

Satellite-trained latent spaces for evaluating km-scale atmospheric simulations

Dwaipayan Chatterjee¹, Peter Knippertz¹, Nina Raabe², Susanne Crewell²,
Benoît Vannière³, and Peter Dueben³

¹Institute of Meteorology and Climate Research Troposphere Research, Karlsruhe Institute of Technology,
Karlsruhe, Germany

²Institute of Geophysics and Meteorology, University of Cologne, Cologne, Germany

³European Centre for Medium-Range Weather Forecasts, Bonn, Germany

Key Points:

- Satellite image embeddings create an observational reference for evaluating cloud-surface radiative patterns in two km-scale weather models
- The novel approach reveals model-specific biases in the timing, duration, and transition pathways of cloud-surface radiative evolution
- Both weather models distort daytime shallow-cloud development toward either cloud-dominated or surface-heating-dominated evolution

Corresponding author: Dwaipayan Chatterjee, dwaipayan.chatterjee@kit.edu

Abstract

As km-scale simulations are increasingly performed globally, new evaluation approaches are needed to assess cloud structure and temporal development against observations. We introduce an observation-centered framework for evaluating two km-scale weather models over central Europe within a latent space learned from geostationary satellite observations. A self-supervised encoder trained on Meteosat Second Generation (MSG) 10.8 μm brightness-temperature imagery yields embeddings partitioned into eight radiative-scene regimes reflecting cloud organization, cloud-top temperature, and near-surface thermal information. MSG-equivalent fields from the Integrated Forecasting System (IFS) and ICOsahedral Non-hydrostatic model over Europe (ICON-EU) are projected into this fixed space and evaluated by regime occurrence, persistence, transitions, scene-by-scene correspondence, and a continuous thermal-scene evolution index. Both models reproduce the broad observed diurnal evolution, including warm daytime development and transitions from surface-dominated broken clouds to cloud-dominated states, but their dominant errors differ. Both under-represent daytime shallow low clouds and pass through them too rapidly, with larger deficits in ICON-EU despite its low-cloud maximum occurring near the observed midday phase. ICON-EU often maps observed low-cloud scenes to the warm-transition regime, which it occupies too broadly and retains too long. IFS also routes departures from daytime low cloud too strongly toward this regime, while distributing corresponding observed scenes among warmer and cooler states. Beyond daytime low clouds, IFS over-retains homogeneous and cloudier nighttime regimes and under-represents widespread deep cloud toward evening, whereas ICON-EU broadens the occurrence of scattered deep convection. By revealing coherent biases in radiative-state evolution, the framework provides pathway-specific targets for evaluating and improving km-scale weather models.

Plain Language Summary

Clouds strongly influence weather, but they are difficult for weather models to reproduce because they involve complex physics, change quickly, and vary from place to place. This can limit forecast accuracy. We develop a machine-learning method that learns recurring combinations of cloud cover and land-surface temperature directly from satellite images. These satellite-based patterns are then used as a reference for evaluating two weather models over central Europe. Both models capture the broad daily shift from warmer daytime conditions, shaped by surface heating, to colder conditions, more strongly influenced by clouds. However, both produce too few shallow, low-level clouds during the day and move through these conditions too quickly. One model more often replaces the observed low-cloud scenes with warmer transitional conditions and spreads scattered deep clouds across too much of the day. The other also moves away from low-cloud conditions too readily, keeps nighttime scenes too cloudy for too long, and produces too few widespread deep-cloud scenes toward evening. The method, therefore, shows not only whether models produce too much or too little cloud, but also whether cloud patterns occur at the right time and evolve in the right way. This provides clearer targets for improving weather models and forecasts.

1 Introduction

Clouds regulate exchanges of radiation, moisture, heat, and momentum across a wide range of scales (Bony et al., 2006; Stephens, 2005; Höjgård-Olsen et al., 2020), yet they remain one of the largest sources of uncertainty in weather and climate prediction (Randall et al., 2007). This uncertainty arises because cloud formation, organization, and dissipation depend on nonlinear interactions among turbulence, convection, microphysics, radiation, surface forcing, and the surrounding thermodynamic environment (IPCC Core Writing Team et al., 2007). Recent advances in kilometer-scale atmospheric modeling

(Hohenegger et al., 2023) have substantially improved the realism of simulated cloud fields (Nordling et al., 2024), allowing models to explicitly represent mesoscale structures that were previously under-resolved, or strongly smoothed (Stevens et al., 2019; Rackow et al., 2025). However, this increasing realism also creates a more demanding evaluation problem. As simulations begin to produce detailed and visually plausible atmospheric scenes, model evaluation must determine not only whether the right amount of cloud or radiation is produced, but also whether the simulated structures occupy the same range of observed mesoscale variability and develop in a physically consistent way (Oware et al., 2025; Bruno et al., 2026).

Traditional model diagnostics remain essential for quantifying mean biases, radiative effects, cloud amount, and bulk cloud properties (Bony et al., 2004; Tselioudis et al., 2000; Murphy et al., 2004). They have also enabled important cloud-regime evaluations based on aggregated distributions or joint histograms of retrieved cloud properties (Gordon et al., 2005; Williams & Webb, 2008). Such approaches quantify which cloud or radiative states are present and how often they occur. However, they often reduce complex atmospheric scenes to a small number of variables, spatial averages, or regime frequencies. This compression can obscure important compensating errors: a model may reproduce a reasonable brightness-temperature distribution or cloud-regime frequency while misrepresenting the spatial arrangement of clouds, the timing of their development, or the pathway by which scenes evolve. These limitations become increasingly important for km-scale simulations, where mesoscale organization is no longer merely a sub-grid effect but an explicit part of the simulated atmospheric state (Brueck et al., 2020). Evaluation frameworks, therefore, need to retain more of the scene structure that km-scale models are designed to represent.

Geostationary satellite observations provide a natural basis for such an evaluation because they observe the same region at high temporal frequency and capture the evolving radiative structure of cloudy, partly cloudy, and clear-sky scenes. Infrared window-channel brightness temperatures (T_b) are particularly useful in this context. In optically thick cloudy scenes, the $10.8\ \mu\text{m}$ brightness temperature is primarily governed by thermal emission from cloud tops and therefore reflects cloud-top temperature and vertical cloud structure (Lensky & Rosenfeld, 2008; Hamann et al., 2014). In clear-sky and partly cloudy scenes, the same channel is also sensitive to land-surface temperature (Trigo et al., 2008; Li et al., 2012), as well as to sub-scene mixtures of surface and cloud emission. Thus, over land, the infrared signal alone does not describe cloud properties. It contains information on a combined surface-cloud radiative scene in which cloud-top thermal structure, cloud organization, clear-sky fraction, land-surface heating, and diurnal variability are intertwined.

Satellite observations are commonly used for numerical weather prediction verification in two complementary ways. One approach retrieves model-like cloud products from observations, for example, cloud masks or derived cloud properties (Derrien & Le Gléau, 2005; Haiden et al., 2011; Prein et al., 2015). The other approach is to forward-simulate satellite-equivalent measurements from model states. Observation simulators such as the Radiative Transfer for TOVS (RTTOV) (Eyre, 1991; Saunders et al., 2018) enable this second approach by converting simulated atmospheric profiles into satellite-equivalent radiances or brightness temperatures (Hamann et al., 2014; Eikenberg et al., 2015; Sokol et al., 2022). These methods allow models and observations to be compared in a more consistent radiative space and have supported evaluations of cloud amount, radiative properties (Gleckler et al., 2008; Pincus et al., 2008), cloud-regime frequencies, and joint distributions of retrieved cloud-top pressure and optical depth (Williams & Webb, 2008). Nevertheless, most such diagnostics still rely on predefined variables or aggregated summaries. They do not provide a compact coordinate system in which the full satellite scene can be compared, tracked, and interpreted as part of the observed variability itself.

Self-supervised representation learning enables the construction of such a coordinate system directly from satellite imagery. By learning from high-dimensional observations without predefined cloud labels, these methods can map atmospheric scenes into a compact latent space that preserves information about spatial structure and radiative variability (Chatterjee et al., 2024; Geiss et al., 2023). In this space, visually and physically similar scenes are placed close together, while dissimilar scenes are separated (Denby, 2020; Happ et al., 2026). Recent studies have shown that representations learned from satellite imagery can organize observed cloud fields along continuous manifolds and reveal meaningful variability in cloud morphology, mesoscale organization, and temporal evolution (Denby, 2023; Chatterjee, Raabe, & Crewell, 2026). This makes latent spaces a promising observation-derived description of atmospheric scene variability.

Observation-derived cloud classifications have also been used for model evaluation. Supervised machine-learning approaches have transferred cloud-type information learned from ground-based or satellite observations into reanalyses and climate model output (Kuma et al., 2023; Kaps et al., 2023, 2024). Related studies have combined satellite-derived cloud regimes with corresponding satellite-simulator output to compare observed and simulated cloud populations in global atmosphere and climate models (Medeiros et al., 2023; Davis & Medeiros, 2024). Unsupervised latent representations have likewise been used to compare storm-resolving simulations, although in that case the representation was learned purely from simulated atmospheric fields rather than from observations (Mooers et al., 2023).

An observation-trained latent space offers a complementary capability by providing a fixed reference manifold. If regimes are defined separately for observations and simulations, similar labels may not correspond to the same radiative structures, and model-specific biases may be absorbed into the diagnostic itself. By contrast, an observation-trained manifold defines the range, neighborhoods, and separations of atmospheric states solely from the observed satellite record. Model simulations can then be projected into this fixed space and evaluated based on how they populate and move through the observed variability.

Here, we develop such an observation-trained latent-space framework for evaluating km-scale atmospheric simulations directly in satellite-observation space. To our knowledge, previous studies have not yet combined a self-supervised representation learning trained exclusively on geostationary satellite observations with its use as a fixed reference manifold for scene-by-scene evaluation of temporal evolution in km-scale numerical weather prediction models. A self-supervised neural network is trained exclusively on Meteosat Second Generation (MSG; Schmetz et al., 2002) 10.8 μm infrared brightness-temperature observations over Central Europe and used to construct a compact latent representation of recurring mesoscale surface-cloud radiative regimes (Section 4.1.1). This observational latent space is then kept fixed and used as a common reference manifold into which forward-simulated brightness temperatures from the ICOSahedral Non-hydrostatic model ICON-EU (Zängl et al., 2015) and the Integrated Forecasting System (IFS) at the kilometer scale are projected. We use this framework to assess how the simulations populate the observed manifold (Section 4.1.2), how their assigned regimes evolve in time (Section 4.2), and where simulated scenes depart from the observed state structure (Section 4.3). In doing so, the study moves considerably beyond evaluating state frequencies alone and provides a scene-based diagnostic framework for assessing whether km-scale models reproduce the organization and temporal development of observed surface-cloud radiative variability.

2 Dataset

This study focuses on Central Europe, spanning 47–53°N and 6–27.5°E, during the boreal spring-to-early-autumn season from April to September in 2020–2024. During this

period, the region is characterized by pronounced cloud and thermal-scene variability associated with land-surface heating, boundary-layer growth, moist convection, frontal disturbances, and their diurnal evolution (Taszarek et al., 2018; Dillerup et al., 2026). We use MSG geostationary infrared observations, available every 15 minutes over Europe at an approximate spatial resolution of ≈ 4 km. The observational latent space (see Section 3) is constructed from the $10.8 \mu\text{m}$ infrared window-channel brightness temperature, which primarily senses thermal emission from the Earth’s surface under clear-sky conditions and from cloud tops under cloudy conditions (Sobrino & Romaguera, 2004). The 6.2 and $7.3 \mu\text{m}$ water-vapor channels provide complementary sensitivity to different depths of the troposphere because their weighting functions generally peak at different atmospheric levels (Koenig & de Coning, 2009). Under typical clear-sky conditions, the $6.2 \mu\text{m}$ channel peaks higher, primarily sampling upper-tropospheric water vapor, whereas the weaker water-vapor absorption in the $7.3 \mu\text{m}$ channel shifts its peak sensitivity downward toward the middle troposphere and, under sufficiently dry conditions, permits contributions from still lower levels. Neither water-vapor channel is used for training; instead, their complementary vertical sensitivities are retained for the physical characterization of the identified latent-space regimes. At each available timestamp, four random 128×128 pixel crops are extracted within the study domain. Each crop spans roughly $512 \text{ km} \times 512 \text{ km}$, providing mesoscale scenes large enough to capture organized cloud structures while retaining regional variability (Chatterjee et al., 2023; Corradini et al., 2026). This sampling strategy, of extracting four random crops per timestep across Europe, yields approximately 350,000 satellite-observation samples.

To compare the observations with numerical weather prediction output, we use IFS and ICON-EU forecasts provided in the MSG-equivalent observation space. These products are generated with the fast radiative transfer model RTTOV, which simulates top-of-atmosphere radiances and brightness temperatures for passive visible, infrared, and microwave satellite sensors. The IFS product is based on IFS Cycle 48r1 and includes MSG-Spinning Enhanced Visible and InfraRed Imager (SEVIRI) equivalent radiances or brightness temperatures for three infrared and two visible channels, including the $6.2 \mu\text{m}$ and $7.3 \mu\text{m}$ water-vapor bands and the $10.8 \mu\text{m}$ infrared window channel. These simulations are part of experimental IFS configurations developed for weather-induced extreme-event applications within the Destination Earth program (Hoffmann et al., 2023) and are available at an approximate spatial resolution of 5 km. We also include the operational ICON-EU MSG-SEVIRI-equivalent brightness temperatures as a regional model for comparison. ICON-EU is a limited-area configuration of the ICON modeling system that is closely coupled to the global ICON forecast through two-way nesting (Zängl et al., 2022) and is available at approximately 6.5 km spatial resolution.

Despite their kilometer-scale grids, both forecast configurations retain parameterizations of shallow and deep convection based on variants of the Tiedtke mass-flux framework (Tiedtke, 1989). They differ in other cloud-relevant physics: ICON-EU uses a prognostic turbulent-kinetic-energy scheme and single-moment bulk microphysics, whereas IFS Cycle 48r1 uses an eddy-diffusivity-mass-flux turbulence scheme and a distinct single-moment bulk cloud and precipitation scheme (Prill et al., 2024; Bechtold et al., 2020).

For both IFS and ICON-EU, the evaluation focuses on the MSG-equivalent $10.8 \mu\text{m}$ infrared window-channel brightness temperature, ensuring consistency with the satellite channel used to train the observational latent space. Both forecast systems are initialized at 00:00 UTC and are available at hourly forecast lead times. We consider the first 24 hours of each forecast to assess the simulated diurnal evolution of mesoscale thermal scenes. For a consistent model intercomparison, the analysis is restricted to 2024, the overlapping period for which both products are available. At each hourly timestamp, four random crops are sampled within the central European study domain: 100×100 pixels for IFS, corresponding to approximately $500 \text{ km} \times 500 \text{ km}$, and 80×80 pixels for ICON-EU, corresponding to approximately $520 \text{ km} \times 520 \text{ km}$. This yields approx-

imately 18,000 thermal-scene samples per model to assess whether the forecast systems occupy and evolve within the same observation-derived latent state space as the satellite observations.

In addition to the random crops, we extract fixed-location crops from the satellite observations and the MSG-equivalent IFS and ICON-EU forecast products. These crops are centered on seven locations across central Europe: Jülich (50.92°N, 6.36°E), Lindenberg (52.21°N, 14.12°E), Warsaw (52.22°N, 21.01°E), Nuremberg (49.45°N, 11.07°E), Dresden (51.05°N, 13.74°E), Vienna (48.20°N, 16.37°E), and Bourges (47.08°N, 2.39°E). Jülich and Lindenberg are included because they host long-term atmospheric observatories with complementary in situ and remote sensing measurements, while the remaining sites provide broader spatial sampling across different parts of the central European domain. These site-centered crops allow us to examine the diurnal evolution of observed and model-derived thermal scenes at fixed geographic locations (Eastman & Warren, 2014; Zhang et al., 2025). For Jülich, we additionally use long-term observations from the JOYRAD-35 cloud radar at the Jülich Observatory for Cloud Evolution (JOYCE; Löhnert et al., 2015). Radar processing also incorporates CloudNet cloud-base height estimates (Illingworth et al., 2007), derived from collocated radar, microwave radiometer, and ceilometer observations at JOYCE, yielding a common 30-second temporal and 30-meter vertical resolution. These measurements are particularly useful for interpreting latent-space regimes dominated by cloudy scenes, where the 10.8 μm satellite signal is primarily controlled by cloud-top thermal emission but does not resolve the vertical cloud structure beneath the emitting layer. The combination of satellite-derived thermal scenes and ground-based radar profiles, therefore, helps relate the learned representation to both the horizontal radiative organization seen from space and the local vertical structure of cloud systems.

To provide additional meteorological context for the selected Jülich case study on 11 May 2024, we use hourly fields from the ERA5 atmospheric reanalysis (Hersbach et al., 2020). Mean sea-level pressure and geopotential height at 500 hPa are used to characterize the synoptic-scale pressure pattern and mid-tropospheric circulation during the case. Cloud occurrence within the case-study domain is additionally quantified using the instantaneous Level-2 binary cloud mask from the third edition of the Satellite Application Facility on Climate Monitoring (CM SAF) CLOUD property dAtAset using SEVIRI (CLAAS-3; Benas et al., 2023). Available every 15 minutes on the native SEVIRI grid, the binary mask distinguishes cloudy from clear pixels and is used to calculate the cloud fraction of each case-study scene and to summarize its evolution over the day. ERA5 and CLAAS-3 are used solely to establish the meteorological context of the case study and do not enter the latent-space training or regime assignment.

3 Method

We construct an observation-centered evaluation framework in three stages. First, a self-supervised encoder learns a continuous compact representation of mesoscale thermal scenes exclusively from MSG observations. Second, this observational continuum is partitioned into a discrete set of physically interpretable reference classes. Finally, MSG-equivalent fields from ICON-EU and IFS are projected into the fixed observational representation and evaluated with respect to class occurrence, persistence, transition pathways, simultaneous correspondence, and continuous thermal-scene evolution.

3.1 Observation-trained representation and reference classes

The encoder is trained on MSG 10.8 μm brightness-temperature scenes using the DINO self-distillation framework (Caron et al., 2021) with a vision-transformer (ViT) backbone (Dosovitskiy et al., 2021). By enforcing consistency among different global and local views of the same scene, DINO learns recurring spatial organization and thermal structure without predefined cloud-regime labels or explicit negative pairs. After train-

ing, the DINO projection head is discarded, and each 128×128 -pixel scene is represented by the 384-dimensional output of the ViT backbone, denoted by $\mathbf{z}$. The feature vectors are L_2 -normalized so that they lie on the unit hypersphere, where cosine distance provides a measure of similarity between mesoscale thermal scenes. The network architecture, multi-crop transformations, training objective, normalization, and optimization settings are described in Appendix A.

After training, the encoder weights are frozen, and embeddings are extracted for all observational samples. We initially partition the normalized observational manifold into ten classes using spherical (k)-means clustering (Hornik et al., 2012), consistent with the cosine geometry of the representation. This initial partition is subsequently refined using only observational information. For each class, we compare the multichannel brightness-temperature distributions from the MSG 6.2, 7.3, and $10.8 \mu\text{m}$ channels together with the corresponding 15-min diurnal occurrence profile. Radiometric similarity is quantified using Jensen–Shannon divergence (JSD), while similarity in diurnal occurrence is evaluated using Earth Mover’s Distance (EMD) and an independent principal-component-analysis-based comparison.

A threshold-sweeping procedure identifies initial clusters that are consistently similar in both their radiometric characteristics and diurnal occurrence. At the most conservative stopping point, three initial clusters representing widespread deep clouds satisfy all similarity criteria and are therefore combined, reducing the partition from ten initial clusters to eight effective observational classes. These classes and their centroids define the fixed observation-derived reference partition used throughout the analysis. Details of the radiometric and diurnal diagnostics, the JSD, EMD, and PCA criteria, the threshold sweep, and the mapping between the initial and final classes are provided in Appendix B.

For visualization, the L_2 -normalized observational features are first reduced to 50 dimensions using PCA, after which a two-dimensional t-distributed Stochastic Neighbor Embedding (t-SNE, van der Maaten & Hinton, 2008) embedding is fitted exclusively to MSG. The t-SNE representation is used only for visualization; all clustering, class assignments, and distance calculations are performed in the original 384-dimensional representation.

3.2 Projection of simulations and physical characterization

Once the observational partition has been established, neither the encoder nor the observation-derived class centroids are modified. MSG-equivalent $10.8 \mu\text{m}$ brightness-temperature crops from ICON-EU and IFS are passed directly through the same frozen encoder at their native array dimensions, without explicit resizing or spatial regridding. Within the encoder, the ViT forward routine automatically adapts the learned positional encodings to the input-dependent patch-token grid.

The input normalization is performed separately for each data source. MSG scenes are standardized using the mean and standard deviation calculated from the observational training data. For the simulations, separate normalization statistics are computed for ICON-EU and IFS, and each dataset is standardized to its respective mean and standard deviation before passing through the encoder. The resulting feature vectors are L_2 -normalized and assigned to the nearest observation-derived centroid using cosine distance. Further details on the model crop dimensions, source-specific normalization, and feature-extraction procedure are provided in Appendix A.

The simulations are therefore not clustered independently, and no simulated data are used to modify the observational class definitions or their centroids. Instead, each simulated scene is interpreted based on its location within the representation learned exclusively from MSG. Differences in class occurrence and evolution can consequently be

interpreted as departures from the observed organization of mesoscale surface–cloud radiative scenes. For the two-dimensional visualizations, the IFS and ICON-EU features are likewise transformed into the fixed observation-fitted t-SNE embedding rather than being embedded independently.

We independently characterize the vertical cloud structure associated with the observational classes using cloud-radar measurements from the JOYRAD-35 instrument at JOYCE. For each class, radar-reflectivity profiles above the CloudNet-derived cloud base height are composited into Contoured Frequency-by-Altitude Diagrams (CFADs). These class-conditioned CFADs complement the satellite-derived thermal representation with vertically resolved information and help identify classes associated with enhanced or vertically developed cloud occurrence.

3.3 Evaluation diagnostics

For all three datasets, the evaluation diagnostics are restricted to 2024, the common evaluation period for which both model products overlap with MSG observations. At the seven fixed locations, class occurrence is summarized as a function of hour of day. Persistence is defined as the uninterrupted duration for which a scene remains assigned to the same class. Class-wise persistence distributions and their site-level means are compared among MSG, ICON-EU, and IFS to determine whether the simulations reproduce the observed residence times of the different surface–cloud radiative regimes.

Transitions are identified at the boundaries of continuous class runs. A non-self transition is counted when the following sample occurs at the expected dataset-specific temporal interval and the destination class persists for at least 20 min. Transition counts are normalized by source class and therefore describe the conditional distribution of destination states after a scene leaves a given class. Model transition biases are expressed as percentage-point differences from the corresponding MSG transition probabilities.

Simultaneous regime correspondence is evaluated using geographically and temporally paired MSG and simulated scenes. Each model crop is paired with the MSG crop representing the same geographic footprint, and each hourly model output is matched to the nearest 15-min MSG observation. The correspondence matrices are normalized by observed class. Their diagonal elements therefore quantify the probability that a simulation occupies the same observation-derived class when that class is observed, whereas off-diagonal elements identify systematic displacement toward other regions of the observational manifold.

For the case-study analysis, the discrete class assignments are complemented by a thermal-scene evolution index (TSEI). It measures the latent-space distance of each scene from its warmest daily reference state, normalized by the separation between the warmest and coldest daily reference states:

$$\text{TSEI}(t) = \frac{d_{\cos}(Z(t), Z_{\text{warmest}, T_b})}{d_{\cos}(Z_{\text{coldest}, T_b}, Z_{\text{warmest}, T_b})}. \quad (1)$$

Here, $d_{\cos}$ denotes cosine distance, while Z_{warmest, T_b} and Z_{coldest, T_b} denote the warm and cold daily reference states. These are selected separately for each data source as the scenes with the daily domain-mean $10.8 \mu\text{m}$ brightness temperatures closest to the 90th and 10th percentiles, respectively. Values close to $\text{TSEI} = 0$ therefore indicate proximity to the warm reference state, whereas values close to $\text{TSEI} = 1$ indicate proximity to the cold reference state. Because the references are selected independently for MSG, IFS, and ICON-EU, the TSEI is used to compare the relative timing and shape of their daily evolution rather than their absolute thermal or morphological states. The class labels provide the complementary information needed to identify the radiative branches that these trajectories follow.

4 Results

4.1 Latent-space representation of combined surface–cloud radiative scenes

Before evaluating the models, Section 4.1.1 establishes the structure and physical meaning of the observation-trained manifold. Section 4.1.2 then passes MSG-equivalent ICON-EU and IFS brightness-temperature fields through the fixed encoder, and assigns the resulting feature vectors to the observation-derived reference classes.

4.1.1 Observational latent space and physical interpretation

The observation-trained latent space organizes MSG 10.8 μm brightness-temperature scenes into eight regions spanning a continuum of coupled surface–cloud radiative states (Figure 1). Because the encoder is trained exclusively on the infrared window channel, these regions should not be interpreted as cloud types alone; their separation reflects the combined influence of land-surface emission, cloud-top temperature, and spatial organization. In the two-dimensional visualization, the classes C1, C3, C5, C7, and C8 form relatively distinct regions, whereas C2 and C4 overlap more with their neighbors. C6 shows both characteristics: a compact, high-density core and a broader density contour extending into adjacent regions. The classes themselves are defined from the original 384-dimensional feature vectors, with t-SNE used only to visualize their organization. We interpret these regions using their diurnal occurrence, thermal evolution, representative satellite scenes, and JOYCE radar CFADs, and use persistence and transition probabilities to characterize their temporal stability and connectivity (Figures 2–3).

The warm, daytime branch of the observational manifold comprises C1 and C2. In the two-dimensional visualization, C1 forms a relatively distinct region, whereas C2 occupies a more overlapping, intermediate part of the manifold (Figure 1). C1 exhibits the highest brightness temperatures and a pronounced daytime occurrence maximum; its comparatively homogeneous thermal structure and limited vertically developed radar echoes indicate that the 10.8 μm radiance is dominated by land-surface heating (Figure 2). We therefore interpret C1 as a daytime clear/weak-cloud regime that anchors the surface-dominated end of the manifold. C2 also peaks around midday but shows greater spatial heterogeneity and radar echoes concentrated at low levels, consistent with shallow or broken boundary-layer clouds embedded within a warm surface background. Because low cloud-top brightness temperatures can be close to those of the underlying surface, the separation of C2 is likely controlled by both spatial organization and absolute temperature. We therefore interpret C2 as a daytime low-cloud regime that retains a substantial surface-radiative contribution.

The nighttime portion of the observational manifold comprises C3–C5. In the two-dimensional visualization, C3 and especially C5 occupy relatively distinct regions, whereas C4 lies closer to the center of the manifold and overlaps more strongly with neighboring states (Figure 1). All three classes occur predominantly between evening and early morning and become nearly absent during the period of strongest daytime heating, but they differ in thermal structure and cloud organization (Figure 2). C3 is the warmest of the three and exhibits a comparatively smooth thermal field and predominantly low-to-moderately developed cloud layers, supporting its interpretation as a nighttime homogeneous cloud regime. C4 shows stronger spatial heterogeneity, consistent with scattered clouds interspersed with warmer surface contributions, and is therefore interpreted as a nighttime scattered-cloud regime. C5 is relatively rare and has markedly lower brightness temperatures than the other two, indicating stronger cloud-top control and supporting its interpretation as a nighttime cloudy regime.

C6 occupies a compact high-density core, while its broader contour overlaps with the daytime clear/weak-cloud and nighttime homogeneous- and scattered-cloud regions (Figure 1). Its occurrence peaks in the morning and evening and becomes negligible around

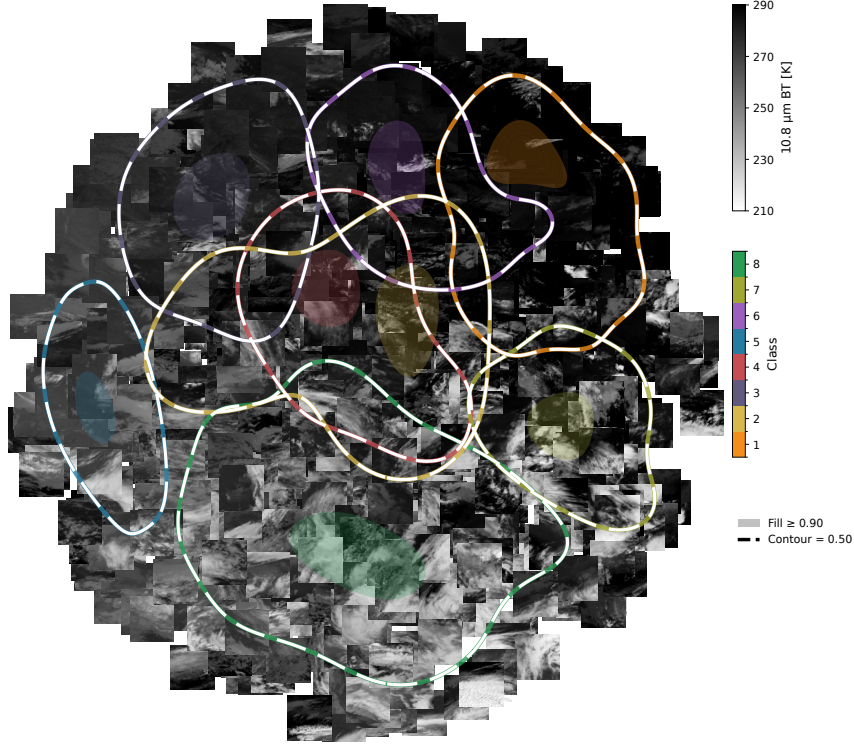


Figure 1. Observation-trained latent space of combined surface–cloud radiative states for 2024. Two-dimensional t-SNE visualization fitted to the latent representations of MSG 10.8 m brightness-temperature scenes. The eight classes are derived from the original 384-dimensional feature vectors produced by the final ViT layer; t-SNE is used only to visualize their organization in two dimensions. The embedded MSG scenes illustrate how their spatial and thermal characteristics vary across the latent space, with brightness temperature indicated by the grayscale color bar. For each class, a kernel density estimate (KDE) is calculated from the embedded MSG samples. Colored shading marks regions with densities of at least 90% of the class-wise maximum, while contours indicate the corresponding 50% levels.

midday. Its comparatively high brightness temperatures, heterogeneous satellite scenes, and predominantly low-to-moderate radar echoes indicate a mixed surface–cloud radiative signature involving shallow or broken clouds (Figure 2). This warm signature may arise from several simultaneous changes, including surface warming, decreasing cloud cover, warmer or lower cloud tops, and reduced cloud optical thickness. Because these contributions cannot be separated using the 10.8 μm channel alone, “warm transition” denotes the combined scene-level radiative signature and the regime’s intermediate position in the latent-space evolution. Over the diurnal cycle, similar scenes may evolve toward warmer surface-dominated regimes or other shallow-cloud states, depending on the development or dissipation of cloud cover. Together, its intermediate latent-space position, early- and late-day occurrence maxima, and mixed surface-cloud radiative signature support the interpretation of C6 as a warm-transition regime connecting the nighttime and daytime portions of the manifold.

The deeper-cloud portion of the observational manifold is represented by C7 and C8, which occupy neighboring but distinct regions in the two-dimensional visualization (Figure 1). C7 contains localized areas of very low brightness temperature embedded within a warmer, spatially heterogeneous background, while its radar echoes indicate substan-

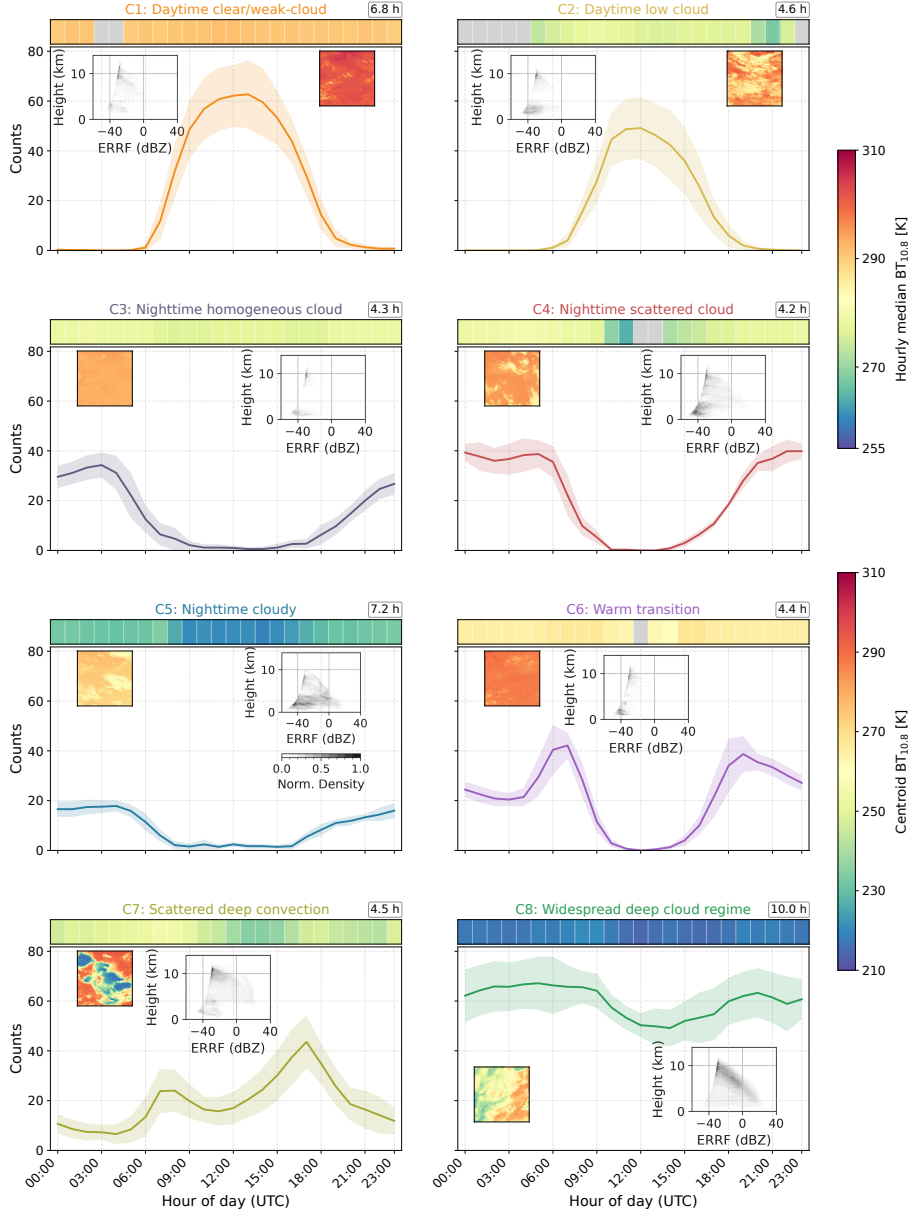


Figure 2. Physical and temporal characteristics of the eight observation-derived radiative classes. For each class, the main curve shows the mean hourly occurrence count across the seven observation sites, with shading indicating one standard deviation across sites. The colored strip above each panel shows the hourly median MSG 10.8 μm brightness temperature; light-grey cells denote class-hour combinations without valid samples. The value in the upper-right corner gives the median across sites of the site-level mean persistence duration for that class. The satellite inset shows a representative 10.8 μm brightness-temperature scene closest to the class centroid, while the Contoured Frequency by Altitude Diagram (CFAD) shows the distribution of radar reflectivity with height from the JOYRAD-35 cloud radar at JOYCE, Jülich.

tial vertical cloud development. Its diurnal occurrence profile further suggests an indirect link to the land-surface cycle, with enhanced occurrence from daytime into the afternoon, consistent with boundary-layer growth and convective triggering over land. We

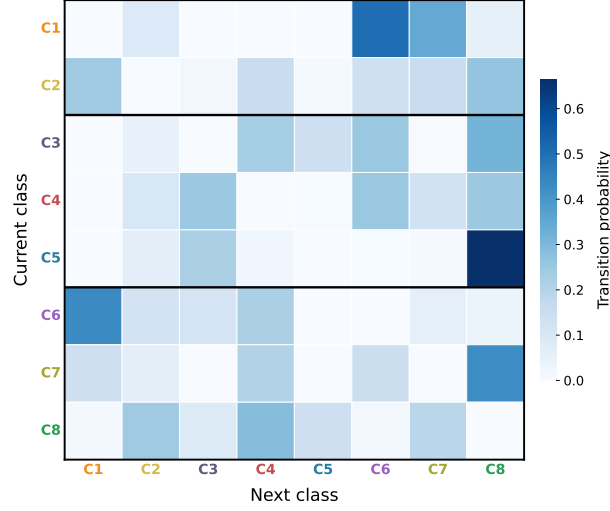


Figure 3. Observed transition pathways between the eight latent radiative classes. Rows indicate the current class and columns the next class assigned after a robust change in radiative state. Self-transitions are excluded, and each row is normalized so that the values represent the conditional probability of entering each destination class after leaving a given source class. A transition is counted only at the boundary of a continuous class run when the next observation occurs at the expected temporal interval, and the destination class persists for at least 20 minutes. This filtering reduces short-lived fluctuations in class assignment and highlights the preferred pathways through the observation-derived latent space.

therefore interpret C7 as a scattered deep-convection regime. Because the classification is based on the $10.8 \mu\text{m}$ radiative field, C7 should not be interpreted as a direct precipitation or convective-core label; rather, it identifies scenes containing localized, very cold, and vertically developed cloud structures.

In contrast, C8 exhibits persistently lower brightness temperatures, indicating that its observed radiance is primarily controlled by cloud-top emission. Its spatially extensive cold-cloud structures and vertically developed radar echoes further indicate that it is a cloud-dominated system. Moreover, its broad diurnal occurrence cycle lacks the sharp daytime maximum expected for surface-heating-driven clear-sky scenes, suggesting only a weak surface-temperature imprint (Figure 2). In a few case-day analyses using MODIS imagery together with JOYCE radar observations (not shown here), this class is frequently associated with frontal cloud systems; within the latent-space interpretation, however, we refer to it more generally as a widespread deep-cloud regime. It is the most frequent of the eight classes.

The persistence durations further distinguish stable radiative states from more transient parts of the observational manifold (Figure 2). The widespread deep-cloud regime has the longest median site-level mean persistence duration, at 10.0 h, consistent with the sustained nature of widespread organized cloud systems. The nighttime cloudy and daytime clear/weak-cloud regimes are also comparatively persistent, with durations of 7.2 and 6.8 h, respectively. These values indicate that both cloud-top-dominated nighttime scenes and strongly surface-heated daytime scenes can remain stable for extended periods. The daytime low-cloud, nighttime homogeneous-cloud, nighttime scattered-cloud, warm-transition, and scattered deep-convection regimes have shorter and relatively similar persistence durations of 4.2–4.6 h. The short residence time of the warm-transition regime is consistent with its role as a bridge between nighttime and daytime states, whereas

the short residence time of scattered deep convection reflects the more episodic character of localized deep-cloud development.

Figure 3 shows the observed non-self transition probabilities, i.e., the preferred pathways through the observational manifold. Within the warm portion of the manifold, the daytime clear/weak-cloud regime most frequently transitions toward the warm-transition regime, with a secondary pathway toward scattered deep convection. The warm-transition regime preferentially returns to daytime clear/weak cloud, supporting a strong exchange between the surface-dominated daytime state and the adjoining mixed surface-cloud state. The nighttime cloudy regime most frequently transitions toward widespread deep cloud and, secondarily, toward nighttime homogeneous cloud, indicating movement toward either a more cloud-dominated state or a related nocturnal cloud configuration. Scattered deep convection also preferentially transitions toward widespread deep cloud; this does not necessarily describe the life cycle of an individual convective system, but likely indicates that the subsequent radiative scene is commonly characterized by more spatially extensive cold cloud tops. Transitions from the daytime low-cloud regime are distributed more broadly among clear/weak-cloud, warm-transition, nighttime scattered-cloud, and widespread deep-cloud states, consistent with its more overlapping, intermediate position in the latent-space visualization. Together, the persistence and transition statistics show that the observational classes have characteristic residence times and preferred evolutionary connections, complementing their radiative, spatial, and vertical-cloud interpretations.

Taken together, these diagnostics show that the eight observation-derived classes are not isolated cloud categories, but connected surface-cloud radiative states shaped jointly by land-surface heating, cloud-top emission, spatial organization, vertical cloud structure, and diurnal timing. Relatively distinct anchor states coexist with more overlapping intermediate regimes, indicating that a scene’s location within the manifold depends on both its cloud organization and the background surface thermal state. The daytime clear/weak-cloud, nighttime cloudy, and widespread deep-cloud regimes are the most persistent, while the others are shorter-lived. Preferred transitions connect daytime clear/weak cloud with the warm-transition regime and nighttime cloudy and scattered deep convection with widespread deep cloud. Together, these characteristics provide a physically and temporally constrained reference for evaluating how ICON-EU and IFS populate and evolve within the observed manifold.

4.1.2 *Model simulations projected into the observational latent space*

Having characterized the physical meaning and evolutionary tendencies of the observational classes, we now use the observation-trained latent space as a fixed reference for evaluating the simulations. MSG-equivalent $10.8\ \mu\text{m}$ brightness-temperature fields from IFS and ICON-EU are passed through the frozen encoder and positioned within the representation learned exclusively from MSG observations. The discrete class assignments identify the observation-defined radiative regimes occupied by the scenes, whereas the underlying latent trajectories retain their gradual evolution within and between these regimes. We illustrate these complementary perspectives with a case study combining the class assignments and continuous TSEI (Section 3.3, Eq. 1).

We consider Jülich on 11 May 2024 (Fig. 4). Satellite imagery shows scattered low-level clouds, with a band of colder, higher cloud entering the area from the northwest during the day. ERA5 mean sea-level pressure and 500-hPa geopotential-height fields (Appendix C) indicate a persistent weak-gradient ridging environment, with no pronounced surface pressure system or mid-tropospheric trough crossing the region. Against this slowly varying large-scale background, the cloud field underwent substantial subdaily variability: within the Jülich-centered 128×128 -pixel domain, CLAAS-3 cloud fraction ranged from 8.5% to 47.1%, with a daily mean of 26.5%.

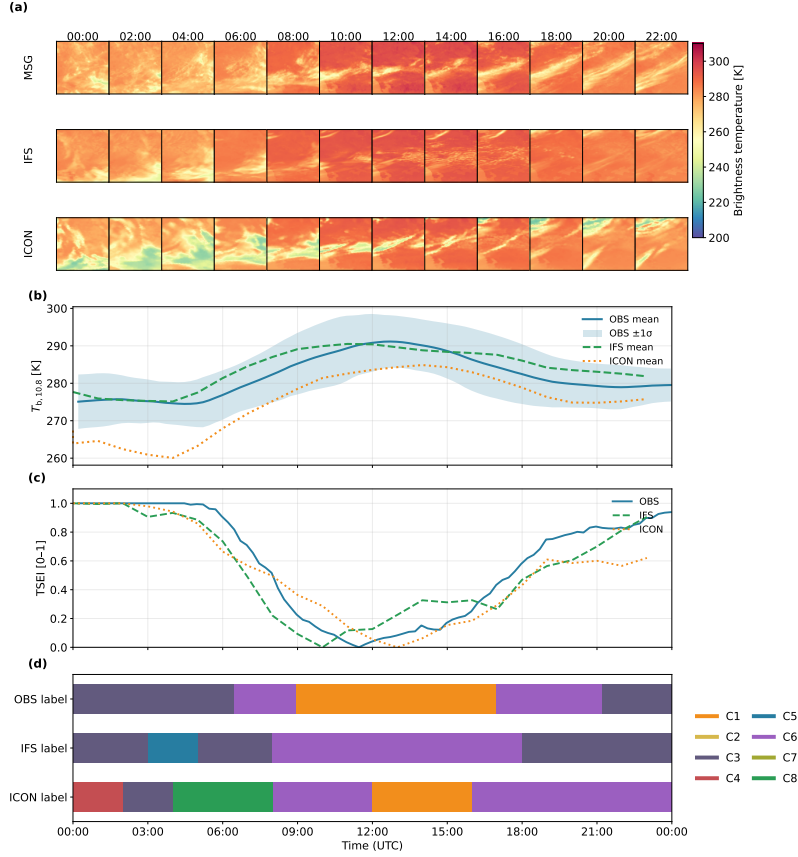


Figure 4. Continuous and discrete thermal-scene evolution for a C6-mediated transition case at Jülich on 11 May 2024. The observed pathway connects the nighttime homogeneous-cloud class (C3) and the daytime clear/weak-cloud class (C1) through the warm-transition class (C6). (a) Two-hourly 10.8 μm brightness-temperature scenes from MSG, IFS, and ICON-EU. (b) Domain-mean 10.8 μm brightness temperature, with shading indicating one spatial standard deviation for MSG. (c) Thermal-scene evolution index (TSEI), calculated as the cosine distance of each latent vector from its warm daily reference state, normalized by the cosine distance between the warm and cold daily reference states. For each data source, these reference states are selected independently from the scenes whose daily domain-mean brightness temperatures are closest to the 90th and 10th percentiles, respectively. (d) Corresponding discrete class labels assigned within the observation-trained latent space.

The case is particularly informative because the observed evolution passes through C6, the warm-transition regime linking the daytime and nighttime portions of the manifold (Section 4.1.1). Consistent with the observation-only transition matrix (Fig. 3), C6 preferentially transitions toward C1, while less frequent transitions toward C3 provide a pathway back to cooler, more homogeneous nighttime conditions. The observed sequence C3 $\rightarrow$ C6 $\rightarrow$ C1 $\rightarrow$ C6 $\rightarrow$ C3 traces a bidirectional diurnal pathway through C6: morning warming shifts scenes from nighttime C3 toward clear or weakly cloud-affected C1, while evening cooling reverses this progression. The case consequently tests whether the simulations reproduce both the warm daytime state and C6’s mediating role during its development and decay.

Read together, the continuous diagnostics show a common warming–cooling evolution but distinct timing and thermal structure (Fig. 4a–c). In MSG, following a slight early-morning decrease, the domain-mean brightness temperature rises from approximately 274 K to 290 K near midday, while the TSEI reaches the warm daily reference at approximately 11:30 UTC. The broad spatial standard-deviation envelope indicates that considerable thermal heterogeneity persists during this warm phase. IFS warms earlier, reaching a comparable domain-mean maximum and its source-specific warm reference at approximately 10:00 UTC. ICON-EU, by contrast, begins with pronounced cold-cloud structures and substantially lower temperatures before reaching its warm reference at approximately 13:00 UTC. This initial cold-cloud excess may partly reflect forecast spin-up from the 00 UTC initialization, although midlatitude cloud fields typically adjust rapidly thereafter. All three trajectories subsequently move back toward colder conditions.

The class assignments reveal that these continuous thermal cycles arise through different radiative pathways (Fig. 4d). MSG reorganizes from C6 into the clear or weakly cloud-affected C1 branch at around 09:00 UTC, remains there until late afternoon, and then returns through C6 to C3. After initially moving among colder regimes, IFS enters C6 and remains there for much of the daytime before returning to C3; despite its early warm excursion, it never develops the observed C1 spatial organization. ICON-EU initially occupies colder, more cloud-dominated regimes before progressing through C6 into C1 around midday, approximately 3 h later than MSG. Its residence in C1 is considerably shorter, and although its TSEI subsequently moves toward the cold side of the trajectory, the class assignment remains in C6 through the end of the day rather than completing the observed return to C3.

The simulations, therefore, deviate from the observed evolution in different ways. IFS reaches its warm reference C6 too early but remains there rather than entering the observed C1 branch, whereas ICON-EU reaches C1 about 3 h late, occupies it only briefly, and remains in C6 through the end of the day instead of returning to C3. The domain-mean brightness temperature describes the bulk thermal evolution, the TSEI resolves its relative phase and shape, and the class labels identify the radiative pathway through which that evolution occurs. This case thus nicely illustrates how the complementary diagnostics can reveal differences that are not evident from the bulk thermal evolution alone.

We next broaden the analysis from the individual case study to the full 2024 sample across all seven sites, evaluating the simulations’ overall representation of the observational regimes. Both weather models populate all eight observational classes and retain the broad thermal organization of the latent space (Fig. 5a). Their differences from MSG, therefore, appear primarily as redistributions among observed radiative regimes rather than as entirely separate states. The clearest shared deficit occurs along the warm daytime branch (C1, C2) (Fig. 5b). Both models reproduce the daytime emergence of the clear/weak-cloud and low-cloud regimes, but assign substantially fewer scenes to these regimes than MSG. For C1, IFS reaches its maximum earlier than observed, whereas the ICON-EU maximum occurs closer to the observed phase but remains considerably weaker. The low-cloud regime C2 is also underrepresented, most strongly in ICON-EU, while the IFS maximum occurs slightly later than in MSG.

The models therefore reproduce the broad response to daytime heating but less frequently develop the observed combination of a warm surface background and clear, weakly cloud-affected, or shallow-cloud spatial structure. The absence of C1 in IFS during the case study is thus consistent with this broader deficit. ICON-EU does reach the warm daytime branch through C1 in that case, but occupies it only briefly, as reflected in its reduced representation across the full sample.

The nighttime branch reveals a different redistribution in the two models. IFS strongly over-represents both the homogeneous-cloud regime C3 and the colder, more cloud-dominated

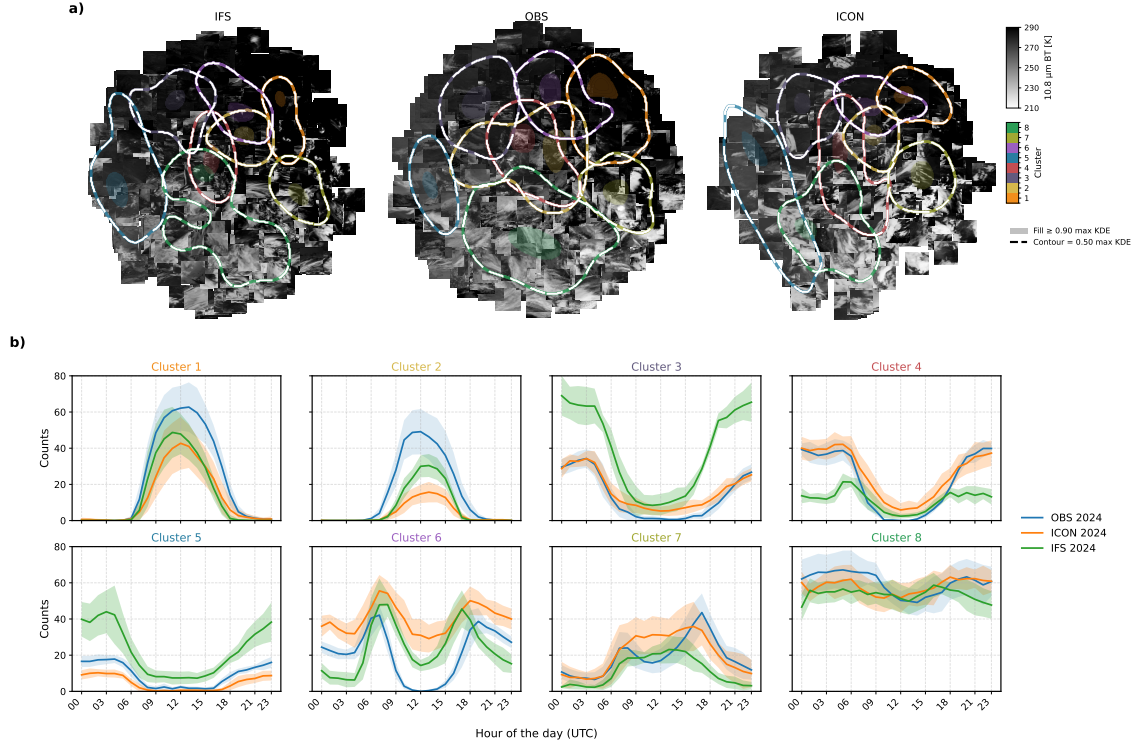


Figure 5. Model representation within the observation-trained latent space. (a) Two-dimensional t-SNE visualization fitted to MSG observations, with IFS and ICON-EU subsequently transformed into the fixed observational embedding. For each data source, kernel density estimates (KDEs) are calculated separately within each of the eight observation-derived classes. Colored shading marks regions with densities of at least 90% of the class-wise maximum, while contours indicate the 50% levels. (b) Mean hourly occurrence count of each class for MSG, ICON-EU, and IFS across the seven sites during 2024, with shading indicating one standard deviation across sites. The MSG t-SNE distribution and diurnal occurrence curves reproduce the observational references shown in Figures 1 and 2, respectively.

C5 state, particularly during the night and the morning and evening shoulders, while assigning comparatively few scenes to the scattered-cloud regime C4. ICON-EU follows the observed C3 and C4 cycles more closely, although it under-represents C5. Thus, IFS preferentially redistributes nocturnal scenes toward more homogeneous or cloud-dominated radiative states, whereas ICON-EU shows smaller departures whose directions differ across the individual nighttime classes.

The transition between the nighttime and daytime branches is also less sharply represented in the simulations. In MSG, C6 forms two relatively confined occurrence maxima during morning warming and evening cooling and becomes almost absent under the strongest midday heating. Both models retain non-negligible C6 occurrence through the middle of the day. This over-representation is especially pronounced in ICON-EU, which maintains elevated occupancy throughout the daily cycle, whereas IFS preserves the two transition-period maxima more clearly but still fails to reproduce the observed midday minimum. The prolonged C6 assignments seen in the case study, therefore, form part of a more general tendency for the simulations to occupy the intermediate warm-transition regime more broadly than observed.

The deeper-cloud branch (C7 and C8) distinguishes errors in the timing of localized development from those in the occurrence of widespread cloud systems. The observed scattered deep-convection regime C7 shows a comparatively concentrated late-afternoon enhancement. ICON-EU distributes C7 occurrence more broadly and earlier in the day, whereas IFS under-represents the regime throughout the day, including at its diurnal maximum. By contrast, all three data sources retain the broad, weakly diurnal occurrence of the widespread deep-cloud regime C8, although IFS becomes increasingly underrepresented toward the evening. The model disagreement is therefore more pronounced when scenes acquire the localized cold-cloud structures associated with C7 than when they occupy the widespread cloud-dominated C8 state.

The weather models reproduce the broad daily progression between warmer and colder radiative conditions but redistribute scenes among the observation-derived regimes through which that progression occurs. IFS is shifted toward homogeneous and cloudier nighttime states and away from warm daytime and scattered deep-cloud regimes. ICON-EU reproduces several nighttime occurrence cycles more closely, but under-populates the warm daytime branch and broadens the occupation of the transitional and scattered deep-cloud states. These occurrence cycles reveal which regimes are over- or under-represented and when, but they do not show whether a bias results from how often a regime is entered, how long it is retained, or where the scene moves after leaving it.

4.2 Evolutionary Tendencies

The analysis in Section 4.1.1 established the characteristic residence times and preferred pathways through the latent manifold for the observations. We now use these properties as a temporal reference to test whether the simulations retain each radiative state for the observed duration and, after leaving it, follow the observed state-to-state pathways.

4.2.1 Persistence

Figure 6 compares the distributions of site-level mean persistence durations for MSG, ICON-EU, and IFS. Because the observational residence times have already been discussed in Section 4.1.1, we focus here on departures of the simulations from that reference.

Across the simulations, persistence agrees most closely with MSG for the daytime clear/weak-cloud regime C1 (6.8 h) and scattered deep convection C7 (4.5 h), with the model distributions largely falling within the observed site-to-site variability (Fig. 6). The largest departures occur in the nighttime homogeneous C3, nighttime cloudy C5, and widespread deep-cloud C8 regimes, although the directions differ between the two models.

IFS over-retains C3 and C5, the same homogeneous and cloudier nighttime regimes that it over-represents in the diurnal occurrence analysis (see Fig. 5). Their enhanced occupancy, therefore, reflects not only more frequent assignment to these states but also longer uninterrupted residence within these nocturnal, cloud-dominated parts of the manifold. By contrast, IFS passes more rapidly through the daytime low-cloud regime C2 and the nighttime scattered-cloud regime C4, while the reduction in persistence for the warm-transition regime C6 is comparatively modest. The under-representation of C2 and C4 is thus accompanied by reduced temporal continuity. For C6, however, the broadened occurrence in Figure 5 is more consistent with more frequent or differently timed entry than with a substantial change in the duration of individual runs. Despite its reduced occurrence, C7 persistence remains close to MSG, indicating that its bias arises primarily from how often or when IFS reaches the regime. C8 remains relatively persis-

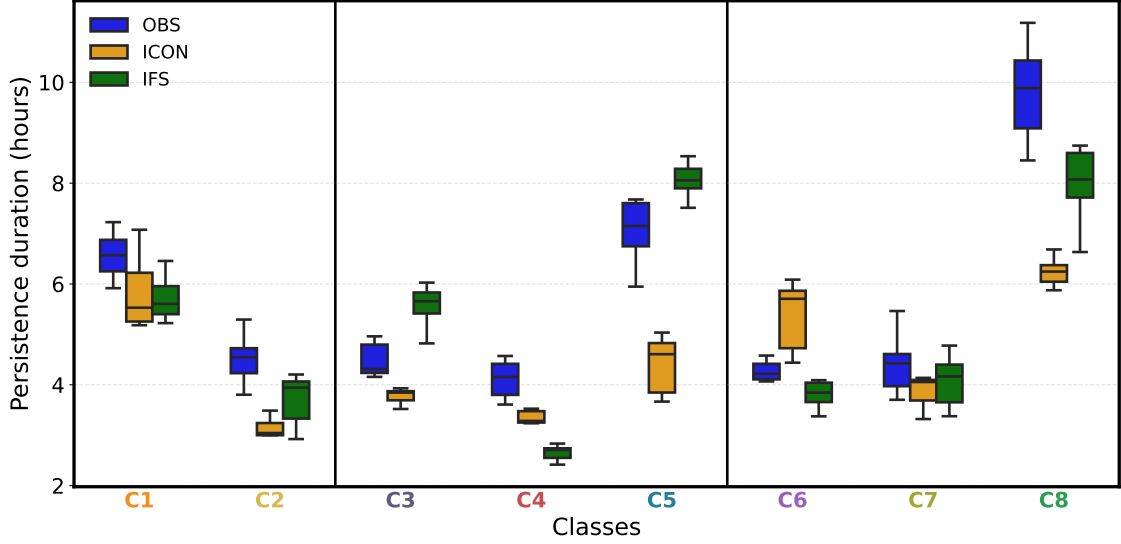


Figure 6. Persistence of the eight radiative classes in MSG observations, ICON-EU, and IFS during 2024. Persistence is the uninterrupted time a scene remains in the same class before transitioning to another class. For each data source and class, the boxplots show the distribution of site-level mean persistence durations across the seven sites, with the central line indicating the median and the box spanning the interquartile range.

tent in IFS but is shorter-lived than in MSG. So its late-day under-representation is likely accompanied by reduced temporal continuity, once it is reached.

ICON-EU exhibits a broader tendency toward shorter persistence than MSG, encompassing both warm daytime regimes C1–C2 and nighttime regimes C3–C5, as well as the widespread deep-cloud regime C8, with the largest reductions occurring in C2, C5, and C8. The shortened residence in C1 and C2 reinforces its under-representation of the warm daytime branch, while the reductions across the nighttime and widespread deep-cloud regimes indicate that these cloud-dominated states are less temporally continuous within the fixed observational partition. The principal exception is C6: ICON-EU persists considerably longer in this warm-transition regime, consistent with its elevated occurrence throughout the day (Figure 5) and showing that the broadened C6 cycle partly reflects overly long uninterrupted assignments. By contrast, the scattered deep convection regime C7 is slightly less persistent, even though its occurrence is distributed too broadly and too early in the day. Its occurrence bias is therefore primarily associated with altered entry timing or frequency rather than excessive residence duration.

Overall, the simulations exhibit contrasting persistence biases: IFS selectively over-retain homogeneous and cloudier nighttime states while passing more rapidly through several daytime and transitional regimes, whereas ICON-EU changes state more readily across most of the observational manifold but over-retain the warm-transition regime C6.

4.2.2 Transition-probability biases

Persistence describes how long a scene remains within a radiative state, but not where it moves after leaving that state. Figure 7, therefore, compares the simulated and observed non-self transition probabilities. Because the observational pathways have already been established in Figure 3, the matrices show here the percentage-point differ-

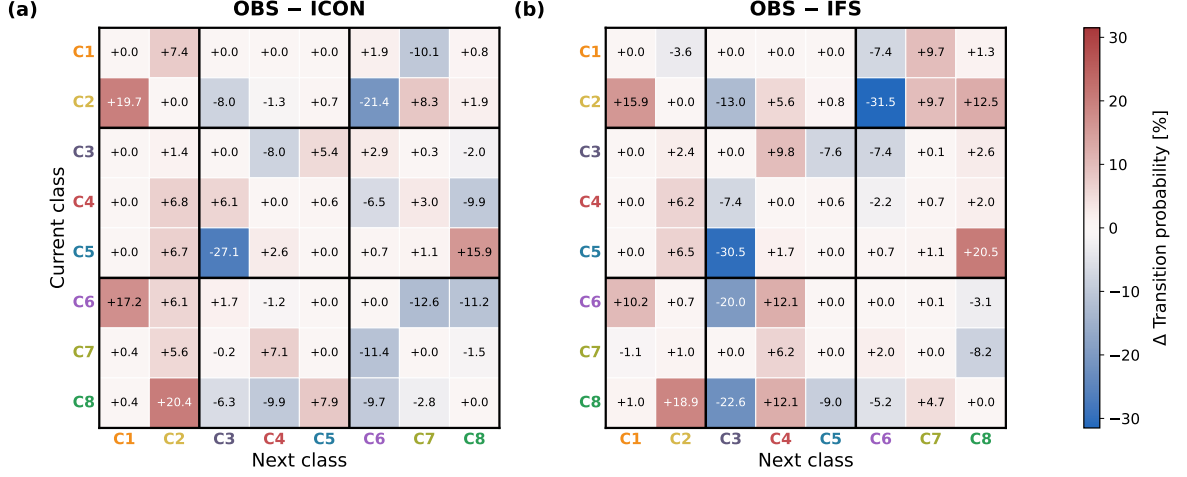


Figure 7. Differences in transition probabilities between the eight observation-derived radiative classes during 2024. The matrices show the percentage-point differences between (a) MSG and ICON-EU ($P_{\text{MSG}} - P_{\text{ICON}}$) and (b) MSG and IFS ($P_{\text{MSG}} - P_{\text{IFS}}$). Rows indicate the current class and columns the next class. Positive values indicate transitions that occur more frequently in MSG, whereas negative values denote transitions favored by the corresponding simulation. Transition probabilities are calculated using the same filtering and row normalization as described in Section 3.3.

ences [$\Delta P_{ij} = 100 (P_{ij}^{\text{MSG}} - P_{ij}^{\text{model}})$], rather than separate transition matrices for each data source.

For ICON-EU, the clearest departures center on the warm-transition regime C6. After leaving the daytime low-cloud regime C2, ICON-EU routes a substantially larger fraction of scenes toward C6 and fewer toward the clear/weak-cloud regime C1 than MSG. The exits from C6 show a complementary bias: The strong observed return to C1 is weakened, while transitions toward scattered deep convection C7 and widespread deep cloud C8 are enhanced. Thus, when scenes leave C6, ICON-EU assigns a lower transition probability to a return to the warm, surface-dominated daytime state and a higher one to evolution toward colder, more cloud-dominated regimes. Together with the enhanced persistence of C6 (Figure 6), this shows that ICON-EU both retains the transitional state too long and alters the direction of evolution after leaving it.

The excessive involvement of C6 extends beyond the C1–C2–C6 exchange. The dominant transition from scattered deep convection C7 toward widespread deep cloud C8 (Figure 3) remains broadly represented in both models, and is somewhat strengthened, particularly in IFS. Relative to MSG, ICON-EU redistributes the remaining C7 exits primarily from C2 and C4 toward C6. Thus, ICON-EU broadly captures the principal progression from scattered convection to widespread deep cloud but misrepresents how other scattered-convective scenes reconnect with the surrounding daytime and mixed-cloud regimes.

The strongest daytime pathway bias in IFS originates from C2. Once this daytime low-cloud regime ends, IFS routes scenes much more frequently toward C6 than MSG, mainly at the expense of transitions back toward C1. This clarifies the temporally broader C6 occurrence identified in the diurnal cycles: IFS populates C6 too frequently, particularly during the early and late parts of the day, even though its C6 persistence differs only modestly from MSG. The excess occupation of C6 is therefore produced mainly by

repeated or differently timed entry into the regime, especially from C2, rather than by substantially longer individual C6 episodes.

Both simulations also share a nighttime-cloud pathway bias. In MSG, scenes leaving the cloudy nighttime regime C5 preferentially evolve toward the widespread deep-cloud regime C8. ICON-EU and IFS redistribute a substantial part of this exit toward the homogeneous nighttime regime C3, with the shift being stronger in IFS. In IFS, this routing bias compounds the enhanced persistence of C3 and C5 (Fig. 6). Cloudy nighttime scenes persist longer and, after leaving C5, are more frequently redirected toward a homogeneous nocturnal state rather than evolving towards widespread deep cloud. ICON-EU shows the same directional redistribution despite its shorter C5 persistence. It therefore leaves C5 too quickly, yet, conditional on leaving, is too likely to enter C3 and insufficiently likely to enter C8.

Overall, the transition probability difference matrices between observation and weather models reveal distinct distortions of radiative-state evolution. ICON-EU involves C6 too strongly in exchanges, connecting low-cloud and scattered-convective scenes with the warm daytime and deeper-cloud branches, whereas IFS more strongly funnels daytime low-cloud scenes into C6. Both models, particularly IFS, favor the homogenization of cloudy nighttime scenes over their evolution toward widespread deep clouds. The biases, therefore, affect not only how frequently and how long regimes are populated, but also the sequence through which surface-heating-dominated and cloud-emission-dominated scenes replace one another. However, because occurrence, persistence, and transition probabilities are each calculated separately in MSG and in the simulations, they do not indicate whether the same radiative regimes are occupied simultaneously. The following regime-correspondence analysis directly examines this scene-by-scene agreement.

4.3 Regime correspondence

Figure 8 compares temporally paired MSG and simulated scenes within the observation-trained latent space. Because each row is normalized by the MSG class, diagonal values quantify whether the simulation occupies the same regime when it is observed, whereas off-diagonal values show how the simulated scene is displaced within the latent manifold.

For ICON-EU, correspondence is strongest for the warm-transition regime (C6) and the widespread deep cloud (C8), with diagonal values of 86.4% and 73.3%, respectively. The daytime clear/weak-cloud regime C1, the nighttime homogeneous- and scattered-cloud regimes C3–C4, and scattered deep convection C7 retain moderate correspondence of approximately 52–59%. The weakest correspondence occurs for daytime low cloud C2, where only 21.0% of the observed C2 scenes remain in this regime. This weak simultaneous agreement is accompanied by a pronounced deficit in C2 occurrence in Fig. 5. Although ICON-EU places its C2 maximum relatively close to the observed midday phase, the peak magnitude is substantially underestimated and disagrees more strongly with MSG than the IFS maximum does.

C6 provides the clearest organizing feature of the ICON-EU matrix: It is accurately recognized when observed, but it is also applied too broadly to other regimes. In addition to retaining 86.4% of observed C6 scenes, ICON-EU assigns 38.7% of observed C2 and 23.6% of C1. During the daytime, it therefore broadens the warm-transition signature across scenes that MSG identifies as low-cloud and surface-heating-dominated. A weaker over-representation occurs at night, when parts of the observed C3 and C4 populations are also displaced toward C6 rather than remaining within their respective homogeneous and scattered-cloud regimes. The particularly strong C2 displacement indicates that daytime low clouds over a warm background frequently lose their distinct identity and are shifted into the broader transitional branch of the manifold. This provides

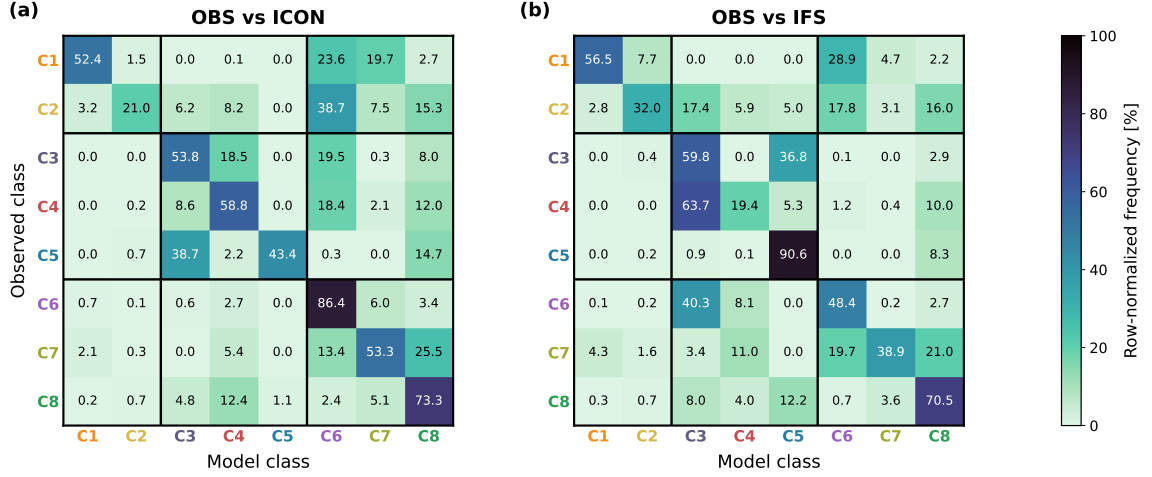


Figure 8. Regime correspondence between observations and simulations in the observation-trained latent manifold. Row-normalized confusion matrices comparing the latent-class labels of MSG observations with the corresponding labels assigned to ICON-EU and IFS simulations. Rows denote the observed coupled surface–cloud radiative state, while columns denote the simulated state for the same paired samples after projection into the fixed observation-trained manifold. Each matrix element gives the percentage of samples from a given observed state that are assigned to each simulated state, with rows summing to 100%. Diagonal values indicate agreement in radiative-state assignment, whereas off-diagonal values reveal systematic shifts toward other parts of the observation-trained latent space.

the instantaneous counterpart to the broadened C6 occurrence, enhanced C6 persistence, and excessive involvement of C6 in the transition pathways.

Within the nighttime branch, the main ICON-EU disagreement concerns the distinction between nighttime cloudy C5 and its neighboring cloud-emission-dominated regimes. Only 43.4% of observed C5 scenes remain in the nighttime cloudy regime, while 38.7% are assigned to the more homogeneous C3 regime. ICON-EU, therefore, represents many cloudy nighttime scenes at the correct time but tends to shift their spatial organization toward more homogeneous clouds.

For IFS, agreement is strongest for C5 and C8, which retain 90.6% and 70.5% of their observed scenes, respectively, while C3 and C1 show moderate correspondence at 59.8% and 56.5%. The dominant disagreement, however, reveals a pronounced homogenization of the nighttime branch: only 19.4% of observed nighttime scattered-cloud scenes remain in C4, whereas 63.7% are assigned to C3, representing the strongest off-diagonal correspondence in either matrix. Physically, IFS replaces scenes containing spatially scattered clouds and a stronger mixture of surface and cloud emission with a more spatially homogeneous nighttime radiative signature. Stratocumulus cover in IFS has been documented to be sensitive to turbulent diffusion under stable boundary-layer conditions (Sandu et al., 2013; Fitch, 2022), making excessive boundary-layer mixing a plausible contributor to this redistribution. This displacement provides a scene-by-scene explanation for the under-representation of C4 during the morning and evening shoulders in Fig. 5, as well as part of the corresponding over-representation of C3.

The attraction toward the homogeneous nighttime branch C3 also extends beyond C4. IFS assigns 40.3% of observed warmer transitional C6 scenes to C3, further populating C3. At the same time, 36.8% of observed C3 scenes are assigned to the cloudier

C5 regime. Because IFS already retains 90.6% of observed C5 scenes, this additional misclassification from C3 helps explain why C5 is over-represented. The combined displacements, therefore, erode the observed distinctions among scattered C4, homogeneous C3, cloudy C5, and transitional C6 scenes by redistributing them toward C3 and C5, consistent with the under-representation of C4 and the over-representation and enhanced persistence of the more homogeneous, cloud-dominated nighttime regimes.

On the daytime side, IFS occupies C6 during 28.9% of the scenes for which MSG has reached the clear/weak-cloud regime C1, a displacement also illustrated by the case study in Figure 4, where MSG progresses to C1 while IFS remains in C6. Scenes observed as C2 have no single dominant alternative assignment and are instead represented as warm transitional C6, homogeneous nighttime C3, or widespread deep-cloud states C8. Together with the C1-to-C6 displacement, this complements the under-representation of C2 and the previously identified strong tendency of IFS to develop daytime low-cloud evolution toward C6. Biases in the simulated land-surface temperature may also contribute to displacements among C1, C2, and C6 by altering the $10.8\ \mu\text{m}$ radiative signature of clear and partly cloudy scenes; however, their contribution cannot be isolated here.

Similarly, the weak correspondence of C2 in ICON-EU and C4 in IFS also concerns regimes occupying intermediate parts of the observational manifold (Fig. 1). Both represent mixed surface–cloud radiative states whose identity depends on the balance among surface emission, cloud fraction, and spatial organization. Relatively modest biases in these properties can therefore shift simulated scenes toward neighboring regimes. However, the resulting displacements are directional rather than diffusely distributed: ICON-EU redirects C2 primarily toward the warm-transition regime C6, whereas IFS redirects C4 overwhelmingly toward the more homogeneous nighttime regime C3. Their intermediate character may therefore make these regimes especially sensitive to model differences, while the preferred direction of displacement reveals the underlying distortion: An expansion of the warm-transition signature in ICON-EU and a homogenization of scattered nighttime cloud in IFS.

Despite their different dominant biases, both simulations show a common disagreement pattern for scattered deep convection (C7). ICON-EU assigns 25.5% of observed C7 scenes to C8 and 13.4% to C6, while the corresponding IFS fractions are 21.0% and 19.7%. Both models, therefore, blur scattered deep convection in the same two directions: toward a warmer, less-developed transitional signature or toward colder cloud that is too spatially widespread. The common displacement toward C8 also points to an organizational error, in which localized deep-convective structures are represented as a more spatially extensive cold-cloud scene.

Overall, the matrices separate errors in recognizing a regime from errors in when and how broadly it is occupied. ICON-EU generates warm transitions (C6) and the widespread deep-cloud regime C8 particularly well, but extends C6 across too much of the latent manifold and blurs the distinction between nighttime cloudy C5 and the neighboring homogeneous C3 regime. IFS preserves most observed C5 and C8 scenes but reduces the distinctions among scattered C4, homogeneous C3, cloudy C5, and transitional structures by preferentially assigning them to C3 and C5. This helps explain the C4 deficit and the C3–C5 excesses in Figure 5. Both models also struggle to preserve the distinct organization of scattered deep convection, which may partly reflect limitations in representing spatial structures approaching the models’ effective resolution (Wahl et al., 2017). Finally, although both simulations misrepresent some of the transition pathways leading into C8, their relatively high C8 correspondence shows that widespread deep-cloud scenes are generally represented correctly once they occur. Selected persistence, transition-probability, and same-time correspondence biases are synthesized into model-specific fingerprints in Fig. 9.

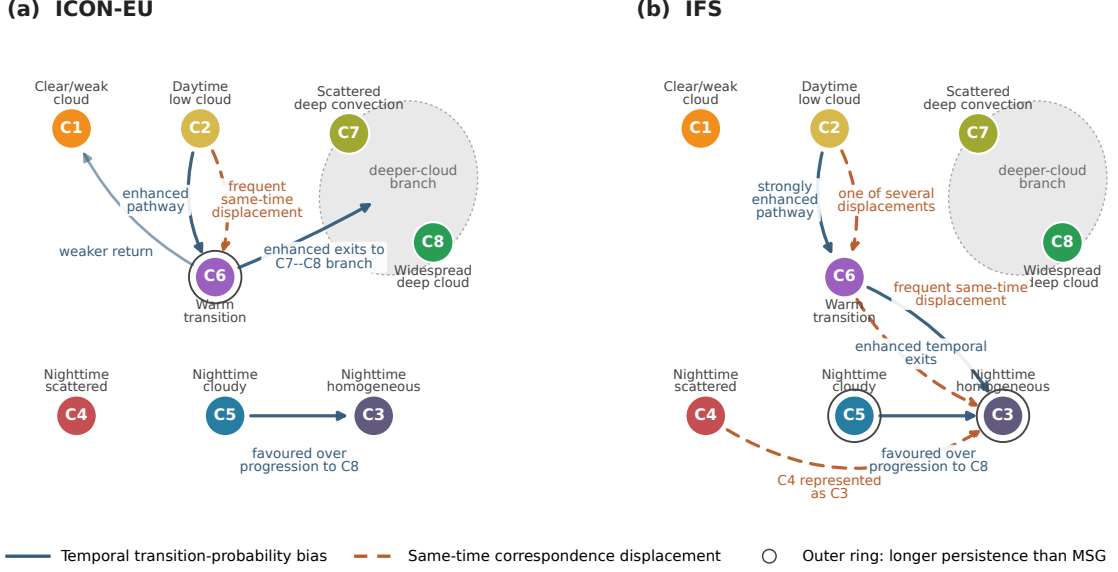


Figure 9. Schematic synthesis of selected (a) ICON-EU and (b) IFS biases relative to MSG within the observation-derived eight-regime framework. Solid blue arrows denote biases in temporal transition probabilities, dashed orange arrows indicate same-time correspondence displacements from an observed regime to the regime represented by the model, and outer rings mark longer persistence than in MSG. The gray-shaded area denotes the deeper-cloud region. Node positions represent conceptual regime relationships rather than quantitative latent-space distances, and arrow widths are not scaled to bias magnitude. Only selected diagnosed biases are shown; the absence of an arrow does not imply agreement, and no causal model process is inferred.

5 Conclusions

We constructed a novel observation-centered benchmark for evaluating km-scale atmospheric simulations by learning a latent representation directly from MSG 10.8 μm brightness-temperature imagery over Central Europe during summer. Spherical k-means initially yielded ten clusters, after which a JSD-based threshold sweep, supported by EMD/PCA diagnostics, consolidated three overlapping widespread deep-cloud clusters, resulting in eight observational regimes. The encoder and this eight-regime partition were subsequently held fixed when projecting MSG-equivalent fields from ICON-EU and IFS into the latent space. Consequently, both simulations were evaluated against the same organization of observed radiative variability rather than against separately derived model regimes. Combining regime occurrence, persistence, transition probabilities, temporally paired correspondence, and continuous latent-space evolution allowed us to examine not only which observed states the simulations produce, but also when they occur, how long they persist, and how the simulated scenes move between them.

The learned manifold describes a continuum of coupled surface–cloud radiative states rather than a classification of cloud types alone. Its eight regimes comprise a warm daytime branch comprising clear or weakly cloud-affected scenes and daytime low cloud; three nighttime states with different degrees of spatial homogeneity and cloud-top control; a warm-transition regime connecting the daytime and nighttime portions of the manifold; and scattered and widespread deep-cloud states. Their physical interpretation (Section 4.1.1) emerges jointly from brightness temperature, spatial organization, diurnal occurrence, cloud radar-derived vertical structure, persistence, and preferred transition path-

ways. Relatively distinct radiative states coexist with more overlapping intermediate regimes, particularly daytime low cloud and nighttime scattered cloud, whose identities depend sensitively on the balance among surface emission, cloud fraction, and cloud organization.

An illustrative case study over Jülich (Section 4.1.2, Fig. 4) combines the discrete regime sequence with the thermal-scene evolution index (TSEI; Eq. 1). While the regime sequence identifies the radiative branches occupied throughout the day, TSEI captures differences in the relative phase and shape of continuous thermal evolution. Together, they show that similarity in domain-mean brightness temperature or proximity to a warm daily reference does not necessarily imply agreement in mesoscale scene organization.

ICON-EU and IFS populate all eight regimes and reproduce the broad daily progression from warmer, surface-influenced conditions to colder, cloud-dominated scenes. Their principal departures stem from the distribution of this progression across the observed regimes (Section 4.1.2). Both models under-represent the clear/weak-cloud and daytime low-cloud branches and too frequently retain the warm-transition signature around midday. IFS additionally shifts nighttime occurrence toward the homogeneous and cloudier regimes at the expense of scattered nighttime cloud, produces a weaker scattered deep-convection maximum, and increasingly under-represents widespread deep cloud toward evening. ICON-EU represents the occurrence cycles of homogeneous and scattered nighttime clouds more closely, but under-represents the nighttime cloudy regime, markedly overextends the warm-transition regime across much of the day, and distributes scattered deep convection too broadly and too early.

The persistence and transition statistics (Sections 4.2.1 and 4.2.2) show that these occurrence biases arise from distinct misrepresentations of radiative-state evolution. IFS over-retains the homogeneous and cloudier nighttime regimes while passing too rapidly through daytime low cloud and scattered nighttime cloud. Its broad warm-transition occurrence instead reflects mainly repeated or differently timed entry, particularly from daytime low cloud (Fig. 7), rather than substantially longer persistence (Fig. 6). ICON-EU changes state more readily than MSG across much of the manifold but over-retains the warm-transition regime, weakens its return toward clear/weak-cloud daytime conditions, and redirects more exits toward scattered and widespread deep cloud. Both models also favor transitions from cloudy nighttime scenes toward the homogeneous nighttime regime at the expense of the observed progression toward widespread deep clouds.

Temporally paired regime correspondence (Section 4.3) further tests whether the observed and simulated regimes occur simultaneously. ICON-EU recognizes warm-transition and widespread deep-cloud scenes well, but extends the warm-transition signature too broadly across daytime and nighttime regimes. Its weakest agreement is for daytime low cloud, which is displaced mainly toward the warm-transition regime, while many nighttime cloudy scenes are represented as homogeneous cloud. IFS preserves nighttime cloudy and widespread deep-cloud scenes more faithfully, but compresses much of the nighttime manifold toward homogeneous and cloudier states. Its daytime low-cloud scenes are instead distributed among several other regimes.

The particularly weak correspondence of daytime low cloud C2 in both models and scattered nighttime cloud in IFS is consistent with their intermediate positions in the observational manifold, which may make them especially susceptible to model differences. Nevertheless, the preferred displacements are strongly directional, revealing an expansion of the warm-transition signature in ICON-EU and a homogenization of nighttime radiative structure in IFS. Both simulations also represent scenes observed as scattered deep convection, either as a less-developed transitional state or as a more spatially widespread cloud state.

The physical characterization of these regimes could be further strengthened by extending the observational representation beyond the 10.8 μm channel. Visible and near-infrared reflectances could distinguish daytime scenes with similar thermal signatures but different cloud optical properties, while additional thermal-infrared and water-vapor channels would support improved characterization throughout the diurnal cycle, including the nighttime regimes.

A further consideration is the mismatch between nominal grid spacing and effective model resolution. Numerical models generally resolve spatial structures only at scales larger than their grid spacing (Wahl et al., 2017), which may contribute to the reduced fidelity of localized cloud organization diagnosed here. Because effective resolution is variable-dependent and has not been quantified for the simulated 10.8 μm radiative fields, its contribution cannot be isolated in the present analysis. Future work should therefore test the sensitivity of the latent-space correspondence to spatial scale, for example, by comparing observations and simulations after filtering or coarse-graining to common effective resolutions.

The new framework diagnoses where these departures occur, when they emerge, and how they are connected through time, but it does not, by itself, identify the physical parameterization or process interaction responsible for them. Establishing such causal links will require targeted sensitivity experiments and additional process-level observations. A complementary extension is to expand the observational reference and its evaluation across the full MSG disk and move from individual latent trajectories to large-sample statistics.

Applying TSEI and developing related trajectory-based metrics across multiple locations, days, seasons, and weather conditions would provide distributions of the phase, excursion magnitude, pathway shape, and return timing of daily radiative evolution. Such statistics would establish whether the phase and organizational differences illustrated at Jülich persist systematically across larger spatial and temporal samples. Therefore, connecting high-frequency satellite imagery with discrete regimes, temporal pathways, paired correspondence, and continuous latent trajectories provides a systematic approach for identifying which aspects of observed mesoscale radiative evolution require improvement in km-scale atmospheric simulations.

Appendix A Training details

A1 Training samples and multi-crop transformations

We adapt the PyTorch implementation of DINO Caron et al. (2021) for geostationary satellite brightness-temperature observations. The original image loader and natural-image augmentations are replaced by a NetCDF-based loader and multi-crop transformations appropriate for single-channel $10.8 \mu\text{m}$ brightness-temperature fields. Four 128×128 -pixel scenes are randomly sampled from the study domain at each available observation time, yielding approximately 350,000 training samples.

From each training scene, two global crops of size 96×96 pixels are generated using a relative crop scale of 0.4–1.0, together with four local crops of size 64×64 pixels using a relative crop scale of 0.05–0.4. All crops are standardized using the mean and standard deviation calculated from the MSG training data:

$$x_{\text{norm}} = \frac{x - \mu_{\text{MSG}}}{\sigma_{\text{MSG}}},$$

where $\mu_{\text{MSG}} = 270.559 \text{ K}$ and $\sigma_{\text{MSG}} = 17.608 \text{ K}$.

A2 Student–teacher architecture and training objective

Both the student and teacher networks use a ViT-S/16 backbone (Dosovitskiy et al., 2021), with a patch size of 16 pixels and a stochastic-depth rate of 0.1. The backbone produces a 384-dimensional class-token representation. During training, this representation is passed through a DINO projection head with an output dimension of 2048. The projection head is discarded after training, and the 384-dimensional backbone representation is retained as the feature vector for the latent-space analysis.

The student network receives all six views of each scene, whereas the teacher receives only the two global views. For an input view x , the student and teacher output distributions are

$$p_s(x) = \text{softmax}\left(\frac{g_s(x)}{\tau_s}\right)$$

and

$$p_t(x) = \text{softmax}\left(\frac{g_t(x) - \mathbf{c}}{\tau_t}\right),$$

where (g_s) and (g_t) denote the outputs of the student and teacher projection heads, (τ_s) and (τ_t) are their respective temperatures, and $(\mathbf{c})$ is the running center of the teacher output. Subtracting the center and applying the teacher temperature, respectively, centers and sharpens the teacher distribution.

For every global view (x_g) passed to the teacher, the training objective compares its output with the student output for every other view (x_v) of the same scene. The cross-entropy for one such pair is

$$H(p_t(x_g), p_s(x_v)) = - \sum_{k=1}^{2048} p_{t,k}(x_g) \log p_{s,k}(x_v),$$

and the total loss is averaged over all cross-view pairs, excluding comparisons between identical views. The network therefore learns representations by making the stu-

dent predictions consistent with the teacher predictions across differently cropped views of the same thermal scene, without predefined class labels or explicit negative samples.

The teacher parameters are not updated through backpropagation. Instead, after each optimization step, they are updated as an exponential moving average of the student parameters,

$$\theta_t \leftarrow m\theta_t + (1 - m)\theta_s,$$

where the teacher momentum (m) increases from 0.998 to 1.0 following a cosine schedule. The center (c) is likewise updated from the batch-mean teacher output using center momenta of 0.85, 0.90, and 0.95 during epochs 0–29, 30–99, and 100–199, respectively.

A3 Optimization

The model is trained for 200 epochs using AdamW in distributed data-parallel mode across four NVIDIA V100 GPUs. A per-GPU batch size of 70 gives an effective batch size of 280. We use a base learning rate of 5×10^{-4} , linearly scaled by the effective batch size relative to a reference batch size of 256. Following a linear warm-up over the first 10 epochs, the learning rate is reduced using a cosine schedule to a minimum of 10^{-6} .

Weight decay is increased from 0.02 to 0.4 using a cosine schedule. Gradients are clipped at a maximum norm of 3.0, and the final layer of the student projection head is frozen during the first 20 epochs. The student temperature is fixed at 0.12. The teacher temperature is warmed to 0.06 during the first 70 epochs and held fixed thereafter. Mixed-precision training is disabled.

A complete 200-epoch training run requires approximately 4.5 h of wall-clock time on four V100 GPUs, corresponding to approximately 18 V100 GPU-hours.

A4 Application of the trained encoder to simulations

After training, the projection head is removed, and the teacher backbone is used as the fixed encoder. Neither its parameters nor the observation-derived class centroids are subsequently updated. For every MSG, ICON-EU, and IFS scene, the 384-dimensional backbone representation is L_2 -normalized before class assignment and distance calculation.

The MSG scenes used to train the encoder have dimensions of 128×128 pixels, whereas the IFS and ICON-EU input crops have dimensions of 100×100 and 80×80 pixels, respectively. The model crops are passed directly through the frozen encoder at these dimensions, without explicit resizing or spatial regridding. The dimensions of each input determine its patch-token grid, to which the encoder’s existing forward routine automatically adapts the learned spatial positional encodings. The class-token positional embedding is retained unchanged.

Input standardization is performed separately for each data source:

$$x_{s,\text{norm}} = \frac{x_s - \mu_s}{\sigma_s}, \quad s \in \{\text{MSG, IFS, ICON-EU}\},$$

where (μ_s) and (σ_s) are calculated independently from the corresponding data source. Thus, MSG inputs use the observational training statistics, whereas IFS and ICON-EU use separate statistics computed from their respective MSG-equivalent $10.8 \mu\text{m}$ brightness-temperature fields. The statistics used in the analysis are summarized in Table A1.

Table A1. Data-source-specific normalization statistics applied to the 10.8 μm brightness-temperature inputs before encoding.

Data source	Mean (K)	Standard deviation (K)
MSG	<i>270.559</i>	<i>17.608</i>
IFS	<i>274.90</i>	<i>17.12</i>
ICON-EU	<i>265.47</i>	<i>21.49</i>

Each normalized model scene is passed through the same frozen teacher encoder used for MSG. Its (L_2)-normalized representation is then assigned to the nearest observation-derived centroid using cosine distance. No simulated scene is used to retrain the encoder, redefine the classes, or update their centroids. Further implementation details and the scripts used for training and feature extraction are provided with the accompanying open-source code.

Appendix B Effective observational systems

The initial spherical k -means partition contains ten clusters. Because clustering divides a continuous latent manifold into discrete regions, some clusters may represent subdivisions of the same physical radiative system rather than distinct regimes. We therefore refine the initial partition using observational diagnostics that are independent of the learned 10.8 μm image embeddings. The refinement compares both the multichannel radiometric characteristics and the diurnal occurrence of the initial clusters (Fig. B1).

B1 Radiometric and temporal fingerprints

For each initial cluster i , a three-dimensional brightness-temperature histogram is constructed from collocated MSG observations in the 6.2, 7.3, and 10.8 μm channels. All clusters use common brightness-temperature bins. The histogram is normalized by the total number of samples assigned to the cluster, yielding the relative-frequency distribution

$$P_i(T_{6.2}, T_{7.3}, T_{10.8})$$

The resulting distribution describes the cluster’s joint radiometric fingerprint. It combines information from the 10.8 μm atmospheric-window channel with the two water-vapor channels, thereby helping to distinguish clusters with similar 10.8 μm appearance but different upper-tropospheric radiometric characteristics. For visualization in Fig. B1a, the full three-dimensional distribution is marginalized over the 7.3 μm dimension and displayed in the 6.2–10.8 μm plane. The similarity calculations, however, use the normalized three-channel distributions rather than the two-dimensional projections.

A temporal fingerprint is constructed independently of the class occurrence at the seven observational sites. For each initial cluster, the assigned MSG scenes are aggregated into the 96 available 15-min bins over the diurnal cycle. The occurrence counts shown in Fig. B1a retain their absolute magnitude for physical interpretation. For the similarity analysis, the counts are normalized by their sum:

$$\frac{n_i(t)}{\sum_{t=1}^{96} n_i(t)}$$

where $n_i(t)$ is the occurrence count of cluster i in time bin t . Consequently, the temporal comparison evaluates the shape and timing of the diurnal cycle without being dominated by differences in the clusters' total frequency.

B2 Pairwise similarity diagnostics

Radiometric similarity is quantified using the Jensen–Shannon divergence (JSD). For two normalized three-channel distributions P_i and P_j , their mean distribution is

$$M_{ij} = \frac{1}{2} (P_i + P_j)$$

and their Jensen-Shannon divergence is

$$\text{JSD}(P_i, P_j) = \frac{1}{2} \text{KL}(P_i \| M_{ij}) + \frac{1}{2} \text{KL}(P_j \| M_{ij})$$

where KL denotes the Kullback–Leibler divergence. JSD is symmetric and remains finite when individual histogram bins have zero probability. A value of zero indicates identical radiometric distributions, while larger values indicate increasing separation between the multichannel fingerprints.

Temporal similarity is assessed using two complementary diagnostics. First, principal component analysis (PCA) is applied jointly to the normalized diurnal profiles of the ten initial clusters. PCA describes the dominant modes through which the profiles differ in their timing and shape. Each profile is represented by its scores in the retained principal-component space, and the PCA-based separation between clusters i and j is calculated as the Euclidean distance between their score vectors:

$$\|\mathbf{a}_i - \mathbf{a}_j\|_2$$

where $\mathbf{a}_i$ and $\mathbf{a}_j$ contain the corresponding principal-component scores. Small distances indicate that the two clusters occupy similar positions with respect to the dominant modes of diurnal variability.

Second, Earth Mover's Distance (EMD) is calculated between the normalized 15-min occurrence profiles. EMD measures the displacement required to transform one temporal distribution into the other and is therefore sensitive to shifts in the timing and breadth of occurrence. Unlike a bin-by-bin difference, it assigns a smaller distance to profiles whose occurrence peaks have similar shapes but are displaced by only a small amount in time.

The PCA- and EMD-based diagnostics provide related but distinct tests. The PCA distance evaluates similarity in the dominant modes of the diurnal-cycle shapes, whereas EMD directly evaluates the temporal redistribution of occurrence. Requiring consistency between them reduces the likelihood that clusters are merged because of a single similarity measure.

B3 Threshold sweep and stopping criterion

The pairwise JSD, PCA-distance, and EMD matrices are shown in Fig. B1b. A pair of initial clusters is treated as a candidate for merging only when it satisfies the radiometric criterion and both temporal criteria:

$$\text{JSD}(P_i, P_j) \leq \tau_{\text{JSD}},$$

$$d_{\text{PCA}}(i, j) \leq \tau_{\text{PCA}}, \quad d_{\text{EMD}}(i, j) \leq \tau_{\text{EMD}},$$

where τ_{JSD} , τ_{PCA} , and τ_{EMD} denote the retained similarity thresholds. Thus, radiometric similarity alone is insufficient: a proposed merger must also be supported independently by both descriptions of diurnal occurrence.

To determine these thresholds, each metric is swept from restrictive values in the lower part of its pairwise distance distribution to the 50th percentile of that distribution (Fig. B1c). The sweep bounds are therefore defined separately for JSD, PCA distance, and EMD according to the empirical spread of their respective distance matrices. The median is used as the upper bound to explore progressively less restrictive solutions without extending into the upper half of the distributions, where comparatively dissimilar cluster pairs would become eligible for merging.

At each tested threshold combination, a pair of clusters is connected only when it simultaneously satisfies the JSD, PCA-distance, and EMD criteria. A group is accepted for merging only when every possible pair within that group satisfies all three criteria. The accepted group, therefore, forms a complete pairwise-connected system rather than being assembled through a sequence of indirect similarities.

The retained stopping point is the most restrictive tested threshold combination at which original clusters 3, 7, and 8 form such a complete group. Further relaxation introduces additional connections and produces broader mergers involving less similar clusters.

B4 Final mapping and physical interpretation

The retained refinement merges original clusters 3, 7, and 8. Their multichannel fingerprints are dominated by similarly cold, cloud-emission-controlled conditions, and their normalized diurnal profiles show comparable temporal behavior. Their separation in the initial k -means solution is therefore interpreted as an overpartitioning of the same widespread deep-cloud system rather than evidence of three independently evolving radiative regimes.

Table B1. Mapping between the ten initial spherical k -means clusters and the eight effective observation-derived classes used in the analysis.

Initial cluster	Final class	Physical interpretation
1	C5	Nighttime cloudy
2	C1	Daytime clear/weak cloud
3, 7, 8	C8	Widespread deep cloud
4	C4	Nighttime scattered cloud
5	C6	Warm transition
6	C3	Nighttime homogeneous cloud
9	C2	Daytime low cloud
10	C7	Scattered deep convection

The other initial clusters are retained separately because they fail at least one of the radiometric or temporal similarity criteria at the selected stopping point. The ten-to-eight mapping is summarized in Table B1.

1085 The final class labels are ordered by their physical interpretation within the obser-
1086 vational manifold rather than by the numerical labels returned by the initial clustering.
1087 After applying the mapping, the eight effective classes and their observation-derived cen-
1088 troids define the fixed reference partition used for MSG, ICON-EU, and IFS. No sim-
1089 ulated information enters either the refinement procedure or the definition of these classes.

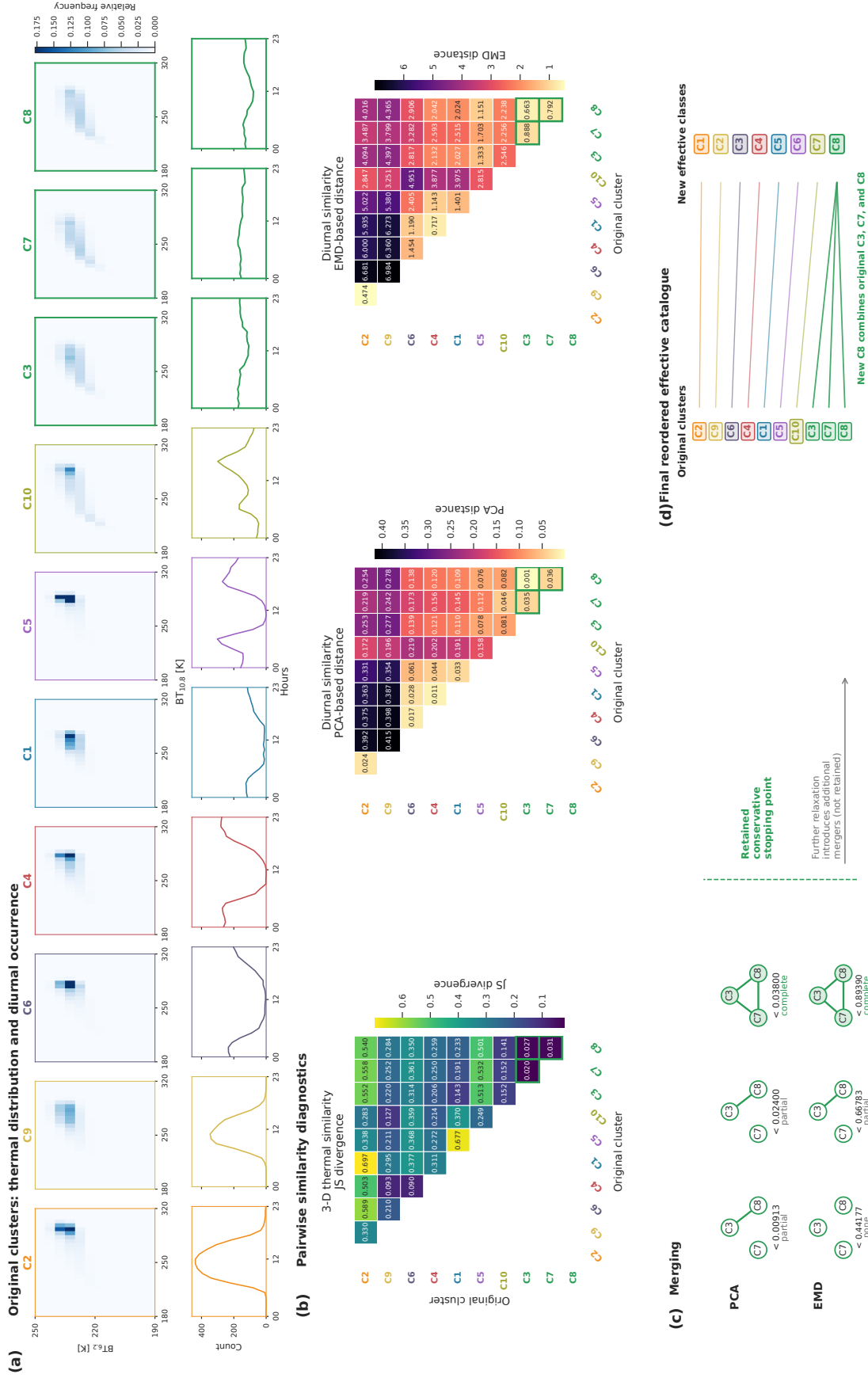


Figure B1. Observation-based refinement of the latent-space partition from ten initial clusters to eight effective radiative systems. (a) Radiometric and temporal fingerprints of the ten initial clusters. The upper row shows class-wise relative-frequency projections of the three-channel MSG brightness-temperature histograms constructed from the 6.2, 7.3, and 10.8 μm channels. For visualization, the full three-dimensional histogram is shown through its 6.2–10.8 μm marginal projection. The lower row shows the corresponding diurnal occurrence counts aggregated over the seven observational sites. Highlighted panels identify original clusters 3, 7, and 8, which constitute the candidate group considered for conservative merging. (b) Pairwise diagnostics used to assess cluster separability. The left matrix shows the Jensen–Shannon divergence between the candidate group considered for conservative merging. The middle and right matrices show the separation of the normalized diurnal occurrence profiles using PCA-based distance and Earth Mover’s Distance, respectively. Outlined cells identify cluster pairs satisfying the retained similarity criteria. (c) Threshold sweep for the temporal similarity criteria applied together with the radiometric Jensen–Shannon threshold. The retained stopping point is the first stage at which the candidate group forms a complete pairwise-connected system and is therefore not retained. (d) Final mapping from the ten initial clusters to eight effective observational systems. Original clusters 3, 7, and 8 are merged into C8, the widespread deep-cloud regime, because they are consistently similar in both their three-channel brightness-temperature fingerprints and their diurnal occurrence structure. All other initial clusters remain distinct.

Appendix C Case study: synoptic environment

The case-study day was characterized by weak and slowly evolving large-scale forcing (Fig. C1). Throughout 11 May 2024, Jülich remained within a broad ridging pattern with weak surface pressure gradients, and neither a pronounced surface pressure system nor a mid-tropospheric trough crossed the region. The substantial subdaily cloud variability examined in the case study was therefore developed without a major change in the synoptic environment.

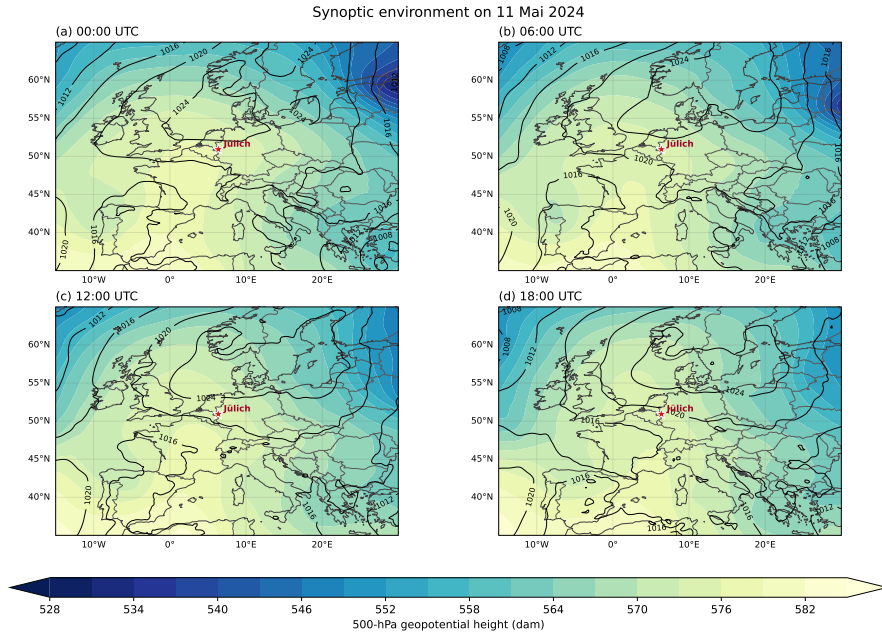


Figure C1. ERA5 synoptic conditions over Europe on 11 May 2024 at 00, 06, 12, and 18 UTC. Shading denotes 500-hPa geopotential height (dam; 1 dam=10 m), and black contours denote mean sea-level pressure (hPa), with a contour interval of 4 hPa. The red star marks Jülich.

Conflict of Interest declaration

The authors declare there are no conflicts of interest for this manuscript.

Acknowledgments

We would like to thank Dr. Vera Schemann from the University of Cologne for providing the ICON-EU data from the German Meteorological Service database. This work was partly supported by the Helmholtz research program “Changing Earth” (<https://earthenvironment.helmholtz.de/changing-earth/program/>) and the WarmWorld project (<https://www.warmworld.de/>), funded by the German Federal Ministry of Education and Research (BMBF) under grant 01LK2202B. The Jülich Observatory for Cloud Evolution (JOYCE) is part of the Aerosol, Clouds and Trace Gases Research Infrastructure (ACTRIS) European Research Infrastructure Consortium (ERIC) and is supported by BMBF under grant numbers 01LK2001G and 01LK2002F.

Open Research Section

The Meteosat Second Generation (MSG) SEVIRI Level-1.5 observations used in this study are available through the EUMETSAT Data Store as the *High Rate SEVIRI Level 1.5 Image Data–MSG–0 degree* collection, with the collection identifier `EO:EUM:DAT:MSG:HRSEVIRI` (European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), 2009). The data were retrieved using version 3.1.1 of the EUMETSAT Data Access Client (EUMDAC; (European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), 2025)).

The MSG-equivalent ICON-EU data analyzed in this study, the trained encoder weights, and the code used for model training, latent-feature extraction, spherical k -means partitioning, and model evaluation are archived together in version 1.0 of the accompanying Zenodo repository (Chatterjee, Knippertz, et al., 2026). The archive is available at <https://doi.org/10.5281/zenodo.21625606>.

The MSG-equivalent IFS data remain stored in the European Centre for Medium-Range Weather Forecasts (ECMWF) Meteorological Archival and Retrieval System (MARS) under research experiment `i4q1` in class `rd` (European Centre for Medium-Range Weather Forecasts (ECMWF), 2024). The experiment can be located at <https://apps.ecmwf.int/mars-catalogue/?class=rd&expver=i4q1>. It is not part of the official Destination Earth data portfolio or ECMWF’s open real-time data catalogue. Access is subject to ECMWF eligibility and redistribution restrictions. In particular, users in the national meteorological services of ECMWF Member and Co-operating States may retrieve these data from MARS but may not redistribute them outside their institutions. The IFS data used in this study are therefore not included in the accompanying Zenodo archive.

The JOYRAD-35 radar-reflectivity observations used in this study are available from the ACTRIS Cloud Remote Sensing Data Centre Unit as a custom collection covering Jülich from 1 April to 30 September 2024 (Pfizenmaier, 2026). The corresponding Cloud-net classification products, from which cloud-base height was obtained, are available as a separate custom collection covering the same site and period (Marke et al., 2026). Both collections are distributed under the Creative Commons Attribution 4.0 licence.

The ERA5 mean-sea-level pressure and 500-hPa geopotential fields used to characterize the case-study synoptic conditions are available from the Copernicus Climate Change Service Climate Data Store as the ERA5 hourly single-level (Hersbach et al., 2023b) and pressure-level (Hersbach et al., 2023a) products, respectively. The instantaneous Level-2 binary cloud-mask product used to quantify the case-study cloud fraction is available as part of the third edition of the CM SAF Cloud property dAtAset using SEVIRI (CLAAS-3; (Meirink et al., 2022)).

Use of generative artificial intelligence tools

OpenAI ChatGPT was used during manuscript preparation to assist with language editing and limited code review. All suggested text changes were independently checked and revised by the authors, who take full responsibility for the final manuscript.

References

- Bechtold, P., Forbes, R., Sandu, I., Lang, S., & Ahlgrim, M. (2020). A major moist physics upgrade for the IFS. *ECMWF Newsletter*(164), 24–32. Retrieved from <https://www.ecmwf.int/en/newsletter/164/meteorology/major-moist-physics-upgrade-ifs> doi: 10.21957/3gt59vx1pb
- Benas, N., Solodovnik, I., Stengel, M., Hüser, L., Karlsson, K.-G., Håkansson, N., ... Meirink, J. F. (2023). Claas-3: the third edition of the cm saf cloud data record based on seviri observations. *Earth System Science Data*, 15(11), 5153–

- 1157 5170. Retrieved from [https://essd.copernicus.org/articles/15/5153/](https://essd.copernicus.org/articles/15/5153/2023/)
1158 2023/ doi: 10.5194/essd-15-5153-2023
- 1159 Bony, S., Colman, R., Kattsov, V. M., Allan, R. P., Bretherton, C. S., Dufresne,
1160 J.-L., ... Webb, M. J. (2006). How well do we understand and evaluate cli-
1161 mate change feedback processes? *Journal of Climate*, 19(15), 3445 - 3482.
1162 Retrieved from [https://journals.ametsoc.org/view/journals/clim/19/](https://journals.ametsoc.org/view/journals/clim/19/15/jcli3819.1.xml)
1163 15/jcli3819.1.xml doi: 10.1175/JCLI3819.1
- 1164 Bony, S., Dufresne, J., Treut, H. L., Morcrette, J. J., & Senior, C. A. (2004). On dy-
1165 namic and thermodynamic components of cloud changes. *Climate Dynamics*,
1166 22, 71-86. Retrieved from [https://api.semanticscholar.org/CorpusID:](https://api.semanticscholar.org/CorpusID:56077074)
1167 56077074
- 1168 Brueck, M., Hohenegger, C., & Stevens, B. (2020). Mesoscale marine tropical pre-
1169 cipitation varies independently from the spatial arrangement of its convective
1170 cells. *Quarterly Journal of the Royal Meteorological Society*, 146(728), 1391-
1171 1402. Retrieved from [https://rmets.onlinelibrary.wiley.com/doi/abs/](https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3742)
1172 10.1002/qj.3742 doi: <https://doi.org/10.1002/qj.3742>
- 1173 Bruno, O., Shaw, J. K., Storelvmo, T., & Hoose, C. (2026). Evaluation of global
1174 climate and storm-resolving model representations of mixed-phase clouds
1175 and their hemispheric contrasts. *Atmosphere*, 17(2). Retrieved from
1176 <https://www.mdpi.com/2073-4433/17/2/156> doi: 10.3390/atmos17020156
- 1177 Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin,
1178 A. (2021). *Emerging properties in self-supervised vision transformers*.
- 1179 Chatterjee, D., Acquitapache, C., Deneke, H., & Crewell, S. (2023). Understanding
1180 cloud systems structure and organization using a machine’s self-learning ap-
1181 proach. *Artificial Intelligence for the Earth Systems*. Retrieved from [https://](https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-22-0096.1/AIES-D-22-0096.1.xml)
1182 [journals.ametsoc.org/view/journals/aies/aop/AIES-D-22-0096.1/](https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-22-0096.1/AIES-D-22-0096.1.xml)
1183 [AIES-D-22-0096.1.xml](https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-22-0096.1/AIES-D-22-0096.1.xml) doi: <https://doi.org/10.1175/AIES-D-22-0096.1>
- 1184 Chatterjee, D., Knippertz, P., Raabe, N., Crewell, S., Vannière, B., & Düben, P.
1185 (2026, July). *Code for training the self-supervised deep learning model, its*
1186 *final checkpoint, and ICON-EU forward-simulated BT forecast data used*
1187 *in the study “satellite-trained latent spaces for evaluating km-scale atmo-*
1188 *spheric simulations”*. Zenodo. Retrieved from [https://doi.org/10.5281/](https://doi.org/10.5281/zenodo.21625606)
1189 [zenodo.21625606](https://doi.org/10.5281/zenodo.21625606) doi: 10.5281/zenodo.21625606
- 1190 Chatterjee, D., Raabe, N., & Crewell, S. (2026, apr). Four low-level cloud regimes
1191 revealed by latent space analysis and their impact on solar energy variability.
1192 *Machine Learning: Earth*, 2(1), 015015. Retrieved from [https://doi.org/](https://doi.org/10.1088/3049-4753/ae4e30)
1193 [10.1088/3049-4753/ae4e30](https://doi.org/10.1088/3049-4753/ae4e30) doi: 10.1088/3049-4753/ae4e30
- 1194 Chatterjee, D., Schnitt, S., Bigalke, P., Acquitapache, C., & Susanne, C. (2024).
1195 Capturing the diversity of mesoscale trade wind cumuli using complementary
1196 approaches from self-supervised deep learning. *Geophysical Research Letters*.
1197 doi: 10.1029/2024GL108889
- 1198 Corradini, D., Cattani, E., Bigalke, P., Chatterjee, D., & Acquistapace, C. (2026,
1199 August). Identifying cloud regimes and their transitions related to extreme
1200 precipitation events over the alps through self-supervised learning. *Preprints*.
1201 Retrieved from <https://doi.org/10.20944/preprints202608.0098.v1> doi:
1202 10.20944/preprints202608.0098.v1
- 1203 Davis, I., & Medeiros, B. (2024). Assessing CESM2 clouds and their response to cli-
1204 mate change using cloud regimes. *Journal of Climate*, 37(10), 2965–2985. doi:
1205 10.1175/JCLI-D-23-0337.1
- 1206 Denby, L. (2020). Discovering the Importance of Mesoscale Cloud Organization
1207 Through Unsupervised Classification. *Geophysical Research Letters*, 47(1),
1208 e2019GL085190. doi: 10.1029/2019GL085190
- 1209 Denby, L. (2023). *Charting the realms of mesoscale cloud organisation using unsu-*
1210 *perervised learning*. doi: 10.48550/arXiv.2309.08567
- 1211 Derrien, M., & Le Gléau, H. (2005). Msg/seviri cloud mask and type from safnwc.

- 1212 *International Journal of Remote Sensing*, 26(21), 4707–4732.
- 1213 Dillerup, I., Lemburg, A., Buschow, S., & Pinto, J. G. (2026). Dynamical system
1214 metrics and weather regimes explain the seasonally-varying link between eu-
1215 ropean heatwaves and the large-scale atmospheric circulation. *Earth System*
1216 *Dynamics*, 17(2), 265–289. Retrieved from [https://esd.copernicus.org/](https://esd.copernicus.org/articles/17/265/2026/)
1217 [articles/17/265/2026/](https://esd.copernicus.org/articles/17/265/2026/) doi: 10.5194/esd-17-265-2026
- 1218 Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner,
1219 T., ... Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for*
1220 *image recognition at scale*.
- 1221 Eastman, R., & Warren, S. G. (2014). Diurnal cycles of cumulus, cumulonim-
1222 bus, stratus, stratocumulus, and fog from surface observations over land and
1223 ocean. *Journal of Climate*, 27(6), 2386 - 2404. Retrieved from [https://](https://journals.ametsoc.org/view/journals/clim/27/6/jcli-d-13-00352.1.xml)
1224 journals.ametsoc.org/view/journals/clim/27/6/jcli-d-13-00352.1.xml
1225 doi: 10.1175/JCLI-D-13-00352.1
- 1226 Eikenberg, S., Köhler, C., Seifert, A., & Crewell, S. (2015). How microphysical
1227 choices affect simulated infrared brightness temperatures. *Atmospheric Re-*
1228 *search*, 156, 67–79.
- 1229 European Centre for Medium-Range Weather Forecasts (ECMWF). (2024). *Inte-*
1230 *grated Forecasting System cycle 48r1 research experiment i4ql*. ECMWF Mete-
1231 orological Archival and Retrieval System (MARS). Retrieved from [https://](https://apps.ecmwf.int/mars-catalogue/?class=rd&expver=i4ql)
1232 apps.ecmwf.int/mars-catalogue/?class=rd&expver=i4ql (Restricted data
1233 set; class rd, experiment i4ql)
- 1234 European Organisation for the Exploitation of Meteorological Satellites (EU-
1235 METSAT). (2009). *High rate SEVIRI level 1.5 image data-MSG-0*
1236 *degree*. EUMETSAT. Retrieved from [https://data.eumetsat.int/](https://data.eumetsat.int/product/EO%3AEUM%3ADAT%3AMSG%3AHRSEVIRI)
1237 [product/EO%3AEUM%3ADAT%3AMSG%3AHRSEVIRI](https://data.eumetsat.int/product/EO%3AEUM%3ADAT%3AMSG%3AHRSEVIRI) (Data set, collection
1238 EO:EUM:DAT:MSG:HRSEVIRI)
- 1239 European Organisation for the Exploitation of Meteorological Satellites (EUMET-
1240 SAT). (2025). *EUMDAC: EUMETSAT data access client*. EUMETSAT.
1241 Retrieved from <https://pypi.org/project/eumdac/3.1.1/> (Software,
1242 version 3.1.1)
- 1243 Eyre, J. (1991, 03/1991). *A fast radiative transfer model for satellite sounding sys-*
1244 *tems* (No. 176). Shinfield Park, Reading: ECMWF. Retrieved from [https://](https://www.ecmwf.int/node/9329)
1245 www.ecmwf.int/node/9329 doi: 10.21957/xsg8d92y3
- 1246 Fitch, A. C. (2022). Improving stratocumulus cloud turbulence and entrainment
1247 parametrizations in OpenIFS. *Quarterly Journal of the Royal Meteorological*
1248 *Society*, 148(745), 1782–1804. doi: 10.1002/qj.4278
- 1249 Geiss, A., Christensen, M., Varble, A., Yuan, T., & Song, H. (2023, 11). Self super-
1250 vised cloud classification. *Artificial Intelligence for the Earth Systems*, 3. doi:
1251 10.1175/AIES-D-23-0036.1
- 1252 Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for
1253 climate models. *Journal of Geophysical Research: Atmospheres*, 113(D6).
1254 Retrieved from [https://agupubs.onlinelibrary.wiley.com/doi/abs/](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007JD008972)
1255 [10.1029/2007JD008972](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007JD008972) doi: <https://doi.org/10.1029/2007JD008972>
- 1256 Gordon, N., Norris, J., Weaver, C., & Klein, S. (2005). Cluster analysis of
1257 cloud regimes and characteristic dynamics of midlatitude synoptic systems
1258 in observations and a model. *Journal of Geophysical Research*, 110. doi:
1259 10.1029/2004JD005027
- 1260 Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B., & Gruber, C. (2011).
1261 The integrated nowcasting through comprehensive analysis (inca) system and
1262 its validation over the eastern alpine region. *Weather and Forecasting*, 26(2),
1263 166–183.
- 1264 Hamann, U., Walther, A., Baum, B., Bennartz, R., Bugliaro, L., Derrien, M., ...
1265 Wind, G. (2014). Remote sensing of cloud top pressure/height from seviri:
1266 analysis of ten current retrieval algorithms. *Atmospheric Measurement Tech-*

- 1267 *niques*, 7(9), 2839–2867. Retrieved from [https://amt.copernicus.org/](https://amt.copernicus.org/articles/7/2839/2014/)
1268 [articles/7/2839/2014/](https://amt.copernicus.org/articles/7/2839/2014/) doi: 10.5194/amt-7-2839-2014
- 1269 Happ, L., Hendricks, S., Albrecht, C., Kaleschke, L., Patil, S., Fellegara, R.,
1270 & Lorenz, D. (2026, 01). *Advancing sea ice surface classification by*
1271 *self-supervised contrastive learning for radar altimetry.* doi: 10.36227/
1272 [techrxiv.176823164.49511515/v1](https://arxiv.org/abs/176823164)
- 1273 Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J.,
1274 ... Thépaut, J.-N. (2023a). *ERA5 hourly data on pressure levels from 1940*
1275 *to present.* Copernicus Climate Change Service Climate Data Store. Re-
1276 trieved from <https://doi.org/10.24381/cds.bd0915c6> (Data set) doi:
1277 10.24381/cds.bd0915c6
- 1278 Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J.,
1279 ... Thépaut, J.-N. (2023b). *ERA5 hourly data on single levels from 1940*
1280 *to present.* Copernicus Climate Change Service Climate Data Store. Re-
1281 trieved from <https://doi.org/10.24381/cds.adbb2d47> (Data set) doi:
1282 10.24381/cds.adbb2d47
- 1283 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater,
1284 J., ... Thépaut, J.-N. (2020). The era5 global reanalysis. *Quarterly Jour-*
1285 *nal of the Royal Meteorological Society*, 146(730), 1999–2049. Retrieved from
1286 <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803> doi:
1287 <https://doi.org/10.1002/qj.3803>
- 1288 Hoffmann, J., Bauer, P., Sandu, I., Wedi, N., Geenen, T., & Thiemert, D. (2023).
1289 Destination earth – a digital twin in support of climate services. *Climate*
1290 *Services*, 30, 100394. Retrieved from [https://www.sciencedirect.com/](https://www.sciencedirect.com/science/article/pii/S2405880723000559)
1291 [science/article/pii/S2405880723000559](https://www.sciencedirect.com/science/article/pii/S2405880723000559) doi: [https://doi.org/10.1016/](https://doi.org/10.1016/j.cliser.2023.100394)
1292 [j.cliser.2023.100394](https://doi.org/10.1016/j.cliser.2023.100394)
- 1293 Hohenegger, C., Korn, P., Linardakis, L., Redler, R., Schnur, R., Adamidis,
1294 P., ... Stevens, B. (2023). Icon-sapphire: simulating the components
1295 of the earth system and their interactions at kilometer and subkilome-
1296 ter scales. *Geoscientific Model Development*, 16(2), 779–811. Retrieved
1297 from <https://gmd.copernicus.org/articles/16/779/2023/> doi:
1298 10.5194/gmd-16-779-2023
- 1299 Hornik, K., Feinerer, I., Kober, M., & Buchta, C. (2012). Spherical k-means clus-
1300 tering. *Journal of Statistical Software*, 50(10), 1–22. Retrieved from [https://](https://www.jstatsoft.org/index.php/jss/article/view/v050i10)
1301 www.jstatsoft.org/index.php/jss/article/view/v050i10 doi: 10.18637/
1302 [jss.v050.i10](https://www.jstatsoft.org/index.php/jss/article/view/v050i10)
- 1303 Höjgård-Olsen, E., Brogniez, H., & Chepfer, H. (2020). Observed evolution of the
1304 tropical atmospheric water cycle with sea surface temperature. *Journal of Cli-*
1305 *mate*, 33(9), 3449 - 3470. Retrieved from [https://journals.ametsoc.org/](https://journals.ametsoc.org/view/journals/clim/33/9/jcli-d-19-0468.1.xml)
1306 [view/journals/clim/33/9/jcli-d-19-0468.1.xml](https://journals.ametsoc.org/view/journals/clim/33/9/jcli-d-19-0468.1.xml) doi: 10.1175/JCLI-D-19-
1307 -0468.1
- 1308 Illingworth, A., Hogan, R., O'Connor, E., Bouniol, D., Brooks, M., Delanoë, J., ...
1309 Wrench, C. (2007, 06). Cloudnet continuous evaluation of cloud profiles in
1310 seven operational models using ground-based observations. *Bulletin of the*
1311 *American Meteorological Society*, 88, 883–898.
- 1312 IPCC Core Writing Team, Pachauri, R., & Reisinger, A. e. (2007). *Climate change*
1313 *2007: Synthesis report. contribution of working groups i, ii and iii to the fourth*
1314 *assessment report of the intergovernmental panel on climate change.* Geneva,
1315 Switzerland: IPCC.
- 1316 Kaps, A., Lauer, A., Camps-Valls, G., Gentine, P., Gómez-Chova, L., & Eyring, V.
1317 (2023). Machine-learned cloud classes from satellite data for process-oriented
1318 climate model evaluation. *IEEE Transactions on Geoscience and Remote*
1319 *Sensing*, 61, 1–15. (Art. no. 4100515) doi: 10.1109/TGRS.2023.3237008
- 1320 Kaps, A., Lauer, A., Kazeroni, R., Stengel, M., & Eyring, V. (2024). Char-
1321 acterizing clouds with the CCCLim dataset, a machine learning cloud

- class climatology. *Earth System Science Data*, 16(6), 3001–3016. doi: 10.5194/essd-16-3001-2024
- Koenig, M., & de Coning, E. (2009). The msg global instability indices product and its use as a nowcasting tool. *Weather and Forecasting*, 24(1), 272–285. Retrieved from <https://journals.ametsoc.org/view/journals/wefo/24/1/2008waf2222141.1.xml> doi: 10.1175/2008WAF2222141.1
- Kuma, P., Bender, F. A.-M., Schuddeboom, A., McDonald, A. J., & Seland, Ø. (2023). Machine learning of cloud types in satellite observations and climate models. *Atmospheric Chemistry and Physics*, 23(1), 523–549. doi: 10.5194/acp-23-523-2023
- Lensky, I. M., & Rosenfeld, D. (2008). Clouds-aerosols-precipitation satellite analysis tool (capsat). *Atmospheric Chemistry and Physics*, 8(22), 6739–6753. Retrieved from <https://acp.copernicus.org/articles/8/6739/2008/> doi: 10.5194/acp-8-6739-2008
- Li, Z., Li, J., Li, Y., Zhang, Y., Schmit, T. J., Zhou, L., ... Menzel, W. P. (2012). Determining diurnal variations of land surface emissivity from geostationary satellites. *Journal of Geophysical Research: Atmospheres*, 117(D23). Retrieved from <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012JD018279> doi: <https://doi.org/10.1029/2012JD018279>
- Löhnert, U., Schween, J. H., Acquistapace, C., Ebell, K., Maahn, M., Barrera-Verdejo, M., ... Crewell, S. (2015). Joyce: Jülich observatory for cloud evolution. *Bulletin of the American Meteorological Society*, 96(7), 1157–1174. Retrieved from <https://journals.ametsoc.org/view/journals/bams/96/7/bams-d-14-00105.1.xml> doi: 10.1175/BAMS-D-14-00105.1
- Marke, T., Pfitzenmaier, L., Pospichal, B., Schween, J. H., & O'Connor, E. J. (2026). Custom collection of classification data from jülich between 1 apr and 30 sep 2024. ACTRIS Cloud Remote Sensing Data Centre Unit (CLU). Retrieved from <https://doi.org/10.60656/cfb4d8e2823945ed> (Data set; licensed under CC BY 4.0) doi: 10.60656/cfb4d8e2823945ed
- Medeiros, B., Shaw, J., Kay, J. E., & Davis, I. (2023). Assessing clouds using satellite observations through three generations of global atmosphere models. *Earth and Space Science*, 10(7), e2023EA002918. doi: 10.1029/2023EA002918
- Meirink, J. F., Karlsson, K.-G., Solodovnik, I., Hüser, I., Benas, N., Johansson, E., ... Hollmann, R. (2022). CLAAS-3: CM SAF CLOUD property dAtAset using SEVIRI-edition 3. Satellite Application Facility on Climate Monitoring. Retrieved from https://doi.org/10.5676/EUM_SAF_CM/CLAAS/V003 (Data set) doi: 10.5676/EUM_SAF_CM/CLAAS/V003
- Mooers, G., Pritchard, M., Beucler, T., Srivastava, P., Mangipudi, H., Peng, L., ... Mandt, S. (2023, Dec 15). Comparing storm resolving models and climates via unsupervised machine learning. *Scientific Reports*, 13(1), 22365. Retrieved from <https://doi.org/10.1038/s41598-023-49455-w> doi: 10.1038/s41598-023-49455-w
- Murphy, J., Sexton, D., Barnett, D., Jones, G., Webb, M., Collins, M., & Stainforth, D. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, 430, 768–772.
- Nordling, K., Keskinen, J.-P., Romakkaniemi, S., Kokkola, H., Räisänen, P., Lipponen, A., ... Raatikainen, T. (2024). Technical note: Emulation of a large-eddy simulator for stratocumulus clouds in a general circulation model. *Atmospheric Chemistry and Physics*, 24(2), 869–890. Retrieved from <https://acp.copernicus.org/articles/24/869/2024/> doi: 10.5194/acp-24-869-2024
- Oware, R. K. T., Zheng, Y., Bogenschütz, P., Zhang, Y., Ma, H.-Y., Xie, S., & Tao, C. (2025). Evaluating e3sm global storm-resolving model simulations of deep convection: Insights from dp-scream during tracer. *Journal of Geophysical Research: Atmospheres*, 130(20), e2025JD044113. Retrieved from <https://>

- agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025JD044113
(e2025JD044113 2025JD044113) doi: <https://doi.org/10.1029/2025JD044113>
- Pfizenmaier, L. (2026). *Custom collection of radar data from jülich between 1 apr and 30 sep 2024*. ACTRIS Cloud Remote Sensing Data Centre Unit (CLU). Retrieved from <https://doi.org/10.60656/cf703d4c1e2346f7> (Data set; licensed under CC BY 4.0) doi: 10.60656/cf703d4c1e2346f7
- Pincus, R., Batstone, C. P., Hofmann, R. J. P., Taylor, K. E., & Glecker, P. J. (2008). Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *Journal of Geophysical Research: Atmospheres*, 113(D14). Retrieved from <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007JD009334> doi: <https://doi.org/10.1029/2007JD009334>
- Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., ... others (2015). A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Reviews of geophysics*, 53(2), 323–361.
- Prill, F., Reinert, D., Rieger, D., & Zängl, G. (2024). *ICON tutorial: Working with the ICON model* (Tech. Rep.). Offenbach am Main, Germany: Deutscher Wetterdienst. Retrieved from https://doi.org/10.5676/DWD_pub/nwv/icon_tutorial2024 doi: 10.5676/DWD_pub/nwv/icon_tutorial2024
- Rackow, T., Pedruzo-Bagazgoitia, X., Becker, T., Milinski, S., Sandu, I., Aguridan, R., ... Ziemen, F. (2025). Multi-year simulations at kilometre scale with the integrated forecasting system coupled to fesom2.5 and nemov3.4. *Geoscientific Model Development*, 18(1), 33–69. Retrieved from <https://gmd.copernicus.org/articles/18/33/2025/> doi: 10.5194/gmd-18-33-2025
- Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., ... Taylor, K. E. (2007). Climate models and their evaluation. In S. Solomon et al. (Eds.), *Climate change 2007: The physical science basis* (pp. 589–662). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.
- Sandu, I., Beljaars, A., Bechtold, P., Mauritsen, T., & Balsamo, G. (2013). Why is it so difficult to represent stably stratified conditions in numerical weather prediction (NWP) models? *Journal of Advances in Modeling Earth Systems*, 5(2), 117–133. doi: 10.1002/jame.20013
- Saunders, R., Hocking, J., Turner, E., Rayer, P., Rundle, D., Brunel, P., & et al. (2018). An update on the rrtov fast radiative transfer model (currently at version 12). *Geoscientific Model Development*, 11(7), 2717–2737. doi: 10.5194/gmd-11-2717-2018
- Schmetz, J., Pili, P., Tjemkes, S., Just, D., Kerkmann, J., Rota, S., & Ratier, A. (2002). An introduction to meteosat second generation (msg). *Bulletin of the American Meteorological Society*, 83(7), 977 - 992. Retrieved from https://journals.ametsoc.org/view/journals/bams/83/7/1520-0477_2002.083.0977_aitmsg_2.3_co_2.xml doi: 10.1175/1520-0477(2002)083<0977:AITMSG>2.3.CO;2
- Sobrino, J., & Romaguera, M. (2004). Land surface temperature retrieval from msg1-seviri data. *Remote Sensing of Environment*, 92(2), 247-254. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0034425704001889> doi: <https://doi.org/10.1016/j.rse.2004.06.009>
- Sokol, Z., Brožková, R., Popová, J., Bobotová, G., & Švábik, F. (2022). Evaluation of aladin nwp model forecasts by ir10.8 μm and wv06.2 μm brightness temperatures measured by the geostationary satellite meteosat second generation. *Atmospheric Research*, 265, 105920. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0169809521004762> doi: 10.1016/j.atmosres.2021.105920
- Stephens. (2005). Cloud feedbacks in the climate system: A critical review. *Journal of Climate*, 18(2), 237 - 273. Retrieved from <https://journals.ametsoc>

- 1432 .org/view/journals/clim/18/2/jcli-3243.1.xml doi: 10.1175/JCLI-3243
- 1433 .1
- 1434 Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., ...
- 1435 Zhou, L. (2019, Sep 30). Dyamond: the dynamics of the atmospheric gen-
- 1436 eral circulation modeled on non-hydrostatic domains. *Progress in Earth and*
- 1437 *Planetary Science*, 6(1), 61. Retrieved from <https://doi.org/10.1186/s40645-019-0304-z> doi: 10.1186/s40645-019-0304-z
- 1438
- 1439 Taszarek, M., Brooks, H. E., Czernecki, B., Szuster, P., & Fortuniak, K. (2018).
- 1440 Climatological aspects of convective parameters over europe: A comparison
- 1441 of era-interim and sounding data. *Journal of Climate*, 31(11), 4281 - 4308.
- 1442 Retrieved from [https://journals.ametsoc.org/view/journals/clim/31/](https://journals.ametsoc.org/view/journals/clim/31/11/jcli-d-17-0596.1.xml)
- 1443 [11/jcli-d-17-0596.1.xml](https://journals.ametsoc.org/view/journals/clim/31/11/jcli-d-17-0596.1.xml) doi: 10.1175/JCLI-D-17-0596.1
- 1444 Tiedtke, M. (1989). A comprehensive mass flux scheme for cumulus parameteriza-
- 1445 tion in large-scale models. *Monthly Weather Review*, 117(8), 1779–1800. doi:
- 1446 10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2
- 1447 Trigo, I. F., Peres, L. F., DaCamara, C. C., & Freitas, S. C. (2008). Thermal
- 1448 land surface emissivity retrieved from sevir/meteosat. *IEEE Transac-*
- 1449 *tions on Geoscience and Remote Sensing*, 46(2), 307-315. doi: 10.1109/
- 1450 TGRS.2007.905197
- 1451 Tselioudis, G., Zhang, Y., & Rossow, W. B. (2000). Cloud and radiation variations
- 1452 associated with northern midlatitude low and high sea level pressure regimes.
- 1453 *Journal of Climate*, 13(2), 312–327.
- 1454 van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal*
- 1455 *of Machine Learning Research*, 9(86), 2579–2605. Retrieved from [http://jmlr](http://jmlr.org/papers/v9/vandermaaten08a.html)
- 1456 [.org/papers/v9/vandermaaten08a.html](http://jmlr.org/papers/v9/vandermaaten08a.html)
- 1457 Wahl, S., Bollmeyer, C., Crewell, S., Figura, C., Friederichs, P., Hense, A., ...
- 1458 Ohlwein, C. (2017). A novel convective-scale regional reanalysis cosmo-rea2:
- 1459 Improving the representation of precipitation.
- 1460 Williams, K. D., & Webb, M. J. (2008, August). A quantitative performance as-
- 1461 sessment of cloud regimes in climate models. *Climate Dynamics*, 33(1), 141–
- 1462 157. Retrieved from <https://doi.org/10.1007/s00382-008-0443-1> doi: 10
- 1463 .1007/s00382-008-0443-1
- 1464 Zängl, G., Reinert, D., & Prill, F. (2022). Grid refinement in icon v2. 6.4. *Geoscientific*
- 1465 *Model Development Discussions*, 2022, 1–32.
- 1466 Zhang, C., Ge, J., Zhang, J., Liu, B., Su, J., Lan, W., ... Qin, Z. (2025). Disen-
- 1467 tangling bimodal and unimodal patterns in the diurnal cycle of low clouds
- 1468 and their meteorological drivers at sacol. *Journal of Geophysical Re-*
- 1469 *search: Atmospheres*, 130(16), e2025JD044125. Retrieved from [https://](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025JD044125)
- 1470 agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025JD044125
- 1471 (e2025JD044125 2025JD044125) doi: <https://doi.org/10.1029/2025JD044125>
- 1472 Zängl, G., Reinert, D., Rípodas, P., & Baldauf, M. (2015). The icon (icosa-
- 1473 hedral non-hydrostatic) modelling framework of dwd and mpi-m: De-
- 1474 scription of the non-hydrostatic dynamical core. *Quarterly Journal of*
- 1475 *the Royal Meteorological Society*, 141(687), 563-579. Retrieved from
- 1476 <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2378>
- 1477 doi: <https://doi.org/10.1002/qj.2378>