Ice Cloud Geometry Retrieval with Calibrated Uncertainty from Passive Satellite Imagery

Ayush Prasad

Space and Earth Observation Centre, Finnish Meteorological Institute, Helsinki, Finland
ayush.prasad@fmi.fi

Abstract. We present a dense prediction system for ice cloud geometry retrieval from extremely sparse supervision, where active satellite sensors (EarthCARE radar/lidar) provide accurate but spatially narrow training labels (∼1.6% of pixels) and passive imagers (VIIRS, Visible Infrared Imaging Radiometer Suite) observe the full globe. A ConvNextUNet trained on co-located tracks predicts eight targets at every pixel with calibrated 90% prediction intervals using conformalized quantile regression (CQR). The best configuration, a five-member quantile ensemble, achieves $R^2 = 0.742$ with prediction quality constant regardless of distance to the supervision track, confirming that the model learns a per-pixel spectral retrieval rather than interpolating from nearby labels. Calibration is robust across latitudes and cloud types, with deep convective clouds as the main failure mode. Deployed globally, the system produces 609M predictions from one day of VIIRS data in 2.4 hours on a single GPU, enabling dense 3D cloud characterization at planetary scale.[1]

Keywords: Sparse-to-dense prediction · Uncertainty quantification · Satellite imagery · Conformalized quantile regression

1 Introduction

Clouds modulate the Earth's energy budget through shortwave reflection and longwave absorption, making their three-dimensional geometry a critical variable for climate modeling and weather prediction [3, 24]. Despite decades of research, cloud feedbacks remain the largest source of uncertainty in climate projections [27]. Accurately characterizing cloud vertical structure is essential for understanding radiative transfer, precipitation processes, and atmospheric dynamics.

Active satellite sensors provide the most accurate measurements of cloud vertical structure. The recently launched EarthCARE mission [8] carries a cloud profiling radar and atmospheric lidar that retrieve ice cloud properties at high vertical resolution. However, like its predecessor CALIPSO [26], EarthCARE observes only a narrow nadir track, covering less than 1% of the globe per day.

[1] Code and data available at https://github.com/ayushprd/sparse2cloud

Passive imagers such as VIIRS [4] observe the full globe daily but lack direct sensitivity to cloud vertical structure.

This coverage gap motivates a fundamental question: *can we learn from the narrow active-sensor tracks and predict dense ice cloud geometry at every pixel of the wide passive swath?* If so, how reliable are such predictions, and how should we quantify their uncertainty? More broadly, this is an instance of a general problem in dense prediction: learning pixel-wise outputs from extremely sparse spatial supervision, where only ∼1% of pixels carry labels.

We address these questions with the following contributions:

1. We present the first dense ice cloud geometry retrieval system with calibrated uncertainty from sparse active-sensor supervision: a ConvNextUNet trained on co-located VIIRS–EarthCARE patches predicts eight targets at every pixel with distribution-free coverage guarantees via CQR.
2. We provide empirical evidence that sparse-to-dense prediction is spectral, not spatial: prediction quality (R^2, coverage, interval width) is constant regardless of pixel distance to the supervision track, consistent with the model learning a per-pixel mapping from spectral features to targets rather than interpolating from nearby labels. This suggests that sparse supervision patterns covering diverse atmospheric conditions may suffice, a finding potentially relevant to other sparse-to-dense retrieval problems (e.g., canopy height from LiDAR tracks [19]).
3. We conduct a systematic comparison of eight uncertainty methods under identical CQR post-hoc calibration, with conditional coverage analysis across latitudes, cloud types, and illumination. We find that raw calibration quality determines post-CQR interval sharpness, and identify deep convective clouds as the main failure mode (3.8% under-coverage).
4. We demonstrate operational feasibility by deploying the model globally, producing 609 million dense predictions with 100% Earth coverage from a single day of VIIRS data in 2.4 hours on one GPU.

2 Related Work

Cloud property retrieval from passive imagery. Operational cloud products from passive sensors rely on physics-based algorithms that match observed radiances to pre-computed lookup tables [17, 20]. These methods estimate cloud-top properties (pressure, temperature, height) but cannot retrieve cloud base, geometric thickness, or ice water content due to limited information content in passive measurements. Learning-based approaches have shown promise: IceCloudNet [9] predicts ice water content profiles from SEVIRI using DARDAR as supervision, achieving $R^2 \approx 0.69$ on cloudy voxels. However, uncertainty quantification has been largely absent from these retrievals.

Uncertainty quantification in deep learning. Deep ensembles [12] remain a strong baseline for predictive uncertainty, though they require training multiple models.

MC-Dropout [6] provides a computationally cheaper alternative through stochastic forward passes. Evidential deep learning [1] places priors on the output distribution to estimate epistemic uncertainty in a single forward pass. Heteroscedastic regression [18] learns input-dependent noise variance. Quantile regression [11] directly estimates conditional quantiles without distributional assumptions.

Conformal prediction. Conformal prediction provides distribution-free coverage guarantees [25]. Conformalized quantile regression (CQR) [21] adapts conformal methods to quantile regression, producing prediction intervals with guaranteed marginal coverage. CQR has seen growing adoption in scientific applications where calibrated uncertainty is critical [2]. However, CQR only guarantees *marginal* coverage. Analyzing whether coverage holds conditionally across subpopulations (e.g., cloud types, latitudes) remains underexplored in remote sensing applications.

Dense prediction from sparse supervision. Learning dense predictions from spatially sparse labels is a well-studied problem. In computer vision, image-guided depth completion from sparse LiDAR points [16] shares a structural parallel with our task: both use dense 2D imagery to predict dense outputs from sparse point-wise supervision. Pauls *et al.* [19] predict global canopy height at 10 m resolution from Sentinel imagery using sparse GEDI LiDAR labels, achieving state-of-the-art accuracy with a shift-resilient loss function. Similar sparse-to-dense approaches appear in ocean remote sensing [13] and soil moisture mapping. Our work differs in explicitly studying the cross-track generalization mechanism and providing calibrated uncertainty for the dense predictions.

3 Method

Notation. We denote input features by x, ground-truth targets by y, predicted quantiles by $\hat{q}_\tau$, and nonconformity scores by s_i. n is the calibration set size for CQR and α is the desired miscoverage rate.

3.1 Problem Setup and Data

Target variables. We extract eight cloud geometry targets from EarthCARE ATL_ICE_2A retrievals [8]: cloud centroid height, cloud top height, cloud base height, peak ice concentration level, geometric thickness, core ice water content (IWC), column-mean IWC, and log ice water path (IWP). The first five describe cloud vertical geometry in pressure level units, while the last three characterize ice content. Each target is defined per atmospheric column from 242 vertical levels spanning 0–16 km altitude.[2]

[2] Heights are reported in pressure levels for training and evaluation. Figures convert to geometric altitude (km) using the US Standard Atmosphere.

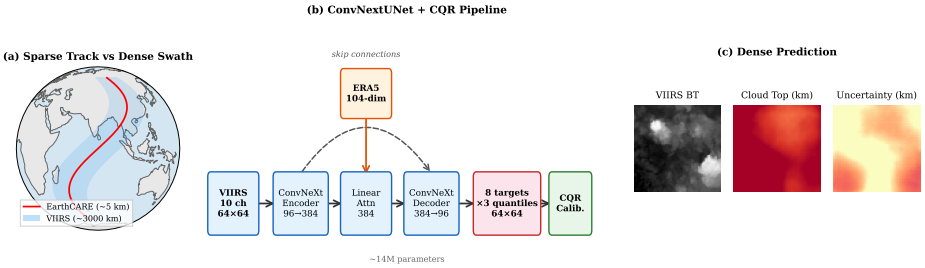

Fig. 1: Overview. **(a)** EarthCARE observes a narrow ∼5 km track while VIIRS covers a ∼3000 km swath. Our goal is dense prediction at every VIIRS pixel. **(b)** ConvNextUNet with ERA5 conditioning predicts 8 cloud geometry targets × 3 quantiles per pixel, followed by CQR calibration. **(c)** Example dense prediction over a tropical deep convective system: VIIRS brightness temperature input, predicted cloud top height (km), and 90% prediction interval width (km). Note that training labels exist only along the narrow EarthCARE track, but predictions are dense at all pixels.

Input features. We use ten VIIRS channels from VNP02MOD [4]: five thermal infrared bands (M12–M16, brightness temperature) providing cloud-top information, four shortwave bands (M07, M08, M10, M11, reflectance) sensitive to cloud optical properties, and the solar zenith angle. VIIRS provides global coverage twice daily at 750 m nadir resolution.

We additionally condition on ERA5 reanalysis profiles [7]: temperature, specific humidity, relative humidity, and geopotential height on 26 pressure levels (104-dimensional vector), providing atmospheric state context. Ablation experiments show ERA5 improves mean R^2 by ∼8 points. Without ERA5, the model retains $R^2 \approx 0.66$.

Co-location and patching. We match VIIRS granules to EarthCARE orbits within ±30 min using a two-stage spatial matching procedure (coarse KDTree, then local refinement). The dataset spans approximately 4,000 EarthCARE orbits across boreal winter and summer conditions. From 100,540 co-located patches (64×64 px at 750 m, ∼48 km^2 each), we extract 6.4M supervised pixel columns along the EarthCARE track (∼1.6% of pixels per patch).

Data splitting. We partition data using a geographic grid (10° cells) to prevent spatial leakage, yielding 4.5M training, 990K validation, and 924K test pixels.

3.2 ConvNextUNet Architecture

We design a fully convolutional encoder-decoder that predicts cloud geometry at every pixel, even those without supervision during training (Fig. 1). We choose ConvNeXt blocks [14] for their large receptive field (7×7 depthwise convolutions) at low parameter cost, with U-Net skip connections preserving spatial detail needed for per-pixel retrieval.

The encoder consists of three stages with channel dimensions [96, 192, 384], each containing two ConvNeXt blocks with 7×7 depthwise separable convolutions and 4× inverted bottleneck expansion. We apply stochastic depth (drop rate 0.25, linearly scheduled) and dropout (0.15) for regularization. Between stages, 2× downsampling reduces spatial resolution.

At the bottleneck, ERA5 profiles are projected by a 2-layer MLP (104→384) and added to the feature maps. A linear attention module [10] with 4 heads captures global dependencies at the coarsest resolution.

The decoder mirrors the encoder with transposed convolutions for upsampling and skip connections from corresponding encoder stages. The output head produces a tensor of shape $(B, 8, Q, 64, 64)$ where $Q=3$ quantiles for uncertainty estimation or $Q=1$ for point prediction. The full model has approximately 14M parameters.

3.3 Quantile Regression and Ensemble

For uncertainty quantification, we train the network to predict three conditional quantiles $\hat{q}_\tau$ at levels $\tau \in \{0.1, 0.5, 0.9\}$ for each target, using the pinball loss:

$$\mathcal{L}_{\text{pinball}}(\hat{q}_\tau, y) = \begin{cases} \tau \cdot (y - \hat{q}_\tau) & \text{if } y \geq \hat{q}_\tau \\ (1-\tau) \cdot (\hat{q}_\tau - y) & \text{otherwise} \end{cases} \tag{1}$$

summed over all targets, quantile levels, and supervised pixels. The loss is computed only at pixels with EarthCARE observations (∼1.6% of each patch, corresponding to 64 columns along the nadir track). At inference, the model predicts at all 64×64 pixels.

Our best method, G-QE, averages the quantile predictions from five independently trained models (different random seeds), analogous to deep ensembles [12] but applied to quantile outputs rather than Gaussian parameters. D4 augmentation (random rotations and flips) is applied during training for all methods. At inference, D4 test-time augmentation (TTA: 4 rotations × 2 flips, averaged after inverse transform) is applied only to G-QE. All other methods use single forward passes.

3.4 Conformalized Quantile Regression

Raw quantile predictions are often miscalibrated: the nominal 80% interval may cover substantially more or fewer than 80% of observations. We apply CQR [21] to obtain distribution-free coverage guarantees.

We split the validation set into a proper validation set (for model selection) and a calibration set. On the calibration set, we compute the nonconformity score:

$$s_i = \max(\hat{q}_{0.1}(x_i) - y_i,\ \ y_i - \hat{q}_{0.9}(x_i)) \tag{2}$$

and obtain the threshold $\hat{q}$ as the $\lceil (n+1)(1-\alpha) \rceil / n$ quantile of the scores, where α is the desired miscoverage rate. At test time, the calibrated prediction interval

is:

$$C(x) = [\hat{q}_{0.1}(x) - \hat{q},\ \ \hat{q}_{0.9}(x) + \hat{q}] \tag{3}$$

This guarantees marginal coverage $P(y \in C(x)) \geq 1 - \alpha$ under exchangeability, without assumptions on the data distribution.

3.5 Physics Constraints

Cloud geometry targets satisfy physical ordering constraints: cloud base $\leq$ centroid $\leq$ cloud top, and cloud base $\leq$ peak level $\leq$ cloud top. We explore two approaches to enforce these:

Soft ordering loss. An auxiliary penalty $\mathcal{L}_{\text{order}} = \lambda \sum \max(0, \hat{y}_{\text{base}} - \hat{y}_{\text{top}})$ added to the training objective, with $\lambda = 0.1$ ramped over the first 10 epochs.

Post-hoc projection. At inference, we project predictions onto the constraint set by clipping: $\hat{y}_{\text{base}} \leftarrow \min(\hat{y}_{\text{base}}, \hat{y}_{\text{top}})$ and similarly for other pairs. This achieves zero violations with no loss in R^2.

4 Experiments

4.1 Experimental Setup

Training. All models are trained for 200 epochs with AdamW [15] (learning rate 10^{-4}, cosine annealing to 10^{-6}, weight decay $5{\times}10^{-4}$), batch size 32, gradient clipping at norm 5.0, and bfloat16 mixed precision on an NVIDIA H200 GPU. We apply D4 augmentation (random rotations and flips) during training. Early stopping with patience 25 selects the best checkpoint by validation loss.

Uncertainty methods. We compare eight methods, all using the same ConvNextUNet backbone:

- **G-QE**: Five-member quantile ensemble (our best method)
- **G-Q**: Single quantile regression model
- **G-DE**: Deep ensemble of five point-estimate models [12]
- **G-HET**: Heteroscedastic regression (learned variance) [18]
- **G-EVI**: Evidential deep learning [1]
- **G-MCD**: MC-Dropout (20 stochastic passes) [6]
- **G-QP**: Quantile regression with soft ordering loss
- **G-QPC**: Quantile regression with physics-constrained head

We also include **G-B**, a deterministic point-estimate baseline (same architecture, L1 loss). All uncertainty methods are post-hoc calibrated with CQR at the same nominal coverage level (90%). Method names follow the convention **G**-{loss}-{strategy}: G = Geometry, Q = Quantile, E = Ensemble, DE = Deep Ensemble, HET = Heteroscedastic, EVI = Evidential, MCD = MC-Dropout, P = Physics-constrained, B = Baseline.

Metrics. We evaluate point prediction quality with R^2 and MAE, and uncertainty quality with: prediction interval coverage probability (PICP, target: 0.90), mean prediction interval width (MPIW, lower is better), and Spearman correlation ρ between interval width and absolute error (higher means uncertainty is more informative).

4.2 Main Results

Table 1 presents the full comparison.

Baselines. The BT regression baseline uses a quadratic ridge regression on 5 thermal brightness temperatures and solar zenith angle, the same physically meaningful inputs used by operational retrievals such as MODIS CO_2 slicing (cloud-top $R^2 = 0.590$). Two MLP baselines quantify the contribution of spatial context: a pixel-only MLP ($R^2 = 0.548$) and one with 5×5 context and ERA5 ($R^2 = 0.641$). The ConvNextUNet (G-B, $R^2 = 0.716$) gains +7.5% over the best MLP, confirming that convolutional processing captures patterns beyond local statistics.

Uncertainty methods. Among eight methods, G-QE achieves the highest R^2 (0.742), strongest error–uncertainty correlation ($\rho = 0.415$), and sharpest intervals (MPIW = 25.1).

Deep ensembles (G-DE) match G-QE in R^2 but are severely miscalibrated before CQR (raw PICP = 0.371 vs. the 80% target). After CQR calibration, all methods reach 90% coverage. However, methods with better raw calibration produce tighter post-CQR intervals, which is why G-QE outperforms G-DE on MPIW despite similar R^2.

Context with existing retrievals. Direct comparison with prior work is difficult due to differences in sensors, targets, and evaluation protocols. IceCloudNet [9] reports $R^2 \approx 0.69$ for ice water content profiles from SEVIRI [22], but evaluates only on cloudy voxels and uses DARDAR [5] rather than EarthCARE as ground truth. Operational MODIS cloud-top products [20] do not estimate cloud base, thickness, or ice content.

For a controlled comparison, our BT regression baseline uses the same inputs and evaluation protocol as G-QE. The +15.2% R^2 gap (0.590 → 0.742) quantifies the benefit of learned spatial features and multi-target joint estimation. Cloud-top height is the easiest of our eight targets ($R^2 = 0.826$).

Table 2 shows per-target R^2 for the top three methods. Geometric height targets (centroid, top, base) achieve $R^2 > 0.82$, while thickness ($R^2 = 0.407$) is hardest. The progression from MLP (0.244) to ConvNextUNet (0.339) to ensemble (0.407) shows diminishing returns, suggesting thickness $R^2 \approx 0.4$ is near the information-theoretic ceiling for passive measurements, consistent with VIIRS having only 2–3 degrees of freedom for vertical cloud structure [20]. Ensembling (G-QE vs. G-Q) improves all targets, with the largest gains on the hardest targets.

Table 1: Comparison of uncertainty methods. All ConvNextUNet methods (top block) share the same backbone and CQR calibration, differing only in loss function and output head (see Sec. 4.1 for method descriptions). Baselines (bottom block): G-B is a deterministic point-estimate ConvNextUNet. †BT regression is a physics-informed reference (quadratic ridge on 5 thermal bands + SZA). Best in **bold**, second-best underlined. G-MCD raw PICP and ρ are undefined as dropout variance does not directly yield quantile intervals.

	Method	Params	R^2 ↑	MAE ↓	PICP	MPIW ↓	Raw PICP	ρ ↑
ConvNextUNet	**G-QE**	70M	**0.742**	**6.08**	0.908	**25.1**	0.796	**0.415**
	G-Q	14M	0.731	6.27	0.912	26.8	0.791	0.388
	G-DE	70M	0.732	6.20	0.911	27.9	0.371	0.242
	G-HET	14M	0.718	6.55	0.901	27.1	**0.811**	0.373
	G-EVI	14M	0.720	6.39	0.912	28.9	0.511	0.328
	G-MCD	14M	0.716	6.41	0.918	31.1	—	—
	G-QP	14M	0.712	6.84	0.907	25.9	0.777	0.383
	G-QPC	14M	0.705	6.95	0.911	30.2	0.683	0.318
Baselines	G-B (point)	14M	0.716	6.41	—	—	—	—
	MLP (pixel+ctx)	0.9M	0.641	—	—	—	—	—
	MLP (pixel only)	0.8M	0.548	—	—	—	—	—
	BT regression†	—	0.590	—	—	—	—	—

4.3 CQR Calibration

Figure 2 demonstrates the effectiveness of CQR calibration. Raw prediction intervals exhibit significant miscalibration: at 80% nominal coverage, empirical coverage ranges from 37.1% (G-DE) to 81.1% (G-HET). After CQR calibration, all methods achieve near-perfect reliability across five coverage levels (50%–95%), with empirical coverage within ±0.2% of the nominal level.

The dramatic improvement for G-DE is notable: despite severely undercovering raw intervals, CQR expands them to achieve exact 90% coverage. However, this comes at the cost of wider intervals (MPIW= 27.9), explaining why well-calibrated raw intervals (G-QE's raw PICP= 0.796) lead to sharper calibrated intervals.

4.4 Conditional Coverage Analysis

CQR guarantees *marginal* coverage but does not ensure uniform coverage across subpopulations. We stratify test predictions by latitude (tropical, subtropical, mid-latitude, polar), illumination (day vs. night), and cloud type (low, mid-level, high-thin, deep convective), computing PICP and MPIW within each group.

Coverage is remarkably stable across most conditions: PICP ranges from 0.898 (polar) to 0.920 (mid-level clouds), all within ±2% of the 90% target. Day and night performance are nearly identical (PICP 0.905 vs. 0.911), confirming that thermal bands dominate the retrieval signal.

Table 2: Per-target R^2 and physics constraint violations. Post-hoc projection achieves zero violations without accuracy loss. MLP baseline uses per-pixel features with 5×5 context and ERA5.

	Cent	Top	Base	Peak	Thick	Core	MeanIWC	IWP	Avg	Violations (%)
G-QE	**.847**	**.826**	**.859**	**.795**	**.407**	**.812**	**.692**	**.697**	**.742**	0.005 → 0
G-Q	.836	.812	.847	.781	.370	.800	.671	.680	.725	0.293 → 0
G-B	.829	.805	.848	.779	.339	.788	.660	.679	.716	—
MLP	.782	.750	.830	.726	.244	.706	.530	.563	.641	—
G-QPC	.819	.791	.831	.763	.344	.774	.640	.650	.702	0 (by design)

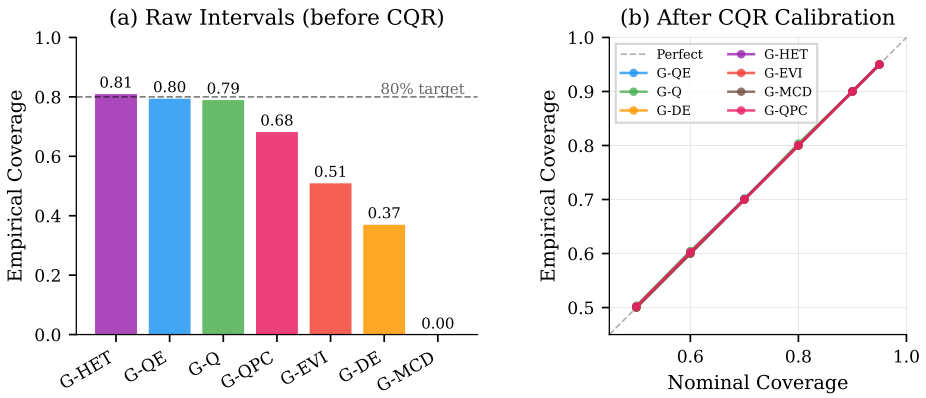

Fig. 2: CQR calibration. **(a)** Raw interval coverage at 80% nominal level, where most methods are significantly miscalibrated. **(b)** After CQR, all methods achieve near-perfect reliability across coverage levels 50%–95%.

The one significant failure mode is **deep convective clouds** (PICP=0.862), where intervals are 3.8% too narrow. Per-target analysis reveals that cloud top and peak level are most under-covered for this type. This is physically expected, as deep convective clouds exhibit the highest spatial variability in vertical extent.

4.5 Cross-Track Generalization

A key question for sparse-to-dense prediction is whether model quality degrades with distance from the supervision track. Figure 3 provides evidence that it does not, at least within the ∼48 km extent of our evaluation patches.

We evaluate G-QE by binning test pixels according to their distance (in pixels) from the nearest EarthCARE track position. Remarkably, R^2, PICP, and prediction interval width are all statistically constant across distance bins from 0–1 px to 5–8 px (Fig. 3a–c). The mean R^2 at 0–1 px (0.730) is indistinguishable from 5–8 px (0.731), and interval widths vary by less than 3% across bins.

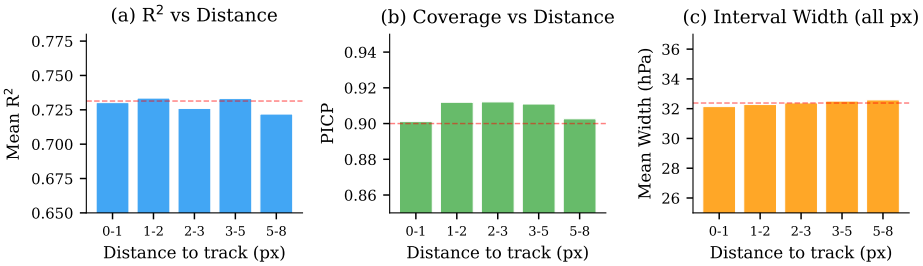

Fig. 3: Cross-track generalization. **(a)** R^2 vs. distance to supervision track, constant across all bins. **(b)** PICP remains at the 90% target. **(c)** Interval width is uniform, confirming the model performs spectral retrieval, not spatial interpolation. Red dashed lines show overall means.

This finding is consistent with the model performing *per-pixel spectral retrieval* rather than spatial interpolation from nearby supervised pixels. The thermal and shortwave signatures at each pixel appear to contain sufficient information to predict cloud geometry independently. We note that this analysis is limited to within-patch distances ($\leq$8 px, $\sim$6 km). Validation at larger distances (e.g., hundreds of km from any track) would require independent ground truth, which the cross-track generalization in Sec. 4.7 addresses qualitatively.

Consistent with this, a label efficiency experiment shows that MLP performance is nearly flat from 5% to 50% of training labels (R^2=0.660−0.673), further confirming that the model learns from spectral features rather than requiring dense spatial coverage.

4.6 Downstream Applications

Calibrated uncertainty enables two downstream applications (Fig. 4).

Selective prediction. By rejecting samples with the widest prediction intervals, we can substantially improve accuracy on the retained subset (Fig. 4a). Retaining the 50% most confident predictions improves R^2 from 0.742 to 0.935 for cloud top height and from 0.407 to 0.501 for thickness. The hardest target benefits most from uncertainty-based filtering.

Cloud type classification. We train a simple classifier to predict cloud type (6 ISCCP categories) from predicted geometry. Adding uncertainty features (interval widths for all targets) to the geometry inputs improves accuracy from 82.1% to 84.9% and macro F1 from 72.9% to 77.6% (Fig. 4b). Uncertainty-based selective classification further improves accuracy to 88.8% when retaining 50% of samples (Fig. 4c).

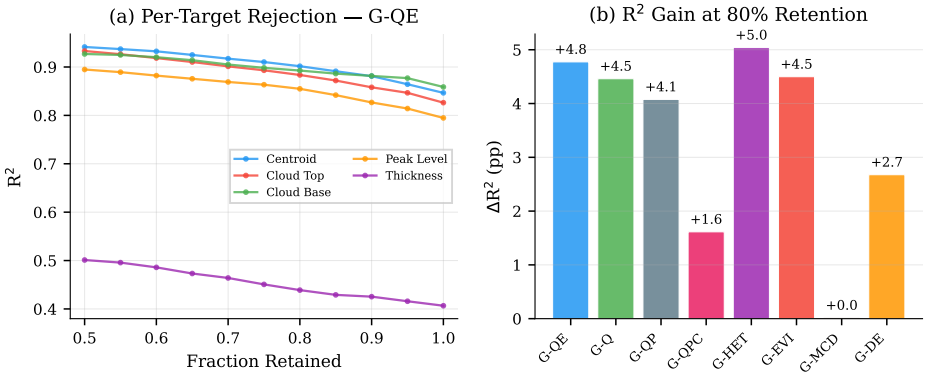

Fig. 4: Downstream applications of calibrated uncertainty. **(a)** Selective prediction: R^2 improves substantially when rejecting high-uncertainty samples. **(b)** Cloud classification accuracy and F1 improve when adding uncertainty features. **(c)** Uncertainty-based selective classification further boosts accuracy to 88.8% at 50% retention.

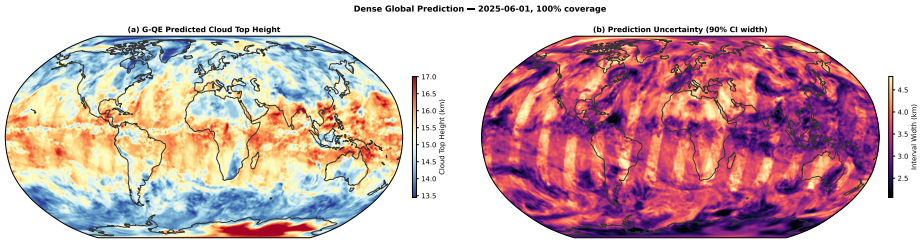

Fig. 5: Global dense prediction from one day of VIIRS data (609M pixels, 100% coverage). **(a)** Predicted cloud top height shows physically consistent patterns: high clouds along the ITCZ, mid-latitude storm tracks, and low subtropical stratus. **(b)** Prediction uncertainty (90% interval width) is highest for deep convective regions.

4.7 Global Dense Prediction

To demonstrate practical applicability, we deploy G-QE on a full day of VIIRS data (241 granules), tiling each granule into overlapping 64×64 patches with stride 32 and averaging predictions in overlapping regions. Results are gridded at 0.25° resolution with Gaussian smoothing (Fig. 5).

The global maps reveal physically plausible patterns: high cloud tops along the Intertropical Convergence Zone (ITCZ), mid-latitude storm tracks with moderate heights, and low clouds in subtropical subsidence regions. Uncertainty maps show wider intervals in the deep tropics (complex multilayer clouds) and narrower intervals for uniform stratus regions. The entire inference pipeline processes 609 million pixel predictions in 2.4 hours on a single GPU.

4.8 Approaches That Did Not Help

In developing this system, we systematically explored several alternative approaches that failed to improve over the baseline. We report these to benefit future work. (i) **Foundation model backbone**: replacing the ConvNextUNet encoder with a 2.6B-parameter SatVision Swin-v2 encoder [23] pretrained on 100M satellite images matched our 14M-parameter model's R^2 despite 185× more parameters, even after end-to-end fine-tuning. This rules out representation capacity as a bottleneck and suggests the task is fundamentally information-limited by the spectral content of passive measurements. (ii) **Spectral consistency regularization**: encouraging predictions at spectrally similar pixels to agree *degraded* performance by 1.2%, likely because convolutional filters already capture spectral consistency. (iii) **Cross-attention ERA5 fusion**: replacing additive ERA5 injection with cross-attention between pressure-level tokens and spatial features reduced R^2 by 0.9%, suggesting the atmospheric state is better used as a simple bias. (iv) **Adversarial training**: PatchGAN discriminators for spatial coherence were numerically unstable. (v) **Physics-constrained architectures**: ordinal regression and differentiable ordering heads provided no accuracy benefit over post-hoc projection.

5 Conclusion

We have presented a complete system for dense cloud geometry retrieval with calibrated uncertainty from passive satellite imagery, trained on sparse active-sensor supervision. By combining a ConvNextUNet with conformalized quantile regression, we achieve $R^2 = 0.742$ with guaranteed 90% coverage and sharp prediction intervals.

Prediction quality is invariant to distance from the supervision track, establishing that this is fundamentally a per-pixel spectral retrieval problem. This insight has practical implications: any sparse supervision pattern covering diverse atmospheric conditions would be equally effective, opening possibilities for more efficient active-passive sensor synergies.

Conditional coverage analysis reveals that CQR provides robust calibration across most conditions but under-covers deep convective clouds by 3.8%, identifying a concrete direction for improvement via subgroup-conditional calibration.

Limitations. Our model is trained exclusively on EarthCARE ATL_ICE_2A ice cloud retrievals and does not predict liquid cloud or warm rain geometry. Extending to all cloud phases would require additional supervision from radar-only products. ERA5 reanalysis has a ∼5-day latency, though ERA5T (∼12 h) or operational NWP forecasts can substitute (Sec. 3.1). Cloud geometric thickness ($R^2 = 0.407$) remains challenging due to the inherently limited vertical information content of passive measurements. Deep convective clouds exhibit under-coverage, suggesting the need for cloud-type-aware calibration.

Future work. Three directions follow from the limitations above. First, the deep convective under-coverage (3.8%) can be addressed by computing separate CQR thresholds per cloud type. Second, the thickness ceiling ($R^2 \approx 0.4$) may yield to temporal compositing of multiple VIIRS overpasses, which adds viewing-angle diversity, or to foundation models pretrained on diverse satellite data as they mature. Third, extending to liquid clouds and mixed-phase systems requires incorporating CloudSat radar-only products as additional supervision. More broadly, integration with operational NWP systems could enable real-time global cloud structure monitoring.

References

1. Amini, A., Schwarting, W., Soleimany, A., Rus, D.: Deep evidential regression. In: Advances in Neural Information Processing Systems. vol. 33, pp. 14927–14937 (2020)
2. Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning **16**(4), 494–591 (2023). https://doi.org/10.1561/2200000101
3. Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.M., Kondo, Y., Liao, H., Lohmann, U., Rasch, P., Satheesh, S.K., Sherwood, S., Stevens, B., Zhang, X.Y.: Clouds and aerosols. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the IPCC, pp. 571–657. Cambridge University Press (2013). https://doi.org/10.1017/CBO9781107415324.016
4. Cao, C., Xiong, X., Blonski, S., Liu, Q., Uprety, S., Shao, X., Bai, Y., Weng, F.: Suomi NPP VIIRS sensor data record verification, validation, and long-term performance monitoring. Journal of Geophysical Research: Atmospheres **118**(20), 11664–11678 (2013). https://doi.org/10.1002/2013JD020418
5. Delanoë, J., Hogan, R.J.: Combined CloudSat-CALIPSO-MODIS retrievals of the properties of ice clouds. Journal of Geophysical Research: Atmospheres **115**(D00H29) (2010). https://doi.org/10.1029/2009JD012346
6. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059 (2016)
7. Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society **146**(730), 1999–2049 (2020). https://doi.org/10.1002/qj.3803
8. Illingworth, A.J., Barker, H.W., Beljaars, A., Ceccaldi, M., Chepfer, H., Clerbaux, N., Cole, J., Delanoë, J., Domenech, C., Donovan, D.P., et al.: The EarthCARE satellite: The next step forward in global measurements of clouds, aerosols, precipitation, and radiation. Bulletin of the American Meteorological Society **96**(8), 1311–1332 (2015). https://doi.org/10.1175/BAMS-D-12-00227.1
9. Jeggle, K., Czerkawski, M., Serva, F., Le Saux, B., Neubauer, D., Lohmann, U.: IceCloudNet: 3D reconstruction of cloud ice from Meteosat SEVIRI. arXiv preprint arXiv:2410.04135 (2024)

10. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are RNNs: Fast autoregressive transformers with linear attention. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 5156–5165 (2020)
11. Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**(1), 33–50 (1978). https://doi.org/10.2307/1913643
12. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
13. Li, X., Liu, B., Zheng, G., Ren, Y., Zhang, S., Liu, Y., Gao, L., Liu, Y., Zhang, B., Wang, F.: Deep-learning-based information mining from ocean remote-sensing imagery. National Science Review **7**(10), 1584–1605 (2020). https://doi.org/10.1093/nsr/nwaa047
14. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11966–11976 (2022). https://doi.org/10.1109/CVPR52688.2022.01167
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
16. Ma, F., Cavalheiro, G.V., Karaman, S.: Self-supervised sparse-to-dense: Self-supervised depth completion from LiDAR and monocular camera. In: ICRA. pp. 3288–3295 (2019)
17. Minnis, P., Sun-Mack, S., Chen, Y., Chang, F.L., Yost, C.R., Smith, W.L., Heck, P.W., Arduini, R.F., Bedka, S.T., Yi, Y., et al.: CERES MODIS cloud product retrievals for Edition 4—Part I: Algorithm changes. IEEE Transactions on Geoscience and Remote Sensing **59**(4), 2744–2780 (2021). https://doi.org/10.1109/TGRS.2020.3008866
18. Nix, D.A., Weigend, A.S.: Estimating the mean and variance of the target probability distribution. In: Proceedings of 1994 IEEE International Conference on Neural Networks. vol. 1, pp. 55–60. IEEE (1994)
19. Pauls, J., Zimmer, M., Kelly, U., Schwartz, M., Saatchi, S., Ciais, P., Pokutta, S., Brandt, M., Gieseke, F.: Estimating canopy height at scale. In: Proceedings of the 41st International Conference on Machine Learning. pp. 39972–39988. Proceedings of Machine Learning Research (2024)
20. Platnick, S., Meyer, K.G., King, M.D., Wind, G., Amarasinghe, N., Marchant, B., Arnold, G.T., Zhang, Z., Hubanks, P.A., Holz, R.E., et al.: The MODIS cloud optical and microphysical products: Collection 6 updates and examples from Terra and Aqua. IEEE Transactions on Geoscience and Remote Sensing **55**(1), 502–525 (2017). https://doi.org/10.1109/TGRS.2016.2610522
21. Romano, Y., Patterson, E., Candès, E.J.: Conformalized quantile regression. In: Advances in Neural Information Processing Systems. vol. 32, pp. 3538–3548 (2019)
22. Schmetz, J., Pili, P., Tjemkes, S., Just, D., Kerkmann, J., Rota, S., Ratier, A.: An introduction to Meteosat Second Generation (MSG). Bulletin of the American Meteorological Society **83**(7), 977–992 (2002)
23. Spradlin, C.S., Caraballo-Vega, J.A., Li, J., Carroll, M.L., Gong, J., Montesano, P.M.: SatVision-TOA: A geospatial foundation model for coarse-resolution all-sky remote sensing imagery. arXiv preprint arXiv:2411.17000 (2024)
24. Stephens, G.L.: Cloud feedbacks in the climate system: A critical review. Journal of Climate **18**(2), 237–273 (2005). https://doi.org/10.1175/JCLI-3243.1
25. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005)

26. Winker, D.M., Vaughan, M.A., Omar, A.H., Hu, Y., Powell, K.A., Liu, Z., Hunt, W.H., Young, S.A.: Overview of the CALIPSO mission and CALIOP data processing algorithms. Journal of Atmospheric and Oceanic Technology **26**(11), 2310–2323 (2009). https://doi.org/10.1175/2009JTECHA1281.1
27. Zelinka, M.D., Myers, T.A., McCoy, D.T., Po-Chedley, S., Caldwell, P.M., Ceppi, P., Klein, S.A., Taylor, K.E.: Causes of higher climate sensitivity in CMIP6 models. Geophysical Research Letters **47**(1), e2019GL085782 (2020). https://doi.org/10.1029/2019GL085782