Colocated MTG - EarthCARE ML-ready dataset for deriving the 3D structure of the atmosphere

Clouds are a critical component of the Earth's climate system, acting as the main modulator of energy and radiation, with a major influence on climate and meteorology. Their complex 3D structure and dynamics pose serious observational and modelling challenges. The recently launched Earth Cloud, Aerosol and Radiation Explorer (EarthCARE) satellite mission carries passive and active instruments that give information on both the horizontal and vertical structure of the atmosphere and was specifically designed to provide improved observations and synergistic products that address these challenges. At the same time, the Meteosat Third Generation (MTG) Flexible Combined Imager (FCI) provides geostationary multispectral observations over Europe, Africa and the surrounding regions in high temporal resolution (every 10 minutes).

In recent years, numerous studies have sought to bridge the gap between active sensors that capture detailed vertical cloud properties but sample sparsely in time (e.g., CALIPSO and CloudSat, with a 16-day repeat cycle, and now EarthCARE, with a 25-day repeat cycle) and geostationary imagers that observe clouds continuously yet resolve only their top (e.g., MSG, Himawari, GOES). This effort is driven by the growing need to incorporate 3D cloud information into meteorological forecasting, energy-yield estimation, radiative transfer modelling, and related applications.

In this work, we introduce the first ML-ready dataset of spatially and temporally colocated MTG-EarthCARE scenes of clouds, designed to bridge these complementary observational capabilities and allow for further atmospheric and machine learning studies. The dataset covers the full 2025 period and includes approximately 1.9 million MTG pretraining scenes and 167,407 finetuning patches across three temporal colocation windows (2, 3, and 4 minutes), over a geographic extent limited to ±45° around the MTG subsatellite point to minimize parallax effects. Each sample consists of fixed-size 128×128 patches of MTG FCI multispectral bands along with the MTG FCI cloud mask and the EarthCARE vertical profile of cloud types and cloud microphysical properties corresponding to the overpass. The dataset is provided in three different versions, each accounting for a different maximum temporal colocation window, enabling the impact of temporal alignment on downstream ML tasks to be systematically investigated. The dataset also supports applications such as vertically resolved retrievals of cloud types and microphysical properties from passive geostationary sensor observations. To establish initial benchmarks, we provide ML baselines for all dataset versions on these tasks, evaluated using F1 score and IoU for cloud type classification and MSE, SSIM, and PSNR for microphysical property retrieval. Beyond these retrieval tasks, the dataset enables applications such as cloud representation learning and the investigation of the relationship between two-dimensional radiances and the underlying vertical cloud structure. Code and datasets will be made publicly available on GitHub.

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