to a Northern/Southern hemisphere split due to the low number of samples a more refined split per region/quadrant would have led to. See comment #3 for additions to the manuscript regarding this topic.

Table 2: this shows clearly that current modeling technique does NOT really change the performance too much.

We think that the improvements, albeit limited, seen across the metrics especially more robust to imbalance are still alright. The analysis per cloud type detailed in the manuscript also brings to light the differences in performance for cloud types. The limitations of the method are mentioned, and ways to improve the architecture and the data inputs are highlighted in the revised manuscript (see answer to comment #3.).

Figure 6: Why different colors but single grayscale bar? It is quite confusing to comprehend.

The single grayscale colorbar option was chosen to make the plot less busy and avoid having a colorbar for each subplot. It can indeed make the understandability of the plot more difficult. To this extent, the figure 6 was updated (cf. revised manuscript).

Figure 7: no truth and no error statistics. What is the point?

Regarding the application of the method to the Apollo17 simulation with ICON-Sapphire, there is no ground truth to compare to as the simulation outputs used here span one single day in 1972 for which no global data record of cloud types is available. The inclusion of this section and figure has the aim to showcase the application of the method to model data and this simulation in particular. It is not included with the aim or claim to evaluate thoroughly the representation of clouds in a simulation from the ICON-Sapphire. For clarification, a few sentences were added in the corresponding section 5.2:

(l 574) "However, due to the time period covered by the simulation, no global data record for cloud types can be used to evaluate the representation of cloud types by the ICON-Sapphire through the CloudViT method. A thorough analysis would be feasible for simulations covering a time period for

which climate data records of cloud types are available, for example the ISCCP H-series climate data record (Young et al., 2018) which starts in 1983. The aim here is rather to present as a proof of concept the transfer of the method to model data outputs, and directly describe the outcome objectively."

Table A1: Indeed, there is high imbalance in the training set. The authors need to address the challenge and produce performant models despite this.

See answer to comment #3 for the answer on performance.