

1 **CloudViT: exploring cloud type classification with vision transformers**  
2 ~~classifying cloud types in global satellite data and in~~  
3 ~~kilometre-resolution simulations using vision transformers.~~

4

5 Julien Lenhardt <sup>1</sup>, Johannes Quaas <sup>1,2</sup>, Dino Sejdinovic <sup>3</sup>, Daniel Klocke <sup>4</sup>

6

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

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

9 Leipzig, Germany

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

11 Australia

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

13 *Correspondence to:* Julien Lenhardt (julien.lenhardt@uni-leipzig.de)

## 14 Abstract

15

16 Clouds constitute, through their interactions with incoming solar radiation and outgoing terrestrial radiation, a fundamental  
17 element of the Earth's climate system. Different cloud types show a variety in cloud microphysical or optical properties, phase,  
18 or vertical extent, and thus disparate radiative effects. Both in observational and model datasets, classifying clouds is important  
19 since different cloud types respond differently to current and future anthropogenic climate change. Cloud types have traditionally  
20 been defined using a simplified partition of ~~a two-dimensional space, e.g.,~~ cloud top pressure and optical thickness, ~~but recently~~  
21 ~~using deep learning.~~ In this study, we present a method called CloudViT (Cloud Vision Transformer) building on surface  
22 observations and spatial extracts of cloud properties from the MODIS instrument to derive cloud types, leveraging spatial  
23 patterns with a vision transformer model. The performance of the model is fair and hampered by the limited number of samples  
24 and the challenging matching between data sources arising during the collocation process. The method is then evaluated through  
25 the distributions of cloud type properties and global spatial patterns of cloud type occurrences. ~~Potential improvements emerge in~~  
26 ~~the reduction in mismatches between data sources, the extension of the colocated dataset, and the refinement of the classification~~  
27 ~~model. CloudViT is applied to data from a global storm-resolving model, showcasing the feasibility of the transfer to model~~  
28 ~~outputs.~~ While the application of the method in its current state comes with apparent uncertainties due to limited performance,  
29 ~~improvements~~ it raises relevant challenges and limitations, from which the community can benefit from discussing for the  
30 ~~development of similar methods to mitigate that~~ ~~Potential improvements emerge in the reduction in mismatches between data~~  
31 ~~sources, the extension of the colocated dataset, and the refinement of the classification model.~~ To foster future CloudViT's  
32 advancements, the dataset and model are available from Zenodo (Lenhardt et al., 2024b).

33

## 34 1 Introduction

35

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

55 Traditional cloud classification methods are built on simple characteristics. The standard classification developed as part of the  
56 International Satellite Cloud Climatology Project (ISCCP) relies on three levels (low, medium, high) of cloud altitude using as  
57 proxy the cloud top pressure (CTP) and three thresholds of cloud optical thickness (COT), defining overall nine cloud types  
58 (Rossow et al., 1991). This classification is performed on scalar fields, setting aside any spatial pattern in the cloud field from  
59 which information could be obtained to better inform the classification process. Relying on the same type of two-dimensional  
60 histograms, recent methods have been developed aiming at refining the created clusters and partially relaxing the constraints on  
61 the pre-defined thresholds (Tzallas et al., 2022). The reason to choose the two parameters is that such a classification lends itself  
62 to the analysis of cloud radiative effects: the cloud radiative effect in the solar is a monotonic function of COT, the one in the  
63 terrestrial spectrum, of CTP. However, one might be interested in sensitivities of cloud thickness or water content to different  
64 drivers (e.g., aerosols) for given cloud types, which is hampered by using CTP and COT to define the types. Also, COT does not  
65 map well onto the distinction between cumuliform and stratiform clouds. For such reasons, Unglaub et al. (2020) defined cloud  
66 regimes from cloud base height and variability in cloud top height, hinting at the added value of some measure of spatial  
67 variability and pattern. However, to leverage spatial structure and textures, cloud classification methods based on artificial  
68 intelligence (AI) have opened new avenues of research built upon vast amounts of remote sensing data. For example, using  
69 convolutional neural networks (CNNs; LeCun et al., 1989; LeCun et al., 1995), Zhang et al. (2018) use ground-based images and  
70 human-labelled cloud types to develop a model for meteorological cloud classification and support weather prediction tasks.  
71 Using a similar architecture, Rasp et al. (2020) classify clouds from expert-labelled satellite images of four different cloud  
72 organisation patterns in the trades. This method further emphasises how expert knowledge to identify cloud patterns can be  
73 learned by CNN models and allow to then better constrain radiative effects of mesoscale convection (Wood and Hartmann, 2006;  
74 Bony et al., 2020; Stevens et al., 2020) which would prove to be too cumbersome manually. The application of deep learning to  
75 the classification of mesoscale cloud patterns in particular (Muhlbauer et al., 2014; Yuan et al. 2020; McCoy et al., 2023)  
76 additionally demonstrates how specific cloud organization patterns, observable by experts in satellite data, can be learned by  
77 machine learning models, and allows a deeper analysis of their radiative effects and characteristics on longer time periods and  
78 larger spatial scales. These studies rely on human observers to initially classify clouds or cloud patterns directly from images,  
79 relying on visual aspects to distinguish clouds, and subsequently linking the identified cloud types to local meteorological  
80 conditions. Kuma et al. (2023) also capitalize on ground-based observations but connect them to shortwave and longwave  
81 radiation satellite retrievals at coarser spatial and temporal resolutions. The method relies on identifying patterns directly in  
82 radiation retrievals to associate them to daily occurrence probabilities of cloud types. This method has the benefit of being able to  
83 be used on outputs from large ensembles of global model simulations and reanalysis datasets which cover extended time-scales  
84 compared to observational datasets. Relying on similar model architectures, Zantedeschi et al. (2019) and Kaps et al. (2023)  
85 classify cloud types derived from active remote sensing labels. The study from Kaps et al. (2023) capitalizes on the model from

86 Zantedeschi et al. (2019) to extrapolate cloud type estimates using global passive remote sensing data, and jointly trains a model  
87 on coarsened data with spatial resolution similar to current global climate models. Other methods have been developed without  
88 the use of cloud type labels, drawing conclusions from clusters appearing in large remote sensing radiation retrievals (Kurihana  
89 et al., 2022). In general, the developed methods rely on identifying characteristic patterns arising in images (related to visible  
90 features of cloud types), radiation retrievals (related to radiative properties of cloud types), or cloud properties retrievals (related  
91 to physical properties of cloud types). Each choice of cloud type labels introduces a certain level of subjectivity in the derived  
92 cloud types. For example, there is less subjectivity in the expert-labelled images than in the produced cloud clusters, which  
93 naturally introduces some subsequent biases. Choosing certain input quantities also physically constrains the variability of cloud  
94 type properties which can hinder the interpretation of the derived cloud type estimates. However, the transferability to global  
95 climate model outputs is a great advantage of some of these methods as they provide a crucial way to diagnose the representation  
96 of clouds in climate models and push towards reducing uncertainties in representing future-climate clouds (Kuma et al. 2023;  
97 Kaps et al. 2023).

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

114 Firstly, we introduce in section 2 the different datasets used in the study alongside the collocation process between the  
115 ground-based and satellite datasets. Subsequently, the different components of the CloudViT method are presented in section 3,  
116 supported by sensitivity studies about the generalisation skill of the models and the benefits of the spatial context. In section 4,  
117 we evaluate the method and investigate the distribution of cloud properties following the predicted cloud types. The results in  
118 section 5 focus on the extension to a global distribution of cloud types ~~and present a first application to climate model data.~~  
119 ~~Challenges, limitations, and lessons learned from CloudViT’s development are highlighted in the following section 6, with the~~  
120 ~~guiding idea of making cloud type classification with vision transformers reliable, capable of achieving notable performance, and~~  
121 ~~potentially applicable to high-resolution climate model simulations. Eventually, we conclude over discuss the benefits of the~~  
122 ~~presented method and challenges of cloud type classification, the potential improvements, and the remaining challenges to make~~  
123 ~~CloudViT’s usage reliable and capable of achieving notable performance on the classification task.~~

## 124 2 Data

125

### 126 2.1 Surface observations

127

128 The cloud type observations used in this study come from two similar global observation datasets maintained by the UK Met  
129 Office, one providing observations made at sea (Met Office, 2006) and the second providing observations made on land (Met  
130 Office, 2008). These observations are performed from weather stations (land or sea) or ships, by trained observers following the  
131 WMO code tables (WMO, 2019). Each cloud level (high, WMO code table 0509; medium, WMO code table 0515; low, WMO  
132 code table 0513; see Table A.1) is separated in 9 different types describing in more detail the aspect and type of the observed  
133 clouds. The labels thus provide a high level of detail regarding the observed cloud scene from the surface. Naturally, the case of  
134 multilayer clouds poses a problem since the field of view and the visibility from the surface are limited, which is why we remove  
135 the potential multilayered cases from the training dataset to focus only on single-layer observed cloud scenes. It induces potential  
136 selection bias issues as some cloud types might more likely be observed in multilayered configurations. The relative amounts of

137 each cloud type before and after the filtering and colocation process are displayed in Figure 2. Similarly, uncertainty is greater for  
138 medium and high clouds as their observation can be more challenging than for low clouds. Furthermore, the spatial distribution  
139 of the labels (Fig. 1, Fig. A.1) can be problematic as the marine observations are distributed mainly along ship routes. On the  
140 other hand, combining that with land observations provides a more complete representation of cloud types, especially for high  
141 level ones, all the while introducing the influence of orography. Other studies like Kuma et al. (2023) and Lenhardt et al. (2024a)  
142 have built estimates of cloud quantities based on these ground-based observation datasets, overcoming limitations pertaining to  
143 incomplete field of view and disparate spatial distribution.

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

150

## 151 2.2 Satellite retrievals

152

153 In addition to the surface observations, we use satellite retrievals from MODIS, in particular from the AQUA satellite. MODIS  
154 retrievals offer a vast amount of data at kilometre-scale resolution with daily overpasses. Each of the supplied granule file  
155 contains cloud microphysical and optical properties across a region with a span of around 2330 km x 2000 km. We make use of  
156 the available CUMULO dataset (Zantedeschi et al., 2019) since it allows access to preprocessed MODIS level 2 satellite data,  
157 with global coverage, and for two full years (2008 and 2016). Among the data variables available, we rely on two unified  
158 products (cf. Table 1) describing either cloud properties (MODIS06 level 2 cloud product, hereafter MYD06; Platnick et al.,  
159 2017) or the cloud cover (MODIS35 level 2 cloud flag mask, hereafter MYD35; Ackerman et al., 2017). The latter's main usage  
160 is to help screen for cloud scenes with a minimum cloud coverage.

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

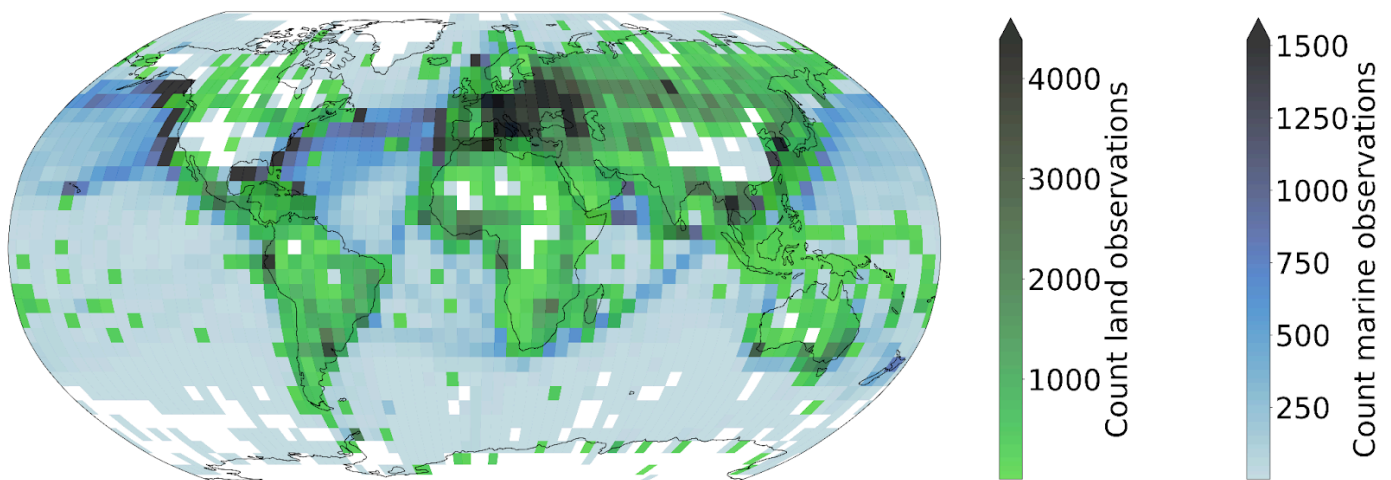
176 As a whole, the MYD06 product has the advantage that, building directly on cloud properties, we can design a classification  
177 model from which the relationship between cloud type and other cloud properties can then be examined. Relying on calibrated  
178 radiances which lie ahead in the retrieval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

184 Alongside the collocated dataset, we build a collection of randomly sampled tiles out of the satellite retrievals from the year 2008.  
185 For each granule, a maximum of 20 tiles are sampled while ensuring the amount of missing data stays limited. This process leads  
186 to the compilation of more than 1.3M single tiles of cloud properties. These tiles are then randomly split temporally into training  
187 (70%), validation (10%) and test (20%) sets. This dataset is the basis for the self-supervised training procedure presented in the  
188 following section.

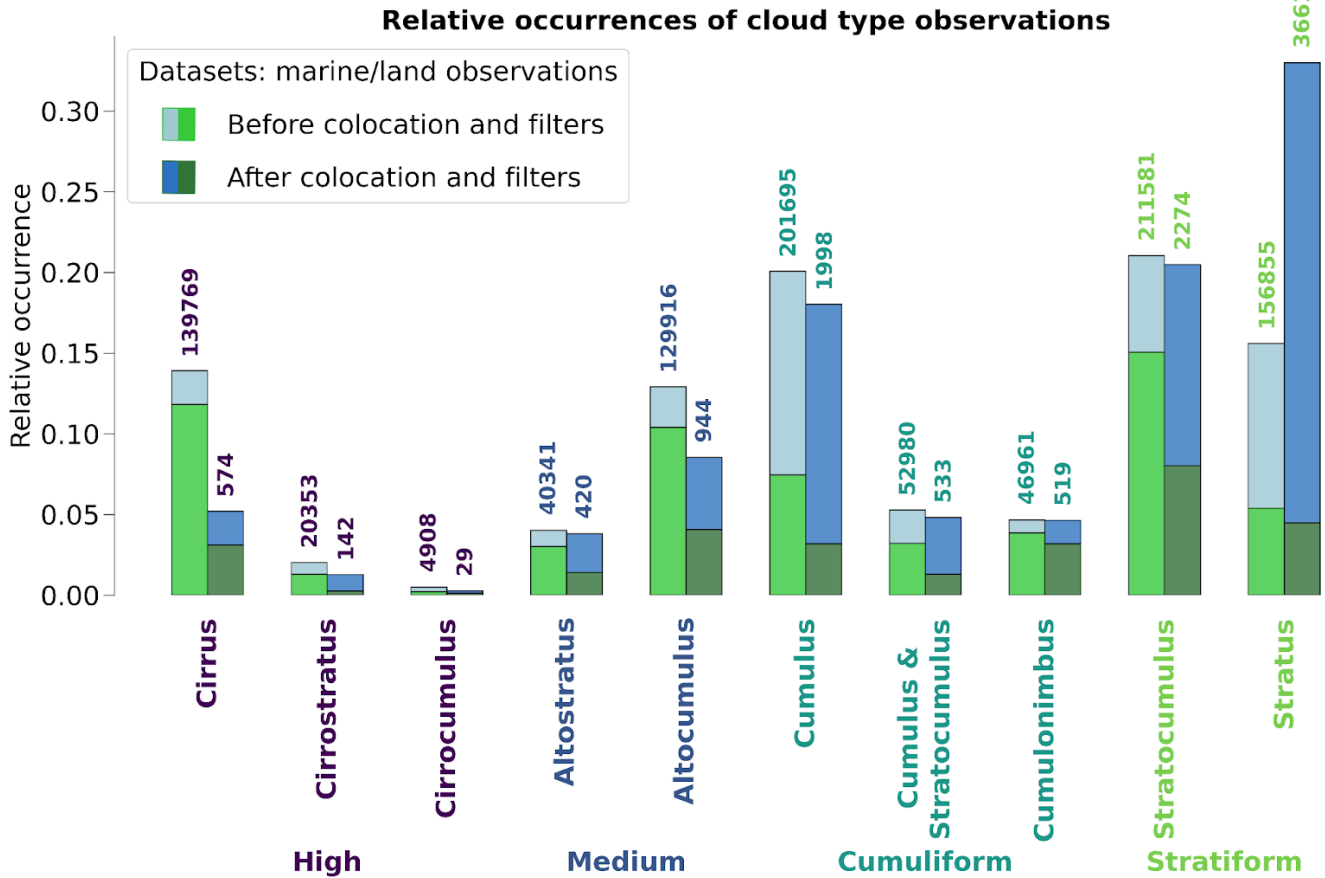
Data product	Description	Variables	Resolution	Usage
Global marine meteorological observations (Met Office, 2006)	Marine surface observations	Cloud type	Latitude/longitude coordinates 0.1° Hourly/daily observations	Labels
Land SYNOP reports (Met Office, 2008)	Land surface observations	Cloud type	Latitude/longitude 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

190 **Table 1 : Datasets description. The surface observations are provided by a worldwide station network available from the**  
 191 **UK MetOffice (Met Office, 2006; Met Office, 2008; see section 2.1). The MODIS data are derived from the collection 6.1**  
 192 **of the datasets (Ackerman, 2017; Platnick et al., 2017; see section 2.2).**  
 193

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



194 **Figure 1: Spatial distribution of cloud type observations for marine (years 2008 and 2016; Met Office, 2006) and land**  
 195 **(year 2016; Met Office, 2008). The corresponding spatial distributions of cloud type observations are included in Figures**  
 196 **A.1 and A.2, for before and after the colocation process, respectively.**  
 197  
 198



199  
 200 **Figure 2: Relative occurrences of cloud types before and after the collocation and filtering process, indicated for both the**  
 201 **marine (blue; Met Office, 2006) and land (green; Met Office, 2008) observational datasets. The x axis corresponds to the**  
 202 **cloud types in the case of 4 and 10 categories. The corresponding numbers of colocated samples for each cloud type are**  
 203 **detailed in Table A.1.**

204 **3 Method**

205

206 **3.1 Method outline**

207

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

215 For the self-supervised task, we train two models to reconstruct 3D data cubes of cloud properties. The first model, which is used  
 216 as a baseline, is a CNN backbone we previously presented in Lenhardt et al. (2024a) to handle satellite retrievals of cloud  
 217 properties for cloud base height prediction. The second model we develop in this study is based on vision transformers  
 218 (Dosovitskiy et al., 2020), a recent type of model compared to the more typical CNNs for computer vision applications. The  
 219 spatial pattern of the cloud properties and their scale provide information about clouds, which can be leveraged to classify them  
 220 for example into more stratiform and more cumuliform types. During the training phase of these models, the samples are images  
 221 of size 128x128 pixels consisting of three different cloud properties: CTH, COT and CWP. We ensure that the models learn to  
 222 distinguish cloud patterns and not to recognise specific geographical locations by extracting samples randomly across global  
 223 satellite retrievals from the year 2008, without adding information about their location. In a second step, a classification model is  
 224 trained on the colocated samples of cloud properties and surface observations. As mentioned in section 2.1, the number of types

225 reported in the observations for clouds is reduced to either 4 or 10 classes (Kuma et al. 2023). The training process follows a  
 226 supervised learning framework, where the classification model outputs a single cloud type (among the 4 or 10 cloud types) for  
 227 the whole extent of the input cloud scene of size 128x128 pixels. The benefit of the presented method using either a CNN or a  
 228 vision transformer, which are models incorporating a certain level of spatial awareness, is that it is consistent with the cloud type  
 229 identified by the human observer. Furthermore, in comparison to conventional methods like the ISCCP, the method benefits from  
 230 a potential ability to distinguish cloud types without using predefined thresholds.

231

### 232 3.2 Vision transformer

233

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

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

$$260 \quad L_r = \frac{1}{N} \sum_{i=1}^N \left\| x_i - D_{\theta}(E_{\theta}(x_i^c)) \right\| \quad (1)$$

261 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  $E_{\theta}$   
 262 are namely the decoder and encoder parts of the model with  $\theta$  designating their learnable parameters.

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

$$271 \quad L_c = -\frac{1}{N} \sum_{i=1}^N \log l_c(x_i, x_i^a, E_{\theta}, D_{\theta}) \quad (2)$$

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

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

$$279 \quad L = \alpha \times L_r + L_c \quad (3)$$

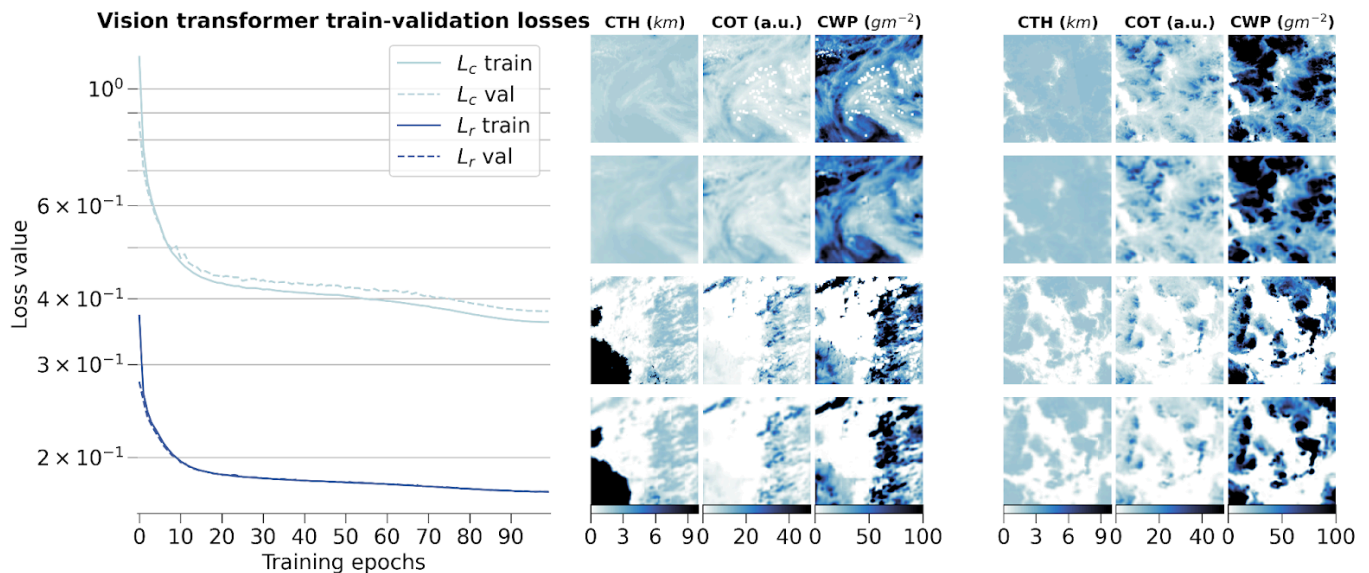
280 where  $\alpha$  is a scaling factor between the two tasks. We follow the recommendation of Atito et al. (2023) to set  $\alpha = 5$  in the case  
281 of small-scale datasets so that the vision transformer can learn enough of the local inductive bias.

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

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

313 Eventually, we decide to use as model the small variant of the vision transformer with a scaling factor  $\alpha$  of 5, as it showcases  
314 good performance in both tasks during the training while having a number of parameters four times smaller than the base variant.  
315 Furthermore, the self-supervised training task on the large unlabelled dataset allows the model to have plenty of data to learn  
316 from, the pre-trained model weights giving only marginal gain for a few epochs at the start. The small variant of the vision  
317 transformer was shown to perform very well on a large variety of tasks as per the results from Atito et al. (2023). The results  
318 across the training, validation and test datasets are shown in Figure 3 for the training process and some examples of reconstructed  
319 samples belonging to all three splits, while Figure 4 highlights the spatial distribution of the reconstruction error per channel and  
320 across splits. On the left panel of Figure 3, the losses show a consistent decreasing trend even at the end of the training epochs.  
321 The training process was halted after 100 epochs due to computational limitations, but would gain to be extended as the vision  
322 transformer’s performance seems to still be improvable. On the right panel of Figure 3, the reconstructions presented for some

323 random samples reveal where the model would benefit from an improved performance: the reconstructions appear realistic, but  
 324 fail to reproduce the exact sharpness that is visible in the satellite retrievals. While this aspect would not guarantee a decisive  
 325 improvement in the downstream task which only relies on the encodings, it would greatly help build more trust in the model. A  
 326 related case that can lead to observed patterns of reconstruction errors in Figure 4 lies in the reconstruction of cloud scenes with  
 327 convective cells. The invigorated core of the convective cell stands much higher and holds more water compared to its  
 328 surroundings which can lead to steep gradients in the cloud quantities when observed from space. As the reconstructions are not  
 329 able to reproduce these features, larger errors can arise from such cloud scenes. This could further propagate to the classification  
 330 performance on the related classes, e.g. mesoscale convection clouds or cumulonimbus, whose intricate patterns are better  
 331 assessed on their own directly (Bony et al., 2020; Rasp et al., 2020; Stevens et al., 2020; Yuan et al. 2020; McCoy et al., 2023).  
 332 The additional patterns in the reconstruction error of Figure 4, in particular for COT, are visible in some consistent areas over  
 333 land. A deeper analysis of the spatial generalisation skill of the model than the one presented in section 3.3.2 covering only the  
 334 colocated dataset might help constrain the spatial generalisation performance of the vision transformer and infer potential  
 335 performance caveats still remaining.  
 336 Ultimately, we can compare the skill of the vision transformer to that of the baseline CNN autoencoder from Lenhardt et al.  
 337 (2024a). The CNN autoencoder was trained using as reconstruction error the mean squared error (MSE) on similar MODIS data  
 338 but only with MODIS granules over the ocean. It was shown to perform similarly with a slightly higher error over land when  
 339 evaluated over a global dataset. The vision transformer model outperforms the CNN autoencoder on all metrics (MSE and  
 340 l1-loss) across all data splits (training, validation and test), displaying consistently across data splits on average an MSE of 0.15  
 341 and a l1-loss of 0.12 compared to 0.3 for both metrics for the CNN. Examples of reconstructed samples additionally show how  
 342 the l1-loss helps produce sharper edges in the reconstruction, a well-known issue with the application of MSE as target metric in  
 343 computer vision (Zhao et al., 2016). The contribution to the error comes mostly from the COT channel for both models and the  
 344 error is concentrated in areas of higher variability for the respective channels. The metrics values are summarised in Table B.1.  
 345 The spatial generalisation skill, alongside the sensitivity to the tile size and the impact of data augmentation on the performance  
 346 on the cloud classification task are analysed in the following section.



347  
 348 **Figure 3: (left) Training and validation losses during model optimization for the small variant of the vision transformer**  
 349 **on the global training dataset. (right) Examples of tiles (first and third rows) with the corresponding reconstructions**  
 350 **(second and fourth rows) for the different cloud property channels.**

351  
 352 **3.3 Cloud type classification**  
 353

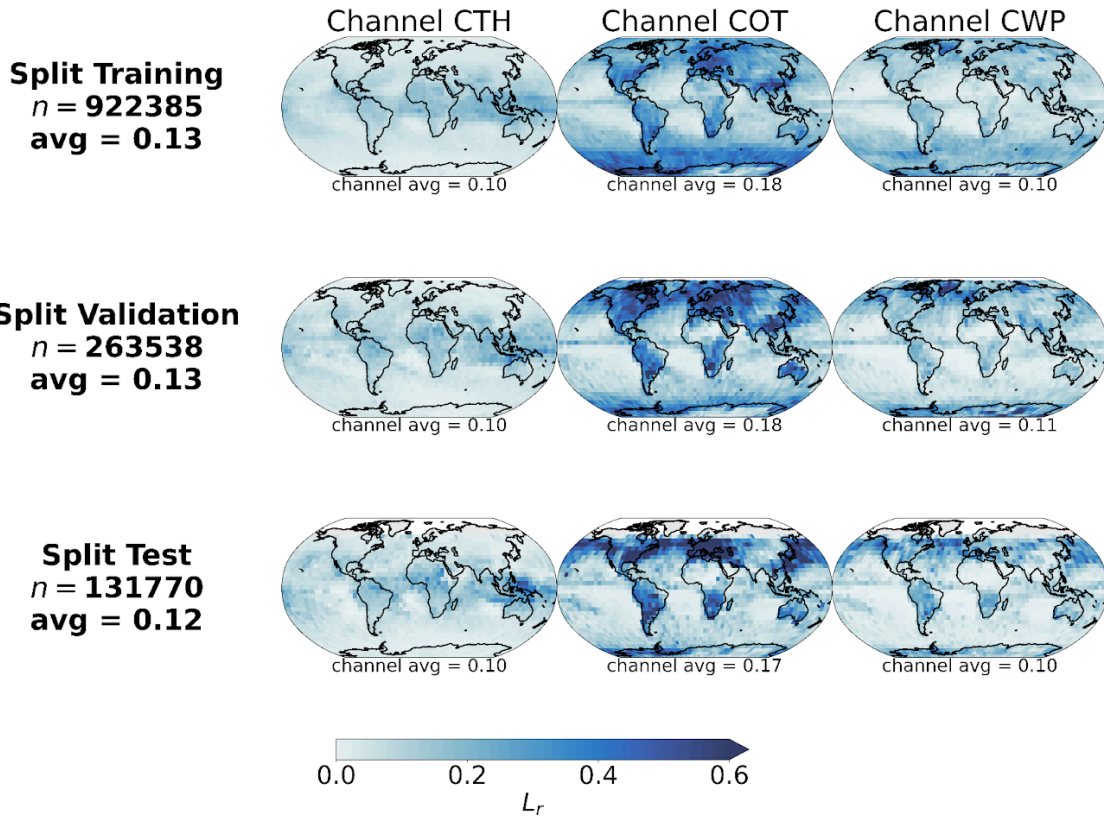
354 The next task at hand is the cloud type classification, building on the colocated samples of satellite retrievals and surface  
 355 observations. For the two years of MODIS AQUA data available, out of 104 823 colocated samples we retain only 11 094 for our  
 356 training and testing datasets after filtering, among others, for missing data - typically 50% of the samples are discarded, mainly  
 357 when the colocated observation lies on the edges of the satellite granule - and single layer cloud observations as reported by the  
 358 observer - around 60% of the previously filtered samples are kept. A main caveat arising from colocating these two data sources  
 359 is the potential mismatch between the actual clouds jointly depicted. Contrarily to methods like Zantedeschi et al. (2019) which

360 relies on joint retrievals of cloud properties and cloud type or Kuma et al. (2023) which aggregates observations at daily time  
361 scales, the presented colocated dataset leaves room for misaligned surface observations and satellite retrievals. As it will be also  
362 highlighted later on, this potential misalignment between data sources constitutes a hurdle in the development of the cloud  
363 classification method. Indeed, if the model needs to learn from satellite data that actually does not visibly fit the surface  
364 observation, then the learning process is hindered. Attempts to reduce this risk have not yielded satisfying results. For example,  
365 decreasing the time-window described in section 2.2 did not ultimately yield improvements in the classification performance,  
366 especially due to generalisation limitations from a lower number of samples. Furthermore, these attempts are mainly limited by  
367 the amount of satellite data that would be necessary to build a substantial and consistent colocated dataset which would span a  
368 larger timeframe than the two years used in this study. After the filtering of the colocated dataset, the cloud type observations are  
369 then regrouped into 4 or 10 types as mentioned previously. The rest of the study will focus on these categories as targets. From  
370 the latent space representations produced by the vision transformer or the CNN autoencoder, we build a classification model  
371 either by attaching a classification head to the encoder network or by using a simpler classification model like a random forest  
372 (RF; Breiman, 2001). To investigate the performance of the classification models on the two classification tasks at hand (4 and  
373 10 cloud types), we use different metrics tailored to unbalanced classification setups as the cloud types are not equally  
374 represented (see Figure 2 and Table A.1). A first method to assign similar weight to all classes regardless of the class' cardinality  
375 is to use macro-averaged metrics. In this framework, the metric of interest is averaged over the samples of each class separately  
376 before being averaged over the classes. This leads to a higher weight for minority classes for which the model might perform  
377 differently, usually worse, compared to the majority classes providing different information over traditional averaging strategies  
378 (micro-averaged for example) where the result will be dominated by the samples from the majority classes. We report several  
379 metrics adapted to an unbalanced setting: the index balanced accuracy (IBA; Garcia et al., 2012) of the geometric mean, the  
380 macro-averaged accuracy and the macro-averaged f1-score.

381 For the classification model we investigate two alternatives: a RF classification model (implementation from *Scikit-learn*  
382 package, Pedregosa et al., 2011) and a MLP classification head (Hinton, 1989; implemented in *PyTorch*, Paszke et al., 2019).  
383 However, a wider diversity of classification models could be implemented based on the backbone provided by the vision  
384 transformer. The base model architecture and weights are made available on Zenodo (Lenhardt et al., 2024b) to foster a more  
385 complete exploration of possibilities regarding the classification model, building on the backbone of the vision transformer. For  
386 example, with a more extensive training dataset, more complex classification models could be explored. In the case of the  
387 architecture presented here, the RF model provides simplicity in the implementation and the training process, while the MLP is  
388 the typical architecture used for the downstream task following a network like a vision transformer or a CNN. The RF model has  
389 10 or 25 trees, for the cases of 4 and 10 cloud types respectively, with a maximum depth of 5. Basic hyper-parameter  
390 optimization showed that with the reduced amount of samples and the limited variety in cloud scenes for some categories (even  
391 more with balanced classes, see section 3.3.3), models displaying limited complexity avoided overfitting and generalised better  
392 on unseen data. The MLP consists of two fully-connected layers (hidden dimension 4096) with a Gaussian Error Linear Unit  
393 (Hendrycks & Gimpel, 2016) in between and is trained using the cross-entropy loss. The sensitivity studies and experiments are  
394 done only using RF models but the evaluation in the subsequent section will be done on both types of classification methods.  
395 Various sensitivities could be explored in the presented setting but we here focus on the potential benefit of the spatial context,  
396 the ability to generalise spatially to unseen locations and the impact of balancing the labelled dataset.

397

## Spatial $L_r$ mean per channel



398

399

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

400

401

### 402 3.3.1 Spatial context and tile size

403 We look at the influence of the input size by training vision transformers (small variant) on different sizes of inputs namely  
404 128x128, 64x64, 32x32 and 16x16. We do not consider larger tile sizes as the cloud scene might then be less representative of the  
405 surface observation, especially since we only consider samples with single labels, and as the assumption that the observed cloud  
406 type occurs on such scales would likely not hold. The losses relative to the vision transformer models trained on the different  
407 input tile sizes are detailed in Figure B.2. Since these models were trained on a reduced dataset as mentioned previously, their  
408 skill cannot be directly compared to the one displayed in Figure 3. While the contrastive losses are similar across input tile sizes,  
409 the reconstruction losses differ. Since we kept the ratio between the patch size and the tile size constant when training the  
410 different models, the difference in reconstruction skill could be attributed to the dimensionality of each patch being much  
411 smaller, for example for a tile of size 16x16 a patch will be 2x2. The reconstruction head being a fairly shallow CNN, the  
412 reconstruction of the spatial patterns inside the patches showcases better skill for smaller input patches after a few epochs, while  
413 for larger patch sizes - and thus tile sizes - a longer training process would be needed as to improve the truthfulness of the  
414 reconstruction to the input. Examples of reconstructions depending on the input tile size are included in Figure B.3 and visually  
415 display how a larger field of view can help capture the larger cloud organisation or even individual sparse clouds. To further  
416 evaluate the potential benefit of the spatial context for the downstream classification task, we consider as an alternate input the  
417 flattened cloud properties of a 9x9 tile centred on the observation location. This yields an input of similar dimensionality  
418 compared to the latent space representation of both the CNN and the vision transformer (3 channels x 9 x 9 = 243). We then train  
419 the same RF classification model on each of the latent representations derived from the trained vision transformers and on the  
420 flattened cloud properties. From the classification metrics, we observe that the smaller the tile size the more prone the model is to  
421 overfitting towards the majority classes (high and stratiform cloud types in the case of 4 types) leading to a decreased  
422 performance on the validation set. For instance, choosing an input tile size of 16x16 results in a decrease of 20% across metrics  
423 from the training to the validation set (compared to around 10-15% across metrics for the larger input tile sizes), and leads to  
424 metrics on the validation set more than 10% lower than with larger input tile sizes. The predictions made using larger spatial  
425 context (tile size greater than 16x16) outperform the method with 9x9 flattened tile inputs across all considered metrics on the

426 validation set. With the input tile size 16x16, the reduced spatial context seems to be limiting for the performance but another  
427 explanation could be a complex latent space compared to the input dimensionality. Overall, even with the vision transformer  
428 backbones being trained only partially, the wider input tile size provides better classification skill and generalisation to unseen  
429 data. In the rest of the study and experiments, if not mentioned specifically, the input tile size is chosen to be 128x128.

430

### 431 3.3.2 Spatial generalisation

432 To investigate the spatial generalisation skill of the cloud classification method, we split our colocated dataset into samples  
433 located in the Northern or Southern hemispheres. Two vision transformer models are additionally trained on samples from only  
434 the respective hemisphere and tested on the other one. The losses relative to the training and testing of both hemispherical  
435 models are included in Figure B.4. Both hemispherical models display similar performance both on the training and testing  
436 datasets, showing that even for a reduced number of training samples, epochs and spatial coverage the vision transformer  
437 architecture generalises well to unseen data. Building on the two trained vision transformers, we set out to evaluate the skill on  
438 the classification tasks. Splitting the labels between the two hemispheres yields 9246 samples for the Northern hemisphere and  
439 1848 samples for the Southern hemisphere. Investigating the different classification metrics for training and testing on both  
440 hemispheres, it is clear that the classification model trained on the Southern hemisphere struggles to generalise from such a low  
441 number of labelled samples and probably overfits since the performance is worsened on the Northern hemisphere samples  
442 (decrease of almost 50% across metrics from the training to the testing set). The classification model trained on the Northern  
443 hemisphere generalises well in the case of the 4 cloud types with consistent metric values between hemispheres (marginal  
444 decrease of around 15% across metrics from the training to the testing set). Overall, the model trained on samples from the  
445 Northern hemisphere and for both cases of number of cloud types, the performance on the Southern hemisphere is similar to  
446 models with larger tile sizes from the previous section, showing consistency across experiments even with limited datasets for  
447 the training of the vision transformer.

448

### 449 3.3.3 Balanced training dataset

450 Balancing the number of samples among classes in the input dataset can be a way to leverage enough information from the  
451 underrepresented classes. We compare here the performance skill of two classification models trained on the colocated dataset or  
452 on a balanced equivalent. To this extent, we use a sampler implementation from the *imbalanced-learn* package (Lemaitre et al.,  
453 2017), namely the Synthetic Minority Oversampling Technique (SMOTE; Chawla et al., 2002) to oversample the minority  
454 classes. Doing so leads to improved classification skill with consistent increases across metrics on the validation set of 3-7% and  
455 15-35% for the cases of 4 or 10 cloud types, respectively. The oversampling impacts mostly the cloud types from the high and  
456 medium classes, and from the cirrocumulus and cirrostratus classes, in the case of 4 cloud types and 10 cloud types, respectively  
457 (see Table A.1). The methods evaluated in the following section will thus include the same over-sampling strategy to overcome  
458 the representation of the minority classes and improve the performance on the classification task.

## 459 4 Evaluation

460

### 461 4.1 Classification evaluation

462

463 In the following section, we detail the classification performance on the test set of the previously mentioned models. Two  
464 baseline models are included, namely a classification model built on the CNN autoencoder from Lenhardt et al. (2024a) and a RF  
465 model built on the flattened 9x9 input tiles as described in section 3.3.1. The method developed in this study is represented by  
466 two models using the aforementioned vision transformer model (see section 3.2) as backbone complemented by either a RF  
467 classifier or a MLP (see section 3.3). In the rest of the study, we denote the trained vision transformer model followed by the  
468 classification model as CloudViT (Cloud Vision Transformer) in its two classification variants (RF or MLP). The classification  
469 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  
470 the 10 cloud types. Since the number of samples is very limited, the performance of the models cannot be only considered as is  
471 but is further evaluated in the subsequent sections through distributions of cloud properties and spatial occurrence distributions.  
472 We emphasize here the need to perform an evaluation beyond the metrics to assess the skill of the model to represent  
473 characteristics expected from different cloud types. These characteristics can relate to the distribution of their physical  
474 parameters and their occurrences, both of which can be assessed thoroughly here only with a more extensive dataset. The  
475 CloudViT/RF method performs the best across all of the three metrics included, despite showing still limited performance overall.  
476 Firstly, the macro-averaged multi-class accuracy does not differ by a large margin between the different methods, but the

477 class-wise accuracies reveal several limitations. The baseline 9x9 RF model largely overfits towards the high and stratiform types  
478 (train and test class accuracies of 0.84/0.81 and 0.63/0.62, respectively), performing poorly on the medium and cumuliform types  
479 (train and test class accuracies of 0.31/0.21 and 0.19/0.15, respectively). The CloudViT/MLP model is biased towards stratiform  
480 clouds (train and test class accuracy of 0.79/0.79) while struggling to identify the other three types (train and test accuracies all  
481 falling between 0.10 and 0.40). The baseline CNN/RF and the CloudViT/RF models are performing quite similarly both on  
482 aggregated and class-wise metrics. However, the CloudViT/RF model showcases improved performance on the stratiform class  
483 (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  
484 accuracies for medium and cumuliform clouds. The performance on the high clouds is similar with slightly higher accuracies for  
485 the CloudViT/RF model. Other metrics like the IBA of the geometric mean and the F1-score further emphasise that the  
486 CloudViT/RF model outperforms the other methods while addressing the imbalance training data to generalise with satisfactory  
487 skill on the unseen test dataset. Nevertheless, the performance detailed here across classes shows apparent limitations as scores  
488 are not ideal. An obvious hurdle of the learning process resides in the overall limited number of samples and the noise present in  
489 particular for cloud types with minimal numbers of samples. Building a dataset with more labels would improve the  
490 classification performance by allowing the classification to more easily converge towards each cloud type’s mean state arising  
491 from a larger number of samples. The simplicity of the classification models chosen here represents a constraint that could be  
492 lifted if more training samples were available as overfitting and balance would then represent lesser issues. Furthermore, the  
493 patterns in the class accuracies can be traced back to shortcomings in the observational dataset. Having only considered  
494 single-layer cloud scenes in the colocated dataset, the high clouds are well predicted in accordance with the observations as a  
495 surface observer would identify with certainty this type of cloud if no other lower cloud is blocking the field of view from the  
496 surface. Stratiform clouds could be more challenging for the observers as they typically display high cloud fraction and high  
497 optical thickness, limiting the ability of the surface observer to quantify with certainty the amount of clouds in other levels.  
498 However, such characteristics can be well captured by computer vision models which build on patterns in the three-dimensional  
499 input data which in particular the baseline 9x9 RF model lacks. This difference between models is in particular apparent for the  
500 cumuliform class which is mostly composed of observations of cumulus. A cloud scene relative to a cumulus observation will  
501 most likely display a lower cloud fraction as the individual clouds are sparsely distributed, extracting only the very near points  
502 around the observation might then be too reductive and limit the accuracy of the classification model. It is confirmed by the  
503 accuracy on this cloud type for which the baseline 9x9 RF model is largely outperformed by all three other models both on  
504 training and test datasets (class accuracy increases between 150% up to 260% on the test dataset). Overall, the classification  
505 model shows fair performance that could be probably improved by widening the scope of the cumbersome collocation process  
506 which requires large amounts of remote sensing data, and by accordingly refining the RF or MLP architectures presented here.  
507 Using the classification model developed here thus comes with apparent uncertainties across the different cloud types. Efforts  
508 were made with the aim to classify all cloud types consistently from the limited training dataset available but to limited  
509 outcomes. The extension of the training dataset appears as an obvious way to purposefully improve the classification  
510 performance of the model. An extended colocated dataset would allow stricter filtering, mainly with respect to the collocation  
511 time-window, which would help improve the representativeness of the samples. The analysis of the classification performance  
512 shows here the limitations of a reduced-size dataset with potential underlying discrepancies between data sources during  
513 collocation. Nonetheless, the evaluation of the predictions in the following section provides insights and reveals relevant features  
514 in the predicted cloud types.

515

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

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

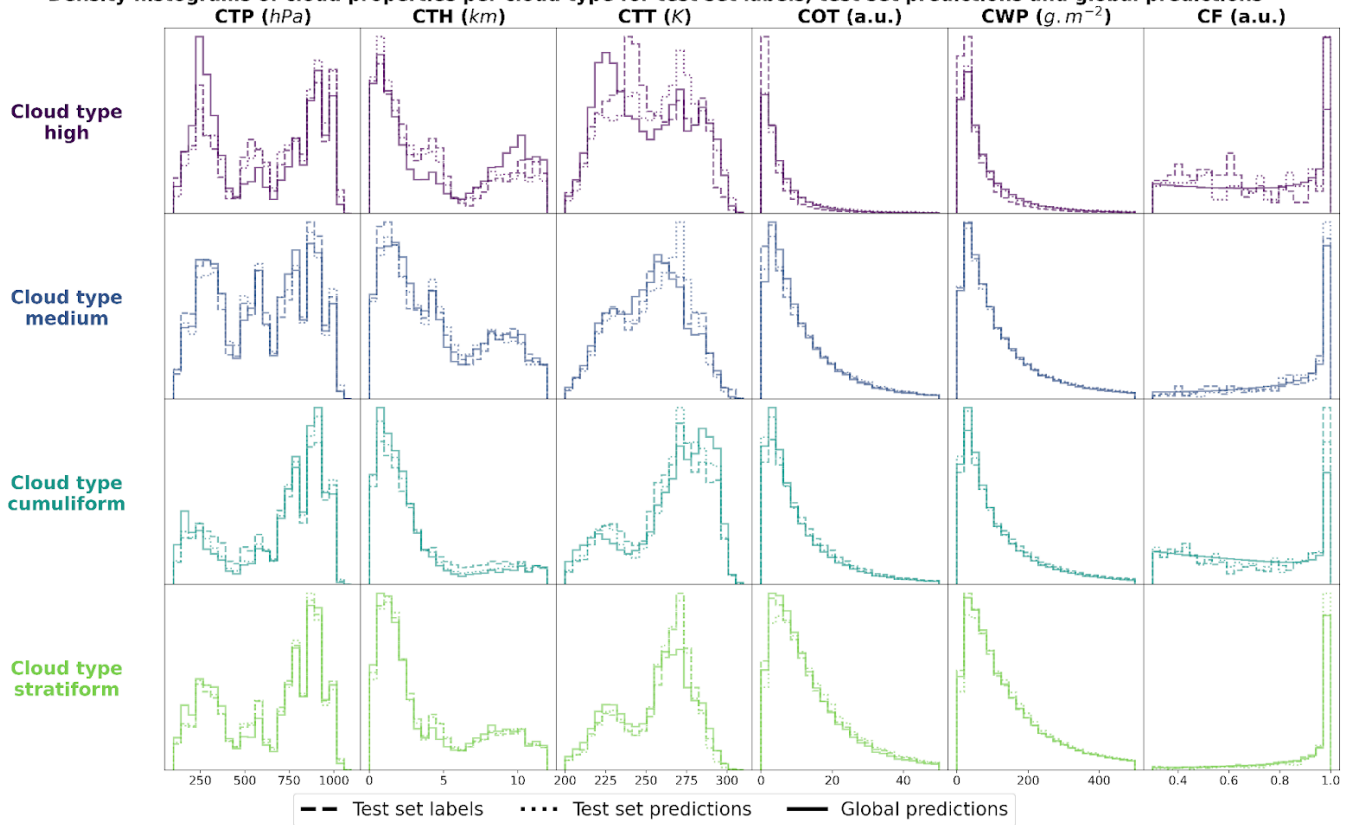
## 520 4.2 Histograms of cloud properties

521

522 In order to evaluate the physical soundness of the predictions made by the CloudViT model, we investigate the distribution of  
523 several cloud properties with respect to the observed and predicted cloud types. In Figure 5, we summarise the distribution of  
524 cloud top pressure (CTP), cloud top height (CTH), cloud top temperature (CTT), cloud optical thickness (COT), cloud water path  
525 (CWP) and cloud fraction (CF) for the 4 cloud types (high, medium, cumuliform, stratiform) and for three different datasets: the  
526 test set labels, the test set predictions and the dataset of global predictions. The latter is built on global MODIS AQUA granules  
527 for the year 2016 - the year is chosen to avoid any overlap with cloud scenes seen during the training of the vision transformer on  
528 data from 2008 - from which we regularly sample tiles in order to build a more comprehensive and global dataset of cloud types  
529 to further evaluate the method. The spatial distribution of cloud types for this dataset is highlighted in the following section and  
530 the global dataset is made available at Lenhardt et al. (2024b). The histograms are built by reporting the respective cloud  
531 properties for all the cloudy pixels in each sampled tile from the dataset apart from the cloud fraction which is computed for the  
532 whole tile from the cloud mask. As a consequence, unless the whole cloud field is composed of only a single cloud type, the  
533 histograms will cover a large range of cloud properties due to multi-layer clouds or multi cloud types scenes (e.g. convective  
534 cells with associated anvils or cumulus/stratocumulus transitions). Even though the trained model only produces fair evaluation  
535 metrics on the test set, the histograms of cloud properties display interesting features consistent with expected characteristics of  
536 the different cloud types. On Figure 5, the histograms pertaining to the test set labels and predictions have distributions close to  
537 identical across cloud types showing a good agreement in the clouds depicted in both datasets while the global dataset histograms  
538 provides a less noisy overview of the distribution of the cloud properties per cloud type. The high clouds are characterised by  
539 low cloud water path and optical thickness, along with colder and higher cloud tops as well as more frequent cloud fractions  
540 smaller than one. All of these aspects are emphasised in the global predictions compared to the limited test set samples, showing  
541 the CloudViT model manages to extract the representative characteristics of the cloud type from the labels. The cumuliform  
542 category encompasses mostly low warm clouds with reduced cloud fractions and moderate cloud water path and optical  
543 thickness. Inside this class, the higher and colder cloud tops are concentrated in the cumulonimbus class, along with larger cloud  
544 water path and cloud optical thickness (see Fig. C.1). The stratiform class includes thick cloud fields with high cloud water path  
545 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  
546 are slightly higher and colder and correspond to stratus/nimbostratus clouds which can also be seen in Figure C.1. The  
547 distributions for medium clouds showcase similarities with several other types and are best evaluated in combination with their  
548 spatial distribution (see Section 5). Examining in more detail the refined cloud types with the 10 cloud types (see Fig. C.1)  
549 reveals slight differences inside broader cloud types. For example, the distinction between the three high cloud types (cirrus,  
550 cirrostratus and cirrocumulus) appears through separations in cloud fraction, cloud optical thickness and cloud water path which  
551 were not obvious from the limited amount of labelled samples. The differences between the three high cloud types further  
552 manifest in distributions of cloud top quantities for which cirrus and cirrostratus display potential multilayered cloud scenes with  
553 a combination of low/warm and high/cold cloud tops. Overall, the CloudViT model seems to generalise well from a few samples  
554 (only around 10 for the cirrocumulus class) by exhibiting in parts physical consistency inside predicted types. Due to the large  
555 cloud scenes considered as input for the classification, the distribution of the cloud properties might not be as representative of  
556 single cloud types as an input tile of, for example, 16 pixels. The main caveat regarding performance on high and medium clouds  
557 from our method is that the ground-based observer identifies these cloud types with higher uncertainty compared to that of low  
558 clouds. Additionally, stratiform clouds with high cloud fraction can hinder the trustworthiness of the surface observation if the  
559 whole field of view is cloudy. Even though the limitations of ground-based observations are evident, they still provide quality  
560 observations on which a classification model can be trained. The collocation between these surface observations and the satellite  
561 retrievals is thus of crucial importance and guides the performance of the later trained model. It partly contributes in the case of  
562 CloudViT to a hurdle to achieve notable classification performance. The model, however, shows its ability to generalise from  
563 limited samples to consistent and physically-relevant distributions of cloud properties among the predicted cloud types. By  
564 refining the training dataset, the improvements can be expected to reflect directly on the classification performance. The  
565 characteristics observed in the histograms across cloud types contribute to an increase in confidence in the ability of CloudViT to  
566 discern various cloud types in large remote sensing datasets despite the method's limited ability described in the previous section.

567

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



568

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

570 **5 Global cloud type distributions in MODIS data** **Results**

571

572 **5.1 Global cloud type distributions in MODIS data**

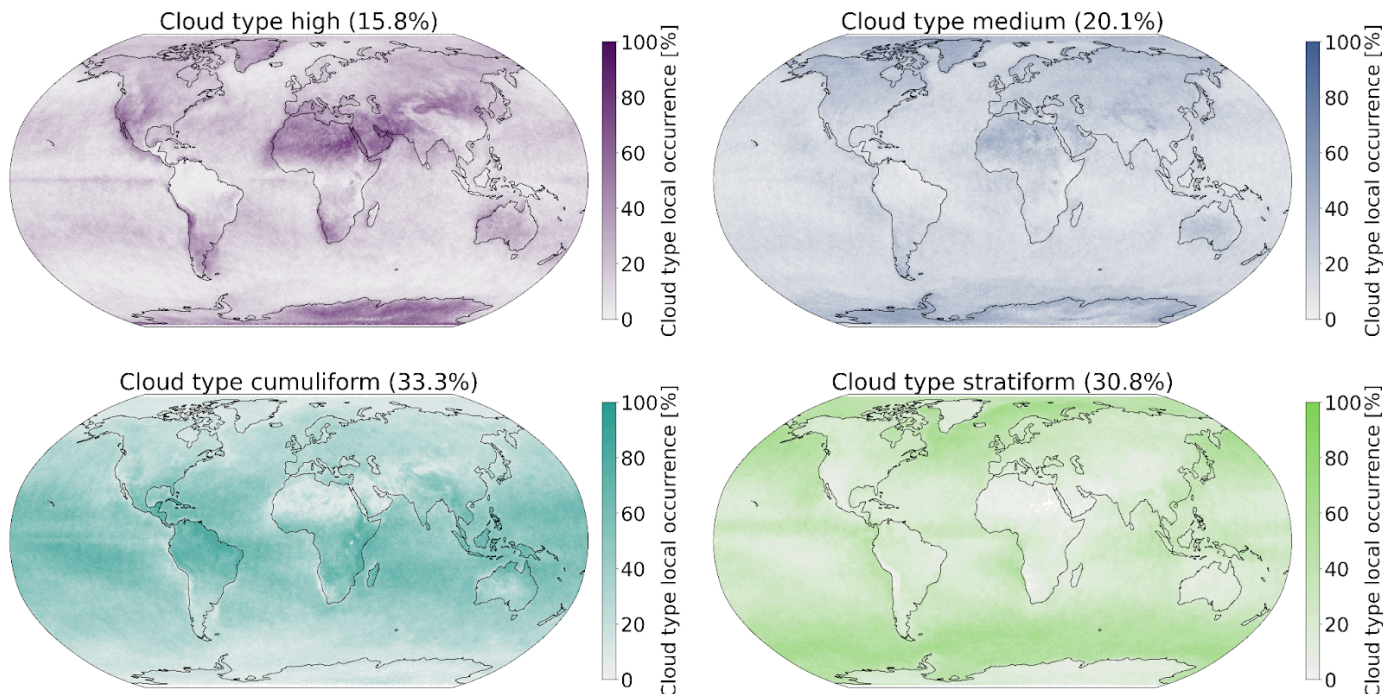
573

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

584 The spatial distributions of the predicted cloud types for the global dataset for the year 2016 are detailed in Figure 6 and Figure  
 585 C.2 for 4 and 10 cloud types, respectively. Firstly, we note that CloudViT predictions capture large scale patterns which are in  
 586 agreement with observational datasets (Sassen et al., 2008; Cesana et al., 2019; Wood, 2012; Pincus et al., 2023). Stratiform  
 587 clouds, and in particular stratocumulus (see Fig C.2), are frequent in the high latitudes and along the western coasts of America  
 588 and Africa. Cumuliform clouds are concentrated in the Tropics apart from the areas where stratocumulus clouds are dominant.  
 589 Medium clouds are concentrated in the polar regions and over land in the higher latitudes. High clouds make up a large portion  
 590 of clouds in the polar regions but also over land. The first notable difference is the low occurrence of high clouds in the Tropics  
 591 which would be expected to be higher (Sassen et al., 2008; Pincus et al., 2023). An explanation could be the frequent occurrence  
 592 of high clouds in multi-layer cloud scenes related to convection in the Tropics. Furthermore, in such cases the model probably  
 593 identifies the cloud types with larger cloud fraction and thus discards potential high clouds in the scene. Incorporating more  
 594 samples of high clouds in that region (see Fig. A.1) could potentially help the performance of the classification model in that

595 regard. The presented spatial distributions may suffer from the somewhat limited performance of the classification model despite  
596 the corresponding reasonable representation of cloud type characteristics showcased in section 4.2. Nevertheless, some  
597 informative features are observed in Figures 6 and C.2 and point towards the good direction for further improving CloudViT.  
598

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



599

600

601

602

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

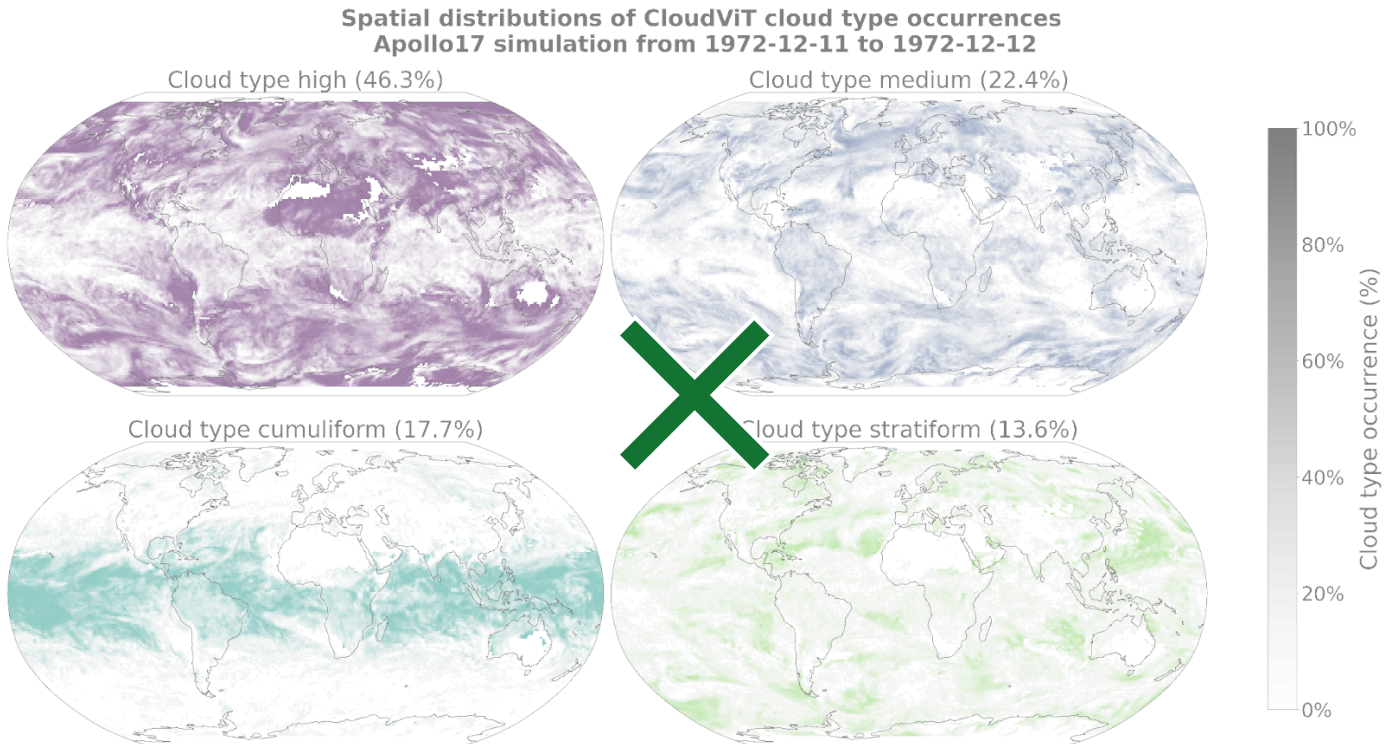
### 603 5.2 Application to a global storm-resolving model simulation ¶

604 ¶

605 As a proof of concept and for probing the potential of CloudViT, we investigate the cloud type representation in general  
606 circulation model (GCM) outputs using our CloudViT method. We build on a new generation of GCMs at kilometre resolution,  
607 namely the ICON-Sapphire (Hohenegger et al., 2023). As the resolution of the simulation increases, some processes like deep  
608 convection can be directly resolved instead of parameterized. Hence, building diagnostics about cloud representation is of  
609 importance to help evaluate the simulations. In particular, we use the simulation run by the Max Planck Institute for Meteorology  
610 (MPI-M) for the period between the 5th and 12th of December 1972, aiming at recreating the Blue Marble picture made during  
611 the Apollo 17 mission on the 7th of December. Here we only use the complete outputs provided for the 11th of December. The  
612 grid used contains 335 544 320 grid points at each level in the atmosphere (R02B11 grid), and outputs are provided every 30  
613 minutes during the simulation for the atmospheric quantities of interest, resulting in overall 48 time steps. As the effective  
614 horizontal resolution of the model simulation and the MODIS data are on similar scales, we can effectively apply CloudViT on  
615 the model outputs. From the model outputs, we derive the cloud properties necessary for the method introduced in this study.  
616 More information about the particular model setup and the derivation of cloud properties is included in Appendix D. However,  
617 the standardisation of the input cloud properties for the vision transformer model is still done based on statistics computed on  
618 MODIS data which could induce a bias in the latent representations and subsequently on the predictions. Extending the method  
619 to other datasets like this GCM simulation thus requires careful investigation that the cloud properties lie in the same range or  
620 display similar distributions. ¶

621 For each 30-minute time step, we proceed to sample tiles, regularly spaced, to reach global coverage of cloud type estimates.  
622 Figure D.37 displays the daily averaged occurrence of the cloud type predictions on a 1° regular grid for the 4 cloud types, the  
623 equivalent for 10 cloud types is presented in Figure D.43. However, due to the time period covered by the simulation, no global  
624 data record for cloud types can be used to evaluate the representation of cloud types by the ICON-Sapphire through the  
625 CloudViT method. A thorough analysis would be feasible for simulations covering a time period for which climate data records  
626 of cloud types are available, for example the ISCCP H-series climate data record (Young et al., 2018) which starts in 1983. The

627 aim here is rather to present as a proof of concept the transfer of the method to model data outputs, and directly describe the  
 628 outcome objectively. A large proportion of the predicted clouds belong to the high cloud type, hinting at the difference in  
 629 sensitivity to clouds retrieved in the climate model data compared to the MODIS retrievals or the mismatch in the training  
 630 process of CloudViT, high clouds being underrepresented and their corresponding classification metrics lower than for some  
 631 other cloud types. However, increasing the cloud ice content threshold by an order of magnitude greatly decreases the amount of  
 632 thin, high and cold clouds in the simulation dataset. This aspect would need further tuning through comparison with remote  
 633 sensing retrievals which are not available for this particular simulated period. On the other hand, the cumuliform class captures  
 634 well the convective systems in the tropics while the stratocumulus decks can be identified (Fig. D.43). Additionally, the medium  
 635 clouds are more present at high latitudes. An important aspect to factor in is that the classification model was only trained on  
 636 daytime satellite observations as the optical cloud properties necessary are only available then. Thus, results on nighttime cloud  
 637 retrievals which is the case for some of the predictions produced from the presented simulation might need more meticulous  
 638 evaluation. Even though it is a limiting factor in the case of the satellite dataset we are using, the simulation outputs provide us  
 639 with the required variables across all timesteps. ¶¶



640  
 641 **Figure 7: Spatial distributions of the CloudViT cloud type occurrences (cloud types high, medium, cumuliform,**  
 642 **stratiform) for the ICON Sapphire Apollo 17 simulation of December 11<sup>th</sup> 1972 aggregated on a 1° regular grid.**

## 643 6 Challenges and limitations

644

645 The method and results highlighted in the previous sections provide useful material to further analyze the developed  
 646 methodology, but also to be critical of its shortcomings. The following section aims to focus on several aspects that we feel are  
 647 relevant for the community when developing cloud type classification methods similar to the one presented here, namely on the  
 648 benefits of such methods, dataset curation and extension, and the potential application to climate model data.

649 Spatially-resolved cloud properties provide usable context for the CloudViT model to improve the cloud classification, as shown  
 650 in the comparison to the baseline method with limited spatial information. Introducing this new transformer model architecture  
 651 additionally improves the classification skill over the CNN backbone mentioned in Lenhardt et al. (2024a). Overall, CloudViT  
 652 achieves passable performance even on sparsely represented classes for both cases of 4 and 10 cloud types. The limited collocated  
 653 dataset proves to be a hurdle for the proper training and evaluation of the method on labelled samples but the generation of an  
 654 extensive global dataset allows deeper investigation into the cloud types. Improvements could come from a more extensive  
 655 training dataset which would encompass a larger variety of cloud type samples to certainly enhance the classification's

656 performance both for the training and testing metrics. The subsequent evaluation exhibits interesting results despite the limited  
657 performance on the colocated dataset. In the global dataset, the predicted cloud types exhibit fairly physically reasonable  
658 distributions of their respective cloud properties, and their global spatial distributions are consistent in parts with other products  
659 (Section 5). Application to climate model data can prove to be technically straightforward and results in insights into how such  
660 methods can be transferred to model data (Appendix D). The necessary cloud quantities can be obtained from common  
661 simulation outputs (cloud liquid water and ice contents, altitude, droplet number) which makes methods using such inputs easily  
662 applicable to climate model simulations. Cloud type diagnostics could be a resourceful addition to the panel of assessment  
663 methods for model data (Kuma et al., 2023; Kaps et al., 2023) given improvements to achieve remarkable performance in the  
664 classification ability as previously described.

665 The hypothesis as to why the model fails to achieve great performance in this study rests heavily on the collocation process  
666 between surface observations and satellite data. The method would benefit from including further ground-based observations  
667 through the collocation process but then much larger storage and computational facilities would be needed as global MODIS data  
668 represents thousands of granules each day. More training samples could simultaneously solve performance issues by providing a  
669 clearer vision of the different cloud types for the classification model to learn from. The improvements through a larger training  
670 dataset will yield relevant benefits only if the potential mismatches occurring during the collocation process are tackled.  
671 Improving the representativeness of the training samples could solve the performance issues faced by the model presented here,  
672 and potentially achieve performance aligned with a wider usage of the method for cloud type analysis. The classification model  
673 could also be refined by finding better alternatives to the RF or MLP presented here. The overall finetuning process involving the  
674 vision transformer and the MLP classification head proved to be cumbersome but holds great promise if the labels and training  
675 process are refined. Transfer learning from a typical ImageNet-trained model did not yield a notable performance difference  
676 which shows the current need for foundation models trained on remote sensing data. The main hurdle here remains the large  
677 diversity in instruments, quantities and resolutions among remote sensing products which hinders the possibility of a unified  
678 model.

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

690 Furthermore, some caveats can appear when applying such a cloud classification method to climate model data. As mentioned in  
691 more details in Appendix D, the input scaling is crucial to ensure proper portability of a method to this other data source. The  
692 absence of nighttime retrievals in the MODIS data also turns the evaluation of predictions on nighttime data points across the  
693 model data into a challenging issue. However, clouds play a role in the climate system both during the day when they cool the  
694 surface by mostly reflecting incoming solar radiation but also at night when they warm the surface by trapping outgoing  
695 terrestrial radiation. Shifts and changes in cloud occurrence and distribution in the current climate but also in future projections  
696 could further influence global climate change (Luo et al., 2024). Applying a cloud classification methodology to a limited  
697 high-resolution climate model simulation is an encouraging direction, but considering more common and computationally less  
698 expensive global km-scale simulations (horizontal resolution of 5 km for example) could be of greater interest to the community  
699 to study longer time scales. To this extent, two conceivable approaches would consist in either retraining a method like CloudViT  
700 on coarser input cloud properties matching the model data resolution - the MODIS Cloud product is also available at a 5 km  
701 resolution even though the 1 km equivalent is recommended for use - or in using such a method as is but with the coarse input  
702 scaled to fit the resolution of the tiles on which it was trained on. The first option could be more interesting as computer vision  
703 models are commonly trained on coarser resolutions first to learn the broad specificity and patterns in the data before fine-tuning  
704 the model on finer resolution (Touvron et al., 2019).

## 705 **76 Conclusion**

706

707 This study introduces a new method called CloudViT to classify cloud types from MODIS cloud properties, specifically CTH,  
708 COT and CWP. CloudViT delivers estimates for either 4 (high, medium, cumuliform, stratiform) or 10 (cirrus, cirrostratus,  
709 cirrocumulus, altostratus, altocumulus, cumulus, cumulus and stratocumulus, cumulonimbus, stratocumulus, stratus) cloud types  
710 with fair performance. The classification model was built on ground-based observations of cloud types (Section 2.1) and  
711 experiments about its generalisation skill and the benefits of spatial information were presented (Section 3). We evaluated the  
712 classification model by examining distributions of cloud properties in Section 4 and the global spatial distribution of cloud types  
713 in Section 5.1. ~~Lastly, we transferred our method to a km-scale climate model simulation made with ICON-Sapphire (Section~~  
714 ~~5.2).~~ ~~Lastly, we pinned down some existing challenges, limitations, and lessons learned from the development of the method for~~  
715 ~~cloud type classification.~~ The global dataset alongside the CloudViT code and weights are made available on Zenodo (Lenhardt  
716 et al., 2024b) to encourage ~~future developments~~ ~~further improvements to the method and the reusability of the model's~~  
717 ~~architecture.~~

718 ~~Spatially resolved cloud properties provide usable context for the CloudViT model to improve the cloud classification, as shown~~  
719 ~~in the comparison to the baseline method with limited spatial information. Introducing this new transformer model architecture~~  
720 ~~additionally improves the classification skill over the CNN backbone mentioned in Lenhardt et al. (2024a). Overall, CloudViT~~  
721 ~~achieves passable performance even on sparsely represented classes for both cases of 4 and 10 cloud types. The limited colocated~~  
722 ~~dataset proves to be a hurdle for the proper training and evaluation of the method on labelled samples but the generation of an~~  
723 ~~extensive global dataset allows deeper investigation into the cloud types. Improvements could come from a more extensive~~  
724 ~~training dataset which would encompass a larger variety of cloud type samples to certainly enhance the classification's~~  
725 ~~performance both for the training and testing metrics. The subsequent evaluation exhibits interesting results despite the limited~~  
726 ~~performance on the colocated dataset. In the global dataset, the predicted cloud types exhibit fairly physically reasonable~~  
727 ~~distributions of their respective cloud properties, and their global spatial distributions are consistent in parts with other products~~  
728 ~~(Section 5.1). Application to climate model data proves to be straightforward and results in insights into how such methods can~~  
729 ~~be transferred to model data, and preliminarily on how clouds are represented in global km-scale simulations. The necessary~~  
730 ~~cloud quantities are obtained from common simulation outputs (cloud liquid water and ice contents, altitude, droplet number)~~  
731 ~~which makes CloudViT easily applicable to other climate model simulations. Cloud type diagnostics such as CloudViT could be~~  
732 ~~a resourceful addition to the panel of assessment methods for model data (Kuma et al., 2023; Kaps et al., 2023) given~~  
733 ~~improvements to achieve remarkable performance in its classification ability as previously described.¶~~

734 ~~The hypothesis as to why the model fails to achieve great performance in this study rests heavily on the collocation process~~  
735 ~~between surface observations and satellite data. The method would benefit from including further ground-based observations~~  
736 ~~through the collocation process but then much larger storage and computational facilities would be needed as global MODIS data~~  
737 ~~represents thousands of granules each day. More training samples could simultaneously solve performance issues by providing a~~  
738 ~~clearer vision of the different cloud types for the classification model to learn from. The improvements through a larger training~~  
739 ~~dataset will yield relevant benefits only if the potential mismatches occurring during the collocation process are tackled.~~  
740 ~~Improving the representativeness of the training samples could solve the performance issues faced by the model presented here,~~  
741 ~~and potentially achieve performance aligned with a wider usage of the method for cloud type analysis. The classification model~~  
742 ~~could also be refined by finding better alternatives to the RF or MLP presented here. The overall finetuning process involving the~~  
743 ~~vision transformer and the MLP classification head proved to be cumbersome but holds great promise if the labels and training~~  
744 ~~process are refined. Transfer learning from a typical ImageNet-trained model did not yield a notable performance difference~~  
745 ~~which shows the current need for foundation models trained on remote sensing data. The main hurdle here remains the large~~  
746 ~~diversity in instruments, quantities and resolutions among remote sensing products which hinders the possibility of a unified~~  
747 ~~model.¶~~

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

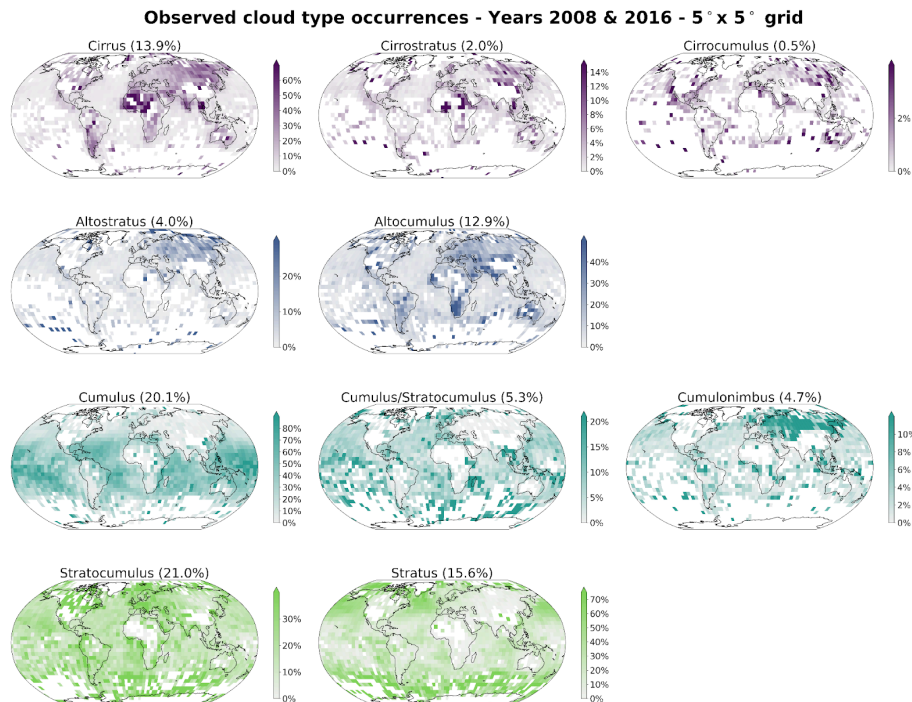
759 Furthermore, some caveats appear when applying CloudViT to climate model data. As mentioned previously, the input scaling is  
760 crucial to ensure proper portability of the method to this other data source. The absence of nighttime retrievals in the MODIS  
761 data also turns the evaluation of predictions on nighttime data points across the model data into a challenging issue. However,  
762 clouds play a role in the climate system both during the day when they cool the surface by mostly reflecting incoming solar  
763 radiation but also at night when they warm the surface by trapping outgoing terrestrial radiation. Shifts and changes in cloud  
764 occurrence and distribution in the current climate but also in future projections could further influence global climate change  
765 (Luo et al., 2024). The proof of concept of applying CloudViT to a limited climate model simulation is encouraging but  
766 considering more common and computationally less expensive global km-scale simulations (horizontal resolution of 5 km for  
767 example) could be of greater interest to the community to study longer time scales. To this extent, two conceivable approaches  
768 would consist in either retraining the CloudViT model on coarser input cloud properties matching the model data resolution - the  
769 MODIS Cloud product is also available at a 5 km resolution even though the 1 km equivalent is recommended for use - or in  
770 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  
771 could be more interesting as computer vision models are commonly trained on coarser resolutions first to learn the broad  
772 specificity and patterns in the data before fine-tuning the model on finer resolution (Touvron et al., 2019).

773 In conclusion, the method presented here showcases and highlights a wide array of potential applications in the study of cloud  
774 types, their characteristics and evolution, and their past, current, and future effects on the Earth's climate: from the extension of  
775 sparse surface observations to global yearly predictions, and existing challenges and limitations in the design of  
776 vision-transformer-based models, and the proof of concept of transferring the method to a global model simulation. Despite the  
777 relatively imbalanced performance assessment of the method which shows both great promise in capturing large scale  
778 characteristics of cloud types distributions but struggles to capture precisely the features in the training dataset, the design and  
779 development of CloudViT is an interesting study in the line of improving existing cloud classification methodologies and its  
780 straightforward application to model data outputs potentially makes it a useful tool for cloud type diagnostics modulo its  
781 performance reliability being improved. Identified challenges and limitations in this particular case can be useful to the  
782 community, both in terms of methodology development and caveat to be avoided. We recommend future further advancements in  
783 cloud classification methods like CloudViT of the method being firstly focused on data curation and followingly on model tuning  
784 once the performance has been raised to desirable levels. To this extent, the necessary datasets and model architecture code are  
785 made available on Zenodo (Lenhardt et al., 2024b).

## 787 Appendix A: Cloud type observations

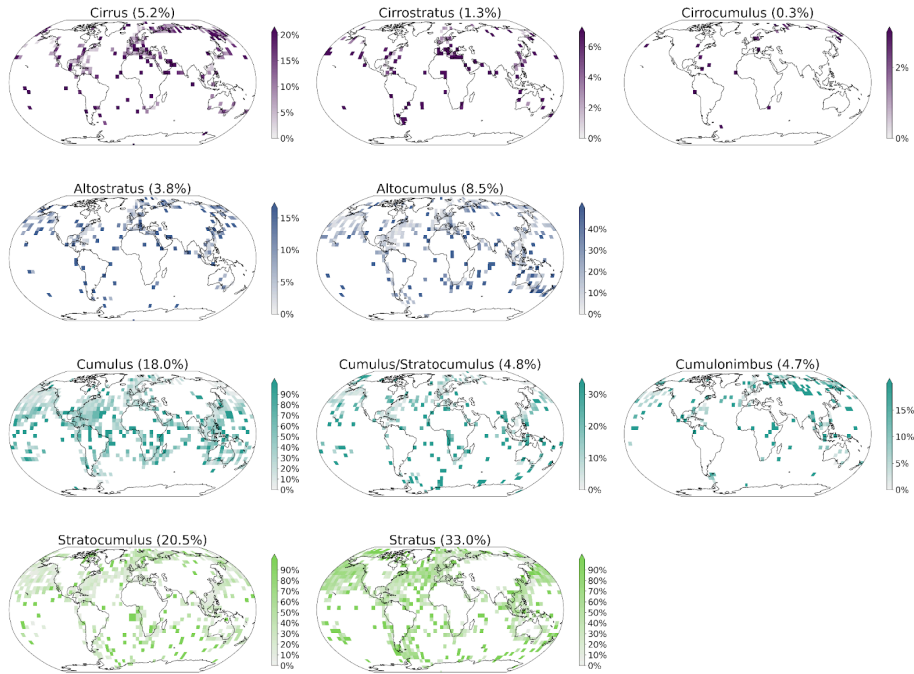
WMO codes	Cloud type: 4 groups	Cloud type: 10 groups	Colocated samples
High clouds 1-6	High	Cirrus	n = 574
High clouds 7-8		Cirrostratus	n = 142
High clouds 9		Cirrocumulus	n = 29
Medium clouds 1-2	Medium	Altostratus	n = 420
Medium clouds 3-9		Alto cumulus	n = 944
Low clouds 1-3	Cumuliform	Cumulus	n = 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

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



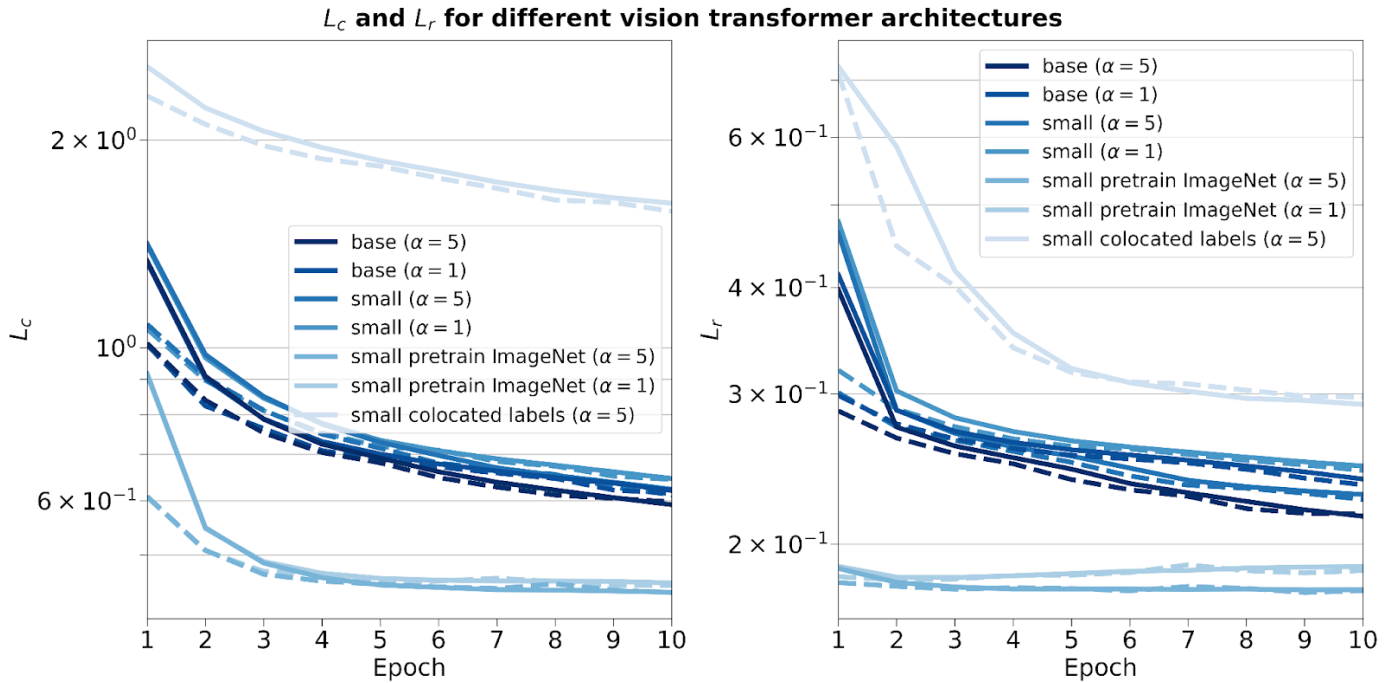
790 **Figure A.1: Spatial distributions of observed cloud types (cloud types cirrus, cirrostratus, cirrocumulus, altostratus,**  
791 **alto cumulus, cumulus, cumulus and stratocumulus, cumulonimbus, stratocumulus, stratus) from the Met Office datasets**  
792 **(Met Office, 2006; Met Office, 2008) for the years 2008 and 2016. Overall percentage of each label in the total dataset is**  
793 **indicated in brackets.**  
794  
795

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



796

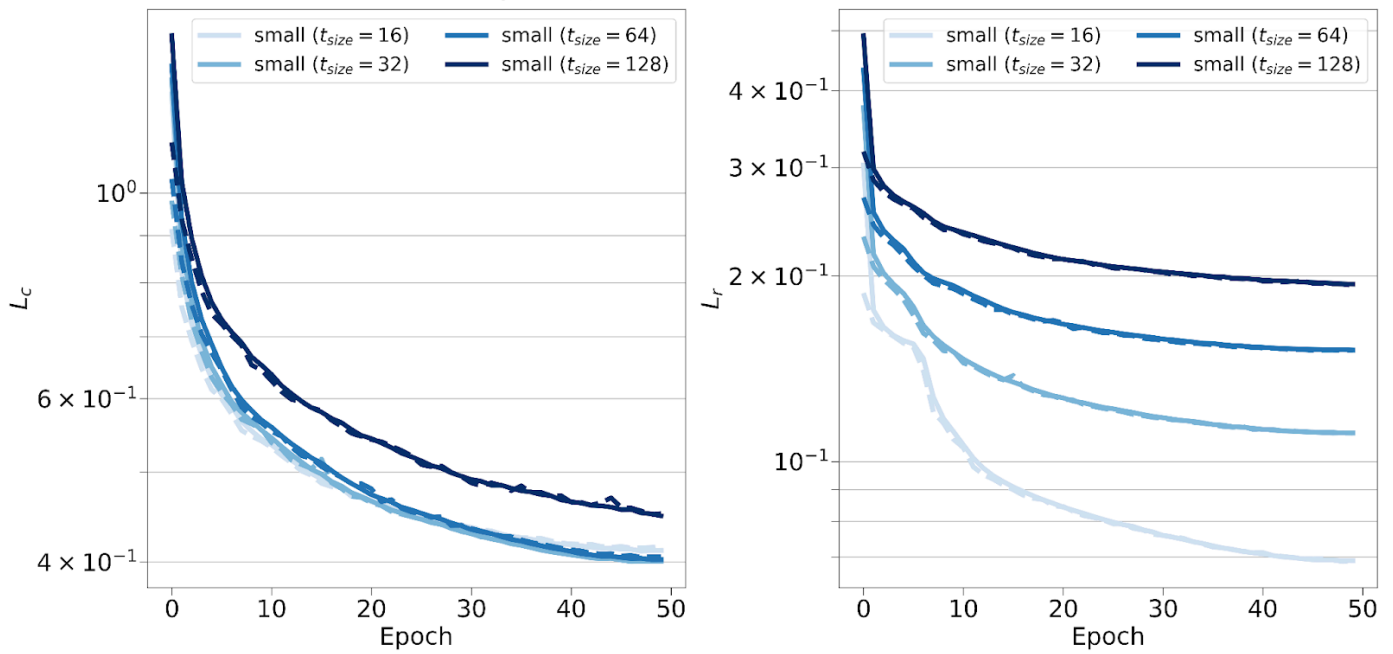
797 **Figure A.2: Spatial distributions of observed cloud types (cloud types cirrus, cirrostratus, cirrocumulus, altostratus,**  
 798 **altocumulus, cumulus, cumulus and stratocumulus, cumulonimbus, stratocumulus, stratus) from the Met Office datasets**  
 799 **(Met Office, 2006; Met Office, 2008)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

807 **Figure B.1: Training and validation contrastive (left) and reconstruction (right) losses for different vision transformer**  
808 **architectures, pretraining weights, training datasets and scaling factor  $\alpha$ .**810 **B.2 Reconstruction errors for the CNN autoencoder and the vision transformer (small variant) on the test set**

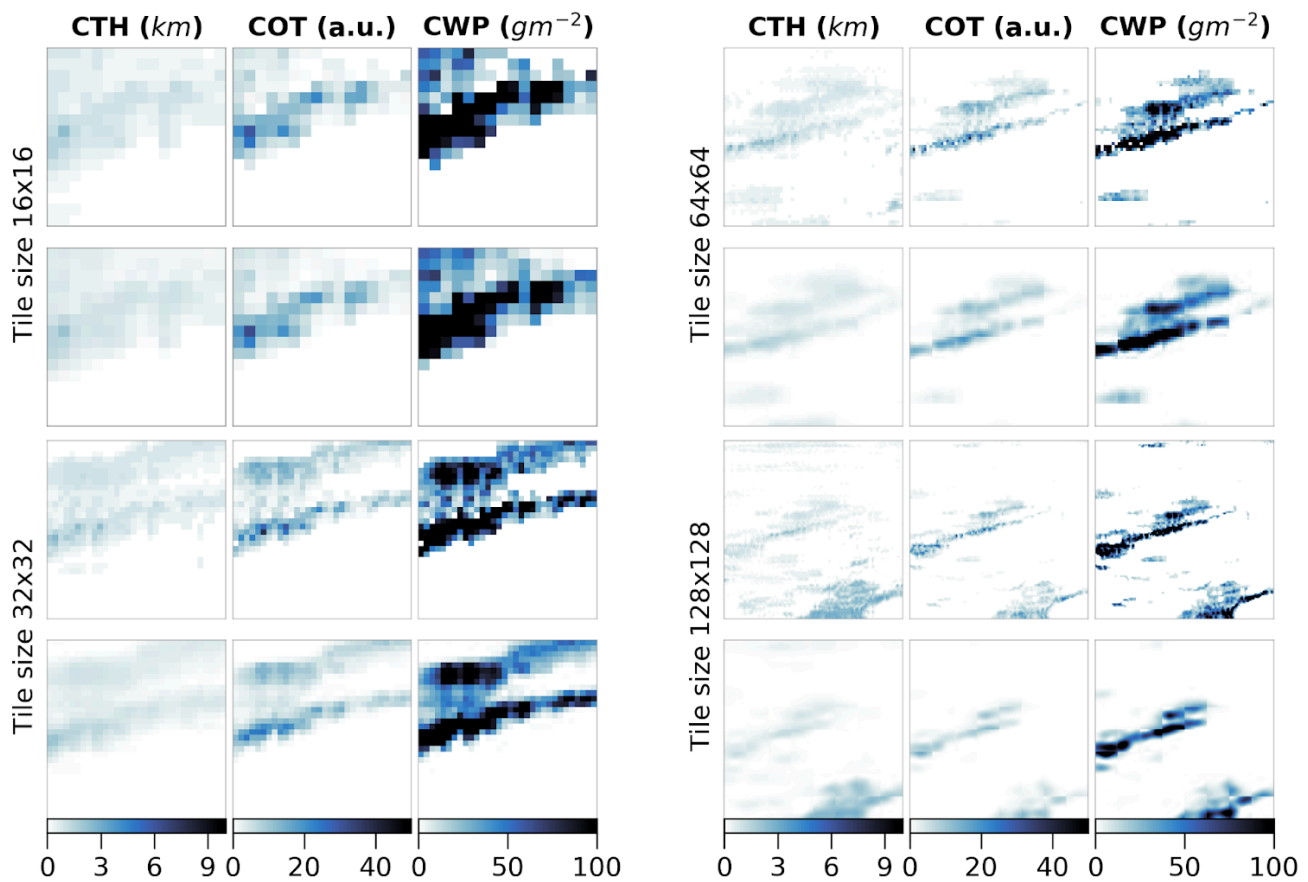
Model type	Reconstruction error	CTH	COT	CWP
CNN autoencoder	MSE	0.27	0.39	0.25
	l1-loss	0.36	0.33	0.21
Vision transformer (small variant)	MSE	0.06	0.25	0.13
	l1-loss	0.10	0.17	0.10

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

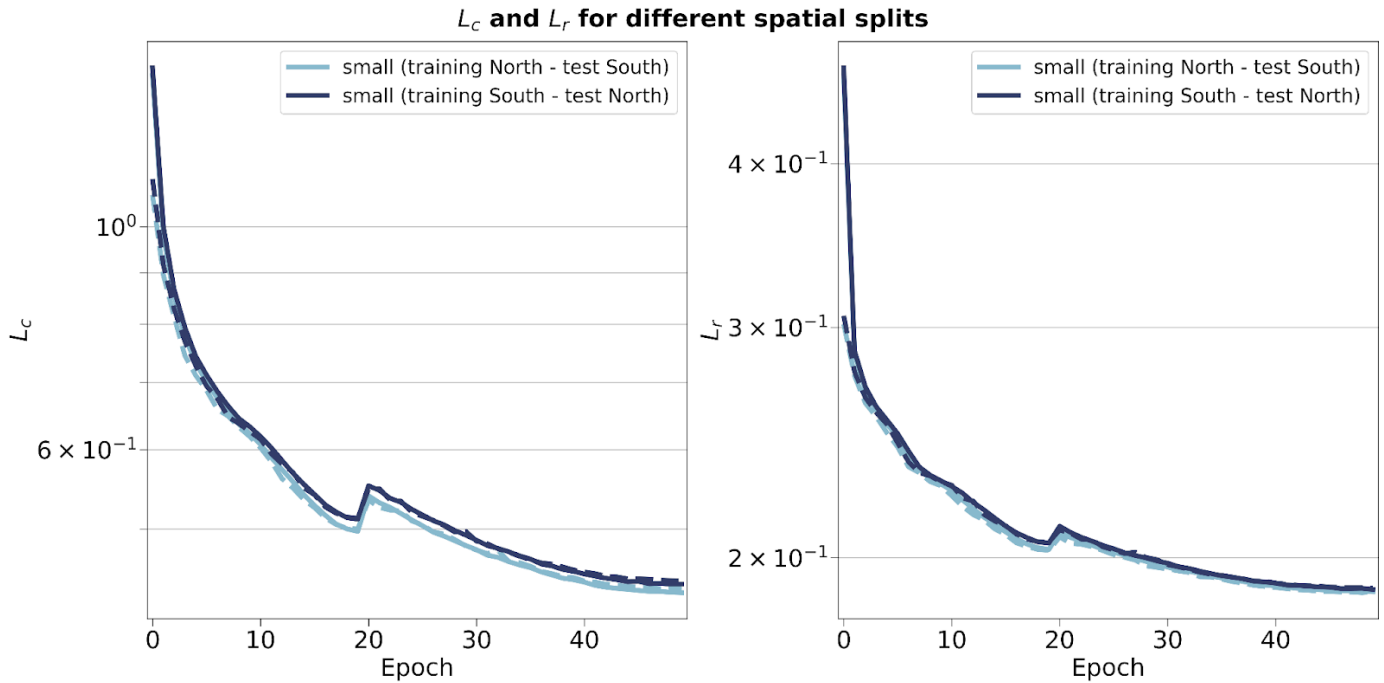
$L_c$  and  $L_r$  for different tile sizes



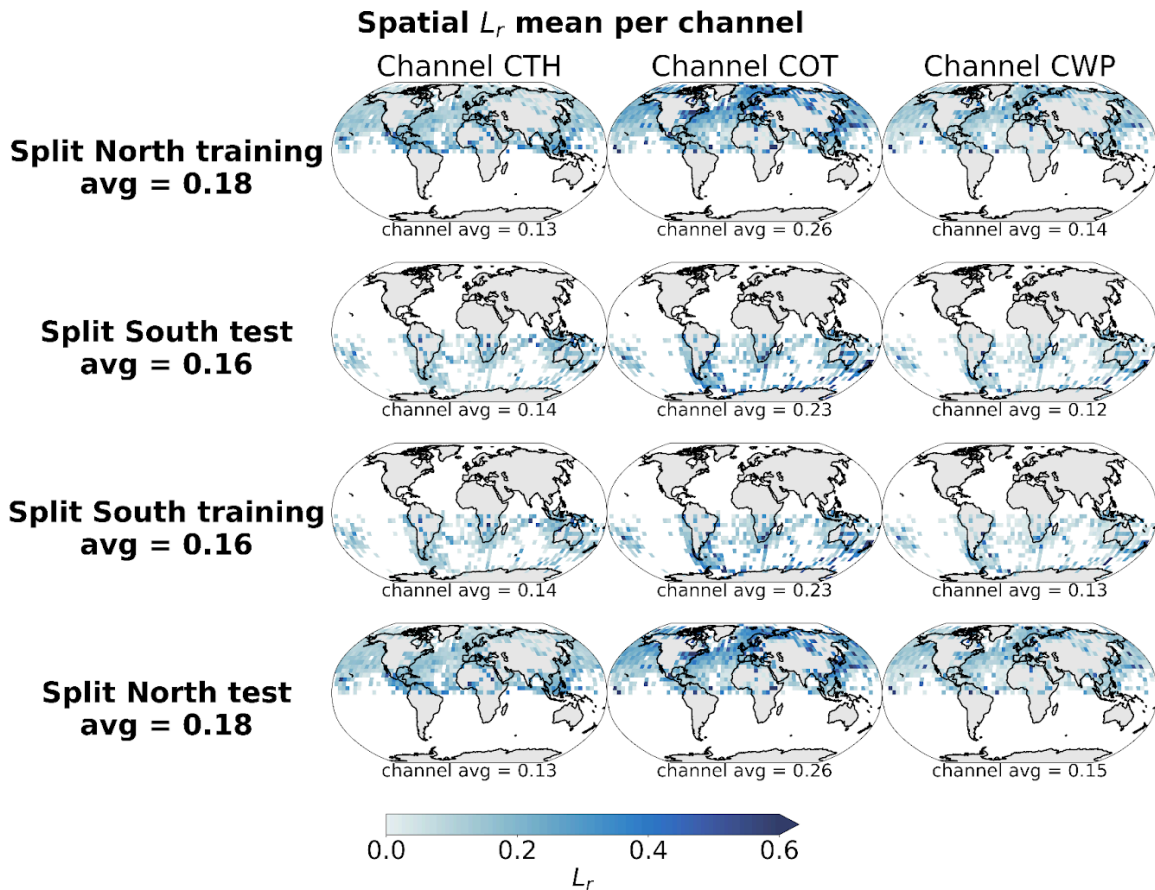
816  
 817 **Figure B.2: Training and validation contrastive (left) and reconstruction (right) losses for vision transformers trained on**  
 818 **different input tile sizes of 16, 32, 64 and 128.**  
 819



820  
 821 **Figure B.3: Input tiles (first and third rows) and corresponding reconstructions (second and fourth rows) for vision**  
 822 **transformers trained on the relevant input tile sizes of 16, 32, 64 and 128.**  
 823



825  
 826 **Figure B.4:** Training (full lines) and validation (dashed lines) metrics for the contrastive (left) and reconstruction (right)  
 827 losses for vision transformers trained on samples from the Northern or Southern hemispheres.



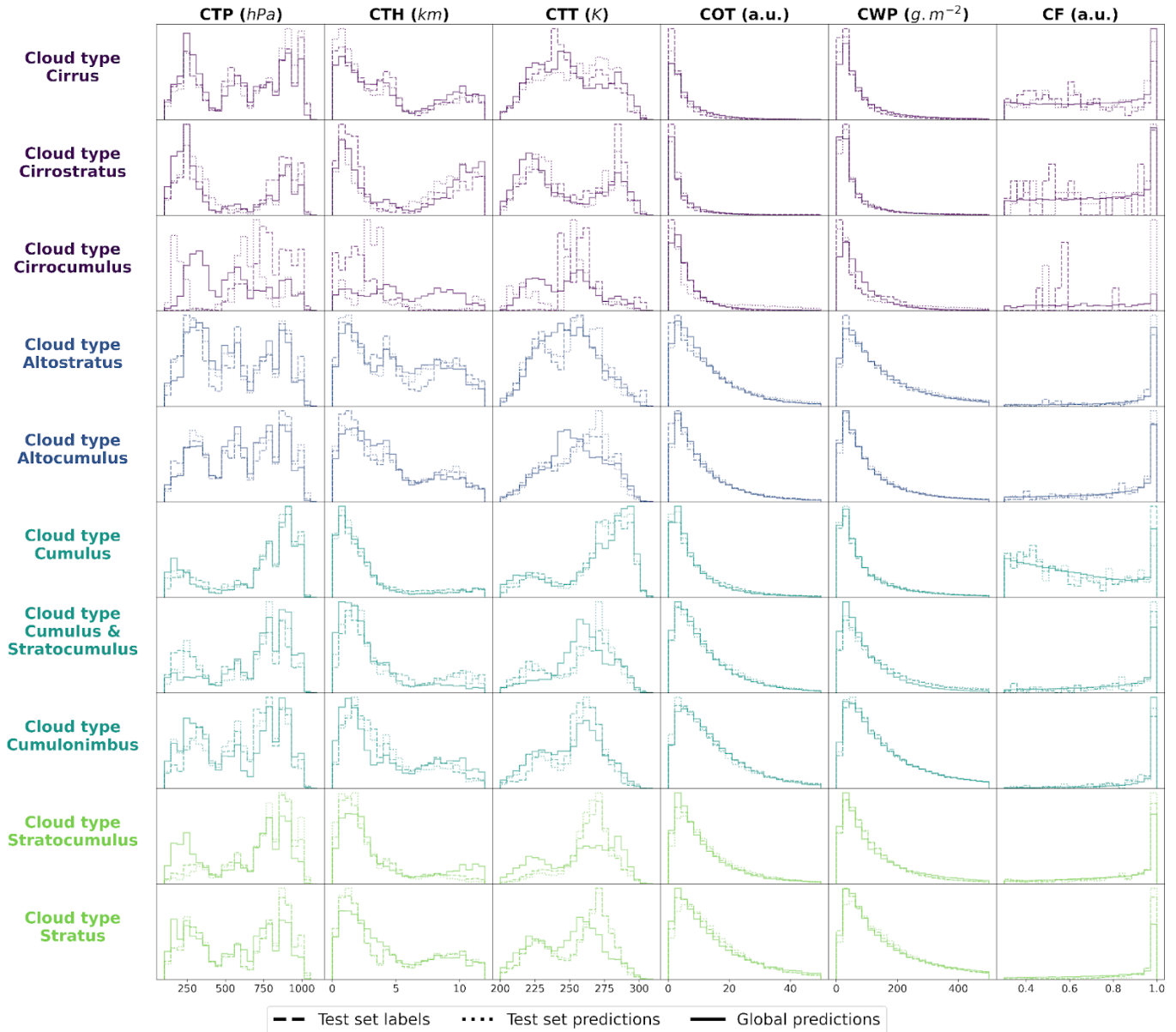
828  
 829 **Figure B.5:** Spatial distributions of the mean channel reconstruction errors for the Northern and Southern hemispheres  
 830 colocated samples. The first two rows correspond to the model trained on the samples from the Northern hemisphere and  
 831 the last two rows to the model trained on the samples from the Southern hemisphere.

832 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>

833 Table C.1: Classification metrics on the test set in the case of 10 cloud types. The metrics noted with a \* are referring to  
 834 their macro-averaged estimate. The baseline CNN/RF refers to the CNN backbone introduced in Lenhardt et al. (2024a).  
 835

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



836

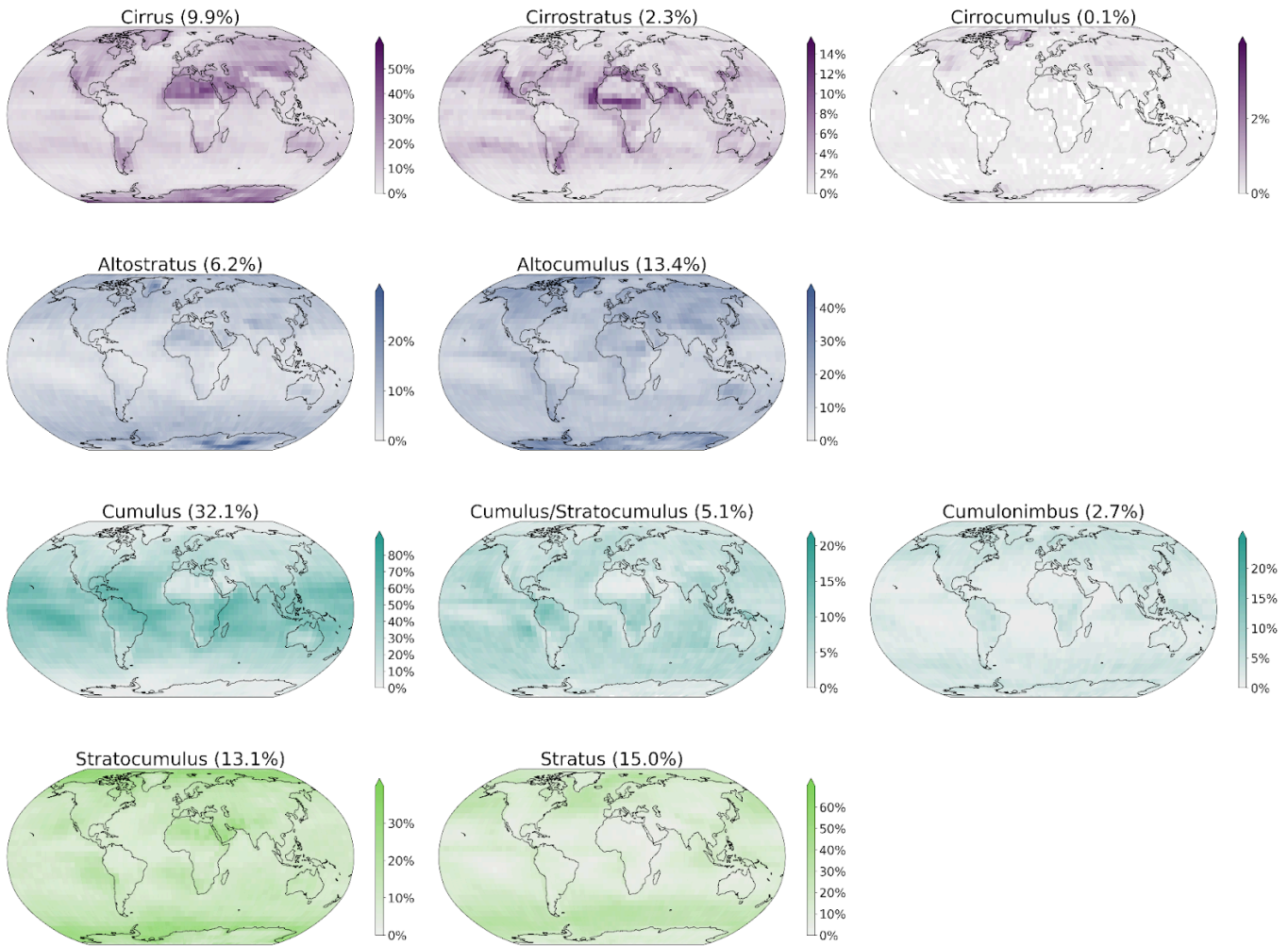
837

838

839

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

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



840

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

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

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

844 **Appendix D: Exploring the technical feasibility of the application to a global storm-resolving model simulation**  
845 **properties computation from model simulation output**

846  
847 As a proof of concept and for probing the potential applicability of methods similar to CloudViT, we technically explore how to  
848 investigate the cloud type representation in general circulation model (GCM) outputs. We build on a new generation of GCMs at  
849 kilometre resolution, namely the ICON-Sapphire (Hohenegger et al., 2023). As the resolution of the simulation increases, some  
850 processes like deep convection can be directly resolved instead of parameterized. Hence, building diagnostics about cloud  
851 representation is of importance to help evaluate the simulations. In particular, we use the simulation run by the Max Planck  
852 Institute for Meteorology (MPI-M) for the period between the 5th and 12th of December 1972, aiming at recreating the Blue  
853 Marble picture made during the Apollo 17 mission on the 7th of December. Here we only use the complete outputs provided for  
854 the 11th of December. The grid used contains 335 544 320 grid points at each level in the atmosphere (R02B11 grid), and outputs  
855 are provided every 30 minutes during the simulation for the atmospheric quantities of interest, resulting in overall 48 time steps.  
856 As the effective horizontal resolution of the model simulation and the MODIS data are on similar scales, we can technically  
857 effectively apply CloudViT on the model outputs. From the model outputs, we derive the cloud properties necessary for the  
858 method introduced in this study.

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

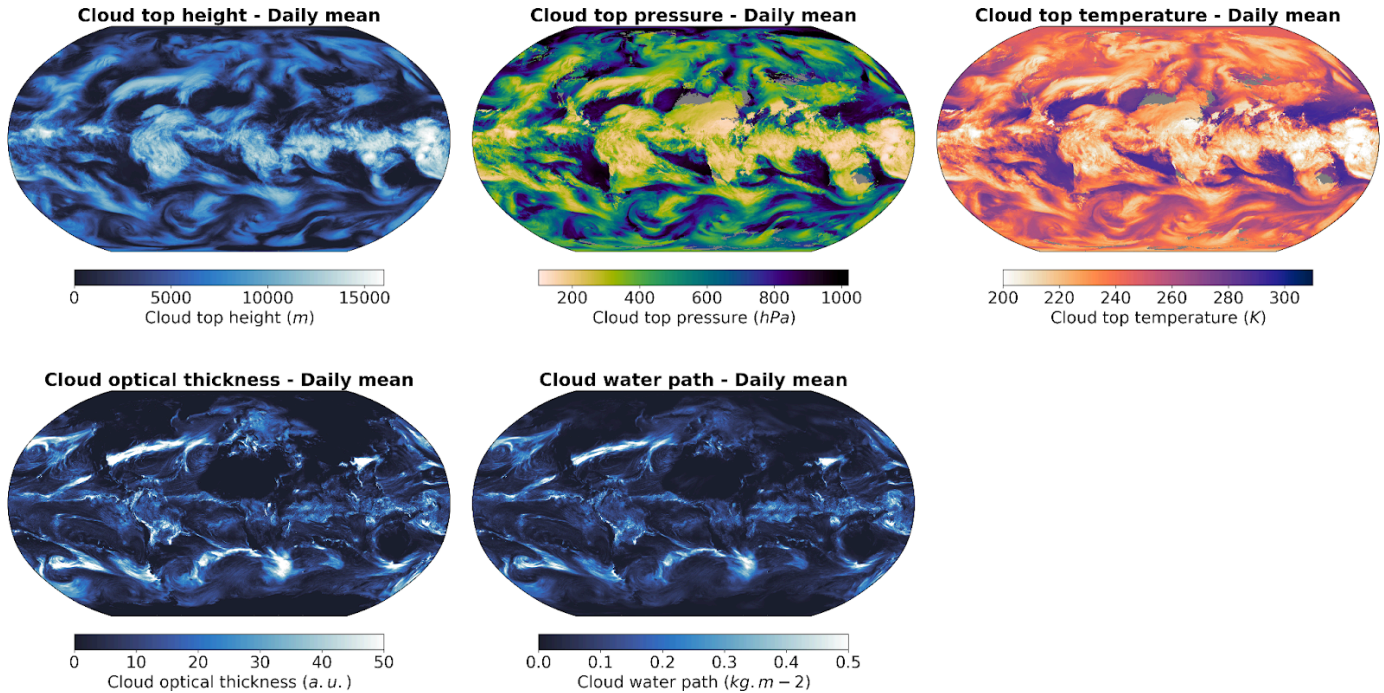
863 The cloud top quantities are retrieved by defining the top-most level where the liquid water content (variable name  $clw$ ) or the  
864 ice content (variable name  $cli$ ) are above a predefined threshold of  $1 \text{ mg.kg}^{-1}$ . This threshold relates to particles of sizes of at least  
865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

$$872 \quad \tau_c = \frac{9}{5} \left( \frac{4\pi}{3\sqrt{2}} \right)^{1/3} \rho_w^{-2/3} (kN_d)^{1/3} c_w^{-1/6} L^{5/6} = 0.2303 \text{ kg}^{-5/6} \text{ m}^{8/3} (kN_d)^{1/3} L^{5/6} \quad (\text{D.1})$$

873 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  
874 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  
875 defined in the simulation by the ECHAM6 parameterization (Equation 6; Stevens et al., 2013). However, the standardisation of  
876 the input cloud properties for the vision transformer model is still done based on statistics computed on MODIS data which could  
877 induce a bias in the latent representations and subsequently on the predictions. Extending the method to other datasets like this  
878 GCM simulation thus requires careful investigation that the cloud properties lie in the same range or display similar distributions.  
879 For each 30-minute time step, we proceed to sample tiles, regularly spaced, to reach global coverage of cloud type estimates.  
880 Figure D.3 displays the daily averaged occurrence of the cloud type predictions on a  $1^\circ$  regular grid for the 4 cloud types, the  
881 equivalent for 10 cloud types is presented in Figure D.4. However, due to the time period covered by the simulation, no global  
882 data record for cloud types can be used to evaluate the representation of cloud types by the ICON-Sapphire through the  
883 CloudViT method. A thorough analysis would be feasible for simulations covering a time period for which climate data records  
884 of cloud types are available, for example the ISCCP H-series climate data record (Young et al., 2018) which starts in 1983. The  
885 aim here is rather to present as a proof of concept the transfer of such a method to model data outputs. A large proportion of the  
886 predicted clouds belong to the high cloud type, hinting at the difference in sensitivity to clouds retrieved in the climate model  
887 data compared to the MODIS retrievals or the mismatch in the training process of CloudViT, high clouds being underrepresented  
888 and their corresponding classification metrics lower than for some other cloud types. However, increasing the cloud ice content  
889 threshold by an order of magnitude greatly decreases the amount of thin, high and cold clouds in the simulation dataset. This  
890 aspect would need further tuning through comparison with remote sensing retrievals which are not available for this particular  
891 simulated period. An important aspect to factor in is that the classification model was only trained on daytime satellite  
892 observations as the optical cloud properties necessary are only available then. Thus, results on nighttime cloud retrievals which is  
893 the case for some of the predictions produced from the presented simulation might need more meticulous evaluation. Even

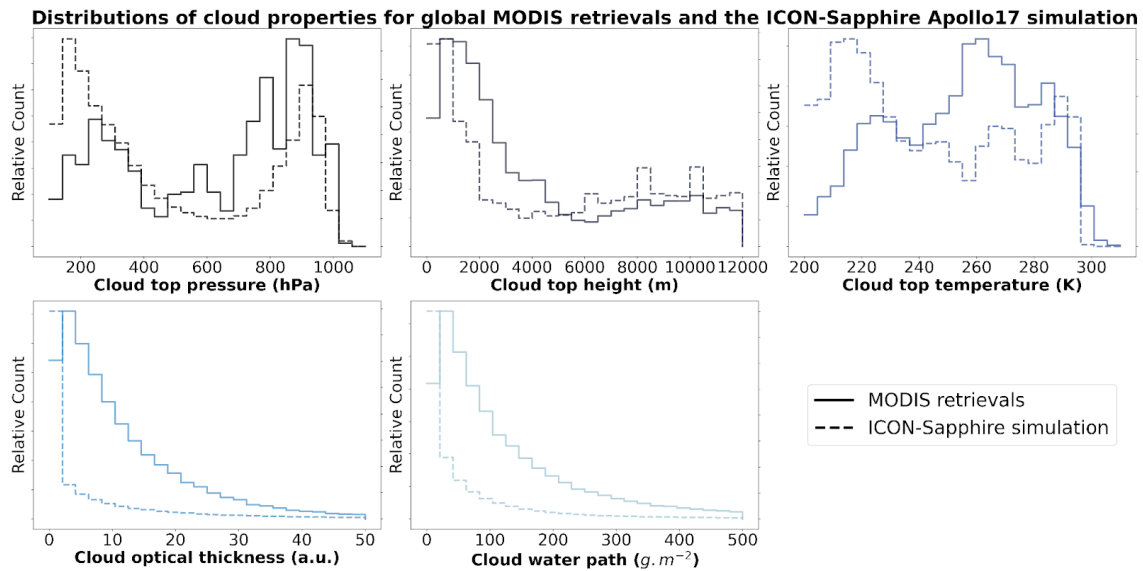
894 though it is a limiting factor in the case of the satellite dataset we are using, the simulation outputs provide us with the required  
 895 variables across all timesteps.

896 ¶  
 897 ¶  
 898



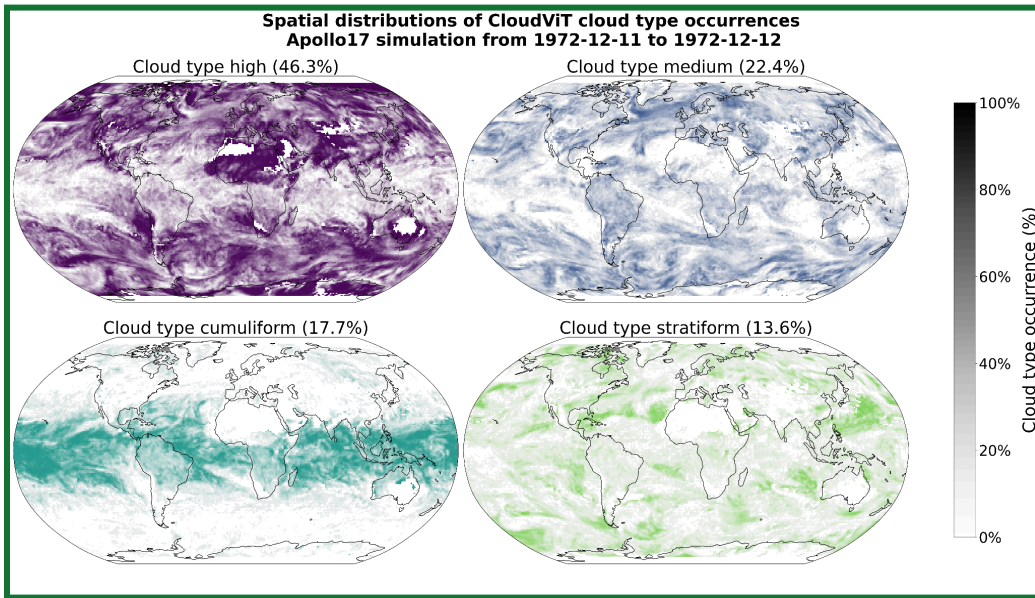
899  
 900 **Figure D.1: Daily averages of cloud top height, cloud top pressure, cloud top temperature, cloud optical thickness and**  
 901 **cloud water path for the 11th of December 1972 from the ICON-Sapphire Apollo 17 simulation.**

902



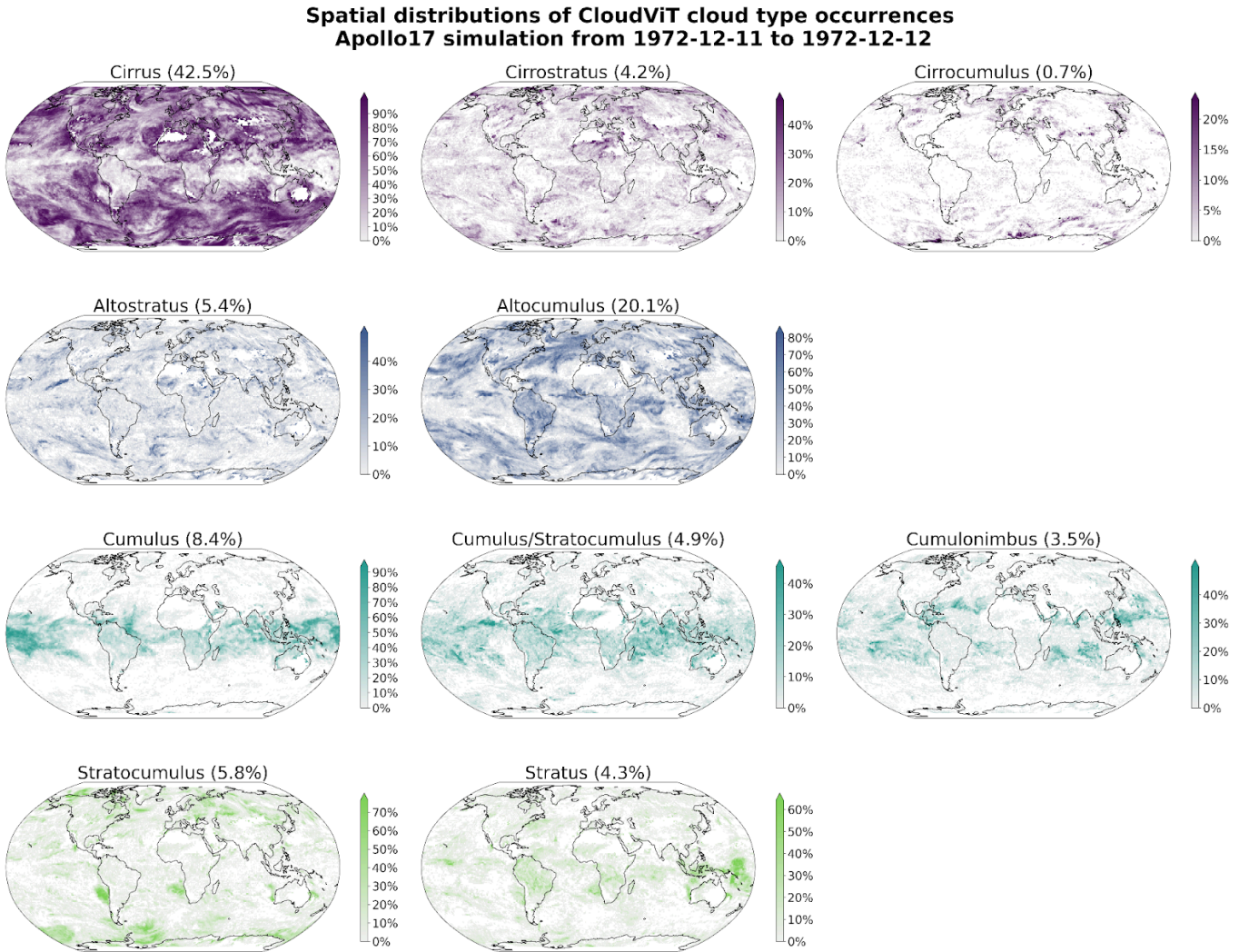
903  
 904 **Figure D.2: Distribution of cloud top pressure, cloud top height, cloud top temperature, cloud optical thickness and cloud**  
 905 **water path for MODIS AQUA retrievals and the ICON-Sapphire Apollo 17 simulation.**

906



907  
908  
909

**Figure D.3: Spatial distributions of the CloudViT cloud type occurrences (cloud types high, medium, cumuliform, stratiform) for the ICON-Sapphire Apollo 17 simulation of December 11<sup>th</sup> 1972 aggregated on a 1° regular grid.**



910  
911  
912  
913

**Figure D.43: Spatial distribution of the CloudViT cloud type occurrences (cloud types cirrus, cirrostratus, cirrocumulus, altostratus, alto cumulus, cumulus, cumulus and stratocumulus, cumulonimbus, stratocumulus, stratus) for the ICON-Sapphire Apollo 17 simulation of December 11<sup>th</sup> 1972 aggregated on a 1° regular grid.**

#### 914 **Code availability**

915  
916 The code used for the method and producing the plots is available on Zenodo (Lenhardt et al., 2024b).

#### 917 **Data availability**

918  
919 The global dataset of the cloud type predictions for the year 2016 is available on Zenodo (Lenhardt et al., 2024b). The dataset is  
920 available as a csv file with corresponding coordinates, MODIS granule file, time of retrieval and predicted cloud type (4 and 10  
921 groups) or in a netCDF file as daily aggregates on a regular grid with a resolution of 1 ° or 5 °. The meteorological observations  
922 from the UK MetOffice (Met Office, 2006; Met Office 2008) are available through the CEDA archive at  
923 <https://catalogue.ceda.ac.uk/uuid/77910bcec71c820d4c92f40d3ed3f249> and  
924 <https://catalogue.ceda.ac.uk/uuid/9f80d42106ba708f92ada730ba321831> for ocean and land observations respectively. The files  
925 from the CUMULO dataset (Zantedeschi et al., 2019) are available at  
926 <https://www.dropbox.com/sh/i3s9q2v2jyjk2it/AACxXnXfMF5wuIqLXqH4NJOra?dl=0>. The simulation outputs are hosted by  
927 the DKRZ (Deutsches Klimarechenzentrum).

#### 928 **Author contribution**

929  
930 JL, JQ, DS and DK designed the study. JL wrote the code. DK provided support regarding the climate model data. JL conducted  
931 the analysis and JL, JQ and DS interpreted the results. JL prepared the manuscript, JQ, DS and DK reviewed the manuscript and  
932 provided comments.

#### 933 **Competing interests**

934  
935 Some authors are members of the editorial board of journal ACP.

#### 936 **Acknowledgements**

937  
938 This work was supported by the European Union's Horizon 2020 research and innovation programme under Marie  
939 Skłodowska-Curie grant agreement No. 860100 (iMIRACLI). We thank the Leipzig University Scientific Computing cluster and  
940 the DKRZ (Deutsches Klimarechenzentrum, projects number bb1036 and bb1153) for computing and data hosting. We  
941 acknowledge the contributors of the CUMULO dataset (Zantedeschi et al., 2019) for providing access to the data files hosted at  
942 <https://www.dropbox.com/sh/i3s9q2v2jyjk2it/AACxXnXfMF5wuIqLXqH4NJOra?dl=0>. Additionally, we acknowledge the  
943 MODIS L2 Cloud product data set from the Level-1 and Atmosphere Archive and Distribution System (LAADS) Distributed  
944 Active Archive Center (DAAC), located in the Goddard Space Flight Center in Greenbelt, Maryland  
945 ([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,  
946 Emilie Fons and Hans Segura for support and discussions in handling the climate model data.

947

## 948 References

949

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

953

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

956

957 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  
958 Cloud-Top Property Refinements for Collection 6, *Journal of Applied Meteorology and Climatology*, 51, 6, 1145-1163,  
959 <https://doi.org/10.1175/JAMC-D-11-0203.1>, 2012.

960

961 Bony, S., Semie, A., Kramer, R.J., Soden, B., Tompkins A.M., and Emanuel, K.A.: Observed modulation of the tropical radiation  
962 budget by deep convective organization and lower-tropospheric stability, *AGU Adv.*, Vol. 1, Issue 3,  
963 <https://doi.org/10.1029/2019av000155>, 2020.

964

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

969

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

971

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

973

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

976

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

979

980 Chen, T., Kornblith, S., Norouzi, M., and Hinton, G.: A simple framework for contrastive learning of visual representations, in:  
981 *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, *Journal of Machine Learning Research*, 119,  
982 1597–1607, <https://dl.acm.org/doi/10.5555/3524938.3525087>, 2020.

983

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

987

988 Dhuria, H. L. and Kyle, H. L.: Cloud Types and the Tropical Earth Radiation Budget, *J. Clim.*, 3, 1409–1434,  
989 [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.

990

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

994

995 Forster, P., T. Storelvmo, K. Armour, W. Collins, J.-L. Dufresne, D. Frame, D.J. Lunt, T. Mauritsen, M.D. Palmer, M. Watanabe,  
996 M. Wild, and H. Zhang: The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity, in *Climate Change 2021: The*  
997 *Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on*  
998 *Climate Change* [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I.  
999 Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)].

1000 Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 923–1054,  
1001 <http://doi.org/10.1017/9781009157896.009>, 2021.

1002

1003 García, V., Sánchez, J. S., and Mollineda, R. A.: On the effectiveness of preprocessing methods when dealing with different  
1004 levels of class imbalance, *Knowledge-Based Systems*, 25, 13–21, <https://doi.org/10.1016/j.knosys.2011.06.013>, 2012.

1005

1006 Hartmann, D. L., Ockert-Bell, M. E., and Michelsen, M. L.: The Effect of Cloud Type on Earth's Energy Balance: Global  
1007 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.

1008

1009 Hendrycks, D., and Gimpel, K.: Gaussian error linear units (gelus), arXiv preprint, <https://doi.org/10.48550/arXiv.1606.08415>,  
1010 2016.

1011

1012 Hinton, G. E.: Connectionist learning procedures, *Artificial intelligence*, 40, 185-234,  
1013 [https://doi.org/10.1016/0004-3702\(89\)90049-0](https://doi.org/10.1016/0004-3702(89)90049-0), 1989.

1014

1015 Hohenegger, C., Korn, P., Linardakis, L., Redler, R., Schnur, R., Adamidis, P., Bao, J., Bastin, S., Behraves, M., Bergemann,  
1016 M., Biercamp, J., Bockelmann, H., Brokopf, R., Brüggemann, N., Casaroli, L., Chegini, F., Datseris, G., Esch, M., George, G.,  
1017 Giorgetta, M., Gutjahr, O., Haak, H., Hanke, M., Ilyina, T., Jahns, T., Jungclaus, J., Kern, M., Klocke, D., Kluft, L., Kölling, T.,  
1018 Kornblueh, L., Kosukhin, S., Kroll, C., Lee, J., Mauritsen, T., Mehlmann, C., Mieslinger, T., Naumann, A. K., Paccini, L.,  
1019 Peinado, A., Praturi, D. S., Putrasahan, D., Rast, S., Riddick, T., Roeber, N., Schmidt, H., Schulzweida, U., Schütte, F., Segura,  
1020 H., Shevchenko, R., Singh, V., Specht, M., Stephan, C. C., von Storch, J.-S., Vogel, R., Wengel, C., Winkler, M., Ziemer, F.,  
1021 Marotzke, J., and Stevens, B.: ICON-Sapphire: simulating the components of the Earth system and their interactions at kilometer  
1022 and subkilometer scales, *Geosci. Model Dev.*, 16, 779–811, <https://doi.org/10.5194/gmd-16-779-2023>, 2023.

1023

1024 Howard, L.: *Essay on the modifications of clouds*, John Churchill & Sons, London, 64 pp., 1803.

1025

1026 Kaps, A., Lauer, A., Camps-Valls, G., Gentine, P., Gómez-Chova, L., and Eyring, V.: Machine-Learned Cloud Classes From  
1027 Satellite Data for Process-Oriented Climate Model Evaluation, *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-15,  
1028 4100515, <https://doi.org/10.1109/TGRS.2023.3237008>, 2023.

1029

1030 Kuma, P., Bender, F. A.-M., Schuddeboom, A., McDonald, A. J., and Seland, Ø.: Machine learning of cloud types in satellite  
1031 observations and climate models, *Atmos. Chem. Phys.*, 23, 523–549, <https://doi.org/10.5194/acp-23-523-2023>, 2023.

1032

1033 Kurihana, T., Moyer, E., Willett, R., Gilton, D.y, and Foster, I.: Data-Driven Cloud Clustering via a Rotationally Invariant  
1034 Autoencoder, *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-25, 4103325,  
1035 <https://doi.org/10.1109/TGRS.2021.3098008>, 2022.

1036

1037 LeCun, Y., Jackel, L. D., Boser, B., Denker, J. S., Graf, H. P., Guyon, I., Henderson, D., Howard, R. E., and Hubbard, W.:  
1038 Handwritten digit recognition: Applications of neural network chips and automatic learning, *IEEE Communications Magazine*,  
1039 Volume 27, Issue 11, 41-46, <https://doi.org/10.1109/35.41400>, 1989.

1040

1041 LeCun, Y., and Bengio, Y.: Convolutional networks for images, speech, and time series, *The handbook of brain theory and neural*  
1042 *networks*, 3361, 10, 1995.

1043

1044 Lemaitre, G., Nogueira, F., and Aridas, C., K.: Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets  
1045 in Machine Learning, *Journal of Machine Learning Research*, 18, 1-5, <http://jmlr.org/papers/v18/16-365.html>, 2017.

1046

1047 Luo, H., Quaas, J., and Han, Y.: Examining cloud vertical structure and radiative effects from satellite retrievals and evaluation of  
1048 CMIP6 scenarios, *Atmos. Chem. Phys.*, 23, 8169–8186, <https://doi.org/10.5194/acp-23-8169-2023>, 2023.

1049

1050 Luo, H., Quaas, J., and Han, J.: Diurnally asymmetric cloud cover trends amplify greenhouse warming, *Science Advances*, 10,  
1051 25, <https://doi.org/10.1126/sciadv.ado5179>, 2024.

1052  
1053 Lenhardt, J., Quaas, J., and Sejdinovic, D.: Marine cloud base height retrieval from MODIS cloud properties using machine  
1054 learning, *Atmos. Meas. Tech.*, 17, 5655–5677, <https://doi.org/10.5194/amt-17-5655-2024>, 2024a.  
1055  
1056 Lenhardt, J., Quaas, J., Sejdinovic, D., and Klocke, D.: CloudViT - Method code and data for the article "CloudViT: classifying  
1057 cloud types in global satellite data and in kilometre-resolution simulations using vision transformers.", Zenodo,  
1058 <https://doi.org/10.5281/zenodo.12731288>, 2024b.  
1059  
1060 McCoy, I. L., McCoy, D. T., Wood, R., Zuidema, P., and Bender, F. A. M.: The role of mesoscale cloud morphology in the  
1061 shortwave cloud feedback, *GRL*, 50, 2, <https://doi.org/10.1029/2022gl101042>, 2023.  
1062  
1063 Met Office: LAND SYNOP reports from land stations collected by the Met Office MetDB System, NCAS British Atmospheric  
1064 Data Centre, <https://catalogue.ceda.ac.uk/uuid/9f80d42106ba708f92ada730ba321831>, 2008.  
1065  
1066 Met Office: MIDAS: Global Marine Meteorological Observations Data, NCAS British Atmospheric Data Centre,  
1067 <https://catalogue.ceda.ac.uk/uuid/77910bcec71c820d4c92f40d3ed3f249>, 2006.  
1068  
1069 Muhlbauer, A., McCoy, I. L., and Wood, R.: Climatology of stratocumulus cloud morphologies: microphysical properties and  
1070 radiative effects, *Atmos. Chem. Phys.*, 14, 6695–6716, <https://doi.org/10.5194/acp-14-6695-2014>, 2014.  
1071  
1072 Oreopoulos, L., Cho, N., and Lee, D.: New insights about cloud vertical structure from CloudSat and CALIPSO observations, *J.*  
1073 *Geophys. Res.-Atmos.*, 122, 9280–9300, <https://doi.org/10.1002/2017JD026629>, 2017.  
1074  
1075 Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison,  
1076 A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S.:  
1077 PyTorch: An Imperative Style, High-Performance Deep Learning Library, in *Advances in Neural Information Processing*  
1078 *Systems* 32 (NeurIPS), 8024–8035,  
1079 <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>, 2019.  
1080  
1081 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg,  
1082 V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in  
1083 Python, *Journal of Machine Learning Research*, 12, 2825–2830, <https://www.jmlr.org/papers/v12/pedregosa11a.html>, 2011.  
1084  
1085 Pincus, R., Hubanks, P. A., Platnick, S., Meyer, K., Holz, R. E., Botambekov, D., and Wall, C. J.: Updated observations of clouds  
1086 by MODIS for global model assessment, *Earth Syst. Sci. Data*, 15, 2483–2497, <https://doi.org/10.5194/essd-15-2483-2023>, 2023.  
1087  
1088 Platnick, S., Ackerman, S. A., King, M. D., Meyer, K., Menzel, W. P., Holz, R. E., Baum, B. A., and Yang, P.: MODIS  
1089 atmosphere L2 cloud product (06\_L2), NASA MODIS Adaptive Processing System, Goddard Space Flight Center,  
1090 [http://doi.org/10.5067/MODIS/MYD06\\_L2.061](http://doi.org/10.5067/MODIS/MYD06_L2.061), 2017.  
1091  
1092 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:  
1093 algorithms and examples from Terra, in: *IEEE Transactions on Geoscience and Remote Sensing*, Volume 41, Number 2, 459–473,  
1094 <http://doi.org/10.1109/TGRS.2002.808301>, 2003.  
1095  
1096 Ramanathan, V., Cess, R. D., Harrison, E. F., Minnis, P., Barkstrom, B. R., Ahmad, E., and Hartmann, D.: Cloud Radiative  
1097 Forcing and Climate: Results from the Earth Radiation Budget Experiment, *Science*, 243, 57–63,  
1098 <https://doi.org/10.1126/science.243.4887.57>, 1989.  
1099  
1100 Rasp, S., Schulz, H., Bony, S., and Stevens, B.: Combining Crowdsourcing and Deep Learning to Explore the Mesoscale  
1101 Organization of Shallow Convection, *Bulletin of the American Meteorological Society*, 101, E1980–E1995,  
1102 <https://doi.org/10.1175/BAMS-D-19-0324.1>, 2020.  
1103

1104 Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N.,  
1105 Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015),  
1106 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.

1107

1108 Rossow, W.B., and Schiffer, R.A.: ISCCP cloud data products, *Bull. Amer. Meteorol. Soc.*, 71, 2-20, 1991.

1109

1110 Sassen, K., Wang, Z., and Liu, D.: Global distribution of cirrus clouds from CloudSat/Cloud-Aerosol Lidar and Infrared  
1111 Pathfinder Satellite Observations (CALIPSO) measurements, *J. Geophys. Res.*, Volume 113, D00A12,  
1112 <https://doi.org/10.1029/2008JD009972>, 2008.

1113

1114 Slingo, A.: Sensitivity of the Earth's radiation budget to changes in low clouds, *Nature*, 343, 49–51  
1115 <https://doi.org/10.1038/343049a0>, 1990.

1116

1117 Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K.,  
1118 Brokopf, R., Fast, I., Kinne, S., Kornblueh, L., Lohmann, U., Pincus, R., Reichler, T., and Roeckner, E.: Atmospheric component  
1119 of the MPI-M Earth System Model: ECHAM6, *Journal of Advances in Modeling Earth Systems*, 5, 2, 146-172,  
1120 <https://doi.org/10.1002/jame.20015>, 2013.

1121

1122 Stevens, B., Bony, S., Brogniez, H., Hentgen, L., Hohenegger, C., Kiemle, C., L'Ecuyer, T. S., Naumann, A. K., Schulz, H.,  
1123 Siebesma, P. A., Vial, J., Winker, D. M., and Zuidema, P.: Sugar, gravel, fish and flowers: Mesoscale cloud patterns in the trade  
1124 winds, *Q. J. R. Meteorol. Soc.*, Vol. 146, Issue 726, <https://doi.org/10.1002/qj.3662>, 2020.

1125

1126 Touvron, H., Vedaldi, A., Douze, M., and Jegou, H.: Fixing the train-test resolution discrepancy, 33rd Conference on Neural  
1127 Information Processing Systems (NeurIPS 2019), Vancouver, Canada, <https://doi.org/10.48550/arXiv.1906.06423>, 2019.

1128

1129 Tzallas, V., Hünerbein, A., Stengel, M., Meirink, J. F., Benas, N., Trentmann, J., Macke, A.: CRAAS: A European Cloud Regime  
1130 dAtAset Based on the CLAAS-2.1 Climate Data Record, *Remote Sensing*, 14, 5548, <https://doi.org/10.3390/rs14215548>, 2022.

1131

1132 Unglaub, C., Block, K., Mülmenstädt, J., Sourdeval, O., and Quaas, J.: A new classification of satellite-derived liquid water  
1133 cloud regimes at cloud scale, *Atmos. Chem. Phys.*, 20, 2407–2418, <https://doi.org/10.5194/acp-20-2407-2020>, 2020.

1134

1135 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I.: Attention Is All You  
1136 Need, *arXiv preprint*, <https://doi.org/10.48550/ARXIV.1706.03762>, 2017.

1137

1138 WMO: Manual on the observation of clouds and other meteors - International Cloud Atlas Volume I (WMO-No. 407), available  
1139 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.

1140

1141 WMO: Manual on the observation of clouds and other meteors - International Cloud Atlas (WMO-No. 407), available at:  
1142 <https://cloudatlas.wmo.int> (last access: 25 February 2025), 2017.

1143

1144 WMO: Manual on Codes, Volume I.1 – International Codes, Annex II to the WMO Technical Regulations, Part A –  
1145 Alphanumeric Codes (WMO-No. 306), ISBN: 978-92-63-10306-2, available at: <https://library.wmo.int/idurl/4/35713>, 2019.

1146

1147 Wood, R.: Stratocumulus clouds, *Monthly Weather Review*, 140, 8, 2373–2423, <https://doi.org/10.1175/MWR-D-11-00121.1>,  
1148 2012.

1149

1150 Wood, R., and Hartmann, D. L.: Spatial variability of Liquid water path in marine low cloud: The importance of mesoscale  
1151 cellular convection, *J Clim*, 19, 9, 1748–1764, <https://doi.org/10.1175/jcli3702.1>, 2006.

1152

1153 Young, A. H., Knapp, K. R., Inamdar, A., Hankins, W., and Rossow, W. B.: The International Satellite Cloud Climatology  
1154 Project H-Series climate data record product, *Earth Syst. Sci. Data*, 10, 583–593, <https://doi.org/10.5194/essd-10-583-2018>,  
1155 2018.

1156

1157 Yuan, T., Song, H., Wood, R., Mohrmann, J., Meyer, K., Oreopoulos, L., and Platnick, S.: Applying deep learning to NASA  
1158 MODIS data to create a community record of marine low-cloud mesoscale morphology, *Atmos. Meas. Tech.*, 13, 6989–6997,  
1159 <https://doi.org/10.5194/amt-13-6989-2020>, 2020.

1160

1161 Zantedeschi, V., Falasca, F., Douglas, A., Strange, R., Kusner, M. J., and Watson-Parris, D.: Cumulo: A Dataset for Learning  
1162 Cloud Classes, *Tackling Climate Change with Machine Learning Workshop*, 33rd Conference on Neural Information Processing  
1163 Systems (NeurIPS 2019), Vancouver, Canada, <https://doi.org/10.48550/arXiv.1911.04227>, 2019.

1164

1165 Zhang, J. L., Liu, P., Zhang, F., & Song, Q. Q.: CloudNet: Ground-based cloud classification with deep convolutional neural  
1166 network, *Geophysical Research Letters*, 45, 8665–8672, <https://doi.org/10.1029/2018GL077787>, 2018.

1167

1168 Zhao, H., Gallo, O., Frosio, I., and Kautz, J.: Loss functions for image restoration with neural networks, *IEEE Transactions on*  
1169 *computational imaging*, 3, 1, 47–57, <https://doi.org/10.1109/TCI.2016.2644865>, 2016.