https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Classification of Cloud Species and Varieties in Ground-Based Images Using Artificial Neural Networks

Szymon Kopeć1, Grzegorz Duniec1, Bogdan Bochenek1

1Institute of Meteorology and Water Management – National Research Institute, Warsaw, Poland

Correspondence to: Bogdan Bochenek (bogdan.bochenek@imgw.pl)

ABSTRACT

Accurate classification of cloud species and varieties is essential for meteorological observation and climate research, yet automated systems utilizing artificial neural networks to address the full World Meteorological Organization (WMO) taxonomy—including 14 species and 9 varieties—remain largely unexplored. Building upon our previous work on cloud genera classification, this study extends that foundation to address the finer taxonomic levels of species and varieties. We present an artificial neural network-based framework evaluated on ground-based images collected across Poland by over 50 experts from the PSHM IMGW-PIB network (2021–2026). Both Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) architectures were evaluated. Multi-class species classification achieved F1 scores ranging from 49% to 79%. Notably, applying a hierarchical approach utilizing genera-specific models—leveraging our earlier findings—yielded improvements exceeding 19 percentage points for certain species. Classification accuracy followed a logarithmic trend with dataset size, where Transformer-based models outperformed CNNs for larger training datasets, while CNNs exhibited an advantage with smaller sample sizes. These findings demonstrate that neural network architectures can effectively classify cloud species and varieties, and that genera-level model stratification significantly enhances classification capability, suggesting potential for integration into operational meteorological workflows.

INTRODUCTION

The World Meteorological Organization (WMO) has established a comprehensive hierarchical taxonomy for cloud classification, comprising ten principal genera subdivided into 14 species (reflecting internal morphology such as fibrous, cumuliform, or lenticular structures) and 9 varieties (describing specific patterns or arrangements) (WMO, 1987),(International cloud atlas, 1956). This systematic classification exhibits considerable complexity, as certain species and varieties may be associated with multiple genera, individual formations may simultaneously display characteristics of several varieties, and supplementary features and accessory clouds introduce additional classificatory layers (Bluestein, 2024; Hamblyn, 2023; Retallack, 1970). The precise characterization afforded by this taxonomic framework is integral to meteorological

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

observation, numerical weather prediction, climate modelling, and advancing understanding of atmospheric processes (Kuma et al., 2023; Rahman, 2023).

Cloud observation methodologies span a wide range of perspectives, including remote sensing from satellites, in situ measurements from aircraft, observations from elevated terrestrial locations, and ground-based monitoring as detailed in this study. Traditionally, the accurate identification of cloud species and varieties has relied on skilled human observers (Kopeć et al., 2024) However, presented approach faces several challenges, including subjectivity in interpretation, inconsistency between observers, and the limited capacity to process large volumes of data continuously (Kalkan et al., 2022). These limitations highlight the potential benefits of developing automated classification systems, which could supplement human expertise while addressing the need for consistent, large-scale cloud observations in modern meteorological operations.

The visual properties and aesthetic qualities of clouds are governed by the scattering of visible light by cloud particles, particularly along the cloud boundaries, with differential scattering efficiencies for various wavelengths of light (Bluestein, 2024) This complexity in cloud appearance further underscores the challenges faced by both human observers and automated systems in accurate classification.

The process of cloud formation is a multifaceted phenomenon, involving not only the fundamental principles of air saturation and adiabatic cooling but also more intricate microphysical processes such as the presence and activation of cloud condensation nuclei (CCN)(Albrecht, 1989; Twomey, 1977) These processes directly influence the physical structure and optical properties of clouds, determining their appearance, including shape, texture, and colour variations (Bohren, 1998) For instance, the size distribution and concentration of water droplets or ice crystals, which are governed by CCN dynamics, significantly affect how clouds scatter and absorb light, thereby impacting their visual characteristics (Twomey, 1977). This intricate relationship between cloud microphysics and macroscopic appearance underscores the complexity in cloud dynamics and the critical importance of accurate and standardized classification methodologies in meteorological research and operational forecasting.

The rigorous categorization of clouds serves a dual purpose: it enables precise communication among meteorologists on a global scale and provides essential quantitative and qualitative data for advancing our comprehension of atmospheric dynamics (Errico et al., 2007). This, in turn, contributes to the ongoing refinement and enhancement of weather forecasting models, ultimately improving prediction accuracy across various temporal and spatial scales (Winkler, 2025).

Our research overcomes the bottleneck of efficiently processing vast amounts of cloud imagery by exploring the application of Artificial Neural Networks (ANNs) to automate the classification of cloud species and varieties in ground-based images. While human experts excel at interpreting complex 3D cloud structures, they are limited by the time-consuming nature of manual classification and the impracticality of analysing large datasets (Wacker et al., 2015). ANNs, on the other hand, can rapidly process enormous volumes of images, potentially surpassing human capabilities in terms of efficiency and consistency (Taravat et al., 2015)

Traditional clouds classification

Traditional methods of cloud classification rely on visual observations conducted by trained meteorological observers. The process includes assessing cloud cover (measured in oktas),

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

identifying cloud genera, species, and varieties based on their shape, structure, brightness, and color, and noting any supplementary features or accessory clouds. Observers use tools such as cloud atlases, refractors, pilot balloons, and benchmarks to support identification and estimate cloud base height. Observations are carried out both during the day and night, with moonlight, starlight, or artificial light aiding visibility under low-light conditions (Kopeć et al., 2024).

Alternative clouds classification

Alternative methods for cloud classification encompass a range of advanced observational and analytical techniques. Automated image recognition systems, based on machine learning algorithms trained on curated datasets of cloud imagery (Lusi et al., 2024) , offer support in cloud type identification. Remote sensing approaches—including satellite observations, radar, and lidar technologies such as ceilometers—enable precise assessment of cloud structure, physical properties, and vertical distribution. Ceilometers, as active lidar instruments, provide continuous measurements of cloud base height, which can be integrated into classification schemes. Furthermore, spectral analysis employing multispectral and hyperspectral sensors facilitates cloud differentiation through distinct radiative signatures across various wavelengths. (Liu et al., 2011)

Comprehensive study

The study by (Nie et al., 2024) provides a comprehensive review of available open datasets containing sky and cloud images that can be used in research on solar energy forecasting and related issues, such as cloud segmentation, classification, and their motion prediction. The authors identified 72 publicly available clouds datasets from around the world that meet the requirements for application in deep learning methods. Detailed information about various aspects of these datasets was collected, creating an extensive collective database. A ranking system based on multiple criteria was also developed, allowing for the evaluation of each dataset based on eight key dimensions that significantly impact data usability. In (Li et al., 2023) contributed to cloud genera classification using a 10-genera dataset, also in (Guzel et al., 2024). However, like most studies, it lacks comprehensive classification of cloud species within each genera. (Shi et al., 2024) While some research classifies certain cloud species, this is not consistent across all genera. Notably, datasets representing cloud varieties are absent. This gap underscores the need for more detailed datasets including genera, species, and varieties to improve our understanding of cloud formations and their meteorological significance. (Rosenberger et al. 2025) present one of the first automated classification system for all 30 WMO-defined cloud classes, including species and varieties beyond the typical 10 genera. Using an ensemble of residual neural networks trained on ground-based RGB images from Vienna, they achieved 83% precision despite extreme class imbalances, with some cloud types appearing only once in three years. While the model excelled at identifying rare clouds through aggressive data augmentation, it struggled with visually similar categories like Stratocumulus at different heights, reflecting challenges that even human observers face in operational cloud classification.

DATA

The dataset, used for supervised multiclass classification, consists of ground-based cloud images collected across Poland by over 50 experts from the PSHM IMGW-PIB network. Due to the inherent class imbalance present in the dataset, undersampling was applied to ensure a more balanced representation of cloud genera and domains during model training. The detailed

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

distribution of training samples across individual classes, after accounting for this imbalance, is provided in Appendix B.

Data collection process

Photographs were taken manually using various smartphone cameras, resulting in a wide range of image quality due to differences in resolution, aspect ratio, and sensor types. Images were captured as needed—mainly during daytime—without automation or fixed intervals. Since 2021, approximately 400 images have been collected daily. Each image was labelled with one cloud genera, species, and varieties (or noted for the absence of such features), and annotated with additional information, including visible objects (e.g., trees, buildings) and weather conditions (e.g., rain, fog).

The time series presented in **Figure 1** reveals pronounced seasonal stratification in image of cloud genera composition over Poland during the 4.5-year observation period (July 2021–January 2026) for Cumulus and Stratocumulus genera.

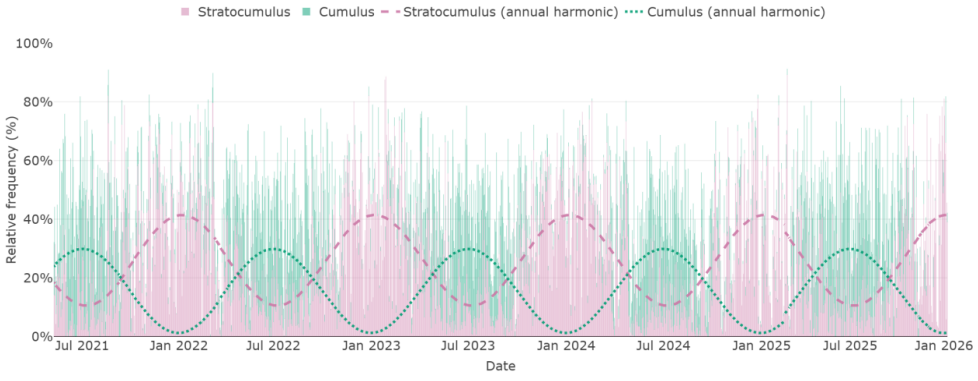


Figure 1. Daily relative frequency of Stratocumulus (pink) and Cumulus (green) clouds. Each bar represents about 400 ground-based images one day. Dashed lines show fitted annual harmonic cycles for each cloud type.

The collected statistics demonstrate a clear bimodal annual pattern characterized by winter dominance of stratiform genera primarily Stratocumulus, Stratus which collectively occupy >50% of sky coverage during November–March periods. Conversely, the warm season (May–September) exhibits a marked shift toward convective cloud types Cumulus, Cumulonimbus with stratiform coverage diminishing to <30% and increased clear-sky frequency with high-level clouds such as Cirrus. Additional statistics, including moving averages, are presented in Appendix C.

Data processing and sorting

The cloud image dataset was hierarchically annotated at the genera level and subsequently at the species and variety levels. Each image was assigned exactly one label at each taxonomic level (genera, species, and varieties). The dataset was partitioned using a 75/25 train–validation split via uniform random assignment, and an independent test set was held out for final evaluation. All images were resized to 384 × 384 pixels while preserving the three RGB channels to match the neural network input requirements. For transfer learning, images were normalized using

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

ImageNet statistics (mean: 0.485, 0.456, 0.406; standard deviation: 0.229, 0.224, 0.225) to ensure compatibility with the pre-trained model weights.

Data requirements and sample size estimation

To determine the minimum required sample size, we employed a precision-based(confidence level, standard deviation, desired precision) calculation. under the following assumptions:

a) features extracted from cloud images through neural network architectures (e.g., convolutional, transformer-based, or hybrid models) can be approximated as having finite variance,
b) model parameter estimation errors follow an approximately normal distribution due to central limit theorem (CLT) effects during gradient-based optimization. These assumptions represent practical simplifications, as comprehensive distributional analysis of extracted features extends beyond the scope of this study. Such an analysis would involve high-dimensional statistical testing, modelling inter-feature dependencies, and validating assumptions across different architectures. Sample size was calculated using the formula:

$$n = \left(Z \cdot \frac{\sigma}{E}\right)^2 \quad (1)$$

where:

- σ = 0.05 (standard deviation of Cohen's κ estimated from preliminary tests using 4-fold stratified cross-validation; see Sect. 4.4)E = 0.02 (desired precision of ±2%)
- z = 1.96 (confidence level 95%)

A safety factor of 10 was applied to the calculated sample size to account for potential violations of distributional assumptions and ensure robust model performance.

Figure 2 shows sample size requirements for comparing two Cohen's kappa values (σ=0.02 vs σ=0.05) across different confidence levels (95-99%) and effect sizes (0.005-0.02). Sample sizes range from ~100 to 5,000 participants per class in model, with larger samples required for higher confidence levels, smaller effect sizes, and σ=0.05 compared to σ=0.02.

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

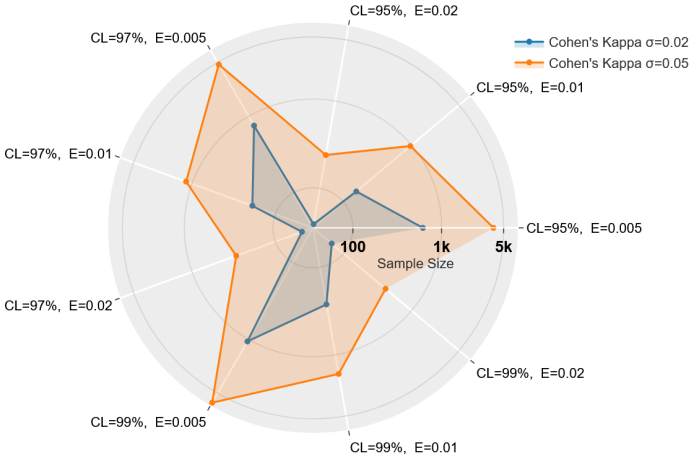


Figure 2. Sample sze by confidence level and effect size for two cohen's kappa values

METHODS

The deep learning methodology utilized for supervised classification involves several critical and interdependent steps. Initially, meticulous dataset preparation is performed, as this phase significantly impacts the overall effectiveness and accuracy of the resulting classification model. The dataset is systematically partitioned into distinct training, validation, and test subsets (Chollet, 2021). To augment dataset diversity and volume, sophisticated augmentation methods, such as AugMix and RandAugment (Cubuk et al., 2019; Hendrycks et al., 2020), are employed.

Deep learning architectures for supervised classification

The selection of an appropriate neural network architecture is critical for image classification tasks. Convolutional Neural Networks (CNNs) and Transformer-based Neural Networks (TNNs) are predominantly employed, typically in conjunction with Fully Connected (FC) layers acting as classifiers. These models often benefit from backbone initialization with pre-trained weights, commonly derived from large-scale benchmark datasets such as ImageNet. This approach supports efficient transfer learning and fine-tuning procedures.

In our study, we primarily utilize the Vision Transformer (ViT-B/16) architecture (Dosovitskiy et al., 2021). The ViT naming convention follows a standard pattern: the first letter denotes the model size — B (Base), L (Large), or H (Huge) — while the number refers to the patch size, e.g., 16×16 or 14×14 pixels. Thus, ViT-B/16 indicates a base-sized model operating on 16×16 pixels patches. The patch size governs the spatial granularity of the input, as self-attention is computed across all resulting patch tokens. Smaller patches (e.g., 14×14 in ViT-H/14) produce more tokens, offering finer spatial resolution but at higher computational cost.

Model training procedure

The training process consists of iterative batch-wise data processing, incorporating forward propagation, backward propagation and systematic parameter updating through optimization algorithms. The model was trained using the categorical cross entropy loss function, which is well

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

suited for multiclass image classification tasks (Goodfellow et al., 2016). The training employed a cosine annealing learning rate schedule, which gradually reduces the learning rate following a cosine decay pattern to promote stable convergence and improved generalization.

Continuous monitoring of the training progression is essential, with critical metrics, including loss values and accuracy measures, being routinely evaluated. Moreover, to mitigate the risk of overfitting, techniques such as regularization, dropout, and layer or batch normalization are rigorously implemented.

Figure 3 provides a simple overview schema of the proposed workflow for data preparation, model training, and optimization in supervised image classification.

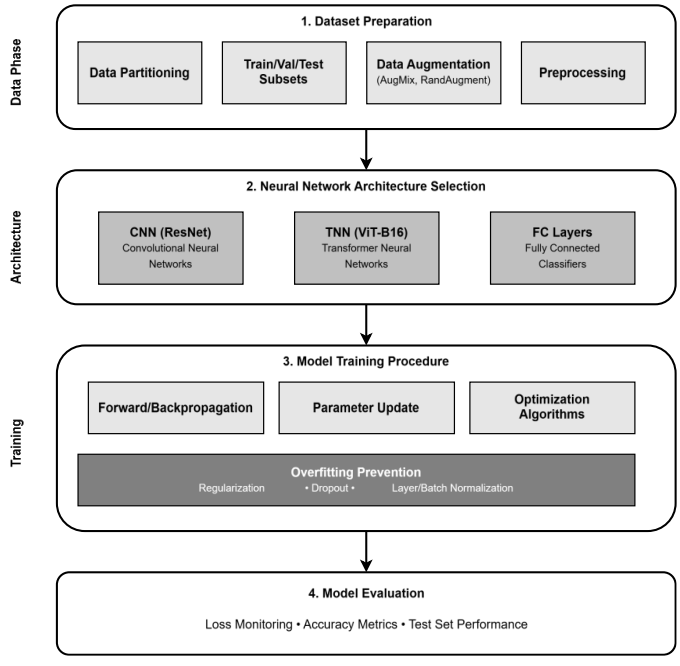


Figure 3. General overview of the model development and training workflow.

Transfer learning

Transfer learning leverages knowledge acquired from one task to enhance performance on related tasks, enabling model training with reduced data requirements and accelerated convergence. This approach exploits the commonality of low-level features across machine learning tasks. In computer vision, fundamental features such as edges, shapes, and textures generalize across object categories; similarly, syntactic and semantic patterns transfer (Pan & Yang, 2010). The methodology involves pre-training models on large-scale datasets to capture generic representations, subsequently adapting this knowledge to target tasks with limited data through techniques including feature extraction, fine-tuning, multi-task learning, and domain adaptation.

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

EXPERIMENTAL RESULTS

The experiments were conducted on balanced datasets, ensuring uniform representation of each class. Two types of artificial neural network (ANN) architectures were used: Convolutional Neural Networks (CNNs) and Transformer-based Neural Networks (TNNs). More detailed information regarding the hyperparameters can be found in (Kopeć et al.,2024), as the *species* and *varieties* models presented in this study were trained using identical or similar configurations. The definitions of business metrics used in this study are provided in Appendix A.

Brute force data collection.

Figure 4 illustrates the relationship between genera cloud classification model accuracy and the number of training samples per class. Both plots use a logarithmic scale on the x-axis, representing the number of samples per class, with time estimates annotated (e.g., 3 days to 4 years) to reflect the practical effort required for data collection. The y-axis shows classification accuracy in percentage.

A clear logarithmic trend is observed in both CNN and TNN performance, modelled by the function:

$$f(x) = a \cdot \ln(x) + b \tag{2}$$

where:

- f(x) is the model accuracy,
- x is the number of samples per class,
- a and b are constants specific to the network type.

The fitted parameters are summarized in the following table X:

Table 1. Model-specific approximation values of the constants $a$and b used in Equation (1) (CNN vs. TNN).

Const/ Model	CNN	TNN
a	2.54	2.98
b	54.75	51.96

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

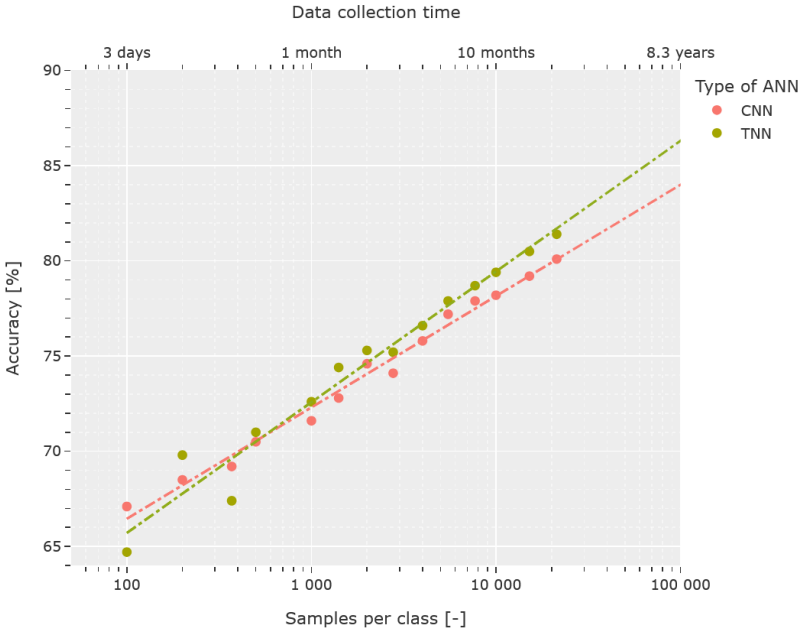


Figure 4. Classification accuracy as a function of available data per genera cloud class (expressed as an equivalent recording duration: 3 days–4 years). CNN and TNN results are compared; dashed lines show trend extrapolation.

Leave-one-year-out (LOYO) validation

To evaluate the temporal robustness of the cloud type classification model, a leave-one-year-out cross-validation (LOYO-CV) scheme was employed. The dataset comprised ground-based sky images collected over five consecutive years (2021–2025), annotated into seven cloud genera and clear sky: Cumulonimbus (Cb), Cumulus (Cu), Stratocumulus (Sc), Altocumulus (Ac), Stratus (St), Cirrus (Ci), and Cirrostratus (Cs). In each fold, data from four years were used for training, and the remaining year was held exclusively for testing. To mitigate the effects of class imbalance, a balanced sampling strategy was applied: 10,000 images per class per year were randomly selected for training, yielding an equal representation of all cloud genera across the training set. The test set for each fold consisted of 1,000 images per class drawn from the excluded year. This setup ensured that the model was never exposed to images from the evaluation year during training, thereby providing a rigorous assessment of its ability to generalize across different annual periods characterized by varying atmospheric and imaging conditions.

Table 2. Leave-one-year-out cross-validation (LOYO) results. Each column corresponds to a model trained on the remaining four years and evaluated on the indicated held-out year. Values represent Cohen's kappa (%).

Model (excluded year)	2021	2022	2023	2024	2025
2021	73.1	—	—	—	—
2022	—	72.4	—	—	—
2023	—	—	73.8	—	—
2024	—	—	—	74.6	—

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

2025	—	—	—	—	74.5

Held-out year	**2021**	**2022**	**2023**	**2024**	**2025**	**Mean ± SD**
Cohen's kappa (%)	73.1	72.4	73.8	74.6	74.5	73.7 ± 0.9

The LOYO-CV results demonstrated consistent classification performance across all folds, with accuracy ranging from 72.4% to 74.6%. The lowest accuracy was observed when 2022 was excluded from training, while the highest was achieved when 2024 was held out. The narrow spread of 2.2 percentage points across folds indicates that the model generalizes well across different annual periods and is not strongly dependent on data from any single year.

Cross-year transfer matrix

The cross-year transfer analysis revealed that the model trained on 2023 data exhibited the best generalization ability (row mean 71.85%), while the model trained on 2025 performed the weakest (69.20%). As in the LOYO-CV experiment, 8-class models were used with balanced sampling — 3,000 images per class for training and 1,000 images per class for testing. Column-wise means showed that classification accuracy increased systematically with the evaluation year — from 69.68% for 2022 to 72.10% for 2025, suggesting an improvement in the quality of acquired images over the study period.

Table 3. Cross-year temporal transfer matrix. Each row represents a model trained on a single year, and each column shows its evaluation Cohen Kappa score (%) when applied to a different year. Diagonal entries (training year = evaluation year) are omitted. Row and column means summarize the average generalization performance of each training year and the average predictability of each evaluation year, respectively.

Train \ Eval	**2021**	**2022**	**2023**	**2024**	**2025**	**Mean**
2021	—	70.40	70.00	70.80	72.00	**70.80**
2022	71.10	—	71.50	71.50	71.00	**71.28**
2023	70.90	71.40	—	72.30	72.80	**71.85**
2024	69.50	68.80	70.80	—	72.60	**70.43**
2025	68.20	68.10	69.30	71.20	—	**69.20**
Mean	**69.93**	**69.68**	**70.40**	**71.45**	**72.10**	

Cross-validation–based variance analysis of Cohen's κ for Sample size determination

In this study, the primary measure of the effectiveness of the developed ground-based cloud image classification system is Cohen's κ coefficient. This coefficient is computed separately for each classifier. To determine the optimal sample size for planned experiments (see Sect. 2.3), it was essential to empirically estimate the variability of Cohen's κ coefficient. For this purpose, a variance analysis of Cohen's κ was performed using a 4-fold stratified cross-validation procedure. The resulting distributions of κ for the species and varieties classification models within each cloud genus are shown in Fig. 5.

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

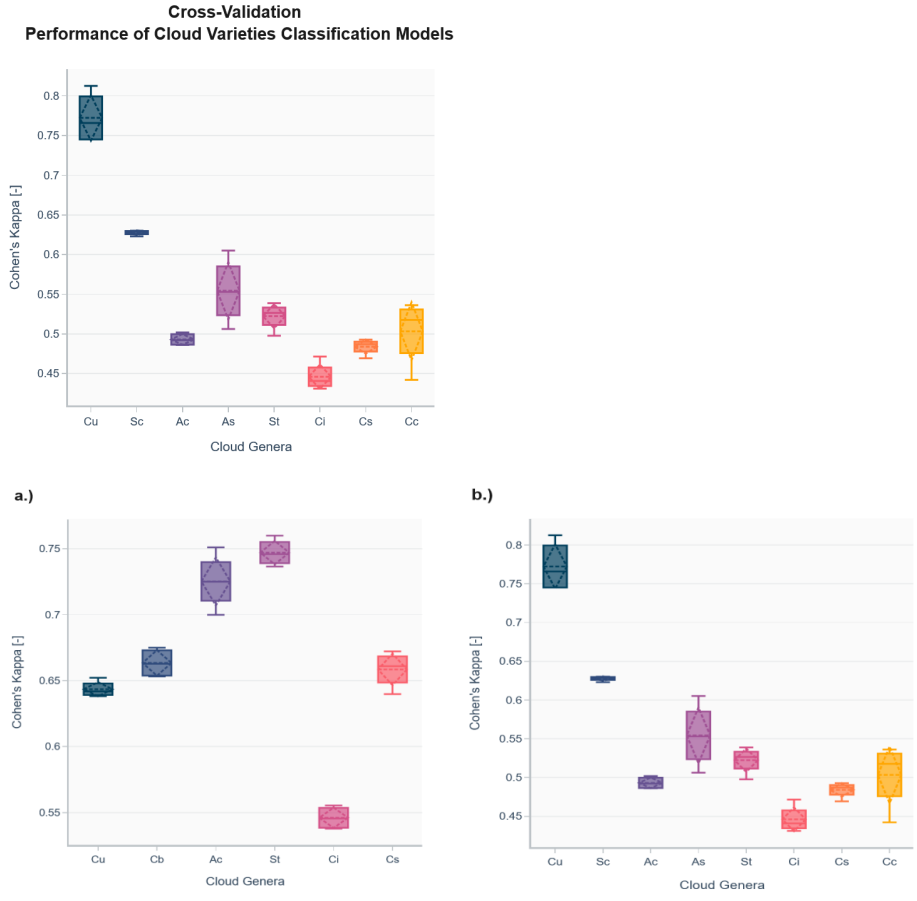


Figure 5. Boxplots of Cohen's κ obtained from 4-fold stratified cross-validation for (a) cloud species and (b) cloud varieties classification models within each cloud genus.The cross-validation method involved dividing the available dataset into four equally sized subsets (folds), stratified by cloud class. Each fold was sequentially used as a validation set, while the remaining three folds formed the training set. This strategy provided four independent estimates of Cohen's κ for each classification model. The variance of Cohen's κ arises primarily from the random partitioning of the dataset and random aspects of the training process (e.g., initialization of model parameters, random data transformations during augmentation). Consequently, the obtained κ estimates vary across folds, enabling assessment of this coefficient's variability under realistic experimental conditions.

This analysis forms the basis for the subsequent stage of the research, aimed at determining the minimum sample size necessary to achieve precise estimation of classifier performance at a specified confidence level and acceptable estimation error.

Cloud genera classification

Figure 4 shows that accuracy improves rapidly with an initial increase in dataset size, particularly during the early phase of data collection (e.g., within the first few days or months). This rapid

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

improvement highlights a practical opportunity: a relatively small investment in data acquisition can yield significant gains in model performance in the early stages of model development.

Transformer-based models (TNNs), while starting with slightly lower baseline accuracy than CNNs, exhibit a steeper growth rate and eventually surpass CNNs as more data becomes available. This suggests that TNNs may be particularly advantageous for large-scale classification tasks when sufficient training data can be acquired.

Overall, the logarithmic relationship confirms diminishing returns as the dataset size increases, emphasizing the importance of identifying a cost-effective balance between data collection effort and desired accuracy. This finding provides a valuable guideline for designing efficient data acquisition strategies, especially in fields like biodiversity monitoring or agricultural phenotyping, where labelled data is expensive and time-consuming to obtain.

Species cloud classification

A model encompassing all 14 classes was developed for cloud species classification, with each class corresponding to a single species. The results obtained for the species *uncinus*, *mediocris*, and *castellanus* did not exceed 60% F1 Score, whereas the highest values were achieved for *lenticularis* and *nebulosus*, reaching 78% and 79%, respectively; all results are presented in **Table 4.** Data for individual species were collected across all cloud genera in which a given species occurs; however, the sample size for certain classes proved insufficient for effective model training. To examine in detail the influence of genera on species classification and to isolate a given species from the entire data pool, separate models were developed for each cloud genera. This approach improved the evaluation metrics but requires prior classification of the cloud genera. The genera classification procedure is described in detail in [K].

Table 4. Precision, Recall, and F1 Score values for the classification of 14 cloud species using the multi-class model, ranked by F1 Score.

Species	Precision (%)	Recall (%)	F1 Score (%)
nebulosus	76	83	79
lenticularis	82	74	78
humilis	75	76	76
fractus	71	65	68
capillatus	68	68	68
fibratus	66	69	68
spissatus	66	69	67
floccus	70	65	67
congestus	65	63	64
stratiformis	67	60	63
calvus	59	64	62
mediocris	60	59	59
castellanus	56	50	53
uncinus	58	43	49

Table 5 presents the F1 scores (%) obtained by a species-level cloud classification model, with species listed as rows and cloud genera as columns. Each entry reflects the model's performance in identifying a given species when it appears in the context of a particular genera.

Legend:

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

- "–" indicates that the species is not taxonomically associated with the given genera or was not present in the dataset under that genera.
- "n.d." (no data) denotes species for which an insufficient number of samples was available to calculate a reliable score.
- "dom." (dominant) indicates that the respective genera was dominated by a single species in the dataset, making it impossible to evaluate the other co-occurring species independently due to class imbalance.

The model shows high performance for several species, particularly those with well-defined structural features. For example, *Fractus* achieves an F1 score of 90.0% when occurring within both *Cu* and *St* genera, while *Calvus* and *Capillatus* score over 80% in *Cb*. These results suggest that such species exhibit strong morphological signatures and are well represented in the dataset.

Performance for *Cirrus* species (columns *Ci*, *Cs*, *Cc*) is more variable. While *Fibratus* (62.3% in *Ci*, 81.9% in *Cs*) and *Floccus* (75.9% in *Ci*, 64.3% in *Cc*) are identified with moderate success, other *Cirrus* types such as *Spissatus* (56.5%) and *Uncinus* (61.5%) show reduced classification accuracy. This reflects the inherent challenges of distinguishing high-level, optically thin cloud forms that often share similar visual characteristics. Moreover, species marked as "dom." were overwhelmingly dominant in their respective genera, limiting the model's exposure to alternative species and thus its ability to learn meaningful discriminative features.

Overall, the results demonstrate the model's ability to distinguish certain cloud species with high F1 score, especially when they exhibit distinct morphology and sufficient representation in the dataset.

Table 5. F1 scores (%) of cloud species classification across genera.

Genera/Spec	Cb	Cu	Sc	Ac	As	Ns	St	Ci	Cs	Cc
Calvus	81	–	–	–	–	–	–	–	–	–
Capillatus	84	–	–	–	–	–	–	–	–	–
Castellanus	–	–	n.d.	77	–	–	–	n.d.	–	n.d.
Congestus	–	76	–	–	–	–	–	–	–	–
Fibratus	–	–	–	–	–	–	–	62	81	–
Floccus	–	–	n.d.	72	–	–	–	75	–	64
Fractus	–	90	–	–	–	–	90	–	–	–
Humilis	–	71	–	–	–	–	–	–	–	–
Lenticularis	–	–	n.d.	83	–	–	–	–	–	n.d.
Mediocris	–	58	–	–	–	–	–	–	–	–
Nebulosus	–	–	–	–	–	–	87	–	74	–
Spissatus	–	–	–	–	–	–	–	56	–	–
Stratiformis	–	–	dom.	67	–	–	–	–	–	75
Uncinus	–	–	–	–	–	–	–	61	–	–

Varieties cloud classification

Table 6 summarizes the performance of a cloud classification model at the variety level, using F1 score (%) as the evaluation metric. Each row corresponds to a specific cloud variety, and each column represents a cloud genera. The entries show the model's ability to correctly classify a given variety when it appears within a specific genera.

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Several varieties show strong classification performance. Notably, *Opacus* achieves high F1 scores in multiple genera, including 72% in *Ac* and 76% in *Sc*, indicating that this variety's dense, light-blocking characteristics are effectively captured by the model. *Radiatus* also performs well, reaching 85% in *Cu*, likely probably due to its distinctive linear patterning.

On the other hand, several rarer or morphologically subtle varieties, such as *Duplicatus*, *Lacunosus*, and *Intortus*, show limited model performance or lack sufficient data across most genera. For instance, *Intortus* achieves only 47% in *Ci*, while *Lacunosus* reaches 68% in *Cc*. These results point to the ongoing challenge of classifying varieties that are underrepresented in the training dataset or exhibit ambiguous visual features.

Overall, the model achieves strong performance for prevalent and morphologically distinctive cloud varieties, whereas rare, visually subtle, or structurally complex varieties remain challenging to discriminate—most likely due to limited and imbalanced training data. Increasing sample counts and improving class balance are therefore expected to enhance variety-level recognition, particularly for high-altitude and structurally intricate cloud types.

Table 6. F1 scores (%) of cloud variety classification across genera

Genera/Varieties	Cb	Cu	Sc	Ac	As	Ns	St	Ci	Cs	Cc
Translucidus	–	–	69	40	64	–	75	–	–	–
Perlucidus	–	–	67	45	–	–	–	–	–	–
Opacus	–	–	76	72	59	–	66	–	–	–
Duplicatus	–	–	n.d.	62	n.d.	–	–	n.d.	65	–
Undulatus	–	–	n.d.	72	n.d.	–	n.d.	–	69	60
Radiatus	–	85	n.d.	–	n.d.	–	–	62	–	–
Lacunosus	–	–	n.d.	n.d.	–	–	–	–	n.d.	68
Intortus	–	–	–	–	–	–	–	47	–	–
Vertebratus	–	–	–	–	–	–	–	62	–	–
Not present	–	87	71	45	71	–	54	50	58	56

CONCLUSION

The results presented in this study have demonstrated that models based on artificial neural networks built using convolutional neural network blocks or based on the self-attention mechanism and Transformer architecture can classify cloud species and varieties with satisfactory business metrics. Nevertheless, for certain cases, image data alone may be insufficient from a synoptic perspective, suggesting that integration with auxiliary meteorological data, such as ceilometer measurements, could further improve classification reliability.

The implementation of separate classification models for species and varieties within each genera resulted in improved metric values. For instance, *calvus* achieved an F1 score of 62% in the multi-class model encompassing all 14 species, compared to 81% when classified within a genera-specific model for Cumulonimbus, representing an improvement of 19 percentage points. Similar improvements were observed for other species, such as *fractus*, which reached 90% F1 score within both Cumulus and Stratus genera, compared to 68% in the general model. These findings demonstrate that genera-level separation significantly enhances species classification

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



performance by reducing inter-class confusion. However, this approach is contingent upon either the availability of a well-performing genera classification model or prior knowledge of the cloud genera.

The relationship between dataset size and classification accuracy follows a logarithmic trend for both CNN and TNN architectures. TNN demonstrates superior performance for larger datasets, suggesting it is more effective at leveraging additional training data. The logarithmic nature of this relationship implies diminishing returns: initial increases in data volume yield substantial accuracy improvements, whereas further expansion provides only marginal gains. From a practical perspective, extending data collection efforts beyond a certain threshold may not justify the additional time and resources required. This underscores the importance of conducting preliminary studies to determine the optimal dataset size for specific classification tasks, enabling more efficient allocation of resources in future research.

The seasonal variability in cloud genera composition confirms the influence of synoptic circulation patterns on regional cloud climatology. Notably, during the winter season, the increased frequency of low-level stratiform clouds (CL genera, particularly Stratus and Stratocumulus) tends to obscure higher-level cloud layers, effectively reducing the observed occurrence of Cirro-type clouds (CH genera, such as Cirrus, Cirrocumulus, and Cirrostratus). This observational bias should be considered when interpreting the seasonal distribution of cloud genera, as the apparent decrease in high-level cloud frequency during winter may partly reflect the screening effect of persistent low-level overcast rather than an actual absence of upper-tropospheric cloudiness. The consistent annual harmonic pattern observed across the 5-year period demonstrates the potential of automated cloud classification systems for long-term climate monitoring and trend detection. However, it should be noted that the 5-year observation period may be insufficient to draw climatological conclusions, which remains beyond the scope of this study.

The LOYO cross-validation demonstrated consistent temporal generalization, with Cohen Kappa scores ranging from 72.4% to 74.6% (spread of 2.2 percentage points), indicating that genera model performance remained stable regardless of the withheld annual period. Cross-year transfer analysis revealed that models trained on 2023 data exhibited optimal generalization ability (mean κ is 71.85%), while 2025-trained models showed the weakest performance (mean κ 69.20%). Column-wise evaluation scores increased systematically from 69.68% in 2022 to 72.10% in 2025, suggesting progressive improvement in image acquisition quality over the study period. The narrow performance variability across validation folds, combined with stable cross-year transfer metrics, confirms that the classifier is not overfitted to temporal artifacts and maintains robust discriminative capacity across varying atmospheric and imaging conditions.

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

APPENDIX A

METRICS

Model performance is evaluated using the confusion matrix and summary metrics (accuracy, precision, recall, and F_β). Cohen's $\kappa$is reported as the primary measure because it accounts for chance agreement and is more informative under class imbalance.

Confusion matrix

In a *k-class* classification problem, the confusion matrix offers a detailed, cell-by-cell view of model performance: each row represents the true label, each column the predicted label, and the diagonal entries C_{ii} record correct predictions.

Tabel A1. Confusion matrix for binary classification.

	Predicted Positive	**Predicted Negative**
Actual Positive	**TP** (True Positive)	**FN** (False Negative)
Actual Negative	**FP** (False Positive)	**TN** (True Negative)

For k-2, binary tasks, the matrix collapses to four canonical outcomes:. With more than two classes, for k>2, you analyse one class at a time (one-vs-all) or look at the whole table for error patterns. We often use row-normalise or column-normalize. For automated cloud-type classification, for instance, *TP* counts images correctly identified as cirrus (or any target class), while *FN* captures cirrus images the algorithm fails to recognise.

Accuracy

Accuracy is defined as the proportion of correctly classified instances to the total number of instances.

$$Accuracy = \frac{1}{N}\sum_{N}^{i} 1(y_i = y_i) \quad \text{(A1)}$$

Its popularity stems from condensing model performance into a single, easily interpretable scalar that—when hard labels are used—remains invariant to the chosen decision threshold. Nevertheless, this metric does not account for class imbalance; in an extreme case, a trivial model that always predicts the majority class can still yield 99 % accuracy. For this reason, accuracy must be reported alongside a confusion matrix or class-specific metrics (precision, recall, F-score) to avoid drawing misleading conclusions about model utility.

Precision — how many of my *alerts* were right?

It shows what proportion of positive predictions is correct. The higher the precision, the lower the Type I error rate (FP → α).

$$Precision = \frac{TP}{TP+FP} \quad \text{(A2)}$$

Boosting precision typically reduces sensitivity (recall) and raises the risk of a Type II error (FN → β). Use precision as the primary metric when minimizing false alarms is critical.

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Recall — how many of the *real* positives did I catch?

It shows what proportion of actual positives the model recovers. The higher the recall, the lower the Type II error rate (FN → β).

$$Recall = \frac{TP}{TP+FN} \quad \text{(A3)}$$

Boosting recall typically lowers precision and increases the risk of a Type I error (FP → α). Choose recall as the primary metric when missing true positives is costly.

F_β Score — balancing the two

The F-beta score is a generalized metric that combines precision and recall into a single measure through a weighted harmonic mean. The parameter β controls the relative importance of recall versus precision in the final score.

$$F_\beta = (1+\beta^2)\frac{precision*recall}{\beta^2*precision+recall} \quad \text{(A4)}$$

Parametr interpretation

- β = 1: Equal weight to precision and recall (F1 score)
- β > 1: Emphasizes recall (e.g., β = 2 for medical diagnostics)
- β < 1: Emphasizes precision (e.g., β = 0.5 for spam detection)

As a harmonic mean, the F-beta score exhibits sharp sensitivity to low values in either component metric. This characteristic prevents models from achieving high scores by optimizing only one metric while neglecting the other, thus providing a more balanced assessment of classifier performance in scenarios where both false positives and false negatives carry different costs.

Cohen's kappa (κ)— agreement beyond chance

Cohen's kappa (κ) measures classification performance while correcting for random agreement, making it particularly valuable for imbalanced cloud type distributions.

$$\kappa = \frac{p_O - p_e}{1-p_e} \quad \text{(A5)}$$

where p_0 is observed accuracy and p_e is chance agreement.

Table A2. Interpretation scale for Cohen's κ. "How much better is the model than random guessing that preserves class frequencies?"
Scale of thumb **(Landis & Koch):**

Value range	Level of agreement
0.81 – 1.00	**Almost perfect**
0.61 – 0.80	**Substantial**
0.41 – 0.60	**Moderate**

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

0.21 – 0.40	**Fair**
0.00 – 0.20	**Slight**
< 0	**Worse than chance**

Because κ corrects for prevalence of class, it is popular for evaluating annotator consistency or classifiers on imbalanced datasets, but it can itself be sensitive when one class dominates.

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

APPENDIX B

Table B1. Sample distribution across cloud genera–species categories in the training dataset for deep learning classifiers. Numerical values indicate the number of images per class.

Genera/ Species	Cb	Cu	Sc	Ac	As	Ns	St	Ci	Cs	Cc
Calvus	9851	–	–	–	–	–	–	–	–	–
Capillatus	10400	–	–	–	–	–	–	–	–	–
Castellanus	–	–	n.d.	328	–	–	–	n.d.	–	n.d.
Congestus	–	7631	–	–	–	–	–	–	–	–
Fibratus	–	–	–	–	–	–	–	800	4800	–
Floccus	–	–	n.d.	1200	–	–	–	506	–	724
Fractus	–	8000	–	–	–	–	12055	–	–	–
Humilis	–	8000	–	–	–	–	–	–	–	–
Lenticularis	–	–	n.d.	1151	–	–	–	–	–	n.d.
Mediocris	–	8000	–	–	–	–	–	–	–	–
Nebulosus	–	–	–	–	–	–	12800	–	4313	–
Spissatus	–	–	–	–	–	–	–	800	–	–
Stratiformis	–	–	dom.	1200	–	–	–	–	–	800
Uncinus	–	–	–	–	–	–	–	796	–	–

Table B2. Number of images used for training the deep-learning models for cloud variety classification across genera (genera/varieties).

Genera/Varieties	Cb	Cu	Sc	Ac	As	Ns	St	Ci	Cs	Cc
Translucidus	–	–	13945	760	800	–	1672	–	–	–
Perlucidus	–	–	14400	760	–	–	–	–	–	–
Opacus	–	–	14400	750	800	–	1680	–	–	–
Duplicatus	–	–	n.d.	760	n.d.	–	–	n.d.	527	–
Undulatus	–	–	n.d.	732	n.d.	–	–	–	560	240
Radiatus	–	782	n.d.	–	n.d.	–	–	693	–	–
Lacunosus	–	–	n.d.	n.d.	–	–	–	–	n.d.	206
Intortus	–	–	–	–	–	–	–	720	–	–
Vertebratus	–	–	–	–	–	–	–	622	–	–
Not present	–	782	14400	760	680	–	1680	720	560	240

Legend:
"–" indicates that the species is not taxonomically associated with the given genera or was not present in the dataset under that genera
"n.d." (no data) denotes species for which an insufficient number of samples was available to calculate a reliable score.
"dom." (dominant) indicates that the respective genera was dominated by a single species in the dataset, making it impossible to evaluate the other co-occurring species independently due to class imbalance.

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

APPENDIX C

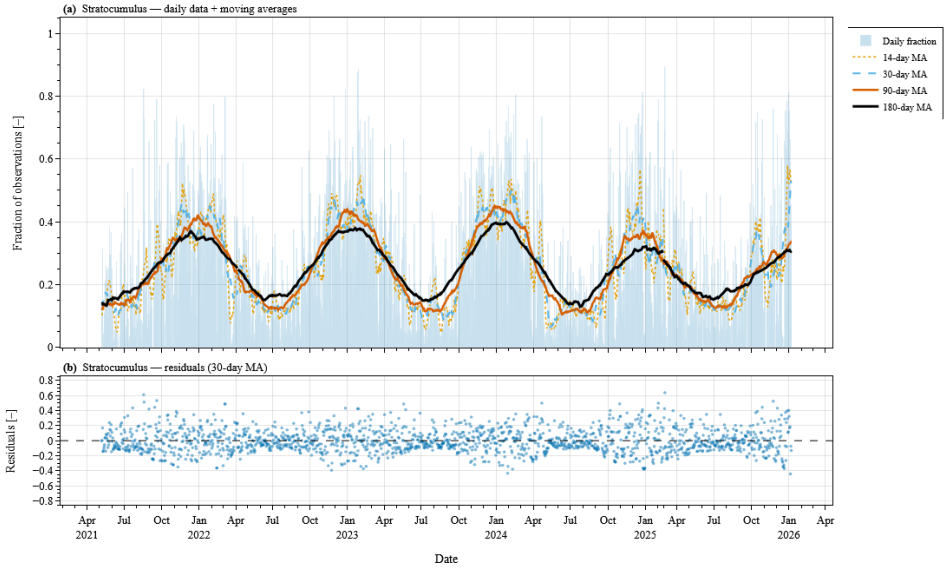


Figure C1. Time series of stratocumulus cloud fraction over the period April 2021 –January2026. Panel (a) shows the daily fraction of observations (light blue bars) along with 14-, 30-, 90-, and 180-day moving averages, revealing a pronounced seasonal cycle with winter maxima and summer minima. Panel (b) presents the residuals from the 30-day moving average, indicating the magnitude of short-term variability around the smoothed trend.

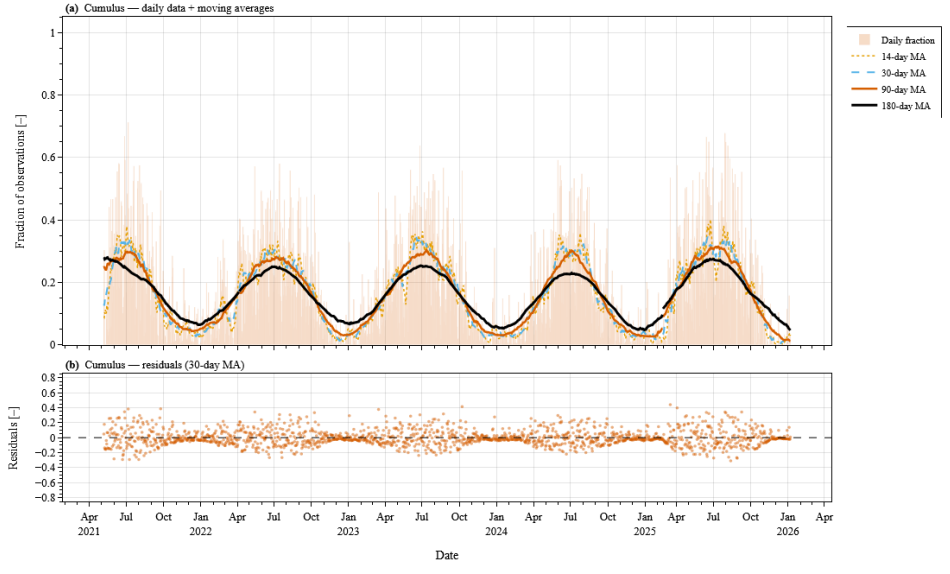


Figure C2. Cumulus cloud fraction over the period April 2021 – March 2026. Panel (a) shows daily values (light orange) with 14-, 30-, 90-, and 180-day moving averages, revealing a clear seasonal cycle with summer maxima and winter minima — opposite in phase to stratocumulus. Panel (b) presents residuals from the 30-day moving average, indicating stationary short-term variability.

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

References

Albrecht, B. A. (1989). Aerosols, Cloud Microphysics, and Fractional Cloudiness. *Science*, *245*(4923), 1227–1230. https://doi.org/10.1126/science.245.4923.1227

Bluestein, H. B. (2024). *The Architecture of Clouds*. Oxford University Press. https://doi.org/10.1093/oso/9780198870548.001.0001

Bohren, C. (1998). Introduction. W *Absorption and Scattering of Light by Small Particles* (s. 1–11). John Wiley & Sons, Ltd. https://doi.org/10.1002/9783527618156.ch1

Chollet, F. (2021). *Deep Learning with Python* (Second). Manning. https://www.manning.com/books/deep-learning-with-python-second-edition?a_aid=keras

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2019). *RandAugment: Practical automated data augmentation with a reduced search space*. https://arxiv.org/abs/1909.13719

Errico, R. M., Ohring, G., Weng, F., Bauer, P., Ferrier, B., Mahfouf, J.-F., & Turk, J. (2007). Assimilation of Satellite Cloud and Precipitation Observations in Numerical Weather Prediction Models: Introduction to the JAS Special Collection. *Journal of the Atmospheric Sciences*, *64*(11), 3737–3741. https://doi.org/10.1175/2007JAS2622.1

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press.

Guzel, M., Kalkan, M., Bostanci, E., Acici, K., & Asuroglu, T. (2024). Cloud type classification using deep learning with cloud images. *PeerJ. Computer Science*, *10*, e1779. https://doi.org/10.7717/peerj-cs.1779

Hamblyn, R. (2023). *Pocket Cloud Book Updated Edition: How to Understand the Skies in association with the Met Office*. David and Charles.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., & Lakshminarayanan, B. (2020). *AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty*. https://arxiv.org/abs/1912.02781

Kalkan, M., Bostancı, G. E., Güzel, M. S., Kalkan, B., Özsarı, Ş., Soysal, Ö., & Köse, G. (2022). Cloudy/clear weather classification using deep learning techniques with cloud images. *Computers and Electrical Engineering*, *102*, 108271. https://doi.org/10.1016/j.compeleceng.2022.108271

Kopeć, S., Duniec, G., Bochenek, B., & Figurski, M. (2024). Artificial neural networks in automatic image classifications of cloud from ground-based observations using deep learning models. *Quarterly Journal of the Royal Meteorological Society*, *150*(765), 5206–5224. https://doi.org/10.1002/qj.4865

Li, Z., Kong, H., & Wong, C.-S. (2023). Neural Network-Based Identification of Cloud Types from Ground-Based Images of Cloud Layers. *Applied Sciences*, *13*(7). https://doi.org/10.3390/app13074470

Liu, L., Sun, X., Chen, F., Zhao, S., & Gao, T. (2011). Cloud Classification Based on Structure Features of Infrared Images. *Journal of Atmospheric and Oceanic Technology*, *28*(3), 410–417. https://doi.org/10.1175/2010JTECHA1385.1

https://doi.org/10.5194/egusphere-2026-1551
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



Lusi, A. R., Orte, P. F., Wolfram, E., & Orlando, J. I. (2024). Cloud classification through machine learning and global horizontal irradiance data analysis. *Quarterly Journal of the Royal Meteorological Society*, *150*(765), 5435–5451. https://doi.org/10.1002/qj.4880

Nie, Y., Li, X., Paletta, Q., Aragon, M., Scott, A., & Brandt, A. (2024). Open-source sky image datasets for solar forecasting with deep learning: A comprehensive survey. *Renewable and Sustainable Energy Reviews*, *189*, 113977. https://doi.org/10.1016/j.rser.2023.113977

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Retallack, B. J. (1970). *Compendium of Lecture Notes for Training Class IV Meteorological Personnel*. Secretariat of the World Meteorological Organization. https://books.google.pl/books?id=gcf6zAEACAAJ

Rosenberger, M., Dorninger, M., and Weissmann, M.: Deriving WMO cloud classes from ground-based RGB pictures with a residual neural network ensemble, Earth Space Sci., 12, e2024EA004112, https://doi.org/10.1029/2024EA004112, 2025.

Shi, C., Han, L., Zhang, K., Xiang, H., Li, X., Su, Z., & Zheng, X. (2024). Improved RepVGG ground-based cloud image classification with attention convolution. *Atmospheric Measurement Techniques*, *17*(3), 979–997. https://doi.org/10.5194/amt-17-979-2024

Taravat, A., Del Frate, F., Cornaro, C., & Vergari, S. (2015). Neural Networks and Support Vector Machine Algorithms for Automatic Cloud Classification of Whole-Sky Ground-Based Images. *IEEE Geoscience and Remote Sensing Letters*, *12*(3), 666–670. https://doi.org/10.1109/LGRS.2014.2356616

Twomey, S. (1977). The Influence of Pollution on the Shortwave Albedo of Clouds. *Journal of Atmospheric Sciences*, *34*(7), 1149–1152. https://doi.org/10.1175/1520-0469(1977)034%253C1149:TIOPOT%253E2.0.CO;2

Wacker, S., Gröbner, J., Zysset, C., Diener, L., Tzoumanikas, P., Kazantzidis, A., Vuilleumier, L., Stöckli, R., Nyeki, S., & Kämpfer, N. (2015). Cloud observations in Switzerland using hemispherical sky cameras. *Journal of Geophysical Research: Atmospheres*, *120*(2), 695–707. https://doi.org/10.1002/2014JD022643

Winkler, P. (2025). History of Research on Cloud Types and Naming of Clouds. *History of Geo- and Space Sciences Discussions*, *2025*, 1–45. https://doi.org/10.5194/hgss-2025-3

WMO, W. M. (1987). *International Cloud Atlas: Manual on the Observation of Clouds and Other Meteors: II*. World Meteorological Organization.