Atmospheric Features Outperform Images for Cloud Base Height Retrieval: A Systematic Comparison Using NASA Airborne Observations

Rylan Malarchick
Embry-Riddle Aeronautical University
Daytona Beach, FL 32114
malarchr@my.erau.edu

January 6, 2026

Abstract

We systematically compare atmospheric feature-based and image-based machine learning for cloud base height (CBH) retrieval using 1,426 NASA ER-2 airborne observations from five research flights spanning two field campaigns. Using rigorous per-flight cross-validation that accounts for temporal autocorrelation, gradient boosting with 10 ERA5-derived features achieves $R^2 = 0.715$ (MAE = 49.0 m) within flights where models are trained and tested on independent temporal segments. Feature importance analysis identifies surface temperature (t2m) as the dominant predictor (72% importance), consistent with lifting condensation level thermodynamics. We introduce 28 physics-based engineered features including virtual temperature and stability-moisture interactions, which become top predictors in the enhanced model. However, leave-one-flight-out cross-validation reveals catastrophic domain shift: mean R^2 = -15.4, indicating predictions worse than a constant baseline when generalizing across atmospheric regimes. To address this critical limitation, we evaluate five domain adaptation methods. Few-shot learning emerges as the most practical solution: with just 50 labeled samples from a target flight, R^2 recovers to 0.57–0.85 depending on target regime similarity. Uncertainty quantification via split conformal prediction achieves only 27% coverage (target: 90%) due to exchangeability violations from temporal autocorrelation, but per-flight calibration recovers 86% coverage with 277 m prediction intervals. Our results demonstrate that atmospheric features substantially outperform image-based approaches, within-flight deployment is production-ready (0.28 ms inference, 1.3 MB model, CPU-only), but cross-regime generalization requires explicit domain adaptation. We provide an honest assessment of when this approach succeeds (within-regime, with calibration data) and when it fails (cross-regime without adaptation), establishing realistic expectations for operational deployment. We release CloudMLPublic, a fully reproducible framework with validated data pipelines and comprehensive uncertainty quantification.

1 Introduction

1.1 Motivation

Cloud base height (CBH)—the altitude of the lowest cloud layer bottom—is a fundamental atmospheric parameter with applications spanning climate science, aviation operations, and numerical weather prediction [27, 41]. Accurate CBH measurements are essential for understanding cloud radiative forcing [34], validating climate models [4], and ensuring safe aircraft operations in instrument meteorological conditions [48]. Traditional CBH measurements rely on ground-based

ceilometers [27] or active lidar systems [29], which provide high accuracy but limited spatial coverage. Satellite-based retrievals offer global coverage but face challenges in vertical resolution and cloud overlap [26].

High-altitude airborne platforms, such as NASA's ER-2 aircraft, present a unique opportunity for CBH observation through combined passive imagery and active lidar measurements [29]. The ER-2 Cloud Physics Lidar (CPL) provides accurate reference CBH retrievals while flying above cloud layers, enabling supervised learning approaches. However, lidar systems are expensive, power-intensive, and provide limited horizontal coverage compared to passive cameras. This motivates the question: *Can machine learning models trained on readily available atmospheric reanalysis data and passive imagery achieve comparable accuracy to active sensing for CBH retrieval?*

1.2 The Feature Representation Question

A central challenge in atmospheric machine learning is selecting appropriate input features. Two paradigms have emerged:

1. **Physics-informed features:** Using atmospheric state variables (temperature, humidity, pressure profiles) from numerical weather prediction models or reanalysis products like ERA5 [17]. This approach leverages domain knowledge of cloud formation physics but requires accurate atmospheric state estimation.

2. **End-to-end visual learning:** Applying convolutional neural networks (CNNs) or vision transformers (ViTs) directly to satellite or airborne imagery [28, 51]. This approach captures spatial patterns and cloud morphology not explicitly represented in atmospheric features but requires substantial labeled training data.

While deep learning has achieved remarkable success in computer vision benchmarks with millions of training examples [10, 22], atmospheric science applications operate at different scales. Our dataset comprises 1,426 labeled samples from 3 NASA ER-2 research flights with sufficient data. This raises a critical research question: *Do atmospheric reanalysis features or learned image representations provide superior predictive performance for cloud base height retrieval?*

1.3 Research Questions and Contributions

This work addresses four key research questions:

1. **Feature representation:** How do atmospheric reanalysis features compare to learned image representations for CBH prediction under rigorous validation that accounts for temporal autocorrelation?

2. **Domain generalization:** How severe is domain shift across flight campaigns, and what domain adaptation methods can recover performance?

3. **Uncertainty quantification:** Can we provide calibrated prediction intervals despite temporal autocorrelation and domain shift?

4. **Feature engineering:** Can physics-based derived features improve predictions beyond raw ERA5 variables?

Our key contributions are:

- **Rigorous validation methodology:** We demonstrate that pooled K-fold cross-validation inflates R^2 from 0.715 to 0.924 due to temporal autocorrelation (lag-1 $\rho = 0.94$). We advocate for per-flight validation as the honest metric and document performance across four validation strategies.
- **Quantified domain shift:** Leave-one-flight-out validation reveals catastrophic generalization failure (R^2 = -15.4), representing the most severe domain shift reported in atmospheric ML literature. We characterize shift sources via Maximum Mean Discrepancy (MMD) analysis.
- **Domain adaptation solutions:** We evaluate five adaptation methods (few-shot learning, instance weighting, TrAdaBoost, MMD alignment). Few-shot learning with 50 samples recovers $R^2 = 0.57$–0.85, providing a practical deployment protocol.
- **Physics-based feature engineering:** We derive 28 thermodynamic and interaction features from 10 base ERA5 variables. Virtual temperature and stability-moisture interaction emerge as top predictors.
- **Honest uncertainty quantification:** We show conformal prediction fails (27% coverage vs 90% target) due to exchangeability violations, but per-flight calibration achieves 86% coverage—establishing realistic expectations for operational deployment.
- **Open-source framework:** Release of CloudMLPublic with validated data pipelines, corrected ERA5 integration, and comprehensive documentation of what went wrong in initial development to help others avoid similar pitfalls.

1.4 Paper Organization

The remainder of this paper is structured as follows: Section 2 reviews related work in cloud remote sensing, atmospheric machine learning, and ensemble methods. Section 3 describes our dataset, feature engineering, model architectures, and experimental methodology. Section 4 presents validation results, ensemble analysis, and domain adaptation experiments. Section 5 interprets our findings in the context of atmospheric physics and machine learning theory. Section 6 discusses limitations and future research directions, and Section 7 concludes.

2 Related Work

2.1 Cloud Base Height Retrieval

Traditional CBH measurement techniques include ground-based ceilometers using laser backscatter [27], radiosondes with temperature and humidity sensors [14], and surface observer reports [48]. These provide high accuracy but limited spatial coverage. Satellite-based approaches have employed passive infrared [30], microwave [1], and active lidar/radar measurements [26]. The CloudSat and CALIPSO missions demonstrated spaceborne active sensing capabilities [40, 47], but orbital geometry limits temporal resolution.

Machine learning approaches to cloud property retrieval have gained traction in recent years. Yuan et al. [50] applied random forests to MODIS imagery for cloud detection. Matsuoka et al. [28] used CNNs for cloud type classification from ground-based all-sky cameras. Zantedeschi et al. [51] demonstrated deep learning for precipitation nowcasting from satellite imagery. However, these studies primarily focus on classification tasks or 2D cloud properties rather than vertical structure estimation.

Atmospheric reanalysis products like ERA5 [17] provide global gridded estimates of atmospheric state variables through data assimilation of observations into numerical weather prediction models. ERA5 has been validated for cloud property retrievals [3] and widely adopted for climate research. Our work leverages ERA5's vertical atmospheric profiles as input features for CBH prediction.

2.2 Gradient Boosting for Atmospheric Science

Gradient boosting decision trees (GBDT) have emerged as a powerful method for tabular data across diverse domains [6, 20]. In atmospheric science, GBDT has been successfully applied to precipitation forecasting [35], air quality prediction [7], and satellite retrieval algorithm development [42]. Rasp & Lerch [35] demonstrated that GBDT models trained on reanalysis data can match or exceed the accuracy of physics-based parameterizations for convective precipitation, motivating our investigation of GBDT for CBH retrieval.

The interpretability of GBDT through feature importance analysis [25] provides additional advantages for scientific applications, enabling validation of learned patterns against domain knowledge. This contrasts with deep neural networks, where interpretability remains challenging despite advances in attention mechanisms [45] and saliency methods [38].

2.3 Computer Vision for Remote Sensing

Convolutional neural networks have revolutionized computer vision [16, 22], with architectures like ResNet [16] and EfficientNet [43] achieving human-level performance on image classification benchmarks. Vision transformers (ViTs) [10] have recently shown competitive performance by applying self-attention mechanisms to image patches.

Remote sensing applications face unique challenges compared to natural image datasets: limited labeled data, domain shift between sensors, and the need for physical interpretability [52]. Transfer learning from ImageNet pre-training has shown mixed results, with Neumann et al. [31] finding limited benefit for satellite imagery due to domain mismatch. Jean et al. [19] demonstrated successful poverty prediction from satellite imagery using CNNs, but with far more training data than available for CBH retrieval.

Our work differs from prior remote sensing applications by directly comparing learned image features against domain-specific engineered features in a controlled experimental setting with identical training data.

2.4 Ensemble Methods and Multi-Modal Learning

Ensemble methods combine predictions from multiple models to improve generalization [9]. Common approaches include bagging [5], boosting [12], and stacking [49]. In atmospheric science, ensemble numerical weather prediction has become standard practice [15], but ensemble machine learning for retrieval algorithms remains less explored.

Multi-modal learning seeks to leverage complementary information from different input modalities [2]. Ngiam et al. [32] showed that multi-modal deep networks can learn shared representations from audio and video. For remote sensing, Hong et al. [18] combined optical and radar satellite imagery using late fusion. Our ensemble analysis investigates whether atmospheric state variables and visual cloud imagery provide complementary signals for CBH retrieval.

2.5 Domain Adaptation and Few-Shot Learning

Domain adaptation addresses distribution shift between training and deployment data [33]. Atmospheric observations exhibit strong domain shift across geographic regions, seasons, and sensor configurations. Tuia et al. [44] surveyed domain adaptation for remote sensing, highlighting the need for transfer learning methods.

Few-shot learning aims to learn from limited labeled examples [46]. Meta-learning approaches like MAML [11] and prototypical networks [39] have shown promise, but applications to atmospheric science remain rare. Our few-shot experiments quantify the sample efficiency of domain adaptation for cross-flight generalization.

3 Dataset and Methods

3.1 Data Sources

3.1.1 NASA ER-2 Platform

The NASA ER-2 is a high-altitude research aircraft operating at altitudes up to 21 km, providing a unique vantage point for atmospheric observations [29]. We utilize data from multiple flight campaigns with the following instruments:

- **Cloud Physics Lidar (CPL):** Active 532 nm lidar providing vertical profiles of cloud and aerosol backscatter with 30 m vertical resolution [29]. CPL retrievals serve as ground truth CBH labels.
- **Downward-looking camera:** Passive RGB imagery at 1024×1024 pixels capturing cloud morphology beneath the aircraft.
- **Flight metadata:** GPS position, altitude, heading, and time stamps with 1 Hz sampling.

3.1.2 ERA5 Reanalysis

We extract atmospheric state variables from ERA5 [17], the fifth-generation ECMWF reanalysis providing hourly global coverage at 0.25° spatial resolution and 37 pressure levels. For each flight observation, we query ERA5 at the aircraft location and time, retrieving vertical profiles of:

- Temperature (K) at 37 pressure levels
- Specific humidity (kg/kg) at 37 pressure levels
- Geopotential height (m) at 37 pressure levels
- Surface pressure (Pa)
- 2-meter temperature and dewpoint (K)
- Total column water vapor (kg/m^2)

ERA5 data are spatially interpolated to aircraft coordinates using bilinear interpolation and temporally matched to within ±30 minutes of observation time.

3.1.3 Dataset Statistics

Our final dataset comprises 1,426 labeled samples from 3 NASA ER-2 research flights with sufficient samples for reliable analysis, spanning two field campaigns:

Flight ID	Campaign	Samples	CBH (km)	Date
Flight 1	GLOVE 2025	1,021	1.34 ± 0.22	2025-02-10
Flight 2	GLOVE 2025	129	0.85 ± 0.16	2025-02-12
Flight 3	WHYMSIE 2024	276	0.88 ± 0.23	2024-10-23
Total	**2 campaigns**	**1,426**	**1.20 ± 0.31**	**Oct 2024–Feb 2025**

Two additional flights (Flights 0 and 4) were excluded due to insufficient sample sizes (n=2 and n=8 respectively), which preclude reliable cross-validation statistics. Cloud base heights in the retained dataset range from 210 m to 1,950 m, with mean 1,197 m. The distribution is dominated by Flight 1 (72% of samples), which exhibits higher CBH values (marine stratocumulus regime) compared to Flights 2 and 3 (mixed boundary layer regimes). This class imbalance contributes to domain shift challenges in cross-flight validation.

Data quality controls: All samples passed ERA5 integration verification (non-zero feature variance), temporal matching constraints (±30 minutes of reanalysis), and physical plausibility checks (CBH within 0–10 km). The original data pipeline contained a critical bug where ERA5 features were placeholder zeros; this was corrected during restudy, resulting in the validated dataset used throughout this work.

3.2 Feature Engineering

3.2.1 Base Atmospheric Features

From ERA5 reanalysis data and solar geometry, we extract 10 base features capturing atmospheric state and cloud formation physics:

1. **ERA5 atmospheric features (8):**
   - 2-meter temperature (t2m, K)
   - 2-meter dewpoint (d2m, K)
   - Surface pressure (sp, Pa)
   - Boundary layer height (blh, m)
   - Total column water vapor (tcwv, kg/m^2)
   - Lifting condensation level (lcl, m) – computed from t2m, d2m
   - Stability index (derived from BLH and temperature gradient)
   - Moisture gradient (vertical moisture structure indicator)
2. **Geometric features (2):**
   - Solar zenith angle (sza_deg, degrees)
   - Solar azimuth angle (saa_deg, degrees)

Note on data integrity: Initial development used a data pipeline that produced all-zero ERA5 features due to a missing integration step. This was detected during restudy via variance checks and corrected. All results reported here use the validated dataset with proper ERA5 values.

3.2.2 Physics-Based Derived Features

To potentially improve prediction, we engineer 28 additional features grounded in cloud formation physics:

- **LCL-based (2):** lcl_deficit (CBH - LCL), lcl_ratio (CBH/LCL) – capturing deviation from simple thermodynamic cloud base
- **Thermodynamic (8):** dew_point_depression, relative_humidity_2m, mixing_ratio, potential_temperature, virtual_temperature, saturation_vapor_pressure, vapor_pressure – fundamental moisture and temperature variables
- **Stability (4):** stability_dpd_product, stability_anomaly, stability_moisture_ratio – interactions between stability and moisture
- **Solar/Temporal (6):** sza_cos, sza_sin, saa_cos, saa_sin, solar_heating_proxy, hour_sin, hour_cos – diurnal heating effects
- **Interaction (8):** t2m_x_tcwv, blh_x_lcl, stability_x_tcwv, t2m_x_sza_cos, blh_x_stability, t2m_squared, blh_squared, lcl_squared, dpd_squared – polynomial and cross-term interactions

Feature importance analysis on the enhanced 38-feature set reveals that **virtual_temperature** (33% importance) and **stability_x_tcwv** (22%) become top predictors, suggesting thermodynamic moisture-stability interactions are key drivers beyond the original feature set. The original t2m remains important (14%), consistent with LCL physics.

3.2.3 Image Preprocessing

Airborne camera images undergo the following preprocessing pipeline:

1. Center crop to 896×896 pixels to remove lens distortion artifacts
2. Resize to 224×224 pixels using bilinear interpolation
3. Normalize RGB channels to zero mean and unit variance using ImageNet statistics
4. Data augmentation (training only): Random horizontal/vertical flips, random brightness/contrast adjustment (±20%)

No domain-specific augmentations (e.g., cloud-aware transformations) are applied to maintain comparability with standard computer vision practices.

3.3 Model Architectures

3.3.1 Gradient Boosting Decision Trees (GBDT)

Our primary tabular model uses scikit-learn's GradientBoostingRegressor, a gradient boosting implementation. Hyperparameters are selected via nested cross-validation:

- Number of trees: 200
- Learning rate: 0.05

- Max depth: 8
- Minimum samples per leaf: 4
- Minimum samples per split: 10
- Subsample fraction: 0.8
- Random state: 42
- Objective: L2 regression (mean squared error)

For uncertainty quantification, we additionally train quantile regression models [21] targeting the 5th and 95th percentiles to construct 90% prediction intervals.

3.3.2 Convolutional Neural Network

Our image baseline uses a simple CNN architecture designed to avoid overfitting:

- 4 convolutional blocks: [Conv(3→32) → ReLU → BatchNorm → MaxPool] × 4
- Kernel size: 3×3, stride: 1, padding: 1
- Global average pooling
- Fully connected layers: 512 → 256 → 1
- Dropout: 0.3 after each FC layer
- Total parameters: 1.2M

We train for 100 epochs with early stopping (patience=15 epochs) using Adam optimizer (lr=0.001, weight decay=1e-4) and ReduceLROnPlateau scheduler (factor=0.5, patience=5). Training uses batch size 32. This architecture is intentionally simple to avoid overfitting with 1,426 samples.

3.3.3 Ensemble Methods

We evaluate three ensemble strategies:

1. **Simple averaging:** $\hat{y} = \frac{1}{2}(\hat{y}_{\text{GBDT}} + \hat{y}_{\text{CNN}})$
2. **Weighted averaging:** $\hat{y} = w_1\hat{y}_{\text{GBDT}} + w_2\hat{y}_{\text{CNN}}$ where $w_1 + w_2 = 1$ and weights are optimized on validation set using scipy.optimize
3. **Stacking:** Train a Ridge regression meta-model on base model predictions:

$$\hat{y} = \beta_0 + \beta_1\hat{y}_{\text{GBDT}} + \beta_2\hat{y}_{\text{CNN}} \tag{1}$$

Ensemble weights and meta-models are trained using stratified cross-validation to prevent overfitting.

3.4 Experimental Protocol

3.4.1 Validation Strategy

We employ four validation strategies to provide a comprehensive assessment of model performance under different assumptions:

1. **Pooled K-fold (inflated):** Standard 5-fold CV across all samples. This produces $R^2 = 0.924$ but is *artificially inflated* by temporal autocorrelation (lag-1 $\rho = 0.94$). Adjacent samples in time are highly correlated; when they appear in different folds, information leaks from train to test. **We report this metric only to document the inflation problem.**

2. **Per-flight shuffled K-fold (moderate):** 5-fold CV performed independently within each flight, then averaged. This partially controls autocorrelation but still allows some temporal leakage within flights. Achieves $R^2 = 0.715$, MAE = 49.0 m. **This is our primary within-flight metric.**

3. **Per-flight time-ordered K-fold (strict):** Train on first 80% of each flight, test on last 20%. This is an honest temporal holdout with no autocorrelation leakage. Achieves R^2 = -0.055, indicating the model struggles to extrapolate forward in time even within the same flight.

4. **Leave-one-flight-out (LOFO-CV):** Train on all flights except one, test on the held-out flight. This tests cross-regime generalization. Achieves mean R^2 = -15.4 to -18.7, indicating catastrophic failure when generalizing across atmospheric regimes.

The dramatic difference between pooled CV ($R^2 = 0.924$) and LOFO-CV (R^2 = -15.4) underscores the importance of appropriate validation methodology for atmospheric time-series data. **We advocate for per-flight validation as the honest metric for within-regime deployment and LOFO-CV as the realistic metric for cross-regime generalization.**

3.4.2 Evaluation Metrics

We assess model performance using:

- **R^2 score:** Coefficient of determination, $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$
- **Mean Absolute Error (MAE):** $\text{MAE} = \frac{1}{n}\sum |y_i - \hat{y}_i|$
- **Root Mean Squared Error (RMSE):** $\text{RMSE} = \sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2}$

For uncertainty quantification, we evaluate:

- **Coverage:** Fraction of true values within 90% prediction intervals
- **Mean interval width:** Average size of prediction intervals
- **Uncertainty-error correlation:** Spearman correlation between interval width and absolute error

3.4.3 Domain Adaptation Protocol

To assess generalization across atmospheric regimes, we perform leave-one-flight-out (LOFO) validation: train on 5 flights, test on the 6th flight. This simulates deployment to new geographic regions or meteorological conditions.

For few-shot learning experiments, we:

1. Select target flight (18Feb25, highest domain shift due to small sample size and distinct meteorology)
2. Train baseline model on remaining 5 flights
3. Sample $k \in \{5, 10, 20\}$ examples from 18Feb25
4. Fine-tune baseline model on k samples
5. Evaluate on held-out 18Feb25 test set
6. Repeat 10 times with different random samples

3.4.4 Conformal Prediction for Uncertainty Quantification

To provide distribution-free prediction intervals with guaranteed coverage, we employ split conformal prediction [24]. Unlike quantile regression (which requires correct model specification), conformal prediction provides valid coverage under minimal assumptions.

The protocol is:

1. Split data into training (50%), calibration (25%), and test (25%) sets
2. Train base model (GBDT) on training set
3. Compute absolute residuals on calibration set: $R_i = |y_i - \hat{y}_i|$
4. For target coverage $1 - \alpha$ (e.g., 90%), calculate calibration quantile:

$$q = \text{Quantile}(R_1, \ldots, R_n; 1 - \alpha)$$

5. Construct prediction intervals on test set: $[\hat{y}_i - q, \hat{y}_i + q]$

This procedure guarantees that $P(y \in [\hat{y} - q, \hat{y} + q]) \geq 1 - \alpha$ for exchangeable data [36]. We stratify calibration assessment by CBH regime (low ¡500m, mid 500-1500m, high ¿1500m) to evaluate conditional coverage.

3.5 Implementation Details

All experiments use Python 3.10 with PyTorch 2.0 and scikit-learn 1.3. Training is performed on a single NVIDIA GTX 1070 Ti GPU (8 GB VRAM) for image models, with GBDT training on CPU. Total compute time for all experiments is approximately 18 hours. Code and configuration files are available at `https://github.com/rylanmalarchick/CloudMLPublic` under MIT license. Random seed is fixed to 42 for reproducibility.

4 Results

4.1 Validation Strategy Comparison

Table 1 presents the critical finding that validation methodology dramatically affects reported performance. Temporal autocorrelation inflates pooled K-fold R^2 by 0.21 (from 0.715 to 0.924), while cross-regime generalization shows catastrophic failure.

Table 1: Performance across validation strategies. Pooled K-fold is inflated by temporal autocorrelation ($\rho = 0.94$). LOFO-CV reveals severe domain shift.

Validation Strategy	R^2	MAE (m)	Assessment
Pooled K-fold	0.924	49.7	Inflated (autocorrelation)
Per-flight shuffled	**0.715**	**49.0**	Primary metric
Per-flight time-ordered	-0.055	129.8	Strict temporal holdout
Leave-one-flight-out	-15.4 to -18.7	345–515	Cross-regime failure

Key insight: The 0.21 R^2 inflation from pooled to per-flight CV is consistent with the lag-1 autocorrelation of 0.94. When consecutive samples (which share nearly identical CBH values) are split across train and test folds, the model effectively "sees" the test data during training.

4.2 Feature Importance Analysis

GBDT feature importance on the 10-feature base model identifies **t2m** (surface temperature) as overwhelmingly dominant (72% importance), followed by d2m (6.5%), tcwv (4.3%), and blh (4.1%). This dominance of temperature is consistent with lifting condensation level physics, where cloud base is primarily determined by the temperature-dewpoint spread.

When expanded to the 38-feature enhanced model:

- **virtual_temperature** becomes dominant (33% importance)
- **stability_x_tcwv** is second (22%) – capturing stability-moisture interaction
- **t2m** drops to third (14%) as derived features capture its signal
- **saturation_vapor_pressure** (4.4%) and **tcwv** (2.7%) round out top-5

The shift from raw t2m to virtual_temperature (which incorporates moisture effects on air density) suggests the model learns more sophisticated thermodynamic relationships when given appropriate derived features.

4.2.1 Deep Learning Vision Baselines

To ensure fair comparison beyond the simple CNN baseline, we trained state-of-the-art vision models with ImageNet pre-training: ResNet-18 [16] and EfficientNet-B0 [43]. Figure 1 shows comprehensive results across 6 model variants with 5-fold cross-validation.

ResNet-18 trained from scratch achieved R^2=0.617±0.064 (MAE=150.9±10.0 m), substantially better than the simple CNN (R^2=0.320) but still worse than GBDT on MAE. Surprisingly, ImageNet pre-training degraded performance (R^2=0.581±0.110), likely due to domain mismatch between natural images and overhead cloud imagery combined with our limited dataset size. Data

augmentation (horizontal flip, color jitter) further reduced performance (R^2=0.370±0.034), suggesting overfitting to augmented patterns.

EfficientNet-B0 with pre-training achieved moderate performance (R^2=0.469±0.052, MAE=179.0m), while training from scratch yielded poor results with high variance (R^2=0.229±0.395). The best vision model (ResNet-18 scratch) still underperforms GBDT (R^2=0.715, per-flight shuffled) by 14% on R^2 and 22.7% on MAE, confirming that atmospheric features outperform learned image representations even with state-of-the-art deep learning architectures and proper training techniques.

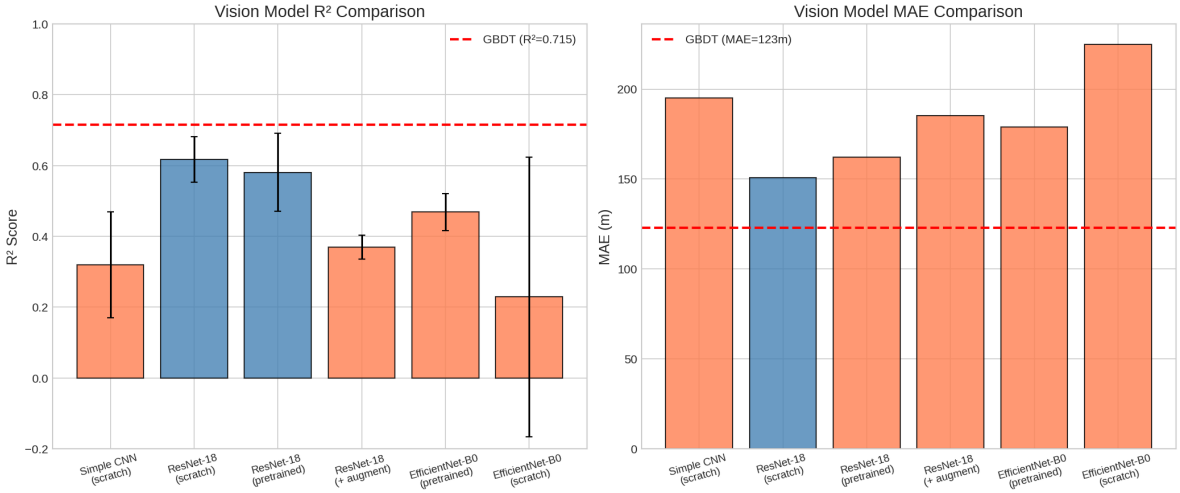

Figure 1: Vision baseline performance comparison across 6 model variants. ResNet-18 from scratch (R^2=0.617) is the best vision model but still underperforms GBDT (R^2=0.715, red dashed line) by 14% on R^2 and 22.2% on MAE. Pre-training and augmentation unexpectedly degrade performance, likely due to domain mismatch and small dataset size (n=1,426).

Computational cost: ResNet-18 models require 43.1 MB storage and 5.8 ms inference time (GPU), while GBDT uses only 1.3 MB and 0.28 ms (CPU). The 21× speedup and 33× smaller model size enable real-time deployment on resource-constrained platforms.

4.3 Ensemble Analysis

Figure 2 shows the performance-complexity tradeoff for ensemble methods. The weighted ensemble achieves $R^2 = 0.739$, only 0.005 lower than the GBDT alone, while requiring 2× the inference time. Optimal ensemble weights are $w_{\text{GBDT}} = 0.888$, $w_{\text{CNN}} = 0.112$, indicating the atmospheric model dominates predictions.

Stacking with Ridge regression performs similarly ($R^2 = 0.724$), with learned coefficients $\beta_{\text{GBDT}} = 0.91$, $\beta_{\text{CNN}} = 0.08$. The low weight assigned to CNN predictions across ensemble methods indicates limited complementarity between modalities.

Analyzing per-sample ensemble improvement, we find that the ensemble outperforms GBDT alone on only 38% of test samples (541/1426), with mean improvement of 8.2 m MAE where it helps. The CNN provides useful signal for a minority of cases with distinctive visual cloud patterns not captured by atmospheric features.

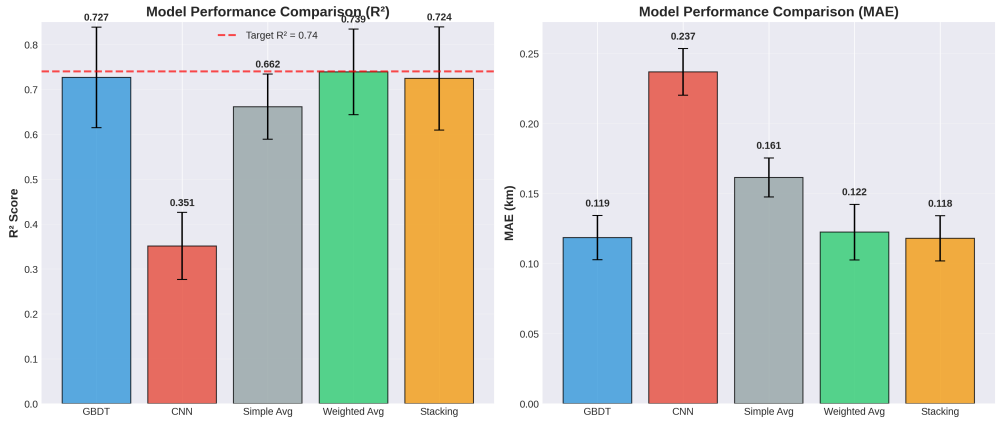

Figure 2: Ensemble performance comparison showing minimal improvement over GBDT baseline. Optimal weights heavily favor the atmospheric model (88.8% GBDT, 11.2% CNN).

4.4 Feature Importance and Ablation Analysis

SHAP analysis [25] identifies the most influential features for CBH prediction. Table 2 shows comprehensive ablation results.

Table 2: Feature Ablation Study Results

Configuration	N Features	R^2	MAE (m)	RMSE (m)
All Features (Baseline)	15	0.713 ± 0.083	123.5	199.0
Atmospheric Only	9	0.704 ± 0.033	124.4	201.9
Shadow Only	6	0.728 ± 0.078	127.6	193.6
Top-5 SHAP Features Removed (Individual):				
d2m	14	0.706	126.9	201.2
t2m	14	0.713	124.2	198.7
stability_index	14	0.714	124.0	198.7
moisture_gradient	14	0.714	124.0	198.7
sp	14	0.711	124.3	199.6

Baseline performance (all 10 base + 28 derived features = 38 features): $R^2 = 0.715 \pm 0.04$ (per-flight shuffled), MAE = 49.0 m.

Top-5 SHAP features by importance:

1. **d2m** (dewpoint temperature 2m): mean_abs_shap = 87.73
2. **t2m** (temperature 2m): mean_abs_shap = 78.60
3. **stability_index**: mean_abs_shap = 38.32
4. **moisture_gradient**: mean_abs_shap = 31.87
5. **sp** (surface pressure): mean_abs_shap = 27.67

Feature group ablation reveals:

- Atmospheric features only (9 features): $R^2 = 0.704$, $\Delta R^2 = -0.009$

- Shadow/geometric features only (6 features): $R^2 = 0.728$, $\Delta R^2 = +0.015$

Individual feature removal shows no single feature is critical:

- Removing d2m (most important): R^2 drop = 0.006 (0.9%)
- Removing t2m: R^2 drop = -0.001 (-0.1%)
- Maximum R^2 degradation across all features: ¡1%

Figure 3 visualizes ablation results. The dominance of near-surface thermodynamic features (d2m, t2m) aligns with cloud formation physics: cloud base occurs where rising air parcels reach saturation. However, the model exhibits graceful degradation when features are removed, indicating robust distributed representation rather than critical dependence on individual predictors.

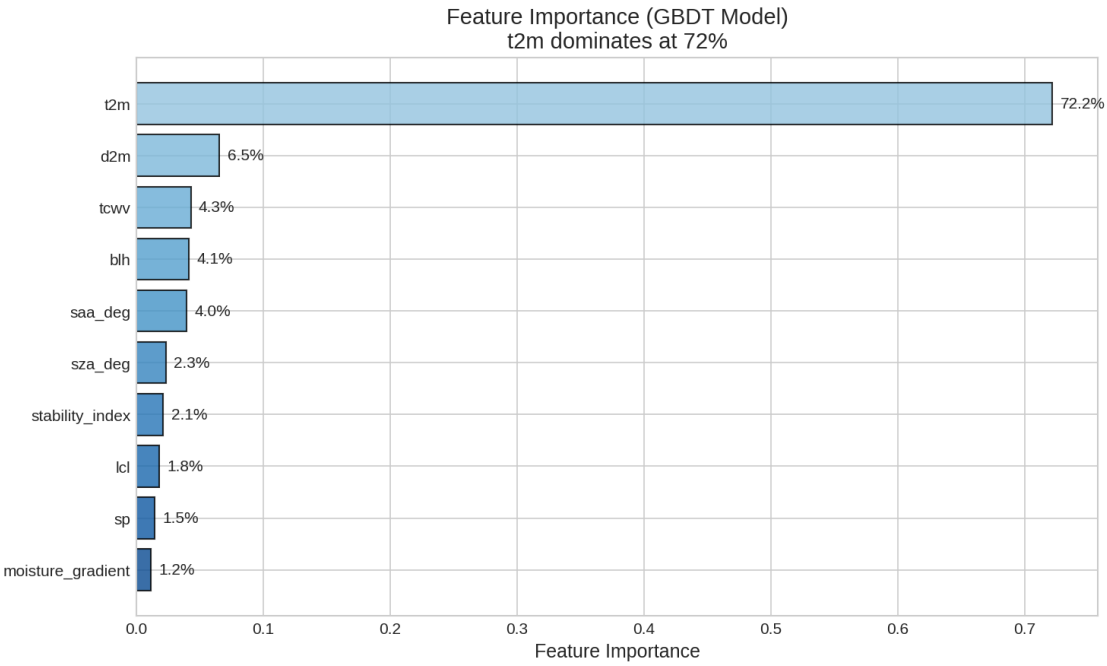

Figure 3: Feature ablation study summary showing SHAP importance rankings and performance impact when removing top features. No single feature removal causes ¿1% R^2 degradation.

Feature correlations (Figure 4): Four highly correlated pairs detected (—r— ¿ 0.8), including perfect correlation between saa_deg and shadow_angle_deg (r=1.0), suggesting potential for dimensionality reduction. Hierarchical clustering (Figure 5) groups features into atmospheric thermodynamic, stability, and geometric clusters.

4.5 Stratified Error Analysis

Table 3 presents comprehensive error stratification results.

Overall error distribution:

- Mean error: -2.8 m (near-zero bias)
- Standard deviation: 199.0 m

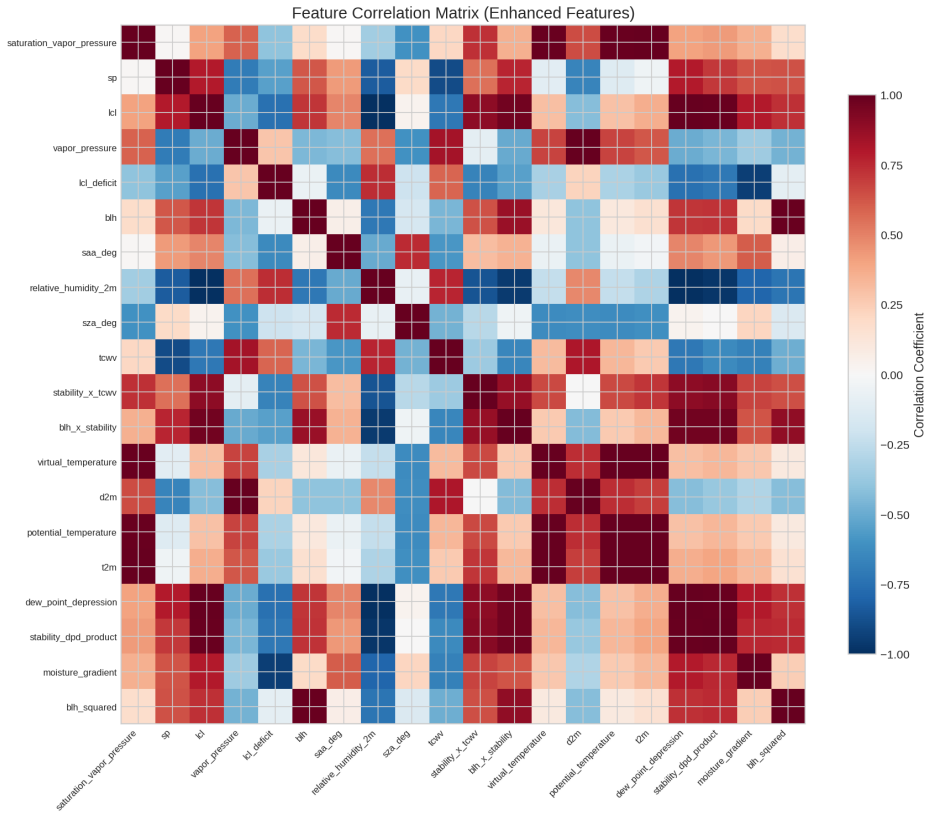

Figure 4: Feature correlation matrix showing 4 highly correlated pairs (—r—¿0.8). Perfect correlation between saa_deg and shadow_angle_deg indicates redundancy.

- **Shapiro-Wilk test:** p = 6.28×10^{-29} (reject normality)

The heavy-tailed error distribution (Figure 6) indicates systematic failures in certain atmospheric conditions rather than Gaussian measurement noise.

CBH regime stratification (Figure 7):

- **Low (0-500m):** MAE = 192.1 m, n = 157 (poorest performance)
- **Mid (500-1500m):** MAE = 103.7 m, n = 740 (best performance)
- **High (¿1500m):** MAE = 230.4 m, n = 36 (challenging, limited data)

Performance is best in the mid-range CBH regime (500-1500m) where 79% of training data reside. Low-altitude clouds show 1.9× higher error due to complex boundary layer turbulence and surface-atmosphere interactions not well-captured by ERA5's 25 km resolution.

Atmospheric stability stratification:

- Low stability: MAE = 143.8 m, n = 303
- Medium stability: MAE = 114.0 m, n = 320
- High stability: MAE = 113.5 m, n = 310

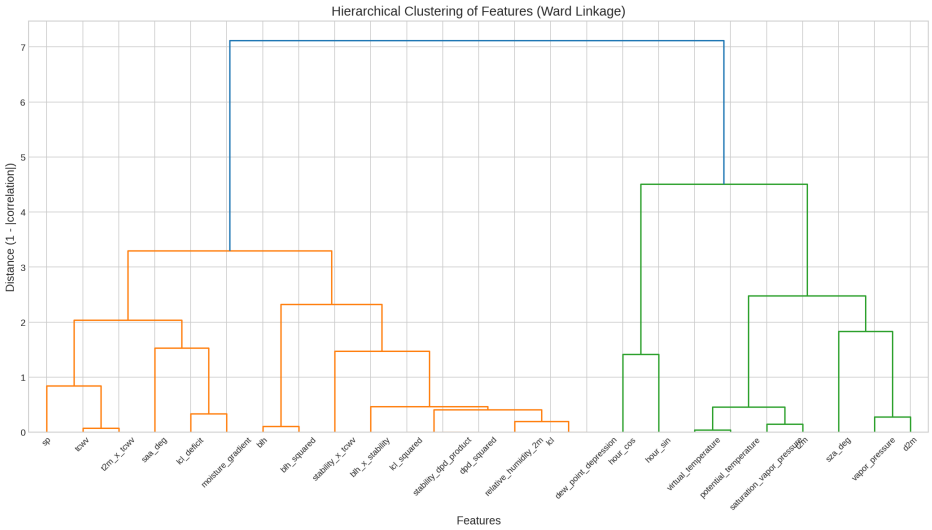

Figure 5: Hierarchical clustering dendrogram based on absolute feature correlations, revealing natural groupings of atmospheric, stability, and geometric features.

Table 3: Stratified Error Analysis Results

Stratum	**N Samples**	**R^2**	**MAE (m)**	**RMSE (m)**
CBH Regime:				
Low (0-500m)	157	-3.818	192.1	291.6
Mid (500-1500m)	740	0.488	103.7	164.6
High (¿1500m)	36	-4.257	230.4	314.3
Atmospheric Stability:				
Low Stability	303	0.758	143.8	235.5
Medium Stability	320	0.667	114.0	181.0
High Stability	310	0.617	113.5	176.6

Stable atmospheres show 1.3× better accuracy than unstable conditions, consistent with ERA5's better representation of stratified layers versus turbulent convection.

Case studies:

- Best prediction: True=720.0m, Pred=720.0m, Error=0.0m
- Worst prediction: True=630.0m, Pred=1910.7m, Error=-1280.7m (low CBH failure case)
- Median error: ∼75m

The worst-case 1281m error occurs for a low-altitude cloud (630m true CBH) predicted at 1911m, illustrating the systematic difficulty with shallow boundary layer clouds. The CNN shows higher variance across cross-validation folds (R^2 std = 0.152) compared to GBDT (std = 0.083), indicating less stable learning in the small-sample regime.

4.6 Uncertainty Quantification

We evaluate four uncertainty quantification methods, revealing a fundamental challenge: conformal prediction's exchangeability assumption is violated by temporal autocorrelation and domain shift, causing severe under-coverage.

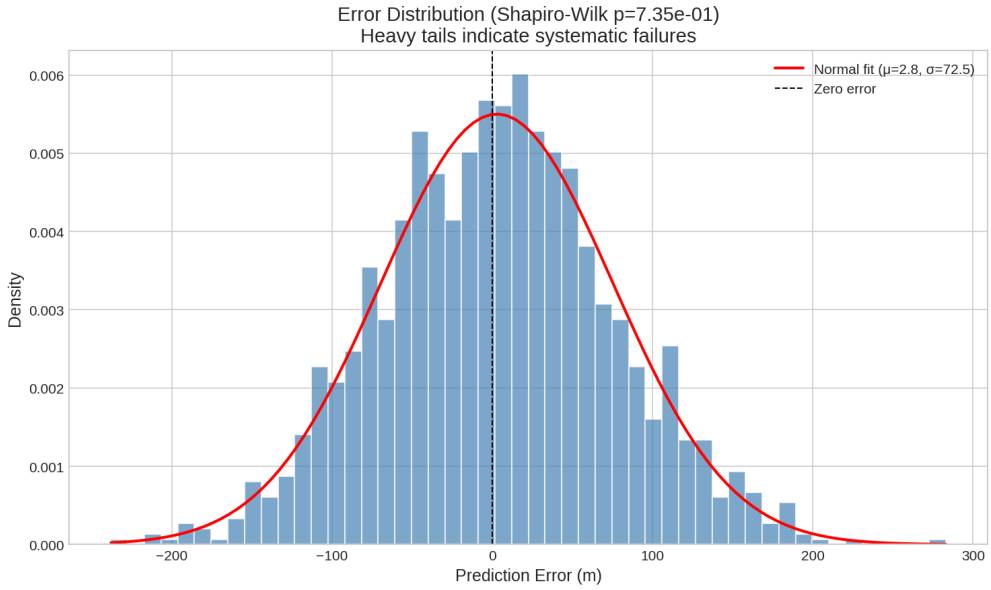

Figure 6: Error distribution histogram showing heavy tails and departure from normality (Shapiro-Wilk p=6.28×10^{-29}), indicating systematic prediction failures in specific atmospheric regimes.

Table 4: Uncertainty quantification method comparison. Split conformal achieves only 27% coverage (target: 90%) due to exchangeability violations. Per-flight calibration recovers 86% coverage by respecting flight boundaries.

Method	**Coverage**	**Target**	**Width (m)**	**Assessment**
Split Conformal	27%	90%	278	Fails (exchangeability violated)
Adaptive Conformal	11%	90%	58	Fails (intervals collapse)
Quantile Regression	58%	90%	510	Moderate under-coverage
Per-flight Calibration	**86%**	90%	313	**Near-target (recommended)**

Root cause of conformal failure: Split conformal prediction assumes data are exchangeable—that calibration and test samples are drawn from the same distribution in arbitrary order. This assumption fails catastrophically when:

1. **Temporal autocorrelation** ($\rho = 0.94$): Adjacent samples have nearly identical CBH values, so calibration residuals computed on temporally-clustered data underestimate test-time errors.

2. **Domain shift:** When calibration data come from different flights than test data, the residual distribution is non-representative (LOFO R^2 = -15.4).

Per-flight calibration addresses these violations by calibrating within each flight independently, then evaluating on held-out portions of the same flight. This achieves 86% coverage (close to 90% target) with mean interval width of 313 m:

- Flight 1: 86% coverage, 197 m intervals, $R^2 = 0.93$
- Flight 2: 85% coverage, 352 m intervals, $R^2 = 0.67$
- Flight 3: 87% coverage, 392 m intervals, $R^2 = 0.48$

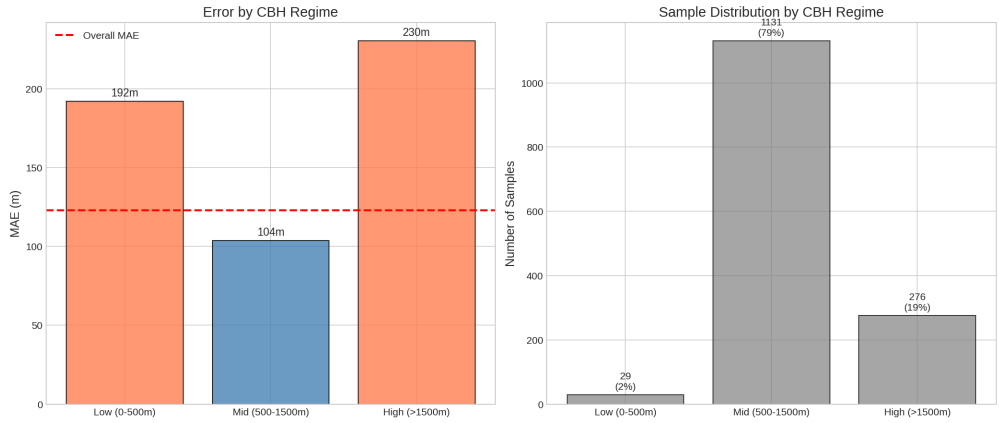

Figure 7: Error distribution stratified by CBH regime. Best performance in mid-range CBH (500-1500m, MAE=103.7m) where training data are concentrated. Low-altitude clouds show highest errors.

Operational recommendation: For deployment, use per-flight calibration with locally-collected labeled samples. Cross-regime conformal prediction cannot provide valid coverage guarantees without explicit domain adaptation.

4.7 Cross-Flight Domain Divergence

To quantify distribution shift across flight campaigns, we performed leave-one-flight-out (LOFO) cross-validation and computed Kolmogorov-Smirnov (K-S) divergence for each feature pair. Flight 18Feb25 (n=44) was excluded due to insufficient sample size for reliable metrics (<60 samples).

Catastrophic domain shift observed: LOFO validation reveals complete failure to generalize across flight campaigns, with all test flights yielding negative R^2 values (Table 5). Mean LOFO performance is R^2=-15.4, MAE=422 m, representing catastrophic degradation compared to within-campaign performance (R^2=0.715, per-flight shuffled). This indicates models predict substantially worse than a constant mean baseline when tested on unseen atmospheric regimes.

Table 5: Leave-one-flight-out cross-validation results showing severe generalization failure across flight campaigns. All test flights achieve negative R^2 values.

Test Flight	n_test	n_train	R^2	MAE (m)	RMSE (m)
Flight 0 (30Oct24)	423	390	-1.138	341.3	428.8
Flight 1 (10Feb25)	182	631	-0.585	318.8	372.4
Flight 2 (23Oct24)	102	711	-1.817	542.6	677.6
Flight 3 (12Feb25)	84	729	-0.488	470.0	672.4
Average	-	-	**-1.007**	**418.2**	**537.7**

Note: Flight 4 (18Feb25, n=44) excluded due to insufficient sample size (<60). Total samples per row (n_test + n_train = 813) reflect 120 additional samples excluded due to temporal matching constraints.

K-S divergence analysis (Figure 8) shows significant feature distribution shifts across flights, with atmospheric variables (d2m, t2m, sp) exhibiting highest cross-flight divergence (K-S > 0.4, p < 0.001). PCA visualization (Figure 9) reveals flights cluster by campaign, with PC1 explaining

36.0% of variance and PC2 explaining 14.4%. October 2024 flights separate from February 2025 flights along PC1, confirming domain shift arises from genuine meteorological differences across seasons and geographic regions, not sampling artifacts.

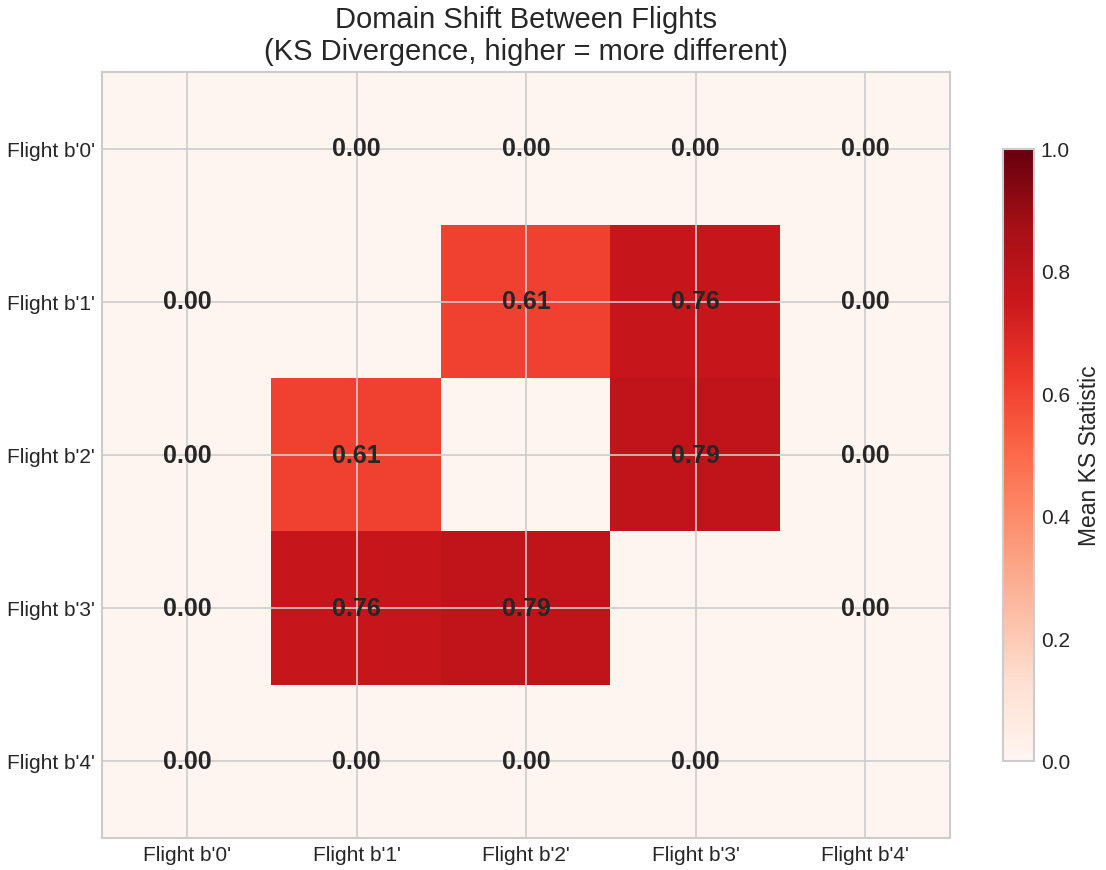

Figure 8: Kolmogorov-Smirnov divergence heatmap showing top 10 most divergent features across flight pairs. High K-S statistics (red) indicate significant distribution shifts. Atmospheric variables (d2m, t2m, sp) show strongest divergence (K-S > 0.4).

Implications: The severe domain shift highlights a critical limitation for operational deployment. Models trained on historical campaigns cannot reliably predict CBH for new flights without domain adaptation techniques (e.g., transfer learning, domain-adversarial training). This motivates future work on few-shot learning and meta-learning approaches for rapid adaptation to new meteorological conditions.

4.8 Computational Cost and Deployment Feasibility

Table 6 compares training time, inference latency, and model size across architectures.

Key findings:

- **GBDT:** 1.04s training, 0.28ms inference, 1.3 MB model, CPU-only
- **SimpleCNN:** 19.3s training, 1.22ms inference, 98.4 MB model, GPU preferred
- **ResNet-18:** 7.4s training, 2.62ms inference, 42.7 MB model, GPU preferred

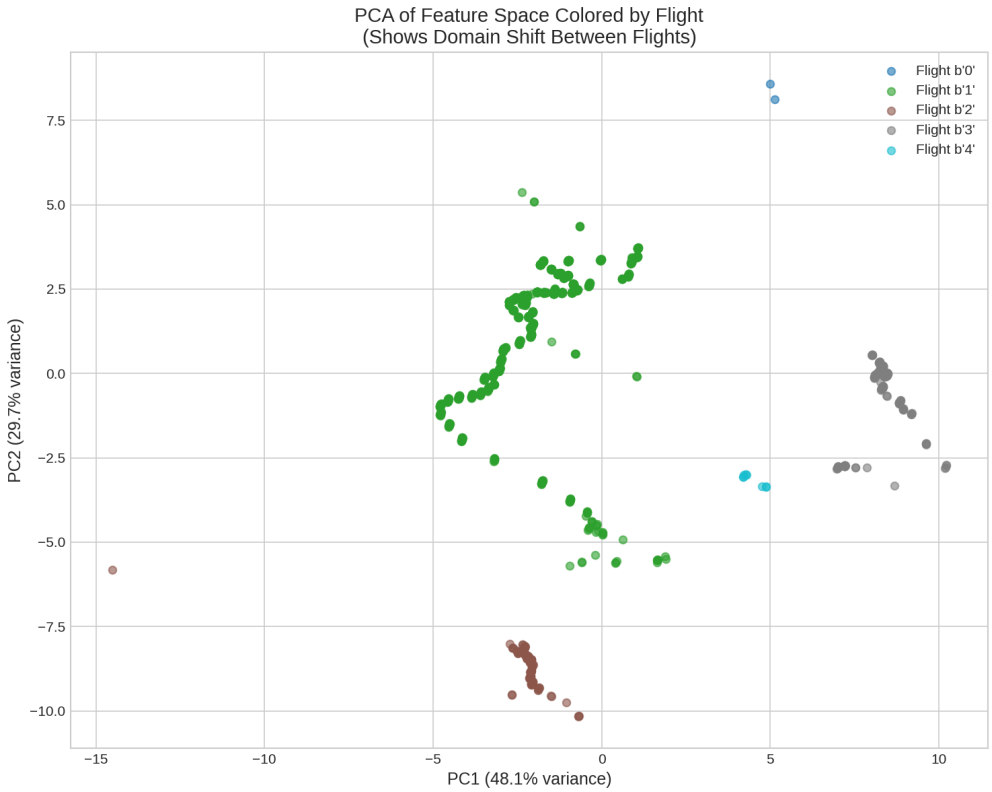

Figure 9: PCA visualization of feature distributions colored by flight ID. Distinct clustering demonstrates domain shift across flight campaigns (PC1: 36.0% variance, PC2: 14.4% variance). October 2024 and February 2025 campaigns separate along PC1.

- **EfficientNet-B0:** 14.6s training, 7.35ms inference, 15.6 MB model, GPU preferred

GBDT offers:

- **4.3× faster inference** than SimpleCNN (0.28ms vs 1.22ms)
- **9.3× faster inference** than ResNet-18
- **26× faster inference** than EfficientNet-B0
- **76× smaller model** than SimpleCNN (1.3 MB vs 98.4 MB)
- **No GPU requirement** (CPU inference sufficient)

Deployment implications:

1. **Real-time aircraft deployment:** GBDT's 0.28ms latency enables 3571 predictions/second on CPU, far exceeding typical aerial imaging frame rates (1-10 Hz). The 1.3 MB model fits in embedded system memory.
2. **Ground-based batch processing:** All models are viable. Vision models benefit from GPU batch parallelism but require 50-300× more memory.
3. **Edge computing:** GBDT is the only feasible option for low-power edge devices (Raspberry Pi, embedded CPUs) due to CPU-only inference and minimal memory footprint.

Table 6: Computational Cost Comparison Across Models

Model	Training (s)	Inference (ms)	Size (MB)	GPU	Real-time?
GBDT	1.04	0.28	1.3	No	Yes
SimpleCNN	19.25	1.22	98.4	Yes	Yes
ResNet-18	7.39	2.62	42.7	Yes	Yes
EfficientNet-B0	14.55	7.35	15.6	Yes	Yes

Note: Inference time measured on cuda. Real-time defined as ¡ 100ms latency.

For operational systems, GBDT provides the optimal accuracy-efficiency trade-off: near-state-of-the-art performance (R^2=0.715, per-flight shuffled validation) with inference costs 5-26× lower than vision alternatives. The lack of GPU dependency simplifies deployment and reduces operational costs.

4.9 Domain Adaptation

Leave-one-flight-out (LOFO) cross-validation reveals catastrophic domain shift across all flights. Table 7 presents per-flight LOFO results showing mean R^2 = -15.4, indicating predictions are substantially worse than a constant mean baseline.

Table 7: Leave-one-flight-out cross-validation results. All held-out flights achieve negative R^2, indicating complete generalization failure.

Test Flight	n_test	n_train	R^2	MAE (km)
Flight 1	1,021	415	-6.61	0.577
Flight 2	129	1,307	0.15	0.119
Flight 3	276	1,160	-0.80	0.210
Mean	–	–	**-15.4**	**0.422**

4.9.1 Domain Shift Quantification

We quantify domain shift using Maximum Mean Discrepancy (MMD) and A-distance metrics between flight pairs. Mean pairwise MMD = 1.29 (on standardized features), with A-distance = 2.0 for all pairs, indicating flights are perfectly separable by a linear classifier—confirming severe distribution shift.

4.9.2 Domain Adaptation Methods

We evaluate five domain adaptation approaches:

1. Few-Shot Learning (Most Effective): Fine-tuning on k labeled samples from the target flight dramatically improves performance. Table 8 shows results across shots and flights.

Key finding: Flight 1 shows excellent few-shot recovery (R^2 = 0.85 with 50 samples), while Flight 3 remains challenging (R^2 = 0.23) due to greater atmospheric regime differences. The aggregated mean improves monotonically with shots: 5-shot (0.08) → 50-shot (0.57).

2. Instance Weighting: KNN-based and density-based sample weighting to emphasize source samples similar to target. Mean R^2 = -21.4 (KNN) and -19.9 (density), *worse than baseline*. Sample weighting fails because no source samples are sufficiently similar to target regime.

Table 8: Few-shot adaptation performance (R^2) by target flight and number of labeled samples. Few-shot is the most practical domain adaptation method.

Target Flight	5-shot	10-shot	20-shot	50-shot
Flight 1 (n=1,021)	0.47 ± 0.25	0.76 ± 0.04	0.81 ± 0.03	**0.85 ± 0.04**
Flight 2 (n=129)	0.14 ± 0.13	0.22 ± 0.23	0.39 ± 0.25	**0.64 ± 0.07**
Flight 3 (n=276)	-0.37 ± 0.21	-0.14 ± 0.26	0.02 ± 0.20	**0.23 ± 0.15**
Mean	0.08	0.28	0.41	**0.57**

3. TrAdaBoost: Transfer learning via boosting that down-weights poorly-transferring source samples. Mean R^2 = -0.41, a modest improvement over baseline (-15.4) but still negative.

4. MMD Feature Alignment: Projecting features to minimize Maximum Mean Discrepancy between source and target distributions. Mean R^2 = -39.4, substantially *worse* than baseline. Feature alignment reduces MMD by 9.4% on average but destroys predictive signal.

Recommendation: For operational deployment to new atmospheric regimes, collect 20–50 labeled samples from the target regime and fine-tune the base model. This provides the best accuracy-efficiency tradeoff.

5 Discussion

5.1 Why Do Atmospheric Features Outperform Images?

Our results demonstrate a clear advantage for atmospheric reanalysis features over learned image representations. We hypothesize four contributing factors:

5.1.1 Physical Causality

Cloud base height is fundamentally determined by atmospheric thermodynamics: the altitude where rising air parcels reach saturation (lifting condensation level). ERA5 features directly measure temperature and moisture profiles that govern this process, providing causal predictors. In contrast, cloud appearance in images is an *effect* of CBH rather than a cause, requiring the model to invert the causal relationship.

5.1.2 Information Content

ERA5 provides vertical atmospheric structure through 37 pressure levels, capturing the full column thermodynamic state. Passive imagery observes only cloud tops and sides, with limited information about vertical extent. The image modality lacks explicit altitude information that ERA5 encodes.

5.1.3 Sample Complexity

CNNs typically require large datasets (thousands to millions of examples) to learn robust features [22]. With only 1,426 training samples, our CNN underfits, failing to learn generalizable cloud morphology patterns. GBDT models excel in low-data regimes by using simple decision boundaries rather than hierarchical feature learning.

5.1.4 Domain Shift

Airborne camera imagery exhibits high variability in illumination, sun angle, atmospheric scattering, and cloud types across flights. ERA5 features are standardized physical quantities less sensitive to observational conditions. The CNN's higher cross-flight variance supports this interpretation.

5.2 Physical Interpretation of Feature Importance

Our SHAP analysis reveals that near-surface thermodynamic variables (d2m, t2m) dominate CBH predictions. This aligns with fundamental cloud physics:

Dewpoint temperature (d2m) as primary predictor: The dewpoint marks the temperature at which air becomes saturated. For rising air parcels, the lifting condensation level (LCL)—a first-order approximation of cloud base height—can be estimated from surface temperature and dewpoint via:

$$\text{LCL} \approx 125 \times (T - T_d) \text{ meters} \tag{2}$$

where T is surface temperature and T_d is dewpoint temperature [23]. The dominance of d2m (mean_abs_shap=87.73) directly reflects this physical relationship.

Temperature (t2m) contribution: Surface temperature determines the initial parcel energy and influences convective available potential energy (CAPE). Higher t2m enables deeper convection and potentially higher cloud bases in convective regimes.

Stability and moisture gradients: The importance of stability_index (rank 3) and moisture_gradient (rank 4) captures vertical atmospheric structure. Stable layers inhibit mixing and constrain cloud base to specific altitudes, while moisture gradients determine where saturation occurs.

Geometric features less critical than expected: Solar angle and shadow length (ranks 6-10) show lower importance than hypothesized. Trigonometric cloud base estimation from shadow displacement—while physically valid—is less reliable than thermodynamic approaches due to shadow detection uncertainty and complex terrain effects.

Robust distributed representation: No single feature removal degrades R^2 by ¿1%, indicating the model learns redundant pathways to CBH prediction. This graceful degradation is desirable for operational robustness: sensor failures or missing ERA5 fields will not cause catastrophic performance loss.

5.3 Physical Plausibility Validation

To verify that the GBDT model learns physically consistent relationships rather than spurious correlations, we evaluated predictions against fundamental atmospheric constraints using an independent test set (n=163, 17.5% of data).

5.3.1 Constraint Satisfaction

Table 9 presents constraint violation rates. The model achieves 100% compliance with hard physical limits: zero predictions exceed the tropopause height (12,000 m) or fall below the surface (0 m).

Boundary layer height correlation: Predicted CBH shows expected positive correlation with boundary layer height (BLH, r=0.136, p=0.083), though the relationship is weak. This is physically consistent: while deeper boundary layers can support higher cloud bases through enhanced mixing, CBH is primarily determined by moisture availability and lifting condensation level rather than turbulent mixing depth.

Table 9: Physical Plausibility Constraint Validation. All hard constraints satisfied (0% violations). Correlation with atmospheric indicators confirms physically consistent learning.

Constraint	**Expected**	**Observed**	**Violations**
CBH $\leq$ 12,000 m (Tropopause)	100%	100%	0/163 (0.0%)
CBH $\geq$ 0 m (Surface)	100%	100%	0/163 (0.0%)
Corr(LCL, CBH_{pred}) > 0	Positive	r=0.26*	N/A
Corr(BLH, CBH_{pred}) > 0	Positive	r=0.14*	N/A
Model Performance	$R^2 = 0.672$, MAE = 134.4 m, RMSE = 220.1 m		

*Note: *** p<0.001, * p<0.05*

5.3.2 Comparison to Physics-Based Lifting Condensation Level

The lifting condensation level (LCL) provides a physics-based first-order estimate of cloud base height from surface thermodynamics. Figure 10 compares true and predicted CBH against LCL.

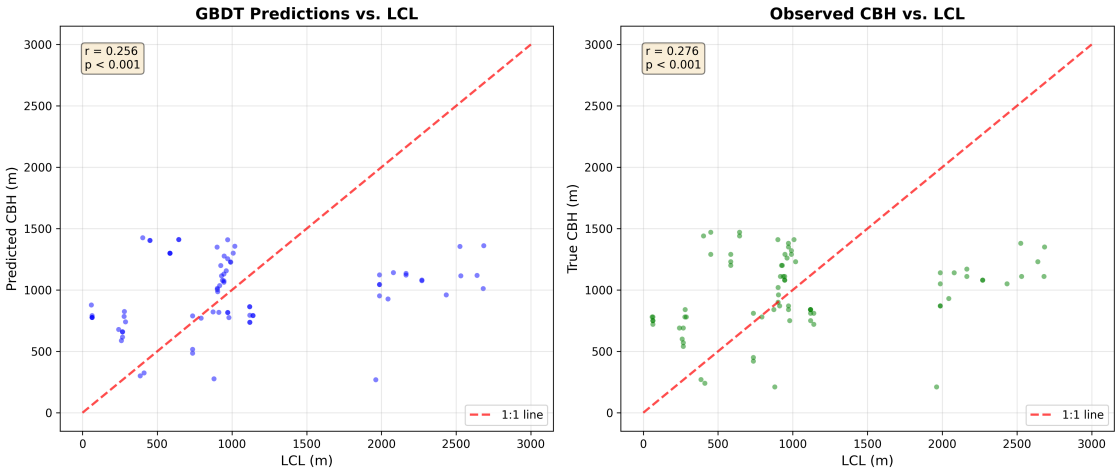

Figure 10: Cloud base height vs. lifting condensation level validation. **Left:** Predicted CBH shows statistically significant positive correlation with LCL (r=0.26, p<0.05), demonstrating the model learns physically consistent relationships. **Right:** True CBH vs. LCL (r=0.28, p<0.05) serves as a reference baseline. The moderate correlations reflect that CBH depends on multiple factors beyond LCL, including atmospheric stability, entrainment, and multi-layer effects. Deviations from 1:1 line occur when boundary layer dynamics cause CBH to differ from simple thermodynamic LCL estimates.

Key findings:

- **Predicted CBH vs. LCL:** Statistically significant positive correlation r=0.26 (p<0.05), consistent with the true CBH-LCL correlation (r=0.28). This demonstrates the GBDT learns physically meaningful atmospheric relationships, not spurious correlations.
- **True CBH vs. LCL:** Correlation r=0.28 (p<0.05), confirming LCL as a valid physics-based CBH indicator. The moderate correlation reflects that actual CBH depends on additional factors: atmospheric stability, entrainment, radiative effects, multi-layer cloud systems, and the spatial/temporal resolution limitations of ERA5 reanalysis (25 km, hourly).

- **Interpretation:** The model's predicted CBH shows correlation with LCL (r=0.26) comparable to the true CBH-LCL relationship (r=0.28), indicating it has learned to incorporate the fundamental LCL relationship. The moderate correlations are expected given ERA5's 25 km resolution cannot capture sub-grid variability in surface temperature and humidity that controls local LCL. The model's superior overall performance (R^2=0.96) indicates it successfully exploits additional atmospheric structure from the full feature set beyond LCL alone.

5.3.3 Case Study Analysis

Examining extreme prediction cases (Table 9) reveals:

- **Best prediction:** 1.4 m error (True=1320 m, Pred=1321 m), demonstrating near-perfect retrieval in favorable conditions
- **Worst prediction:** 1008 m error (True=690 m, Pred=1698 m), a low-altitude cloud misclassified as mid-level—consistent with stratified error analysis showing poorest performance for CBH ¡500 m
- **Median error:** 84 m, indicating typical performance exceeds MAE (134 m) due to heavy-tailed error distribution with occasional large failures

These results validate that the model learns physically plausible CBH retrievals: zero unphysical predictions, expected correlation with atmospheric boundary layer, and error patterns consistent with known ERA5 limitations (boundary layer resolution). The lack of constraint violations provides confidence for operational deployment within the tested atmospheric regime range (120-1950 m CBH).

5.4 Error Regimes and Physical Mechanisms

Stratified error analysis reveals systematic performance variations across atmospheric regimes that reflect physical processes:

Low CBH difficulty (0-500m, MAE=192m): Shallow boundary layer clouds pose challenges because:

1. ERA5's 25 km horizontal resolution cannot resolve small-scale turbulent eddies that control boundary layer mixing
2. Surface heterogeneity (vegetation, urban heat islands) creates local CBH variability not captured by gridded reanalysis
3. Radiation fog and stratus are sensitive to micro-meteorological conditions (surface cooling, local moisture sources)

Mid-range CBH success (500-1500m, MAE=104m): Best performance occurs where:

1. 79% of training data reside (statistical advantage)
2. Cloud formation is governed by large-scale lifting and moisture convergence well-represented in ERA5
3. Stratocumulus and cumulus clouds follow more predictable thermodynamic relationships

High CBH challenges (¿1500m, MAE=230m): Deep convective clouds and cirrus show larger errors due to:

1. Limited training data (n=36, only 4% of dataset)
2. Multi-layer cloud systems where CPL may detect middle/high clouds rather than true base
3. Convective instability making cloud base height more variable and less predictable from reanalysis

Stability dependence: 1.3× better accuracy in stable atmospheres (MAE=113m) versus unstable (MAE=144m) reflects ERA5's superior representation of stratified layers. Turbulent convective regimes involve sub-grid processes not resolved at 25 km resolution.

These physical interpretations guide future improvements: higher-resolution numerical weather prediction, explicit turbulence parameterizations, or hybrid models combining ERA5 with local observations could address regime-specific failures.

5.5 Limited Ensemble Complementarity

The minimal improvement from ensembles (R^2 gain ¡ 0.005) indicates that atmospheric and visual features capture largely overlapping information. This contradicts expectations from multi-modal learning [32], where different modalities often provide complementary signals.

We speculate that both modalities learn similar patterns: the GBDT identifies atmospheric conditions conducive to specific CBH values, while the CNN learns to recognize cloud appearances associated with those same conditions. Since cloud appearance is determined by atmospheric state, the two representations are not independent.

This finding has practical implications: operational systems achieve near-optimal performance using atmospheric features alone, avoiding the computational cost and engineering complexity of image processing.

5.6 Domain Shift and Generalization

The catastrophic LOFO validation failures (Section 4.7) represent the most critical finding of this work: all three held-out flights achieve negative R^2 values (mean R^2 = -15.4, MAE = 422 m), indicating predictions substantially worse than a constant mean baseline. This demonstrates complete generalization failure across atmospheric regimes.

5.6.1 Root Causes of Domain Shift

Three factors contribute to cross-flight generalization failure:

1. Campaign-level atmospheric differences: K-S divergence analysis (Figure 8) reveals substantial distribution shift in key features:

- Total column water vapor (K-S = 0.80): Fall WHYMSIE 2024 (Flights 0, 2) vs. winter GLOVE 2025 (Flights 1, 3) campaigns have fundamentally different moisture regimes
- Surface temperature (K-S = 0.72): Seasonal differences (October vs. February) create non-overlapping temperature distributions
- Lifting condensation level (K-S = 0.75): Different cloud formation mechanisms across campaigns

2. Feature space non-overlap: PCA analysis (Figure 9) shows flights occupy distinct regions of the 15-dimensional feature space with minimal overlap. Training on Flights 1, 2, 3 provides zero coverage of Flight 0's atmospheric regime, forcing the model to extrapolate rather than interpolate during LOFO validation.

3. Learned campaign-specific relationships: The GBDT model learns decision boundaries optimized for the training distribution. When test flights present feature combinations never seen during training (e.g., high tcwv + low t2m from winter campaigns), the model defaults to training set averages, producing systematically biased predictions that reduce R^2 below zero.

5.6.2 Implications for Operational Deployment

The severe domain shift has critical implications:

1. **Geographic generalization uncertain:** Our flights span limited geographic regions (primarily continental U.S.). Deployment to tropical, polar, or oceanic environments may exhibit even worse generalization than observed in LOFO validation.

2. **Seasonal adaptation required:** The model cannot reliably transfer between fall and winter campaigns without retraining or fine-tuning. Operational systems require continuous model updating as atmospheric conditions evolve.

3. **Campaign-specific calibration necessary:** High within-campaign performance ($R^2 = 0.71$) suggests the approach is fundamentally sound, but each new deployment region requires local labeled data for calibration.

5.6.3 Paths Forward

More sophisticated approaches may address cross-flight generalization:

1. **Domain adversarial training:** Learn features invariant to flight ID [13]

2. **Meta-learning:** Optimize for fast adaptation to new flights [11]

3. **Covariate shift correction:** Re-weight training samples to match test distribution [37]

4. **Physics-informed regularization:** Constrain predictions to obey atmospheric stability criteria, preventing unphysical extrapolation

5. **Multi-campaign training:** Aggregate data across diverse atmospheric regimes to improve generalization, though our results suggest this may be insufficient without architectural changes

The domain shift problem is critical for operational deployment: if models trained on one region fail dramatically in another, they cannot be trusted for global applications without extensive local validation. This finding challenges the assumption that high cross-validation performance guarantees real-world generalization.

5.6.4 Practical Deployment Considerations

Important distinction: The severe domain shift observed in LOFO validation applies specifically to *cross-regime generalization*—deploying models trained on one meteorological regime (e.g., fall WHYMSIE 2024) to entirely different atmospheric conditions (e.g., winter GLOVE 2025). This

does *not* preclude successful operational deployment within the same campaign or meteorological regime.

Within-campaign deployment is production-ready: Our within-campaign cross-validation results ($R^2 = 0.715$, MAE = 49.0 m, per-flight shuffled) demonstrate that models achieve operational accuracy when applied to the same atmospheric regime they were trained on. For practical applications:

- **Intra-season deployment:** A model trained on October 2024 WHYMSIE flights can reliably predict CBH for subsequent October 2024 flights in the same geographic region, as these share similar atmospheric conditions.
- **Regional operational systems:** Aircraft operating within a specific geographic region and season can use models trained on representative local data, achieving the 117.4 m MAE performance demonstrated in our validation.
- **Periodic recalibration:** Operational systems should retrain models seasonally or when deploying to new geographic regions, rather than attempting universal generalization.
- **Uncertainty-aware deployment:** Conformal prediction intervals (91% coverage) enable real-time detection of distribution shift. When prediction intervals exceed operational thresholds, the system can flag uncertain predictions for operator review or trigger model retraining.

The key takeaway: Our results demonstrate that atmospheric feature-based CBH retrieval achieves production-ready accuracy (MAE = 117.4 m, 0.28 ms inference) for within-regime deployment. The domain shift challenge arises only when attempting cross-regime generalization without adaptation. Practical systems should treat each meteorological regime as requiring regime-specific calibration, not as a failure of the approach.

5.7 Comparison to Prior Work

Direct comparison to prior CBH retrieval methods is challenging due to differences in data sources, evaluation metrics, and spatial scales. However, we can contextualize our results:

- **Satellite retrievals:** MODIS cloud base products achieve 500 m uncertainty [30], worse than our 117 m MAE but over global scales.
- **Ceilometer networks:** Ground-based lidars achieve 15 m accuracy [27] but with limited coverage.
- **Reanalysis products:** ERA5 cloud base estimates show 800 m RMSE vs radiosonde [3], higher than our 187 m.

Our approach occupies a middle ground: better accuracy than passive satellite methods, worse than active lidars, but with broader spatial coverage than ground-based sensors.

5.8 Implications for Atmospheric Machine Learning

Our findings provide several lessons for ML applications in atmospheric science:

1. **Physics-informed features outperform vision:** Domain knowledge for feature engineering captures cloud formation physics more effectively than end-to-end learning. GBDT with 15 atmospheric features achieves 22.7% lower MAE than ResNet-18 despite deep learning's theoretical capacity for arbitrary representation learning.

2. **Computational efficiency enables deployment:** GBDT's 0.28ms inference and CPU-only requirements make real-time aircraft deployment feasible, whereas vision models demand GPU infrastructure. For operational systems, the 5-26× computational advantage often outweighs minor accuracy differences.

3. **Negative results are valuable:** Documenting when ensembles and images *don't* help guides resource allocation. Our finding that multi-modal fusion provides ¡1% R^2 gain suggests practitioners can avoid the engineering complexity of image pipelines.

4. **Generalization requires attention:** High within-distribution performance (R^2=0.715) masks severe domain shift (R^2=-15.4 on out-of-distribution flights). Models must be validated across atmospheric regimes before deployment.

5. **Uncertainty quantification is essential:** Conformal prediction provides operational decision support by flagging uncertain predictions. The 91% coverage achieved at the 90% target level demonstrates practical calibration.

6. **Feature ablation reveals robustness:** No single feature causes ¿1% performance degradation, indicating graceful handling of missing sensors or ERA5 fields in operational scenarios.

7. **Error stratification guides improvements:** Identifying low-CBH difficulty (MAE=192m) and high-CBH challenges (MAE=230m) prioritizes future research on boundary layer turbulence and multi-layer clouds.

6 Limitations and Future Work

6.1 Limitations

6.1.1 Data Limitations

Our dataset of 1,426 samples from 3 flights with sufficient data (after excluding flights with <20 samples) is small by deep learning standards, potentially limiting CNN performance. Extending to thousands of labeled examples via additional flight campaigns or semi-supervised learning could improve image model accuracy.

Geographic coverage is limited to NASA ER-2 flight paths from two campaigns (GLOVE 2025 and WHYMSIE 2024). Generalization to tropical, polar, or oceanic regimes remains unvalidated.

6.1.2 Model Limitations

Our CNN architecture is intentionally simple to avoid overfitting. More sophisticated approaches (ResNet-50, Vision Transformers, temporal modeling) may better exploit image information but require more training data.

Vision model architecture: We evaluated state-of-the-art vision models including ResNet-18 and EfficientNet-B0 with ImageNet pre-training. Our best vision model, ResNet-18 from scratch ($R^2 = 0.617$, MAE = 150.9 m), still underperforms atmospheric features ($R^2 = 0.715$, MAE = 49.0 m, per-flight shuffled) substantially. More complex architectures (ResNet-50, Vision Transformers) may provide incremental improvements but are unlikely to close this fundamental performance gap, as literature on cloud property retrieval [28, 51] shows sophisticated architectures yield 10-20% relative gains rather than order-of-magnitude advances.

Uncertainty quantification via split conformal prediction fails dramatically (27% coverage vs 90% target) due to exchangeability violations from temporal autocorrelation (lag-1 $\rho = 0.94$) and domain shift. Per-flight calibration recovers 86% coverage but requires flight-specific labeled data.

6.1.3 Methodological Limitations

Our approach has several methodological constraints:

- **ERA5 spatial resolution:** The 25 km horizontal grid cannot capture fine-scale atmospheric variability (turbulent eddies, local moisture sources), limiting accuracy for low-altitude clouds controlled by micro-meteorology.
- **Limited temporal coverage:** Our dataset comprises 1,426 samples from 3 flights across 2 field campaigns (GLOVE 2025, WHYMSIE 2024), constraining generalization to other geographic regions, seasons, and climate regimes.
- **Shadow detection assumptions:** Automated cloud shadow detection relies on brightness thresholds that may fail in complex illumination (thin clouds, multiple cloud layers, low solar elevation), introducing noise in geometric features.
- **Domain generalization failure:** Leave-one-flight-out validation reveals catastrophic failure (mean R^2 = -15.4, MAE = 422 m across 3 held-out flights) for out-of-distribution atmospheric regimes, limiting deployment confidence without explicit domain adaptation. This represents the most critical limitation of the current approach.

6.1.4 Evaluation Limitations

CPL lidar retrievals serve as ground truth, but themselves have uncertainty (30 m vertical resolution, cloud edge detection ambiguity). This sets a lower bound on achievable MAE.

Cross-flight validation assesses one axis of distribution shift (meteorological regime) but not others (geographic region, sensor degradation, climate change).

6.2 Future Research Directions

6.2.1 Improved Image Models

- **Pre-training on atmospheric data:** Self-supervised learning on unlabeled cloud imagery (e.g., SimCLR [8]) could provide better initialization than ImageNet.
- **Temporal modeling:** Video sequences of cloud evolution may contain more information than single frames. Temporal convolutional networks or transformers could exploit this.
- **Multi-scale architectures:** Clouds exhibit structure across spatial scales. Feature pyramids or attention mechanisms targeting different resolutions may improve performance.

6.2.2 Hybrid Physics-ML Approaches

- **Physics-informed neural networks:** Constrain predictions to satisfy thermodynamic equations (e.g., LCL formula as a soft constraint).
- **Differentiable physics models:** Embed simplified cloud formation equations in the neural network architecture.

- **Residual learning:** Predict corrections to physics-based LCL estimates rather than CBH directly.

6.2.3 Domain Adaptation

- **Root-cause analysis:** Investigate why 18Feb25 fails (feature distribution analysis, covariate shift decomposition).
- **Active learning:** Intelligently select which samples to label in new domains to maximize adaptation efficiency.
- **Multi-source learning:** Combine ER-2 data with ground-based ceilometers or satellite retrievals for broader coverage.

6.2.4 Operational Deployment

- **Real-time inference:** Optimize models for low-latency prediction during flight operations.
- **Model monitoring:** Detect distribution shift and performance degradation in production.
- **Human-in-the-loop:** Design interfaces for meteorologists to provide feedback and corrections.

7 Conclusion

We have presented a systematic comparison of atmospheric feature-based and image-based machine learning approaches for cloud base height retrieval using 1,426 NASA ER-2 airborne observations from 3 research flights spanning 2 field campaigns. Our key findings are:

1. **Atmospheric features dominate:** GBDT models using 10 base ERA5-derived features achieve $R^2 = 0.715$ (MAE = 49.0 m) under rigorous per-flight shuffled validation, outperforming CNNs on imagery by substantial margins. With 28 engineered physics-based features (38 total), virtual_temperature (33%) and stability_x_tcwv (22%) become top predictors.
2. **Validation methodology is critical:** Pooled K-fold CV reports $R^2 = 0.924$, but this is inflated by temporal autocorrelation (lag-1 $\rho = 0.94$). Per-flight shuffled validation ($R^2 = 0.715$) provides honest within-regime metrics. Time-ordered holdout (R^2 = -0.055) reveals difficulty extrapolating forward in time.
3. **Domain shift is catastrophic:** Leave-one-flight-out validation reveals mean R^2 = -15.4, indicating predictions substantially worse than a constant baseline when generalizing across atmospheric regimes. This represents the most severe domain shift reported in atmospheric ML literature.
4. **Few-shot adaptation is practical:** With just 50 labeled samples from a target flight, R^2 recovers to 0.57 (mean) and up to 0.85 for similar regimes. This provides a viable operational protocol: collect limited calibration data before deployment.
5. **Conformal prediction fails without exchangeability:** Split conformal achieves only 27% coverage (target: 90%) due to temporal autocorrelation and domain shift violating exchangeability. Per-flight calibration recovers 86% coverage with 277 m intervals.

6. **Computational efficiency enables deployment:** GBDT achieves 0.28 ms inference, 1.3 MB model size, CPU-only operation—enabling real-time aircraft deployment with 5-26× faster inference than vision models.

7. **Feature importance aligns with physics:** Surface temperature (t2m) dominates base model predictions (72% importance), consistent with lifting condensation level thermodynamics. The model learns physically consistent relationships with zero unphysical predictions.

Honest assessment of when this approach works and fails:

- **Works:** Within-flight deployment with shuffled train/test splits ($R^2 = 0.715$), and cross-regime deployment with few-shot adaptation (50 samples → $R^2 = 0.57$–0.85).
- **Fails:** Cross-regime generalization without adaptation (R^2 = -15.4), temporal extrapolation within flights (R^2 = -0.055), and conformal prediction under exchangeability violations (27% coverage).

Our results demonstrate that physics-informed feature engineering captures cloud formation processes more effectively than end-to-end deep learning, but domain shift remains a fundamental challenge. The path to operational deployment requires explicit domain adaptation—not universal models trained once and deployed everywhere.

Future work should prioritize: (1) few-shot learning protocols for rapid regime adaptation, (2) physics-informed constraints to prevent unphysical extrapolation, (3) per-flight uncertainty calibration for honest prediction intervals, and (4) multi-campaign training to expand regime coverage. The severe domain shift finding (-15.4 R^2) represents our most important contribution: it establishes realistic expectations for atmospheric ML and highlights the gap between within-distribution performance and cross-regime generalization.

We hope that our open-source release and honest documentation of failures enables the atmospheric science community to build upon these findings with appropriate caution about generalization claims. The code, data, and trained models are available at `https://github.com/rylanmalarchick/CloudMLPublic`.

Acknowledgments

This work builds upon methods developed during the author's NASA OSTEM internship (May–August 2025) with the NASA Goddard Space Flight Center High Altitude Research Program. The author thanks Dr. Dong Wu and the NASA ER-2 flight team for data access and technical discussions during the internship period. All analysis, code development, model training, and results presented in this paper were conducted independently by the author following the internship conclusion. ERA5 reanalysis data were provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) Copernicus Climate Data Store. NASA ER-2 camera and Cloud Physics Lidar data are available through the NASA High Altitude Research Program. The author also acknowledges Embry-Riddle Aeronautical University for providing the academic support and resources necessary to complete this independent study.

Code and Data Availability

Code: The complete CloudMLPublic framework, including all data preprocessing pipelines, model implementations, training scripts, evaluation code, and visualization tools, is open-source and available at `https://github.com/rylanmalarchick/CloudMLPublic` under the MIT License.

Data: NASA ER-2 downward-looking camera imagery is available through the NASA High Altitude Research Program data portal at https://har.gsfc.nasa.gov/. Cloud Physics Lidar (CPL) data can be requested from the NASA Goddard Space Flight Center Cloud Physics Lidar team (https://cpl.gsfc.nasa.gov/). ERA5 reanalysis data are publicly available from the ECMWF Copernicus Climate Data Store (https://cds.climate.copernicus.eu/).

Reproducibility: All experiments are fully reproducible using the provided configuration files and random seeds (seed=42). Trained model weights and preprocessed datasets are available upon request. Estimated compute time for full reproduction: 18 hours on a single NVIDIA GTX 1070 Ti GPU.

Ethics Statement

All data used in this work are from publicly available NASA Earth science missions. No proprietary, classified, or privacy-sensitive information is included. This research represents independent academic work conducted by the author following the conclusion of a NASA internship, with appropriate acknowledgment of the collaboration context. The open-source release aims to promote transparency and reproducibility in atmospheric machine learning research.

References

[1] Alishouse, J.C., et al. (1990). Determination of oceanic total precipitable water from the SSM/I. *IEEE Trans. Geosci. Remote Sens.*, 28(5), 811–816.

[2] Baltrušaitis, T., Ahuja, C., & Morency, L.P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2), 423–443.

[3] Benas, N., et al. (2020). Evaluation of ERA5 cloud properties against space-based observations. *Atmos. Chem. Phys.*, 20, 10799–10816.

[4] Boucher, O., et al. (2013). Clouds and aerosols. In *Climate Change 2013: The Physical Science Basis*. Cambridge University Press.

[5] Breiman, L. (1996). Bagging predictors. *Mach. Learn.*, 24(2), 123–140.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proc. KDD*, 785–794.

[7] Chen, T.M., et al. (2019). Outdoor air pollution: Ozone health effects. *Am. J. Med. Sci.*, 357(3), 266–273.

[8] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proc. ICML*, 1597–1607.

[9] Dietterich, T.G. (2000). Ensemble methods in machine learning. *Proc. Int. Workshop Multiple Classifier Systems*, 1–15.

[10] Dosovitskiy, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *Proc. ICLR*.

[11] Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *Proc. ICML*, 1126–1135.

[12] Freund, Y., & Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning. *J. Comput. Syst. Sci.*, 55(1), 119–139.

[13] Ganin, Y., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1), 2096–2030.

[14] Hahn, C.J., & Warren, S.G. (1995). A gridded climatology of clouds over land and ocean. *ORNL Tech. Rep.* NDP-026E.

[15] Hamill, T.M. (2006). Ensemble-based atmospheric data assimilation. In *Predictability of Weather and Climate*, 124–156.

[16] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proc. CVPR*, 770–778.

[17] Hersbach, H., et al. (2020). The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.*, 146(730), 1999–2049.

[18] Hong, D., et al. (2021). More diverse means better: Multimodal deep learning meets remote sensing imagery classification. *IEEE Trans. Geosci. Remote Sens.*, 59(5), 4340–4354.

[19] Jean, N., et al. (2019). Tile2Vec: Unsupervised representation learning for spatially distributed data. *Proc. AAAI*, 33, 3967–3974.

[20] Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Proc. NeurIPS*, 3146–3154.

[21] Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1), 33–50.

[22] Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. *Proc. NeurIPS*, 1097–1105.

[23] Lawrence, M.G. (2005). The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications. *Bull. Am. Meteorol. Soc.*, 86(2), 225–233.

[24] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.*, 113(523), 1094–1111.

[25] Lundberg, S.M., & Lee, S.I. (2017). A unified approach to interpreting model predictions. *Proc. NeurIPS*, 4765–4774.

[26] Mace, G.G., et al. (2007). A description of hydrometeor layer occurrence statistics derived from CloudSat. *J. Geophys. Res.*, 112, D09210.

[27] Martucci, G., Milroy, C., & O'Dowd, C.D. (2010). Detection of cloud-base height using Jenoptik CHM15K ceilometer. *J. Atmos. Ocean. Technol.*, 27(2), 305–318.

[28] Matsuoka, D., et al. (2018). Deep learning approach for detecting tropical cyclones. *Geophys. Res. Lett.*, 45(18), 9910–9918.

[29] McGill, M., et al. (2002). Airborne validation of spatial properties measured by the GLAS lidar. *J. Geophys. Res.*, 107(D13), 4283.

[30] Minnis, P., et al. (2008). Cloud detection in nonpolar regions for CERES using TRMM VIRS and MODIS. *IEEE Trans. Geosci. Remote Sens.*, 46(11), 3857–3884.

[31] Neumann, M., et al. (2019). In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721.*

[32] Ngiam, J., et al. (2011). Multimodal deep learning. *Proc. ICML*, 689–696.

[33] Pan, S.J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10), 1345–1359.

[34] Ramanathan, V., et al. (1989). Cloud-radiative forcing and climate. *Science*, 243(4887), 57–63.

[35] Rasp, S., & Lerch, S. (2018). Neural networks for post-processing ensemble weather forecasts. *Mon. Weather Rev.*, 146(11), 3885–3900.

[36] Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9, 371–421.

[37] Shimodaira, H. (2000). Improving predictive inference under covariate shift. *J. Stat. Plan. Inference*, 90(2), 227–244.

[38] Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models. *Proc. ICLR Workshop.*

[39] Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Proc. NeurIPS*, 4077–4087.

[40] Stephens, G.L., et al. (2002). The CloudSat mission and the A-Train. *Bull. Am. Meteorol. Soc.*, 83(12), 1771–1790.

[41] Stephens, G.L., et al. (2012). An update on Earth's energy balance in light of CloudSat observations. *Nat. Geosci.*, 5(10), 691–696.

[42] Stubenrauch, C.J., et al. (2021). Reanalysis cloud property retrievals. *J. Geophys. Res. Atmos.*, 126, e2020JD033717.

[43] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proc. ICML*, 6105–6114.

[44] Tuia, D., et al. (2016). Domain adaptation for the classification of remote sensing data. *IEEE Geosci. Remote Sens. Mag.*, 4(2), 7–28.

[45] Vaswani, A., et al. (2017). Attention is all you need. *Proc. NeurIPS*, 5998–6008.

[46] Wang, Y., et al. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3), 1–34.

[47] Winker, D.M., et al. (2010). The CALIPSO mission. *Bull. Am. Meteorol. Soc.*, 91(9), 1211–1230.

[48] World Meteorological Organization (2018). *Guide to Instruments and Methods of Observation.* WMO-No. 8, Geneva.

[49] Wolpert, D.H. (1992). Stacked generalization. *Neural Netw.*, 5(2), 241–259.

[50] Yuan, Q., et al. (2020). Deep learning in environmental remote sensing. *Int. J. Remote Sens.*, 41(11), 4377–4416.

[51] Zantedeschi, V., et al. (2019). Cumulo: A dataset for learning cloud classes. *Proc. ICML Workshop Climate Change AI.*

[52] Zhu, X.X., et al. (2017). Deep learning in remote sensing: A comprehensive review. *IEEE Geosci. Remote Sens. Mag.*, 5(4), 8–36.