CERBERUS: A THREE-HEADED DECODER FOR VERTICAL CLOUD PROFILES

Emily K. de Jong, Nipun Gunawardena, Kevin Smalley, Hassan Beydoun, & Peter Caldwell
Lawrence Livermore National Laboratory
Livermore, CA 94550 USA
{dejong5, gunawardena1, smalley5, beydoun1, caldwell19}@llnl.gov

ABSTRACT

Atmospheric clouds exhibit complex three-dimensional structure and microphysical details that are poorly constrained by the predominantly two-dimensional satellite observations available at global scales. This mismatch complicates data-driven learning and evaluation of cloud processes in weather and climate models, contributing to ongoing uncertainty in atmospheric physics. We introduce CERBERUS, a probabilistic inference framework for generating vertical radar reflectivity profiles from geostationary satellite brightness temperatures, near-surface meteorological variables, and temporal context. CERBERUS employs a three-headed encoder–decoder architecture to predict a zero-inflated (ZI) vertically-resolved distribution of radar reflectivity. Trained and evaluated using ground-based Ka-band radar observations at the ARM Southern Great Plains site, CERBERUS recovers coherent structures across cloud regimes, generalizes to withheld test periods, and provides uncertainty estimates that reflect physical ambiguity, particularly in multilayer and dynamically complex clouds. These results demonstrate the value of distribution-based learning targets for bridging observational scales, introducing a path toward model-relevant synthetic observations of clouds.

1 INTRODUCTION

Atmospheric cloud processes occur at scales that are poorly resolved by a majority of observations available for model learning and validation. The satellite record is extensive in spatiotemporal coverage but primarily consists of 2D top-of-atmosphere perspectives. Meanwhile, vertically-resolved measurements of cloud properties are confined to sparse ground-based sites and *in situ* aircraft measurements (Lamb et al., 2026). This scale mismatch is one reason why clouds continue to drive uncertainty in both weather and climate predictions (Boucher et al., 2013; Morrison et al., 2020).

Recent work has leveraged the polar-orbiting radar CloudSat as a target for conditionally-generated cloud structures using GANs (Leinonen et al., 2019), U-Nets (Brüning et al., 2024), and masked-autoencoders (MAEs) (Girtsou et al., 2025; Ermis et al., 2025). However, currently these approaches target only daytime clouds and may be deterministic, leaving ground-based measurements unexploited and uncertainty unquantified. This work introduces CERBERUS (**C**loud **E**stimation with vertically-**R**esolved **Be**ta-distributed **R**etrievals of **U**ncertainty and **S**tructure), a probabilistic data-driven framework for inferring vertical radar reflectivity conditioned on both space-based imagery and near-surface meteorological context. CERBERUS uses three prediction heads to output the probability of reflective cloud and two parameters of the reflectivity distribution at each altitude. This work illustrates CERBERUS at the Atmospheric Radiation Measurement (ARM) site in Oklahoma, USA, demonstrating scalable estimation and uncertainty quantification of cloud vertical profiles that can facilitate atmospheric model evaluation and calibration.

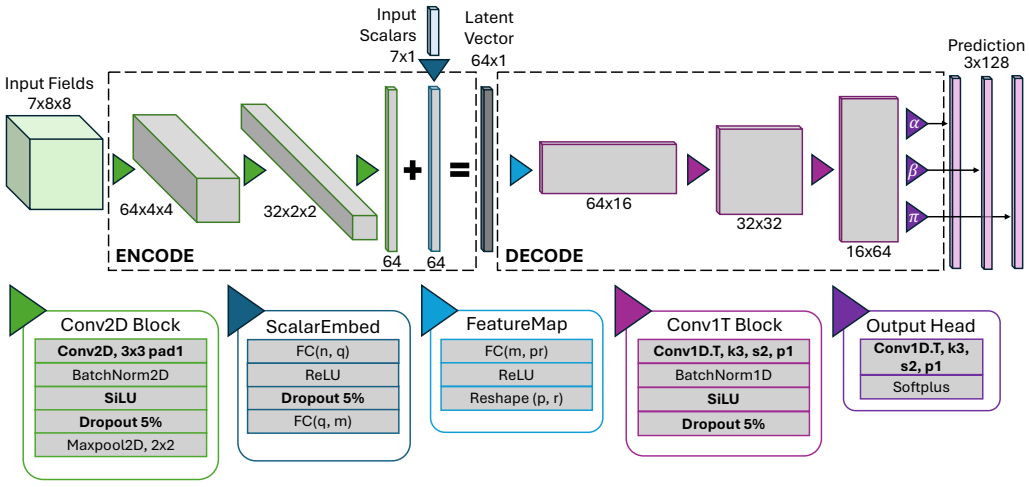

Figure 1: Illustration of the CERBERUS three-headed architecture and neural network operations. Bold text indicates hyperparameters that were optimized.

2 Data & Methods

2.1 Dataset & Preprocessing

Target data: Reflectivities and cloud-top products from the ARM Ka-band zenith-pointing radar (KAZR) at the Southern Great Plains (SGP) site are collected from January 2020–March 2025 (KAZ; Kollias et al., 2007). Data are quality-controlled based on signal-to-noise, resampled to 5-minute averages, and interpolated to 128 equispaced altitudes from 160 m to 15 km and smoothed with a Savitzky-Golay filter (Savitzky & Golay, 1964) with order 3 and window length 50.

Input data: Inputs to the inference model include 30-minutely 2D brightness temperatures from the GOES-16 satellite at 13.3 μm, 11.9 μm (SW), 11.2 μm (IR), 8.4 μm, 6.8 μm, and 3.9 μm (SIR), plus the visible reflectance (0.65 μm) (GOE). These seven fields are remapped to an 8×8 grid at $0.02°$-resolution centered at the ARM SGP site. Cloud-top-height (CTH) is retrieved using the VISST algorithm from the same datastream to confirm consistency of observed clouds between the radiometer (GOES) and radar (KAZR) measurements. In addition we consider five near-surface meteorological variables from hourly MERRA-2 reanalysis (MER) sampled at the SGP KAZR site: 10 m temperature (T), winds (u, v), and relative humidity (rh), as well as surface pressure (P0). Time-of-day and day-of-year are sine-encoded to account for seasonality and diurnal cycle, but no positional encoding is included as this experiment targets a single location.

Data selection: To ensure consistency and detectability in the observed clouds between the radar reflectivity target and the satellite fields, we impose 4 filters on the 30 min paired KAZR-GOES retrievals used for training and evaluation (see section A.1, Figure A.1). We utilize an 80/20 training/validation random data split over 2020–2024 data (18,000 target-profile pairs), reserving the remaining 3 months of data from winter 2025 for testing (1600 pairs).

Normalization: All data utilize a min-max normalization. Non-cloudy values of the normalized reflectivity target dataset are zero-filled, corresponding to a detection threshold of -60 dBZ. This leads to a zero-inflated (ZI) target distribution with all values between 0 and 1 inclusive (Figure A.2), motivating the choice of a zero-inflated beta distribution (ZIB) to model these data.

2.2 The CERBERUS Model Structure & Training

The structure of the 3-headed CERBERUS encoder-decoder architecture is illustrated in Figure 1. The brightness temperature fields are encoded via convolutions and then summed with embedded near-surface scalars in a FiLM-like approach (Perez et al., 2017). The decoder then projects, reshapes, and transforms the resulting latent vector, with the final layer utilizing three separate con-

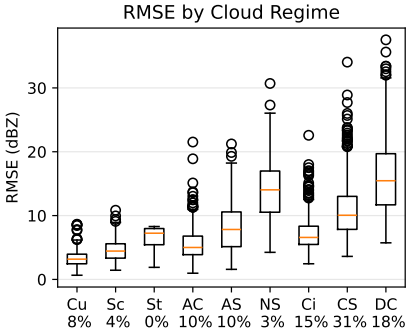

Figure 2: Statistics of per-sample RMSE (dBZ) for each cloud regime (see Table A.1) over the validation set: median (orange), interquartile-range (IQR; box), $Q_1/Q_3 \pm 1.5IQR$ (whiskers), and outliers (circles). Regime abbreviations defined in Table A.1.

volutional output heads to predict the three parameters of a zero-inflated beta (ZIB) reflectivity distribution at each of the 128 target altitudes. Hyperparameters including dropout, CNN activation function, and kernel size were optimized (Akiba et al., 2019).

CERBERUS has similarities to related UNet and SatMAE models for reconstructing 3D clouds (Girtsou et al., 2025; Ermis et al., 2025; Brüning et al., 2024), such as mapping 2D satellite fields to a latent space. However, the three-headed decoder of CERBERUS allows for simultaneous learning of uncertainty in the reflectivity profiles, unlike prior deterministic approaches. A comparison against deterministic and non-ZI baselines (section A.3, Figures A.3 and A.4) supports the added value of this three-headed probabilistic structure. While previous models used RMSE as their training objective, this probabilistic approach uses the negative log-likelihood of the observation y in the predicted distribution:

$$\mathcal{L}(y, (\alpha, \beta, \pi)) = \begin{cases} -\log(\pi), & y = 0 \\ -\log\big(1 - \pi - \mathcal{B}[\alpha, \beta](y)\big), & y > 0 \end{cases} \tag{1}$$

where π is the predicted probability of non-cloudiness and $\mathcal{B}[\alpha, \beta]$ is the beta distribution with predicted parameters α, β. We add a small scalar $\epsilon = 10^{-3}$ to all $\log$-arguments for stability. Model weights are trained using the Adam optimizer with batch size 100 and initial learning rate 10^{-3} for a maximum of 50 epochs, selecting the model with the smallest validation loss for evaluation (Figure A.5).

3 Results

CERBERUS demonstrates robust performance across both conditional classification of cloudy altitudes (ROC-AUC=0.957) and regression ($R^2 > 0.6$; Figure A.6). IR and near-IR brightness temperatures contribute most to model accuracy, with the visible reflectance (only available during daytime measurements) being the least important GOES field (Figure A.7). Out of the scalar conditions, the 10 m temperature contributes most to model performance and is indicative of thermodynamic and boundary layer characteristics. Near-surface winds and humidity, by contrast, may be redundant with information already embedded in the IR observations.

Error between measured reflectivity and theß mean predicted reflectivity vary across cloud regimes, with predictions of low and thin clouds attaining the lowest RMSE, and larger RMSE in deep clouds (Figure 2). Among the test set (Figure 3), CERBERUS displays the most confident and correct predictions of stratiform and low clouds (bottom row), but consistently struggles to predict complex multilayer clouds (top left). These results mirror the performance of SatMAE in Girtsou et al. (2025): nimbostratus and the prevalent deep convective clouds over the SGP are most challenging to predict. By weighting RMSE according to cloud regime prevalence in the European geostationary satellite record (Girtsou et al., 2025), we find equivalent or improved performance (depending on the metric) across cloudy scenes relative to previous 3D cloud predictions (Table A.1).

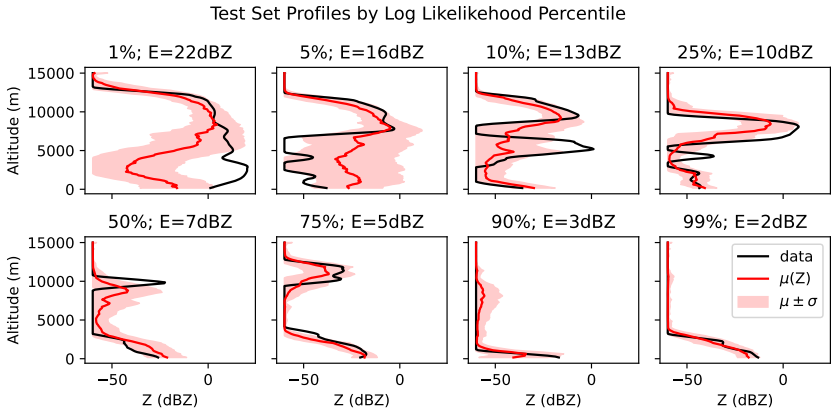

Figure 3: Test data and predicted reflectivity mean and uncertainty as a function of altitude, ranked according to quantile using altitude-scale energy (analogous to RMSE; see section A.2).

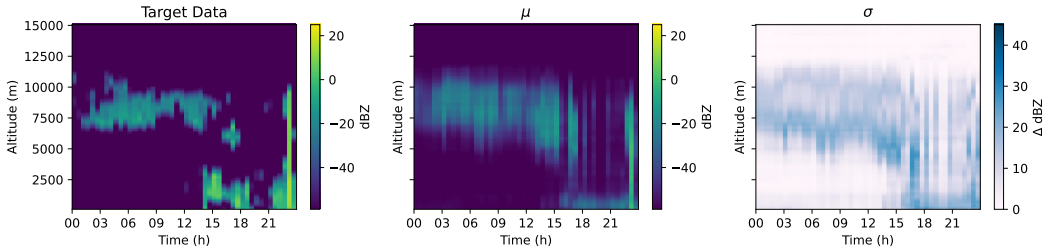

Figure 4: Illustration of the time-evolving reflectivity on Jan 29, 2025 (test set): true KAZR data (far left); and CERBERUS model mean and uncertainty (panels 2-3).

Time-resolved composites (Figures 3, A.8, A.9) reveal that CERBERUS captures coherent cloud evolution across the test period, for instance capturing the development of a decoupled precipitating and later convective cloud beneath an initial anvil on Jan 29, 2025. Satellite measurements often saturate in the upper cloud layer, providing limited conditioning on the decoupled clouds underneath. However, CERBERUS does tend to predict a broader reflectivity distribution at these challenging altitudes (model spread in Figure 3 top left; right panel in Figure 4), indicating that learned uncertainty reflects physically meaningful ambiguity. As in (Brüning et al., 2024), the ZIB mean predictions exhibit overly smooth cloud boundaries, but the distribution spread predicted by CERBERUS add value by indicating these structural uncertainties, with larger spread at and below cloud base and in multilayer clouds, and with less uncertainty in the convective core.

4 Conclusions & Future Work

CERBERUS uses a three-headed encoder-decoder architecture to produce probabilistic estimates of vertically-resolved cloud reflectivity conditioned on 2D satellite fields and near-surface meteorological variables. Despite its simplicity and 1D-profile predictions, CERBERUS produces coherent reflectivity fields with uncertainty estimates that reflect the non-uniqueness of mapping a 2D satellite image to a 3D cloud. Future work will extend this framework toward predictions of model-relevant cloud microphysical quantities such as cloud water content and droplet size distributions, incorporating additional data from both ground-based Doppler radar as well as from global high-resolution models (Donahue et al., 2024, e.g.). We anticipate that this extension to global microphysical data will necessitate more expressive architectures such as mixture-density predictions (Bishop, 1994) or transformer-based encoding Cong et al. (2023), as well as additional conditional inputs like location embedding or broader satellite horizontal context.

ACKNOWLEDGMENTS

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory (LLNL) under Contract DE-AC52-07NA27344 and supported by the Laboratory Directed Research and Development Program (LDRD), project number 25-ERD-045. The authors have declared that none of them have any competing interests. Released under IM number LLNL-CONF-2015468.

Claude and ChatGPT were used to assist in debugging code, result visualization, and editorial support. All concepts related to model structure, datasets, training strategy, and qualitative analysis were developed by the authors.

All source code, analysis notebooks, and post-processed data used to produce the results in this paper are archived at https://zenodo.org/records/19242435.

REFERENCES

VISSTGRIDG16V4MINNIS, SGP, 2020-2024. https://adc.arm.gov/discovery//results. Accessed: 2025-04-24.

ARSCLKAZRBND1KOLLIAS, SGP, 2020-2024. https://adc.arm.gov/discovery//results. Accessed: 2025-04-22.

M2I1NXASM.5.12.4:inst1_2d_asm_Nx. https://disc.gsfc.nasa.gov/datasets?project=MERRA-2. Accessed: 2025-04-22.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework, July 2019. URL http://arxiv.org/abs/1907.10902. arXiv:1907.10902 [cs].

Christopher M. Bishop. Mixture density networks, 1994. URL https://publications.aston.ac.uk/id/eprint/373/. Num Pages: 438543.

Olivier Boucher, D. Randall, C. Bretherton, Graham Feingold, P. Forster, V.-M. Kerminen, Y. Kondo, H. Liao, U. Lohmann, Philip J Rasch, S.K. Satheesh, S. Sherwood, B. Stevens, and X.Y. Zhang. *Clouds and Aerosols. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK and New York, NY, USA, 2013.

Sarah Brüning, Stefan Niebler, and Holger Tost. Artificial intelligence (AI)-derived 3D cloud tomography from geostationary 2D satellite data. *Atmospheric Measurement Techniques*, 17(3): 961–978, February 2024. ISSN 1867-1381. doi: 10.5194/amt-17-961-2024. URL https://amt.copernicus.org/articles/17/961/2024/. Publisher: Copernicus GmbH.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery, January 2023. URL http://arxiv.org/abs/2207.08051. arXiv:2207.08051 [cs].

A. S. Donahue, P. M. Caldwell, L. Bertagna, H. Beydoun, P. A. Bogenschutz, A. M. Bradley, T. C. Clevenger, J. Foucar, C. Golaz, O. Guba, W. Hannah, B. R. Hillman, J. N. Johnson, N. Keen, W. Lin, B. Singh, S. Sreepathi, M. A. Taylor, J. Tian, C. R. Terai, P. A. Ullrich, X. Yuan, and Y. Zhang. To Exascale and Beyond—The Simple Cloud-Resolving E3SM Atmosphere Model (SCREAM), a Performance Portable Global Atmosphere Model for Cloud-Resolving Scales. *Journal of Advances in Modeling Earth Systems*, 16(7):e2024MS004314, 2024. ISSN 1942-2466. doi: 10.1029/2024MS004314. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2024MS004314. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2024MS004314.

Shirin Ermis, Cesar Aybar, Lilli Freischem, Stella Girtsou, Kyriaki-Margarita Bintsi, Emiliano Diaz Salas-Porras, Michael Eisinger, William Jones, Anna Jungbluth, and Benoit Tremblay. Global 3D Reconstruction of Clouds & Tropical Cyclones, November 2025. URL http://arxiv.org/abs/2511.04773. arXiv:2511.04773 [cs].

Stella Girtsou, Emiliano Diaz Salas-Porras, Lilli Freischem, Joppe Massant, Kyriaki-Margarita Bintsi, Guiseppe Castiglione, William Jones, Michael Eisinger, Emmanuel Johnson, and Anna Jungbluth. 3D Cloud reconstruction through geospatially-aware Masked Autoencoders, January 2025. URL http://arxiv.org/abs/2501.02035. arXiv:2501.02035 [cs].

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437. _eprint: https://doi.org/10.1198/016214506000001437.

P. Kollias, E. E. Clothiaux, M. A. Miller, B. A. Albrecht, G. L. Stephens, and T. P. Ackerman. Millimeter-Wavelength Radars: New Frontier in Atmospheric Cloud and Precipitation Research. October 2007. doi: 10.1175/BAMS-88-10-1608. URL https://journals.ametsoc.org/view/journals/bams/88/10/bams-88-10-1608.xml.

Kara D. Lamb, Clare E. Singer, Kaitlyn Loftus, Hugh Morrison, Margaret Powell, Joseph Ko, Jatan Buch, Arthur Z. Hu, Marcus van Lier Walqui, and Pierre Gentine. Perspectives on Systematic Cloud Microphysics Scheme Development With Machine Learning. *Journal of Advances in Modeling Earth Systems*, 18(1): e2025MS005341, 2026. ISSN 1942-2466. doi: 10.1029/2025MS005341. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2025MS005341. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2025MS005341.

Jussi Leinonen, Alexandre Guillaume, and Tianle Yuan. Reconstruction of Cloud Vertical Structure With a Generative Adversarial Network. *Geophysical Research Letters*, 46(12):7035–7044, 2019. ISSN 1944-8007. doi: 10.1029/2019GL082532. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2019GL082532. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019GL082532.

Hugh Morrison, Marcus van Lier-Walqui, Ann M. Fridlind, Wojciech W. Grabowski, Jerry Y. Harrington, Corinna Hoose, Alexei Korolev, Matthew R. Kumjian, Jason A. Milbrandt, Hanna Pawlowska, Derek J. Posselt, Olivier P. Prat, Karly J. Reimel, Shin-Ichiro Shima, Bastiaan van Diedenhoven, and Lulin Xue. Confronting the Challenge of Modeling Cloud and Precipitation Microphysics. *Journal of Advances in Modeling Earth Systems*, 12(8), 2020. ISSN 1942-2466. doi: 10.1029/2019MS001689.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer, December 2017. URL http://arxiv.org/abs/1709.07871. arXiv:1709.07871 [cs].

Abraham. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, July 1964. ISSN 0003-2700. doi: 10.1021/ac60214a047. URL https://doi.org/10.1021/ac60214a047.

A Appendix

A.1 Data Filtering Criteria

The follow criteria are applied to remove radar-radiometer pairs which are not cloudy, or are not consistent between the measuring instruments (KAZR, GOES). All 4 criteria are applied to the training and validation datasets, but the 4th criterion is not applied to data in the test set in order to mimic the case where a ground-truth reflectivity measurement is not available for comparison.

1. $\max_z Z(z) > -40$dBZ, where Z is KAZR reflectivity (in dBZ) and z is altitude; removes artifacts.
2. Cloud thickness (from KAZR) $> 200m$, removes very thin clouds.
3. Both KAZR and GOES have valid (non-NaN) measurements at the SGP site (see Figure A.1).
4. $|CTH_{GOES} - CTH_{KAZR}| < \sigma(|CTH_{GOES} - CTH_{KAZR}|)$ where CTH is cloud-top height and σ is the standard deviation across the training/validation dataset; removes inconsistent scenes where GOES and KAZR may not be measuring the same cloud.

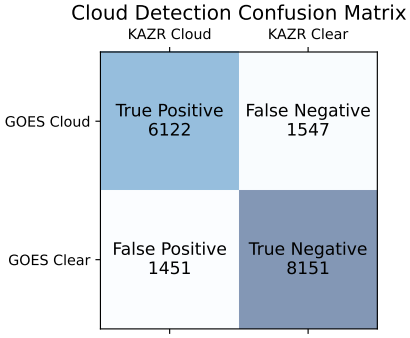

Figure A.1: Clear-sky vs cloud detection from GOES and KAZR at the ARM SGP site in 2022, aggregated across altitudes. Only the "True Positive" data are included in the training and validation datasets.

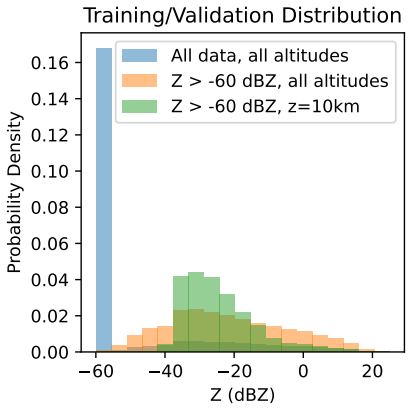

Figure A.2: Distribution of aggregated target reflectivities from the training and validation KAZR data, aggregated across altitude (blue, orange) and timestamps (all). Blue includes all training/validation data with NaN-reflectivities filled with the -60dBZ detection threshold; orange indicates the PDF with these non-cloudy (≤ -60 dBZ) data removed; and green indicates the PDF of all reflectivities at 10km altitude aggregated across training/validation.

A.2 Additional Metrics

In addition to the negative log-likelihood (NLL) used to train CERBERUS (Equation 1) we report additional traditional metrics including the Receiver Operating Curve (ROC) and Area Under the Curve (AUC), the correlation coefficient R^2, and the root-mean squared error (RMSE) between individual reflectivity predictions per altitude (Figure A.6). Specifically, RMSE is evaluated as the difference between the measurement y and the mean of the distribution prediction.

In addition to these traditional metrics, we further consider the Continuous Ranked Probability Score (CRPS) and the Energy of our model (Gneiting & Raftery, 2007), which are scalar and vector (respectively) scoring functions that treat the difference between a distribution and an observation. Specifically, the CRPS measures the elementwise difference between the predicted CDF F and empirical CDF corresponding to the measurement y:

$$\texttt{CRPS}(F, y) = \int \big(F(x) - H(x - y)\big)^2 dx \tag{2}$$

$$= \mathbb{E}_{X\sim F}|X - y| - \frac{1}{2}\mathbb{E}_{X,X'\sim F}|X - X'|, \tag{3}$$

where H denotes the Heaviside function and $\mathbb{E}$ denotes expectation. For our beta-type distributions, we take 100 samples from the predicted CDF to compute the empirical CRPS. CRPS reduces to mean absolute error in the case of a deterministic prediction.

Whereas CRPS is computed independently for each altitude of a predicted cloud profile, the Energy is a score on a multivariate prediction using vector norms:

$$\texttt{Energy}(F, y) = \mathbb{E}_{X\sim F}||X - y|| - \frac{1}{2}\mathbb{E}_{X,X'\sim F}||X - X'||. \tag{4}$$

We adopt the 2-norm convention for this work. In that case, the Energy is proportional to the root-mean squared error in the case of a deterministic prediction, differing by a factor of the square root of the vector dimension ($\sqrt{128}$ in our case). In Figures 3 and A.4 and Table A.1, Energy is reported with this vector-length scaling taken into account in order to make it directly comparable to RMSE.

A.3 Comparison to Non-Zero-Inflated Baselines

To verify improvement in model skill from the probabilistic ZIB prediction target of CERBERUS, we compare its performance to two additional baseline models. The first is a purely deterministic model, which uses a single-headed output layer to predict the reflectivity and is trained using mean-squared error as the loss function. The second is a beta distribution target *without* a ZI component: this two headed model predicts only the α and β parameters of the beta distributed reflectivity using two prediction heads, and is trained using the negative log likelihood with the standard beta distribution. All three cases (deterministic, beta, and ZIB) utilize the same model structure otherwise and differ only in the number of output heads.

Analysis of the CRPS (or mean-absolute error in the deterministic case) in Figure A.3 reveals that indeed, the probabilistic predictions from the ZIB and beta models outperforms the deterministic model across altitudes. Furthermore, the ZIB predictions add value on the order of a 1dBZ reduction in error below 8 km altitude relative to a non-ZI counterpart. Aggregated results reported in Figure A.4 tell a similar story: in deterministic metrics such as RMSE and correlation coefficient, the expectation value of the ZIB model consistently outperforms that of the 2-headed beta model, achieving comparable performance to the deterministic case. In the probabilistic Energy metric (RMSE for deterministic case), the ZIB model consistently displays the smallest distance from measurements. These results validate and motivate the choice to use a zero-inflated and probabilistic prediction target over deterministic and non-ZI alternatives for this particular reflectivity application.

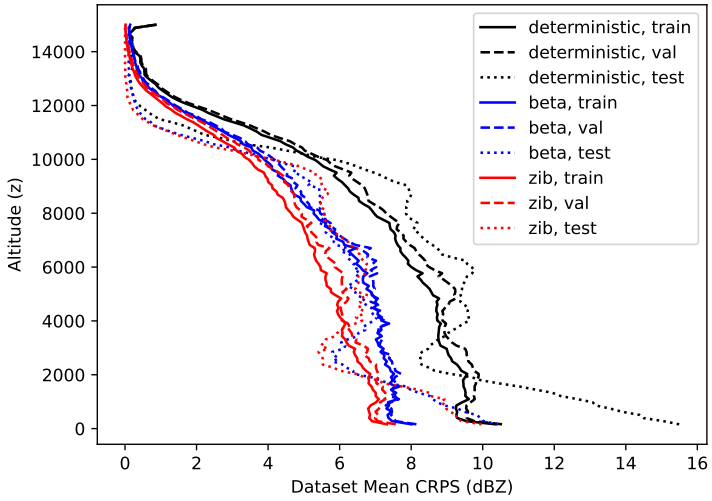

Figure A.3: Continuously ranked probability score (or MAE, for deterministic case) for the ZIB CERBERUS model and the non-ZI beta and deterministic baselines. CRPS is computed independently at each altitude for a given reflectivity measurement-prediction pair, and we report the mean CRPS across the training, validation, and testing datasets, respectively.

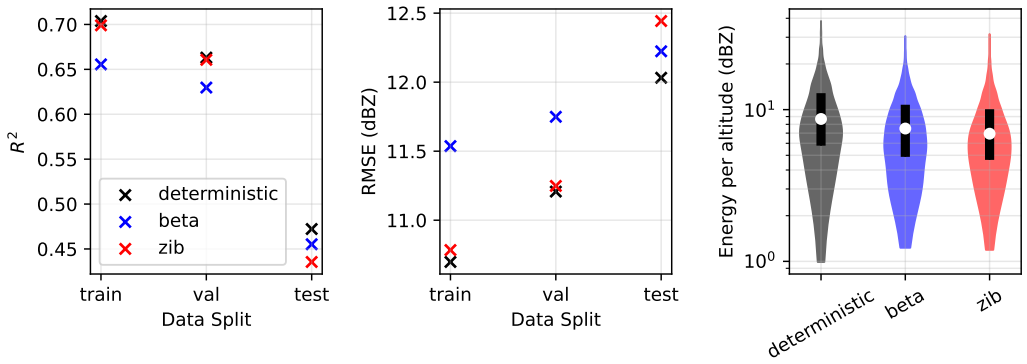

Figure A.4: Correlation coefficient R^2 (left), RMSE (middle), and scaled Energy (RMSE) for the ZIB model and baselines. The first two metrics are reported for training, validation, and testing data splits. The Energy / RMSE is reported only for the validation dataset; Energy is scaled by the vector size ($\times 1/\sqrt{128}$) to make it directly comparable to the RMSE that is reported for the deterministic model.

A.4 TRAINING CHARACTERISTICS

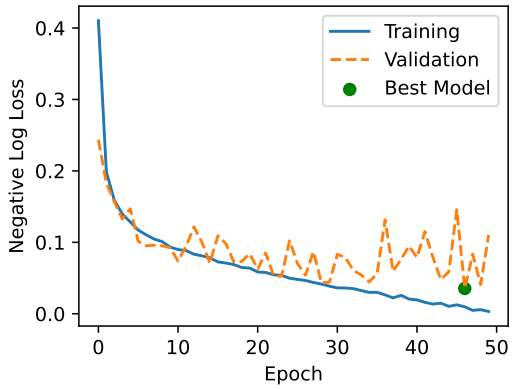

Figure A.5: Training and validation loss for the presented ZIB configuration of CERBERUS.

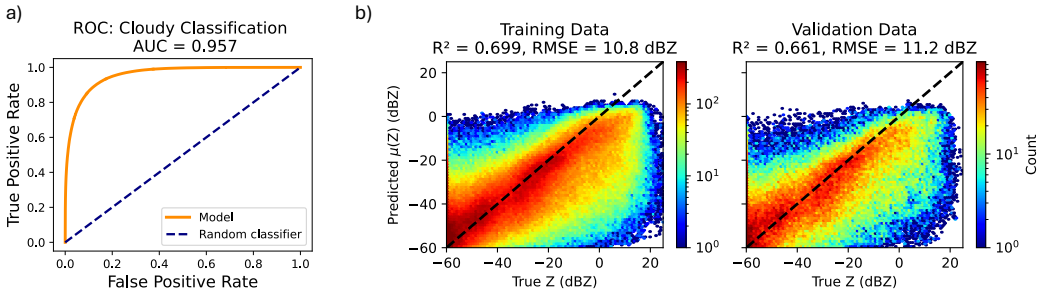

Figure A.6: (a) Receiver operating characteristic (ROC) curve for the classification head π of CERBERUS evaluated on validation data; (b) parity between the true reflectivity values and the mean of the ZIB prediction, aggregated across altitudes and timestamps in the training and validation sets, respectively.

A.5 Performance by Cloud Regime

Table A.1 reports both RMSE (between data and predicted mean) and the scaled Energy score for CERBERUS. The results reported for SatMAE are taken from Girtsou et al. (2025). Cloud regimes in the SGP dataset are determined according to cloud-top-height thresholds (2km, 6km) and maximum cloud reflectivity thresholds (-20dBZ, 0dBZ) from the KAZR satellite to permit inclusion of nighttime clouds that would be excluded from an optical depth threshold.

Cloud Regime (Abbreviation)	SatMAE RMSE in dBZ (SEVIRI prevalence)	CERBERUS RMSE / Energy in dBZ (CERBERUS prevalence)
Cumulus (Cu)	7.0 (0.8%)	3.4 / 2.9 (7.8%)
Stratocumulus (Sc)	7.5 (1.4%)	4.7 / 3.9 (4.1%)
Stratus (St)	4.8 (3.0%)	6.2 / 5.1 (0.1%)
Altocumulus (AC)	8.4 (0.6%)	5.6 / 4.7 (9.6%)
Altostratus (AS)	9.8 (0.4%)	8.1 / 6.5 (10.1%)
Nimbostratus (NS)	12.5 (0.0%)	14.5 / 11.4 (3.4%)
Cirrus (Ci)	4.8 (0.6%)	7.1 / 5.8 (15.0%)
Cirrostratus (CS)	N/A	10.8 / 8.6 (31.3%)
Deep Convection (DC)	10.3 (0.6%)	16.2 / 12.8 (18.5%)
Native-mean (non-clear sky)	6.6	9.7 / 7.8
SEVIRI-weighted mean	6.6	6.5 / 5.3

Table A.1: RMSE (mean across validation set samples), scaled Energy (for CEBERUS), and frequency of various cloud regimes, including results from CERBERUS at the SGP and results reported in Girtsou et al. (2025) for the SatMAE model over the MSG/SEVIRI satellite range. Clear-sky scenes are excluded from this analysis. Overall RMSE is reported across cloud regimes according to raw data weighting and according to SEVIRI cloud regime weighting. These data correspond to Figure 2.

A.6 Feature Importance

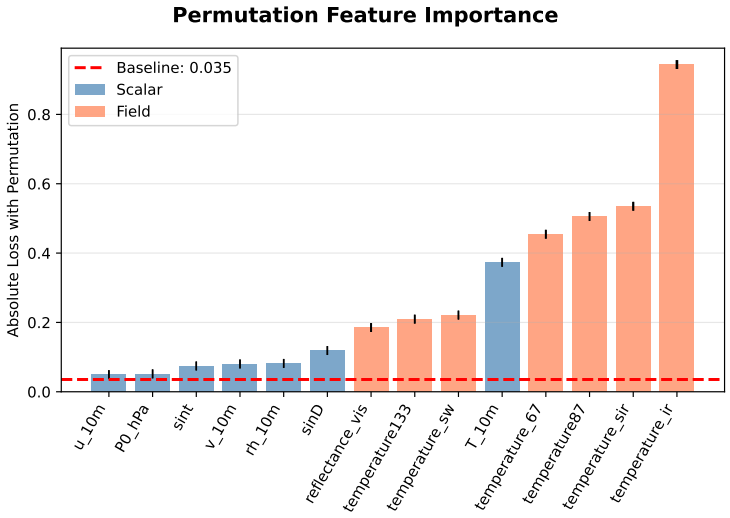

Figure A.7: Permutation feature importance with the original model loss (negative log likelihood) plotted as a red-dashed line. Features are plotting in order of least important/smallest change to loss, to most important, and are colored according to field type (GOES) or scalar type (meteorology/reanalysis).

B Additional Time-height Reflectivity Demonstrations

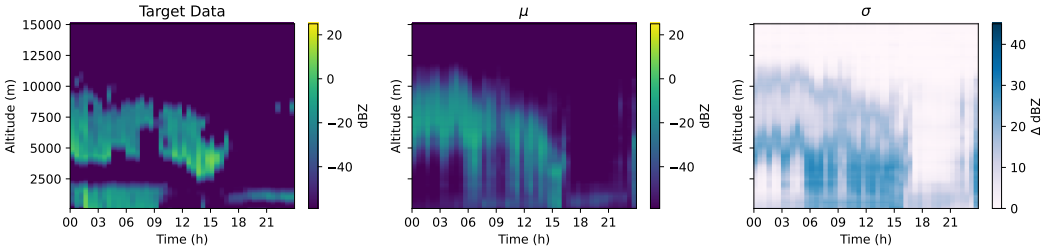

Figure A.8: As in Figure 4, for February 11, 2025 (test set).

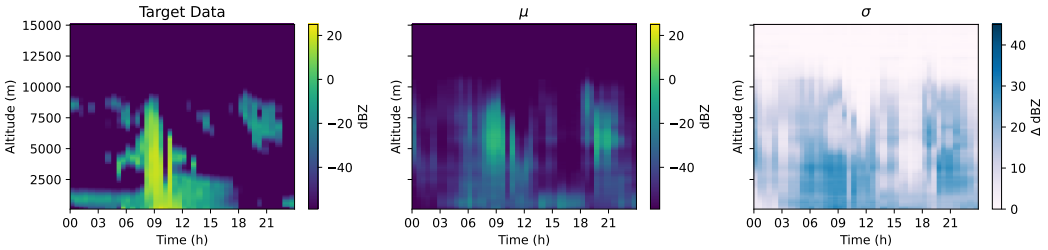

Figure A.9: As in Figure 4, for February 12, 2025 (test set).