# EARTHCAREBENCH: A COLOCATED MTG - EARTHCARE ML-READY DATASET FOR DERIVING THE 3D STRUCTURE OF THE ATMOSPHERE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Clouds are a critical component of the Earth's climate system, yet their complex 3D structure and temporal variability pose serious observational and modeling challenges. Observations of the EarthCARE and MTG satellite missions may be utilized in a synergistic way to address these challenges. EarthCARE satellite acquires 2D top-of-atmosphere (TOA) images, along with information of the vertical structure of the atmosphere along its ground track. MTG satellite provides geostationary 2D TOA imager observations, with high temporal resolution, over a wide area.

We introduce EarthCAREBench, the first ML-ready dataset of spatially and temporally colocated MTG-EarthCARE scenes, for the full year of 2025, available in three different versions that correspond to three maximum temporal colocation windows (2, 3, and 4 minutes). The dataset is provided for a geographic extent of ±45° around the MTG subsatellite point, in order to minimize the parallax effects. Each sample consists of fixed size of 128×128 pixels of MTG FCI multispectral observations and cloud mask product, along with the colocated overpass of EarthCARE, with profiles of cloud types and cloud microphysical properties. EarthCAREBench enables cloud representation learning, vertically-resolved atmospheric retrievals through the combination of passive and active satellite observations, along with investigating the effect of temporal colocation in these retrievals. We establish cloud type segmentation and microphysical retrieval baselines, evaluated using F1 score and IoU for classification, and SSIM and PSNR for the regression task. The code and EarthCAREBench dataset will be made publicly available.

## 1 INTRODUCTION

Clouds cover roughly two thirds of the globe (King et al., 2013) and are the main modulator of Earth's energy budget. Yet their response to cooling or warming remains the largest source of uncertainty in estimates of climate sensitivity (Boucher et al., 2013; Zelinka et al., 2020). That response depends not only on horizontal extent but also on the *vertical* distribution of cloud phase and water content, which also matters for aviation icing, precipitation estimation and surface solar energy forecasting (Cao et al., 2018; Luo et al., 2023; Sun et al., 2025; Papachristopoulou et al., 2024).

No single instrument observes this 3D structure at scale. Active remote sensing profilers such as CloudSat (Stephens et al., 2002), CALIPSO (Winker et al., 2009) and, since 2024, EarthCARE (Illingworth et al., 2015; Wehr et al., 2023) resolve clouds vertically at $\sim$100 m, but only along a $\sim$1 km-wide nadir curtain. Geostationary imagers such as the Flexible Combined Imager (FCI) on Meteosat Third Generation (MTG) (Holmlund et al., 2021) capture a large part of the Earth every 10 minutes, providing 2D TOA observations in 16 spectral channels. These observations do not directly resolve the vertical structure of clouds, containing information mainly for their top part. FCI's 1-2 km sampling is close to EarthCARE's 1 km along-track grid, which enables pixel-to-profile matching with little resampling. The matched pairs provide supervision for learning to map 2D geostationary radiances to the vertical cloud fields, in order to provide the 3D structure and microphysical properties of cloudy scenes.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

From a machine-learning perspective, this mapping poses three challenges. First, **supervision is extremely sparse**: the model predicts a dense volume ($8$ classes $\times$ $242$ levels $\times$ $128^2$ pixels, with the classes to contain the cloud types and microphysical properties, and the levels to correspond to the height levels of the vertical profiles), but labels exist only along the EarthCARE track, on about $130$ of $16{,}384$ pixels per patch ($<1\%$). Second, **labels are misaligned in time**: each label carries an MTG-EarthCARE time offset $\Delta t$ during which clouds may evolve, so the tolerance on $\Delta t$ trades dataset size against label uncertainty. Third, **targets are heavily imbalanced**: clear sky dominates the classification, and only $\sim 7\%$ and $\sim 2\%$ of ice and liquid microphysical property retrievals, respectively, are non-zero.

Existing data-driven 3D cloud retrievals pair observations acquired with older imagers (e.g. MODIS or MSG/SEVIRI) with observations from CloudSat and/or CALIPSO (Leinonen et al., 2019; Brüning et al., 2024; Jeggle et al., 2025; Girtsou et al., 2025). Usually, these retrievals target a single quantity, such as the reflectivity or the ice water content of the clouds, and do not release a standardized benchmark. Public datasets such as CUMULO (Zantedeschi et al., 2022) and 3S-GEOPROF-COMB (Bertrand et al., 2023) address 2D cloud typing or gridded climatologies instead. C3DIR (White et al., 2026) utilizes for its training observations from GOES-R and MSG geostationary satellites, along with EarthCARE products, but it does not releases its training data. In summary, no public benchmark yet pairs MTG/FCI observations with EarthCARE products, and none couples cloud classification with microphysics.

We introduce EarthCAREBench, the first public, ML-ready MTG-EarthCARE dataset, including cloud classification and microphysical properties. Our contributions are:

- **Dataset.** Patches containing $128\times128$ FCI pixels (radiances in 14 channels and the cloud mask product). Each patch is paired with the coincident EarthCARE products of cloud classification and microphysical property profiles (AC-TC and ACM-COM L2b products, respectively). The dataset includes all available data of 2025, within $\pm 45^\circ$ of the MTG sub-satellite point (section 3).
- **Controlled temporal alignment.** We used three different MTG/EarthCARE colocation times, of $|\Delta t| \leq 2$, 3 and 4 minutes (resulting in 31,380 / 46,272 / 61,487 patches, respectively), enabling the assessment of the effect of the temporal difference between MTG and EarthCARE observations on learning (section 3.1).
- **Benchmark.** We provide benchmark performances for the 3D target classification (F1, IoU) and cloud microphysical property regression (SSIM, PSNR), under a leakage-free temporal split, with baselines analyzed across $\Delta t$ and minority classes (section 4).
- **Open pipeline.** The available code regenerates the dataset for new MTG and EarthCARE data, along for 2025 EarthCARE data processed with new processing baselines.

## 2 Related work

Volumetric reconstruction from 2D measurements is most mature in medical tomography (Cline et al., 1987; Khan et al., 2018; Ying et al., 2019); for clouds, multi-angle stereo, ray-casting and ML-based scattering tomography recover cloud-top geometry and, in some cases, the internal volume (Seiz & Davies, 2006; Lee et al., 2018; Yu et al., 2019; Dandini et al., 2022; Sde-Chen et al., 2021; Ronen et al., 2025). All of these methods, however, require simultaneous multi-angle views, which are rarely available from space.

A second line of work extends the narrow active-sensor curtains into the surrounding passive imagery. Early approaches were statistical or physics-guided, such as the EarthCARE scene-construction algorithm (Barker et al., 2011), and the cloud-type-dependent CloudSat-MODIS extrapolation (Miller et al., 2014). Deep learning dominates the more recent approaches. Leinonen et al. (2019) generated CloudSat reflectivity curtains from MODIS with a conditional GAN. Later work utilized geostationary imagers: retrieving ice water content from MSG/SEVIRI observations, supervised by combined CALIPSO/Cloudsat retrievals (Jeggle et al., 2025); cloud reflectivity with a Res-UNet (Brüning et al., 2024); geospatially-aware MAE pretraining (Girtsou et al., 2025); Himawari-8 cloud-phase profiles (Yang et al., 2025; 2026); Fengyun-4A cloud amount (Lin et al., 2025); and GOES-R reflectivity with Transformers (Zhang et al., 2026). Ermis et al. (2026) combined three geostationary imagers with CloudSat and provided a tropical-cyclone benchmark. Most recently, White et al. (2026) trained geostationary observations from GOES-R and MSG on

Table 1: Comparison of ML-ready EO datasets relevant to 3D atmospheric reconstruction. "cls." denotes classification task, as the specific label space varies with dataset (i.e., cloud type, cloud phase, or hydrometeor detection). "reg." denotes regression.

| Dataset | Passive imager | Active / labels | Task | Public |
|---|---|---|---|---|
| CUMULO (Zantedeschi et al., 2022) | MODIS | CloudSat 2B-CLDCLASS | Cloud-type cls. | ✓ |
| SEN12MS-CR-TS (Ebel et al., 2022) | Sentinel-2 | Sentinel-1 SAR | Cloud removal (reg.) | ✓ |
| 3S-GEOPROF-COMB (Bertrand et al., 2023) | — | CloudSat + CALIPSO | Cloud fraction (grid) | ✓ |
| Brüning et al. (Brüning et al., 2024) | MSG/SEVIRI | CloudSat/CPR ($Z$) | Reflectivity reg. | ✗ |
| SGMAGNet (Yang et al., 2025) | Himawari-8 | CALIOP + CPR | Cloud-phase cls. | ✗ |
| Girtsou et al. (Girtsou et al., 2025) | MSG/SEVIRI | CloudSat/CPR ($Z$) | Reflectivity reg. | ✗ |
| IceCloudNet (Jeggle et al., 2025) | MSG/SEVIRI | DARDAR (radar-lidar) | Detection cls. + IWC/$N_{\text{ice}}$ reg. | ✓ |
| Ermis et al. (Ermis et al., 2026) | SEVIRI/AHI/ABI | CloudSat + CALIPSO | Microphysics reg. | ✓ |
| C3DIR (White et al., 2026) | Multi-imager | EarthCARE | Detection cls. + water-content reg. | ✗ |
| **EarthCAREBench** | MTG | EarthCARE (AC-TC + ACM-COM) | **Cloud-type cls. + microphysics reg.** | ✓ |

EarthCARE cloud products, marking the shift from CloudSat/CALIPSO to EarthCARE as the label source.

Despite this methodological progress, training data remain a bottleneck. Existing public datasets address narrower or related tasks, rely on pre-EarthCARE active sensors, or they are not released as reusable ML-ready benchmarks

EarthCARE offers the most complete observational suite yet for cloud and aerosol studies and has been identified as the key platform for future 3D cloud products (Noh et al., 2022; Girtsou et al., 2025). Table 1 positions EarthCAREBench among these datasets.

## 3 Dataset Description

EarthCAREBench pairs $128 \times 128$ MTG/FCI pixels with the vertically-resolved EarthCARE products of cloud-type classification and cloud, precipitation and aerosol microphysical properties along the satellite track. Three versions are provided, with maximum MTG-EarthCARE time differences of 2, 3 and 4 minutes. This section describes the source products, quality control, colocation, sample format and dataset statistics. Note that the cloud-type classification (with cloud types following EarthCARE's naming convention e.g. "liquid", "ice cloud", "supercooled liquid", etc) corresponds to cloud-phase classification, with the latter more commonly used in the literature.

### 3.1 Source Data

MTG/FCI. The Flexible Combined Imager (FCI) aboard the Meteosat Third Generation (MTG) satellites scans the full disk, covering Europe, Africa, the Middle East and the surrounding oceans every 10 minutes, providing observations in 16 spectral channels. Its nominal sampling distance at nadir is 1-2 km (Holmlund et al., 2021). EarthCAREBench uses the FCI Level-1c product, which provides radiometrically calibrated radiances and brightness temperatures on a common reference grid, for 14 of the 16 FCI channels (table 5 in section A). The two water-vapour channels (WV6.3, WV7.3) are excluded, since WV6.3 contained daily periods of invalid data until November 2024 (EUMETSAT, 2024), and both channels were unavailable in many scenes examined in 2025. Channels with a native sampling distance of 2 km are upsampled to the common 1 km grid with bilinear interpolation. The Level-2 FCI cloud mask is added in EarthCAREBench as an auxiliary field, upsampled with nearest-neighbour interpolation. All fields are then extracted in patches of $128 \times 128$ pixels.

EarthCARE. The EarthCARE mission was launched on May 29, 2024 and is a Low Earth Orbit (LEO) satellite, in a sun-synchronous orbit, providing global coverage with a 25-day repetition cycle. EarthCARE carries the ATmospheric LIDar (ATLID), the Cloud-Profiling Radar (CPR), the MultiSpectral Imager (MSI) and the BroadBand Radiometer (BBR), whose observations are processed individually and synergistically and produce cloud, aerosol, precipitation and radiation products (Wehr et al., 2023; Eisinger et al., 2024). EarthCAREBench includes the AC-TC and ACM-COM Level-2b synergistic products as reference targets. **AC-TC** utilizes ATLID and CPR observations to classify each target in the atmospheric profile into cloud, aerosol and precipitation types, with accompanying quality flags (Irbah et al., 2023). The product is provided on the Joint

Standard Grid (JSG), with $\sim$1 km horizontal sampling and $\sim$0.1 km vertical sampling up to 20 km ($\sim$0.5 km up to 40 km). It is also provided in medium- and low-resolution versions that smooth weak features horizontally over 40 and 150 km, respectively (Donovan et al., 2024). **ACM-COM** provides aerosol/cloud/precipitation products using (a) observations from single instruments, and (b) the synergy of the lidar/radar/imager observations (i.e., the ACM-CAP L2b product). Specifically, the ACM-CAP product provides the best-estimate variational retrieval (CAPTIVATE) of clouds, aerosols and precipitation from ATLID, CPR and MSI that accounts for observational and retrieval uncertainties. The retained AC-TC and ACM-CAP variables are listed in tables 6 and 7 in Appendix B.

**Quality Control** EarthCARE Level-2 products are subject to observational and retrieval uncertainties. EarthCAREBench retains the quality flags of the AC-TC and ACM-CAP products, which mark missing and unreliable values. In the baseline experiments, only EarthCARE-observed voxels with valid values contribute to the loss and evaluation metrics. The distributed flags allow users to apply stricter filtering as needed.[1]

**Colocation** Geostationary viewing geometry becomes increasingly oblique towards the edge of the MTG disk, which displaces the apparent position of clouds relative to the surface (parallax effect). To limit this effect, EarthCAREBench is restricted to $\pm 45^\circ$ latitude and longitude around the MTG sub-satellite point. The parallax effect is not corrected explicitly, but this allows also for colocations of better quality. At the edge of this domain, residual misalignment remains for high clouds.

The alignment between MTG/FCI imagery and EarthCARE products exploits the near-simultaneous view of the same atmospheric column from the two platforms. FCI scans the disk in chunks over 10 minutes, so different latitudes are acquired at different times. Each FCI pixel is therefore assigned the mid-scan time of its chunk, while each EarthCARE along-track pixel keeps its own acquisition time. The EarthCARE ground track is projected onto the FCI geostationary grid with pyproj, and an FCI grid cell is marked as "colocated" only if its acquisition time lies within $\Delta t$ of the matching EarthCARE pixel. The colocated cells form the overpass mask, which defines where supervision is available, and their time offsets are stored alongside. We release three variants with $\Delta t = 2$, 3 and 4 minutes, which trade sample count against temporal consistency: larger time offsets may introduce different evolution of clouds between the two dataset and produce mismatched labels (Fig. 2, Appendix C.1).

Patches of $128 \times 128$ pixels are then extracted along each overpass. Patch centers are placed along the EarthCARE track every 103 pixels, giving a 20% overlap between consecutive patches, and shifted across the EarthCARE track by a random offset of up to $\pm 25\%$ of the patch width, so the EarthCARE track does not always cross the patch centre.

**ML-ready Format** Each sample preserves the native geometry of both sensors. The FCI component is a complete $128 \times 128$ patch, $\mathbf{X}_{\mathrm{FCI}} \in \mathbb{R}^{14\times 128\times 128}$, stored with its latitude and longitude, acquisition time, FCI cloud state, the `overpass_mask` of colocated pixels and the `overpass_time_diff` holding their MTG-EarthCARE time offsets. EarthCARE retrievals are stored only along the track, as $N_{\mathrm{pix}}$ vertical profiles of size $N_{\mathrm{pix}} \times N_{\mathrm{height}}$ on each variable's native vertical grid. Each profile is indexed by its position $(x_i, y_i)$ within the patch, and $N_{\mathrm{pix}}$ varies between samples with the crossing geometry. This compact curtain representation avoids storing mostly empty 3D volumes. All retained AC-TC and ACM-CAP variables, including quality and uncertainty fields, are distributed with each sample, independently of the subset used by the baselines.

Samples are serialized in Zarr for chunked, parallel access. Each sample is an independent DataTree with four groups (table 2): `/mtg` (FCI patch and ancillary fields), `/ac_tc` and `/com` (AC-TC and ACM-CAP profiles), and `/metadata` (source file identifiers, patch position and geographic bounds, so that every sample can be traced back to the original products). The dataset and code will be released upon publication.

[1] For the known quality limitations of the products, see the EUMETSAT FCI validation report (EUMETSAT, 2024) and the ESA processor disclaimers for AC-TC, ACM-CAP, ATLID and CPR (European Space Agency, n.d.b;n;n;n).

EarthCAREBench therefore provides the complete set of variables, which are not restricted to a particular machine-learning task. The complete set of retained AC-TC and ACM-CAP variables, including atmospheric target classification, microphysical properties, quality-control, and uncertainty fields, is distributed with each sample. The dataset and the code for loading, preprocessing, visualization and the benchmark experiments will be released upon publication.

Table 2: Organization of the main groups contained in each EarthCAREBench sample.

| DataTree group | Main contents | Main dimensions |
|---|---|---|
| `/mtg` | 14-band FCI data, cloud state, latitude/longitude, acquisition time, overpass mask, and temporal offset | band $\times\, y \times x$, with $y = x = 128$ |
| `/ac_tc` | AC-TC target classification, height, time, and quality information | Npix $\times$ JSG_height, with JSG_height $= 242$ |
| `/com` | Retained ACM-CAP cloud and aerosol optical and microphysical properties, and quality information | Npix $\times$ level_number or Npix $\times$ layer_number |
| `/metadata` | Source-product provenance, patch indices, spatial bounds, and sample metadata | Sample-level/scalar |

### 3.2 Statistical Analysis

Relaxing the colocation window from 2 to 4 minutes nearly doubles the dataset, from 31,380 to 61,487 patches (46,272 for 3 minutes). The variants are built independently rather than nested, so a smaller-window variant is not a strict subset of a larger one; each is populated separately to ensure adequate spatial and temporal coverage. Samples span the whole year of 2025, although the period of October to December contains slightly fewer patches, due to the lower availability of the EarthCARE BA processing baseline. he statistics below refer to the 4-minute variant, the largest of the three.

**Geographical coverage** The dataset covers Africa, the eastern Atlantic Ocean, southern Europe, and adjacent regions, with the highest sample densities occurring across tropical and subtropical Africa and the adjacent Atlantic. The non-uniform distribution reflects EarthCARE orbital sampling, the MTG viewing geometry, and the spatial and temporal sampling criteria. A relevant visualization is provided in Fig. 1, Appendix B. The dataset contains $52.3\%$ daytime and $47.7\%$ nighttime observations, providing a balanced representation of illumination conditions.

**Temporal colocation and EarthCARE coverage** The absolute MTG-EarthCARE acquisition-time differences are approximately uniformly distributed within the permitted 240 s interval. The number of colocated EarthCARE profiles per patch, $N_{\mathrm{pix}}$, varies with the overpass geometry and is concentrated around 120-140 profiles. The corresponding distributions are provided in Appendix C.1.

**AC-TC product class distribution** The native EarthCARE AC-TC product classification exhibits a pronounced class imbalance, with clear atmosphere dominating the atmospheric bins and several cloud, aerosol, and precipitation categories occurring much less frequently. This imbalance motivates the reduced eight-class scheme adopted for the benchmark (Table 9). Nevertheless, substantial imbalance remains after the mapping and is accounted for in the benchmark loss formulation and class-wise evaluation. The distributions of the original standard- and low-resolution AC-TC product classifications are presented in Appendix C.2.

**ACM-CAP product microphysical properties** The ACM-CAP retrieved microphysical properties exhibit substantial sparsity and zero inflation. Among valid retrievals, approximately $92.94\%$ of ice-effective-radius and ice-water-content values are zero, compared with $97.93\%$ for the corresponding liquid-cloud properties. Approximately $4.56\%$ of the values are invalid (i.e., "NaN") for these four variables.

This distinction between invalid and zero-valued retrievals is essential for the regression benchmark, detailed statistics are provided in Appendix C.3.

# 4 BASELINE

To demonstrate the use of EarthCAREBench for vertically-resolved atmospheric retrievals, we define two supervised benchmark tasks: (a) the atmospheric target classification, and (b) the cloud microphysical property retrieval. Both tasks use the multispectral MTG FCI observations as input and the EarthCARE AC-TC and ACM-CAP products as reference targets. A common residual U-Net (Ronneberger et al., 2015) baseline is used to provide an initial performance reference for the dataset.

## 4.1 TASKS

**Segmentation task** The first benchmark task focuses on predicting the vertically-resolved atmospheric classification, using as labels the AC-TC product, which is reduced in EarthCAREBench in eight classes (Table 9, Appendix D), resulting in a classification scheme tailored to cloud retrieval applications.

For an input FCI patch $\mathbf{X}$, the model estimates class logits $\hat{\mathbf{Y}}_{\text{seg}} \in \mathbb{R}^{C\times Z\times 128\times 128}$, where $C = 8$ is the number of atmospheric classes. The predicted atmospheric class at each voxel is obtained as $\hat{c}(z, y, x) = \arg\max_c \hat{Y}_{\text{seg}}(c, z, y, x)$.

**Regression task** The second benchmark task focuses on retrieving vertically-resolved cloud microphysical properties from the FCI observations. Four cloud microphysical properties provided by the ACM-CAP product in EarthCAREBench are used as benchmark regression targets, $\mathbf{v} = [r_{\text{e,ice}}, \text{IWC}, r_{\text{e,liq}}, \text{LWC}]$, corresponding to ice effective radius, ice water content, liquid effective radius, and liquid water content, respectively. For $V = 4$ regression variables, the model predicts $\hat{\mathbf{Y}}_{\text{reg}} \in \mathbb{R}^{V\times Z\times 128\times 128}$.

Each ACM-CAP variable is aligned to $Z = 242$ vertical positions on its native vertical grid (layer midpoints), so segmentation and regression outputs are interpreted on the AC-TC level heights and ACM-CAP layer heights, respectively.

## 4.2 BASELINE ARCHITECTURE

Our baseline is a two-dimensional residual U-Net (ResUNet), following the design used for 3D cloud reconstruction from MSG/SEVIRI by Brüning et al. (2024) and Girtsou et al. (2025). The network encodes a single FCI observation of shape $[B, 14, 128, 128]$ into a shared spatial feature map, and lightweight task-specific heads expand this map along the vertical dimension. Because the vertical structure is predicted as output channels rather than through 3D convolutions, the model produces a dense volume for the whole patch at the cost of a 2D network.

**Shared ResUNet backbone.** The encoder has three residual downsampling stages with $2 \times 2$ max pooling. The first stage maps the $14$ input channels to $32$ feature maps, and each subsequent stage doubles the number of filters. A residual bridge connects the encoder to a symmetric decoder, which restores the input resolution and receives the corresponding encoder features through skip connections. With a dropout probability of $0.5$, the backbone outputs a shared feature tensor $\mathbf{F} \in \mathbb{R}^{32\times 128\times 128}$.

**Prediction heads.** Each task head is a $1 \times 1$ convolution on $\mathbf{F}$ whose output channels are reshaped into a vertical dimension. The classification head produces $C \times Z = 8 \times 242 = 1936$ channels per pixel, reshaped to $[B, C, Z, H, W]$. The regression head produces $V \times Z = 4 \times 242 = 968$ channels, reshaped to $[B, V, Z, H, W]$ and followed by a ReLU that keeps the ice/liquid effective radius and water content non-negative. An optional auxiliary head ($3 \times 3$ convolution, batch normalization, ReLU and a $1 \times 1$ classifier) predicts the 2D FCI cloud mask, $\hat{\mathbf{Y}}_{\text{CM}} \in \mathbb{R}^{2\times 128\times 128}$. Following White et al. (2026), this head adds supervision over the full patch, including pixels where EarthCARE

labels are unavailable. The heads share the backbone and can be trained jointly or individually (segmentation-only, regression-only, or multitask, each with or without the cloud-mask head).

**Pretrained ViT.** As an additional, preliminary baseline, we fine-tune a ViT-B encoder (patch size 8) pretrained on unlabelled FCI imagery with the masked-autoencoder strategy of Girtsou et al. (2025). Its multi-layer features are decoded to $128 \times 128$ and passed to the same prediction heads as the ResUNet. A full description of the pretraining and a systematic evaluation will be presented in a follow-up study.

### 4.3 Experimental Setup

**Data Splits** Satellite imagery exhibits strongly spatiotemporal autocorrelations. Patches from the same EarthCARE overpass observe the same cloud systems, and scenes acquired a few hours apart can share the same mesoscale weather. A random split would therefore place highly-correlated samples in both the training and the evaluation sets, which would lead to overly optimistic scores. We instead use a temporal split by calendar day. In every month of 2025, days 1-16 go to training ($\sim$53% of days), days 17-23 to validation ($\sim$23%) and days 25 to the end of the month to testing ($\sim$21%). Day 24 is left out as a buffer. All partitions cover every season and the full annual cycle of illumination and cloud regimes. The same day-based split is applied to the 2-, 3- and 4-minute variants, and the split definitions are distributed with the dataset so that future methods can be evaluated on identical partitions. Table 3 lists the resulting partition sizes.

Table 3: Dataset partitions and storage volumes for the three temporal colocation variants of EarthCAREBench.

| Colocation window | Total samples | Train | Validation | Test | Volume (GB) |
|---|---|---|---|---|---|
| 2 min | 31,380 | 21,259 | 5,778 | 3,680 | – |
| 3 min | 46,272 | 31,329 | 8,642 | 5,288 | – |
| 4 min | 61,487 | 41,193 | 11,244 | 7,023 | – |

**Supervision.** FCI inputs are standardized per channel using the mean and standard deviation of the training data, and missing values are replaced by the patch mean. EarthCARE provides labels only along its ground track, so all losses and metrics are computed on the supervised voxels. These are the voxels that lie inside the overpass mask and hold a valid EarthCARE retrieval. All other voxels are assigned an ignore label for classification and masked out for regression. The microphysical targets span several orders of magnitude, so each variable is transformed with $\log(1 + y/s)$ and min-max scaled to $[0, 1]$ using training-set statistics. Physical zeros therefore remain exactly zero and stay distinct from missing retrievals.

**Segmentation loss.** To address the strong class imbalance of the AC-TC product target classification (Appendix C.2), the segmentation head is trained with an equally-weighted sum of a focal loss (Lin et al., 2017) and a Tversky loss (Salehi et al., 2017). Both losses use the same per-class weights, $w = (0.3, 1.5, 5.5, 3.8, 1.5, 0.7, 0.6, 0.5)$ for the eight classes in table 9. These were initialised from the inverse class frequencies in the training split and then adjusted manually based on validation performance in preliminary experiments. They were not tuned systematically, so a principled weighting scheme may further improve performance on the minority classes. The focal loss ($\gamma = 2$) down-weights the easy well-classified voxels, which are mostly clear sky. The Tversky loss ($\alpha = 0.3$, $\beta = 0.7$) penalises false negatives more than false positives, which favors recall of the rare cloud classes.

**Regression loss.** More than 90% of valid microphysical values are zero, so a plain MSE is minimised by predicting zero everywhere. Each of the four cloud microphysical properties is therefore trained with two terms: a weighted MSE that fits the values, and a soft Dice term that fits the spatial extent of cloudy regions. In the weighted MSE, non-zero targets are weighted five times more than zeros. The soft Dice term compares a sigmoid-relaxed prediction mask with the target mask above a threshold of $0.1$ in normalised units. The two terms are combined as $10\,\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{Dice}}$. The loss is computed per sample and averaged over variables.

**Multi-task retrieval.** The auxiliary cloud-mask head is trained with a pixel-wise cross-entropy against the binary FCI cloud mask over the full $128\times128$ patch. The total loss is the unweighted sum of the segmentation, regression and cloud-mask losses. Single-task models drop the unused terms.

**Training and checkpoint selection.** All models are trained with AdamW (Loshchilov & Hutter, 2019) at a constant learning rate of $1.5 \times 10^{-4}$. The ResUNet is trained for 60 epochs with batch size 10. After every epoch we evaluate on the validation split and keep the checkpoints with the lowest validation loss and with the highest validation mIoU, F1, PSNR and SSIM, as well as the last epoch. Test results are reported for the checkpoint with the highest highest PSNR. All experiments were run on a single NVIDIA GeForce RTX 3090 Ti GPU (24 GB). The training of the multitask ResUNet takes approximately three days.

**Evaluation** To isolate the effect of the colocation window, all models, whether trained on the 2-, 3- or 4-minute variant, are evaluated on the same held-out test set: the test split of the 2-minute variant, whose small temporal offsets make its labels the least affected by cloud evolution in time. Because all variants share the same calendar-day split, this test set contains no days seen during training by any model.

Segmentation is evaluated with overall accuracy, macro-averaged F1 and mean IoU (mIoU) over the eight classes, together with per-class F1 and IoU. Counts are accumulated over the entire split before averaging, so every class contributes equally regardless of its frequency. Regression is evaluated with SSIM (Wang et al., 2004) and PSNR, computed for the different cloud microphysical properties on the EarthCARE curtain (height $\times$ along-track profiles) of each sample in normalised units, and then averaged.

Because the microphysical targets are dominated by zero values, which SSIM and PSNR largely reward, we additionally evaluate the regression more thoroughly with the detection F1 score of cloudy voxels ($\tau = 0.1$ in normalized units) and the RMSE computed only on voxels with non-zero targets; these results are reported in Appendix section F.

### 4.4 Results

Performance differs strongly across classes and largely reflects two factors: how frequent each class is in the training data, and how directly it can be observed by a passive imager. Clear sky, the majority class, is recovered reliably by all models (F1 $\geq 0.88$). Among the cloud classes, ice clouds reach the highest F1 ($0.32$-$0.40$ for the ResUNet). They usually form the uppermost layer, so FCI observes them directly through their cold thermal-infrared signature. Liquid clouds are considerably harder ($0.10$-$0.21$), as they lie lower in the atmosphere and are frequently partly or completely hidden beneath ice layers, so their presence must be inferred rather than observed. The mixed-phase classes, supercooled liquid and ice with supercooled liquid, remain challenging ($0.14$-$0.17$): they are among the rarest classes and are defined by the thermodynamic phase inside the cloud rather than at its top. Both aerosol classes score comparatively well ($0.69$-$0.71$ for tropospheric and $0.49$-$0.64$ for stratospheric aerosols). Tropospheric and stratospheric aerosols occupy a narrow, persistent altitude range that the model can learn largely from the vertical coordinate.

The ResUNet reaches SSIM values of $0.85$-$0.89$ and PSNR values of $28.4$-$29.2$ dB. Because more than 90% of valid microphysical values are zero, these scores are strongly influenced by empty bins. We therefore additionally report the detection F1 of cloudy voxels and the RMSE on non-zero targets in table 10.

Differences between the colocation windows are moderate. The 3-minute model reaches the highest PSNR ($29.18$ dB, vs. $28.40$ and $28.43$ dB for 2 and 4 minutes) and the highest F1 for ice ($0.40$) and liquid clouds ($0.21$), while the 2-minute model performs best for the mixed-phase and stratospheric-aerosol classes. SSIM increases slightly with window size ($0.853$, $0.860$, $0.890$). The 3-minute window thus appears to offer the best overall compromise between sample count and label consistency: the 2-minute variant provides fewer training samples, whereas the 4-minute variant admits labels with larger temporal offsets, during which clouds may advect or evolve (Fig. 2). This is most visible for liquid clouds, whose F1 drops from $0.21$ to $0.10$ between the 3- and 4-minute models,

Table 4: Benchmark results on the held-out 2-minute test set. $\Delta t$ denotes the colocation window of the training data. All results are reported for the checkpoint with the highest validation PSNR; columns 3-9 give per-class F1 for atmospheric target classification, and SSIM and PSNR evaluate the microphysical retrieval. Best values in bold. ViT results are preliminary.

| Model | Train $\Delta t$ | Clear sky | Ice cloud | Liquid cloud | Supercooled liquid | Ice & supercooled liquid | Tropospheric aerosol | Stratospheric aerosols | SSIM | PSNR |
|---|---|---|---|---|---|---|---|---|---|---|
| ResUNet | 2 min | 0.895 | 0.320 | 0.180 | 0.140 | 0.170 | 0.707 | 0.641 | 0.853 | 28.40 |
| | 3 min | 0.896 | 0.400 | 0.210 | 0.150 | 0.150 | 0.694 | 0.485 | 0.860 | 29.18 |
| | 4 min | 0.882 | 0.364 | 0.101 | 0.143 | 0.155 | 0.709 | 0.613 | 0.890 | 28.43 |
| ViT + MAE pretraining | 3 min | **0.930** | **0.589** | **0.352** | **0.207** | **0.286** | **0.722** | **0.666** | **0.910** | **30.53** |

consistent with low, small-scale liquid clouds being particularly sensitive to temporal misalignment. All results are from single training runs.

The ViT initialized with masked-autoencoder pretraining outperforms the ResUNet trained on the same 3-minute variant on every class and metric. It improves F1 for ice clouds from $0.40$ to $0.59$, for liquid clouds from $0.21$ to $0.35$, for supercooled liquid from $0.15$ to $0.21$ and for ice with supercooled liquid from $0.15$ to $0.29$, and raises PSNR from $29.2$ to $30.5$ dB and SSIM from $0.86$ to $0.91$. The largest relative gains occur for the mixed-phase and liquid classes, suggesting that representations learned from unlabelled geostationary imagery can partly compensate for the sparse active-sensor supervision. A systematic study of pretraining strategies and a detailed statistical analysis of the reconstructions are left for future work. Example retrievals of the three models for the same scene are shown in section F.

## 5 Summary and Next Steps

We introduced EarthCAREBench, the first public, ML-ready dataset that pairs MTG/FCI imagery with colocated EarthCARE synergistic retrievals of cloud type and cloud microphysical properties. The dataset covers the full year of 2025 within $\pm 45^\circ$ of the MTG sub-satellite point and is released in three variants with colocation windows of 2, 3 and 4 minutes. Each sample in the dataset stores 128x128 pixels of FCI observations together with the sparse EarthCARE curtain, its overpass mask and per-pixel time offsets, so that supervision can be restricted to observed locations and the effect of temporal misalignment can be studied directly. Baseline experiments show that the vertically-resolved cloud structure can be partly recovered from geostationary radiances alone. Classification performance follows the ability of the passive imager to detect the corresponding cloud class: high ice clouds are recovered best, while low-liquid and mixed-phase clouds remain challenging, due to their weak contribution in the FCI observations. Cloud microphysical property retrieval is considerably more challenging, and the strong zero inflation of the targets calls for evaluation metrics that focus on cloudy voxels.

Several directions remain open to investigation. First, preliminary results with masked-autoencoder pretraining on unlabelled FCI imagery indicate that self-supervised learning can partly compensate for the sparse active-sensor supervision. We plan to extend this analysis with a systematic search over architectures, loss formulations and class weightings. Second, we plan to analyze the performance in more detail, including results per height level and per cloud type, the retrieval of clouds located beneath other cloud layers, and the uncertainty of the predicted profiles. Third, beyond evaluation with the EarthCARE products, the reconstructed 3D cloud fields will be validated externally with independent airborne observations from EarthCARE validation campaigns, such as PERCUSION campaign (Groß et al., 2026), part of the ORCESTRA campaign over the tropical Atlantic (Stevens et al., 2026). Moreover 3D radiative-transfer closure studies will be performed, comparing radiance and flux simulations using the derived 3D cloud scenes, with corresponding EarthCARE BBR products (Barker et al., 2011). Finally, the dataset will be extended with new data, as further years and new EarthCARE processing baselines become available.

## REFERENCES

H. W. Barker, M. P. Jerg, T. Wehr, S. Kato, D. P. Donovan, and R. J. Hogan. A 3d cloud-construction algorithm for the earthcare satellite mission. *Quarterly Journal of the Royal Meteorological Society, 137(657), 1042–1058*, 2011. doi: 10.1002/qj.824.

W. Bertrand, J. E. Kay, J. Haynes, and G. de Boer. 3s-geoprof-comb: A global gridded dataset for cloud vertical structure from combined cloudsat and calipso observations. 2023. doi: 10.5281/ZENODO.8057790.

O. Boucher, D. Randall, P. Artaxo, C. Bretherton, G. Feingold, P. Forster, V.-M. Kerminen, Y. Kondo, H. Liao, U. Lohmann, P. Rasch, S. K. Satheesh, S. Sherwood, B. Stevens, and X. Y. Zhang. Clouds and aerosols. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley (eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

S. Brüning, S. Niebler, and H. Tost. Artificial intelligence (ai)-derived 3d cloud tomography from geostationary 2d satellite data. *Atmospheric Measurement Techniques, 17(3), 961–978*, 2024. doi: 10.5194/amt-17-961-2024.

Y. Cao, W. Tan, and Z. Wu. Aircraft icing: An ongoing threat to aviation safety. *Aerospace Science and Technology, 75, 353–385*, 2018. doi: 10.1016/j.ast.2017.12.028.

H. E. Cline, C. L. Dumoulin, H. R. Hart, W. E. Lorensen, and S. Ludke. 3d reconstruction of the brain from magnetic resonance images using a connectivity algorithm. *Magnetic Resonance Imaging, 5(5), 345–352*, 1987. doi: 10.1016/0730-725X(87)90124-X.

P. Dandini, C. Cornet, R. Binet, L. Fenouil, V. Holodovsky, Y. Schechner, D. Ricard, and D. Rosenfeld. 3d cloud envelope and cloud development velocity from simulated cloud (c3iel) stereo images. *Atmospheric Measurement Techniques, 15(20), 6221–6242*, 2022. doi: 10.5194/amt-15-6221-2022.

D. P. Donovan, G.-J. van Zadelhoff, and P. Wang. The earthcare lidar cloud and aerosol profile processor (a-pro): The a-aer, a-ebd, a-tc, and a-ice products. *Atmospheric Measurement Techniques, 17, 5301–5340*, 2024. doi: 10.5194/amt-17-5301-2024.

P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu. Sen12ms-cr-ts: A remote-sensing data set for multimodal multitemporal cloud removal. *IEEE Transactions on Geoscience and Remote Sensing, 60, 1–14*, 2022. doi: 10.1109/TGRS.2022.3146246.

M. Eisinger, F. Marnas, K. Wallace, T. Kubota, N. Tomiyama, Y. Ohno, T. Tanaka, E. Tomita, T. Wehr, and D. Bernaerts. The earthcare mission: science data processing chain overview. *Atmospheric Measurement Techniques, 17, 839–862*, 2024. doi: 10.5194/amt-17-839-2024.

Shirin Ermis, Cesar Aybar, Lilli Freischem, Stella Girtsou, Kyriaki-Margarita Bintsi, Emiliano Diaz Salas-Porras, Michael Eisinger, William Jones, Anna Jungbluth, and Benoit Tremblay. Global 3d reconstruction of clouds tropical cyclones, 2026. URL https://arxiv.org/abs/2511.04773.

EUMETSAT. FCI L1 Operational Product Validation Report. Technical Report EUM2009502, EUMETSAT, December 2024. URL https://user.eumetsat.int/s3/eup-strapi-media/FCI_L1_Operational_Product_Validation_Report_eb9e6db47c.pdf. Version 1A.

European Space Agency. EarthCARE ACM-CAP Level 2B Processor Disclaimer, n.d.a. URL https://earthcarehandbook.earth.esa.int/documents/d/earthcare-data-handbook/earthcare-disclaimer-l2-acm-cap. Accessed: 24 September 2026.

European Space Agency. EarthCARE AC-TC Level 2B Processor Disclaimer, n.d.b. URL https://earthcarehandbook.earth.esa.int/documents/d/earthcare-data-handbook/earthcare-ac-tc-level-2b-processor-disclaimer. Accessed: 24 September 2026.

European Space Agency. EarthCARE A-PRO Level 2A Processor Disclaimer, n.d.c. URL https://earthcarehandbook.earth.esa.int/documents/d/earthcare-data-handbook/earthcare-product-disclaimer-atlid-level-2a-a-pro. Accessed: 24 September 2026.

European Space Agency. EarthCARE C-PRO Level 2A Processor Disclaimer, n.d.d. URL https://earthcarehandbook.earth.esa.int/documents/d/earthcare-data-handbook/earthcare-disclaimer-cpr-level-2a-c-fmr-c-cd-c-tc. Accessed: 24 September 2026.

S. Girtsou, E. Diaz Salas-Porras, L. Freischem, J. Massant, K.-M. Bintsi, G. Castiglione, W. Jones, M. Eisinger, E. Johnson, and A. Jungbluth. 3d cloud reconstruction through geospatially-aware masked autoencoders, 2025. URL https://arxiv.org/abs/2501.02035.

Silke Groß, Florian Ewald, Bjorn Stevens, Martin Wirth, Georgios Dekoutsidis, André Ehrlich, Dimitra Kouklaki, Kevin Krüger, Sophie Rosenburg, Lea Volkmer, Jörg von Bismarck, Lutz Hirsch, Anna E. Luebke, Eleni Marinou, Bernhard Mayer, Montserrat Pinol Sole, Manfred Wendisch, Julia Windmiller, Vassilis Amiridis, Rob Koopman, Takuji Kubota, and Markus Rapp. Persistent EarthCARE underflight studies of the ITCZ and organized convection (PERCUSION): contribution to EarthCARE validation. *Atmospheric Measurement Techniques*, 19:3933–3959, 2026. doi: 10.5194/amt-19-3933-2026.

K. Holmlund, J. Grandell, J. Schmetz, R. Stuhlmann, B. Bojkov, R. Munro, M. Lekouara, D. Coppens, B. Viticchie, T. August, B. Theodore, P. Watts, M. Dobber, G. Fowler, S. Bojinski, A. Schmid, K. Salonen, S. Tjemkes, D. Aminou, and P. Blythe. Meteosat third generation (mtg): Continuation and innovation of observations from geostationary orbit. *Bulletin of the American Meteorological Society, 102(5), E990–E1015*, 2021. doi: 10.1175/BAMS-D-19-0304.1.

A. J. Illingworth, H. W. Barker, A. Beljaars, M. Ceccaldi, H. Chepfer, N. Clerbaux, J. Cole, J. Delanoë, C. Domenech, D. P. Donovan, S. Fukuda, M. Hirakata, R. J. Hogan, A. Huenerbein, P. Kollias, T. Kubota, T. Nakajima, T. Y. Nakajima, T. Nishizawa, Y. Ohno, H. Okamoto, R. Oki, K. Sato, M. Satoh, M. W. Shephard, A. Velázquez-Blázquez, U. Wandinger, T. Wehr, and G.-J. van Zadelhoff. The earthcare satellite: The next step forward in global measurements of clouds, aerosols, precipitation, and radiation. *Bulletin of the American Meteorological Society, 96(8), 1311–1332*, 2015. doi: 10.1175/BAMS-D-12-00227.1.

A. Irbah, J. Delanoë, G.-J. van Zadelhoff, D. P. Donovan, P. Kollias, B. Puigdomènech Treserras, S. Mason, R. J. Hogan, and A. Tatarevic. The classification of atmospheric hydrometeors and aerosols from the earthcare radar and lidar: The a-tc, c-tc, and ac-tc products. *Atmospheric Measurement Techniques, 16, 2795–2820*, 2023. doi: 10.5194/amt-16-2795-2023.

K. Jeggle, M. Czerkawski, F. Serva, B. Le Saux, D. Neubauer, and U. Lohmann. Icecloudnet: 3d reconstruction of cloud ice from meteosat seviri. *Artificial Intelligence for the Earth Systems, 4(4), 240098*, 2025. doi: 10.1175/AIES-D-24-0098.1.

U. Khan, A. Yasin, M. Abid, I. Shafi, and S. A. Khan. A methodological review of 3d reconstruction techniques in tomographic imaging. *Journal of Medical Systems, 42(10), 190*, 2018. doi: 10.1007/s10916-018-1042-2.

M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks. Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites. *IEEE Transactions on Geoscience and Remote Sensing, 51(7), 3826–3852*, 2013. doi: 10.1109/TGRS.2012.2227333.

B. Lee, L. Di Girolamo, G. Zhao, and Y. Zhan. Three-dimensional cloud volume reconstruction from the multi-angle imaging spectroradiometer. *Remote Sensing, 10(11), 1858*, 2018. doi: 10.3390/rs10111858.

J. Leinonen, A. Guillaume, and T. Yuan. Reconstruction of cloud vertical structure with a generative adversarial network. *Geophysical Research Letters, 46(12), 7035–7044*, 2019. doi: 10.1029/2019GL082532.

H. Lin, J. Li, M. Min, F. Zhang, K. Wang, and Q. Wu. Clann: Cloud amount neural network for estimating 3d cloud from geostationary satellite imager. *Remote Sensing of Environment, 318, 114600*, 2025. doi: 10.1016/j.rse.2025.114600.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017. doi: 10.1109/ICCV.2017.324.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

H. Luo, J. Quaas, and Y. Han. Examining cloud vertical structure and radiative effects from satellite retrievals and evaluation of cmip6 scenarios. *Atmospheric Chemistry and Physics, 23(14), 8169–8186*, 2023. doi: 10.5194/acp-23-8169-2023.

S. D. Miller, J. M. Forsythe, P. T. Partain, J. M. Haynes, R. L. Bankert, M. Sengupta, C. Mitrescu, J. D. Hawkins, and T. H. Vonder Haar. Estimating three-dimensional cloud structure via statistically blended satellite observations. *Journal of Applied Meteorology and Climatology, 53(2), 437–455*, 2014. doi: 10.1175/JAMC-D-13-070.1.

Y.-J. Noh, J. M. Haynes, S. D. Miller, C. J. Seaman, A. K. Heidinger, J. Weinrich, M. S. Kulie, M. Niznik, and B. J. Daub. A framework for satellite-based 3d cloud data: An overview of the viirs cloud base height retrieval and user engagement for aviation applications. *Remote Sensing, 14(21), 5524*, 2022. doi: 10.3390/rs14215524.

K. Papachristopoulou, I. Fountoulakis, A. F. Bais, B. E. Psiloglou, N. Papadimitriou, I.-P. Raptis, A. Kazantzidis, C. Kontoes, M. Hatzaki, and S. Kazadzis. Effects of clouds and aerosols on downwelling surface solar irradiance nowcasting and short-term forecasting. *Atmospheric Measurement Techniques, 17(7), 1851–1877*, 2024. doi: 10.5194/amt-17-1851-2024.

Roi Ronen, Vadim Holodovsky, and Yoav Y. Schechner. Variable imaging projection cloud scattering tomography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8): 6097–6109, 2025. doi: 10.1109/TPAMI.2022.3195920.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *Machine Learning in Medical Imaging (MLMI)*, volume 10541 of *Lecture Notes in Computer Science*, pp. 379–387. Springer, 2017. doi: 10.1007/978-3-319-67389-9_44.

Yael Sde-Chen, Yoav Y. Schechner, Vadim Holodovsky, and Eshkol Eytan. 3DeepCT: Learning volumetric scattering tomography of clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5671–5682, 2021. doi: 10.1109/ICCV48922.2021.00562.

G. Seiz and R. Davies. Reconstruction of cloud geometry from multi-view satellite images. *Remote Sensing of Environment, 100(2), 143–149*, 2006. doi: 10.1016/j.rse.2005.09.016.

Graeme L. Stephens, Deborah G. Vane, Ronald J. Boain, Gerald G. Mace, Kenneth Sassen, Zhien Wang, Anthony J. Illingworth, Ewan J. O'Connor, William B. Rossow, Stephen L. Durden, Steven D. Miller, Richard T. Austin, Angela Benedetti, and Cristian Mitrescu. The CloudSat mission and the A-Train: A new dimension of space-based observations of clouds and precipitation. *Bulletin of the American Meteorological Society*, 83(12):1771–1790, 2002. doi: 10.1175/BAMS-83-12-1771.

Bjorn Stevens, Sandrine Bony, Silke Gross, Daniel Klocke, Julia M. Windmiller, Allison A. Wing, Jörg von Bismarck, E. Brito, R. O. David, Julien Delanoë, David Farrell, and Y. Wu. ORCESTRA: Organized convection and EarthCARE studies over the tropical Atlantic. *Tellus*, 79(1):35–52, 2026. doi: 10.16993/tellus.4123.

J. Sun, Y. Li, H. Hu, Q. Li, C. Ye, Y. Shi, and Z. Chen. Impact of cloud vertical structure perturbations on the retrieval of cloud optical thickness and effective radius from fy4a/agri. *Atmospheric Chemistry and Physics, 25(22), 16347–16361*, 2025. doi: 10.5194/acp-25-16347-2025.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

T. Wehr, T. Kubota, G. Tzeremes, K. Wallace, H. Nakatsuka, Y. Ohno, R. Koopman, S. Rusli, M. Kikuchi, M. Eisinger, T. Tanaka, M. Taga, P. Deghaye, E. Tomita, and D. Bernaerts. The earthcare mission: Science and system overview. *EGUsphere*, 2023. doi: 10.5194/egusphere-2022-1476.

Charles H. White, Yoo-Jeong Noh, John M. Haynes, and Imme Ebert-Uphoff. C3DIR: A deep learning 3-dimensional cloud property retrieval scheme for passive satellite imagers. *arXiv preprint arXiv:2607.16929*, 2026. doi: 10.48550/arXiv.2607.16929.

D. M. Winker, M. A. Vaughan, A. H. Omar, Y. Hu, K. Powell, Z. Liu, W. H. Hunt, and S. A. Young. Overview of the calipso mission and caliop data processing algorithms. *Journal of Atmospheric and Oceanic Technology, 26, 2310–2323*, 2009. URL https://api.semanticscholar.org/CorpusID:122020200.

C. Yang, F. Wang, X. Yang, H. Huang, W. Cao, and X. Chu. Sgmagnet: A baseline model for 3d cloud phase structure reconstruction on a new passive active satellite benchmark, 2025. URL https://arxiv.org/abs/2509.15706.

C. Yang, F. Wang, Q. Lu, R. Liu, X. Yang, J. Cheng, Y. Cao, and Z. Ni. Cnn-based retrieval of 3d cloud structures solely from geostationary satellite imagery. *Geophysical Research Letters, 53(5), e2025GL121014*, 2026. doi: 10.1029/2025GL121014.

X. Ying, H. Guo, K. Ma, J. Wu, Z. Weng, and Y. Zheng. X2CT-GAN: Reconstructing CT from biplanar x-rays with generative adversarial networks. *arXiv*, 2019. URL https://arxiv.org/abs/1905.06902.

H. Yu, J. Ma, S. Ahmad, E. Sun, C. Li, Z. Li, and J. Hong. Three-dimensional cloud structure reconstruction from the directional polarimetric camera. *Remote Sensing, 11(24), 2894*, 2019. doi: 10.3390/rs11242894.

V. Zantedeschi, F. Falasca, A. Douglas, R. Strange, M. J. Kusner, and D. Watson-Parris. Cumulo: A dataset for learning cloud classes, 2022. URL https://arxiv.org/abs/1911.04227.

M. D. Zelinka, T. A. Myers, D. T. McCoy, S. Po-Chedley, P. M. Caldwell, P. Ceppi, S. A. Klein, and K. E. Taylor. Causes of higher climate sensitivity in cmip6 models. *Geophysical Research Letters, 47(1), e2019GL085782*, 2020. doi: 10.1029/2019GL085782.

S. Zhang, S. Zhang, Y. Zhou, H. Liu, and M. Duan. Three-dimensional cloud radar reflectivity reconstruction from geostationary multispectral imagery using a context-aware transformer. *International Journal of Applied Earth Observation and Geoinformation, 149, 105275*, 2026. doi: 10.1016/j.jag.2026.105275.

## A MTG FCI SPECTRAL CHANNELS

This Appendix provides information on the spectral characteristics and native spatial sampling distances of the 14 MTG FCI channels that are included in the EarthCAREBench. Some channels are provided also in high-resolution, however the normal resolution observations are retained in EarthCAREBench.

Table 5: Spectral characteristics and native spatial sampling distances of the 14 MTG FCI channels retained in EarthCAREBench.

| **FCI channel** | **Central wavelength** $\lambda_0$ ($\mu$m) | **Spectral width** $\Delta\lambda_0$ ($\mu$m) | **SSD** (km) |
|---|---|---|---|
| VIS 0.4 (`vis_04`) | 0.444 | 0.060 | 1.0 |
| VIS 0.5 (`vis_05`) | 0.510 | 0.040 | 1.0 |
| VIS 0.6 (`vis_06`) | 0.640 | 0.050 | 1.0 / 0.5 (HR) |
| VIS 0.8 (`vis_08`) | 0.865 | 0.050 | 1.0 |
| VIS 0.9 (`vis_09`) | 0.914 | 0.020 | 1.0 |
| NIR 1.3 (`nir_13`) | 1.380 | 0.030 | 1.0 |
| NIR 1.6 (`nir_16`) | 1.610 | 0.050 | 1.0 |
| NIR 2.2 (`nir_22`) | 2.250 | 0.050 | 1.0 / 0.5 (HR) |
| IR 3.8 (`ir_38`) | 3.800 | 0.400 | 2.0 / 1.0 (HR) |
| IR 8.7 (`ir_87`) | 8.700 | 0.400 | 2.0 |
| IR 9.7 (`ir_97`) | 9.660 | 0.300 | 2.0 |
| IR 10.5 (`ir_105`) | 10.500 | 0.700 | 2.0 / 1.0 (HR) |
| IR 12.3 (`ir_123`) | 12.300 | 0.500 | 2.0 |
| IR 13.3 (`ir_133`) | 13.300 | 0.600 | 2.0 |

*Note:* SSD denotes spatial sampling distance. HR denotes the high-resolution sampling mode available for selected FCI channels. The listed distances refer to the native FCI products, before co-registration onto the common EarthCAREBench patch grid and the normal resolution is used.

## B EARTHCARE PRODUCTS

This Appendix provides information on the geographical distribution of the EarthCAREBench samples as well as on the two synergistic EarthCARE AC-TC and ACM-CAP L2b products that are included in EarthCAREBench. Figure 1 shows the geographical distribution of the colocated samples using $5^\circ \times 5^\circ$ latitude-longitude bins and Tables 6, 7 include the original variable names, their dimensions and a short description for each one of them.

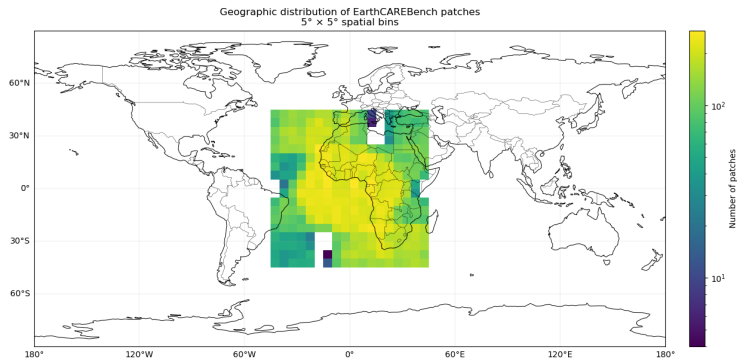



Figure 1: Geographical distribution of the EarthCAREBench samples for the 4-minute colocation dataset. Sample counts are aggregated into $5^\circ \times 5^\circ$ latitude-longitude bins.

Table 6: Variables retained from the EarthCARE AC-TC L2b product.

| Variable | Description | Dimensions |
|---|---|---|
| height | Height of pixel centres above the WGS84 ellipsoid (m) | along_track × JSG_height |
| time | UTC time at the JSG pixel | along_track |
| synergetic_target_classification | High-resolution ATLID-CPR target classification | along_track × JSG_height |
| synergetic_target_classification_medium_resolution | Medium-resolution ATLID-CPR target classification | along_track × JSG_height |
| synergetic_target_classification_low_resolution | Low-resolution ATLID-CPR target classification | along_track × JSG_height |
| quality_status | Confidence of the high-resolution classification based on the target and availability of observations from the two instruments | along_track × JSG_height |
| quality_status_medium_resolution | Confidence of the medium-resolution classification based on the target and availability of observations from the two instruments | along_track × JSG_height |
| quality_status_low_resolution | Confidence of the low-resolution classification based on the target and availability of observations from the two instruments | along_track × JSG_height |

Table 7: Variables retained from the ACM-CAP component of the EarthCARE ACM-COM L2b product.

| Variable | Description | Dimensions |
|---|---|---|
| **Coordinates and atmospheric state** | | |
| time | UTC time at the JSG pixel | along_track |
| height_level | Height of pixel centres above the WGS84 ellipsoid (m) | level_number × along_track |
| latitude_active | Latitude beneath the active sensors | along_track |
| longitude_active | Longitude beneath the active sensors | along_track |
| pressure_layer_mean | Mean atmospheric pressure within each layer | layer_number × along_track |
| pressure_level | Atmospheric pressure at each vertical level | level_number × along_track |
| specific_humidity_layer_mean | Layer-mean specific humidity | layer_number × along_track |
| temperature_layer_mean | Layer-mean atmospheric temperature | layer_number × along_track |
| temperature_level | Atmospheric temperature at each vertical level | level_number × along_track |
| **Aerosol properties** | | |
| aerosol_classification | Aerosol type classification | layer_number × along_track |
| aerosol_extinction | Aerosol extinction coefficient at $0.355\,\mu\mathrm{m}$ | layer_number × along_track |
| aerosol_extinction_error | Relative $1\sigma$ uncertainty of the aerosol extinction coefficient at $0.355\,\mu\mathrm{m}$ | layer_number × along_track |

*Continued on next page*

*Table 7 continued from previous page*

| Variable | Description | Dimensions |
|---|---|---|
| aerosol_extinction_quality_status | Quality status of the aerosol-extinction retrieval | along_track |
| **Cloud and precipitation properties** | | |
| cloud_flag | Flag indicating the presence or absence of cloud | layer_number × along_track |
| cloud_retrieval_quality_status | Quality status of the cloud retrieval | layer_number × along_track |
| rain_water_content | Rain water content | layer_number × along_track |
| rain_water_content_quality_status | Quality status of the rain-water-content retrieval | layer_number × along_track |
| snow_water_content | Snow water content | layer_number × along_track |
| snow_water_content_quality_status | Quality status of the snow-water-content retrieval | layer_number × along_track |
| **Cloud microphysical properties** | | |
| liquid_water_content | Liquid water content | layer_number × along_track |
| liquid_effective_radius | Liquid effective radius | layer_number × along_track |
| liquid_water_content_error | $1\sigma$ random uncertainty in the natural logarithm of liquid water content | layer_number × along_track |
| liquid_effective_radius_error | $1\sigma$ random uncertainty in the natural logarithm of liquid effective radius | layer_number × along_track |
| ice_water_content | Ice water content | layer_number × along_track |
| ice_effective_radius | Ice effective radius | layer_number × along_track |
| ice_water_content_error | $1\sigma$ random uncertainty in the natural logarithm of ice water content | layer_number × along_track |
| ice_effective_radius_error | $1\sigma$ random uncertainty in the natural logarithm of ice effective radius | layer_number × along_track |
| **Quality information** | | |
| atmospheric_profile_layer_mean_quality_status | Quality status of the layer-mean atmospheric profile | along_track |
| atmospheric_profile_level_quality_status | Quality status of the atmospheric profile at model levels | along_track |
| quality_status | Generic product quality status | along_track |

# C Statistical Analysis Appendix

## C.1 Temporal and spatial colocation statistics

Figure 2 illustrates why the colocation window matters: with large time differences, the EarthCARE profile may no longer describe the clouds seen in the MTG image.

Figure 3 presents the distributions of the absolute acquisition-time differences and the number of colocated EarthCARE profiles per MTG patch.

## C.2 AC-TC product Class distribution

Here, in Fig. 4 we present the percentage of pixels classified in one of the native AC-TC product classes inside the 4-min variation of the EarthCAREBench dataset using both the standard- and the low-resolution variables.

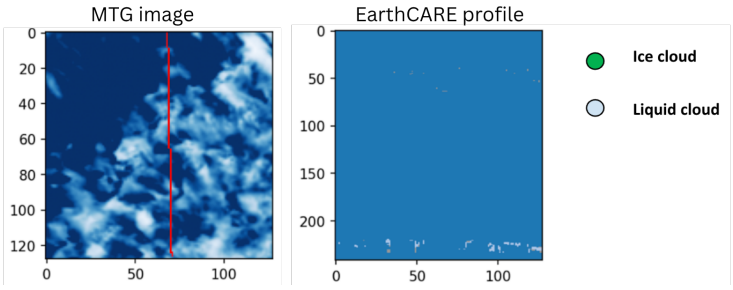
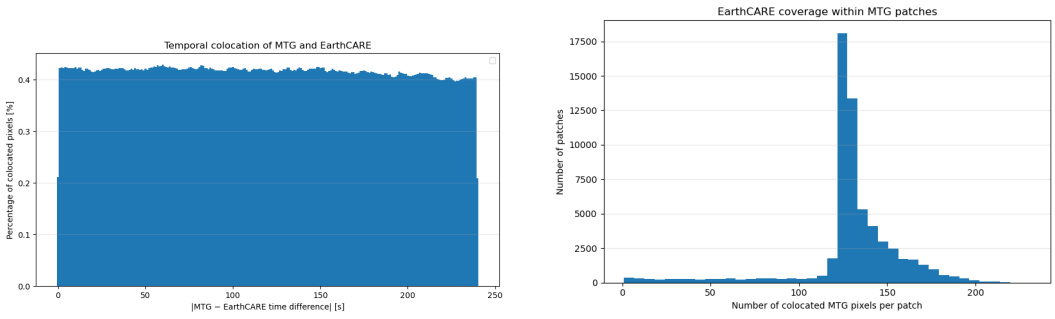



Figure 2: Example of a colocated scene with an 8-minute window, larger than those used in EarthCAREBench. Because of the time difference between the two observations, the clouds visible in the FCI image are not captured by the EarthCARE AC-TC profile. Such mismatched pairs artificially increase the training loss and push the model towards incorrect mappings between FCI images and EarthCARE vertical cloud profiles.




(a) Temporal colocation

(b) Colocated EarthCARE profiles

Figure 3: Temporal and spatial colocation characteristics of the 4-minute EarthCAREBench dataset. (a) Distribution of absolute MTG-EarthCARE acquisition-time differences. (b) Distribution of the number of colocated EarthCARE profiles, $N_{\mathrm{pix}}$, per MTG patch.

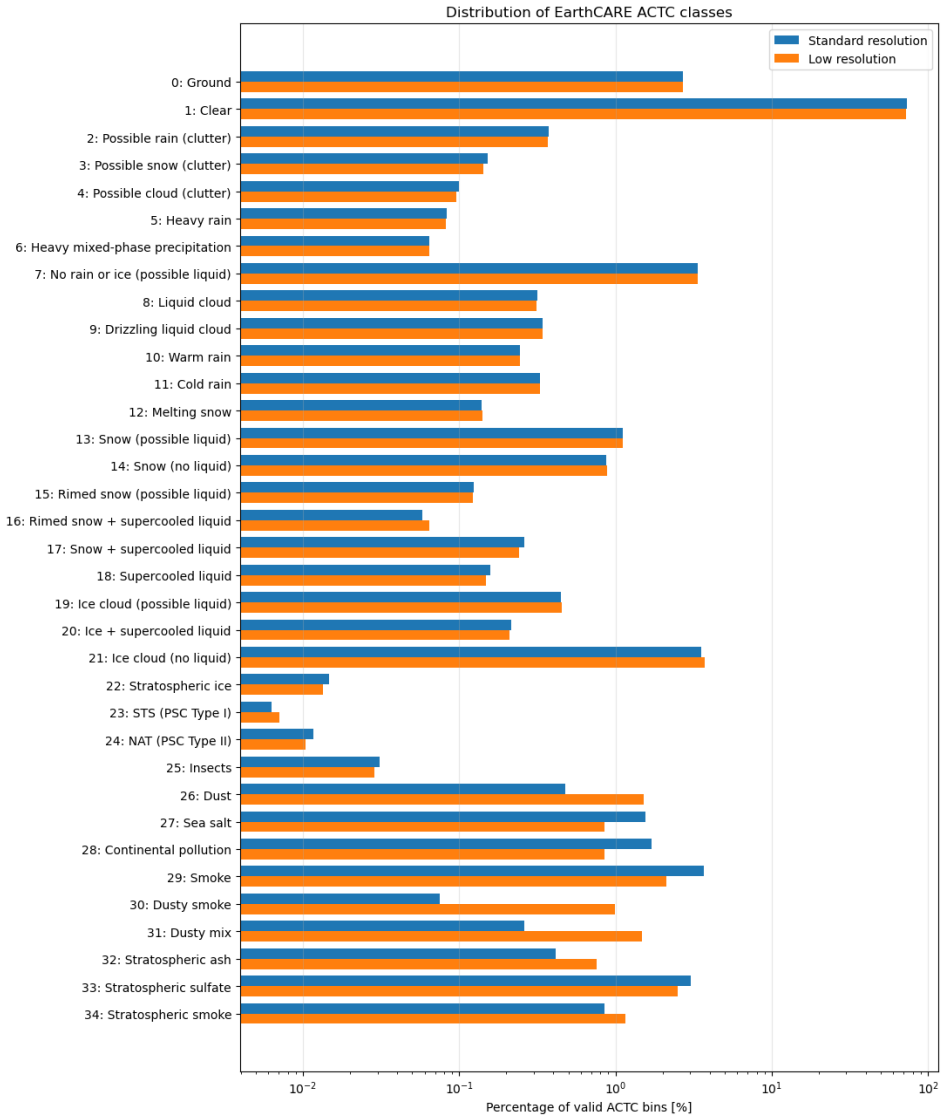



Figure 4: Distribution of the native EarthCARE AC-TC product classes in the 4-minute EarthCAREBench dataset. The relative frequencies of valid atmospheric bins are shown for the standard- and low-resolution classifications on a logarithmic scale.

Among the hydrometeor categories, liquid- and ice-cloud classes account for substantial fractions of the valid atmospheric bins, whereas several mixed-phase and precipitation-related classes are considerably less common. Some of the rarest categories, including polar stratospheric cloud classes, contribute only a small fraction of the complete dataset.

The standard- and low-resolution classifications exhibit broadly similar distributions at the dataset scale. Nevertheless, differences are visible for individual classes, particularly for some of the aerosol and mixed-phase categories. These differences are associated with the different horizontal integration scales used by the AC-TC product classifications, which affect the representation of weak atmospheric features.

### C.3 ACM-CAP PRODUCT STATISTICAL ANALYSIS

Here we present in detail the statistical analysis of the variables concerning the microphysical properties of clouds and precipitation, along with the aerosol extinction coefficient, provided by the ACM-CAP product.

Table 8: Statistical characteristics of the retained ACM-CAP variables in the 4-minute Earth-CAREBench dataset. Zero fractions are calculated relative to valid retrievals, whereas invalid fractions are calculated relative to all atmospheric bins. The mean, standard deviation, minimum, and maximum are calculated using only valid non-zero retrievals. The effective radius is expressed in m, the water content in $\mathrm{kg\,m^{-3}}$, and the aerosol extinction coefficient ("aerosol extinction") in $\mathrm{m^{-1}}$.

| Variable | Zero (%) | Invalid (%) | Non-zero mean | Non-zero std | Non-zero min | Non-zero max |
|---|---|---|---|---|---|---|
| Ice effective radius | 92.94 | 4.56 | 0.000044 | 0.000030 | 7.641025e-08 | 0.005861 |
| Ice water content | 92.94 | 4.56 | 0.000087 | 0.000550 | 1.172509e-13 | 1.274242 |
| Liquid effective radius | 97.93 | 4.56 | 0.000007 | 0.000005 | 6.653435e-08 | 0.000103 |
| Liquid water content | 97.93 | 4.56 | 0.000117 | 0.000252 | 3.594217e-10 | 0.005000 |
| Rain water content | 98.86 | 4.56 | 0.000060 | 0.000130 | 1.492907e-09 | 0.039102 |
| Aerosol extinction | 85.05 | 18.56 | 0.000041 | 0.000101 | 1.079923e-09 | 0.002792 |

An important distinction is made between invalid, valid zero, and valid non-zero values. An invalid value indicates that no usable retrieval is available, whereas a valid zero is retained as a physically-meaningful absence of the corresponding atmospheric constituent. The zero-percentage values reported in Table 8 are therefore calculated relative to the valid values rather than to all atmospheric bins.

For the four cloud microphysical properties used in the baseline benchmark, approximately $4.56\%$ of the values are invalid. Among the remaining valid retrievals, $92.94\%$ of both ice effective radius and ice water content are zero. Thus, only approximately $7.06\%$ of valid ice-property retrievals are non-zero.

The liquid-cloud variables are even more sparse: $97.93\%$ of valid liquid effective-radius and liquid-water-content values are zero, leaving only approximately $2.07\%$ of valid values as non-zero retrievals.

Expressed relative to all atmospheric bins, including invalid retrievals, the corresponding non-zero fractions are approximately $6.74\%$ for each ice variable and $1.97\%$ for each liquid variable. This extreme zero inflation represents an important property of the regression benchmark: a global error metric can be dominated by the much more frequent zero-valued atmospheric bins unless target-positive retrievals are evaluated separately.

Rain water content is even more sparse, with approximately $98.86\%$ of valid values equal to zero. In contrast, the aerosol extinction coefficient contains a larger non-zero fraction: approximately $14.95\%$ of valid retrievals are non-zero, although its missing-data fraction is also considerably higher, at approximately $18.56\%$.

Figure 5 shows the percentage of zero values per height along with the mean non-zero values for each variable. Liquid-cloud and rain properties are concentrated in lower levels with their largest

mean non-zero values occuring below 5 km. In contrast, non-zero ice retrievals can be found in middle and upper troposhere.

The aerosol extinction coefficient is quite variant throughout the vertical column with higher values generally occuring below 10 km. Overall, the vertical distributions further highlight the strong sparsity of the microphysical retrievals and their dependence on atmospheric height.

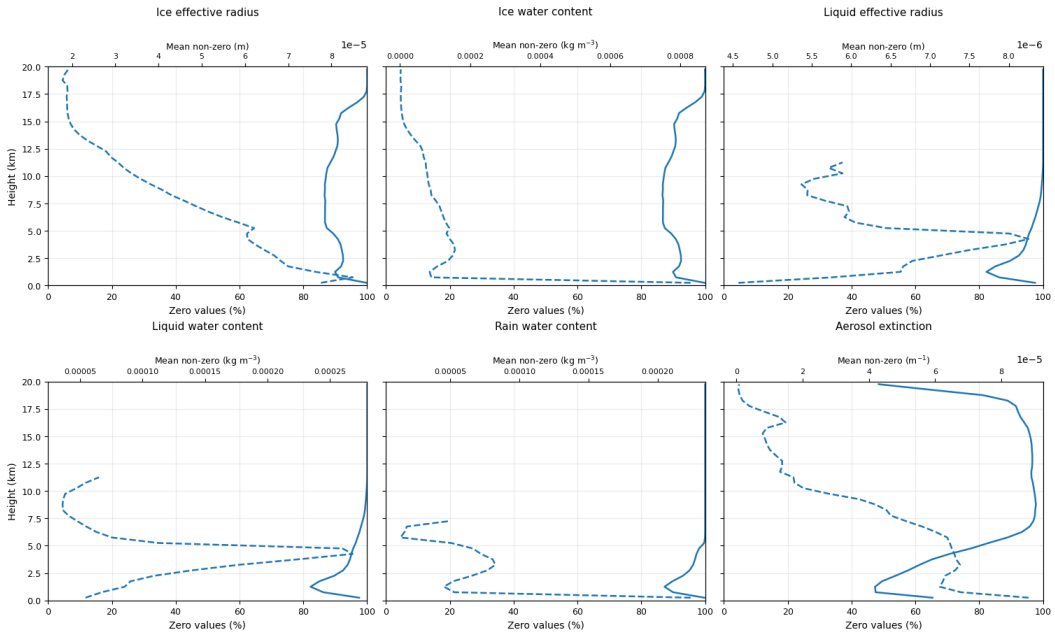



Figure 5: Vertical distribution of retrieval sparsity and non-zero magnitude for six ACM-CAP product variables in the 4-minute EarthCAREBench dataset. Each panel shows the percentage of valid retrievals equal to zero (solid line) and the mean of valid non-zero retrievals (dashed line) as a function of height. The mean non-zero values are expressed in the corresponding physical units: m for effective radius, $\mathrm{kg\,m^{-3}}$ for water content, and $\mathrm{m^{-1}}$ for the aerosol extinction coefficient.

## D AC-TC PRODUCT CLASSIFICATION MAPPING

Here we present the mapping of the original AC-TC classes to eight grouped categories for the EarthCAREBench segmentation benchmark. Table 9 presents the adopted mapping which is relevant to the task of cloud retrievals and also addresses the imbalance of the classes.

## E SUPPLEMENTARY BENCHMARK METRICS

In Table 10 we present egression results on the held-out 2-minute test set. Detection F1 is computed for voxels exceeding t = 0.1 in normalized units; cloudy RMSE is computed on voxels with non-zero targets, in normalized units. t denotes the colocation window of the training data

## F RETRIEVAL EXAMPLES

We present example results of the retrievals in the curtain where the target values are available.

We observe that the low metrics are not associated with a complete failure of the model to detect the atmospheric structure, but more with the difficulty in distinguishing the exact boundaries and vertical structure and height with fine detail, especially in neighboring cloud categories, and more systematically for the minority liquid and mixed phase cloud classes that are found usually embedded inside other, more dominant classes.

Table 9: Mapping from the original EarthCARE AC-TC product classes to the EarthCAREBench atmospheric classes.

| Class ID | EarthCAREBench class | Original AC-TC class IDs |
|---|---|---|
| $-1$ | `unknown` | $-1$ |
| 0 | `clear_sky` | 1 |
| 1 | `liquid_cloud` | 8 |
| 2 | `supercooled_liquid` | 18 |
| 3 | `ice_and_supercooled_liquid` | 20 |
| 4 | `ice_cloud` | 21 |
| 5 | `tropospheric_aerosols` | 26, 27, 28, 29, 30, 31 |
| 6 | `stratospheric_aerosols` | 32, 33, 34 |
| 7 | `other` | 0, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 22, 23, 24, 25 |

Table 10: Regression results on the held-out 2-minute test set. Detection F1 is computed for voxels exceeding $\tau = 0.1$ in normalized units; cloudy RMSE is computed on voxels with non-zero targets, in normalized units. $\Delta t$ denotes the colocation window of the training data.

| | | Detection F1 ↑ | | | | | Cloudy RMSE ↓ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Train** $\Delta t$ | $r_{\mathrm{e,ice}}$ | IWC | $r_{\mathrm{e,liq}}$ | LWC | Mean | $r_{\mathrm{e,ice}}$ | IWC | $r_{\mathrm{e,liq}}$ | LWC | Mean |
| ResUNet | 2 min | 0.387 | 0.354 | 0.297 | 0.281 | 0.330 | 0.289 | 0.197 | 0.323 | 0.366 | 0.294 |
| | 3 min | 0.414 | 0.439 | 0.291 | 0.283 | 0.357 | 0.265 | 0.183 | 0.314 | 0.333 | 0.274 |
| | 4 min | 0.412 | 0.440 | 0.293 | 0.285 | 0.357 | 0.266 | 0.185 | 0.313 | 0.333 | 0.274 |

The SatMAE-based model demonstrates improved retrieval performance, capturing both the magnitude and the fine-scale vertical structure of cloud microphysical properties more accurately. It also better preserves cloud boundaries and resolves finer vertical details, resulting in a more faithful representation of the observed cloud structure.

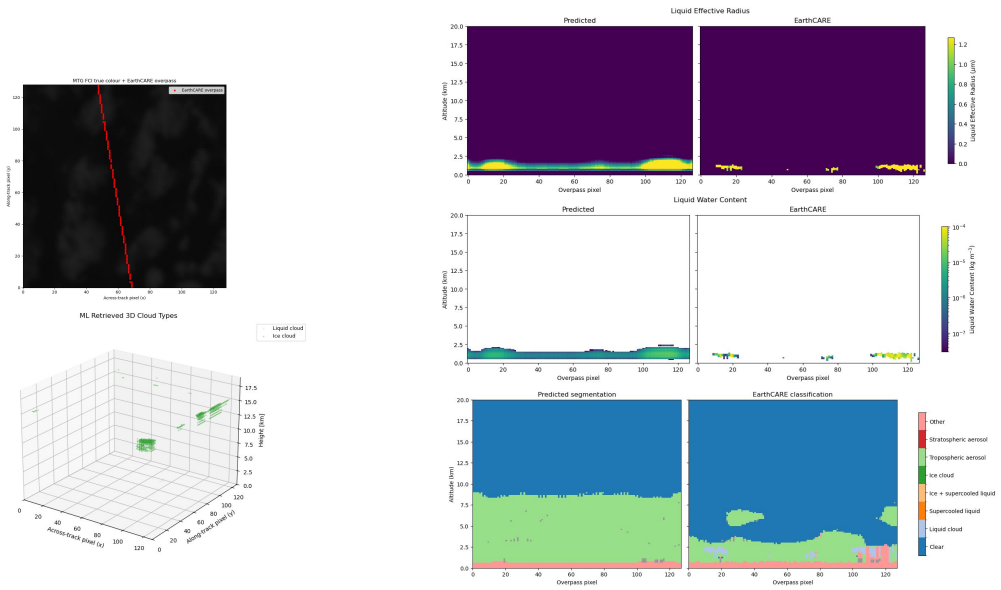



Figure 6: Night-time retrieval example of the ResUNet trained on the 2-minute variant. Top left: FCI image with the EarthCARE ground track (red). Bottom left: retrieved 3D cloud types over the full patch. Right: predicted (left column) and EarthCARE (right column) liquid effective radius, liquid water content and atmospheric target classification along the track.

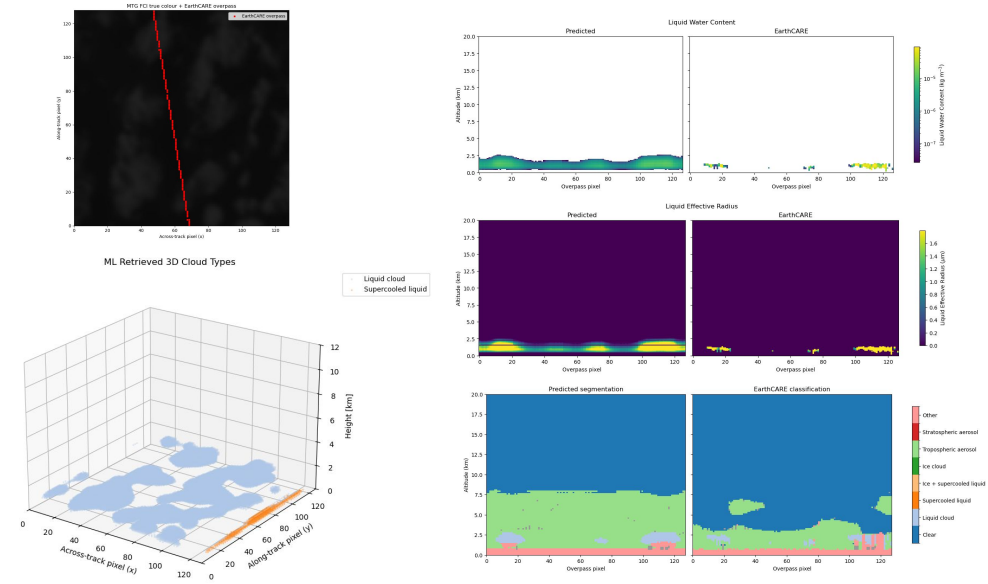



Figure 7: As fig. 6, for the ResUNet trained on the 3-minute variant.

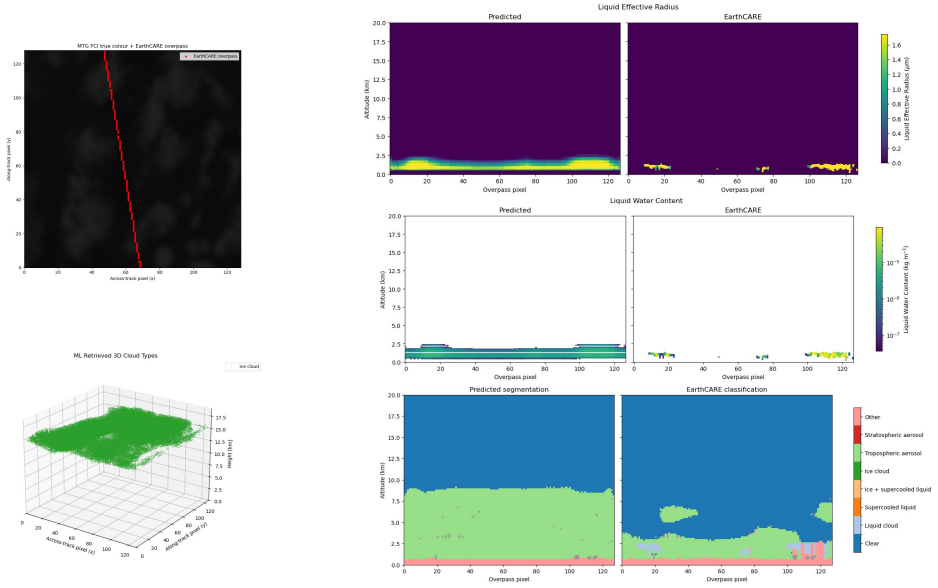



Figure 8: As fig. 6, for the ResUNet trained on the 4-minute variant.

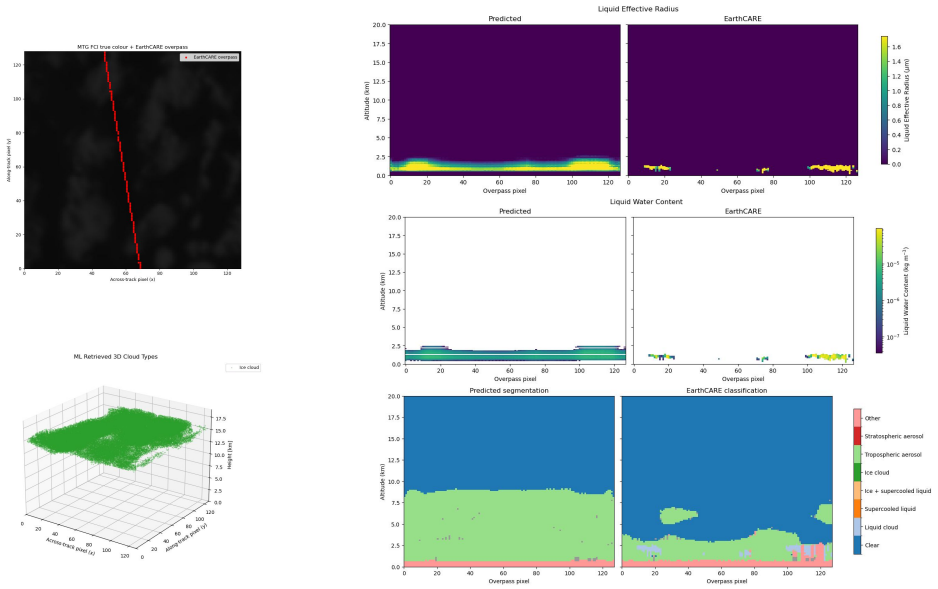



Figure 9: As fig. 6, for the ViT + pretrained SatMAE trained on the 3-minute variant.

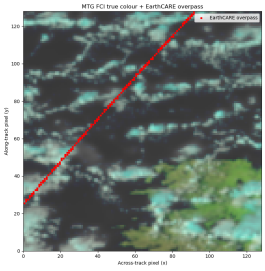



Figure 10: Daytime cloud retrieval example: FCI/MTG patch input.

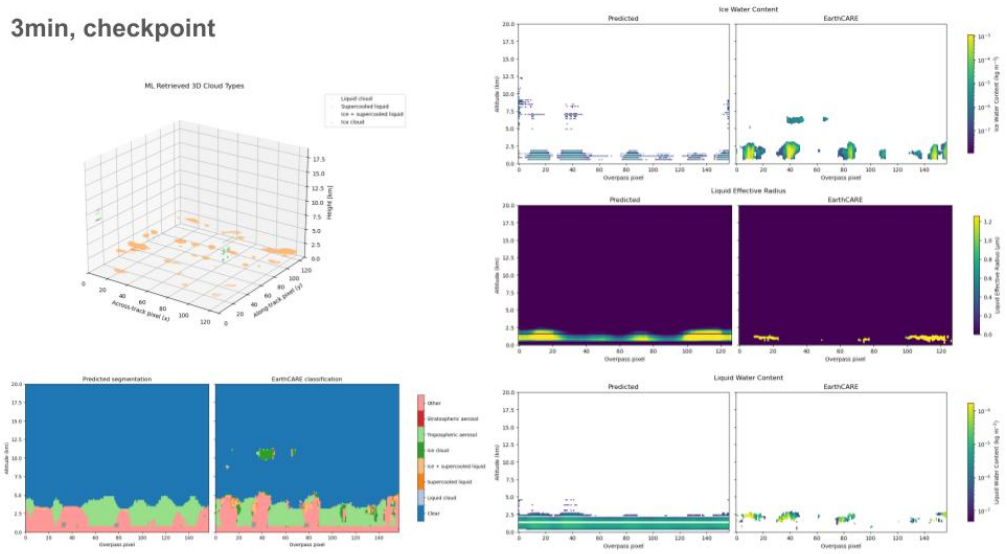



Figure 11: Daytime retrieval example of the ResUNet trained on the 2-minute variant. Top left: retrieved 3D cloud types over the full patch. Bottom left: predicted (left column) and EarthCARE (right column) target classification along the track. Right: predicted (left column) and EarthCARE (right column) ice water content(top), liquid effective radius (middle), liquid water content (down) along the track.

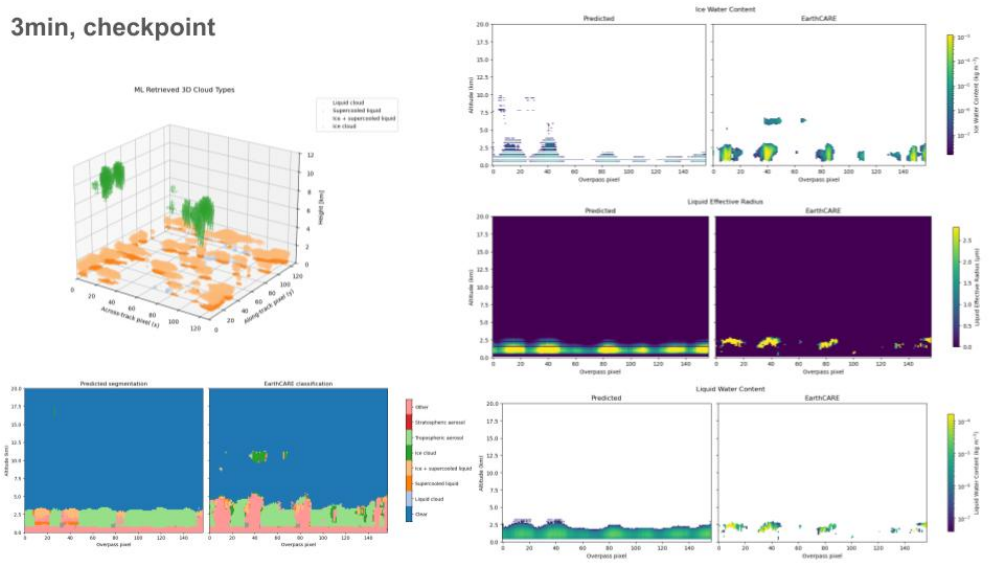



Figure 12: Daytime retrieval example of the ResUNet trained on the 3-minute variant. Top left: retrieved 3D cloud types over the full patch. Bottom left: predicted (left column) and EarthCARE (right column) target classification along the track. Right: predicted (left column) and EarthCARE (right column) ice water content(top), liquid effective radius (middle), liquid water content (down) along the track.

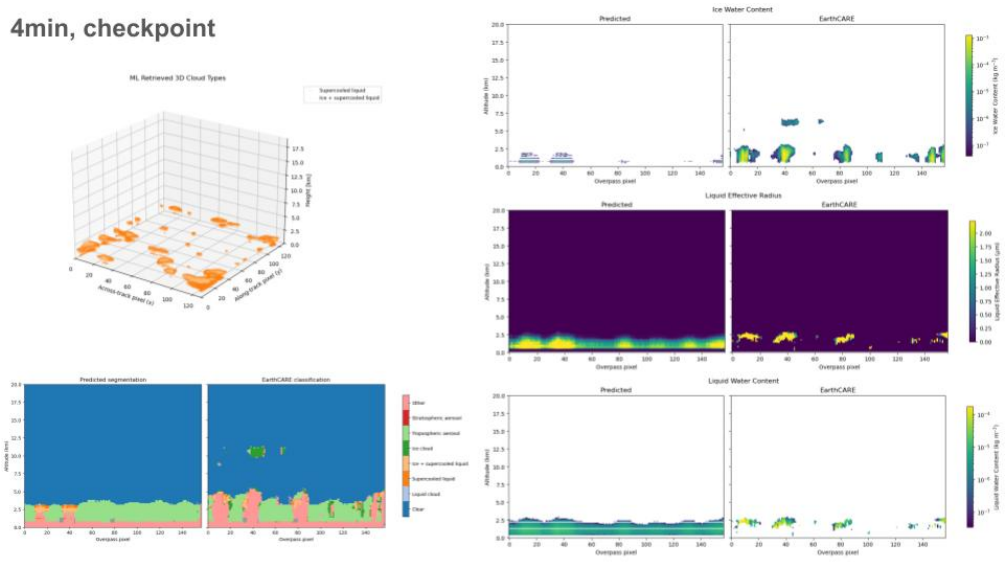



Figure 13: Daytime retrieval example of the ResUNet trained on the 4-minute variant. Top left: retrieved 3D cloud types over the full patch. Bottom left: predicted (left column) and EarthCARE (right column) target classification along the track. Right: predicted (left column) and EarthCARE (right column) ice water content(top), liquid effective radius (middle), liquid water content (down) along the track.

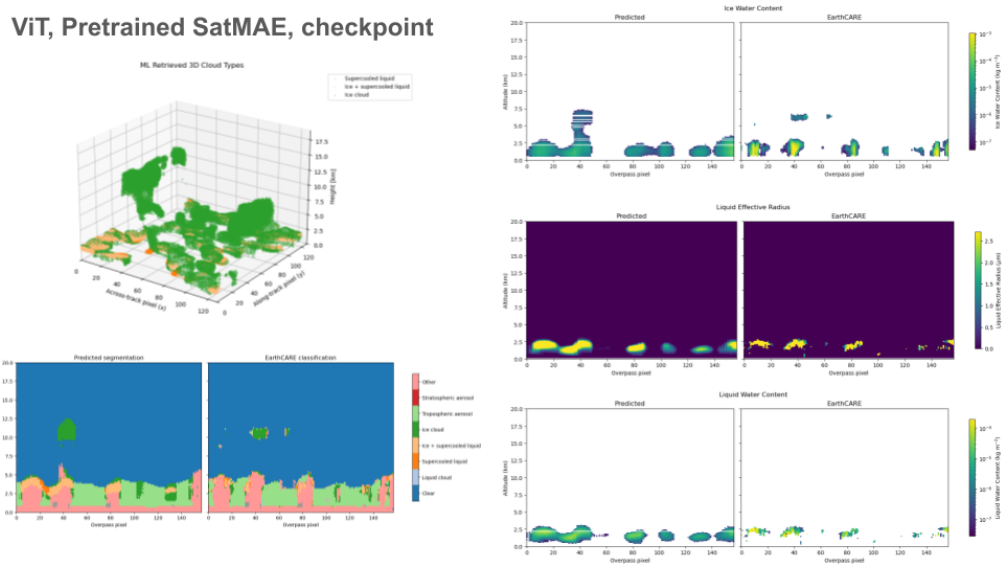



Figure 14: Daytime retrieval example of the ViT + pretrained SatMAE trained on the 3-minute variant. Top left: retrieved 3D cloud types over the full patch. Bottom left: predicted (left column) and EarthCARE (right column) target classification along the track. Right: predicted (left column) and EarthCARE (right column) ice water content(top), liquid effective radius (middle), liquid water content (down) along the track.