https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


# Assessment of tomography-derived three-dimensional cloud information and its suitability for downstream applications

Alessia Boccalatte[1,2], Yves-Marie Saint-Drenan[1], Benoit Gschwind[1], and Gabriel Chesnoiu[1]

[1]Centre Observation, Impacts, Energie (O.I.E), MINES Paris - PSL University, Sophia Antipolis, 06904, France
[2]LOCIE EMR 5002, Université Savoie Mont Blanc, CNRS, 60 Avenue du Lac Léman, Le Bourget-du-Lac, 73370, France

**Correspondence:** Alessia Boccalatte (alessia.boccalatte@univ-smb.fr)

**Abstract.** Meteorological geostationary satellites carry passive imagers that provide multispectral observations over a large fraction of the Earth disk every 5-15 min. Cloud properties retrieved from these observations are primarily two-dimensional, as they are inferred from top-of-atmosphere radiances, and contain only limited information on vertical cloud structure. Recent SEVIRI-CloudSat tomography studies have shown that it is possible to infer CloudSat-like three-dimensional cloud structure from multispectral passive imagery. The main open question is therefore no longer feasibility of such approach, but the reliability of the reconstructed cloud information for downstream applications. The present study addresses this gap through a systematic performance evaluation of a SEVIRI-CloudSat tomography method on an independent year-long (2019) dataset of 17,110 scenes. The analysis assesses the fidelity of the estimated reflectivity, but also evaluates cloud occurrence, cloud geometry, vertically aggregated summaries derived from the 3D information, and their dependence on different variables. Scene-wise reflectivity RMSE over all CloudSat-valid voxels, including cloudy and clear-sky voxels is consistent with previous studies (2.76 $\mathrm{dBZ}$). However, evaluation restricted to radar-detectable cloudy voxels yields a much larger median RMSE (14.48 $\mathrm{dBZ}$) and bias (-11.16 $\mathrm{dBZ}$), indicating systematic attenuation of moderate and strong echoes. Cloudy-voxel fraction is only weakly biased (-2.0%), whereas cloudy-column fraction is underestimated by -17.4%, consistent with missed observed-cloudy columns (POD $= 0.63$) and few false detections (FAR $= 0.031$). Cloud top and base heights remain informative when evaluated only where cloud geometry is defined in both observation and prediction (about 60% of observed-cloudy columns) with median errors of 1.38 and 1.41 $\mathrm{km}$. The scene-level vertical cloud centroid provides a complementary aggregate summary, with 96% cloudy-scene recall, $r = 0.89$, and an RMSE of 1.04 $\mathrm{km}$, while 97.4% of observed multilayer columns collapse to single-layer reconstructions. The results show that SEVIRI-CloudSat tomography is limited in resolving fine-scale vertical structure and multilayer organization, but remains informative for bulk descriptors such as cloud fraction, vertical cloud centroid, and cloud top and base heights.

## 1 Introduction

Clouds strongly affect Earth's radiative energy balance and atmospheric circulation and remain a leading source of uncertainty for climate projections and modeling (Lamer et al., 2020; IPCC, 2023). Applications ranging from climate research and numerical weather prediction to environmental monitoring and solar radiation assessment depend on satellite-derived cloud

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


information. Current operational geostationary products mainly use two-dimensional (2D), top-of-atmosphere (TOA) imager quantities such as cloud-top pressure or height, cloud fraction, and spectral or broadband optical properties, sometimes supplemented by column-integrated water or ice content products (Sayer et al., 2025; Mcgarragh et al., 2018). These products are used both as stand-alone diagnostics and as inputs to downstream schemes for TOA and surface flux estimation, long-term cloud climatologies, and process-oriented model evaluation of cloud representation and radiative effects in atmospheric models (Wiltink et al., 2025a; Karlsson and Devasthale, 2018).

In many cases, however, uncertainties are dominated by cloud vertical structure. This applies for example to multi-layer systems, deep convection, and mixed-phase clouds, where layer positioning and the vertical extent of optically thick regions strongly affect radiative fluxes, latent-heating profiles, and convective organization (Yu et al., 2022; Massie et al., 2021). Cloud vertical position is also relevant for the geometric correction of geostationary observations, since parallax displacement depends on cloud height and can be difficult to represent with a single height in vertically extended or multilayer scenes (Wiltink et al., 2025b). For these applications, knowing where the cloud is located in the atmospheric column, how vertically extended it is, and whether distinct layers coexist, can be as important as knowing that cloud is present at all. Yet different vertical cloud configurations can produce similar TOA radiances, so the problem is intrinsically ill-posed with respect to vertical cloud structure. In this context, cloud tomography, i.e. the reconstruction of three-dimensional (3D) cloud fields from multispectral satellite imagery and vertically resolved cloud profiles from active sensors on polar-orbiting satellites, offers a way to enrich geostationary observations with information on cloud vertical organization.

The observing systems that provide vertical cloud information have well-known limitations. Spaceborne active sensors such as the CloudSat Cloud Profiling Radar (CPR) and similar radars or lidars deliver vertically resolved profiles, but only along narrow nadir swaths and at fixed local solar times, with long revisit intervals. Additionally, their finite sensitivity and near-surface blind zone imply that a substantial fraction of boundary-layer clouds, light precipitation and shallow mixed-phase layers is missed (Schulte et al., 2023; Schirmacher et al., 2023). By contrast, geostationary imagers such as the Spinning Enhanced Visible and InfraRed Imager (SEVIRI) provide frequent multispectral two-dimensional observations over a full-disk domain, but they do not directly resolve the vertical distribution of cloud within the atmospheric column. Evaluations against active-sensor references show that passive geostationary retrieval errors depend strongly on cloud scene characteristics. For example, SEVIRI cloud-top height retrievals generally degrade for optically thin and multilayer clouds, while deep-convective cloud detection from geostationary imagery also shows frequent misclassifications in complex cloud scenes (Kotarba and Wojciechowska, 2025).

The complementary information provided by active and passive satellite observations has recently motivated statistical and deep-learning methods that fuse both data types within a tomographic framework (Leinonen et al., 2019; Girtsou et al., 2025; Jeggle et al., 2024; Brüning et al., 2024). Collocated pairs of geostationary radiances and vertical profiles are used to train a model that maps multispectral 2D data to vertically resolved cloud variables or 3D cloud fields (reflectivity as measured by the radar). Most current implementations use CloudSat measurements or CloudSat-CALIPSO fusion product (DARDAR) as the vertical reference and TOA radiances from instruments such as SEVIRI, MODIS or AGRI. Depending on the study, the target may be radar reflectivity, cloud occurrence or cloud mask, cloud amount, or microphysical quantities such as ice water content

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


(IWC). The present work focuses on the all-cloud reconstruction of CloudSat-like 94 GHz radar reflectivity from SEVIRI observations.

Radar reflectivity is a useful intermediate target for all-cloud 3D cloud reconstruction. At millimetre wavelengths (94 GHz for CloudSat), reflectivity is sensitive to the hydrometeor size distribution and concentration, with a strong weighting toward larger particles. In ice-cloud regions above the freezing level, it can be related to ice water content (IWC) through microphysical assumptions and empirical parameterizations. Operational products, such as CloudSat-CALIPSO 2C-ICE, therefore combine reflectivity with temperature and lidar information to infer IWC vertical profiles and particle size (Deng et al., 2015; Aubry et al., 2024), while joint radar-lidar classification schemes use reflectivity and backscatter to determine phase and cloud type (Ceccaldi et al., 2013).

For the purposes of the present study, the main interest of reflectivity is that it allows a hierarchy of cloud-structure descriptors to be derived from a single reconstructed 3D field. These include threshold-based cloud occurrence and vertical occurrence profiles, cloud-top and cloud-base height, cloud thickness, multilayer structure, and vertically aggregated summaries of cloud placement. At geostationary sampling, such descriptors could support applications including radiative interpretation, monitoring of rapidly evolving cloud systems, process-oriented model evaluation, and studies linked to the surface solar resource (Qu et al., 2017; Schroedter-Homscheidt et al., 2022). These applications, however, do not all require the same level of structural detail. Some depend mainly on cloud occurrence, cloud-top height, or the broad vertical placement of cloud layers, whereas others require more reliable estimates of cloud-base height, vertical extent, and the preservation of multilayer structure. In addition, the usefulness of a reconstructed cloud field depends not only on the descriptors themselves, but also on how they are validated.

In the present context, the relevant question is therefore not whether a SEVIRI-to-CloudSat mapping can be learned, but which variables derived from the reconstructed reflectivity field remain reliable enough to support downstream interpretation. Existing work addresses this issue only partially: most studies have focused on demonstrating reconstruction feasibility, evaluating the primary target variable, or reporting global reconstruction scores, rather than on systematically assessing how reconstruction errors propagate into derived descriptors relevant for downstream interpretation.

### 1.1 Related studies

Several recent studies have used machine learning to infer vertical cloud structure from passive satellite observations. These studies differ not only in the passive sensor used, but also in the target variable, the spatial support of the prediction, and the validation strategy. They can be broadly grouped into three families: curtain- or profile-based reconstructions, domain-wide products targeting cloud variables other than all-cloud reflectivity, and domain-wide reflectivity reconstructions from geostationary imagery.

A first group of studies demonstrates that passive observations contain information on vertical cloud structure, but remains restricted to along-track or curtain-based products. Leinonen et al. (2019) showed with a conditional GAN that collocated MODIS cloud-property inputs can be used to generate CloudSat-like two-dimensional radar scenes. This demonstrated the feasibility of passive-to-radar reconstruction, but did not produce domain-wide three-dimensional fields. Spradlin et al. (2024)

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


addressed a related problem from a foundation-model perspective: SatVision-TOA was pre-trained on all-sky MODIS TOA imagery and then fine-tuned for a downstream curtain-based 3D cloud retrieval task, in which vertical cloud masks are predicted along CloudSat-CALIPSO overpasses. This work therefore evaluates transfer to a profile-based cloud-mask retrieval task rather than full-volume reflectivity reconstruction.

A second group produces domain-wide vertical cloud information, but targets variables different from all-cloud radar reflectivity. Lin et al. (2025) developed CLANN (Cloud Amount Neural Network), which combines FY-4A AGRI radiances, CALIPSO profiles and ERA5 meteorological predictors (relative humidity and temperature at upper and lower layers, surface pressure and skin temperature) to retrieve 3D cloud amount on a vertical grid. Their results show that the neural network reproduces the main features of CALIPSO-derived cloud amount and cloud-amount-weighted height with denser sampling, but the target is cloud amount rather than radar reflectivity. IceCloudNet (Jeggle et al., 2024) uses SEVIRI infrared channels and auxiliary meta-data (latitude, longitude, month, binary day/night mask, and land/water mask) to predict 3D ice water content and ice crystal number concentration from DARDAR over a tropical domain. Its evaluation is broader and more application-oriented, spanning voxel-scale, overpass-scale, and climatological diagnostics, but its target remains restricted to ice-cloud microphysics and cannot be interpreted as an all-cloud reflectivity benchmark.

The most direct line of work for the present study is therefore the reconstruction of CloudSat-like reflectivity from geostationary imagery. Brüning et al. (2024) introduced a Res-UNet that maps eleven SEVIRI channels to CloudSat 94 GHz reflectivity on a full Meteosat disk grid, targeting all radar-detectable clouds. Their evaluation relies mainly on global and height-dependent RMSE and on derived cloud-top height, compared with the CLAAS operational product (Benas et al., 2017), with validation restricted to a single test month (May 2016). More recently, Girtsou et al. (2025) proposed geospatially aware masked autoencoders for MSG/SEVIRI, pre-training transformers on large volumes of unlabeled imagery and then fine-tuning them for 3D reflectivity reconstruction. Compared with studies relying only on global reflectivity summaries, their evaluation is broader, including perceptual, segmentation, cloud-type, and regional diagnostics. However, the analysis remains primarily centered on reconstruction skill itself rather than on the systematic validation of derived variables such as CTH and CBH.

Taken together, these studies establish the feasibility of inferring vertical cloud information from passive observations. However, they do not yet determine systematically which derived descriptors remain robust once a 3D field has been reconstructed. This distinction is important for downstream interpretation: scene-based radar reconstructions, curtain-based cloud-mask metrics, cloud-amount validation, global reflectivity scores, and climatological diagnostics do not necessarily indicate whether threshold-derived cloud occurrence, cloud-top and cloud-base height, cloud thickness, multilayer structure, or vertically aggregated summaries of cloud vertical placement can be used reliably. Table 1 summarizes this positioning.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


**Table 1.** Summary of recent machine-learning studies inferring vertical cloud structure from passive satellite observations. The comparison highlights the target variable, evaluation support, validation focus, and remaining gap relative to the descriptor-oriented evaluation proposed here.

| Study | Target and output | Evaluation support | Main validation focus | Gap relative to the present study |
|---|---|---|---|---|
| Leinonen et al. (2019) | CloudSat reflectivity; 2D CloudSat-like radar scenes from MODIS cloud-property inputs | Collocated CloudSat-MODIS scenes; curtain-scale validation | RMSE; reflectivity-altitude distributions; qualitative and probabilistic examples | No full-domain 3D reconstruction; no occurrence, geometry, thickness, layer-count, or centroid diagnostics |
| Spradlin et al. (2024) | CloudSat-CALIPSO vertical cloud mask; curtain-based cloud-mask retrieval after foundation-model pre-training | Large-scale pre-training; downstream validation on active-sensor curtains | SSIM for pre-training; mIOU; accuracy; AUC | Cloud-mask task, not reflectivity reconstruction; no reflectivity-amplitude or geometry-derived diagnostics |
| Lin et al. (2025) | CALIPSO/CALIOP cloud amount; 3D cloud amount from FY-4A AGRI and ERA5 predictors | Profile-level validation; seasonal and monthly analyses | MAE; MBE; RMSE; Pearson's $r$; uncertainty; cloud-amount-weighted height | Cloud amount, not reflectivity; no reflectivity-derived CTH/CBH, thickness, intensity-error, or multilayer diagnostics |
| Jeggle et al. (2024) | DARDAR ice-cloud microphysics; 3D IWC and $N_{\mathrm{ice}}$ from SEVIRI IR and auxiliary metadata | Independent-year tropical-domain evaluation; pixel-, overpass-, and climatology-scale checks | Ice-cloud occurrence; IWC and $N_{\mathrm{ice}}$ errors; climatological consistency; case studies | Rich evaluation, but ice-cloud microphysics only; not an all-cloud reflectivity or geometry-descriptor assessment |
| Brüning et al. (2024) | CloudSat reflectivity; all-cloud 3D SEVIRI-CloudSat reflectivity reconstruction | Most direct baseline; independent test restricted to May 2016 | Global RMSE; height-dependent RMSE; reflectivity distributions; monthly CTH against CLAAS | Direct target, but one-month evaluation; limited descriptor-level validation beyond reflectivity and aggregated CTH |
| Girtsou et al. (2025) | CloudSat reflectivity; 3D SEVIRI-CloudSat reconstruction with masked autoencoders | One-year SEVIRI–CloudSat setting; regional, monthly, and cloud-type analyses | RMSE; percentage error; PSNR; SSIM; segmentation F1; cloud-type diagnostics | Broad reconstruction metrics, but no separated validation of occurrence, geometry, thickness, multilayer collapse, and bulk placement |

### 1.2 Research novelty, contributions, and paper structure

The present study revisits SEVIRI-CloudSat cloud tomography by taking the all-cloud reflectivity reconstruction framework of Brüning et al. (2024) as the reference baseline. The objective is not to introduce a new retrieval architecture, but to assess more systematically what information on vertical cloud structure is actually preserved by this type of reconstruction. The general

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


learning problem remains the reconstruction of CloudSat-like 94 GHz radar reflectivity from eleven SEVIRI channels, with implementation and pre-processing adaptations described in Sect. 2.

The evaluation strategy is therefore structured around three limitations of existing SEVIRI-CloudSat tomography studies: the limited temporal support of previous validation, the dominance of global reflectivity metrics, and the lack of systematic assessment of derived cloud descriptors. First, the reconstruction is tested on an independent validation year rather than on a single test month, allowing a broader sampling of seasons, latitude bands, and cloud regimes. Second, direct reflectivity fidelity is separated from diagnostics derived after thresholding and aggregation, including cloud occurrence, CTH, CBH, thickness, multilayer preservation, and vertically aggregated summaries of cloud vertical placement. Third, all diagnostics are interpreted on the CloudSat-verifiable support, so that the reconstruction is assessed as a CloudSat-like surrogate of radar-detectable cloud structure rather than as an unbiased 3D cloud truth.

This framework allows to distinguish descriptors that remain informative from descriptors that are systematically degraded, and thus to clarify the level of physical interpretation that current SEVIRI-based cloud tomography can reasonably support.

The remainder of the paper describes the data, pre-processing and model configurations (Sect. 2), the evaluation framework (Sect. 3), the results (Sect. 4), and the implications for SEVIRI-based cloud tomography and downstream applications (Sect. 5), before concluding in Sect. 6.

## 2 Data and methods

This section describes the datasets (Sect. 2.1), the construction of the SEVIRI-CloudSat matched samples (Sect. 2.2), the data split strategy used for the reference-compatible consistency check and for the main independent evaluation protocol (Sect. 2.3), the baseline neural-network configuration and hyperparameter tuning (Sect. 2.4), and the cloud properties derived from the reconstructed reflectivity field (Sect. 2.4.1). The reconstruction task and baseline architecture follow the all-cloud SEVIRI-CloudSat framework of Brüning et al. (2024). The reconstruction is evaluated on an independent year using a descriptor-oriented framework that quantifies how reflectivity errors propagate into the selected cloud descriptors on the CloudSat-verifiable support. The training, inference, and support-restricted evaluation workflow is summarized in Fig. 1.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


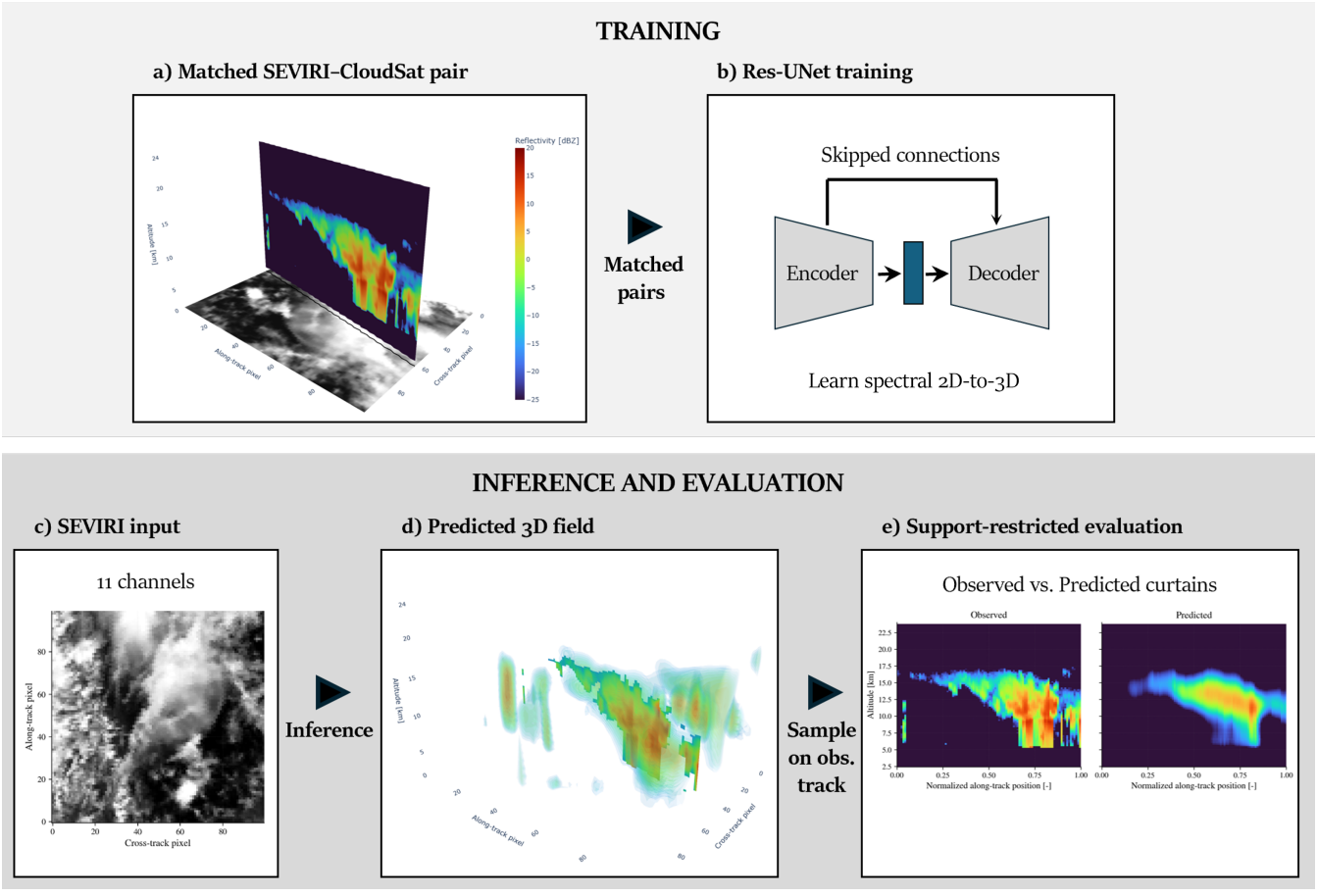


**Figure 1.** Overview of the SEVIRI-CloudSat tomography workflow. During training (top row), collocated SEVIRI tiles and CloudSat reflectivity curtains are used as matched pairs to train the Res-UNet. At inference (bottom row), only the SEVIRI tile is provided, and the network reconstructs a CloudSat-like 3D reflectivity field on the common grid. For quantitative evaluation, the predicted field is sampled along the observed CloudSat track, and the comparison is restricted to the CloudSat-observed curtain.

## 2.1 Data sources

The reconstruction task is formulated as a supervised mapping from SEVIRI spectral images to CloudSat 94 GHz radar reflectivity on a fixed three-dimensional grid.

The geostationary predictors are derived from the Spinning Enhanced Visible and InfraRed Imager (SEVIRI) on board Meteosat Second Generation (MSG). SEVIRI provides 12 spectral channels at a nominal nadir resolution of about 3 km and a full-disk repeat cycle of 15 min (Schmetz et al., 2002). Following Brüning et al. (2024), 11 channels are used as input, excluding the high-resolution visible (HRV) channel because of its different sampling. These include three solar-reflective bands (VIS0.6, VIS0.8, NIR1.6), one mixed solar–thermal band (IR3.9), two water-vapour bands (WV6.2, WV7.3), and five thermal infrared bands (IR8.7, IR9.7, IR10.8, IR12.0, IR13.4). In contrast to Brüning et al. (2024), which relied on SEVIRI

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


level-1.5 data in ICARE HDF format, native EUMETSAT level-1.5 files in SEVIRI geostationary projection are used here and resampled onto a regular $2400 \times 2400$ latitude-longitude grid covering approximately $60°$ S–$60°$ N and the Meteosat full disk.

The vertical reference is provided by CloudSat 2B-GEOPROF, which contains nadir-looking 94 GHz radar reflectivity profiles on 125 vertical bins with a nominal spacing of about 240 m from the surface to roughly 30 km (Marchand et al., 2008). To remain consistent with Brüning et al. (2024), the lowest levels affected by ground clutter and attenuation and the upper predominantly cloud-free levels are discarded, and the analysis is restricted to 90 levels between approximately 2.4 and 24 km. Reflectivity is expressed in dBZ, converted from centi-dBZ when read from file, and clipped to the range $[-35, 20]$ dBZ. Because the Cloud Profiling Radar has finite sensitivity and a near-surface blind zone caused by ground clutter and attenuation, thin cirrus, tenuous mixed-phase layers, drizzle, and many shallow boundary-layer clouds are often not captured in 2B-GEOPROF, especially below about 2-3 km (Marchand et al., 2008). The learning target therefore represents radar-detectable clouds within the 2.4-24 km layer rather than total cloud occurrence.

Samples for which the corresponding SEVIRI acquisition is unavailable are excluded. Within otherwise available SEVIRI input tiles, remaining non-finite values are imputed using a distance-weighted K-nearest-neighbour method with four neighbours. Each channel is then standardized using channel-wise $z$-score normalization based on training-set means and standard deviations.

### 2.2 Matching routine

The SEVIRI-CloudSat matching procedure follows the general strategy introduced by Brüning et al. (2024). CloudSat 2B-GEOPROF profiles are first screened using the CPR cloud mask, and only samples with mask quality values of at least 6 are retained. Profile times are converted to UTC and rounded to the nearest 15 min, and each profile is associated with the nearest SEVIRI scan.

For each time slot within the MSG reference domain, CloudSat profiles inside the SEVIRI full-disk domain are mapped onto the fixed $2400 \times 2400$ MSG grid and grouped into $128 \times 128$ tiles following the ground track. The corresponding SEVIRI channels are cropped to the same tile window, and CloudSat reflectivity is stored on that grid as a sparse three-dimensional target array. Valid reflectivity values are available only along the CloudSat curtain, while all other voxels are treated as missing. If several CloudSat samples fall into the same MSG pixel, the maximum reflectivity is retained, yielding a one-pixel-wide transect at MSG resolution.

To reduce the dominance of clear-sky scenes in the training archive, an empirical filtering step similar to the one described in Brüning et al. (2024) is introduced. Reflectivity is thresholded at $-15$ dBZ, and a vertical level is considered active if at least 10 voxels exceed this threshold. Only tiles with more than 10% active vertical levels are retained. The resulting archive still contains clear-sky voxels, but each tile is required to include a non-negligible amount of radar-detectable cloud signal. These matched SEVIRI tiles and CloudSat curtains constitute the paired training samples used to learn the 2D-to-3D mapping illustrated in Fig. 1.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


### 2.3 Dataset split and sampling strategy

Two training/validation split configurations are considered in this study, but only one is used for the main experiments and for all results reported in the remainder of the paper. Throughout, validation data are used during model development and checkpoint selection, whereas test or evaluation data are kept independent and used only for final performance assessment. The first configuration is used only as a reference-compatible verification step against the SEVIRI-CloudSat setup of Brüning et al. (2024). Its purpose is to check that the present implementation reproduces the performance range of the reference framework under the same temporal split. The second, seasonally distributed configuration defines the actual training, internal-validation, and independent evaluation protocol adopted for the present analysis.

The reference-compatible configuration follows the chronological split used by Brüning et al. (2024): matched samples from January-September 2017 are used for training, while matched samples from October-December 2017 are used for validation. This yields 12,567 training samples and 3,685 validation samples. The corresponding May 2016 reference test set contains 1,500 samples. Under this setup, the present implementation yields a clipped profile-mean RMSE of 2.78 dBZ on the May 2016 test set, close to the 2.99 dBZ reported by Brüning et al. (2024). This agreement is used as an implementation consistency check, supporting the comparability of the present pipeline with the reference SEVIRI-CloudSat framework.

The main experiments use a seasonally distributed split designed to provide a more balanced basis for internal validation and for the final independent performance assessment. Most matched samples from 2017 are assigned to training, while a smaller subset is reserved for internal validation by selecting samples from three target days per month, around the 1st, 10th, and 20th, with minor adjustments when CloudSat coverage is unavailable. This strategy yields 14,106 training samples and 2,146 internal-validation samples distributed across all months of the year.

All available matched samples from 2019 are assigned to an independent evaluation year. Year 2018 is excluded because the CloudSat record is incomplete. The final protocol used in this study therefore combines a continuous training year, a sparse but seasonally distributed internal-validation subset within 2017, and a fully independent 2019 evaluation year of 17,110 matched samples.

Because samples extracted from the same CloudSat overpass are strongly correlated, splitting is performed at the overpass level, so that all samples from a given overpass are assigned to the same subset. The advantage of the seasonally distributed split over the chronological October-December validation is quantified by training a lightweight classifier to distinguish training from validation samples using day of year, hour, mean longitude, and mean latitude. The resulting area under the curve (AUC) is 0.53 for the seasonally distributed split and 1.00 for the chronological split, confirming the much stronger seasonal separation of the latter.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


**Table 2.** Summary of the data partitions used in this study for the reference-compatible consistency check and for the main protocol.

| Split | Use | $N_{\text{train}}$ | $N_{\text{val}}$ | Selected months | $N_{\text{val}_{\text{days}}}$ | 2019 eval. | 2016 test | AUC |
|---|---|---|---|---|---|---|---|---|
| Chronological split | Consistency check | 12,567 | 3,685 | 10-12 | 61 | - | 1,500 | 1.00 |
| Seasonally distributed | Main protocol | 14,106 | 2,146 | 1-12 | 36 | 17,110 | - | 0.53 |

## 2.4 Neural-network baseline architecture and training protocol

The baseline model is a residual U-Net (Res-UNet) following Brüning et al. (2024). Input tiles are $128 \times 128$ SEVIRI patches with 11 radiance channels, and the network predicts 94 GHz radar reflectivity on 90 vertical levels over an inner $100 \times 100$ patch. The architecture follows a standard encoder-decoder design with skip connections and residual blocks. As in Brüning et al. (2024), border pixels are cropped so that only the central $100 \times 100$ region is used for loss computation and evaluation.

Inputs and targets are normalised as described in Sect. 2.1: SEVIRI channels are standardised with channel-wise $z$-scores, while CloudSat reflectivities are clipped to the range $[-35, 20]\,\text{dBZ}$ and linearly rescaled to $[-1, 1]$. Training and internal validation are restricted to daytime scenes (07:00-18:00 UTC) to avoid very low solar-elevation conditions when visible channels are used.

The learning objective is formulated as a voxel-wise regression of reflectivity. Voxels where CloudSat has no valid measurement are excluded from the loss. The baseline loss is the mean-squared error:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N_{\text{valid}}} \sum_{i \in \Omega_{\text{valid}}} (\hat{r}_i - r_i)^2 \tag{1}$$

where $\hat{r}$ and $r$ denote the predicted and target reflectivity fields on the inner $100 \times 100$ patch, expressed in the same normalised space and defined on 90 vertical levels, $\Omega_{\text{valid}}$ is the set of voxels with finite CloudSat reflectivity and $N_{\text{valid}}$ its cardinality.

The final training configuration used in the main experiments is summarized in Table 3.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


**Table 3.** Baseline model architecture and training configuration used in the main experiments.

| Setting | Value |
|---|---|
| Backbone | Res-UNet |
| Input channels | 11 |
| Input size | $128 \times 128$ |
| Output levels | 90 |
| Output crop | $100 \times 100$ |
| Start filters | 32 |
| Depth | 4 |
| Optimizer | Adam |
| Learning rate | $10^{-4}$ |
| Weight decay | $3 \times 10^{-4}$ |
| Batch size | 8 |
| Maximum epochs | 50 |
| Loss | MSE |
| Reflectivity range | $[-35, 20]$ dBZ |
| Normalised target range | $[-1, 1]$ |
| Daytime restriction | 07:00-18:00 UTC |
| Data augmentation | Random $90^\circ$ rotation ($p = 0.25$) |

#### 2.4.1 Derived products from reconstructed reflectivity

Several cloud properties are derived from the reconstructed three-dimensional reflectivity field for the performance. A binary cloud mask is defined using a fixed threshold of $-15$ dBZ, consistent with the filtering applied during dataset construction. For each vertical column within the 2.4-24 km layer, cloud-top height (CTH) and cloud-base height (CBH) are defined as the highest and lowest levels, respectively, at which reflectivity exceeds $-15$ dBZ. Columns without any exceedance are treated as clear sky. The choice of the $-15$ dBZ threshold it is commonly adopted in CloudSat-based analyses to isolate meaningful radar-detectable cloud or precipitation echoes from weak background returns, and it is also consistent with the baseline study used here for comparison (Marchand et al., 2008; Brüning et al., 2024).

Additional column-integrated descriptors are also extracted, including maximum reflectivity, geometric cloud thickness, and the number of distinct cloud layers, defined as contiguous cloudy segments separated by at least one clear level. These quantities provide a direct link between voxel-level reflectivity skill and more aggregated descriptors of cloud structure.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


## 3 Evaluation framework

Evaluating a mapping from geostationary radiances to a three-dimensional CloudSat-like reflectivity field requires a support-aware and descriptor-oriented strategy. The model predicts a full 3D reflectivity volume on the common grid, whereas CloudSat provides direct observational support only along a narrow curtain. All quantitative diagnostics are therefore restricted to the CloudSat-verifiable support, and the resulting scores should be interpreted as an evaluation of the CloudSat-like reconstruction where an active-sensor reference is available, not as a validation of the full predicted 3D tile.

A single global RMSE on valid CloudSat voxels remains useful as a compact reference metric and for comparison with previous work, but it is not sufficient to characterize reconstruction quality. The valid support is strongly imbalanced: weak or near-background reflectivity values dominate, while radar-detectable cloud occupies a smaller fraction of the sampled voxels. Global reflectivity errors must therefore be complemented by diagnostics that isolate cloud-conditioned performance and test how errors propagate into threshold-derived and vertically aggregated cloud descriptors.

Table 4 summarizes the evaluation logic adopted in this study. The framework separates direct reflectivity fidelity from cloud occurrence and detection, threshold-derived cloud geometry, bulk vertical organisation, and stratified interpretation by cloud regime, season, and latitude band.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


**Table 4.** Overview of the evaluation framework. Diagnostics are computed only on CloudSat-verifiable support and are organised from direct reflectivity errors to threshold-derived descriptors, aggregated vertical summaries, and stratified interpretation.

| Evaluation level | Support / operation | Main diagnostics | Interpretive role |
|---|---|---|---|
| Reflectivity fidelity | CloudSat-valid voxels; direct comparison of observed and predicted reflectivity | All-valid RMSE and bias; cloud-conditioned RMSE and bias; observed–predicted distributions; error versus observed reflectivity | Tests whether reflectivity amplitude is preserved, and separates background-dominated scores from errors within radar-detectable cloud |
| Cloud occurrence and detection | Thresholded fields on verifiable voxels and columns, using $\tau = -15\,\mathrm{dBZ}$ | Voxel cloud fraction; cloudy-column fraction; vertical occurrence profile; POD; FAR; CSI | Assesses whether the amount, vertical distribution, and column-level detection of radar-detectable cloud are reproduced |
| Cloud geometry | Threshold-derived boundaries on verifiable cloudy columns | CTH; CBH; envelope thickness; total cloudy thickness; geometry coverage | Tests whether cloud boundaries and vertical extent remain reliable after thresholding; coverage indicates how much of the observed cloudy population can be compared |
| Vertical organisation | Scene-level aggregation of thresholded cloud occurrence | Vertical centroid; cloudy-scene recall; correlation; bias; MAE | Evaluates whether the bulk vertical placement of cloud is preserved even when amplitude, boundaries, or layer separation are imperfect |
| Validity domain | Previous diagnostics stratified by scene characteristics | Cloud-fraction/intensity regimes; season; latitude band | Identifies the conditions under which each descriptor remains informative, rather than relying only on global averages |

### 3.1 Reference support and thresholds

All diagnostics reported in this study are restricted to CloudSat-verifiable support. Voxel-wise quantities are computed only where CloudSat provides a valid reflectivity value and where a corresponding prediction is available. For column-based diagnostics, a column is considered verifiable if CloudSat contains at least one valid voxel in that column. Column-based occurrence

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


labels and threshold-derived geometry products are then evaluated only on the observation-defined support of those verifiable columns. Consequently, the reported metrics quantify performance only where CloudSat provides direct observational support and should not be interpreted as a validation of the full reconstructed 3D tile.

This distinction is important because predicted cloud outside the CloudSat-verifiable support cannot be scored as either correct or incorrect. Off-support predictions are therefore excluded from the formal performance assessment. They may still be analysed descriptively or through other products, but they are not part of the reference-based skill evaluation presented here.

In the present study, a voxel is classified as cloudy when reflectivity exceeds $\tau = -15\,\mathrm{dBZ}$. As mentioned in Sect. 2.4.1, this threshold is chosen to be consistent with both sample selection and the baseline study used here for comparison (Brüning et al., 2024), and because it is commonly adopted in CloudSat-based analyses to isolate radar-detectable cloud echoes (Marchand et al., 2008).

A second threshold, $\tau_{\mathrm{core}} = 0\ \mathrm{dBZ}$, is used only in diagnostics targeting the stronger part of the reflectivity distribution. The main goal in this case is to test whether the reconstruction preserves the more intense echoes, which are generally associated with optically thicker or more vertically developed structures and are expected to be harder to recover from passive radiances.

### 3.2 Reflectivity fidelity

The first level of evaluation concerns reflectivity amplitude. The reflectivity scores used in this study serve two purposes: a first set is retained for consistency with the reference SEVIRI-CloudSat evaluation of Brüning et al. (2024), while a second set is used as the main diagnostic to interpret reconstruction errors.

For reference compatibility, two clipped RMSE scores are reported. Here, *clipped* means that observed and predicted reflectivity values are limited to the interval $[-25, 20]\ \mathrm{dBZ}$ before computing the error: values below $-25\,\mathrm{dBZ}$ are set to $-25\,\mathrm{dBZ}$, and values above $20\ \mathrm{dBZ}$ are set to $20\ \mathrm{dBZ}$. This clipping reproduces the reflectivity range used in the reference evaluation and ensures comparability with earlier work. These scores are therefore reported for consistency, but they are not the main basis for interpreting reconstruction quality.

The first reference-compatible score is a clipped curtain RMSE, where *curtain* denotes the vertical cross-section sampled by CloudSat within each predicted 3D tile. The *clipped curtain* RMSE thus pools all CloudSat-valid voxels along it and computes one average error per sample. The second score is a *clipped profile-mean* RMSE, where the RMSE is first computed at each analysed altitude level using the available CloudSat-valid voxels, and then averaged vertically. Compared with the curtain RMSE, the profile-mean RMSE gives a more balanced weight to different altitude levels.

The main diagnostics used here are instead based on the unmodified reflectivity field on CloudSat-valid voxels. Bias and RMSE are computed as:

$$\mathrm{Bias} = \langle Z_{\mathrm{pred}} - Z_{\mathrm{obs}} \rangle \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\left\langle \left(Z_{\mathrm{pred}} - Z_{\mathrm{obs}}\right)^2 \right\rangle} \tag{3}$$

Here, $Z_{\mathrm{pred}}$ and $Z_{\mathrm{obs}}$ are the predicted and observed radar reflectivity values in $\mathrm{dBZ}$ on the common grid, and $\langle \cdot \rangle$ denotes averaging over CloudSat-valid voxels. In the Results section, this quantity is referred to as the *all-valid voxel RMSE*.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


An additional *cloud-conditioned voxel RMSE* is computed as:

$$\mathrm{RMSE}_{\mathrm{cld}} = \sqrt{\left\langle \left(Z_{\mathrm{pred}} - Z_{\mathrm{obs}}\right)^2 \right\rangle_{Z_{\mathrm{obs}} > \tau}} \tag{4}$$

In this expression, $\tau = -15\,\mathrm{dBZ}$, and $\langle \cdot \rangle_{Z_{\mathrm{obs}} > \tau}$ denotes averaging over the subset of CloudSat-valid voxels for which the observed reflectivity exceeds the cloud-detection threshold.

The all-valid voxel RMSE and the cloud-conditioned voxel RMSE answer different questions: the former summarizes the error over the full CloudSat-valid support, whereas the latter isolates performance within radar-detectable cloud. Their comparison is important because a reconstruction can show moderate all-valid errors while still underestimating the amplitude of actual cloud echoes.

To characterize intensity-dependent behaviour, the signed error $Z_{\mathrm{pred}} - Z_{\mathrm{obs}}$ is also analysed as a function of observed reflectivity on CloudSat-valid voxels, and observed-predicted joint distributions are examined at representative altitudes on the same verifiable support. These diagnostics are intended to reveal systematic attenuation of moderate and strong echoes that may remain hidden in scalar summary statistics.

### 3.3 Threshold-derived cloud occurrence

Once thresholded at $\tau = -15\,\mathrm{dBZ}$, the reconstructed reflectivity field is also interpreted as a binary radar-detectable cloud mask on the verifiable CloudSat support. Many downstream quantities, including cloudy-column fraction, cloud-top height, cloud-base height, can be derived from this thresholded field rather than from reflectivity amplitude directly.

Three complementary occurrence diagnostics are considered here.

The first is the voxel cloudy fraction, defined as:

$$CF_{\mathrm{vox}} = \frac{\sum_{z,y,x} \mathbb{1}[Z(z,y,x) > \tau]}{N_{\mathrm{vox,valid}}} \tag{5}$$

Here, $Z$ denotes either the observed or predicted reflectivity field in $\mathrm{dBZ}$ on the common grid, $\mathbb{1}[\cdot]$ is the indicator function, equal to 1 when the condition is satisfied and 0 otherwise, $\tau$ is the cloud-detection threshold (i.e. $-15\,\mathrm{dBZ}$), and $N_{\mathrm{vox,valid}}$ is the number of CloudSat-valid voxels.

The second is the cloudy-column fraction, defined as:

$$CF_{\mathrm{col}} = \frac{\sum_{y,x} \mathbb{1}[\max_z Z(z,y,x) > \tau]}{N_{\mathrm{col,verif}}} \tag{6}$$

Here, $N_{\mathrm{col,verif}}$ is the number of verifiable columns, that is, columns containing at least one CloudSat-valid voxel. The vertical maximum is evaluated only over valid levels within each verifiable column. This quantity therefore measures the fraction of verifiable columns containing at least one cloudy voxel.

The third is the vertical cloud-occurrence profile, defined as:

$$p_{\mathrm{cld}}(z) = \frac{\sum_{y,x} \mathbb{1}[Z(z,y,x) > \tau]}{N_{\mathrm{valid}}(z)} \tag{7}$$

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


Here, $N_{\mathrm{valid}}(z)$ is the number of CloudSat-valid voxels at altitude level $z$. This profile describes how cloudy occurrence is distributed with height.

These diagnostics capture different aspects of the thresholded field. $CF_{\mathrm{vox}}$ measures how much of the sampled support is occupied by radar-detectable cloud, $CF_{\mathrm{col}}$ is more sensitive to the prevalence of cloudy columns, and $p_{\mathrm{cld}}(z)$ describes the vertical allocation of cloud occurrence.

Scene-level occurrence biases are also reported as:

$$\Delta CF_{\mathrm{vox}} = CF_{\mathrm{vox,pred}} - CF_{\mathrm{vox,obs}} \tag{8}$$

$$\Delta CF_{\mathrm{col}} = CF_{\mathrm{col,pred}} - CF_{\mathrm{col,obs}} \tag{9}$$

Here, the subscripts $\mathrm{pred}$ and $\mathrm{obs}$ denote quantities derived from the predicted and observed reflectivity fields, respectively. An error in $CF_{\mathrm{col}}$ points primarily to a bias in the number of cloudy columns, whereas an error in $CF_{\mathrm{vox}}$ may arise from either horizontal or vertical underfilling.

### 3.4 Column-level cloud detection skill

The thresholded field also allows a classical detection-oriented evaluation on verifiable CloudSat columns. At column level, both observation and prediction are converted into binary labels according to whether at least one voxel within the CloudSat-supported part of the column exceeds $\tau$.

From the resulting contingency table, the probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI) are computed:

$$\mathrm{POD} = \frac{TP}{TP+FN} \tag{10}$$

$$\mathrm{FAR} = \frac{FP}{TP+FP} \tag{11}$$

$$\mathrm{CSI} = \frac{TP}{TP+FP+FN} \tag{12}$$

Here, $TP$, $FP$, and $FN$ denote true positives, false positives, and false negatives, respectively. These metrics complement RMSE because they evaluate whether radar-detectable cloudy columns are identified at all, regardless of the exact reflectivity amplitude within them. They are computed only on verifiable CloudSat columns and therefore do not penalize predicted cloud outside the CloudSat-observed support.

### 3.5 Threshold-derived cloud geometry

Cloud-top height (CTH), cloud-base height (CBH), and cloud thickness are derived from the thresholded reflectivity field on verifiable CloudSat support. These quantities are more directly interpretable than voxel-level reflectivity, but they are also more sensitive to thresholding and to the loss of weak echoes near cloud boundaries.

For each verifiable column classified as cloudy, CTH and CBH are defined as the highest and lowest altitude levels, respectively, at which $Z > \tau$, with $\tau = -15\,\mathrm{dBZ}$. Two thickness measures are considered. The first is the envelope thickness, defined

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


as:

$$T_{\mathrm{env}} = \mathrm{CTH} - \mathrm{CBH} + \Delta z \tag{13}$$

where $\Delta z$ is the vertical grid spacing. This quantity spans the full cloudy envelope between base and top and therefore includes possible gaps between layers.

The second is the total cloudy thickness, defined as:

$$T_{\mathrm{tot}} = \Delta z \sum_{z} \mathbb{1}[Z(z,y,x) > \tau] \tag{14}$$

Here, $Z(z,y,x)$ is the reflectivity value at altitude level $z$ in column $(y,x)$, $\tau$ is the cloud-detection threshold ($-15\,\mathrm{dBZ}$), $\mathbb{1}[\cdot]$ is the indicator function, and the summation is evaluated over the analysed levels of the CloudSat-supported part of the column. This quantity counts only the cloudy levels themselves and therefore excludes clear gaps between layers.

Both quantities are retained because they describe different aspects of the vertical structure. $T_{\mathrm{env}}$ is sensitive to the overall extent of the cloudy envelope, whereas $T_{\mathrm{tot}}$ is less affected by gaps and therefore remains informative even when multilayer structure is not fully preserved.

The evaluation of geometry is inherently conditional, because CTH, CBH, and thickness are defined only in columns classified as cloudy. This point is particularly important for CBH. In the present setup, the lower part of the CloudSat profile is excluded to reduce ground-clutter contamination, so CBH should not be interpreted as the true physical cloud base. It is the lowest radar-detectable cloudy level within the truncated 2.4-24 km analysis range. For this reason, CBH is expected to be less realistic and more difficult to interpret than CTH, especially for shallow or lower-tropospheric clouds.

Geometry errors are therefore evaluated only where the corresponding quantity is defined in both observation and prediction. This conditional evaluation is necessary, but it can make geometry scores appear better than they are if many observed-cloudy columns are missed entirely. To make this explicit, a geometry coverage is also reported:

$$C_{\mathrm{geom}} = \frac{\sum_{y,x} \mathbb{1}[y_{\mathrm{obs}} = 1 \wedge y_{\mathrm{pred}} = 1]}{\sum_{y,x} \mathbb{1}[y_{\mathrm{obs}} = 1]} \tag{15}$$

Here, $y_{\mathrm{obs}}$ and $y_{\mathrm{pred}}$ are the binary cloudy-column labels derived from the observed and predicted fields, and the summations are evaluated over verifiable CloudSat columns. This quantity measures the fraction of observed-cloudy columns for which geometry is also defined in the prediction. In other words, it indicates how much of the observed cloudy population survives the thresholding step well enough for top, base, and thickness to be compared.

### 3.6 Scene-level vertical cloud organisation

Threshold-derived top and base heights are informative, but they remain sensitive to local edge errors, missing weak echoes, and the imperfect reconstruction of multilayer structure. An additional aggregated diagnostic is therefore introduced to characterize the bulk vertical allocation of cloud occurrence.

The centroid height of cloud occurrence is defined as:

$$z_c = \frac{\sum_z z\, p_{\mathrm{cld}}(z)}{\sum_z p_{\mathrm{cld}}(z)} \tag{16}$$

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


Here, $p_{\mathrm{cld}}(z)$ is the vertical cloud-occurrence profile defined in Sect. 3.3, $z$ is altitude, and the summation is performed over the analysed vertical levels. This quantity represents the mean vertical location of cloudy occurrence within the verifiable support of the evaluation sample.

The centroid is introduced because many applications depend more on the dominant vertical placement of cloud than on the exact reconstruction of every local boundary. This is the case, for example, in cloud-regime characterization, process-oriented model evaluation, and broader radiative interpretation, and potentially in some solar-radiation applications, where it may be more informative to know whether cloud occurrence is concentrated mainly in the lower, middle, or upper troposphere than to recover the exact top, base, or internal separation of each layer (Garnier et al., 2021; Kim et al., 2024).

### 3.7 Regime, seasonal, and latitudinal stratification

Global averages are not sufficient to characterize the usefulness of the reconstruction, because different evaluation samples do not pose the same level of difficulty. The analysis is therefore stratified in order to identify the conditions under which the reconstruction remains informative.

Samples are grouped according to the observed cloudy-column fraction $CF_{\mathrm{col,obs}}$ and the maximum observed reflectivity on CloudSat-valid voxels ($\max(Z_{\mathrm{obs}})$). The observed cloudy-column fraction is partitioned into three regimes, $CF_{\mathrm{col,obs}} < 0.2$, $0.2 \leq CF_{\mathrm{col,obs}} \leq 0.6$, and $CF_{\mathrm{col,obs}} > 0.6$, corresponding to low, medium, and high cloudy-column conditions on the verifiable support. Sample intensity is divided into weak and strong regimes using a threshold of $5\,\mathrm{dBZ}$ on the maximum observed reflectivity. Unlike the $-15\,\mathrm{dBZ}$ threshold used to define radar-detectable cloud occurrence, this $5\,\mathrm{dBZ}$ threshold is not intended as a universal physical threshold. It is introduced here as a practical stratification criterion to separate scenes dominated by weak returns from scenes containing clearly stronger echoes.

Additional stratification is performed by season (DJF, MAM, JJA, SON) and by broad latitude bands, defined as tropics ($|\phi| < 20°$), midlatitudes ($20° \leq |\phi| < 50°$), and high latitudes ($|\phi| \geq 50°$). The corresponding labels and thresholds are summarized in Table 5.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


**Table 5.** Labels and thresholds used for the regime, seasonal, and latitudinal stratifications.

| Category | Label | Definition |
|---|---|---|
| Cloudy-column fraction | Low | $CF_{\mathrm{col,obs}} < 0.2$ |
| Cloudy-column fraction | Medium | $0.2 \leq CF_{\mathrm{col,obs}} \leq 0.6$ |
| Cloudy-column fraction | High | $CF_{\mathrm{col,obs}} > 0.6$ |
| Scene intensity | Weak | $\max(Z_{\mathrm{obs}}) < 5\,\mathrm{dBZ}$ |
| Scene intensity | Strong | $\max(Z_{\mathrm{obs}}) \geq 5\,\mathrm{dBZ}$ |
| Season | DJF | December-January-February |
| Season | MAM | March-April-May |
| Season | JJA | June-July-August |
| Season | SON | September-October-November |
| Latitude band | Tropics | $\lvert\phi\rvert < 20°$ |
| Latitude band | Midlatitudes | $20° \leq \lvert\phi\rvert < 50°$ |
| Latitude band | High latitudes | $\lvert\phi\rvert \geq 50°$ |

This final step is intended to determine not only whether the reconstruction performs well on average, but also under which conditions it remains most informative and which aspects of the cloud field can be interpreted with greater confidence.

## 4 Results

All results are obtained on the independent 2019 evaluation year, restricted to daytime samples (07:00-18:00 UTC) and to CloudSat-supported voxels within the analysed vertical range (2.4-24 km). Unless otherwise stated, reflectivity statistics are computed on the valid CloudSat support, whereas occurrence and geometry diagnostics are evaluated from threshold-derived products on verifiable CloudSat support, as defined in Sect. 3.

The results are organised according to the hierarchy introduced in Sect. 3. The analysis starts from reflectivity fidelity itself and then moves to quantities derived after thresholding or vertical aggregation.

### 4.1 Reflectivity reconstruction and illustrative examples

The first step is to assess how faithfully the reconstructed field reproduces CloudSat reflectivity on the valid CloudSat support. Table 6 summarizes the main reflectivity metrics on the independent 2019 evaluation year. The table includes two types of scores: reference-compatible clipped scores, retained for comparison with previous SEVIRI-CloudSat reflectivity studies, and non-clipped diagnostics used to interpret the reconstruction errors.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


**Table 6.** Summary reflectivity metrics on the independent 2019 evaluation year.

| Metric | Median [IQR] (dBZ) | Mean (dBZ) |
|---|---|---|
| Reference-compatible curtain RMSE (clipped) | 4.45 [3.43, 5.62] | 4.59 |
| Reference-compatible profile-mean RMSE (clipped) | 2.76 [1.98, 3.72] | 2.95 |
| All-valid voxel RMSE | 5.81 [4.68, 7.08] | 5.98 |
| Cloud-conditioned voxel RMSE | 14.48 [11.71, 17.66] | 14.93 |
| All-valid voxel bias | +0.13 [-0.27, +0.49] | 0.07 |
| Cloud-conditioned voxel bias | -11.16 [-14.83, -7.87] | -11.59 |

The reference-compatible scores are based on reflectivity values clipped to the interval $[-25, 20]$ dBZ. The clipped curtain RMSE is computed over all valid voxels along the CloudSat-observed vertical section within each predicted tile, whereas the clipped profile-mean RMSE first computes errors separately at each altitude level and then averages them vertically. These scores remain moderate, with a median clipped curtain RMSE of 4.45 dBZ and a median clipped profile-mean RMSE of 2.76 dBZ. They indicate that, using the scoring conventions of earlier SEVIRI-CloudSat reflectivity studies, the reconstruction has comparable overall error levels.

However, the non-clipped diagnostics reveal a more contrasted behaviour. The all-valid voxel RMSE is higher, with a median of 5.81 dBZ, while the corresponding median bias remains close to zero, at (+0.13 dBZ). This near-zero bias over all valid voxels does not imply that cloudy regions are well reconstructed, because the valid support contains many weak or near-background reflectivity values. When the analysis is restricted to voxels that are cloudy in the CloudSat observation, the error increases strongly: the cloud-conditioned voxel RMSE reaches 14.48 dBZ, and the cloud-conditioned bias is (-11.16 dBZ). The dominant reflectivity failure mode therefore becomes apparent when the evaluation focuses specifically on radar-detectable cloud.

Figure 2 makes this point explicit by showing prediction error as a function of observed reflectivity, for all valid voxels and for the cloud-conditioned subset. In both panels, the bin-wise median shifts downward as observed reflectivity increases. This behaviour is already visible when all valid voxels are included, but it becomes much clearer once the background is removed. In the cloud-conditioned panel, the negative shift is systematic across the full observed range and becomes strongest for high reflectivity values. This pattern suggests an attenuation of moderate and strong echoes in the reconstructed field, but may also reflect the tendency of models trained with an MSE loss to regress toward the conditional mean and underestimate extreme values.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


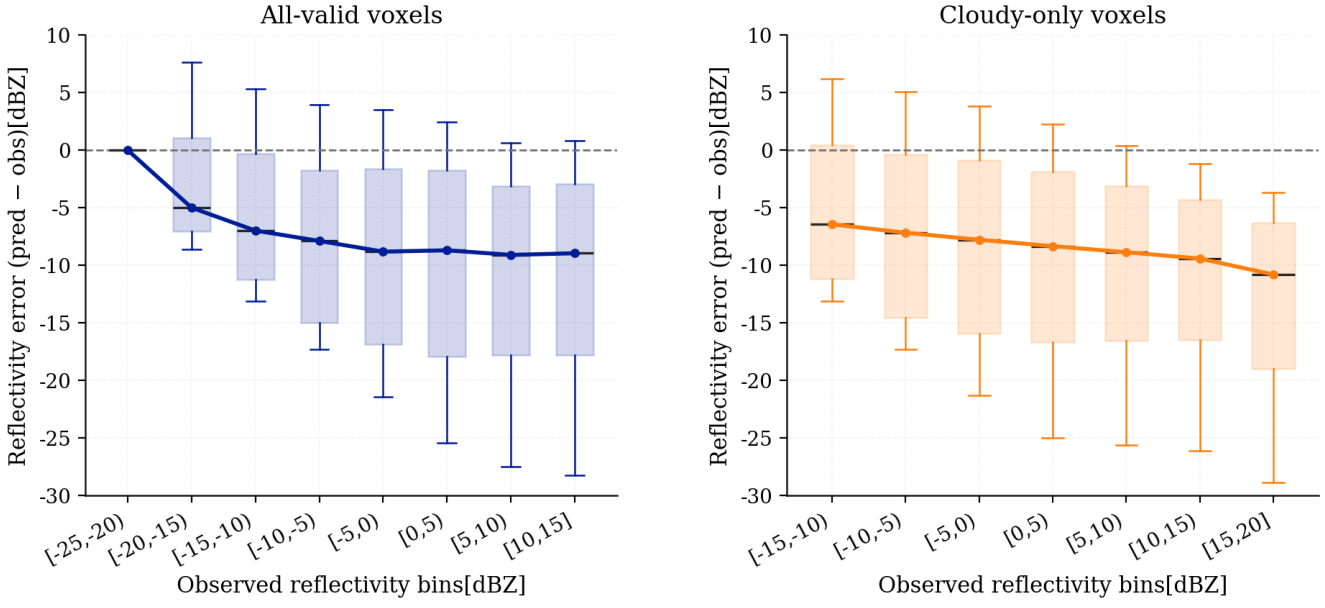


**Figure 2.** Reflectivity error as a function of observed reflectivity on CloudSat-valid support. Left: all valid voxels. Right: cloud-conditioned voxels ($Z_{\mathrm{obs}} > \tau$, with $\tau = -15\,\mathrm{dBZ}$). Boxes indicate the interquartile range, whiskers the 10th-90th percentiles, and lines the bin-wise median. Reflectivities are clipped to $[-25, 20]$ dBZ.

Figure 3 complements the vertically aggregated error analysis in Fig. 2 by resolving the observed-predicted relationship at three selected altitudes (6, 10, and 14 km). The plots show the distribution of all valid observed-predicted voxel pairs at each level. The orange line represents the mean predicted reflectivity within 5 dBZ classes of observed reflectivity. $R^2_{\mathrm{all}}$ is computed from all individual valid voxel pairs and measures how closely the overall observed-predicted distribution follows a linear relationship. At 6 km, the observed-predicted relationship is the weakest ($R^2_{\mathrm{all}} = 0.48$), and the class-mean response increasingly falls below $x = y$ as observed reflectivity increases. At 10 and 14 km, the voxel-wise relationship is stronger ($R^2_{\mathrm{all}} = 0.65$ and 0.64), but the same departure of the class mean from $x = y$ remains for moderate and strong echoes. Figure 3 therefore confirms, at all three altitudes, that predicted reflectivity generally increases with observed reflectivity, but stronger echoes are systematically underestimated, consistently with the increasingly negative errors shown in Fig. 2.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


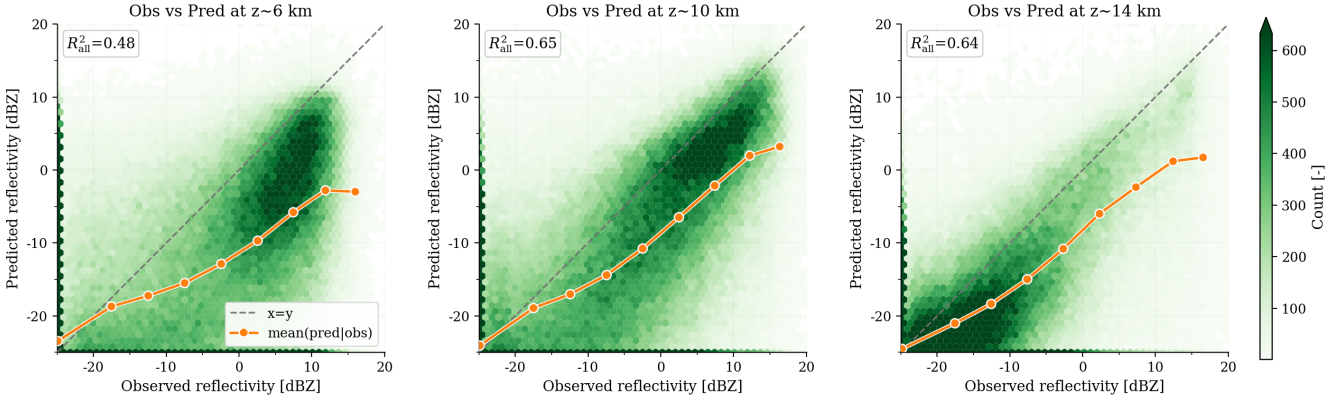


**Figure 3.** Observed versus predicted reflectivity at 6, 10, and 14 km on CloudSat-valid support, using all valid voxels at each level. Hexagons indicate sample density on a linear scale, with the upper colour limit set to the 95th percentile. The dashed line denotes $x = y$. The orange line shows the mean predicted reflectivity within 5 dBZ classes of observed reflectivity. $R^2_{\mathrm{all}}$ is computed from all valid voxel pairs at each level. Reflectivities are clipped to $[-25, 20]$ dBZ.

This interpretation is also visible in Fig. 4, which shows three illustrative samples selected from contrasted cloud regimes. These examples are included to qualitatively support the reading of the metrics discussed above, not to replace them. A broader set of 200 additional examples, selected by stratified random sampling across all cloud-cover and intensity regimes, is provided in the Supplement (Figs. S1–S200). Low, medium, and high cloud-cover scenes are defined by $CF_{\mathrm{col,obs}} < 0.2$, $0.2 \leq CF_{\mathrm{col,obs}} \leq 0.6$, and $CF_{\mathrm{col,obs}} > 0.6$, respectively, while weak and strong regimes are separated using a maximum observed reflectivity threshold of $5\,\mathrm{dBZ}$. In addition to the reflectivity curtains, Fig. 4 also shows the observed and predicted cloud-top and cloud-base heights derived from thresholded reflectivity ($\tau = -15\,\mathrm{dBZ}$). These curves are included here only as a qualitative visual indication of the reconstructed cloud envelope; their systematic evaluation is presented later.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


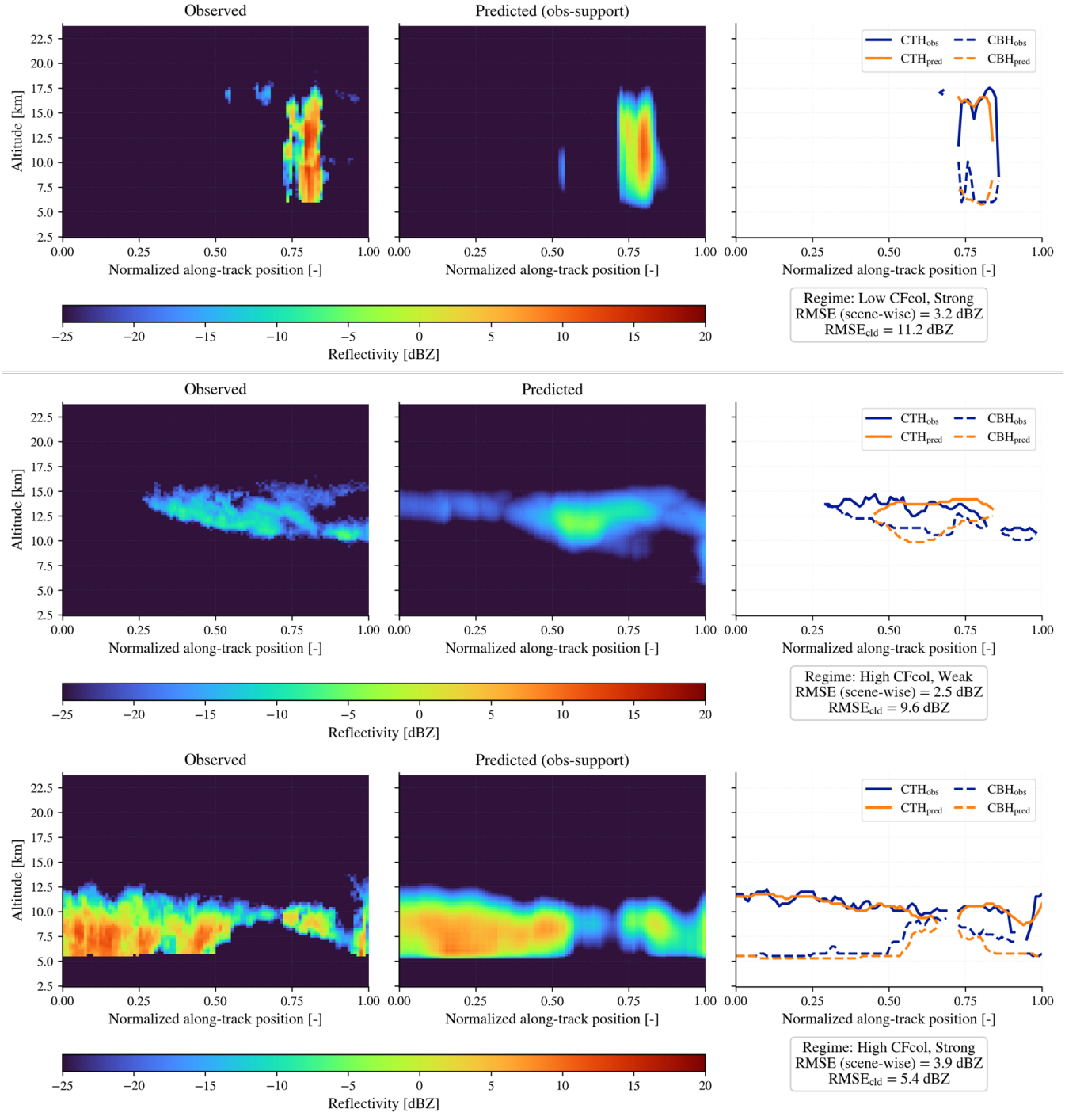


**Figure 4.** Representative observed and predicted CloudSat-like reflectivity samples. The first two panels show observed and predicted reflectivity on the CloudSat-valid curtain; the rightmost panel shows cloud-top height (CTH) and cloud-base height (CBH) derived from thresholded reflectivity ($\tau = -15\,\mathrm{dBZ}$). Reported scores are sample-wise reflectivity RMSE on valid voxels and on the cloud-conditioned subset.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


Across the selected cases, the dominant cloudy layer is generally reconstructed at the correct altitude and the broad distinction between lower, middle, and higher-level cloud is preserved. This confirms that the model does recover meaningful information on the large-scale vertical organisation of cloud. However, the same samples also show the limitations already identified in the statistical diagnostics. Intense echoes are weakened and spread over broader regions, and internally structured scenes tend to lose part of their separation between layers. These behaviours are consistent with a passive-to-3D inversion and with a voxel-wise squared-error objective, both of which tend to favour smoothed conditional-mean reconstructions.

Taken together, the results show that the model tends to reconstruct a physically meaningful but smoothed version of the CloudSat reflectivity field. It preserves the broad vertical placement of cloud and part of the large-scale reflectivity organisation, but systematically underestimates the reflectivity of moderate and strong echoes.

### 4.2 Threshold-derived cloud occurrence and vertical distribution

After thresholding the observed and reconstructed reflectivity fields at$\tau = -15$ dBZ, both fields can be interpreted as binary masks of radar-detectable clouds. This section focuses on the outcome of this thresholding step, evaluating how well the reconstruction reproduces the amount of detected cloud, the fraction of cloudy columns, and the altitude range where cloud occurrence is concentrated.

Three diagnostics are used to evaluate the impact of thresholding on the reconstructed cloud mask. The voxel cloud fraction $CF_{\mathrm{vox}}$ measures the fraction of voxels classified as cloudy within the sampled vertical section. The cloudy-column fraction $CF_{\mathrm{col}}$ measures the fraction of columns containing at least one cloudy voxel. This distinction is important because $CF_{\mathrm{col}}$ can decrease strongly when marginal or vertically thin cloudy columns are missed, whereas $CF_{\mathrm{vox}}$ may change less if those columns contain relatively few cloudy voxels or if thicker columns remain partly above the threshold. Conversely, errors in cloud thickness or vertical extent can affect $CF_{\mathrm{vox}}$ without necessarily changing $CF_{\mathrm{col}}$. The vertical occurrence profile $p_{\mathrm{cld}}(z)$ then describes where this thresholded cloud occurrence is located in altitude. Column-level POD, FAR, and CSI complement $CF_{\mathrm{col}}$ by evaluating detection at the column scale. These metrics indicate whether errors in $CF_{\mathrm{col}}$ arise from missed cloudy columns or from false detections.

Table 7 summarizes the main occurrence diagnostics and column-detection scores. The median bias in voxel cloud fraction is modest, with $\Delta CF_{\mathrm{vox}} = -0.020$ and a median absolute error of 0.022. By contrast, the bias in cloudy-column fraction is much more negative, with $\Delta CF_{\mathrm{col}} = -0.174$ and a median absolute error of 0.175. The detection scores clarify this error: the probability of detection is moderate (POD = 0.63), the false alarm ratio is very small (FAR = 0.031), and the critical success index remains moderate (CSI = 0.62). The thresholded reconstruction is therefore conservative: it generates few spurious cloudy columns, but it misses a non-negligible fraction of the observed ones.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


**Table 7.** Summary of threshold-derived cloud-occurrence diagnostics on the independent 2019 evaluation year. Bias and MAE are reported as sample-level medians with interquartile ranges. Column-detection scores are computed globally over CloudSat-verifiable columns using $\tau = -15\,\mathrm{dBZ}$.

| Product | Bias median [IQR] | MAE median [IQR] | POD | FAR | CSI |
|---|---|---|---|---|---|
| Voxel cloud fraction | $-0.020$ [$-0.035$, $-0.008$] | 0.022 [0.011, 0.036] | – | – | – |
| Cloudy-column fraction | $-0.174$ [$-0.286$, $-0.091$] | 0.175 [0.092, 0.287] | – | – | – |
| Column cloud detection | – | – | 0.63 | 0.031 | 0.62 |

Fig. 5 further examines whether these biases depend on the observed amount of cloud in each sample. The figure shows binned observed-predicted relationships: samples are grouped according to the observed value of $CF_{\mathrm{col}}$ or $CF_{\mathrm{vox}}$, and the solid line shows the conditional median of the predicted value in each bin. The right panel shows that $CF_{\mathrm{vox}}$ remains relatively close to the one-to-one line over most of its observed range, although a weak negative bias persists throughout. On the other hand, the left panel shows that for $CF_{\mathrm{col}}$ the underestimation is stronger and increases as the observed cloudy-column fraction grows. Near the upper end of the range, the predicted conditional median remains well below unity even when the observed support is almost fully cloudy. The numerical values reported in the inset boxes of Fig. 5 are global mean bias and MAE over all samples, whereas Table 7 reports medians and interquartile ranges.

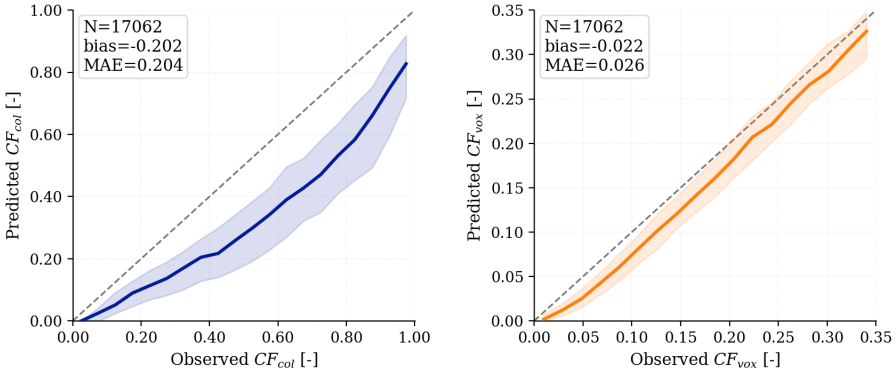


**Figure 5.** Binned observed-predicted relationships for threshold-derived cloud occurrence on CloudSat-verifiable support. Left: cloudy-column fraction $CF_{\mathrm{col}}$. Right: voxel cloud fraction $CF_{\mathrm{vox}}$. Shaded areas indicate the interquartile range and solid lines the conditional median.

Taken together, the two panels show that the reconstruction errors affect the two threshold-derived occurrence measures differently. The model loses cloud much more strongly in terms of the number of cloudy columns than in terms of the total number of cloudy voxels. This difference is consistent with the definitions of the two metrics: each cloudy column contributes equally to $CF_{\mathrm{col}}$, whereas its contribution to $CF_{\mathrm{vox}}$ depends on the number of cloudy voxels it contains. For example, missing

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


a column containing only a few cloudy voxels reduces $CF_{\mathrm{col}}$ by one full column count, while having only a small effect on $CF_{\mathrm{vox}}$. The two biases therefore are expected not to vary proportionally.

The mean vertical cloud-occurrence profile is shown in Fig. 6. Although the analysed vertical range begins at about 2.4 km, mean occurrence remains nearly zero below about 5 km because very few valid CloudSat observations remain at these levels after the applied quality screening. Above this level, the predicted curve reproduces the overall shape of the observed profile: cloud occurrence increases rapidly, reaches a maximum in the lower-to-mid troposphere, and then decreases progressively with altitude. This indicates that the reconstruction preserves the broad vertical band in which cloud occurrence is concentrated.

At the same time, the predicted profile remains below the observed one at almost all heights, indicating a systematic underestimation of threshold-derived cloud occurrence. The main occurrence maximum is still located in the same broad altitude range, but it is weaker and slightly shifted toward lower levels. Above the maximum, the predicted profile decays more rapidly than the observed one, showing that occurrence is also underestimated in the middle and upper part of the column.

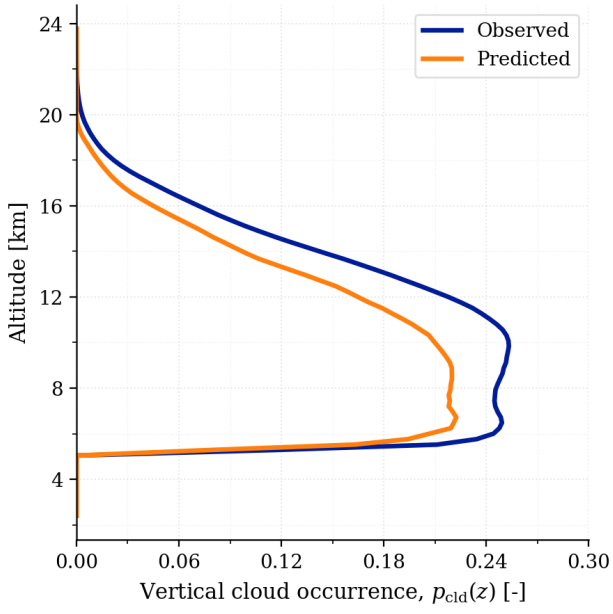


**Figure 6.** Mean vertical cloud-occurrence profile $p_{\mathrm{cld}}(z)$ for the observed and reconstructed fields on CloudSat-valid support ($\tau = -15$ dBZ). The evaluation grid extends from 2.4 to 24 km. The near-zero occurrence below approximately 5 km reflects the quality screening applied to the processed CloudSat reference.

These results suggest that the thresholded reconstruction is therefore more reliable as an indicator of the vertical region in which cloud occurrence is concentrated than as an unbiased estimate of the total amount of cloudy support throughout the profile. This distinction is important because applications based on cloud presence, cloud screening, or surface-radiation attenuation are more sensitive to missed cloudy columns, whereas applications focused on vertical cloud placement may still benefit from the reconstructed occurrence profile despite its negative bias.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


### 4.3 Conditional skill of threshold-derived cloud boundaries and thickness

This paragraph is aimed to assess how the thresholded reflectivity field translates into geometric quantities such as cloud-top height (CTH), cloud-base height (CBH), and cloud thickness. These diagnostics are conditional by construction: CTH, CBH, and thickness can only be compared in columns where cloud is detected in both the CloudSat observation and the reconstruction.

Table 8 summarizes the threshold-derived geometry diagnostics. The conditional RMSE values for CTH and CBH are of similar magnitude, with median values of 1.38 and 1.41 km, respectively. This indicates that the broad altitude of the upper and lower cloud boundaries is often recovered with reasonable accuracy. For CTH, this order of magnitude is qualitatively consistent with the CTH behaviour reported by Brüning et al. (2024), who found an overall mean difference of 1.28 km against monthly aggregated CLAAS-V002E1 cloud-top heights, together with a tendency of the Res-UNet to predict lower clouds and to underestimate higher clouds. A direct numerical comparison is not appropriate, however, because their CTH analysis was performed on monthly full-disk aggregates against an external SEVIRI-based product, whereas the present CTH scores are conditional and restricted to CloudSat-verifiable columns.

The median RMSE is 2.05 km for envelope thickness and 1.65 km for total cloudy thickness. The larger value obtained for envelope thickness should not, however, be interpreted directly as evidence that thickness is intrinsically more difficult to reconstruct than CTH or CBH. Since envelope thickness is defined as the difference between the two boundaries, its error combines errors in both CTH and CBH. As a simple propagation benchmark, assuming independent and unbiased boundary errors gives an expected RMSE of $\sqrt{1.38^2 + 1.41^2} = 1.97\,\mathrm{km}$, which is close to the observed 2.05 km.

**Table 8.** Summary of threshold-derived cloud-geometry diagnostics on CloudSat-verifiable support. RMSE values are conditional on jointly defined geometry, and geometry coverage denotes the fraction of observed-cloudy columns included in this comparison.

| Metric | Median [IQR] | Mean |
|---|---|---|
| CTH RMSE (conditional, km) | 1.38 [0.95, 2.09] | 1.66 |
| CBH RMSE (conditional, km) | 1.41 [0.84, 2.18] | 1.64 |
| Envelope thickness RMSE (conditional, km) | 2.05 [1.42, 2.92] | 2.30 |
| Total cloudy thickness RMSE (conditional, km) | 1.65 [1.21, 2.24] | 1.83 |
| Geometry coverage on observed-cloudy columns | 0.60 [0.39, 0.78] | 0.57 |

Figure 7 provides a column-level view of the observed-predicted relationship for the 767,086 columns in which both observed and predicted boundaries are defined. CTH shows the tighter relationship ($r = 0.86$). The conditional median remains close to $x = y$ over most of the central observed range, but departs below it for the highest cloud tops. This behaviour is consistent with the overall negative CTH bias of $-0.72\,\mathrm{km}$, indicating that the most elevated tops tend to be predicted too low.

CBH shows a weaker observed-predicted relationship than CTH ($r = 0.77$). The conditional median remains relatively close to $x = y$ over the lower part of the observed range but increasingly departs below it for higher observed cloud bases. This tendency is confirmed by the conditional bias, which changes from $+0.06\,\mathrm{km}$ over all jointly defined columns to $-0.40\,\mathrm{km}$

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


for $CBH_{\mathrm{obs}} \geq 6\,\mathrm{km}$, and becomes progressively more negative at higher observed bases. The near-zero overall bias therefore results from compensation between positive errors near the lower end of the observed CBH distribution and negative errors at higher altitudes. Its interpretation should also account for the effective lower truncation of the processed CloudSat cloud-occurrence profile, which contains almost no detected cloud below approximately 5 km.

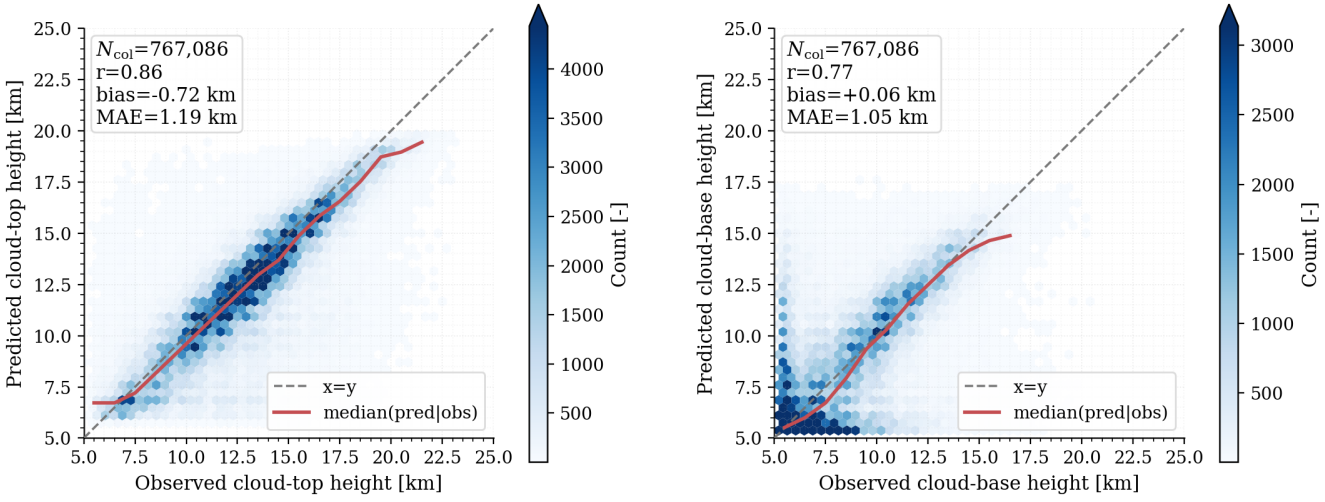


**Figure 7.** Scatterplot of column-level observed-predicted threshold-derived cloud-top height (CTH) and cloud-base height (CBH) on CloudSat-verifiable support. Hexagons indicate sample density, the dashed line represents x=y line, and the solid line the conditional median of the prediction given the observation.

A further limitation concerns the preservation of multilayer vertical structure. Using the threshold-derived layer count at column level, the confusion matrix between single-layer and multilayer columns is strongly asymmetric. Observed single-layer columns are almost always reconstructed as single-layer, with only 0.5% classified as multilayer. Observed multilayer columns, by contrast, are collapsed into single-layer outcomes: only 2.6% are reconstructed as multilayer, whereas 97.4% are mapped to a single-layer structure. The limitation is therefore not a generic ambiguity between the two classes, but a pronounced collapse of vertical complexity. Once the cloud field becomes internally structured, the reconstruction tends to preserve the bulk cloudy region while losing the separation between distinct layers.

Taken together, these results show that the threshold-derived boundaries retain useful information on the broad vertical location of cloud, but only on the subset of columns where cloud is detected in both observation and prediction. The envelope-thickness error is broadly consistent with the propagation of CTH and CBH errors. The clearest structural limitation is instead the failure to preserve multilayer organisation, with the large majority of observed multilayer columns reconstructed as single-layer.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


### 4.4 Scene-level vertical centroid summary

The previous subsection focused on threshold-derived cloud boundaries. However, CTH, CBH and thickness remain sensitive to edge effects, missed weak echoes and loss of internal layering. For this reason, it is useful to introduce a more aggregated descriptor of the vertical cloud distribution, namely the centroid height derived from the full vertical cloud-occurrence profile. This quantity does not describe the outer cloud envelope, but rather the altitude at which cloud occurrence is concentrated on average within a scene.

Table 9 reports the corresponding scene-level agreement. Out of 16 979 scenes that are cloudy in the observation, 16 243 are also cloudy in the prediction, corresponding to a cloudy-scene recall of 96%. On this common subset, the centroid shows a strong scene-level relationship, with $r = 0.89$, a bias of $-0.33\,\mathrm{km}$, an MAE of $0.72\,\mathrm{km}$, and an RMSE of $1.04\,\mathrm{km}$. These values should not be interpreted as direct evidence that the centroid is reconstructed more accurately than CTH, CBH, or thickness, because the centroid is a scene-level aggregate whereas the boundary diagnostics are evaluated column by column. As a simple reference, if the centroid is approximated by the midpoint between CTH and CBH and the two boundary errors are assumed to be independent and unbiased, the expected RMSE is $\frac{1}{2}\sqrt{1.38^2 + 1.41^2} = 0.99\,\mathrm{km}$. This is close to the observed centroid RMSE of $1.04\,\mathrm{km}$. The comparison is only heuristic, since the centroid used here is derived from the full scene-level cloud-occurrence profile rather than directly from CTH and CBH, and the boundary values are medians of scene-wise RMSEs. Nevertheless, it shows that the lower centroid error is broadly consistent with the expected effect of aggregation and should not be interpreted by itself as evidence of superior reconstruction skill.

**Table 9.** Scene-level agreement for the vertical centroid height derived from the full vertical cloud-occurrence profile. Agreement metrics are reported only for scenes classified as cloudy in both observation and prediction.

| $N_{\mathrm{obs\,cloudy}}$ | $N_{\mathrm{both}}$ | $\mathrm{Recall}_{\mathrm{cloudy}}$ | $r$ | Bias (km) | MAE (km) | RMSE (km) |
|---|---|---|---|---|---|---|
| 16,979 | 16,243 | 0.96 | 0.89 | -0.33 | 0.72 | 1.04 |

Figure 8 confirms this behaviour. The highest density of points is distributed close to the x=y line and the conditional median follows the diagonal rather well over most of the observed range. A systematic deviation appears mainly for the highest centroid values, for which the predicted centroid tends to remain slightly lower than the observed one. In other words, scenes whose cloud occurrence is concentrated higher in the column are reconstructed with a moderate downward shift, but without losing the overall vertical ordering of the scenes.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


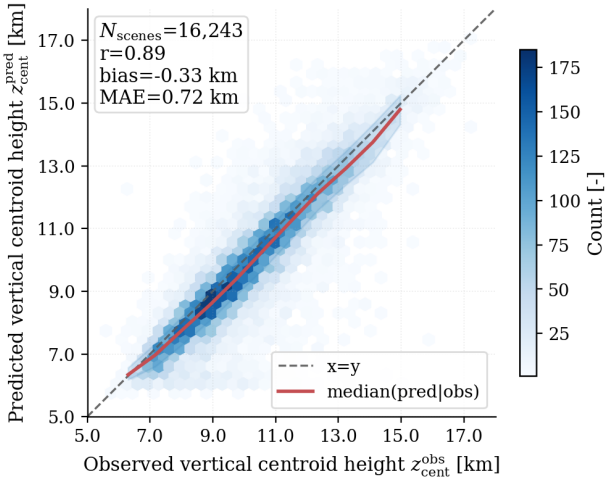


**Figure 8.** Observed vs. predicted scene-level vertical centroid height derived from the full vertical cloud-occurrence profile. The dashed line is x=y and the solid line the conditional median of the prediction given the observation.

The centroid should therefore be interpreted as a complementary aggregate descriptor rather than as evidence of superior reconstruction of vertical geometry. Its high cloudy-scene recall and strong observed-predicted relationship show that the model retains useful information on the altitude at which cloud occurrence is concentrated within a scene. The close agreement between the observed RMSE and the simple error-propagation benchmark, however, indicates that its lower error magnitude is largely consistent with the expected effect of aggregation.

### 4.5 Performance across cloud regimes, seasons, and latitude bands

The previous subsections described the average behaviour of the reconstruction. The next question is whether this behaviour is uniform across scenes, or whether some cloud regimes are reconstructed more reliably than others. This point is important, because low-cloud-cover scenes, broad stratiform scenes and scenes containing stronger echoes do not present the same level of difficulty.

Regimes are defined using two observed CloudSat-based quantities. The observed cloudy-column fraction $CF_{\mathrm{col,obs}}$ is divided into low ($CF_{\mathrm{col,obs}}$<0.2), medium ($0.2 \leq CF_{\mathrm{col,obs}} \leq 0.6$), and high ($CF_{\mathrm{col,obs}}$>0.6) cloud-cover regimes. Scene intensity is divided into weak and strong cases according to the maximum observed reflectivity, using 5 dBZ as a practical threshold.

Figure 9 shows the distribution of the scene-median error on cloudy voxels after stratification by observed column cloud cover and scene intensity. The median error remains negative in all six regimes, confirming that the tendency to underestimate reflectivity on cloudy voxels is systematic. At the same time, the magnitude of this underestimation is not the same across regimes. Low-($CF_{\mathrm{col}}$) scenes show the most negative errors, while high-($CF_{\mathrm{col}}$) scenes are less degraded. This indicates that

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


sparse or broken cloudy scenes are harder to reconstruct than broad, spatially extended ones. The effect of intensity is visible, but the clearest separation is associated with cloudy-column fraction.

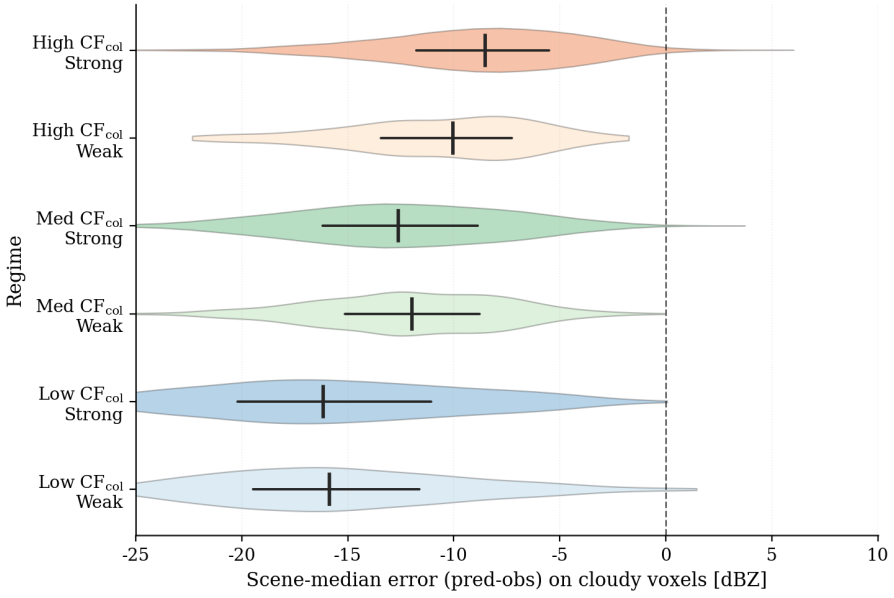


**Figure 9.** Distribution of scene-median reflectivity error on cloud-conditioned voxels, stratified by observed cloudy-column fraction and scene intensity. Boxes indicate the interquartile range and central lines the median.

Table 10 shows how this regime dependence propagates from reflectivity error to detection and geometry diagnostics. The most interpretable regime is the high-$CF_{\mathrm{col}}$, strong regime. It contains 7776 scenes, nearly half of the stratified evaluation set, and combines the best column-detection scores, with POD=0.73 and CSI=0.72, with moderate conditional geometry errors, 1.39 km for CTH and 1.41 km for CBH. The cloud-conditioned RMSE remains high, (12.9 dBZ), so reflectivity amplitude is still underestimated, but the broad cloudy structure is more consistently detected.

The least reliable cases are the low-$CF_{\mathrm{col}}$ regimes. In the low-$CF_{\mathrm{col}}$, weak regime, (POD=0.00) and (CSI=0.00), meaning that the observed cloudy columns are essentially not retained after thresholding. In the low-$CF_{\mathrm{col}}$, strong regime, some cloudy columns are detected (POD=0.45, CSI=0.43), but the cloud-conditioned RMSE is the largest of all regimes, reaching (18.7 dBZ). These results indicate that sparse scenes are problematic in two different ways: weak sparse clouds can disappear after thresholding, while stronger sparse structures may be detected but with large amplitude errors.

The low-$CF_{\mathrm{col}}$ regimes are relatively rare in this evaluation set, with 315 weak scenes and 891 strong scenes, corresponding together to about (7.1 %) of the stratified samples. Nevertheless, they are important for interpretation because they identify the main failure domain of the reconstruction.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


**Table 10.** Regime-stratified median performance for cloud-conditioned reflectivity error, conditional geometry error, and column-detection skill. Regimes are defined by observed column cloud cover and scene intensity.

| $CF_{col}$ | Intensity | N | $RMSE_{cld}$ | $CTH_{RMSE}$ | $CBH_{RMSE}$ | POD | FAR | CSI |
|---|---|---|---|---|---|---|---|---|
| Low | Weak | 315 | 16.8 | 0.89 | 1.05 | 0.00 | 0.00 | 0.00 |
| Low | Strong | 891 | 18.7 | 1.52 | 1.35 | 0.45 | 0.00 | 0.43 |
| Med | Weak | 657 | 13.7 | 0.96 | 1.12 | 0.27 | 0.00 | 0.26 |
| Med | Strong | 7171 | 16.1 | 1.42 | 1.45 | 0.53 | 0.00 | 0.50 |
| High | Weak | 252 | 12.5 | 0.96 | 1.35 | 0.39 | 0.00 | 0.38 |
| High | Strong | 7776 | 12.9 | 1.39 | 1.41 | 0.73 | 0.00 | 0.72 |

This regime-wise interpretation nevertheless requires caution for the geometry metrics. Because CTH and CBH errors are conditional on jointly detected cloudy columns, they cannot be interpreted independently of the detection skill. For instance, the low-$CF_{col}$, weak regime shows a relatively small median CTH RMSE, but this does not imply that the geometry is well reconstructed in that regime, since the corresponding POD and CSI are both zero. Rather, it indicates that geometry can only be compared on a very small residual subset of cases. Regime-wise geometry scores are therefore informative mainly when the corresponding regime also retains a non-negligible fraction of detected cloudy columns.

This regime hierarchy is also consistent with the additional stratified examples provided in the Supplement (Figs. S1–S200). In low-$CF_{col}$ scenes, the reconstruction often misses isolated cloudy structures or strongly attenuates them. In medium-$CF_{col}$ scenes, the dominant cloudy region is usually retained, but with reduced internal contrast. In high-$CF_{col}$ scenes, the broad vertical placement is generally preserved, although reflectivity gradients and multilayer complexity are smoothed.

The seasonal and latitudinal dependence is shown in Fig. 10. The seasonal effect is visible in both panels. In all latitude bands, JJA corresponds to the highest values of $RMSE_{cld}$, reaching 15.3-15.4 dBZ, and also to the highest CTH errors, especially in the tropics and at high latitudes. Outside JJA, the values remain lower and more homogeneous. The best conditions are found in mid-latitude MAM for CTH and in high-latitude MAM for cloud-conditioned reflectivity error.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


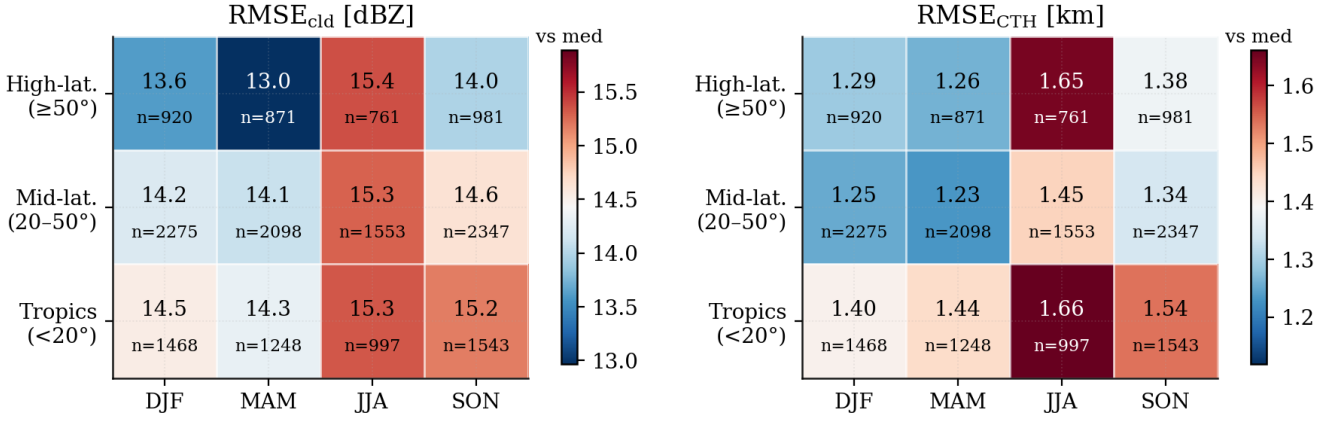


**Figure 10.** Season × latitude-band summaries of cloud-conditioned reflectivity RMSE and conditional CTH RMSE. Cell values denote the regime median and the annotation below each value indicates the number of samples in the corresponding bin.

These seasonal and latitudinal contrasts are present, but they are weaker than the regime dependence. The main performance differences are associated with cloud-scene structure, especially cloudy-column fraction, rather than with latitude or season alone. Seasonal and latitudinal effects should therefore be interpreted partly as changes in the composition of cloud regimes within each bin.

Overall, the stratified analysis shows that the reconstruction remains informative over a substantial part of the evaluation set, but its reliability is not uniform. The most favorable conditions are broad and spatially extended cloud regimes, whereas sparse scenes and scenes with stronger local structures remain the main source of degradation. This behaviour is also qualitatively consistent with experience from SEVIRI-based solar-radiation retrievals, where scattered or broken cloud fields are known to be particularly challenging because of their strong spatial heterogeneity, rapid temporal variability, and the limited spatial resolution of the geostationary observations (Qu et al., 2017). This does not imply that performance is spatially or temporally invariant, nor that it remains unchanged at large SEVIRI viewing angles. Rather, it indicates that, in the present evaluation, the main observed source of variability is the structure of the cloud scene itself.

## 5 Discussion, limitations, and future perspectives

The present results suggest that the main value of SEVIRI-CloudSat tomography lies less in reconstructing a radar-equivalent 3D cloud field than in recovering a smoothed, support-dependent representation of bulk vertical cloud organisation. This interpretation is consistent across the diagnostics. The global reflectivity benchmark remains moderate because it mixes cloudy and cloud-free support, whereas the cloud-conditioned evaluation reveals the dominant failure mode, namely the attenuation of moderate and strong echoes. After thresholding, this attenuation affects the derived products unevenly: bulk cloudy occupancy remains overall stable, but cloudy-column occurrence and threshold-derived geometry degrade more strongly. At the same time, the vertical diagnostics show a stable scene-level cloud centroid allocation, but a very low multilayer structure detection.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


The reconstructed field should therefore be interpreted as a CloudSat-like surrogate of where cloud occurrence is vertically concentrated than as a faithful retrieval of full internal cloud structure.

The present analysis also highlight several important limitations. First, the performance evaluation is support-restricted by construction: quantitative verification is only possible where CloudSat provides an observable reference, so the skill of the full extrapolated 3D field away from the verifiable curtain cannot be established. Second, the target itself is not a completely reliable cloud truth. CloudSat is sun-synchronous and samples only a subset of cloud conditions, while the present framework is restricted to CloudSat-supported, daytime scenes within the analysed vertical range. The reconstructions should therefore be interpreted as pseudo-CloudSat fields at CloudSat overpass times, not as an unbiased 3D cloud truth valid at any local time. This point is important because some of the most attractive potential uses of geostationary tomography implicitly suggest continuous volumetric cloud information, whereas the present validation only supports a more limited interpretation. Third, several diagnostics are conditional and should be read as such: low cloud-top or cloud-base errors are informative only for the subset of columns in which geometry is retained in both observation and reconstruction. More broadly, part of the degradation is likely intrinsic to the problem itself, since strong reflectivity echoes and multilayer separation are among the aspects least directly constrained by passive SEVIRI radiances.

These limitations also point to more specific future developments. A first direction is to provide the network with additional physical context, for instance through meteorological conditioning from reanalysis fields such as temperature, humidity, or vertical motion, which may help constrain quantities that are only weakly encoded in instantaneous SEVIRI radiances, including layer separation, cloud vertical extent, and the altitude of phase transitions. A second direction is architectural rather than purely incremental: the present results suggest that a single continuous reflectivity output is not necessarily the most informative formulation for all applications. Multi-task or multi-head networks, jointly predicting reflectivity together with occurrence, cloud top and base height, or bulk vertical allocation, could help align the learning problem more directly with the descriptors that remain informative in the present evaluation. Additional temporal context from successive SEVIRI images is another interesting extension. Finally, off-support analyses remain important, not as formal validation, but to characterize what type of three-dimensional cloud field the model tends to generate beyond the CloudSat curtain. In this perspective, future progress will likely come not only from improving voxel-wise fidelity, but from combining stronger physical conditioning with output formulations better matched to the intended downstream products.

## 6 Conclusions

Using an all-cloud SEVIRI-CloudSat reflectivity baseline, this study assessed cloud tomography on an independent 2019 daytime year (17,110 matched scenes) with an extended evaluation framework covering reflectivity fidelity, threshold-derived occurrence and geometry, bulk vertical allocation, and regime dependence. The results show that the reconstruction remains informative, but not uniformly across cloud scenes. Threshold-derived geometry remains informative only conditionally, with median cloud-top and cloud-base errors of 1.38 and 1.41 km on the subset of columns that remain geometrically comparable, corresponding to about 60% of observed-cloudy columns. The envelope-thickness RMSE of 2.05 km is close to the expected

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


simple propagation of the CTH and CBH errors and therefore does not indicate a clear additional loss of skill. Similarly, the scene-level vertical centroid provides a useful aggregate summary, with 96% cloudy-scene recall, $r = 0.89$, and an RMSE of 1.04 km; however, this value is also close to the expected error propagation. The clearest structural limitation is instead the preservation of multilayer organisation: only 2.6% of observed multilayer columns are reconstructed as multilayer, while 97.4% collapse to single-layer outcomes.

A second conclusion is that the main limitation is a smoothing of the reconstructed field rather than a complete loss of cloud signal. This effect weakens stronger echoes and propagates most strongly in sparse scenes, whereas broad and spatially extended cloud regimes remain the most reliable. Among strong scenes, for example, the cloud-conditioned RMSE decreases from 18.7 dBZ in low-cloud-cover cases to 12.9 dBZ in high-cloud-cover cases, while POD increases from 0.45 to 0.73. Seasonal and latitudinal contrasts are visible, but remain secondary compared with regime dependence. SEVIRI-CloudSat tomography is therefore best interpreted as a pseudo-CloudSat product at CloudSat overpass times, most credible for vertically informed cloud-occurrence diagnostics and bulk cloud allocation, while its structural limits must remain explicit in downstream use.

*Code and data availability.* The original Res-UNet model framework and SEVIRI-CloudSat data-matching scheme on which the present implementation is based are publicly available in the Zenodo archive of Brüning (2023) at https://doi.org/10.5281/zenodo.8238110. The scripts developed specifically for the analyses and figures presented in this study are available from the corresponding author upon reasonable request. The Meteosat Second Generation SEVIRI High Rate Level 1.5 Image Data used for 2016, 2017, and 2019 are available from the EUMETSAT Data Store under product ID EO:EUM:DAT:MSG:HRSEVIRI at https://data.eumetsat.int/data/map/EO:EUM:DAT:MSG:HRSEVIRI (last access: 21 July 2026). The CloudSat 2B-GEOPROF R05 data used for the same periods are available from the CloudSat Data Processing Center at https://www.cloudsat.cira.colostate.edu/data-products/2b-geoprof (last access: 21 July 2026).

*Author contributions.* AB: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing – original draft, and writing – review and editing. YMSD: conceptualization, supervision, and writing – review and editing. BG: software and data curation. GC: conceptualization, writing – original draft, and writing – review and editing. All authors participated in regular project discussions and approved the submitted version of the manuscript.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* The authors acknowledge EUMETSAT for providing access to the Meteosat Second Generation SEVIRI Level 1.5 data and the CloudSat Data Processing Center for providing access to the 2B-GEOPROF data. The computations were performed using computing resources provided by MINES Paris – PSL.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


## References

Aubry, C., Delanoë, J., Groß, S., Ewald, F., Tridon, F., Jourdan, O., and Mioche, G.: Lidar-radar synergistic method to retrieve ice, supercooled water and mixed-phase cloud properties, Atmospheric Measurement Techniques, 17, 3863–3881, https://doi.org/10.5194/amt-17-3863-2024, 2024.

Benas, N., Finkensieper, S., Stengel, M., Zadelhoff, G. J. V., Hanschmann, T., Hollmann, R., and Meirink, J. F.: The MSG-SEVIRI-based cloud property data record CLAAS-2, Earth System Science Data, 9, 415–434, https://doi.org/10.5194/essd-9-415-2017, 2017.

Brüning, S.: AI-derived 3D cloud tomography, https://doi.org/10.5281/zenodo.8238110, code, 2023.

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

Ceccaldi, M., Delanoë, J., Hogan, R. J., Pounder, N. L., Protat, A., and Pelon, J.: From CloudSat-CALIPSO to EarthCare: Evolution of the DARDAR cloud classification and its comparison to airborne radar-lidar observations, Journal of Geophysical Research Atmospheres, 118, 7962–7981, https://doi.org/10.1002/jgrd.50579, 2013.

Deng, M., Mace, G. G., Wang, Z., and Berry, E.: Cloudsat 2C-ICE product update with a new Ze parameterization in lidar-only region, Journal of Geophysical Research, 120, 12,198–12,208, https://doi.org/10.1002/2015JD023600, 2015.

Garnier, A., Pelon, J., Pascal, N., Vaughan, M. A., Dubuisson, P., Yang, P., and Mitchell, D. L.: Version 4 CALIPSO Imaging Infrared Radiometer ice and liquid water cloud microphysical properties – Part II: Results over oceans, Atmospheric Measurement Techniques, 14, 3277–3299, https://doi.org/10.5194/amt-14-3277-2021, 2021.

Girtsou, S., Salas-Porras, E. D., Freischem, L., Massant, J., Bintsi, K.-M., Castiglione, G., Jones, W., Eisinger, M., Johnson, E., and Jungbluth, A.: 3D Cloud reconstruction through geospatially-aware Masked Autoencoders, http://arxiv.org/abs/2501.02035, 2025.

IPCC: The Earth's Energy Budget, Climate Feedbacks and Climate Sensitivity, p. 923–1054, Cambridge University Press, https://doi.org/10.1017/9781009157896.009, 2023.

Jeggle, K., Czerkawski, M., Serva, F., Saux, B. L., Neubauer, D., and Lohmann, U.: IceCloudNet: 3D reconstruction of cloud ice from Meteosat SEVIRI, pp. 1–39, http://arxiv.org/abs/2410.04135, 2024.

Karlsson, K. G. and Devasthale, A.: Inter-comparison and evaluation of the four longest satellite-derived cloud climate data records: CLARA-A2, ESA Cloud CCI V3, ISCCP-HGM, and PATMOS-x, Remote Sensing, 10, https://doi.org/10.3390/rs10101567, 2018.

Kim, B.-R., Kim, G., Cho, M., Choi, Y.-S., and Kim, J.: First results of cloud retrieval from the Geostationary Environmental Monitoring Spectrometer, Atmospheric Measurement Techniques, 17, 453–470, https://doi.org/10.5194/amt-17-453-2024, 2024.

Kotarba, A. Z. and Wojciechowska, I.: Satellite-based detection of deep-convective clouds: the sensitivity of infrared methods and implications for cloud climatology, Atmospheric Measurement Techniques, 18, 2721–2738, https://doi.org/10.5194/amt-18-2721-2025, 2025.

Lamer, K., Kollias, P., Battaglia, A., and Preval, S.: Mind the gap - Part 1: Accurately locating warm marine boundary layer clouds and precipitation using spaceborne radars, Atmospheric Measurement Techniques, 13, 2363–2379, https://doi.org/10.5194/amt-13-2363-2020, 2020.

Leinonen, J., Guillaume, A., and Yuan, T.: Reconstruction of Cloud Vertical Structure With a Generative Adversarial Network, Geophysical Research Letters, 46, 7035–7044, https://doi.org/10.1029/2019GL082532, 2019.

Lin, H., Li, J., Min, M., Zhang, F., Wang, K., and Wu, Q.: CLANN: Cloud amount neural network for estimating 3D cloud from geostationary satellite imager, Remote Sensing of Environment, 318, 114 600, https://doi.org/10.1016/j.rse.2025.114600, 2025.

https://doi.org/10.5194/egusphere-2026-4441
Preprint. Discussion started: 28 July 2026


Marchand, R. T., Mace, G. G., Ackerman, T. P., and Stephens, G. L.: Hydrometeor Detection using CloudSat – an Earth-Orbiting 94-GHz Cloud Radar, Journal of Atmospheric and Oceanic Technology, 25, 519–533, https://doi.org/10.1175/2007JTECHA1006.1, 2008.

Massie, S. T., Cronk, H., Merrelli, A., O'Dell, C., Schmidt, K. S., Chen, H., and Baker, D.: Analysis of 3D cloud effects in OCO-2 XCO2 retrievals, Atmospheric Measurement Techniques, 14, 1475–1499, https://doi.org/10.5194/amt-14-1475-2021, 2021.

Mcgarragh, G. R., Poulsen, C. A., Thomas, G. E., Povey, A. C., Sus, O., Stapelberg, S., Schlundt, C., Proud, S., Christensen, M. W., Stengel, M., Hollmann, R., and Grainger, R. G.: The Community Cloud retrieval for CLimate (CC4CL)-Part 2: The optimal estimation approach, Atmospheric Measurement Techniques, 11, 3397–3431, https://doi.org/10.5194/amt-11-3397-2018, 2018.

Qu, Z., Oumbe, A., Blanc, P., Espinar, B., Gesell, G., Gschwind, B., Klüser, L., Lefèvre, M., Saboret, L., Schroedter-Homscheidt, M., and Wald, L.: Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat?4 method, Meteorologische Zeitschrift, 26, 33–57, https://doi.org/10.1127/metz/2016/0781, 2017.

Sayer, A. M., Cairns, B., Knobelspiesse, K. D., Lelli, L., Rajapakshe, C., Giangrande, S. E., Thomas, G. E., and Zhang, D.: Evaluation of cloud height, optical thickness, and phase retrievals from the CHROMA algorithm applied to Sentinel-3 OLCI data, Atmospheric Measurement Techniques, 18, 6681–6703, https://doi.org/10.5194/amt-18-6681-2025, 2025.

Schirmacher, I., Kollias, P., Lamer, K., Mech, M., Pfitzenmaier, L., Wendisch, M., and Crewell, S.: Assessing Arctic low-level clouds and precipitation from above - a radar perspective, Atmospheric Measurement Techniques, 16, 4081–4100, https://doi.org/10.5194/amt-16-4081-2023, 2023.

Schmetz, J., Pili, P., Tjemkes, S., Just, D., Kerkmann, J., Rota, S., and Ratier, A.: AN INTRODUCTION TO METEOSAT SECOND GENERATION (MSG), Bulletin of the American Meteorological Society, 83, 977 – 992, https://journals.ametsoc.org/view/journals/bams/83/7/1520-0477_2002_083_0977_aitmsg_2_3_co_2.xml, 2002.

Schroedter-Homscheidt, M., Azam, F., Betcke, J., Hanrieder, N., Lefèvre, M., Saboret, L., and Saint-Drenan, Y.: Surface solar irradiation retrieval from MSG/SEVIRI based on APOLLO Next Generation and HELIOSAT4 methods, Meteorologische Zeitschrift, 31, 455–476, https://doi.org/10.1127/metz/2022/1132, 2022.

Schulte, R. M., Lebsock, M. D., and Haynes, J. M.: What CloudSat cannot see: liquid water content profiles inferred from MODIS and CALIOP observations, Atmospheric Measurement Techniques, 16, 3531–3546, https://doi.org/10.5194/amt-16-3531-2023, 2023.

Spradlin, C. S., Caraballo-Vega, J. A., Li, J., Carroll, M. L., Gong, J., and Montesano, P. M.: SatVision-TOA: A Geospatial Foundation Model for Coarse-Resolution All-Sky Remote Sensing Imagery, pp. 1–19, http://arxiv.org/abs/2411.17000, 2024.

Wiltink, J. I., Deneke, H., van Heerwaarden, C. C., and Meirink, J. F.: Evaluating parallax and shadow correction methods for global horizontal irradiance retrievals from Meteosat SEVIRI, Atmospheric Measurement Techniques, 18, 3917–3936, https://doi.org/10.5194/amt-18-3917-2025, 2025a.

Wiltink, J. I., Deneke, H., van Heerwaarden, C. C., and Meirink, J. F.: Evaluating parallax and shadow correction methods for global horizontal irradiance retrievals from Meteosat SEVIRI, EGUsphere, 2025, 1–29, https://doi.org/10.5194/egusphere-2024-4139, 2025b.

Yu, H., Emde, C., Kylling, A., Veihelmann, B., Mayer, B., Stebel, K., and Roozendael, M. V.: Impact of 3D cloud structures on the atmospheric trace gas products from UV-Vis sounders - Part 2: Impact on NO2retrieval and mitigation strategies, Atmospheric Measurement Techniques, 15, 5743–5768, https://doi.org/10.5194/amt-15-5743-2022, 2022.