

# **1 CloudViT: classifying cloud types in global satellite data and in 2 kilometre-resolution simulations using vision transformers.**

3

4 Julien LENHARDT <sup>1</sup>, Johannes QUAAS <sup>1,2</sup>, Dino SEJDINOVIC <sup>3</sup>, Daniel KLOCKE <sup>4</sup>

5

6 <sup>1</sup>Leipzig Institute for Meteorology, Universität Leipzig, Leipzig, Germany

7 <sup>2</sup>ScaDS.AI - Center for Scalable Data Analytics and Artificial Intelligence, Leipzig University, Humboldtstraße 25, 04105

8 Leipzig, Germany

9 <sup>3</sup>School of Computer and Mathematical Sciences & Australian Institute for Machine Learning, University of Adelaide, Adelaide,

10 Australia

11 <sup>4</sup>Max Planck Institute for Meteorology (MPI-M), Hamburg, Germany

12 *Correspondence to:* Julien LENHARDT (julien.lenhardt@uni-leipzig.de)

### 13 Abstract

14

15 Clouds constitute, through their interactions with incoming solar radiation and outgoing terrestrial radiation, a fundamental  
16 element of the Earth's climate system. Different cloud types show a variety in cloud microphysical or optical properties, phase,  
17 or vertical extent, and thus disparate radiative effects. Both in observational and model datasets, classifying clouds is important  
18 since different cloud types respond differently to current and future anthropogenic climate change. Cloud types have traditionally  
19 been defined using a simplified partition of a two-dimensional space, e.g., cloud top pressure and optical thickness. In this study,  
20 we present a method called CloudViT (Cloud Vision Transformer) building on surface observations and spatial extracts of cloud  
21 properties from the MODIS instrument to derive cloud types, leveraging spatial patterns with a vision transformer model. The  
22 performance of the model is fair and somewhat hampered by the limited number of samples and the challenging matching  
23 between data sources arising during the collocation process. The method is then evaluated through the distributions of cloud type  
24 properties and global spatial patterns of cloud type occurrences. Subsequently, CloudViT is applied to data from a global  
25 storm-resolving model, showcasing the feasibility of transferring CloudViT to model outputs and providing insights into the  
26 representation of clouds. Future work to enhance the classification performance of CloudViT could be achieved by a larger and  
27 of better-quality training dataset, and by refining the classification model. The global cloud type dataset and the method code  
28 constituting CloudViT are available from Zenodo (Lenhardt et al., 2024b).

29

## 30 1 Introduction

31

32 Clouds form an essential component in the Earth's climate, by impacting the atmospheric energy budget and water cycle, and by  
33 influencing the reflected solar radiation as well as the outgoing terrestrial radiation fluxes. Clouds are highly variable spatially  
34 and temporally, and occur in a large variety of types (Howard, 1803; WMO, 2017). Typically, separating clouds between low and  
35 high (WMO, 1975), and between stratiform and cumuliform (WMO, 1975, 2017), reveals different and complex cloud effects on  
36 processes such as radiation and precipitation formation (Hartmann et al., 1992; Dhuria and Kyle, 1990). The high variability and  
37 complexity of clouds are some of the causes for the uncertainties in estimates of their response to anthropogenic climate change  
38 both currently and in the future (Boucher et al., 2013; Forster et al., 2021). These uncertainties manifest both in observational  
39 datasets for which the aim is to constrain past and current effects, and in climate models where cloud representation is of utmost  
40 importance to properly constrain future scenarios. Through the phase (liquid, ice or mixed), the droplet size distribution, the  
41 vertical structure or other micro- and macro-physical properties, different cloud types can lead to drastically diverse radiative  
42 effects making the cloud type a property of interest to help describe their involvement in the weather and climate system  
43 (Ramanathan et al., 1989; Slingo, 1990; Oreopoulos et al., 2017; Luo et al., 2023). Unravelling and understanding trends in  
44 clouds has become more tractable in recent decades due to the large amount of remote sensing data made available globally on a  
45 daily basis. However, analysing such extensive datasets manually becomes challenging, especially with the goal of extracting  
46 meaningful information about different cloud types based on their patterns, microphysical properties or radiative effects.  
47 Algorithms have taken over this complex task but still struggle to provide objective groupings out of the intricate spatio-temporal  
48 patterns observed in remote sensing data. At the same time, applying methods which are engineered on remote sensing data to  
49 climate models could become more viable as new global climate models are bridging the gap in resolution by reaching km-scale  
50 resolutions, though this transfer to climate model data comes with its own challenges.

51 Traditional cloud classification methods are built on simple characteristics. The standard classification developed as part of the  
52 International Satellite Cloud Climatology Project (ISCCP) relies on three levels (low, medium, high) of cloud altitude using as  
53 proxy the cloud top pressure (CTP) and three thresholds of cloud optical thickness (COT), defining overall nine cloud types  
54 (Rossow et al., 1991). This classification is performed on scalar fields, setting aside any spatial pattern in the cloud field from  
55 which information could be obtained to better inform the classification process. Relying on the same type of two-dimensional  
56 histograms, recent methods have been developed aiming at refining the created clusters and partially relaxing the constraints on  
57 the pre-defined thresholds (Tzallas et al., 2022). The reason to choose the two parameters is that such a classification lends itself  
58 to the analysis of cloud radiative effects: the cloud radiative effect in the solar is a monotonic function of COT, the one in the  
59 terrestrial spectrum, of CTP. However, one might be interested in sensitivities of cloud thickness or water content to different  
60 drivers (e.g., aerosols) for given cloud types, which is hampered by using CTP and COT to define the types. Also, COT does not  
61 map well onto the distinction between cumuliform and stratiform clouds. For such reasons, Unglaub et al. (2020) defined cloud  
62 regimes from cloud base height and variability in cloud top height, hinting at the added value of some measure of spatial  
63 variability and pattern. However, to leverage spatial structure and textures, cloud classification methods based on artificial  
64 intelligence (AI) have opened new avenues of research built upon vast amounts of remote sensing data. For example, using  
65 convolutional neural networks (CNNs; LeCun et al., 1989; LeCun et al., 1995), Zhang et al. (2018) use ground-based images and  
66 human-labelled cloud types to develop a model for meteorological cloud classification and support weather prediction tasks.  
67 Using a similar architecture, Rasp et al. (2020) classify clouds from expert-labelled satellite images of four different cloud  
68 organisation patterns in the trades. This method further emphasises how expert knowledge to identify cloud patterns can be  
69 learned by CNN models and allow to then better constrain radiative effects of mesoscale convection (Bony et al., 2019) which  
70 would prove to be too cumbersome manually. These studies both rely on human observers to initially classify clouds or cloud  
71 patterns directly from images, relying on visual aspects to distinguish clouds, and subsequently linking the identified cloud types  
72 to local meteorological conditions. Kuma et al. (2023) also capitalize on ground-based observations but connect them to  
73 shortwave and longwave radiation satellite retrievals at coarser spatial and temporal resolutions. The method relies on identifying  
74 patterns directly in radiation retrievals to associate them to daily occurrence probabilities of cloud types. This method has the  
75 benefit of being able to be used on outputs from large ensembles of global model simulations and reanalysis datasets which cover  
76 extended time-scales compared to observational datasets. Relying on similar model architectures, Zantedeschi et al. (2019) and  
77 Kaps et al. (2023) classify cloud types derived from active remote sensing labels. The study from Kaps et al. (2023) capitalizes  
78 on the model from Zantedeschi et al. (2019) to extrapolate cloud type estimates using global passive remote sensing data, and  
79 jointly trains a model on coarsened data with spatial resolution similar to current global climate models. Other methods have  
80 been developed without the use of cloud type labels, drawing conclusions from clusters appearing in large remote sensing  
81 radiation retrievals (Kurihana et al., 2022). In general, the developed methods rely on identifying characteristic patterns arising in

82 images (related to visible features of cloud types), radiation retrievals (related to radiative properties of cloud types), or cloud  
83 properties retrievals (related to physical properties of cloud types). Each choice of cloud type labels introduces a certain level of  
84 subjectivity in the derived cloud types. For example, there is less subjectivity in the expert-labelled images than in the produced  
85 cloud clusters, which naturally introduces some subsequent biases. Choosing certain input quantities also physically constrains  
86 the variability of cloud type properties which can hinder the interpretation of the derived cloud type estimates. However, the  
87 transferability to global climate model outputs is a great advantage of some of these methods as they provide a crucial way to  
88 diagnose the representation of clouds in climate models and push towards reducing uncertainties in representing future-climate  
89 clouds (Kuma et al. 2023; Kaps et al. 2023).

90 In this study, we investigate the classification of clouds by merging surface observations of cloud types and passive satellite  
91 retrievals of cloud properties, building a method called CloudViT (Cloud Vision Transformer). Following a similar methodology  
92 from previous work (Lenhardt et al., 2024a), we define cloud scenes as tiles of 128x128 pixels which encompass cloud  
93 microphysical and optical properties at a 1 km horizontal resolution. The employed cloud properties are from the MODerate  
94 Resolution Imaging Spectroradiometer (MODIS, Platnick et al. (2017)), and more particularly the cloud top height (CTH), the  
95 cloud optical thickness (COT) and the cloud water path (CWP), which are paired with surface network observations of cloud  
96 types (cf. Table 1). To harness the spatial aspect of the cloud scene and extract relevant features from the input cloud properties,  
97 we resort to computer vision models based on CNNs and transformers (Dosovitskiy et al., 2020). Firstly, a vision transformer  
98 model is trained in a self-supervised setting to create a condensed latent representation of the input cloud field. Subsequently, a  
99 simpler classification model is fitted to predict the cloud type corresponding to the cloud scene, learning from the labels of a  
100 wide ground-based observation network. The formulated method has the goal to produce estimates of cloud types while  
101 generalising from the local ground observations to global distributions, increasing both the temporal and spatial coverage. The  
102 method relies partly on the assumption that the observed cloud types exist on scales similar to the extent of the tiles, and  
103 additionally builds on the spatial patterns characteristic of different cloud types. Moreover, as the ground-based cloud type  
104 observations provide consistent labels which are only available at sparse locations, we can leverage long-standing instruments  
105 like MODIS to design an algorithm based on satellite retrievals suited to generalisation to global distributions.

106 Firstly, we introduce in section 2 the different datasets used in the study alongside the collocation process between the  
107 ground-based and satellite datasets. Subsequently, the different components of the CloudViT method are presented in section 3,  
108 supported by sensitivity studies about the generalisation skill of the models and the benefits of the spatial context. In section 4,  
109 we evaluate the method and investigate the distribution of cloud properties following the predicted cloud types. The results in  
110 section 5 focus on the extension to a global distribution of cloud types and present a first application to climate model data.  
111 Eventually, we discuss the benefits of the presented method, the potential improvements and the remaining challenges.

## 112 2 Data

113

### 114 2.1 Surface observations

115

116 The cloud type observations used in this study come from two similar global observation datasets maintained by the UK Met  
117 Office, one providing observations made at sea (Met Office, 2006) and the second providing observations made on land (Met  
118 Office, 2008). These observations are performed from weather stations (land or sea) or ships, by trained observers following the  
119 WMO code tables (WMO, 2019). Each cloud level (high, WMO code table 0509; medium, WMO code table 0515; low, WMO  
120 code table 0513; see Table A.1) is separated in 9 different types describing in more detail the aspect and type of the observed  
121 clouds. The labels thus provide a high level of detail regarding the observed cloud scene from the surface. Naturally, the case of  
122 multilayer clouds poses a problem since the field of view and the visibility from the surface are limited, which is why we remove  
123 the potential multilayered cases from the training dataset to focus only on single-layer observed cloud scenes. It induces potential  
124 selection bias issues as some cloud types might more likely be observed in multilayered configurations. The relative amounts of  
125 each cloud type before and after the filtering and collocation process are displayed in Figure 2. Similarly, uncertainty is greater for  
126 medium and high clouds as their observation can be more challenging than for low clouds. Furthermore, the spatial distribution  
127 of the labels (Fig. 1, Fig. A.1) can be problematic as the marine observations are distributed mainly along ship routes. On the  
128 other hand, combining that with land observations provides a more complete representation of cloud types, especially for high  
129 level ones, all the while introducing the influence of orography. Other studies like Kuma et al. (2023) and Lenhardt et al. (2024a)  
130 have built estimates of cloud quantities based on these ground-based observation datasets, overcoming limitations pertaining to  
131 incomplete field of view and disparate spatial distribution.

132 For simplifying the analysis but also the training of the classification model, we group the 27 reported WMO cloud types into 4  
133 and 10 categories, similarly to Kuma et al. (2023). The first categorisation allows for broad classification by dividing the cloud  
134 species into high, medium, cumuliform and stratiform types. The second categorisation provides a more detailed classification  
135 while still limiting the subdivision of similar cloud types. This prevents a too pronounced unbalance in the cloud type labels  
136 while possibly removing some of the subjective biases and uncertainty stemming from the human observers. The detailed  
137 categories corresponding to the WMO codes are available in Table A.1 and shown in Figure 2.

138

## 139 2.2 Satellite retrievals

140

141 In addition to the surface observations, we use satellite retrievals from MODIS, in particular from the AQUA satellite. MODIS  
142 retrievals offer a vast amount of data at kilometre-scale resolution with daily overpasses. Each of the supplied granule file  
143 contains cloud microphysical and optical properties across a region with a span of around 2330 km x 2000 km. We make use of  
144 the available CUMULO dataset (Zantedeschi et al., 2019) since it allows access to preprocessed MODIS level 2 satellite data,  
145 with global coverage, and for two full years (2008 and 2016). Among the data variables available, we rely on two unified  
146 products (cf. Table 1) describing either cloud properties (MODIS06 level 2 cloud product, hereafter MYD06; Platnick et al.,  
147 2017) or the cloud cover (MODIS35 level 2 cloud flag mask, hereafter MYD35; Ackerman et al., 2017). The latter's main usage  
148 is to help screen for cloud scenes with a minimum cloud coverage.

149 The MYD06 data product incorporates miscellaneous properties pertaining to the cloud top (temperature, pressure, height)  
150 alongside some microphysical and optical properties (effective radius, water path, optical depth). As mentioned previously, our  
151 method builds upon level 2 data which are typically obtained from calibrated radiances through methods described in Platnick et  
152 al. (2017). More specifically, cloud top properties are retrieved using several radiance channels: harnessing the opacity of CO<sub>2</sub>,  
153 the CTP of high clouds is retrieved with wavelengths in the CO<sub>2</sub> absorption range, while infrared wavelengths combined with  
154 simulated brightness temperatures are used for lower and thicker clouds. The related CTH retrieval can thus suffer from regional  
155 biases as the brightness temperatures are based on vertical profiles from reanalysis using regional and monthly averaged lapse  
156 rate data along with surface temperature (Baum et al., 2012). The method introduced here can thus incorporate said biases from  
157 the input data into the learning process. The microphysical and optical properties of clouds - COT and cloud effective radius  
158 (CER) - are retrieved concurrently from multispectral reflectances, CTP values, surface types and cloud masks. Lastly, the CWP  
159 is also retrieved as part of the cloud optical properties algorithm detailed in Platnick et al. (2017). The additional input quantities  
160 needed to derive and retrieve the mentioned cloud properties (e.g. water vapour and ozone vertical profiles from reanalysis;  
161 Platnick et al., 2003; Baum et al., 2012) can result in subsequent uncertainties where only sparse observations like in remote  
162 marine areas are available for the data assimilation. Eventually, from the entirety of available MYD06 retrievals, we select three  
163 cloud properties in particular, namely the CTH, COT, and CWP.

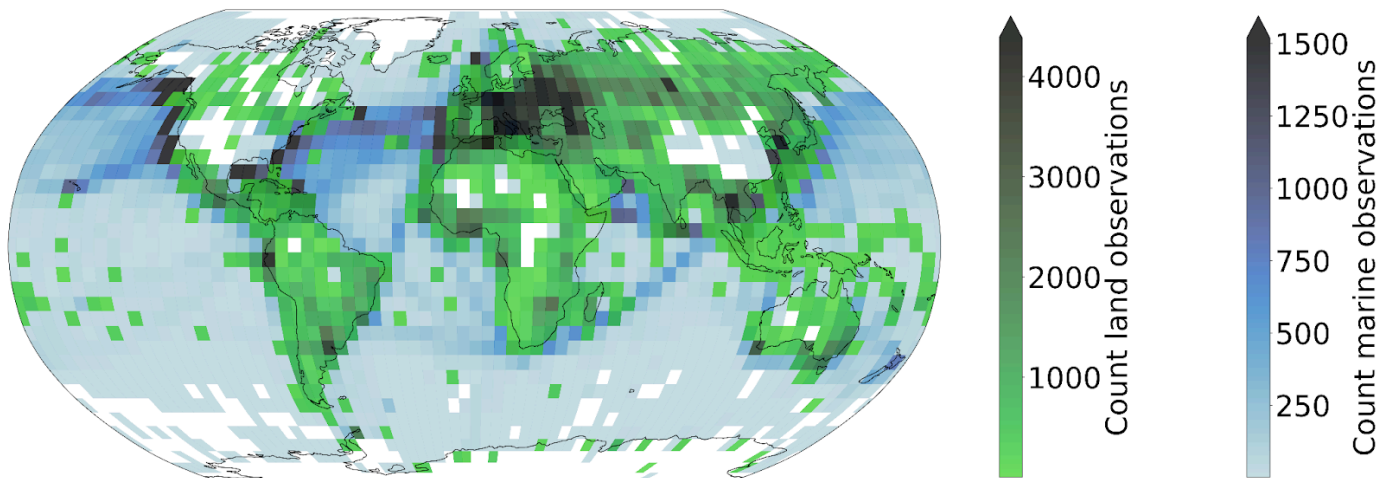
164 As a whole, the MYD06 product has the advantage that, building directly on cloud properties, we can design a classification  
165 model from which the relationship between cloud type and other cloud properties can then be examined. Relying on calibrated  
166 radiances which lie ahead in the retrieval process could offer a more neutral input but due to the large associated dimensionality,  
167 extracting information about clouds might become more challenging. Additionally, basing the method on commonly used cloud  
168 properties allows us to directly associate the results with other derived cloud classifications, making the comparison and  
169 understanding of the predictions more straightforward. Nevertheless, the biases introduced by using level 2 data in comparison to  
170 level 1 calibrated radiances and reflectances should be properly characterised and taken into account in the behaviour of the  
171 statistical model.

172 Alongside the collocated dataset, we build a collection of randomly sampled tiles out of the satellite retrievals from the year 2008.  
173 For each granule, a maximum of 20 tiles are sampled while ensuring the amount of missing data stays limited. This process leads  
174 to the compilation of more than 1.3M single tiles of cloud properties. These tiles are then randomly split temporally into training  
175 (70%), validation (10%) and test (20%) sets. This dataset is the basis for the self-supervised training procedure presented in the  
176 following section.

Data product	Description	Variables	Resolution	Usage
Global marine meteorological observations (Met Office, 2006)	Marine surface observations	Cloud type	Latitude/longitude coordinates 0.1° Hourly/daily observations	Labels
Land SYNOP reports (Met Office, 2008)	Land surface observations	Cloud type	Latitude/longitude coordinates 0.1° Hourly/daily observations	Labels
MODIS Atmosphere L2 Cloud Product (MYD06) (Platnick, 2017)	Cloud-top properties, cloud optical and microphysical properties	Cloud top height, CTH (m) Cloud optical thickness, COT (a.u.) Cloud water path, CWP (g.m <sup>-2</sup> )	1-km pixel resolution Daily overpass	Input features
MODIS Atmosphere L2 Cloud Mask Product (MYD35) (Ackerman, 2017)	Cloud pixel flag	Cloud mask	1-km resolution Daily overpass	Used for cloud scene filtering

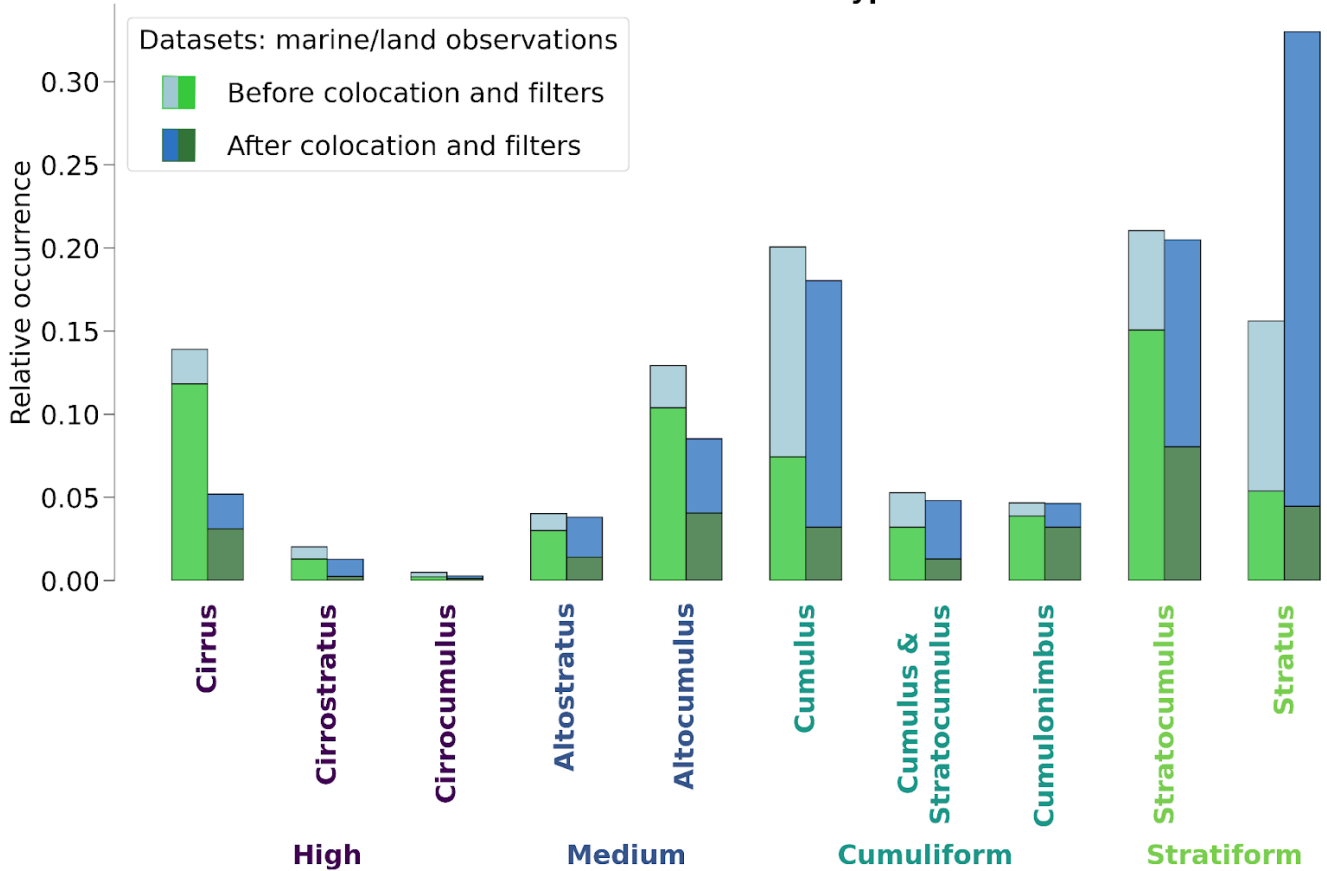
178 **Table 1 : Datasets description. The surface observations are provided by a worldwide station network available from the**  
 179 **UK MetOffice (Met Office, 2006; Met Office, 2008; see section 2.1). The MODIS data are derived from the collection 6.1**  
 180 **of the datasets (Ackerman, 2017; Platnick et al., 2017; see section 2.2).**  
 181

### Marine (2008, 2016) & land (2016) cloud type observations count



182 **Figure 1: Spatial distribution of cloud type observations for marine (years 2008 and 2016; Met Office, 2006) and land**  
 183 **(year 2016; Met Office, 2008). The corresponding spatial distributions of cloud type observations are included in Figures**  
 184 **A.1 and A.2, for before and after the colocation process, respectively.**  
 185  
 186

## Relative occurrences of cloud type observations



187

188 **Figure 2: Relative occurrences of cloud types before and after the collocation and filtering process, indicated for both the**  
 189 **marine (blue; Met Office, 2006) and land (green; Met Office, 2008) observational datasets. The x axis corresponds to the**  
 190 **cloud types in the case of 4 and 10 categories. The corresponding numbers of collocated samples for each cloud type are**  
 191 **detailed in Table A.1.**

192

193

### 194 3 Method

195

#### 196 3.1 Method outline

197

198 Relying on computer vision models and their large number of trainable parameters usually requires adapting the training strategy,  
 199 particularly when the training dataset is of modest size. In the presented study, the amount of labels available is greatly reduced  
 200 during the collocation process (see Table A.1 for the number of samples per cloud type) but still contains useful and exploitable  
 201 information about the observed cloud types. We thus introduce a self-supervised learning process which allows us to draw on the  
 202 larger amount of satellite data available before addressing the more complex task of cloud classification. The larger purpose of  
 203 this methodology is to be able to classify clouds on a global scale, outside of the areas where surface observations were made and  
 204 outside of the typical coverage of human observation stations.

205 For the self-supervised task, we train two models to reconstruct 3D data cubes of cloud properties. The first model, which is used  
 206 as a baseline, is a CNN backbone we previously presented in Lenhardt et al. (2024a) to handle satellite retrievals of cloud  
 207 properties for cloud base height prediction. The second model we develop in this study is based on vision transformers  
 208 (Dosovitskiy et al., 2020), a recent type of model compared to the more typical CNNs for computer vision applications. The  
 209 spatial pattern of the cloud properties and their scale provide information about clouds, which can be leveraged to classify them  
 210 for example into more stratiform and more cumuliform types. During the training phase of these models, the samples are images  
 211 of size 128x128 pixels consisting of three different cloud properties: CTH, COT and CWP. We ensure that the models learn to  
 212 distinguish cloud patterns and not to recognise specific geographical locations by extracting samples randomly across global

213 satellite retrievals from the year 2008, without adding information about their location. In a second step, a classification model is  
 214 trained on the collocated samples of cloud properties and surface observations. As mentioned in section 2.1, the number of types  
 215 reported in the observations for clouds is reduced to either 4 or 10 classes (Kuma et al. 2023). The training process follows a  
 216 supervised learning framework, where the classification model outputs a single cloud type (among the 4 or 10 cloud types) for  
 217 the whole extent of the input cloud scene of size 128x128 pixels. The benefit of the presented method using either a CNN or a  
 218 vision transformer, which are models incorporating a certain level of spatial awareness, is that it is consistent with the cloud type  
 219 identified by the human observer. Furthermore, in comparison to conventional methods like the ISCCP, the method benefits from  
 220 a potential ability to distinguish cloud types without using predefined thresholds.

221

### 222 3.2 Vision transformer

223

224 Vision transformers were introduced by Dosovitskiy et al. (2020), building on the transformer architecture previously presented  
 225 in Vaswani et al. (2017) which was mainly applied to natural language processing (NLP) tasks. The adaptation to images was  
 226 made by splitting images into patches of a certain size, 16 pixels in the case of the seminal paper, and providing the sequence of  
 227 embeddings of these patches to a transformer. The patches from the images are then treated as words would be in a NLP  
 228 application. The transformer can then be trained in a supervised fashion to classify the input images. They have been shown to  
 229 perform at the same level or even outperform classical computer vision models like ResNets on tasks like classification (e.g. see  
 230 Section 4 of Dosovitskiy et al., 2020). However, as mentioned in section 3.1, this type of model, alongside CNNs, is data hungry  
 231 and requires a large number of labelled samples to be trained from scratch in a supervised fashion. In this setting, self-supervised  
 232 pretraining can lead to highly performant models while not requiring a larger training dataset. We train a vision transformer  
 233 following the self-supervised pretraining methodology presented in Atito et al. (2023), named Self-supervised vision  
 234 Transformer (SiT). This methodology allows to train vision transformers in a self-supervised fashion building on the concept of  
 235 Group Masked Model Learning (GMML), additionally using the same autoencoder framework as with traditional CNNs like the  
 236 commonly used U-Net (Ronneberger et al., 2015) or our baseline model from Lenhardt et al. (2024a). The SiT architecture used  
 237 in this study is adapted from the seminal vision transformer architecture (Dosovitskiy et al., 2020) by setting the latent dimension  
 238 to 256, similarly to the CNN architecture introduced in Lenhardt et al. (2024a).

239 One strength of the transformer architecture is the possibility to easily include several simultaneous learning tasks. We can use  
 240 this ability to our advantage and incorporate two objectives for the self-supervised training process: input reconstruction  
 241 following GMML and contrastive learning. The input reconstruction is achieved by adapting the transformer into an autoencoder  
 242 architecture. Like with traditional CNN autoencoders, the task is for the model to reconstruct the provided input. We benefit  
 243 further from another advantage of vision transformers as they showcase a reduced complexity compared to CNNs since they rely  
 244 to a much lesser degree on convolution operations. The methodology of Atito et al. (2023) additionally uses recent results in  
 245 GMML to further help in the self-supervised learning task. The framework of GMML is integrated in the reconstruction task by  
 246 replacing random parts of the input image with noise. The overarching goal of this image modification is to train the model to  
 247 learn semantic representations of the input data, allowing reconstruction of masked areas only with knowledge of some other  
 248 patches in the input image. The objective for this reconstruction task hence takes the form of the l1-loss, a commonly used metric  
 249 (Zhao et al., 2016) between the standardised input and the reconstructed output:

$$250 \quad L_r = \frac{1}{N} \sum_{i=1}^N \left\| x_i - D_{\theta}(E_{\theta}(x_i^c)) \right\| \quad (1)$$

251 where  $x_i$  is the input standardised image,  $x_i^c$  is the corrupted standardised image,  $\|\cdot\|$  is the l1-loss,  $N$  is the batch size,  $D_{\theta}$  and  $E_{\theta}$   
 252 are namely the decoder and encoder parts of the model with  $\theta$  designating their learnable parameters.

253 The second learning task included in the training process is based on contrastive learning. Since the presented self-supervised  
 254 process does not rely on labels for the training data contrary to the vision transformer from Dosovitskiy et al. (2020), the learning  
 255 task needs to be adapted. To this extent, several geometric transformations and perturbations are applied to the training samples  
 256 for which the transformer should produce similar outputs. The synthetic pairs can then be used as matching pairs and a metric  
 257 can be built measuring their similarity. The contrastive task is thus training the model to minimise the distance between matching  
 258 pairs of sample and corresponding augmented sample, while maximising the distance between different samples in the batch.  
 259 Atito et al. (2023) propose to use as a contrastive metric the arithmetic mean over the matching pairs in the batch of the cross  
 260 entropy of their normalised similarities:

$$261 \quad L_c = -\frac{1}{N} \sum_{i=1}^N \log l_c(x_i, x_i^a, E_{\theta'}, D_{\theta'}) \quad (2)$$

262 where the similarity metric between a sample  $x_i$  and its augmented version  $x_i^a$  is the normalised temperature-scaled softmax  
263 similarity (Chen et al., 2020). The actual process of the contrastive learning further requires the use of a momentum encoder to  
264 generate different versions for the pairs of samples and their corresponding augmented samples.

265 The integral self-supervised training process consists in a combination of the two previously presented learning tasks. For each  
266 batch of samples, we create augmented versions of the samples which together constitute matching pairs. GMMML corruptions are  
267 applied to both samples and the model is subsequently trained to reconstruct the original inputs from these corrupted samples. At  
268 the same time, the similarity between matching pairs of samples is maximised. The complete loss function thus takes the form of:

$$269 \quad L = \alpha \times L_r + L_c \quad (3)$$

270 where  $\alpha$  is a scaling factor between the two tasks. We follow the recommendation of Atito et al. (2023) to set  $\alpha = 5$  in the case  
271 of small-scale datasets so that the vision transformer can learn enough of the local inductive bias.

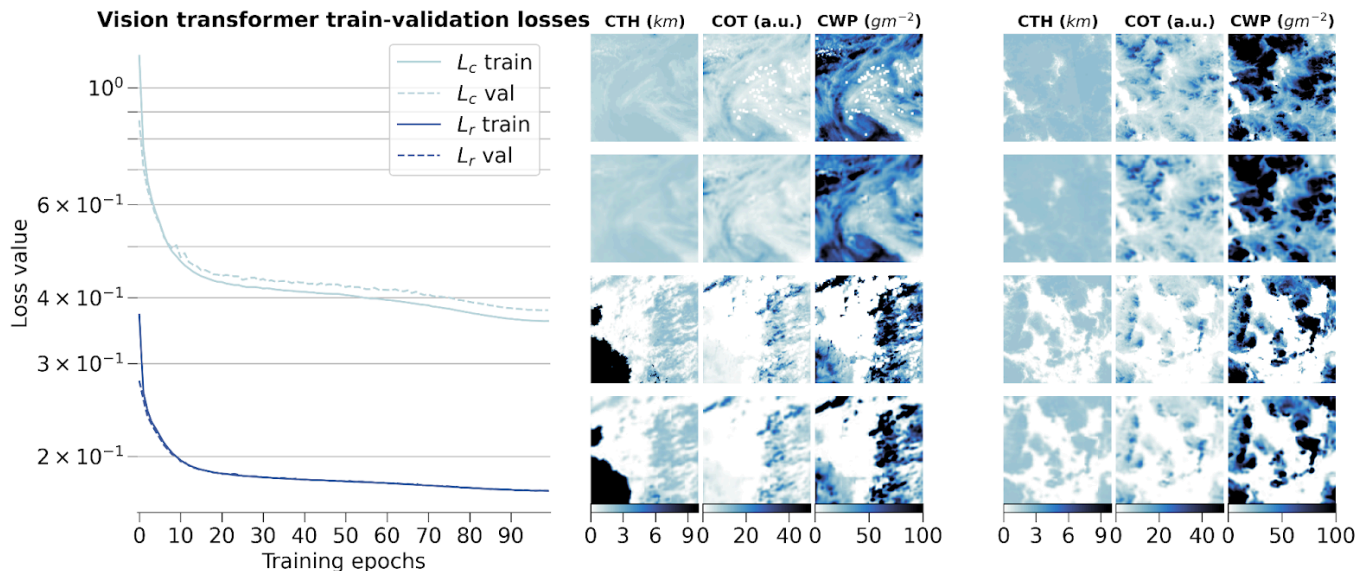
272 We set out to examine in further detail the ability of the vision transformer and of the self-supervised training methodology by  
273 evaluating how different configurations of the input data and of the model architecture can impact the quality of the learnt  
274 representations and the transfer to cloud classification. We mainly discuss in this section the reconstruction skill of the vision  
275 transformer and the potential influence of contrastive learning. The transfer to the cloud classification task will be described in  
276 the following section where fine-tuning to the downstream task or the use of external models are surveyed. Since training vision  
277 transformers requires large computing resources, we limit ourselves for all the pretraining processes to only 10% of the initial  
278 dataset mentioned in section 2.2, similar to what is done in Atito et al. (2023) regarding ablation studies.

279 To begin with, we investigate how the two architectures of vision transformers fare during the self-supervised training and how  
280 the scaling factor between the contrastive loss and the reconstruction loss impacts the learning process. The two architectures  
281 tested correspond to the small variant of the vision transformer from Atito et al. (2023) and the base variant from Dosovitskiy et  
282 al. (2020). To offer an overview on each model’s complexity, their respective numbers of parameters are 21M and 86M, the main  
283 difference originating from the number of heads in the self-attention layers, the size of the multi-layer perceptron (MLP) and the  
284 hidden dimension. We additionally investigate the self-supervised training process by using pre-trained weights made available  
285 in Atito et al. (2023) for which the pretraining was done on a computer vision task, the ImageNet-1K dataset (Deng et al., 2009).  
286 However, the pretrained weights of the ImageNet-1K dataset are only made available for the small variant of the vision  
287 transformer. An additional comparison is done with a model trained only on the colocated dataset using the small variant. The  
288 contrastive and reconstruction losses for the different model setups are detailed in Figure B.1. Firstly, we notice that the model  
289 trained solely on the colocated dataset would need more epochs to reach similar performance compared to all the other setups. As  
290 the colocated dataset contains two orders of magnitude less samples than the training dataset, the model has also seen much less  
291 data after 10 epochs, hindering the training process most notably for the contrastive loss. Even after further training the model on  
292 the colocated dataset for 150 epochs, it is struggling to match the other models trained on the complete training dataset with best  
293 contrastive and reconstruction losses of 0.95 and 0.23, respectively. On the other hand, the other setups reach similar  
294 performance in both contrastive and reconstruction losses after 10 epochs. The model with pretrained weights displays better  
295 performance right from the start of the training process but improves only marginally thereafter. This could be explained by the  
296 fact that using the pretrained weights allows the model to capture already well the structure and patterns of the clouds in the  
297 remote sensing data even though their modality is different from the one seen in the ImageNet-1K dataset. It thus shows the  
298 strength of transfer learning in computer vision tasks. Nevertheless, we can observe that for the pretrained model both the  
299 contrastive and reconstruction losses are reaching a plateau after only a few epochs while the other model setups display a  
300 negative gradient indicating further learning capabilities. Focusing on the different variants trained with scaling factors of 1 or 5,  
301 we notice that the choice of a larger scaling factor leads to better reconstruction skill while losing almost no performance with  
302 respect to the contrastive loss.

303 Eventually, we decide to use as model the small variant of the vision transformer with a scaling factor  $\alpha$  of 5, as it showcases  
304 good performance in both tasks during the training while having a number of parameters four times smaller than the base variant.  
305 Furthermore, the self-supervised training task on the large unlabelled dataset allows the model to have plenty of data to learn  
306 from, the pre-trained model weights giving only marginal gain for a few epochs at the start. The small variant of the vision  
307 transformer was shown to perform very well on a large variety of tasks as per the results from Atito et al. (2023). The results  
308 across the training, validation and test datasets are shown in Figure 3 for the training process and some examples of reconstructed  
309 samples belonging to all three splits, while Figure 4 highlights the spatial distribution of the reconstruction error per channel and  
310 across splits.

311 Ultimately, we can compare the skill of the vision transformer to that of the baseline CNN autoencoder from Lenhardt et al.  
312 (2024a). The CNN autoencoder was trained using as reconstruction error the mean squared error (MSE) on similar MODIS data

313 but only with MODIS granules over the ocean. It was shown to perform similarly with a slightly higher error over land when  
 314 evaluated over a global dataset. The vision transformer model outperforms the CNN autoencoder on all metrics (MSE and  
 315 l1-loss) across all data splits (training, validation and test), displaying consistently across data splits on average an MSE of 0.15  
 316 and a l1-loss of 0.12 compared to 0.3 for both metrics for the CNN. Examples of reconstructed samples additionally show how  
 317 the l1-loss helps produce sharper edges in the reconstruction, a well-known issue with the application of MSE as target metric in  
 318 computer vision (Zhao et al., 2016). The contribution to the error comes mostly from the COT channel for both models and the  
 319 error is concentrated in areas of higher variability for the respective channels. The metrics values are summarised in Table B.1.  
 320 The spatial generalisation skill, alongside the sensitivity to the tile size and the impact of data augmentation on the performance  
 321 on the cloud classification task are analysed in the following section.

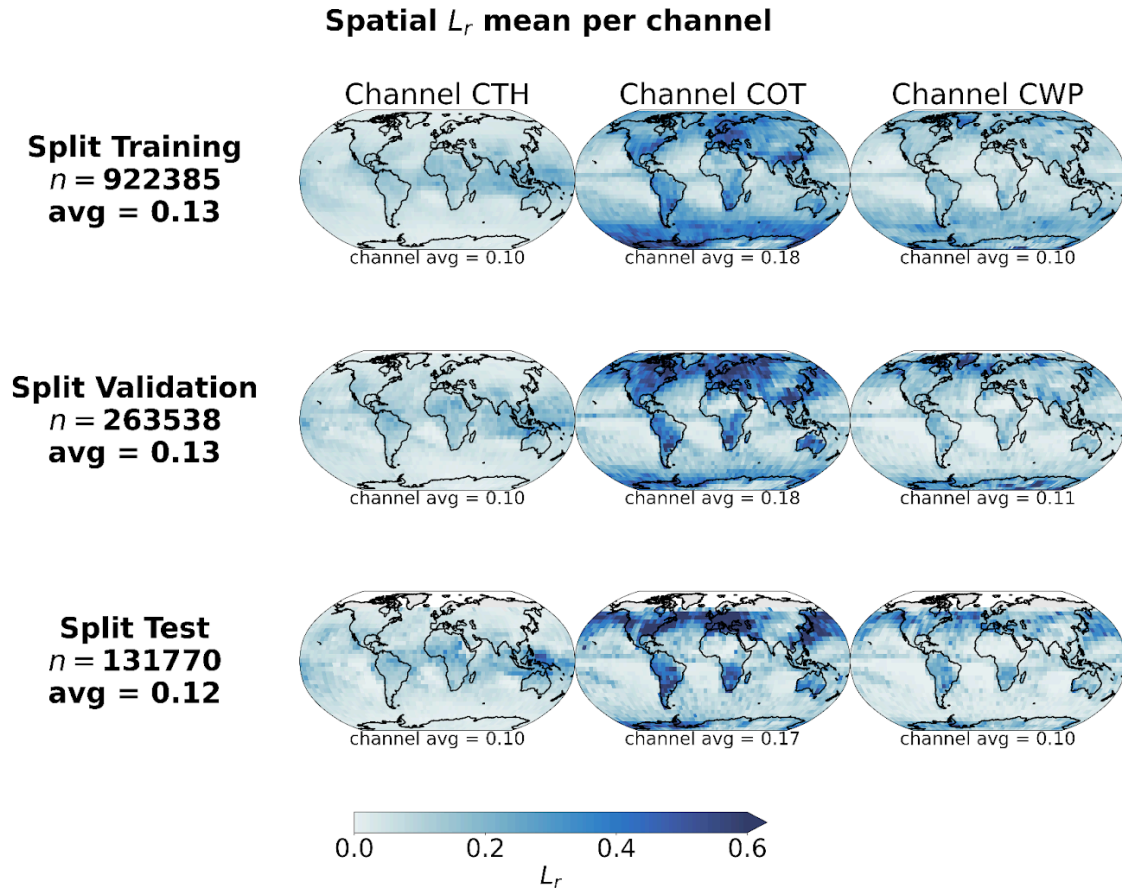


322  
 323 **Figure 3: (left) Training and validation losses during model optimization for the small variant of the vision transformer**  
 324 **on the global training dataset. (right) Examples of tiles (first and third rows) with the corresponding reconstructions**  
 325 **(second and fourth rows) for the different cloud property channels.**

326  
 327 **3.3 Cloud type classification**  
 328

329 The next task at hand is the cloud type classification, building on the colocated samples of satellite retrievals and surface  
 330 observations. For the two years of MODIS AQUA data available, out of 104 823 colocated samples we retain only 11 094 for our  
 331 training and testing datasets after filtering, among others, for missing data - typically 50% of the samples are discarded, mainly  
 332 when the colocated observation lies on the edges of the satellite granule - and single layer cloud observations as reported by the  
 333 observer - around 60% of the previously filtered samples are kept. The cloud type observations are then regrouped into 4 or 10  
 334 types as mentioned previously. A main caveat arising from collocating these two data sources is the potential mismatch between  
 335 the actual clouds jointly depicted. Contrarily to methods like Zantedeschi et al. (2019) which rely on joint retrievals of cloud  
 336 properties and cloud type or Kuma et al. (2023) which aggregates observations at daily time scales, the presented colocated  
 337 dataset leaves room for misaligned surface observations and satellite retrievals. The rest of the study will focus on these  
 338 categories as targets. From the latent space representations produced by the vision transformer or the CNN autoencoder, we build  
 339 a classification model either by attaching a classification head to the encoder network or by using a simpler classification model  
 340 like a random forest (RF; Breiman, 2001). To investigate the performance of the classification models on the two classification  
 341 tasks at hand (4 and 10 cloud types), we use different metrics tailored to unbalanced classification setups as the cloud types are  
 342 not equally represented (see Figure 2 and Table A.1). A first method to assign similar weight to all classes regardless of the class'  
 343 cardinality is to use macro-averaged metrics. In this framework, the metric of interest is averaged over the samples of each class  
 344 separately before being averaged over the classes. This leads to a higher weight for minority classes for which the model might  
 345 perform differently, usually worse, compared to the majority classes providing different information over traditional averaging  
 346 strategies (micro-averaged for example) where the result will be dominated by the samples from the majority classes. We report  
 347 several metrics adapted to an unbalanced setting: the index balanced accuracy (IBA; Garcia et al., 2012) of the geometric mean,  
 348 the macro-averaged accuracy and the macro-averaged f1-score.

349 For the classification model we investigate two alternatives: a RF classification model (implementation from *Scikit-learn*  
 350 package, Pedregosa et al., 2011) and a MLP classification head (Hinton, 1989; implemented in *PyTorch*, Paszke et al., 2019).  
 351 However, a wider diversity of classification models could be implemented based on the backbone provided by the vision  
 352 transformer. The RF model provides simplicity in the implementation and the training process, while the MLP is the typical  
 353 architecture used for the downstream task following a network like a vision transformer or a CNN. The RF model has 10 or 25  
 354 trees, for the cases of 4 and 10 cloud types respectively, with a maximum depth of 5. Basic hyper-parameter optimization showed  
 355 that with the reduced amount of samples and the limited variety in cloud scenes for some categories (even more with balanced  
 356 classes, see section 3.3.3), models displaying limited complexity avoided overfitting and generalised better on unseen data. The  
 357 MLP consists of two fully-connected layers (hidden dimension 4096) with a Gaussian Error Linear Unit (Hendrycks & Gimpel,  
 358 2016) in between and is trained using the cross-entropy loss. The sensitivity studies and experiments are done only using RF  
 359 models but the evaluation in the subsequent section will be done on both types of classification methods. Various sensitivities  
 360 could be explored in the presented setting but we here focus on the potential benefit of the spatial context, the ability to  
 361 generalise spatially to unseen locations and the impact of balancing the labelled dataset.  
 362



363  
 364 **Figure 4: Spatial distributions of mean channel reconstruction errors for CTH, COT and CWP, aggregated on a  $5^\circ$**   
 365 **regular grid for the training, validation and test datasets.**  
 366

### 367 3.3.1 Spatial context and tile size

368 We look at the influence of the input size by training vision transformers (small variant) on different sizes of inputs namely  
 369 128x128, 64x64, 32x32 and 16x16. We do not consider larger tile sizes as the cloud scene might then be less representative of the  
 370 surface observation, especially since we only consider samples with single labels, and as the assumption that the observed cloud  
 371 type occurs on such scales would likely not hold. The losses relative to the vision transformer models trained on the different  
 372 input tile sizes are detailed in Figure B.2. Since these models were trained on a reduced dataset as mentioned previously, their  
 373 skill cannot be directly compared to the one displayed in Figure 3. While the contrastive losses are similar across input tile sizes,  
 374 the reconstruction losses differ. Since we kept the ratio between the patch size and the tile size constant when training the  
 375 different models, the difference in reconstruction skill could be attributed to the dimensionality of each patch being much  
 376 smaller, for example for a tile of size 16x16 a patch will be 2x2. The reconstruction head being a fairly shallow CNN, the

377 reconstruction of the spatial patterns inside the patches showcases better skill for smaller input patches after a few epochs, while  
378 for larger patch sizes - and thus tile sizes - a longer training process would be needed as to improve the truthfulness of the  
379 reconstruction to the input. Examples of reconstructions depending on the input tile size are included in Figure B.3 and visually  
380 display how a larger field of view can help capture the larger cloud organisation or even individual sparse clouds. To further  
381 evaluate the potential benefit of the spatial context for the downstream classification task, we consider as an alternate input the  
382 flattened cloud properties of a 9x9 tile centred on the observation location. This yields an input of similar dimensionality  
383 compared to the latent space representation of both the CNN and the vision transformer (3 channels x 9 x 9 = 243). We then train  
384 the same RF classification model on each of the latent representations derived from the trained vision transformers and on the  
385 flattened cloud properties. From the classification metrics, we observe that the smaller the tile size the more prone the model is to  
386 overfitting towards the majority classes (high and stratiform cloud types in the case of 4 types) leading to a decreased  
387 performance on the validation set. For instance, choosing an input tile size of 16x16 results in a decrease of 20% across metrics  
388 from the training to the validation set (compared to around 10-15% across metrics for the larger input tile sizes), and leads to  
389 metrics on the validation set more than 10% lower than with larger input tile sizes. The predictions made using larger spatial  
390 context (tile size greater than 16x16) outperform the method with 9x9 flattened tile inputs across all considered metrics on the  
391 validation set. With the input tile size 16x16, the reduced spatial context seems to be limiting for the performance but another  
392 explanation could be a complex latent space compared to the input dimensionality. Overall, even with the vision transformer  
393 backbones being trained only partially, the wider input tile size provides better classification skill and generalisation to unseen  
394 data. In the rest of the study and experiments, if not mentioned specifically, the input tile size is chosen to be 128x128.

395

### 396 3.3.2 Spatial generalisation

397 To investigate the spatial generalisation skill of the cloud classification method, we split our colocated dataset into samples  
398 located in the Northern or Southern hemispheres. Two vision transformer models are additionally trained on samples from only  
399 the respective hemisphere and tested on the other one. The losses relative to the training and testing of both hemispherical  
400 models are included in Figure B.4. Both hemispherical models display similar performance both on the training and testing  
401 datasets, showing that even for a reduced number of training samples, epochs and spatial coverage the vision transformer  
402 architecture generalises well to unseen data. Building on the two trained vision transformers, we set out to evaluate the skill on  
403 the classification tasks. Splitting the labels between the two hemispheres yields 9246 samples for the Northern hemisphere and  
404 1848 samples for the Southern hemisphere. Investigating the different classification metrics for training and testing on both  
405 hemispheres, it is clear that the classification model trained on the Southern hemisphere struggles to generalise from such a low  
406 number of labelled samples and probably overfits since the performance is worsened on the Northern hemisphere samples  
407 (decrease of almost 50% across metrics from the training to the testing set). The classification model trained on the Northern  
408 hemisphere generalises well in the case of the 4 cloud types with consistent metric values between hemispheres (marginal  
409 decrease of around 15% across metrics from the training to the testing set). Overall, the model trained on samples from the  
410 Northern hemisphere and for both cases of number of cloud types, the performance on the Southern hemisphere is similar to  
411 models with larger tile sizes from the previous section, showing consistency across experiments even with limited datasets for  
412 the training of the vision transformer.

413

### 414 3.3.3 Balanced training dataset

415 Balancing the number of samples among classes in the input dataset can be a way to leverage enough information from the  
416 underrepresented classes. We compare here the performance skill of two classification models trained on the colocated dataset or  
417 on a balanced equivalent. To this extent, we use a sampler implementation from the *imbalanced-learn* package (Lemaitre et al.,  
418 2017), namely the Synthetic Minority Oversampling Technique (SMOTE; Chawla et al., 2002) to oversample the minority  
419 classes. Doing so leads to improved classification skill with consistent increases across metrics on the validation set of 3-7% and  
420 15-35% for the cases of 4 or 10 cloud types, respectively. The oversampling impacts mostly the cloud types from the high and  
421 medium classes, and from the cirrocumulus and cirrostratus classes, in the case of 4 cloud types and 10 cloud types, respectively  
422 (see Table A.1). The methods evaluated in the following section will thus include the same over-sampling strategy to overcome  
423 the representation of the minority classes and improve the performance on the classification task.

## 424 4 Evaluation

425

### 426 4.1 Classification evaluation

427

428 In the following section, we detail the classification performance on the test set of the previously mentioned models. Two  
429 baseline models are included, namely a classification model built on the CNN autoencoder from Lenhardt et al. (2024a) and a RF  
430 model built on the flattened 9x9 input tiles as described in section 3.3.1. The method developed in this study is represented by  
431 two models using the aforementioned vision transformer model (see section 3.2) as backbone complemented by either a RF  
432 classifier or a MLP (see section 3.3). In the rest of the study, we denote the trained vision transformer model followed by the  
433 classification model as CloudViT (Cloud Vision Transformer) in its two classification variants (RF or MLP). The classification  
434 metrics on the test dataset for these four models are summarised in Table 2 for the case of the 4 cloud types and in Table C.1 for  
435 the 10 cloud types. Since the number of samples is very limited, the performance of the models cannot be only considered as is  
436 but is further evaluated in the subsequent sections through distributions of cloud properties and spatial occurrence distributions.  
437 The CloudViT/RF method performs the best across all of the three metrics included, despite showing still limited performance  
438 overall. Firstly, the macro-averaged multi-class accuracy does not differ by a large margin between the different methods, but the  
439 class-wise accuracies reveal several limitations. The baseline 9x9 RF model largely overfits towards the high and stratiform types  
440 (train and test class accuracies of 0.84/0.81 and 0.63/0.62, respectively), performing poorly on the medium and cumuliform types  
441 (train and test class accuracies of 0.31/0.21 and 0.19/0.15, respectively). The CloudViT/MLP model is biased towards stratiform  
442 clouds (train and test class accuracy of 0.79/0.79) while struggling to identify the other three types (train and test accuracies all  
443 falling between 0.10 and 0.40). The baseline CNN/RF and the CloudViT/RF models are performing quite similarly both on  
444 aggregated and class-wise metrics. However, the CloudViT/RF model showcases improved performance on the stratiform class  
445 (increase of 0.13 in the class accuracy both on the train and test datasets) and only a marginal decrease (0.03) on the class  
446 accuracies for medium and cumuliform clouds. The performance on the high clouds is similar with slightly higher accuracies for  
447 the CloudViT/RF model. Other metrics like the IBA of the geometric mean and the F1-score further emphasise that the  
448 CloudViT/RF model outperforms the other methods while addressing the imbalance training data to generalise with satisfactory  
449 skill on the unseen test dataset. Nevertheless, the performance detailed here across classes shows apparent limitations as scores  
450 are not ideal. An obvious hurdle of the learning process resides in the overall limited number of samples and the noise present in  
451 particular for cloud types with minimal numbers of samples. Building a dataset with more labels would improve the  
452 classification performance by allowing the classification to more easily converge towards each cloud type's mean state arising  
453 from a larger number of samples. The simplicity of the classification models chosen here represents a constraint that could be  
454 lifted if more training samples were available as overfitting and balance would then represent lesser issues. Furthermore, the  
455 patterns in the class accuracies can be traced back to shortcomings in the observational dataset. Having only considered  
456 single-layer cloud scenes in the colocated dataset, the high clouds are well predicted in accordance with the observations as a  
457 surface observer would identify with certainty this type of cloud if no other lower cloud is blocking the field of view from the  
458 surface. Stratiform clouds could be more challenging for the observers as they typically display high cloud fraction and high  
459 optical thickness, limiting the ability of the surface observer to quantify with certainty the amount of clouds in other levels.  
460 However, such characteristics can be well captured by computer vision models which build on patterns in the three-dimensional  
461 input data which in particular the baseline 9x9 RF model lacks. This difference between models is in particular apparent for the  
462 cumuliform class which is mostly composed of observations of cumulus. A cloud scene relative to a cumulus observation will  
463 most likely display a lower cloud fraction as the individual clouds are sparsely distributed, extracting only the very near points  
464 around the observation might then be too reductive and limit the accuracy of the classification model. It is confirmed by the  
465 accuracy on this cloud type for which the baseline 9x9 RF model is largely outperformed by all three other models both on  
466 training and test datasets (class accuracy increases between 150% up to 260% on the test dataset). Overall, the classification  
467 model shows fair performance that could be probably improved by widening the scope of the cumbersome colocation process  
468 which requires large amounts of remote sensing data, and by accordingly refining the RF or MLP architectures presented here.  
469 Nonetheless, the evaluation of the predictions in the following section provides insights and reveals relevant features in the  
470 predicted cloud types.

Method	Multi-class accuracy *	IBA geometric mean	F1-score *
Baseline 9x9 RF	0.45	0.32	0.35
Baseline CNN/RF	0.45	0.32	0.40
CloudViT/MLP	0.40	0.32	0.42
<b>CloudViT/RF</b>	<b>0.46</b>	<b>0.36</b>	<b>0.43</b>
CloudViT/RF (train)	0.55	0.41	0.49

472 **Table 2: Classification metrics on the test set in the case of 4 cloud types. The metrics noted with a \* are referring to their**  
473 **macro-averaged estimate. The method on which the rest of the study is based is highlighted in bold. The baseline**  
474 **CNN/RF refers to the CNN backbone introduced in Lenhardt et al. (2024a).**

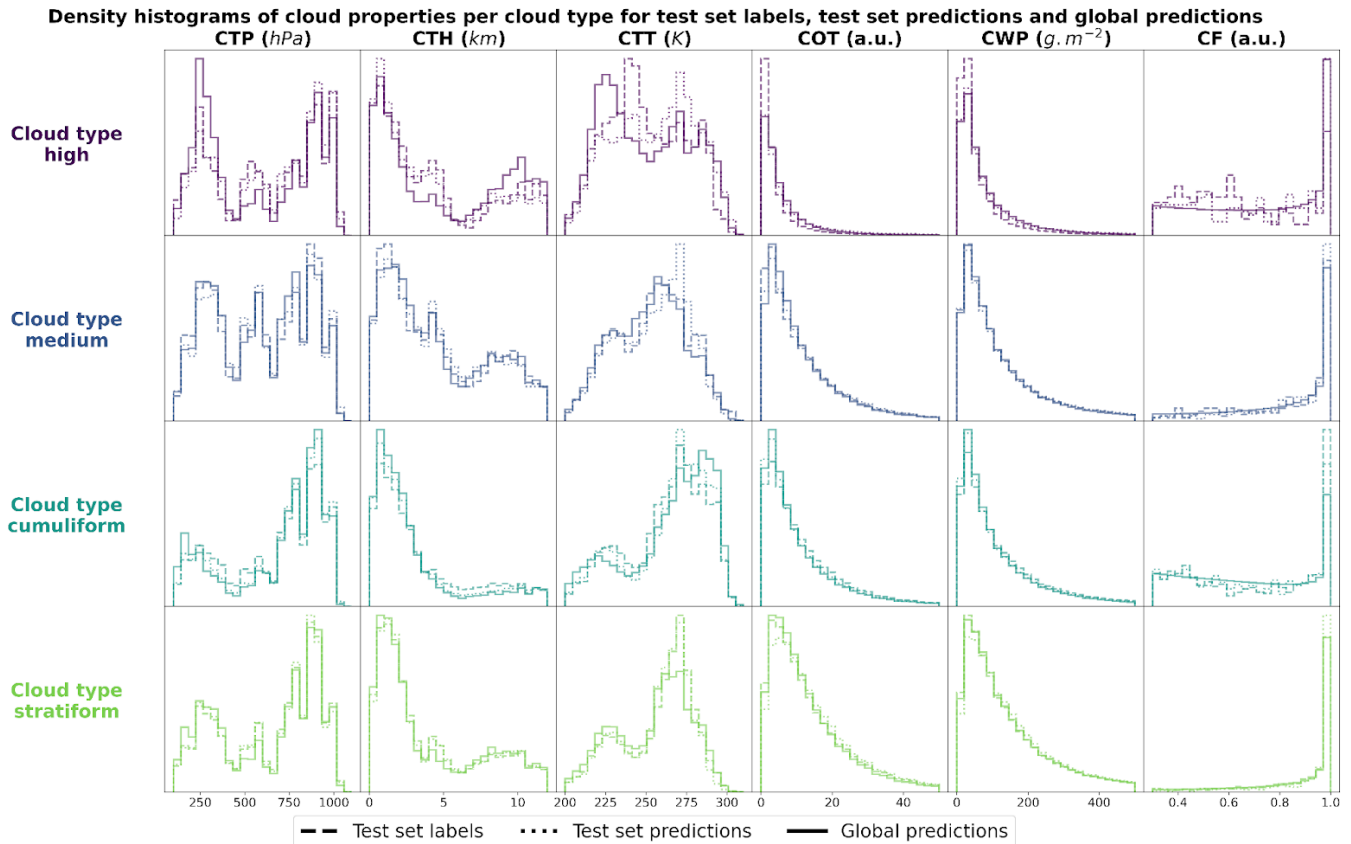
475

## 476 4.2 Histograms of cloud properties

477

478 In order to evaluate the physical soundness of the predictions made by the CloudViT model, we investigate the distribution of  
479 several cloud properties with respect to the observed and predicted cloud types. In Figure 5, we summarise the distribution of  
480 cloud top pressure (CTP), cloud top height (CTH), cloud top temperature (CTT), cloud optical thickness (COT), cloud water path  
481 (CWP) and cloud fraction (CF) for the 4 cloud types (high, medium, cumuliform, stratiform) and for three different datasets: the  
482 test set labels, the test set predictions and the dataset of global predictions. The latter is built on global MODIS AQUA granules  
483 for the year 2016 - the year is chosen to avoid any overlap with cloud scenes seen during the training of the vision transformer on  
484 data from 2008 - from which we regularly sample tiles in order to build a more comprehensive and global dataset of cloud types  
485 to further evaluate the method. The spatial distribution of cloud types for this dataset is highlighted in the following section and  
486 the global dataset is made available at Lenhardt et al. (2024b). The histograms are built by reporting the respective cloud  
487 properties for all the cloudy pixels in each sampled tile from the dataset apart from the cloud fraction which is computed for the  
488 whole tile from the cloud mask. As a consequence, unless the whole cloud field is composed of only a single cloud type, the  
489 histograms will cover a large range of cloud properties due to multi-layer clouds or multi cloud types scenes (e.g. convective  
490 cells with associated anvils or cumulus/stratocumulus transitions). Even though the trained model only produces fair evaluation  
491 metrics on the test set, the histograms of cloud properties display interesting features consistent with expected characteristics of  
492 the different cloud types. On Figure 5, the histograms pertaining to the test set labels and predictions have distributions close to  
493 identical across cloud types showing a good agreement in the clouds depicted in both datasets while the global dataset histograms  
494 provides a less noisy overview of the distribution of the cloud properties per cloud type. The high clouds are characterised by  
495 low cloud water path and optical thickness, along with colder and higher cloud tops as well as more frequent cloud fractions  
496 smaller than one. All of these aspects are emphasised in the global predictions compared to the limited test set samples, showing  
497 the CloudViT model manages to extract the representative characteristics of the cloud type from the labels. The cumuliform  
498 category encompasses mostly low warm clouds with reduced cloud fractions and moderate cloud water path and optical  
499 thickness. Inside this class, the higher and colder cloud tops are concentrated in the cumulonimbus class, along with larger cloud  
500 water path and cloud optical thickness (see Fig. C.1). The stratiform class includes thick cloud fields with high cloud water path  
501 and almost full spatial coverage of the cloud scenes (cloud fraction close to 1 in most cases). A fraction of the clouds in this class  
502 are slightly higher and colder and correspond to stratus/nimbostratus clouds which can also be seen in Figure C.1. The  
503 distributions for medium clouds showcase similarities with several other types and are best evaluated in combination with their  
504 spatial distribution (see Section 5). Examining in more detail the refined cloud types with the 10 cloud types (see Fig. C.1)  
505 reveals slight differences inside broader cloud types. For example, the distinction between the three high cloud types (cirrus,  
506 cirrostratus and cirrocumulus) appears through separations in cloud fraction, cloud optical thickness and cloud water path which  
507 were not obvious from the limited amount of labelled samples. The differences between the three high cloud types further  
508 manifest in distributions of cloud top quantities for which cirrus and cirrostratus display potential multilayered cloud scenes with  
509 a combination of low/warm and high/cold cloud tops. Overall, the CloudViT model seems to generalise well from a few samples  
510 (only around 10 for the cirrocumulus class) by exhibiting in parts physical consistency inside predicted types. Due to the large  
511 cloud scenes considered as input for the classification, the distribution of the cloud properties might not be as representative of  
512 single cloud types as an input tile of, for example, 16 pixels. The main caveat regarding performance on high and medium clouds

513 from our method is that the ground-based observer identifies these cloud types with higher uncertainty compared to that of low  
 514 clouds. Additionally, stratiform clouds with high cloud fraction can hinder the trustworthiness of the surface observation if the  
 515 whole field of view is cloudy. Even though the limitations of ground-based observations are evident, they still provide quality  
 516 observations on which a classification model can be trained. The characteristics observed in the histograms across cloud types  
 517 contribute to an increase in confidence in the ability of CloudViT to discern various cloud types in large remote sensing datasets.  
 518



519

520 **Figure 5: Density histograms of cloud properties for each cloud type from high, medium, cumuliform and stratiform.**

## 521 5 Results

522

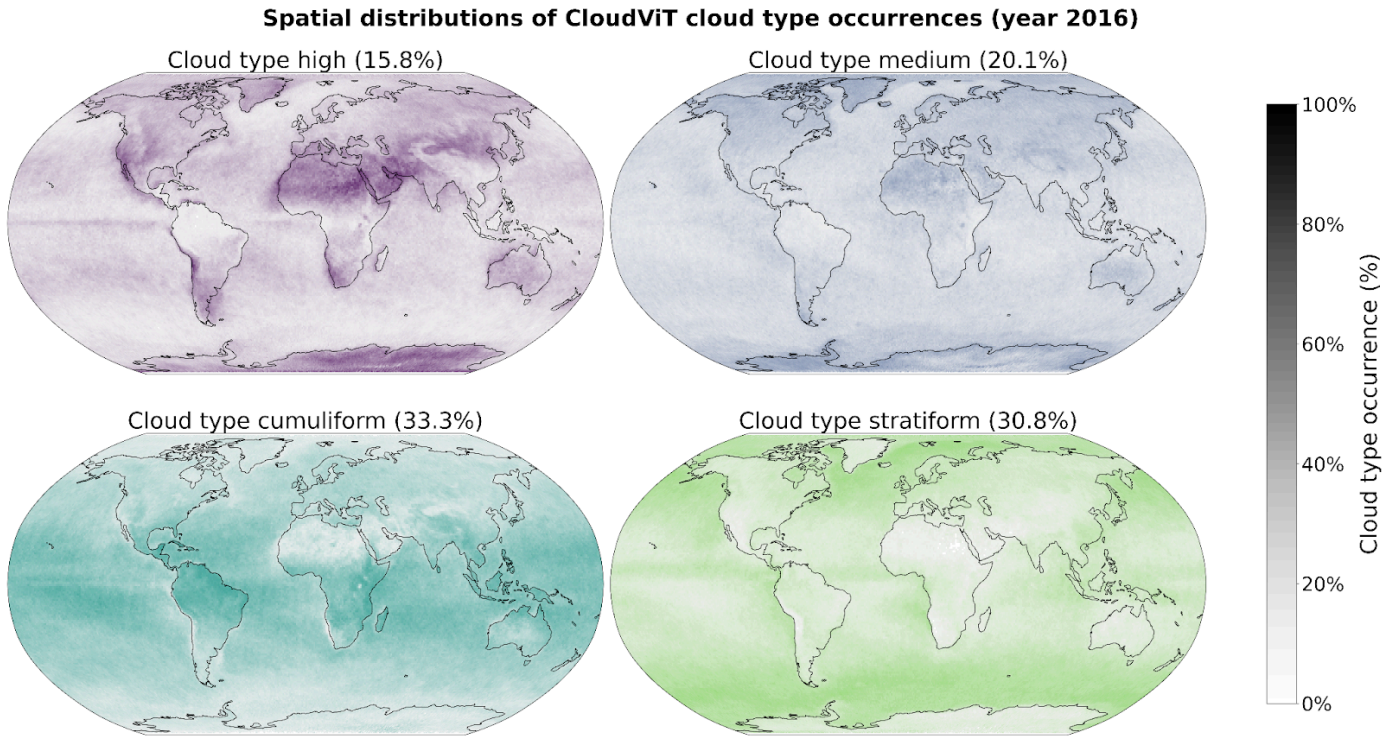
### 523 5.1 Global cloud type distributions in MODIS data

524

525 Additionally to the physical and microphysical characteristics of the different cloud types, their global spatial distribution can  
 526 help us further understand in which regions they are more or less frequent and qualitatively assess the presented classification  
 527 method compared to other remote sensing products. To this extent, as mentioned in the previous evaluation section (see Section  
 528 4), we build an extensive cloud type dataset for the year 2016 from MODIS AQUA granules which are regularly sampled for  
 529 tiles of 128x128 pixels. The sampling step (64) is chosen for computational efficiency and memory purposes to be not too small  
 530 to avoid large overlap between neighbouring tiles but large enough to ensure representativeness in the later aggregated  
 531 predictions of the MODIS granules. Furthermore, as the area covered by each tile is rather wide, the spatial distribution of cloud  
 532 types might be less smooth than other products (e.g. Sassen et al., 2008) or other methods (Zantedeschi et al., 2020) which are  
 533 providing cloud types for smaller cloud fields. Additionally, the dataset is built on single daily overpasses of the MODIS  
 534 instrument and can thus be biased towards the local retrieval time (13:30 h, early afternoon for AQUA).

535 The spatial distributions of the predicted cloud types for the global dataset for the year 2016 are detailed in Figure 6 and Figure  
 536 C.2 for 4 and 10 cloud types, respectively. Firstly, we note that CloudViT predictions capture large scale patterns which are in  
 537 agreement with observational datasets (Sassen et al., 2008; Cesana et al., 2019; Wood, 2012; Pincus et al., 2023). Stratiform  
 538 clouds, and in particular stratocumulus (see Fig C.2), are frequent in the high latitudes and along the western coasts of America  
 539 and Africa. Cumuliform clouds are concentrated in the Tropics apart from the areas where stratocumulus clouds are dominant.

540 Medium clouds are concentrated in the polar regions and over land in the higher latitudes. High clouds make up a large portion  
 541 of clouds in the polar regions but also over land. The first notable difference is the low occurrence of high clouds in the Tropics  
 542 which would be expected to be higher (Sassen et al., 2008; Pincus et al., 2023). An explanation could be the frequent occurrence  
 543 of high clouds in multi-layer cloud scenes related to convection in the Tropics. Furthermore, in such cases the model probably  
 544 identifies the cloud types with larger cloud fraction and thus discards potential high clouds in the scene. Incorporating more  
 545 samples of high clouds in that region (see Fig. A.1) could potentially help the performance of the classification model in that  
 546 regard. The presented spatial distributions may suffer from the somewhat limited performance of the classification model despite  
 547 the corresponding reasonable representation of cloud type characteristics showcased in section 4.2. Nevertheless, some  
 548 informative features are observed in Figures 6 and C.2 and point towards the good direction for further improving CloudViT.  
 549

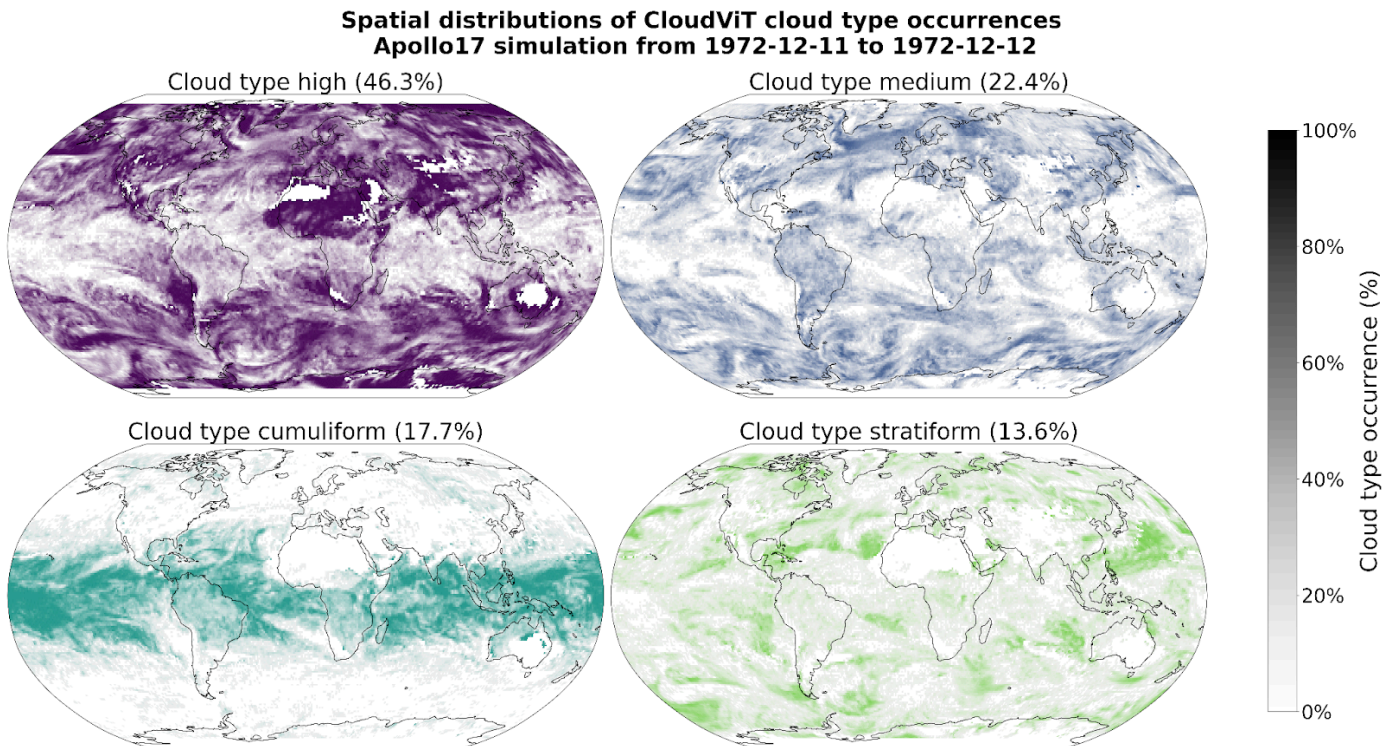


550  
 551 **Figure 6: Spatial distributions of the CloudViT cloud type occurrences (cloud types high, medium, cumuliform,**  
 552 **stratiform) for MYD06 granules for the year 2016 aggregated on a 1° regular grid.**  
 553

## 554 5.2 Application to a global storm-resolving model simulation

555  
 556 As a proof of concept and for probing the potential of CloudViT, we investigate the cloud type representation in general  
 557 circulation model (GCM) outputs using our CloudViT method. We build on a new generation of GCMs at kilometre resolution,  
 558 namely the ICON-Sapphire (Hohenegger et al., 2023). As the resolution of the simulation increases, some processes like deep  
 559 convection can be directly resolved instead of parameterized. Hence, building diagnostics about cloud representation is of  
 560 importance to help evaluate the simulations. In particular, we use the simulation run by the Max Planck Institute for Meteorology  
 561 (MPI-M) for the period between the 5th and 12th of December 1972, aiming at recreating the Blue Marble picture made during  
 562 the Apollo 17 mission on the 7th of December. Here we only use the complete outputs provided for the 11th of December. The  
 563 grid used contains 335 544 320 grid points at each level in the atmosphere (R02B11 grid), and outputs are provided every 30  
 564 minutes during the simulation for the atmospheric quantities of interest, resulting in overall 48 time steps. As the effective  
 565 horizontal resolution of the model simulation and the MODIS data are on similar scales, we can effectively apply CloudViT on  
 566 the model outputs. From the model outputs, we derive the cloud properties necessary for the method introduced in this study.  
 567 More information about the particular model setup and the derivation of cloud properties is included in Appendix D. However,  
 568 the standardisation of the input cloud properties for the vision transformer model is still done based on statistics computed on  
 569 MODIS data which could induce a bias in the latent representations and subsequently on the predictions. Extending the method

570 to other datasets like this GCM simulation thus requires careful investigation that the cloud properties lie in the same range or  
 571 display similar distributions.  
 572 For each 30-minute time step, we proceed to sample tiles, regularly spaced, to reach global coverage of cloud type estimates.  
 573 Figure 7 displays the daily averaged occurrence of the cloud type predictions on a 1° regular grid for the 4 cloud types, the  
 574 equivalent for 10 cloud types is presented in Figure D.3. A large proportion of the predicted clouds belong to the high cloud type,  
 575 hinting at the difference in sensitivity to clouds retrieved in the climate model data compared to the MODIS retrievals or the  
 576 mismatch in the training process of CloudViT, high clouds being underrepresented and their corresponding classification metrics  
 577 lower than for some other cloud types. However, increasing the cloud ice content threshold by an order of magnitude greatly  
 578 decreases the amount of thin, high and cold clouds in the simulation dataset. This aspect would need further tuning through  
 579 comparison with remote sensing retrievals which are not available for this particular simulated period. On the other hand, the  
 580 cumuliform class captures well the convective systems in the tropics while the stratocumulus decks can be identified (Fig. D.3).  
 581 Additionally, the medium clouds are more present at high latitudes. An important aspect to factor in is that the classification  
 582 model was only trained on daytime satellite observations as the optical cloud properties necessary are only available then. Thus,  
 583 results on nighttime cloud retrievals which is the case for some of the predictions produced from the presented simulation might  
 584 need more meticulous evaluation. Even though it is a limiting factor in the case of the satellite dataset we are using, the  
 585 simulation outputs provide us with the required variables across all timesteps.



586  
 587 **Figure 7: Spatial distributions of the CloudViT cloud type occurrences (cloud types high, medium, cumuliform,**  
 588 **stratiform) for the ICON-Sapphire Apollo 17 simulation of December 11<sup>th</sup> 1972 aggregated on a 1° regular grid.**

## 589 6 Conclusion

590  
 591 This study introduces a new method called CloudViT to classify cloud types from MODIS cloud properties, specifically CTH,  
 592 COT and CWP. CloudViT delivers estimates for either 4 (high, medium, cumuliform, stratiform) or 10 (cirrus, cirrostratus,  
 593 cirrocumulus, altostratus, altocumulus, cumulus, cumulus and stratocumulus, cumulonimbus, stratocumulus, stratus) cloud types  
 594 with fair performance. The classification model was built on ground-based observations of cloud types (Section 2.1) and  
 595 experiments about its generalisation skill and the benefits of spatial information were presented (Section 3). We evaluated the  
 596 classification model by examining distributions of cloud properties in Section 4 and the global spatial distribution of cloud types  
 597 in Section 5.1. Lastly, we transferred our method to a km-scale climate model simulation made with ICON-Sapphire (Section  
 598 5.2). The global dataset alongside the CloudViT code and weights are made available on Zenodo (Lenhardt et al., 2024b).

599 Spatially-resolved cloud properties provide usable context for the CloudViT model to improve the cloud classification, as shown  
600 in the comparison to the baseline method with limited spatial information. Introducing this new transformer model architecture  
601 additionally improves the classification skill over the CNN backbone mentioned in Lenhardt et al. (2024a). Overall, CloudViT  
602 achieves passable performance even on sparsely represented classes for both cases of 4 and 10 cloud types. The limited colocated  
603 dataset proves to be a hurdle for the proper training and evaluation of the method on labelled samples but the generation of an  
604 extensive global dataset allows deeper investigation into the cloud types. Improvements could come from a more extensive  
605 training dataset which would encompass a larger variety of cloud type samples to certainly enhance the classification's  
606 performance both for the training and testing metrics. The subsequent evaluation exhibits interesting results despite the limited  
607 performance on the colocated dataset. In the global dataset, the predicted cloud types exhibit fairly physically reasonable  
608 distributions of their respective cloud properties, and their global spatial distributions are consistent in parts with other products  
609 (Section 5.1). Application to climate model data proves to be straightforward and results in insights into how such methods can  
610 be transferred to model data, and preliminarily on how clouds are represented in global km-scale simulations. The necessary  
611 cloud quantities are obtained from common simulation outputs (cloud liquid water and ice contents, altitude, droplet number)  
612 which makes CloudViT easily applicable to other climate model simulations. Cloud type diagnostics such as CloudViT could be  
613 a resourceful addition to the panel of assessment methods for model data (Kuma et al., 2023; Kaps et al., 2023).

614 Overall, the method would benefit from including further ground-based observations through the collocation process but then  
615 much larger storage and computational facilities would be needed as global MODIS data represents thousands of granules each  
616 day. More training samples could simultaneously solve performance issues by providing a clearer vision of the different cloud  
617 types for the classification model to learn from. The classification model could also be refined by finding better alternatives to  
618 the RF or MLP presented here. The overall finetuning process involving the vision transformer and the MLP classification head  
619 proved to be cumbersome but holds great promise if the labels and training process are refined. Transfer learning from a typical  
620 ImageNet-trained model did not yield a notable performance difference which shows the current need for foundation models  
621 trained on remote sensing data. The main hurdle here remains the large diversity in instruments, quantities and resolutions among  
622 remote sensing products which hinders the possibility of a unified model.

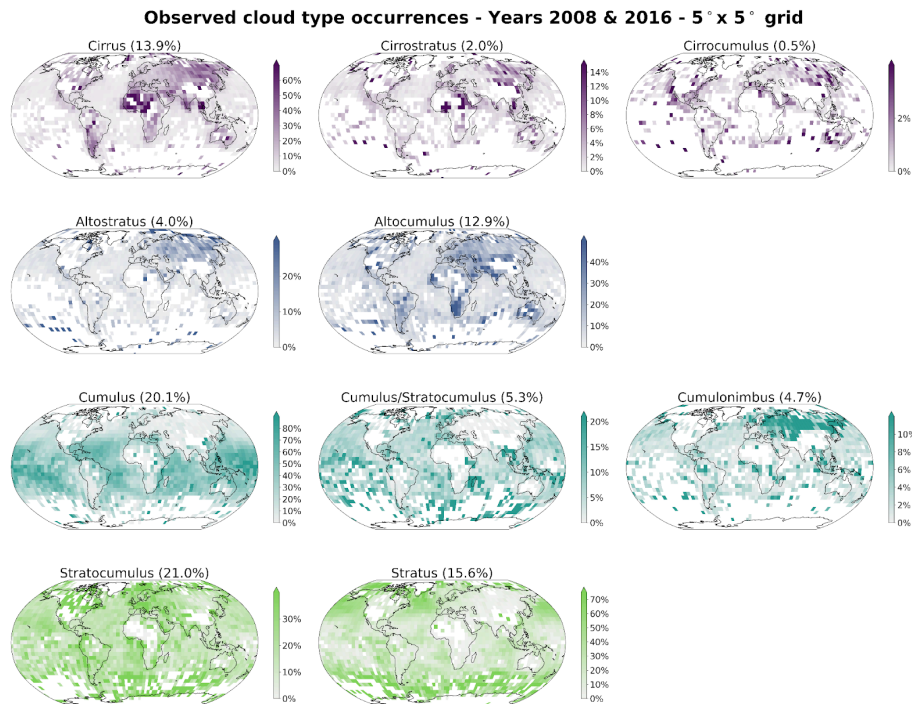
623 To improve the spatial coverage of the CloudViT predictions, the direct application to granules from MODIS TERRA would  
624 technically not require much more work as the instruments are similar and provide the same cloud properties. An additional  
625 benefit would come after the upcoming decommissioning of the CloudSat mission which was providing cloud type retrievals  
626 along its track aligned with MODIS. We would then be able to still offer information about cloud types over the same areas even  
627 though no vertical information is available and used from our predictions on MODIS level 2 data. As for other satellite cloud  
628 products, the main difference would arise, similarly to climate model data, from the potentially different distributions and ranges  
629 in the input cloud properties which would need either retraining of the vision transformer or careful scaling to match the  
630 distributions seen in MODIS data. Some limitations due to satellite retrieval shortcomings should be taken into account when  
631 applying the described method to certain areas. Indeed, since MODIS data is collected through near-nadir scanning, observations  
632 in high-latitude regions become oblique, leading to distortions and potential errors in cloud property retrievals, such as cloud top  
633 height and optical thickness.

634 Furthermore, some caveats appear when applying CloudViT to climate model data. As mentioned previously, the input scaling is  
635 crucial to ensure proper portability of the method to this other data source. The absence of nighttime retrievals in the MODIS  
636 data also turns the evaluation of predictions on nighttime data points across the model data into a challenging issue. However,  
637 clouds play a role in the climate system both during the day when they cool the surface by mostly reflecting incoming solar  
638 radiation but also at night when they warm the surface by trapping outgoing terrestrial radiation. Shifts and changes in cloud  
639 occurrence and distribution in the current climate but also in future projections could further influence global climate change  
640 (Luo et al., 2024). The proof of concept of applying CloudViT to a limited climate model simulation is encouraging but  
641 considering more common and computationally less expensive global km-scale simulations (horizontal resolution of 5 km for  
642 example) could be of greater interest to the community to study longer time scales. To this extent, two conceivable approaches  
643 would consist in either retraining the CloudViT model on coarser input cloud properties matching the model data resolution - the  
644 MODIS Cloud product is also available at a 5 km resolution even though the 1 km equivalent is recommended for use - or in  
645 using CloudViT as is but with the coarse input scaled to fit the resolution of the tiles on which it was trained on. The first option  
646 could be more interesting as computer vision models are commonly trained on coarser resolutions first to learn the broad  
647 specificity and patterns in the data before fine-tuning the model on finer resolution (Touvron et al., 2019).

## 649 Appendix A: Cloud type observations

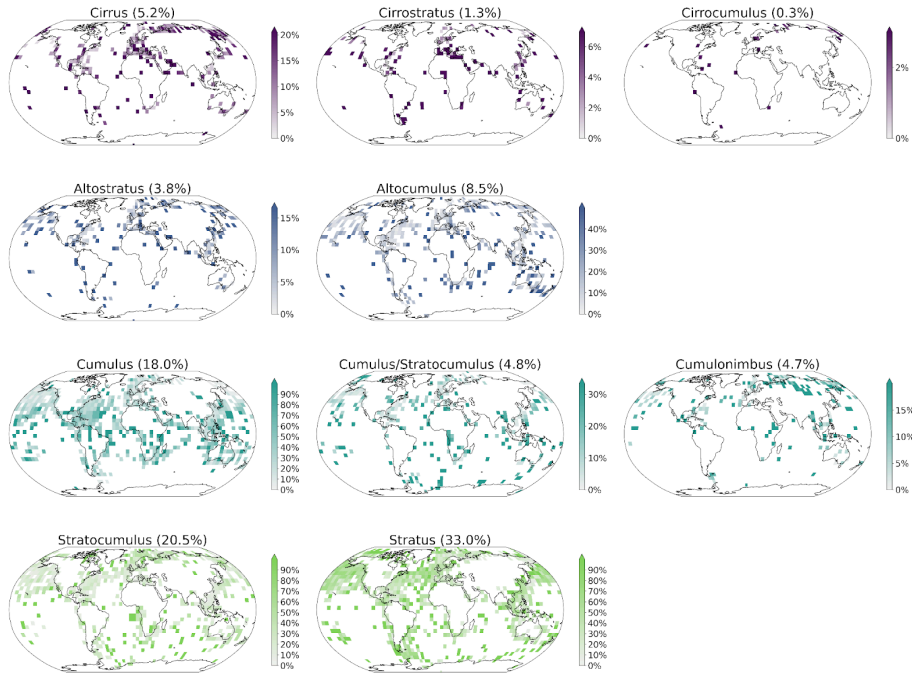
WMO codes	Cloud type: 4 groups	Cloud type: 10 groups	Colocated samples
High clouds 1-6	High	Cirrus	n = 574
High clouds 7-8		Cirrostratus	n = 142
High clouds 9		Cirrocumulus	n = 29
Medium clouds 1-2	Medium	Altostratus	n = 420
Medium clouds 3-9		Alto cumulus	n = 944
Low clouds 1-3	Cumuliform	Cumulus	n = 1998
Low clouds 8		Cumulus and stratocumulus	n = 533
Low clouds 9	Stratiform	Cumulonimbus	n = 519
Low clouds 4-5		Stratocumulus	n = 2274
Low clouds 6-7		Stratus	n = 3661
Total			n = 11 094

650 **Table A.1: Cloud types from the WMO observational datasets, their groups following Kuma et al. (2023) and the**  
651 **corresponding number of samples in the colocated dataset. The WMO codes correspond to the 9 types for each level.**



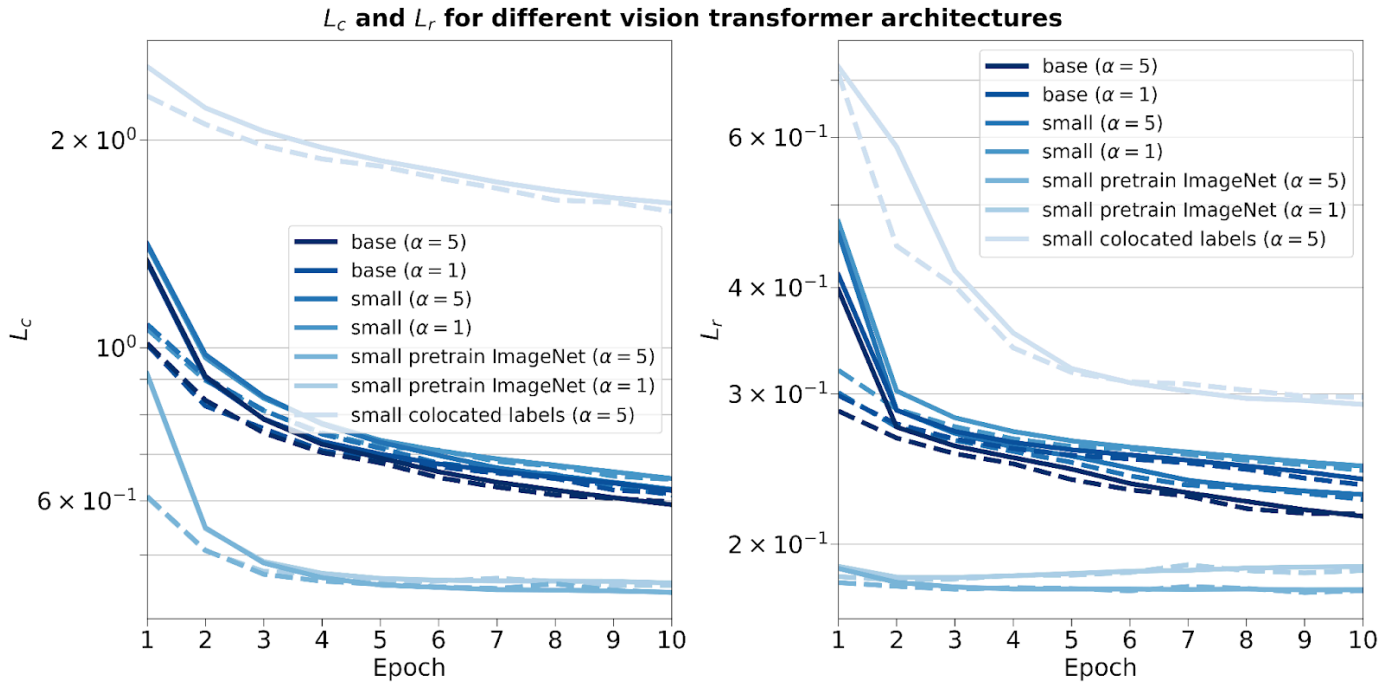
652 **Figure A.1: Spatial distributions of observed cloud types (cloud types cirrus, cirrostratus, cirrocumulus, altostratus,**  
653 **alto cumulus, cumulus, cumulus and stratocumulus, cumulonimbus, stratocumulus, stratus) from the Met Office datasets**  
654 **(Met Office, 2006; Met Office, 2008) for the years 2008 and 2016. Overall percentage of each label in the total dataset is**  
655 **indicated in brackets.**  
656  
657

Observed cloud type occurrences - Colocated dataset - 5° x 5° grid



658

659 **Figure A.2: Spatial distributions of observed cloud types (cloud types cirrus, cirrostratus, cirrocumulus, altostratus,**  
 660 **altocumulus, cumulus, cumulus and stratocumulus, cumulonimbus, stratocumulus, stratus) from the Met Office datasets**  
 661 **(Met Office, 2006; Met Office, 2008) for the years 2008 and 2016 colocated with the satellite cloud retrievals (Platnick et**  
 662 **al., 2017) used for training the classification model. Overall percentage of each label in the total dataset is indicated in**  
 663 **brackets.**



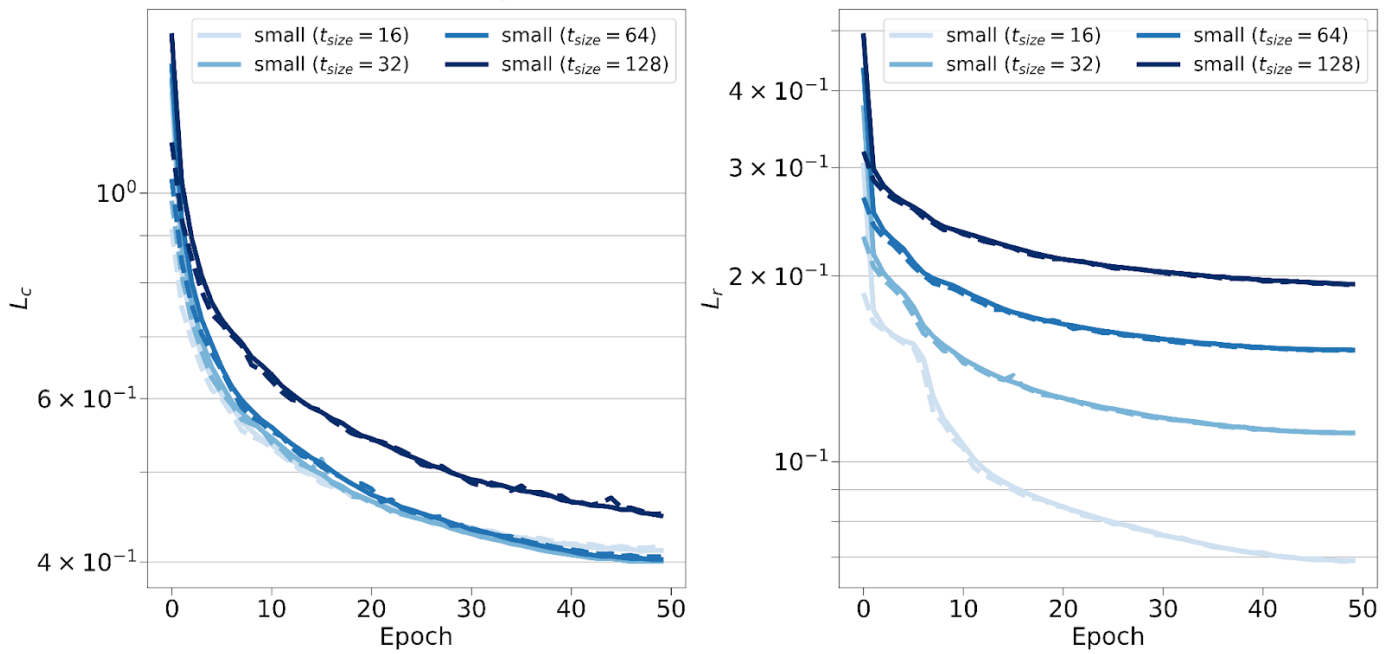
669 **Figure B.1: Training and validation contrastive (left) and reconstruction (right) losses for different vision transformer**  
 670 **architectures, pretraining weights, training datasets and scaling factor  $\alpha$ .**

672 **B.2 Reconstruction errors for the CNN autoencoder and the vision transformer (small variant) on the test set**

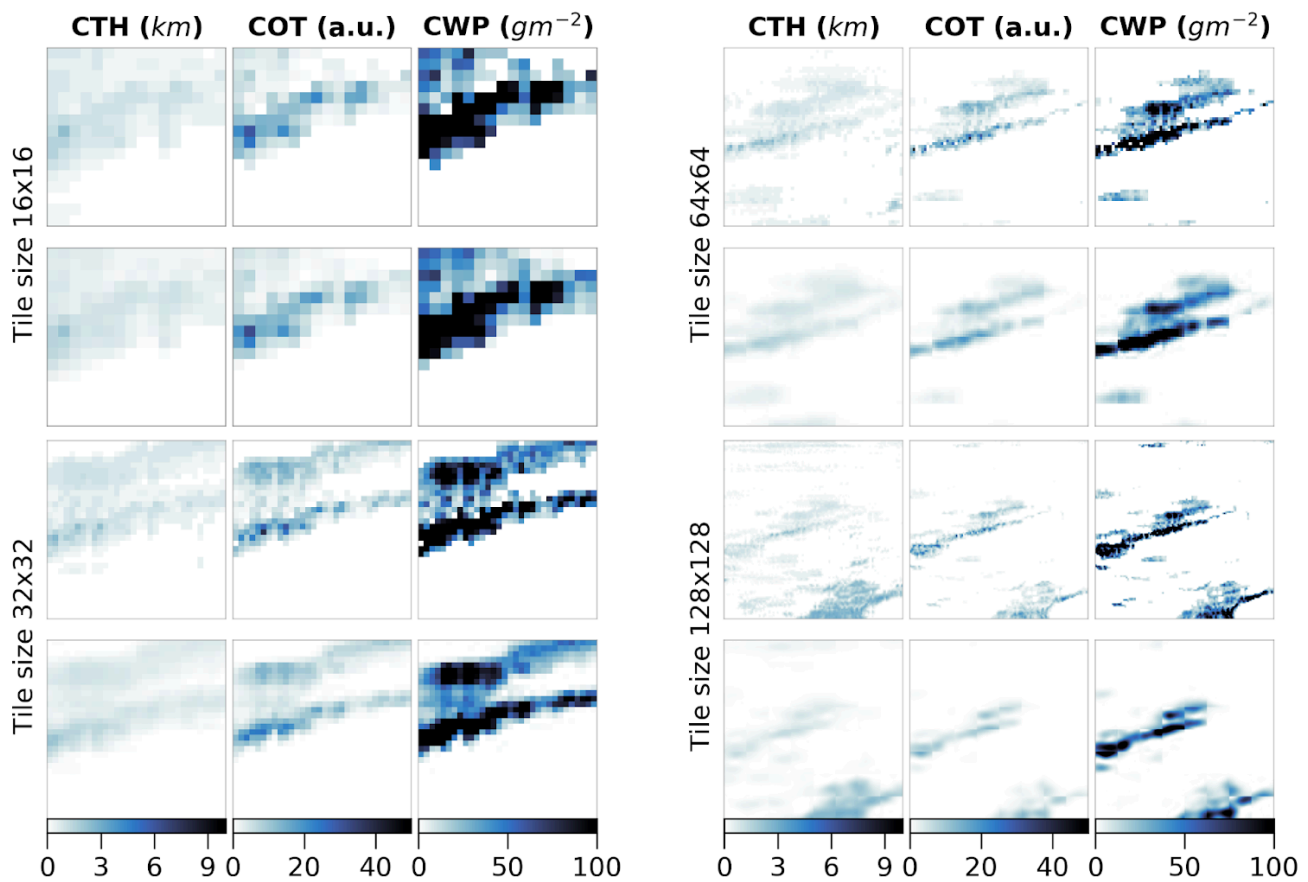
Model type	Reconstruction error	CTH	COT	CWP
CNN autoencoder	MSE	0.27	0.39	0.25
	l1-loss	0.36	0.33	0.21
Vision transformer (small variant)	MSE	0.06	0.25	0.13
	l1-loss	0.10	0.17	0.10

674 **Table B.1: Reconstruction relative errors of the CNN (Lenhardt et al., 2024a) and the vision transformer models across**  
 675 **channels (CTH, COT and CWP) on the test dataset.**

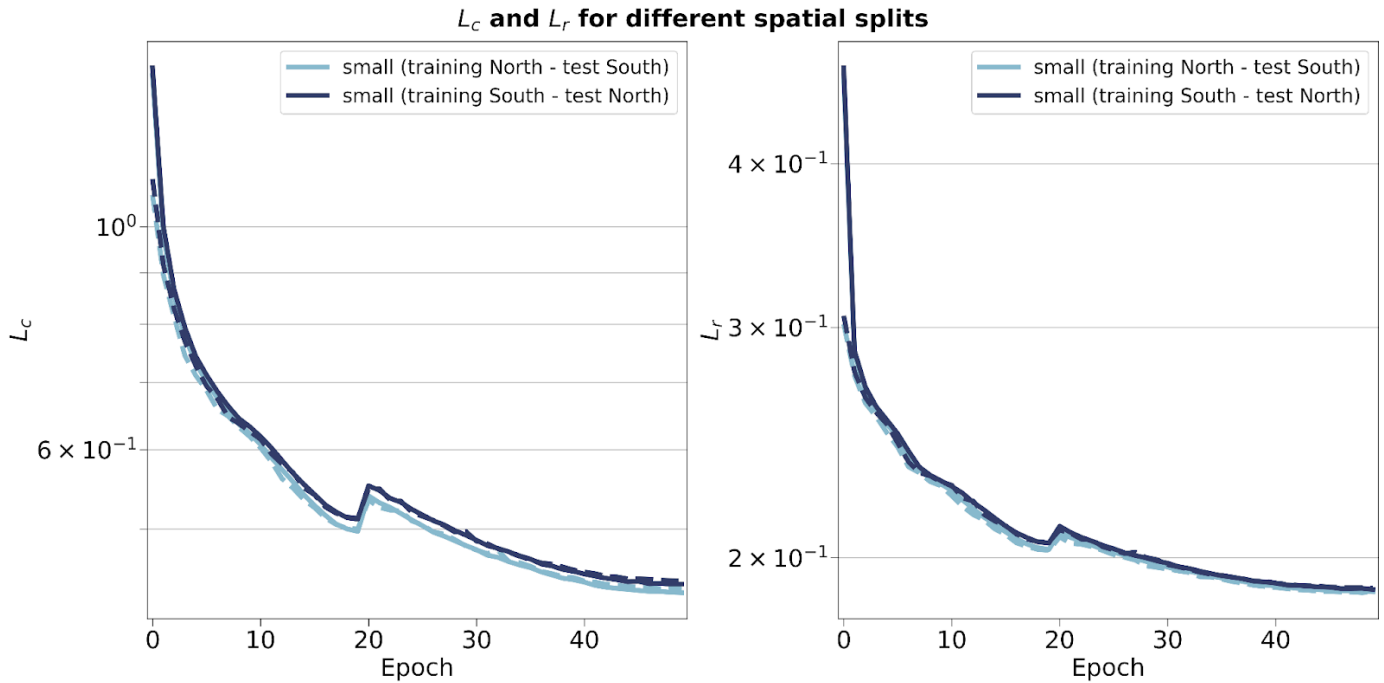
$L_c$  and  $L_r$  for different tile sizes



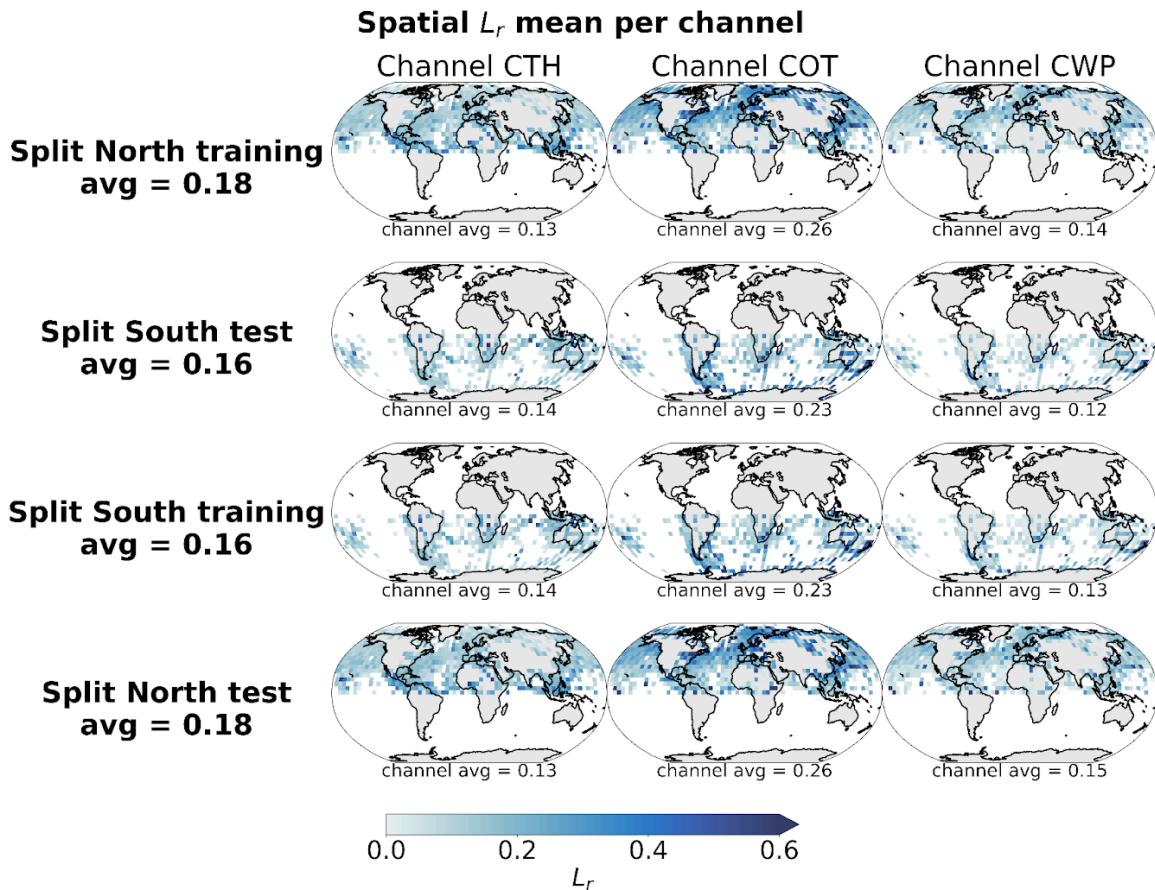
678  
 679 **Figure B.2: Training and validation contrastive (left) and reconstruction (right) losses for vision transformers trained on**  
 680 **different input tile sizes of 16, 32, 64 and 128.**  
 681



682  
 683 **Figure B.3: Input tiles (first and third rows) and corresponding reconstructions (second and fourth rows) for vision**  
 684 **transformers trained on the relevant input tile sizes of 16, 32, 64 and 128.**  
 685



687  
 688 **Figure B.4:** Training (full lines) and validation (dashed lines) metrics for the contrastive (left) and reconstruction (right)  
 689 losses for vision transformers trained on samples from the Northern or Southern hemispheres.



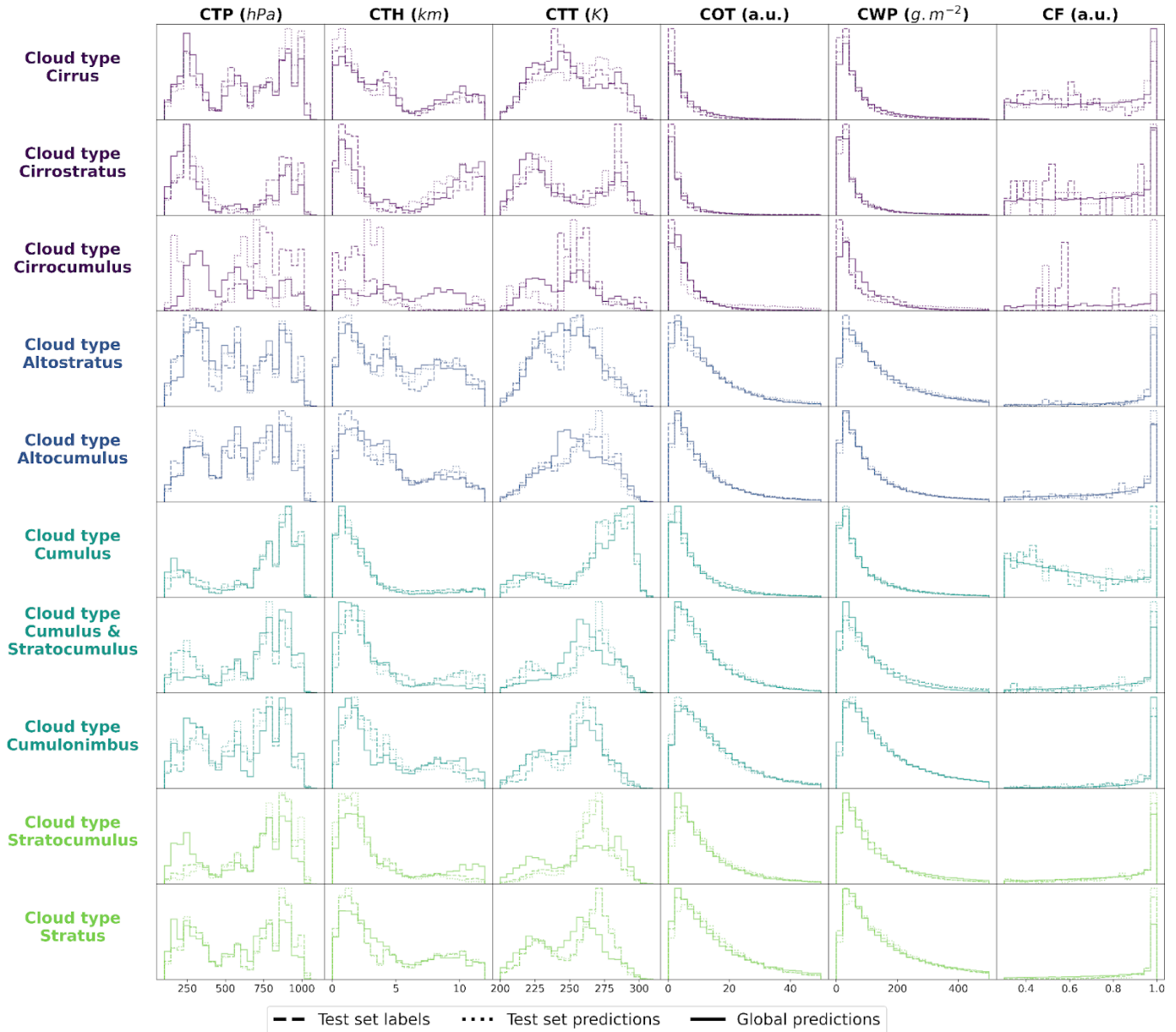
690  
 691 **Figure B.5:** Spatial distributions of the mean channel reconstruction errors for the Northern and Southern hemispheres  
 692 colocated samples. The first two rows correspond to the model trained on the samples from the Northern hemisphere and  
 693 the last two rows to the model trained on the samples from the Southern hemisphere.

694 Appendix C: Cloud type classification for 10 types

Method	Multi-class accuracy *	IBA geometric mean	F1-score *
Baseline 9x9 RF	0.19	0.26	0.16
Baseline CNN/RF	0.22	0.18	0.17
CloudViT/MLP	0.22	0.20	0.16
<b>CloudViT/RF</b>	<b>0.23</b>	<b>0.26</b>	<b>0.21</b>

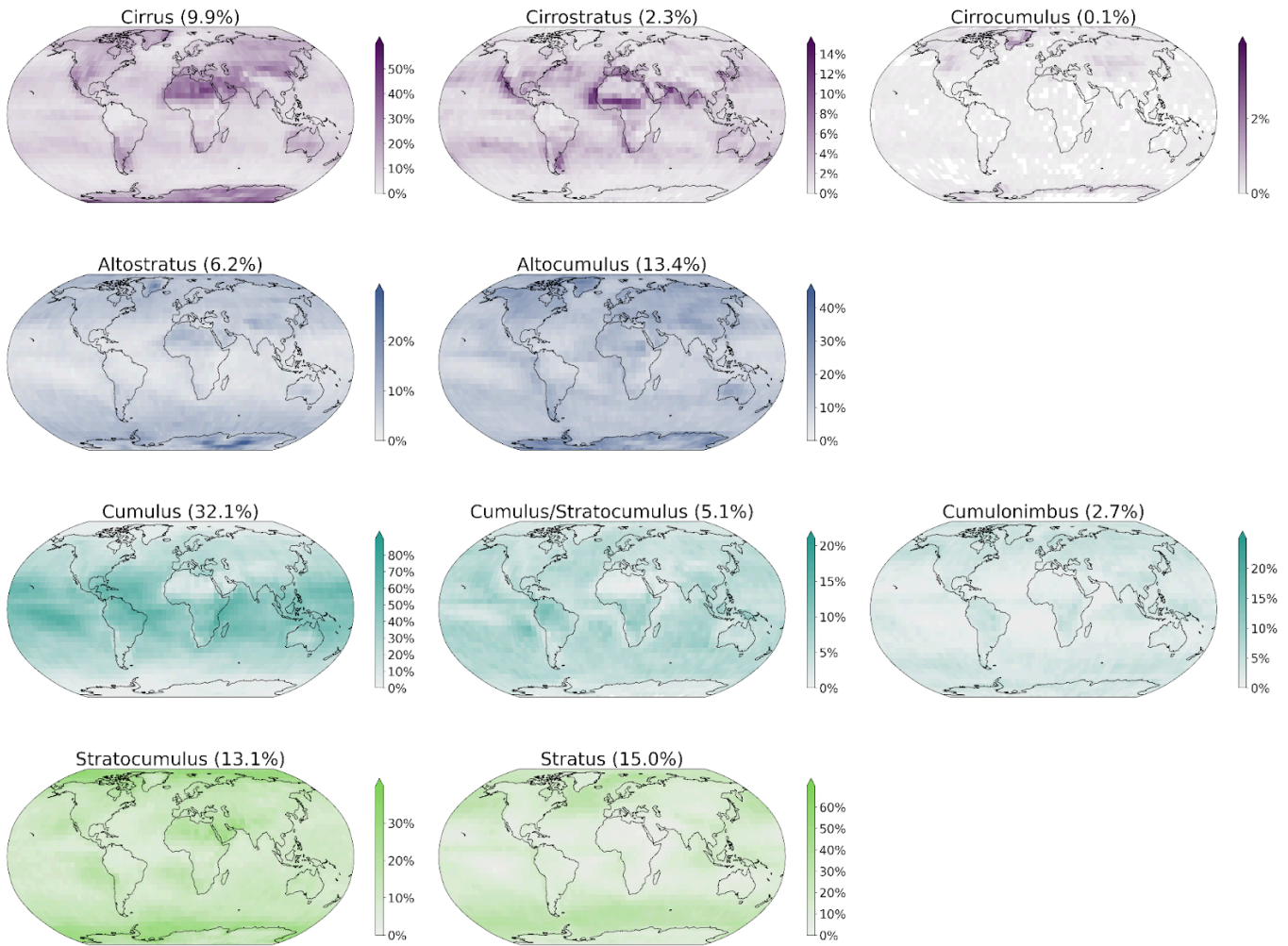
695 Table C.1: Classification metrics on the test set in the case of 10 cloud types. The metrics noted with a \* are referring to  
 696 their macro-averaged estimate. The baseline CNN/RF refers to the CNN backbone introduced in Lenhardt et al. (2024a).  
 697

Density histograms of cloud properties per cloud class for test set labels, test set predictions and global predictions



698  
 699 Figure C.1: Density histograms of cloud properties for each cloud type from cirrus, cirrostratus, cirrocumulus,  
 700 altostratus, alto cumulus, cumulus, cumulus and stratocumulus, cumulonimbus, stratocumulus, stratus.  
 701

### Spatial distributions of CloudViT cloud type occurrences (year 2016)



702

703 Figure C.2: Spatial distributions of the CloudViT cloud type occurrences (cloud types cirrus, cirrostratus, cirrocumulus,

704 altostratus, alto cumulus, cumulus, cumulus and stratocumulus, cumulonimbus, stratocumulus, stratus) for MYD06

705 granules for the year 2016 aggregated on a 1° regular grid.

706 **Appendix D: Cloud properties computation from model simulation output**

707

708 In order to compute the different cloud properties used in our method (Table 1), we use the available atmospheric outputs from  
 709 the model simulation. The simulation was made using the ICON-2.6.6-rc version in R02B11 grid resolution with 90 vertical  
 710 levels in the atmosphere (335544320 grid points per level) and 128 vertical levels in the ocean (237102291 surface grid points).  
 711 Observed aerosols and greenhouse gas concentrations of December 1972 were used for the atmosphere.

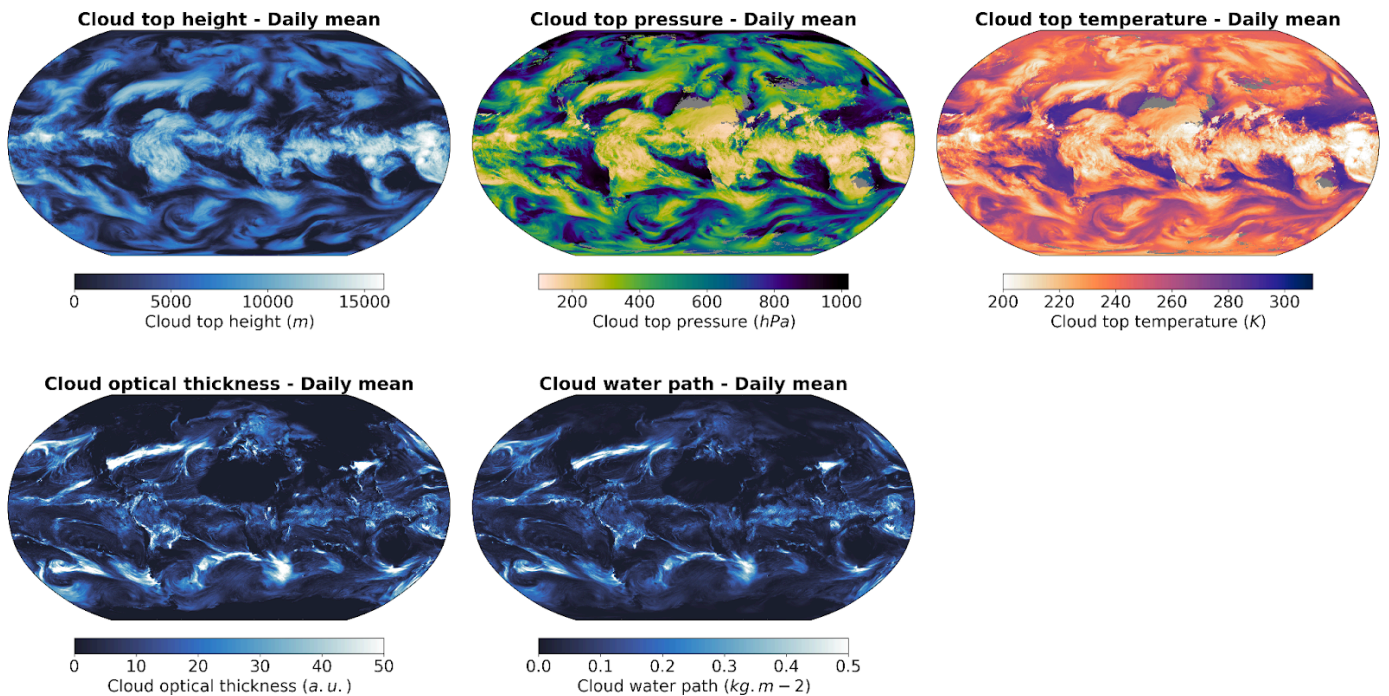
712 The cloud top quantities are retrieved by defining the top-most level where the liquid water content (variable name *clw*) or the ice  
 713 content (variable name *cli*) are above a predefined threshold of  $1 \text{ mg.kg}^{-1}$ . This threshold relates to particles of sizes of at least a  
 714 few micrometres which is similar to what the sensors on the MODIS AQUA instrument are able to retrieve. Using 3D outputs of  
 715 atmospheric quantities like temperature (variable name *ta*) and pressure (variable name *pfull*), we derive the cloud top properties  
 716 also present in the MODIS MOD/MYD06 level 2 cloud properties product. The CTH is derived using the altitude in the  
 717 corresponding vertical level in the grid. Secondly, the CWP is computed by summing the vertically integrated cloud liquid water  
 718 path (variable name *cllvi*) and cloud ice path (variable name *clivi*) which are already provided as simulation outputs. Lastly, we  
 719 computed the COT by vertically summing the layer-wise COT computed from the following equation, detailed in Carslaw  
 720 (2022), equation 12.49 (Chapter 12.3, page 515):

721 
$$\tau_c = \frac{9}{5} \left(\frac{4\pi}{3\sqrt{2}}\right)^{1/3} \rho_w^{-2/3} (kN_d)^{1/3} c_w^{-1/6} L^{5/6} = 0.2303 \text{ kg}^{-5/6} \text{ m}^{8/3} (kN_d)^{1/3} L^{5/6} \quad (\text{D.1})$$

722 Where  $L = clw * \rho_{air} * \delta z$  the layer liquid water path,  $\rho_w = 1000 \text{ kg.m}^{-3}$  density of water,  $k = 1$  a factor to account for  
 723 the width of the droplet size distribution,  $c_w = 2e^{-6} \text{ kg.m}^{-4}$  the adiabatic condensation rate and  $N_d$  the vertical droplet number  
 724 defined in the simulation by the ECHAM6 parameterization (Equation 6; Stevens et al., 2013).

725

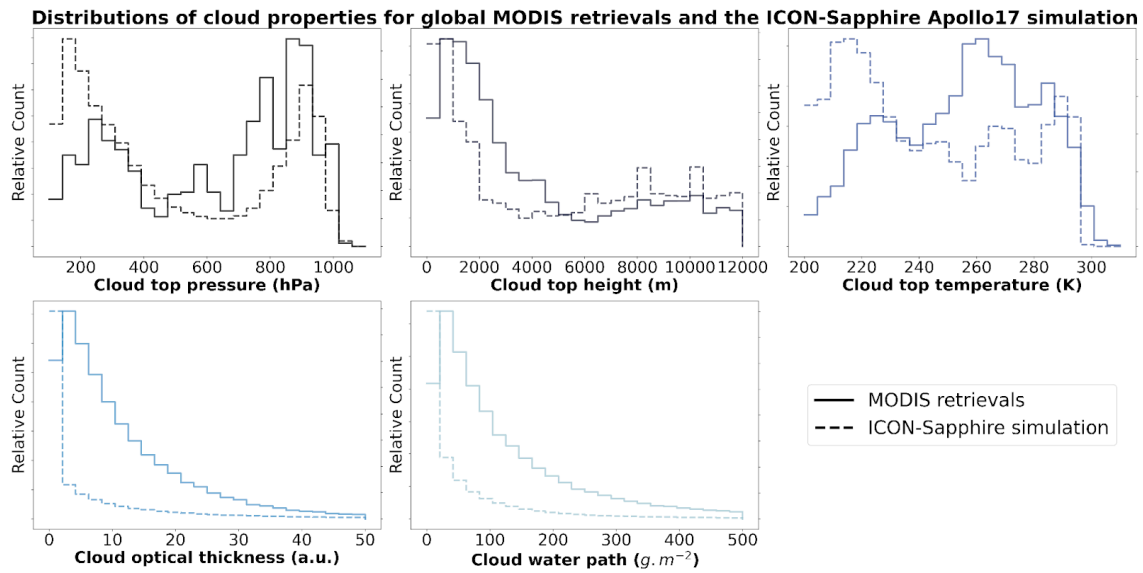
726



727

728 **Figure D.1: Daily averages of cloud top height, cloud top pressure, cloud top temperature, cloud optical thickness and**  
 729 **cloud water path for the 11th of December 1972 from the ICON-Sapphire Apollo 17 simulation.**

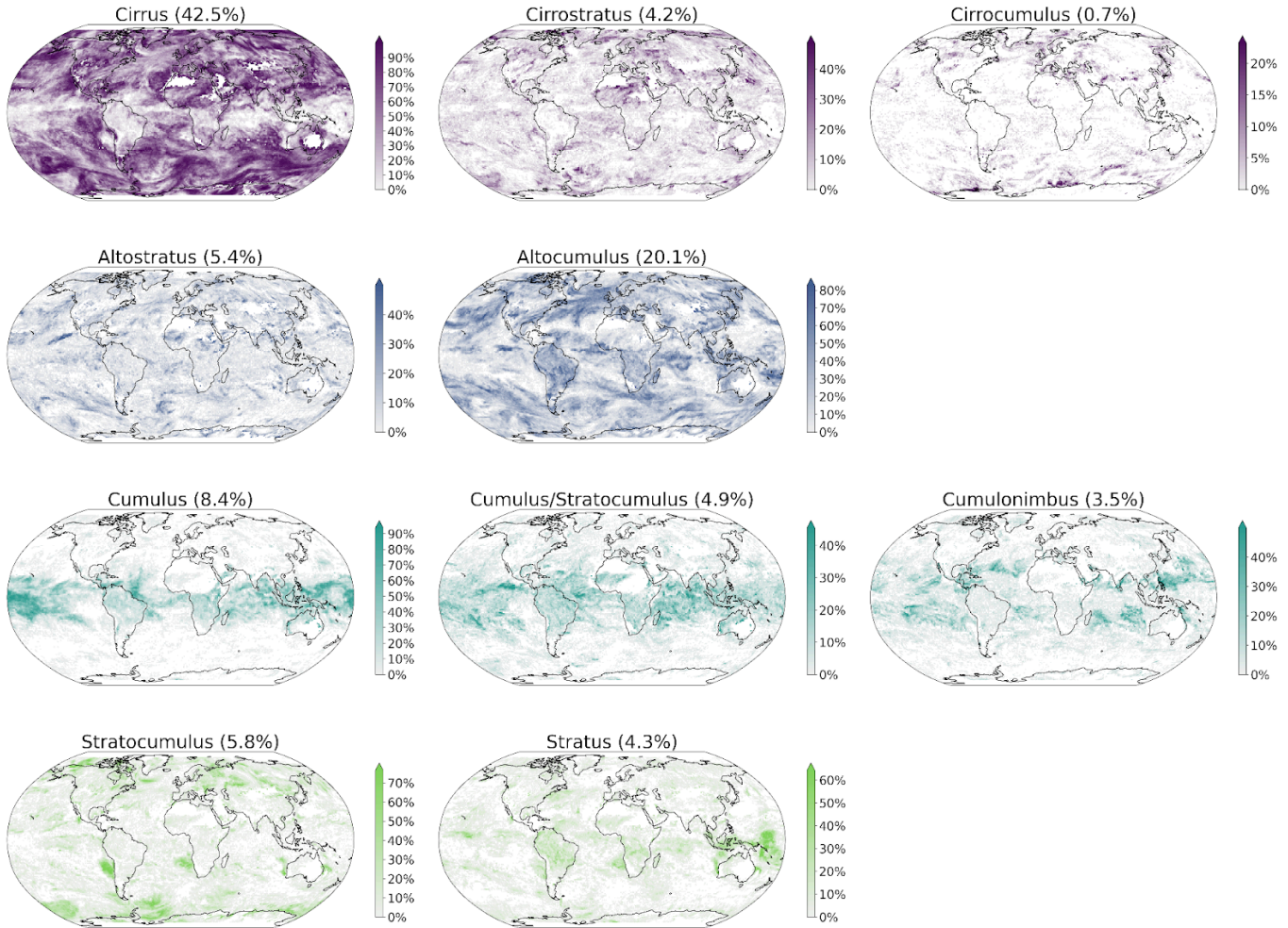
730



731

732 **Figure D.2: Distribution of cloud top pressure, cloud top height, cloud top temperature, cloud optical thickness and cloud**  
 733 **water path for MODIS AQUA retrievals and the ICON-Sapphire Apollo 17 simulation.**

**Spatial distributions of CloudViT cloud type occurrences**  
**Apollo17 simulation from 1972-12-11 to 1972-12-12**



734

735 **Figure D.3: Spatial distribution of the CloudViT cloud type occurrences (cloud types cirrus, cirrostratus, cirrocumulus,**  
 736 **altostratus, altopcumulus, cumulus, cumulus and stratocumulus, cumulonimbus, stratocumulus, stratus) for the**  
 737 **ICON-Sapphire Apollo 17 simulation of December 11th 1972 aggregated on a 1° regular grid.**

### 738 **Code availability**

739  
740 The code used for the method and producing the plots is available on Zenodo (Lenhardt et al., 2024b).

### 741 **Data availability**

742  
743 The global dataset of the cloud type predictions for the year 2016 is available on Zenodo (Lenhardt et al., 2024b). The dataset is  
744 available as a csv file with corresponding coordinates, MODIS granule file, time of retrieval and predicted cloud type (4 and 10  
745 groups) or in a netCDF file as daily aggregates on a regular grid with a resolution of 1 ° or 5 °. The meteorological observations  
746 from the UK MetOffice (Met Office, 2006; Met Office 2008) are available through the CEDA archive at  
747 <https://catalogue.ceda.ac.uk/uuid/77910bcec71c820d4c92f40d3ed3f249> and  
748 <https://catalogue.ceda.ac.uk/uuid/9f80d42106ba708f92ada730ba321831> for ocean and land observations respectively. The files  
749 from the CUMULO dataset (Zantedeschi et al., 2019) are available at  
750 <https://www.dropbox.com/sh/i3s9q2v2jyjk2it/AACxXnXfMF5wuIqLXqH4NJOra?dl=0>. The simulation outputs are hosted by  
751 the DKRZ (Deutsches Klimarechenzentrum).

### 752 **Author contribution**

753  
754 JL, JQ, DS and DK designed the study. JL wrote the code. DK provided support regarding the climate model data. JL conducted  
755 the analysis and JL, JQ and DS interpreted the results. JL prepared the manuscript, JQ, DS and DK reviewed the manuscript and  
756 provided comments.

### 757 **Competing interests**

758  
759 Some authors are members of the editorial board of journal ACP.

### 760 **Acknowledgements**

761  
762 This work was supported by the European Union's Horizon 2020 research and innovation programme under Marie  
763 Skłodowska-Curie grant agreement No. 860100 (iMIRACLI). We thank the Leipzig University Scientific Computing cluster and  
764 the DKRZ (Deutsches Klimarechenzentrum, projects number bb1036 and bb1153) for computing and data hosting. We  
765 acknowledge the contributors of the CUMULO dataset (Zantedeschi et al., 2019) for providing access to the data files hosted at  
766 <https://www.dropbox.com/sh/i3s9q2v2jyjk2it/AACxXnXfMF5wuIqLXqH4NJOra?dl=0>. Additionally, we acknowledge the  
767 MODIS L2 Cloud product data set from the Level-1 and Atmosphere Archive and Distribution System (LAADS) Distributed  
768 Active Archive Center (DAAC), located in the Goddard Space Flight Center in Greenbelt, Maryland  
769 ([https://ladsweb.modaps.eosdis.nasa.gov/archive/allData/61/MYD06\\_L2/](https://ladsweb.modaps.eosdis.nasa.gov/archive/allData/61/MYD06_L2/)). We would like to also acknowledge Monika Esch,  
770 Emilie Fons and Hans Segura for support and discussions in handling the climate model data.

771

## 772 References

773

774 Ackerman, S. A., and Frey, R.: MODIS Atmosphere L2 Cloud Mask Product (35\_L2), NASA MODIS Adaptive Processing  
775 System, Goddard Space Flight Center, [http://doi.org/10.5067/MODIS/MOD35\\_L2.061](http://doi.org/10.5067/MODIS/MOD35_L2.061),  
776 [http://doi.org/10.5067/MODIS/MYD35\\_L2.061](http://doi.org/10.5067/MODIS/MYD35_L2.061), 2017.

777

778 Atito, S., Awais, M., & Kittler, J.: Sit: Self-supervised vision transformer, arXiv preprint,  
779 <https://doi.org/10.48550/arXiv.2104.03602>, 2021.

780

781 Baum, B.A., Menzel, W. P., Frey, R. A., Tobin, D. C., Holz, R. E., Ackerman, S. A., Heidinger, A. K., and Yang, P.: MODIS  
782 Cloud-Top Property Refinements for Collection 6, *Journal of Applied Meteorology and Climatology*, 51, 6, 1145–1163,  
783 <https://doi.org/10.1175/JAMC-D-11-0203.1>, 2012.

784

785 Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M., Kondo, Y., Liao, H., Lohmann,  
786 U., Rasch, P., Satheesh, S. K., Sherwood, S., Stevens, B. and Zhang, X. Y.: Clouds and aerosols, *Climate Change 2013: The*  
787 *Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on*  
788 *Climate Change*, 571–657, <https://doi.org/10.1017/CBO9781107415324.016>, 2013.

789

790 Breiman, L.: Random Forests. *Machine Learning*, 45 (1), 5–32, <https://doi.org/10.1023/A:1010933404324>, 2001.

791

792 Carslaw, K.: *Aerosols and Climate*, 1st Edition, Elsevier, ISBN 9780128197660, 2022.

793

794 Cesana, G., Del Genio, A. D., and Chepfer, H.: The Cumulus And Stratocumulus CloudSat-CALIPSO Dataset (CASCCAD),  
795 *Earth Syst. Sci. Data*, 11, 1745–1764, <https://doi.org/10.5194/essd-11-1745-2019>, 2019.

796

797 Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: Smote: synthetic minority over-sampling technique, *Journal*  
798 *of artificial intelligence research*, 16, 321–357, <https://doi.org/10.1613/jair.953>, 2002.

799

800 Chen, T., Kornblith, S., Norouzi, M., and Hinton, G.: A simple framework for contrastive learning of visual representations, in:  
801 *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, *Journal of Machine Learning Research*, 119,  
802 1597–1607, <https://dl.acm.org/doi/10.5555/3524938.3525087>, 2020.

803

804 Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database, in: 2009  
805 *IEEE conference on computer vision and pattern recognition*, Miami, FL, USA, 248–255,  
806 <https://doi.org/10.1109/CVPR.2009.5206848>, 2009.

807

808 Dhuria, H. L. and Kyle, H. L.: Cloud Types and the Tropical Earth Radiation Budget, *J. Clim.*, 3, 1409–1434,  
809 [https://doi.org/10.1175/1520-0442\(1990\)003<1409:CTATTE>2.0.CO;2](https://doi.org/10.1175/1520-0442(1990)003<1409:CTATTE>2.0.CO;2), 1990.

810

811 Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G.,  
812 Gelly, S., Uszkoreit, J., and Houshy, N. : An image is worth 16x16 words: Transformers for image recognition at scale, arXiv  
813 preprint, <https://doi.org/10.48550/arXiv.2010.11929>, 2020.

814

815 Forster, P., T. Storelvmo, K. Armour, W. Collins, J.-L. Dufresne, D. Frame, D.J. Lunt, T. Mauritsen, M.D. Palmer, M. Watanabe,  
816 M. Wild, and H. Zhang: The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity, in *Climate Change 2021: The*  
817 *Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on*  
818 *Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I.*  
819 *Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)].*  
820 *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 923–1054,*  
821 <http://doi.org/10.1017/9781009157896.009>, 2021.

822

823 García, V., Sánchez, J. S., and Mollineda, R. A.: On the effectiveness of preprocessing methods when dealing with different  
824 levels of class imbalance, *Knowledge-Based Systems*, 25, 13–21, <https://doi.org/10.1016/j.knosys.2011.06.013>, 2012.

825

826 Hartmann, D. L., Ockert-Bell, M. E., and Michelsen, M. L.: The Effect of Cloud Type on Earth's Energy Balance: Global  
827 Analysis, *J. Clim.*, 5, 1281–1304, [https://doi.org/10.1175/1520-0442\(1992\)005<1281:TEOCTO>2.0.CO;2](https://doi.org/10.1175/1520-0442(1992)005<1281:TEOCTO>2.0.CO;2), 1992.

828

829 Hendrycks, D., and Gimpel, K.: Gaussian error linear units (gelus), arXiv preprint, <https://doi.org/10.48550/arXiv.1606.08415>,  
830 2016.

831

832 Hinton, G. E.: Connectionist learning procedures, *Artificial intelligence*, 40, 185-234,  
833 [https://doi.org/10.1016/0004-3702\(89\)90049-0](https://doi.org/10.1016/0004-3702(89)90049-0), 1989.

834

835 Hohenegger, C., Korn, P., Linardakis, L., Redler, R., Schnur, R., Adamidis, P., Bao, J., Bastin, S., Behraves, M., Bergemann,  
836 M., Biercamp, J., Bockelmann, H., Brokopf, R., Brüggemann, N., Casaroli, L., Chegini, F., Datsaris, G., Esch, M., George, G.,  
837 Giorgetta, M., Gutjahr, O., Haak, H., Hanke, M., Ilyina, T., Jahns, T., Jungclaus, J., Kern, M., Klocke, D., Kluft, L., Kölling, T.,  
838 Kornbluh, L., Kosukhin, S., Kroll, C., Lee, J., Mauritsen, T., Mehlmann, C., Mieslinger, T., Naumann, A. K., Paccini, L.,  
839 Peinado, A., Praturi, D. S., Putrasahan, D., Rast, S., Riddick, T., Roeber, N., Schmidt, H., Schulzweida, U., Schütte, F., Segura,  
840 H., Shevchenko, R., Singh, V., Specht, M., Stephan, C. C., von Storch, J.-S., Vogel, R., Wengel, C., Winkler, M., Ziemann, F.,  
841 Marotzke, J., and Stevens, B.: ICON-Sapphire: simulating the components of the Earth system and their interactions at kilometer  
842 and subkilometer scales, *Geosci. Model Dev.*, 16, 779–811, <https://doi.org/10.5194/gmd-16-779-2023>, 2023.

843

844 Howard, L.: *Essay on the modifications of clouds*, John Churchill & Sons, London, 64 pp., 1803.

845

846 Kaps, A., Lauer, A., Camps-Valls, G., Gentine, P., Gómez-Chova, L., and Eyring, V.: Machine-Learned Cloud Classes From  
847 Satellite Data for Process-Oriented Climate Model Evaluation, *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-15,  
848 4100515, <https://doi.org/10.1109/TGRS.2023.3237008>, 2023.

849

850 Kuma, P., Bender, F. A.-M., Schuddeboom, A., McDonald, A. J., and Seland, Ø.: Machine learning of cloud types in satellite  
851 observations and climate models, *Atmos. Chem. Phys.*, 23, 523–549, <https://doi.org/10.5194/acp-23-523-2023>, 2023.

852

853 Kurihana, T., Moyer, E., Willett, R., Gilton, D., and Foster, I.: Data-Driven Cloud Clustering via a Rotationally Invariant  
854 Autoencoder, *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-25, 4103325,  
855 <https://doi.org/10.1109/TGRS.2021.3098008>, 2022.

856

857 LeCun, Y., Jackel, L. D., Boser, B., Denker, J. S., Graf, H. P., Guyon, I., Henderson, D., Howard, R. E., and Hubbard, W.:  
858 Handwritten digit recognition: Applications of neural network chips and automatic learning, *IEEE Communications Magazine*,  
859 Volume 27, Issue 11, 41-46, <https://doi.org/10.1109/35.41400>, 1989.

860

861 LeCun, Y., and Bengio, Y.: Convolutional networks for images, speech, and time series, *The handbook of brain theory and neural*  
862 *networks*, 3361, 10, 1995.

863

864 Lemaitre, G., Nogueira, F., and Aridas, C., K.: Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets  
865 in Machine Learning, *Journal of Machine Learning Research*, 18, 1-5, <http://jmlr.org/papers/v18/16-365.html>, 2017.

866

867 Luo, H., Quaas, J., and Han, Y.: Examining cloud vertical structure and radiative effects from satellite retrievals and evaluation of  
868 CMIP6 scenarios, *Atmos. Chem. Phys.*, 23, 8169–8186, <https://doi.org/10.5194/acp-23-8169-2023>, 2023.

869

870 Luo, H., Quaas, J., and Han, J.: Diurnally asymmetric cloud cover trends amplify greenhouse warming, *Science Advances*, 10,  
871 25, <https://doi.org/10.1126/sciadv.ado5179>, 2024.

872

873 Lenhardt, J., Quaas, J., and Sejdinovic, D.: Marine cloud base height retrieval from MODIS cloud properties using machine  
874 learning, *Atmos. Meas. Tech.*, 17, 5655–5677, <https://doi.org/10.5194/amt-17-5655-2024>, 2024a.

875

876 Lenhardt, J., Quaas, J., Sejdinovic, D., and Klocke, D.: CloudViT - Method code and data for the article "CloudViT: classifying  
877 cloud types in global satellite data and in kilometre-resolution simulations using vision transformers.", Zenodo,  
878 <https://doi.org/10.5281/zenodo.12731288>, 2024b.

879

880 Met Office: LAND SYNOP reports from land stations collected by the Met Office MetDB System, NCAS British Atmospheric  
881 Data Centre, <https://catalogue.ceda.ac.uk/uuid/9f80d42106ba708f92ada730ba321831>, 2008.

882

883 Met Office: MIDAS: Global Marine Meteorological Observations Data, NCAS British Atmospheric Data Centre,  
884 <https://catalogue.ceda.ac.uk/uuid/77910bcec71c820d4c92f40d3ed3f249>, 2006.

885

886 Oreopoulos, L., Cho, N., and Lee, D.: New insights about cloud vertical structure from CloudSat and CALIPSO observations, *J.*  
887 *Geophys. Res.-Atmos.*, 122, 9280–9300, <https://doi.org/10.1002/2017JD026629>, 2017.

888

889 Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison,  
890 A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S.:  
891 PyTorch: An Imperative Style, High-Performance Deep Learning Library, in *Advances in Neural Information Processing*  
892 *Systems* 32 (NeurIPS), 8024–8035,  
893 <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>, 2019.

894

895 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg,  
896 V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in  
897 Python, *Journal of Machine Learning Research*, 12, 2825-2830, <https://www.jmlr.org/papers/v12/pedregosa11a.html>, 2011.

898

899 Pincus, R., Hubanks, P. A., Platnick, S., Meyer, K., Holz, R. E., Botambekov, D., and Wall, C. J.: Updated observations of clouds  
900 by MODIS for global model assessment, *Earth Syst. Sci. Data*, 15, 2483–2497, <https://doi.org/10.5194/essd-15-2483-2023>, 2023.

901

902 Platnick, S., Ackerman, S. A., King, M. D., Meyer, K., Menzel, W. P., Holz, R. E., Baum, B. A., and Yang, P.: MODIS  
903 atmosphere L2 cloud product (06\_L2), NASA MODIS Adaptive Processing System, Goddard Space Flight Center,  
904 [http://doi.org/10.5067/MODIS/MYD06\\_L2.061](http://doi.org/10.5067/MODIS/MYD06_L2.061), 2017.

905

906 Platnick, S., King, M.D., Ackerman, S.A., Menzel, W.P., Baum, B.A., Riedi, J.C., and Frey, R.A.: The MODIS cloud products:  
907 algorithms and examples from Terra, in: *IEEE Transactions on Geoscience and Remote Sensing*, Volume 41, Number 2, 459-473,  
908 <http://doi.org/10.1109/TGRS.2002.808301>, 2003.

909

910 Ramanathan, V., Cess, R. D., Harrison, E. F., Minnis, P., Barkstrom, B. R., Ahmad, E., and Hartmann, D.: Cloud Radiative  
911 Forcing and Climate: Results from the Earth Radiation Budget Experiment, *Science*, 243, 57–63,  
912 <https://doi.org/10.1126/science.243.4887.57>, 1989.

913

914 Rasp, S., Schulz, H., Bony, S., and Stevens, B.: Combining Crowdsourcing and Deep Learning to Explore the Mesoscale  
915 Organization of Shallow Convection, *Bulletin of the American Meteorological Society*, 101, E1980-E1995,  
916 <https://doi.org/10.1175/BAMS-D-19-0324.1>, 2020.

917

918 Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N.,  
919 Hornegger, J., Wells, W., Frangi, A. (eds) *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*,  
920 *Lecture Notes in Computer Science*, Volume 9351, Springer, Cham., [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28), 2015.

921

922 Rossow, W.B., and Schiffer, R.A.: ISCCP cloud data products, *Bull. Amer. Meteorol. Soc.*, 71, 2-20, 1991.

923

924 Sassen, K., Wang, Z., and Liu, D.: Global distribution of cirrus clouds from CloudSat/Cloud-Aerosol Lidar and Infrared  
925 Pathfinder Satellite Observations (CALIPSO) measurements, *J. Geophys. Res.*, Volume 113, D00A12,  
926 <https://doi.org/10.1029/2008JD009972>, 2008.

927  
928 Slingo, A.: Sensitivity of the Earth's radiation budget to changes in low clouds, *Nature*, 343, 49–51  
929 <https://doi.org/10.1038/343049a0>, 1990.

930  
931 Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K.,  
932 Brokopf, R., Fast, I., Kinne, S., Kornblueh, L., Lohmann, U., Pincus, R., Reichler, T., and Roeckner, E.: Atmospheric component  
933 of the MPI-M Earth System Model: ECHAM6, *Journal of Advances in Modeling Earth Systems*, 5, 2, 146-172,  
934 <https://doi.org/10.1002/jame.20015>, 2013.

935  
936 Touvron, H., Vedaldi, A., Douze, M., and Jegou, H.: Fixing the train-test resolution discrepancy, 33rd Conference on Neural  
937 Information Processing Systems (NeurIPS 2019), Vancouver, Canada, <https://doi.org/10.48550/arXiv.1906.06423>, 2019.

938  
939 Tzallas, V., Hünerbein, A., Stengel, M., Meirink, J. F., Benas, N., Trentmann, J., Macke, A.: CRAAS: A European Cloud Regime  
940 dAtAset Based on the CLAAS-2.1 Climate Data Record, *Remote Sensing*, 14, 5548, <https://doi.org/10.3390/rs14215548>, 2022.

941  
942 Unglaub, C., Block, K., Mülmenstädt, J., Sourdeval, O., and Quaas, J.: A new classification of satellite-derived liquid water  
943 cloud regimes at cloud scale, *Atmos. Chem. Phys.*, 20, 2407–2418, <https://doi.org/10.5194/acp-20-2407-2020>, 2020.

944  
945 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I.: Attention Is All You  
946 Need, arXiv preprint, <https://doi.org/10.48550/ARXIV.1706.03762>, 2017.

947  
948 WMO: Manual on the observation of clouds and other meteors - International Cloud Atlas Volume I (WMO-No. 407), available  
949 at: [https://cloudatlas.wmo.int/docs/wmo\\_407\\_en-v1.pdf](https://cloudatlas.wmo.int/docs/wmo_407_en-v1.pdf) (last access: 25 February 2025), 1975.

950  
951 WMO: Manual on the observation of clouds and other meteors - International Cloud Atlas (WMO-No. 407), available at:  
952 <https://cloudatlas.wmo.int> (last access: 25 February 2025), 2017.

953  
954 WMO: Manual on Codes, Volume I.1 – International Codes, Annex II to the WMO Technical Regulations, Part A –  
955 Alphanumeric Codes (WMO-No. 306), ISBN: 978-92-63-10306-2, available at: <https://library.wmo.int/idurl/4/35713>, 2019.

956  
957 Wood, R.: Stratocumulus clouds, *Monthly Weather Review*, 140, 8, 2373–2423, <https://doi.org/10.1175/MWR-D-11-00121.1>,  
958 2012.

959  
960 Zantedeschi, V., Falasca, F., Douglas, A., Strange, R., Kusner, M. J., and Watson-Parris, D.: Cumulo: A Dataset for Learning  
961 Cloud Classes, Tackling Climate Change with Machine Learning Workshop, 33rd Conference on Neural Information Processing  
962 Systems (NeurIPS 2019), Vancouver, Canada, <https://doi.org/10.48550/arXiv.1911.04227>, 2019.

963  
964 Zhang, J. L., Liu, P., Zhang, F., & Song, Q. Q.: CloudNet: Ground-based cloud classification with deep convolutional neural  
965 network, *Geophysical Research Letters*, 45, 8665–8672, <https://doi.org/10.1029/2018GL077787>, 2018.

966  
967 Zhao, H., Gallo, O., Frosio, I., and Kautz, J.: Loss functions for image restoration with neural networks, *IEEE Transactions on*  
968 *computational imaging*, 3, 1, 47–57, <https://doi.org/10.1109/TCI.2016.2644865>, 2016.