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.

Response to Referee #1 on the manuscript "CloudViT: classifying cloud types in global satellite data and in kilometre-resolution simulations using vision transformers" https://doi.org/10.5194/egusphere-2024-2724-RC1

Julien Lenhardt, Johannes Quaas, Dino Sejdinovic, and Daniel Klocke
25.02.2025

Dear anonymous referee,
We would like to thank you for the insightful 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.

Best regards,
Julien Lenhardt on behalf of the authors

Summary:

This paper presents CloudViT, a novel cloud classification method based on Vision Transformers (ViTs) and cloud properties derived from MODIS satellite data. The authors aim to classify cloud types across global datasets using spatial patterns of cloud properties such as cloud top height (CTH), cloud optical thickness (COT), and cloud water path (CWP). The method is evaluated on co-located ground-based observations and satellite data, producing accurate classifications of different cloud types. The approach is further tested with applications to General Circulation Models (GCMs), notably ICON-Sapphire, showcasing CloudViT's ability to generalize cloud type retrievals at kilometer-scale resolution.
Many thanks for this supportive summary of our study.

General comments:

CloudViT leverages self-supervised learning for pretraining and contrastive learning to overcome the limited number of labeled cloud observations. The method is robust, showing competitive performance when compared to traditional methods and CNN-based approaches, and effectively captures global cloud distributions, including complex cloud types like cumuliform and stratiform clouds. I think the paper is suitable for acceptance with minor revisions.

We thank the reviewer for the evaluation of the manuscript. We hope that through the modifications made in the revised manuscript we addressed the minor comments about language and scientific discussion.

Minor Comments:

L142: Change "retrieved" to the verb form "retrieve."
The reviewer is correct, this has been updated.

L177: Replace "requires" with "require" to agree with the plural subject.
The subject is in this case the action of building so it is followed by the verb in singular form. We replaced the “substantial amount” by “large number” for better readability.

L209: In the sentence "this type of model, alongside CNNs, are," replace "are" with the singular verb "is" to agree with the subject "this type of model."
Indeed, the following verbs “are” and “requires” were replaced to fit the singular subject.

L323: Change "cardinal" to "cardinality" to correctly refer to the size or number of elements in a set.
This correction was adapted in the manuscript.

L587-L593: I believe it would be beneficial to discuss the limitations, such as follows : Since MODIS data is collected through near-nadir scanning, observations in high-latitude regions become oblique, leading to distortions and errors in cloud property retrievals, such as cloud top height and optical thickness. This could potentially affect the model’s performance in polar regions.
We thank the reviewer for the valuable suggestion about the limitation in high-latitude MODIS data for the usage of the presented model. This discussion was added in the relevant paragraph.

Response to Referee #2 on the manuscript “CloudViT: classifying cloud types in global satellite data and in kilometre-resolution simulations using vision transformers” https://doi.org/10.5194/egusphere-2024-2724-RC2

Julien Lenhardt, Johannes Quaas, Dino Sejdinovic, and Daniel Klocke
25.02.2025

Dear anonymous referee,
We would like to thank you for the insightful 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.

Best regards,
Julien Lenhardt on behalf of the authors

Summary:

The paper shows results of using a ViT model that is pretrained on MODIS data to classify cloud scenes into 4/10 cloud types as defined by WMO.

General comments:

The authors go into great details at times on model training choices etc. The paper is, however, light on physics. The classification performance of the models shown in the paper is honestly quite poor. Accuracy of 0.46 and F1 score of 0.43 for the best model on test data cannot be treated as state-of-the-art. Note that the statistics for the training data are not that much better either. There is something not quite right about this paper, either the choice of training data, the training procedure, or something else because ViT models are quite capable as the author wrote in the intro, yet the resulting performance is so poor. We urge the authors to investigate this glaring mismatch and improve the model's performance. Otherwise, results from application of such a model are highly unreliable, which defeats the purpose. I therefore suggest a major revision.
Technically, I do not see anything wrong with the general approach in terms of engineering. The authors described how they approached the problem, and given enough data and if the approach is sound, the models used in this paper should give us highly performing models.
The authors also did not do a thorough job at reviewing the literature on cloud type classification using machine learning/ deep learning. Their introduction to the subject seems a bit vague. I suggest the authors pay more attention to the actual physics instead of details of engineering and implementation because this is not an applied machine learning journal.
We thank the reviewer for the crucial comments about the manuscript. The several general points mentioned here are addressed in the following. Overall, the comments and reservations expressed by the reviewer have been addressed in several ways: by adapting the introduction to cloud type classification, by presenting the results in a more critical way, by emphasizing more clearly the

limitations and hurdles remaining in the method’s development/results. The modifications are best presented in the track changes as they span large parts of the manuscript and sometimes consist of entirely new paragraphs.

- The addition of more physical context and analysis is indeed welcomed in the paper. Several sentences/sections have been added throughout the manuscript. In particular, attention was given to the introduction to cloud type classification, giving a larger focus to the physical processes and quantities considered. The corresponding section was adapted in lines 68-94 of the revised manuscript.
- The topic of the model’s performance is worth discussing as the overall classification performance could indeed be improved. In the manuscript, the investigation of the cloud type predictions was done through two different scopes. A focus was put on detailing and analysing the classification performance not only as an aggregated metric but also as class-wise metrics. The dataset of cloud type observations being largely unbalanced, using such metrics allows to shed light on disparities in the model’s performance across cloud type classes without limiting the evaluation to the general aggregated metric (Garcia et al., 2012). In section 4.1, we strived for a new class-focused discussion of the model’s performance. The performance could indeed be improved to have better trust in the model’s prediction but the limited training/testing dataset alongside the potentially misaligned samples present in it contribute to uncertainties about evaluating the model only through these classification metrics. The second focus is to evaluate the model through characteristics outside of the classification metrics on the colocated dataset. The subsequent section 4.2 displays how the global predictions made with the model (despite its fair performance on the colocated dataset) broadly exhibit in their histograms of cloud properties several features relevant for the different cloud types. The conclusions are also encouraging when investigating the global spatial distributions in section 5.1 and push us to think that the classification model is on a good track. The limiting reach of the conclusions made with this version of the classification model was emphasized across the revised manuscript.
- On a related note, the main reason for the train metrics not being very high is that we tried to avoid overfitting as much as possible due to the limited number of samples as indicated at lines 359-361 The synthetic balancing of the dataset appears to provide limited improvements in this case. Overall, better performance could be obtained by complementing the training dataset with more samples in order to also get rid of the potentially noisy cloud type labels. This statement was added in the evaluation (lines 454-460) and conclusion (lines 614-616, 627-628) sections.
- Additionally, the limitations arising from the training dataset are evident and constrain the resulting performance on the test dataset. The colocation process alongside the nature of the data sources (surface observations and satellite retrievals) is intrinsically a limiting factor as both input streams might not always represent the same objects. Misaligned samples might be present in the colocated dataset which render the learning task more difficult for simple classification models. Being aware of this limitation, a choice is made to guarantee the amount of samples available and their representativeness. Nevertheless, the following analysis of the physical soundness of the classified cloud types alongside their spatial distributions is reassuring for the viability of the method.

- The evaluation of the cloud types through distributions of cloud properties and spatial distributions of cloud type occurrences provides a supporting argument to the classification method presented in the manuscript. The evaluation made in section 4.2 describes how main expected features of the cloud types are identified in the cloud properties histograms (Figures 5 and C.1). We argue that, despite the limited number of training samples and the fair performance of the model on the colocated dataset, the identified types in the global dataset are fairly consistent with expected distributions. The noise still present is mainly due to the fact that all cloudy pixels from each scene are used to produce the histograms. A similar comment follows for the spatial distributions of cloud type occurrences (Figures 6 and C.2). Clarifications were added in the revised manuscript to detail the limitations still present in this evaluation due to the initial limited performance of the classification method on the training dataset. To give a bit more details about the model's evaluation and in particular how the types relate to labels coming from another data source, we compare here predictions and colocated Calipso/CloudSat retrievals (2B-CLDCLASS-lidar product; Sassen et al., 2008). A caveat of comparing such datasets is that the spatial scales they represent are quite different. It is probably best to keep comparing climatological distributions like in section 5.1. Only 1093 samples with information from both this dataset and the CloudViT predictions are processed. The number of samples to compare with being fairly low, no overarching conclusions can be made. For the 2B-CLDCLASS-lidar retrievals, the most frequent cloud type across vertical layer is kept and followed by an average over the Calipso/CloudSat track contained in the cloud scene (similar to how the cloud type labels are defined in Zantedeschi et al., 2019, in their CUMULO dataset). Comparing the 9 classes from the 2B-CLDCLASS-lidar product to the 10 cloud types CloudViT predictions, we have first an overall abundance of stratocumulus labels across all retrieved samples as they might be often represented when aggregating over a cloud scene due to their large cloud cover and provide some sort of background class for the size of cloud scene we are considering here. The predicted stratiform clouds from CloudViT are mostly corresponding to stratocumulus (around 50%), nimbostratus, and altostratus in the 2B-CLDCLASS-lidar product. The few 2B-CLDCLASS-lidar cumulus retrievals are also broadly identified in CloudViT predictions. The high cloud class from 2B-CLDCLASS-lidar (for cirrus and cirrostratus clouds) corresponds fairly well to CloudViT predictions with more than 50% of cases matching, in particular for CloudViT predictions of cirrostratus (probably due to the larger cloud cover).
- The application to model data remains of interest in the scope of this manuscript. We find that, despite the performance on the training dataset which could be improved, the spatial distributions presented in FIgures 7 and D.3 reveal interesting features. This section is rather aimed at showcasing the feasibility of transferring the CloudViT method to this particular high-resolution global simulation as a proof of concept as further emphasised in the sections 5.2 and 6. Rather than discarding this whole topic across the paper, we adapted the language, hoping to make the goal of this application clearer.

Minor comments:

Line 34-35: references are needed. This sentence is also a bit disconnected from previous ones.

The sentence was modified to better fit to the rest of the passage and references were added l39-43:

“Typically, separating clouds between low and high (WMO, 1975), and between stratiform and cumuliform (WMO, 1975, 2017), reveals different and complex cloud effects on radiation and precipitation formation (Hartmann et al., 1992; Dhuria and Kyle, 1990). The high variability and complexity of clouds are some of the causes for the uncertainties in estimates [...]”
The following references were added:

- Dhuria, H. L. and Kyle, H. L.: Cloud Types and the Tropical Earth Radiation Budget, J. Clim., 3, 1409–1434, https://doi.org/10.1175/1520-0442(1990)003<1409:CTATTE>2.0.CO;2, 1990.
- Hartmann, D. L., Ockert-Bell, M. E., and Michelsen, M. L.: The Effect of Cloud Type on Earth's Energy Balance: Global Analysis, J. Clim., 5, 1281–1304, https://doi.org/10.1175/1520-0442(1992)005<1281:TEOCTO>2.0.CO;2, 1992.
- WMO: Manual on the observation of clouds and other meteors - International Cloud Atlas Volume I (WMO-No. 407), available at: https://cloudatlas.wmo.int/docs/wmo_407_en-v1.pdf (last access: 30 August 2024), 1975.

lines 39-41: references are needed

The following references were added:

- Ramanathan, V., Cess, R. D., Harrison, E. F., Minnis, P., Barkstrom, B. R., Ahmad, E., and Hartmann, D.: Cloud Radiative Forcing and Climate: Results from the Earth Radiation Budget Experiment, Science, 243, 57–63, https://doi.org/10.1126/science.243.4887.57, 1989.
- Slingo, A.: Sensitivity of the Earth's radiation budget to changes in low clouds, Nature, 343, 49–51 https://doi.org/10.1038/343049a0, 1990.
- Oreopoulos, L., Cho, N., and Lee, D.: New insights about cloud vertical structure from CloudSat and CALIPSO observations, J. Geophys. Res.-Atmos., 122, 9280–9300, https://doi.org/10.1002/2017JD026629, 2017.
- Luo, H., Quaas, J., and Han, Y.: Examining cloud vertical structure and radiative effects from satellite retrievals and evaluation of CMIP6 scenarios, Atmos. Chem. Phys., 23, 8169–8186, https://doi.org/10.5194/acp-23-8169-2023, 2023.

Lines 46-47: please rewrite this sentence because it is confusing to read.

The sentence was reformulated to clarify the meaning l53-55:
“At the same time, applying methods which are engineered on remote sensing data to climate models could become more viable as new global climate models are bridging the gap in resolution by reaching km-scale resolutions.”

Line 51: the classification is not done pixel-wise as far as I'm aware.

Indeed. What is meant here is that the classification is done over spatially-aggregated values of the two cloud properties resulting in using scalar values (in opposition to later presented methods which use spatial information). This was modified to l59-60:
“This classification is performed on scalar fields, setting aside any spatial pattern [...]”

Lines 62-68: incomplete review of cloud type classification

This comment was addressed in the answer to general comments.

Line 82: 'robust retrievals': this is usually not considered a retrieval since in cloud remote sensing community retrievals have specific meaning.

The term “retrievals” is indeed misused here. It was replaced by “estimates”.

Line 103: I'm dubious on the point that such classification provides 'a high level of precision'.
“Precision” was certainly the wrong term to be used here. It was replaced by “detail”.

Line 112: again, I'm not sure why 'retrievals' are used here. It reads off.
There are misuses of the term “retrieval” in some places throughout the manuscript which were replaced by “estimates” as mentioned in a previous comment.

Figure 1: should at least contain panels that show the actual number of training data.
Plots containing sample numbers per cloud type are presented in Figure A.1 (before colocation) and Figure A.2 (after colocation). References to these figures were added in the caption of Figure 1. The number of colocated training samples is detailed in Table A.1 for the different cloud types and is mentioned in section 3.3. A reference to this table was added in the caption of Figure 2 for more clarity.

Lines 177-178: not clear what this sentence means.
This sentence aims at introducing why the choices of method and training were made, in particular since the amount of samples available is minimal. The sentence was modified to l201-202:
“Relying on computer vision models and their large number of trainable parameters usually requires adapting the training strategy, particularly when the training dataset is of modest size.”

Line 178-179: present the actual number of training data samples to give readers a clear idea.
This is indeed not clearly stated or referenced here so a mention of the Table A.1 containing this information was added for clarity l202-204:
“In the presented study, the amount of labels available is greatly reduced during the colocation process (see Table A.1 for the number of samples per cloud type) [...]”

Line 190: 128x128 pixels are not the same as 128kmx128km.
This was clarified in all the instances throughout the manuscript by keeping only the pixels designation as, indeed, there is no 1:1 equivalence in the case of the MODIS retrievals used in this manuscript (for instance distorted retrievals at the edges of the swath, distorted viewing angle close to polar regions, to name some).

Line 197: sloppy language use. Suggest to change.
The sentence was modified to l220-222:
“The benefit of the presented method using either a CNN or a vision transformer, which are models incorporating a certain level of spatial awareness, is that it is consistent with the cloud type identified by the human observer.”

Line 199: is 4&10 much more detailed than 9 types?
The “more detailed” was removed.

Lines 215-217: what? This sentence is quite confusing. Please rewrite.
The sentence was convoluted and did not make a clear statement. It was adapted to l239-241:

“The SiT architecture used in this study is adapted from the seminal vision transformer architecture (Dosovitskiy et al., 2020) by setting the latent dimension to 256, similarly to the CNN architecture introduced in Lenhardt et al. (2024a).”

Line 243: please use terms consistently. Do not use different terms to refer to the same thing.
The sentence was modified to l266-267:
“The actual process of the contrastive learning further requires the use of a momentum encoder to generate different versions for the pairs of samples and their corresponding augmented samples.”

Anything after Table 2 and Figure 5 is not worth discussing too much because the performance is just not acceptable. The authors need to dig deeper into their data, approach, or something else to find ways to improve the results before application.
This comment was addressed in the answer to general comments.