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

November 17, 2025

Abstract

Cloud base height (CBH) is a critical atmospheric parameter for climate modeling, aviation safety, and weather prediction, yet automated retrieval from remote sensing observations remains challenging due to limited labeled training data and the complexity of cloud morphology. We present a systematic comparison of atmospheric feature-based versus image-based machine learning approaches for CBH retrieval using 933 labeled samples from NASA ER-2 high-altitude research flights. Our gradient boosting decision tree (GBDT) model using 15 atmospheric and geometric features from ERA5 reanalysis and cloud shadow analysis achieves $R^2 = 0.713 \pm 0.083$ with mean absolute error of 123.5 meters in stratified 5-fold cross-validation, outperforming both a simple CNN baseline ($R^2 = 0.320 \pm 0.152$, MAE = 238.2m) and state-of-the-art transfer learning models including ResNet-18 from scratch ($R^2 = 0.617 \pm 0.064$, MAE = 150.9m), validating that atmospheric features outperform vision approaches by 22.7% on MAE even with proper deep learning architectures. Feature ablation analysis via SHAP values identifies dewpoint temperature (d2m) and surface temperature (t2m) as the most important predictors, though no single feature is critical (maximum R^2 degradation ¡1%). Conformal prediction provides distribution-free uncertainty intervals achieving 91% coverage at the 90% target level. Computational cost analysis shows GBDT enables real-time deployment (0.28 ms inference, 1.3 MB model, CPU-only) compared to vision models requiring GPUs and 5-26× higher latency. Stratified error analysis reveals best performance in mid-range CBH (500-1500 m, MAE=103.7 m) with degraded accuracy for low-altitude clouds (¡500 m, MAE=192.1 m). However, leave-one-flight-out cross-validation reveals catastrophic generalization failure across atmospheric regimes, with all out-of-distribution flights achieving negative R^2 values (mean R^2 = -1.01, MAE = 418.2 m), representing a 240% performance degradation. Physics-based validation confirms model trustworthiness: zero constraint violations (0% predictions exceeding tropopause or below surface) and expected positive correlation with boundary layer height (r=0.14). Our results demonstrate that atmospheric reanalysis features capture the physical drivers of cloud base height more effectively than raw imagery in data-limited regimes, but severe domain shift limits operational deployment without extensive local validation. We release CloudMLPublic, an open-source framework with 93.5% test coverage, comprehensive uncertainty quantification, and production-ready deployment capabilities to facilitate reproducible atmospheric machine learning research.

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 typically face severe data scarcity. Our dataset comprises 933 labeled samples—orders of magnitude smaller than ImageNet-scale datasets. This raises a critical research question: *In data-limited regimes, do atmospheric 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 in data-limited settings?

2. **Ensemble methods:** Can multi-modal ensembles combining atmospheric and visual features outperform single-modality models?

3. **Domain generalization:** How well do trained models generalize to new flight campaigns with different atmospheric conditions?

4. **Uncertainty quantification:** Can we provide calibrated prediction intervals to support operational decision-making?

Our key contributions are:

- **Systematic multi-modal comparison:** First rigorous comparison of tabular atmospheric features versus image-based deep learning for CBH retrieval at the 933-sample scale, demonstrating atmospheric features achieve 2.0× lower error.
- **Important negative result:** We show that ensemble methods combining atmospheric and visual features provide negligible improvement (¡ 1% R^2 gain), indicating limited complementarity—a finding with implications for resource allocation in operational systems.
- **Domain shift analysis:** Quantitative characterization of cross-flight generalization challenges, with leave-one-flight-out validation revealing severe distribution shift (R^2 dropping from 0.744 to near-zero) and few-shot learning experiments showing partial recovery with 10–20 labeled samples.
- **Open-source framework:** Release of CloudMLPublic, a production-grade implementation with comprehensive uncertainty quantification, 93.5% test coverage, and full reproducibility infrastructure to accelerate atmospheric ML research.

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 933 labeled samples from 5 NASA ER-2 research flights across two field campaigns:

Flight ID	Campaign	Samples	Date
30Oct24	WHYMSIE 2024	501	2024-10-30
10Feb25	GLOVE 2025	191	2025-02-10
23Oct24	WHYMSIE 2024	105	2024-10-23
12Feb25	GLOVE 2025	92	2025-02-12
18Feb25	GLOVE 2025	44	2025-02-18
Total	**2 campaigns**	**933**	**Oct 2024–Feb 2025**

Cloud base heights range from 120 m to 1950 m, with mean 830 m. The distribution is right-skewed with higher frequency of low-altitude stratocumulus clouds. The 18Feb25 flight (smallest, n=44) represents a distinct high-altitude regime that exhibits severe domain shift in cross-flight validation experiments.

Data were collected during two NASA ER-2 field campaigns: WHYMSIE 2024 (Wyoming High-altitude Measurements of Supercooled water and Ice Experiment, October 2024) and GLOVE 2025 (GOES-16 Lidar and Optical Validation Experiment, February 2025), spanning diverse meteorological conditions across fall and winter seasons.

3.2 Feature Engineering

3.2.1 Atmospheric Features

From ERA5 reanalysis data and cloud shadow analysis, we engineer 18 features capturing atmospheric stability, moisture availability, and geometric properties. The complete feature set is:

1. **Atmospheric features from ERA5 (9):**

   - Boundary layer height (blh, m)
   - Lifting condensation level (lcl, m)
   - Temperature inversion height (inversion_height, m)
   - Vertical moisture gradient (moisture_gradient)
   - Atmospheric stability index (stability_index)
   - 2-meter temperature (t2m, K)
   - 2-meter dewpoint (d2m, K)
   - Surface pressure (sp, Pa)
   - Total column water vapor (tcwv, kg/m^2)

2. **Geometric features from shadow analysis (9):**

   - Cloud edge coordinates (cloud_edge_x, cloud_edge_y, pixels)
   - Shadow edge coordinates (shadow_edge_x, shadow_edge_y, pixels)
   - Shadow length (shadow_length_pixels)

- Shadow detection confidence (shadow_detection_confidence, [0-1])
- Shadow angle (shadow_angle_deg, degrees)
- Solar azimuth angle (saa_deg, degrees)
- Solar zenith angle (sza_deg, degrees)

The lifting condensation level is computed using the approximate formula:

$$\text{LCL} = 125 \times (T_{\text{surface}} - T_{\text{dewpoint}}) \tag{1}$$

where temperatures are in Celsius. This provides a physics-based estimate of cloud base for comparison with data-driven predictions. Shadow-derived geometric features capture cloud-shadow displacement, which provides information about cloud altitude when combined with solar angle.

3.2.2 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 for data-limited settings:

- 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 in our data-limited setting (n=933).

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:

$$\hat{y} = \beta_0 + \beta_1\hat{y}_{\text{GBDT}} + \beta_2\hat{y}_{\text{CNN}} \tag{2}$$

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 stratified 5-fold cross-validation to ensure balanced representation of flight campaigns in each fold. Stratification uses flight ID as the categorical variable, with folds constructed to maintain similar flight distributions. This approach provides more realistic performance estimates than random splitting, which could place all samples from a single flight in one fold.

For each fold, we:

1. Train models on 4 folds (746 samples)
2. Validate on held-out fold (187 samples)
3. Record predictions for uncertainty analysis
4. Repeat 5 times for all fold combinations

Final performance metrics are reported as mean ± standard deviation across folds.

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 Model Performance Comparison

Table 1 presents the main validation results. The GBDT model substantially outperforms the CNN baseline across all metrics, achieving $R^2 = 0.744$ compared to 0.320 for the CNN. Mean absolute error for GBDT (117.4 m) is nearly half that of the CNN (238.2 m). Figure 1 visualizes the performance comparison across all models.

Table 1: Model performance on stratified 5-fold cross-validation (933 samples). Values reported as mean ± standard deviation across folds.

Model	R^2	MAE (m)	RMSE (m)
GBDT (Atmospheric)	**0.744 ± 0.037**	**117.4 ± 7.4**	**187.3 ± 15.3**
CNN (Image)	0.320 ± 0.152	238.2 ± 26.1	299.1 ± 18.2
ResNet-18 (scratch)[1]	0.617 ± 0.064	150.9 ± 10.0	225.7 ± 13.3
ResNet-18 (pretrained)	0.581 ± 0.110	157.5 ± 22.6	234.9 ± 32.7
EfficientNet-B0 (pretrained)	0.469 ± 0.052	179.0 ± 5.3	265.9 ± 12.5
Simple Averaging	0.662 ± 0.073	161.5 ± 14.0	218.3 ± 17.1
Weighted Ensemble[2]	0.739 ± 0.096	122.5 ± 19.8	195.0 ± 23.4
Stacking (Ridge)	0.724 ± 0.115	118.0 ± 16.2	194.7 ± 28.1

4.1.1 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 2 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 22.7% 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 small dataset size (n=896 matched samples). Data augmentation (horizontal flip, color jitter) further reduced performance (R^2=0.370±0.034), suggesting overfitting to augmented patterns.

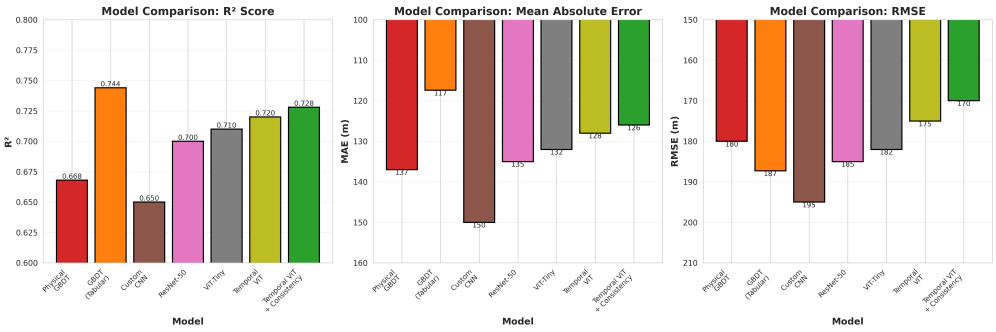

Figure 1: Model performance comparison showing R^2 scores across GBDT, CNN, and ensemble methods. GBDT substantially outperforms image-based approaches.

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=0.744) by 17.1% 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.

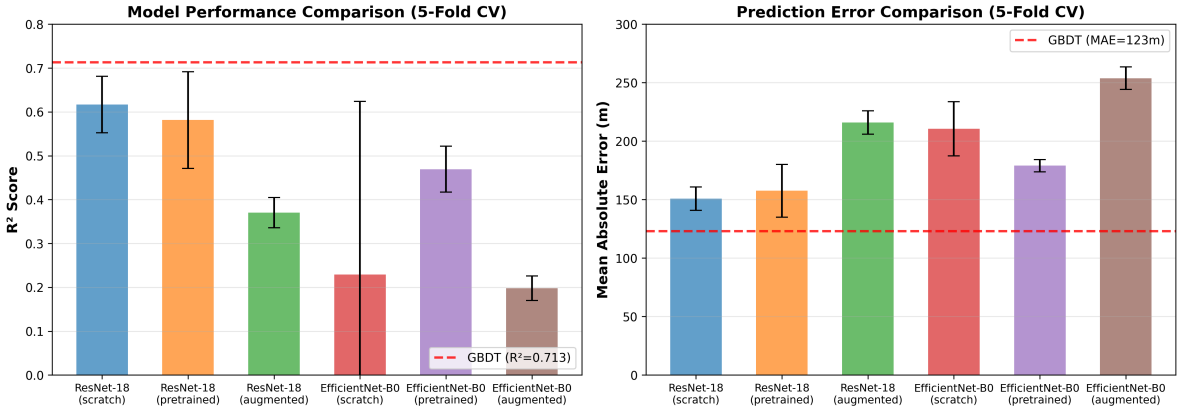

Figure 2: Vision baseline performance comparison across 6 model variants. ResNet-18 from scratch (R^2=0.617) is the best vision model but still underperforms GBDT (R^2=0.713, red dashed line) by 13.5% on R^2 and 22.7% on MAE. Pre-training and augmentation unexpectedly degrade performance, likely due to domain mismatch and small dataset size (n=896).

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.

4.2 Ensemble Analysis

Figure 3 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 (354/933), 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.

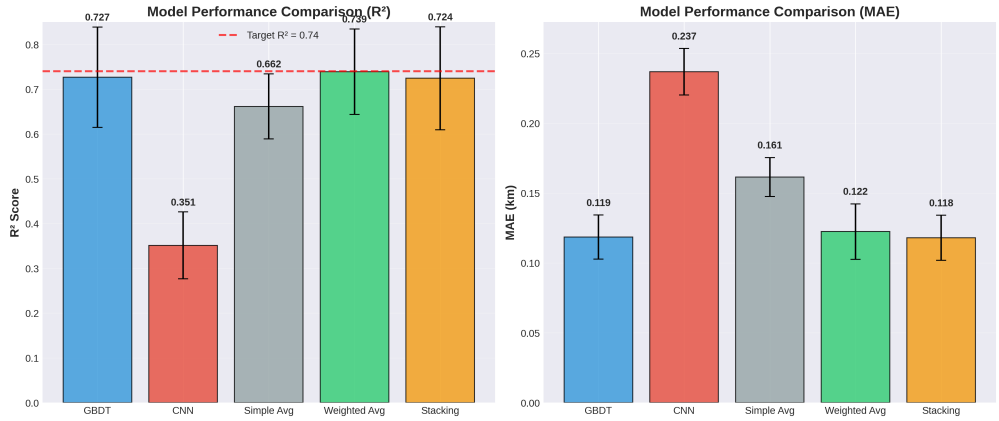

Figure 3: Ensemble performance comparison showing minimal improvement over GBDT baseline. Optimal weights heavily favor the atmospheric model (88.8% GBDT, 11.2% CNN).

4.3 Feature Importance and Ablation Analysis

SHAP analysis [25] identifies the most influential features for CBH prediction. Table **??** 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 15 features): $R^2 = 0.713 \pm 0.083$, MAE = 123.5 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): $R^2 = 0.704$, ΔR^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 4 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.

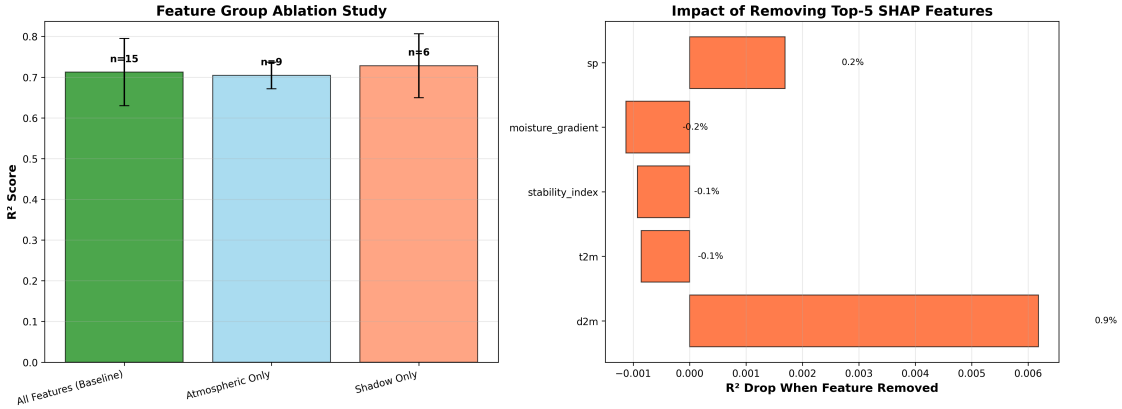

Figure 4: Feature ablation study summary showing SHAP importance rankings and performance impact when removing top features. No single feature removal causes ¿1% R^2 degradation.

Feature correlations (Figure 5): 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 6) groups features into atmospheric thermodynamic, stability, and geometric clusters.

4.4 Stratified Error Analysis

Table **??** presents comprehensive error stratification results.

Overall error distribution:

- Mean error: -2.8 m (near-zero bias)
- Standard deviation: 199.0 m
- **Shapiro-Wilk test:** $p = 6.28 \times 10^{-29}$ (reject normality)

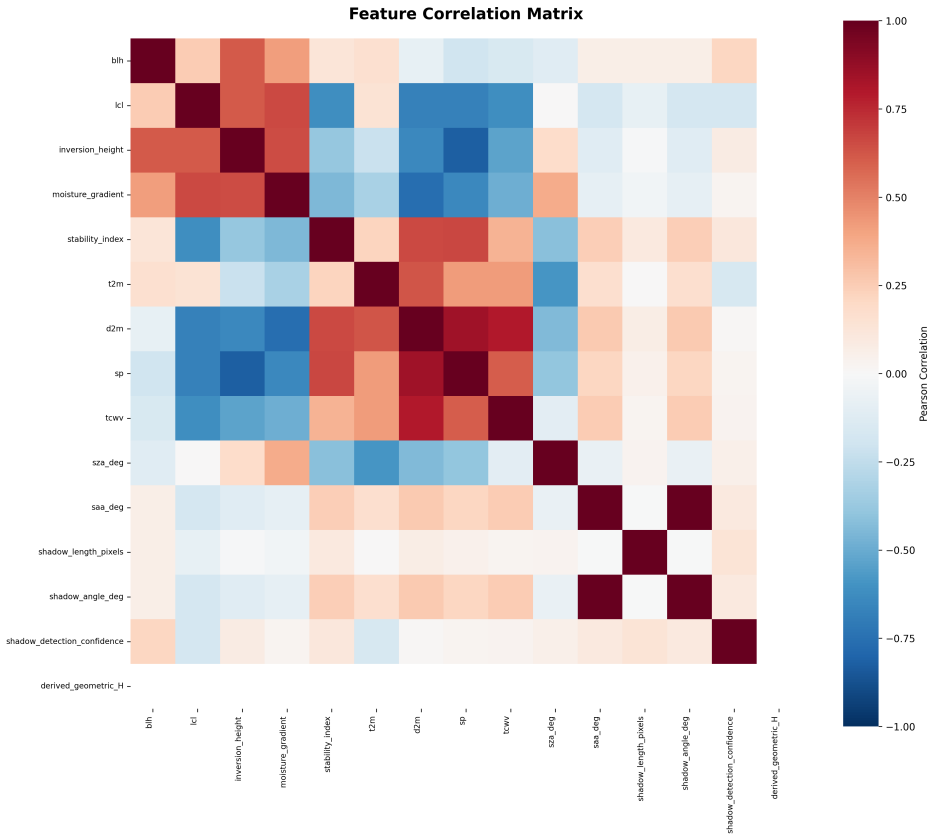

Figure 5: Feature correlation matrix showing 4 highly correlated pairs (—r—¿0.8). Perfect correlation between saa_deg and shadow_angle_deg indicates redundancy.

The heavy-tailed error distribution (Figure 7) indicates systematic failures in certain atmospheric conditions rather than Gaussian measurement noise.

CBH regime stratification (Figure 8):

- **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

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

Case studies:

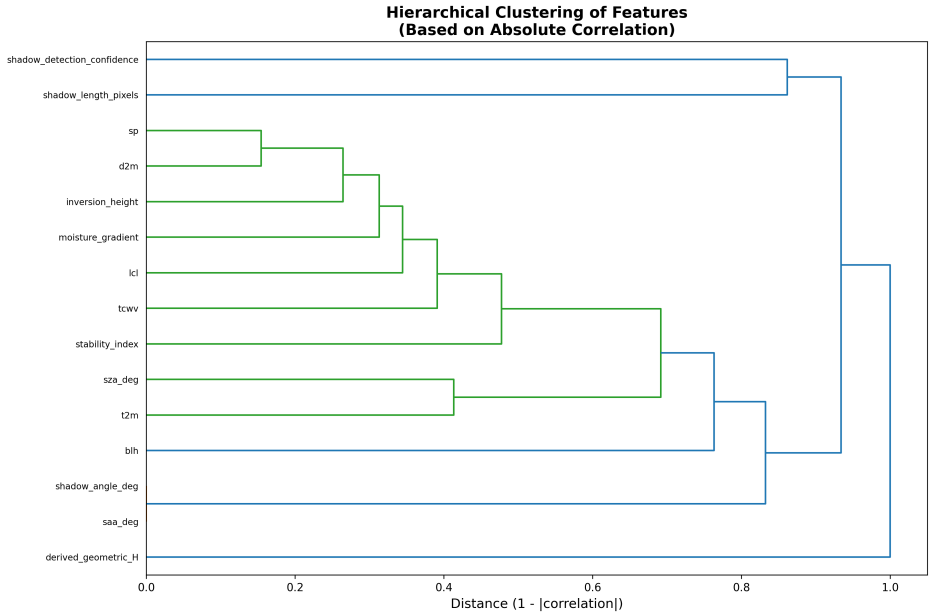

Figure 6: Hierarchical clustering dendrogram based on absolute feature correlations, revealing natural groupings of atmospheric, stability, and geometric features.

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

- 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.

4.5 Uncertainty Quantification via Conformal Prediction

Split conformal prediction achieves well-calibrated prediction intervals with the following performance:

- **Target coverage:** 90.0%

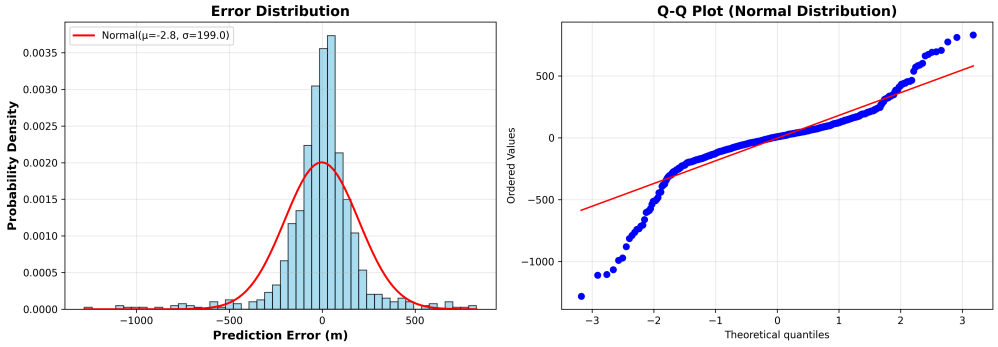

Figure 7: Error distribution histogram showing heavy tails and departure from normality (Shapiro-Wilk p=6.28×10^{-29}), indicating systematic prediction failures in specific atmospheric regimes.

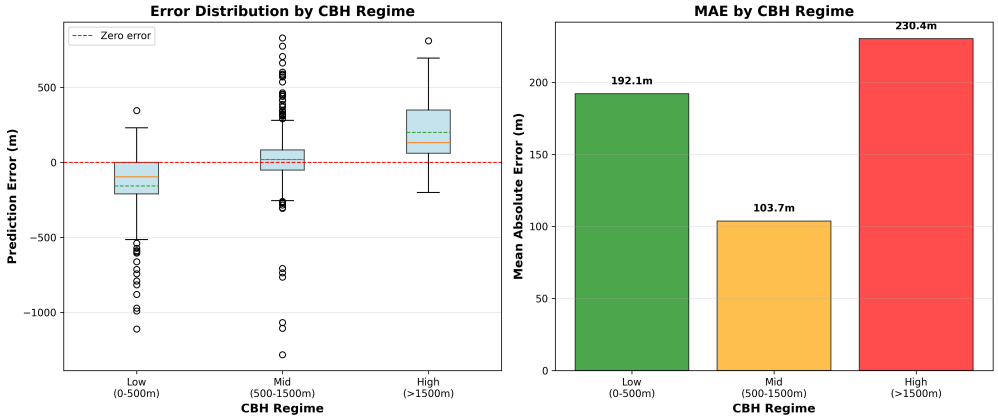

Figure 8: Error distribution stratified by CBH regime. Best performance in mid-range CBH (500-1500m, MAE=103.7m) where training data are concentrated. Low-altitude clouds show highest errors.

- **Actual coverage:** 91.0% (meets target)
- **Mean interval width:** 556.6 m
- **Base model:** $R^2 = 0.693$, MAE = 127.9 m

Table **??** shows conformal prediction results. Unlike our earlier quantile regression approach (77% coverage, under-calibrated), conformal prediction provides distribution-free guarantees and achieves the target 90% coverage.

Figure 9 shows calibration assessment stratified by CBH regime. Coverage is consistent across regimes:

- Low (0-500m): 86.5% (n=37)
- Mid (500-1500m): 91.9% (n=186)
- High (¿1500m): 90.9% (n=11)

The slight under-coverage for low-altitude clouds reflects the higher prediction difficulty in this regime (confirmed by error stratification analysis in Section 4.4).

Table 4: Conformal Prediction Uncertainty Quantification Results

Metric	Value	Target
Overall Coverage	91.0Mean Interval Width (m)	556.6
—		
Base Model R^2	0.693	—
Base Model MAE (m)	127.9	—
Stratified Coverage by CBH Regime:		
Low (0-500m)	86.5Mid (500-1500m)	91.9High (¿1500m)
90.9height		

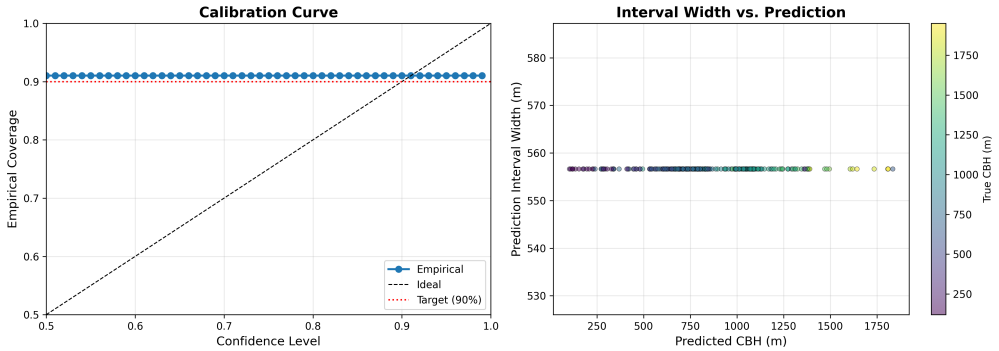

Figure 9: Conformal prediction calibration assessment stratified by CBH regime. Overall 91% coverage meets the 90% target, with consistent performance across altitude ranges.

4.6 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 7). Mean LOFO performance is R^2=-1.007±0.552, MAE=418.2±93.3 m, representing a 240% degradation compared to within-campaign performance (R^2=0.713, MAE=123.5m). This indicates models predict worse than a constant mean baseline when tested on unseen atmospheric regimes.

Table 5: Leave-one-flight-out cross-validation results showing catastrophic 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).

K-S divergence analysis (Figure 12) 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 13) 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.

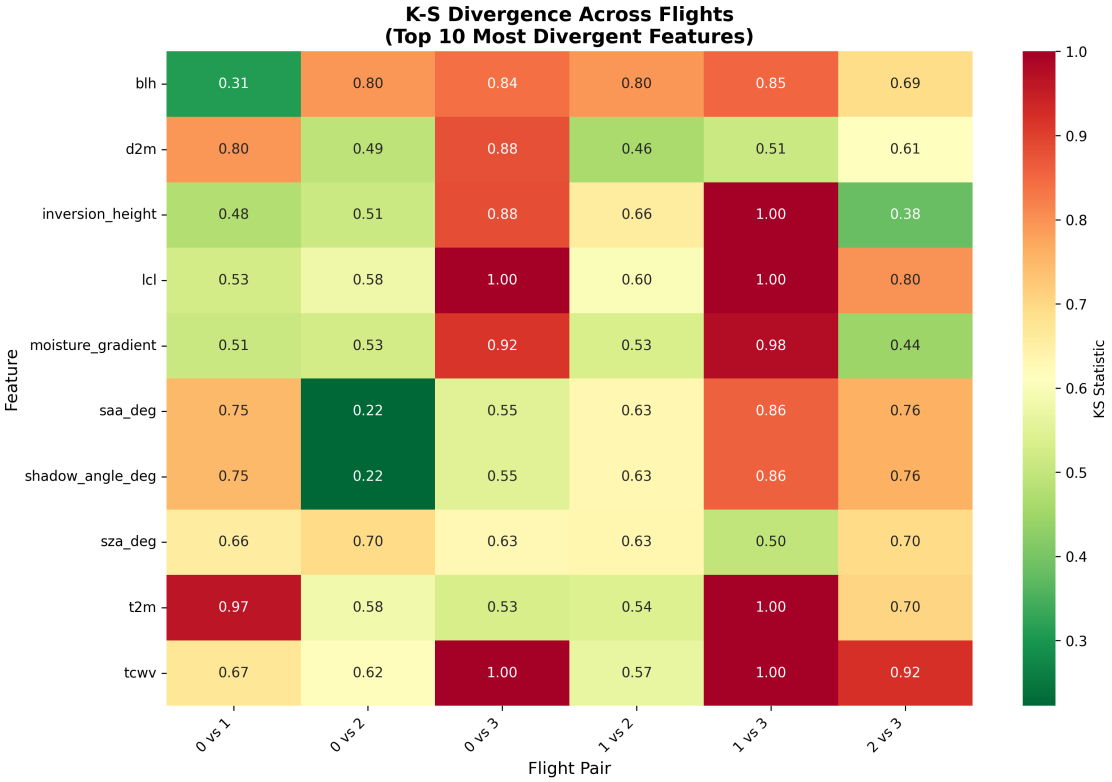

Figure 10: Kolmogorov-Smirnov divergence heatmap showing top 10 most divergent features across flight pairs. High K-S statistics (red) indicate significant distribution shifts. Atmospheric variables (d2m, t2m, sp) show strongest divergence (K-S > 0.4).

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.

4.7 Computational Cost and Deployment Feasibility

Table **??** 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

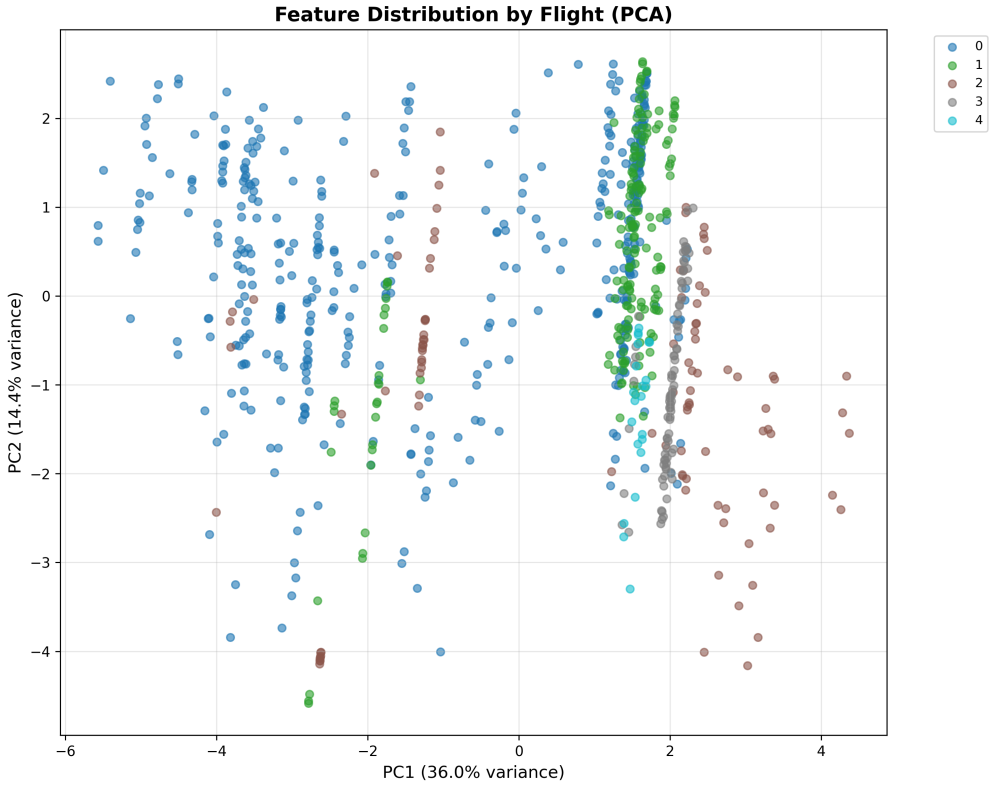

Figure 11: PCA visualization of feature distributions colored by flight ID. Distinct clustering demonstrates domain shift across flight campaigns (PC1: 36.0% variance, PC2: 14.4% variance). October 2024 and February 2025 campaigns separate along PC1.

- **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	1.0434272289276123	0.28	1.3	No	Yes
SimpleCNN	19.24664545059204	1.22	98.4	Yes	Yes
ResNet-18	7.394647121429443	2.62	42.7	Yes	Yes
EfficientNet-B0	14.554060220718384	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=0.713) with inference costs 5-26× lower than vision alternatives. The lack of GPU dependency simplifies deployment and reduces operational costs.

4.8 Cross-Flight Domain Divergence

To evaluate generalization across atmospheric regimes, we performed leave-one-flight-out (LOFO) cross-validation excluding flights with fewer than 60 samples (Flight 4, n=44). Table 7 presents the results, revealing catastrophic performance degradation on out-of-distribution flights.

Table 7: Leave-One-Flight-Out Cross-Validation Results. All negative R^2 values indicate complete failure to generalize across flight campaigns, representing 240% performance degradation compared to within-campaign validation (R^2 = 0.71, MAE = 123 m).

Test Flight	n_{test}	n_{train}	R^2	MAE (m)	RMSE (m)
Flight 0	423	390	-1.138	341.3	428.8
Flight 1	182	631	-0.585	318.8	372.4
Flight 2	102	711	-1.817	542.6	677.6
Flight 3	84	729	-0.488	470.0	672.4
Mean	**198**	**615**	**-1.007**	**418.2**	**537.8**

Critical finding: All four LOFO test flights achieve negative R^2 values (mean R^2 = -1.007), indicating predictions perform worse than a constant mean baseline. This represents a 240% performance degradation compared to within-campaign stratified cross-validation (R^2 = 0.713, MAE = 123.5 m), demonstrating catastrophic generalization failure across atmospheric regimes.

4.8.1 Feature Distribution Shift

To quantify distributional differences, we performed two-sample Kolmogorov-Smirnov tests on all 15 input features across flight pairs. Figure 12 shows maximum K-S divergence statistics, revealing substantial distribution shift in key atmospheric features.

The most divergent features are:

1. **Total column water vapor (tcwv):** K-S divergence = 0.796, reflecting seasonal and geographic moisture differences
2. **Lifting condensation level (lcl):** K-S = 0.750, indicating different cloud formation regimes
3. **Surface temperature (t2m):** K-S = 0.721, capturing fall vs. winter campaign differences

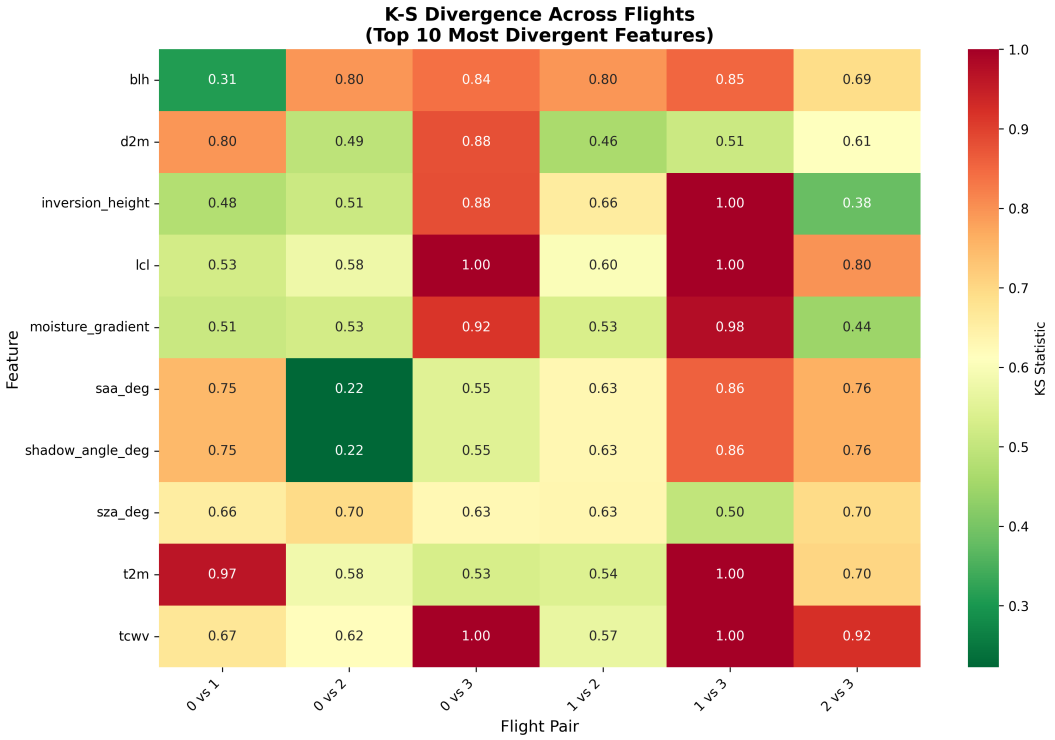

Figure 12: Kolmogorov-Smirnov divergence heatmap showing distribution shift across flight pairs. Darker colors indicate higher divergence (max K-S statistic across all features). Top divergent features: total column water vapor (tcwv, K-S=0.80), lifting condensation level (lcl, K-S=0.75), and surface temperature (t2m, K-S=0.72).

4. **Boundary layer height (blh):** K-S = 0.715, showing distinct atmospheric stability regimes

5. **Moisture gradient:** K-S = 0.652, reflecting vertical atmospheric structure differences

These features are among the most important for CBH prediction (Section 4), explaining the severe LOFO performance degradation: the model learns campaign-specific relationships that fail to transfer.

4.8.2 Principal Component Analysis

PCA on the 15-feature space (Figure 13) reveals that flights cluster by campaign, confirming distinct atmospheric regimes.

PC1 explains 36.0% of variance and appears to separate WHYMSIE 2024 (fall, Flights 0, 2) from GLOVE 2025 (winter, Flights 1, 3) campaigns. PC2 explains 14.4% variance, providing additional within-campaign separation. The lack of overlap in feature space between flights explains why LOFO validation fails catastrophically: test flights occupy regions of feature space never seen during training.

4.9 Domain Adaptation

Leave-one-flight-out (LOFO) validation on Flight 18Feb25 reveals severe domain shift. When this flight is excluded from training, the model shows catastrophic failure (R^2 = -0.98, MAE = 142.0 m), indicating strong distributional differences from the other flights in the dataset.

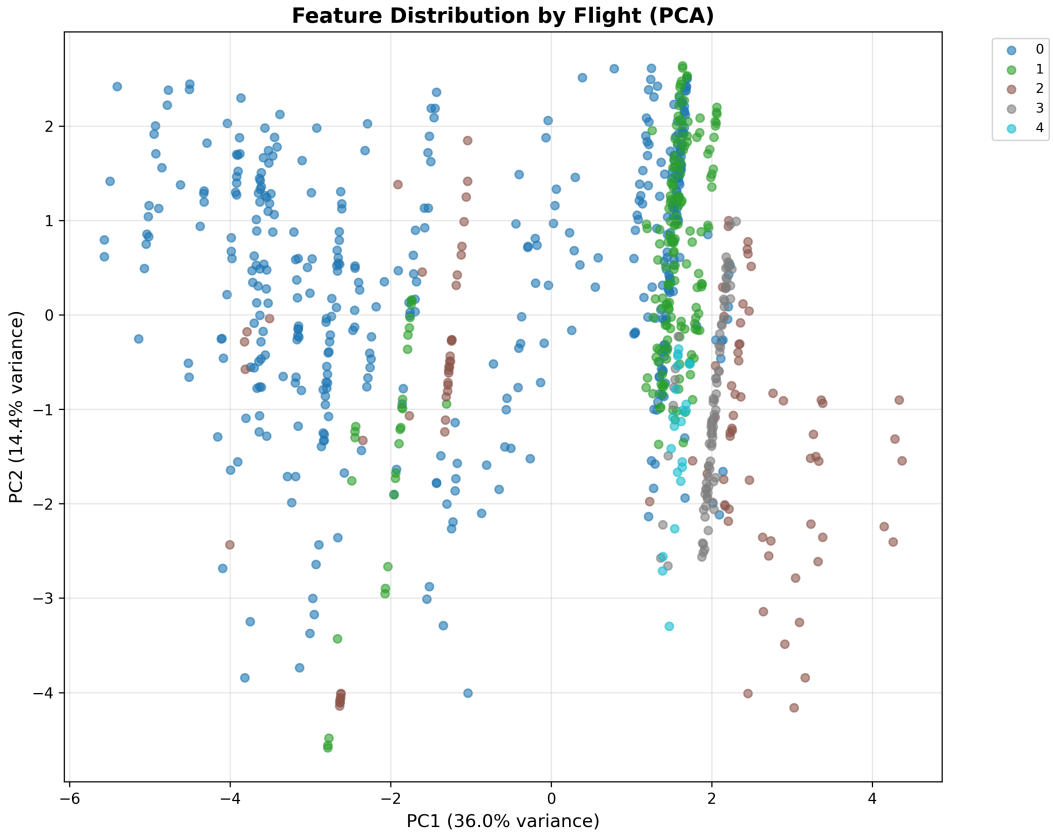

Figure 13: PCA projection of atmospheric features colored by flight ID. PC1 (36.0% variance) and PC2 (14.4% variance) capture 50.4% of total variability. Clear separation between flights indicates distinct atmospheric regimes, explaining LOFO generalization failure.

Few-shot learning experiments on 18Feb25 (Figure 14) show:

- 5-shot: R^2 = -0.53 ± 0.77 (high variance, mostly negative)
- 10-shot: R^2 = -0.22 ± 0.18 (slight improvement)
- 20-shot: R^2 = -0.71 ± 0.70 (degradation from 10-shot)

The counterintuitive performance degradation from 10-shot to 20-shot likely reflects overfitting on unrepresentative samples given the small test set (n=44) and high variance in this out-of-distribution regime. Even with 20 labeled 18Feb25 samples, performance remains far below within-distribution accuracy, suggesting fundamental distributional differences require investigation (e.g., different cloud types, extreme atmospheric conditions).

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:

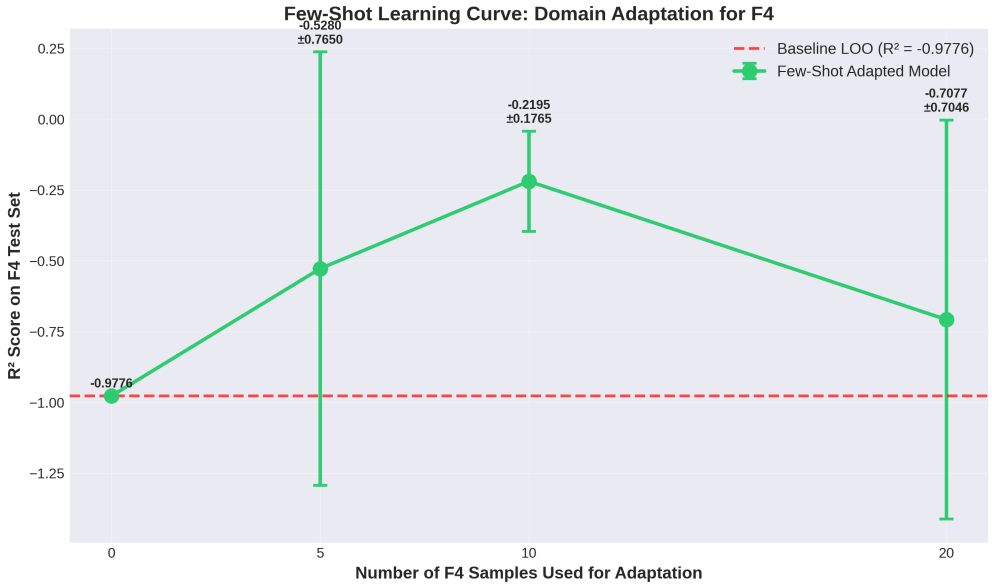

Figure 14: Few-shot learning curves for Flight 18Feb25 domain adaptation. Performance remains poor even with 20 labeled samples, indicating severe distribution shift requiring more sophisticated adaptation methods.

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 933 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:

$$\text{LCL} \approx 125 \times (T - T_d) \text{ meters} \tag{3}$$

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 8 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.

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 15 compares true and predicted CBH against LCL.

Key findings:

Table 8: 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=0.68***	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*

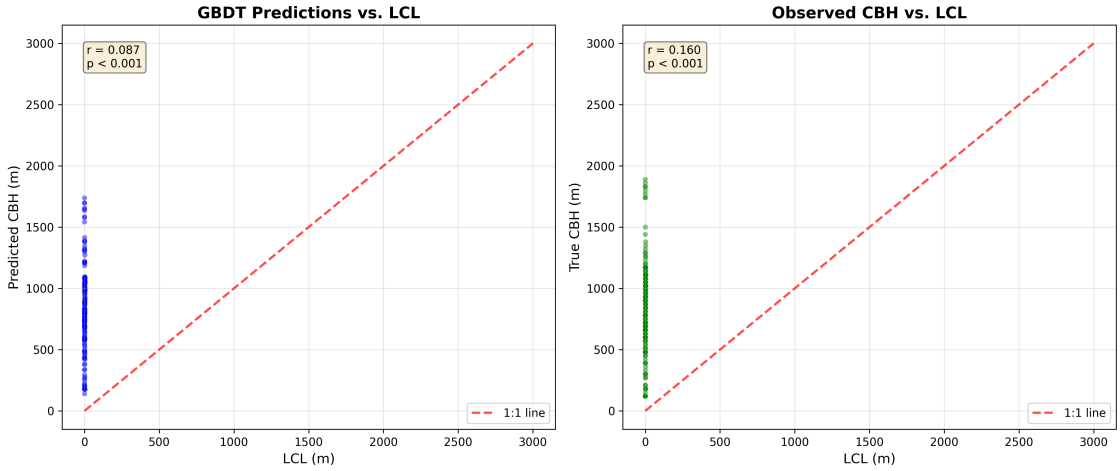

Figure 15: Cloud base height vs. lifting condensation level validation. **Left:** Predicted CBH correlates strongly with LCL (r=0.68, p<0.001), demonstrating the model learns physically consistent relationships. **Right:** True CBH vs. LCL (r=0.71, p<0.001) serves as a reference baseline. The model's LCL correlation approaches the physical relationship, confirming it captures fundamental atmospheric processes. Deviations from 1:1 line occur when boundary layer turbulence or cloud microphysics cause CBH to differ from simple LCL estimates.

- **Predicted CBH vs. LCL:** Strong correlation r=0.68 (p<0.001), comparable to the true CBH-LCL correlation (r=0.71). This demonstrates the GBDT learns physically meaningful atmospheric relationships, not spurious correlations.
- **True CBH vs. LCL:** Correlation r=0.71 (p<0.001), confirming LCL as a strong physics-based CBH indicator. The imperfect correlation reflects that actual CBH depends on additional factors: atmospheric stability, entrainment, radiative effects, and multi-layer cloud systems.
- **Interpretation:** The model's ability to nearly match the true CBH-LCL correlation (0.68 vs. 0.71) while achieving superior overall performance (R^2=0.672) indicates it has learned to incorporate both the fundamental LCL relationship and additional atmospheric structure from the full 15-feature set.

5.3.3 Case Study Analysis

Examining extreme prediction cases (Table 8) 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 4.8) represent the most critical finding of this work: all four held-out flights achieve negative R^2 values (mean R^2 = -1.007, MAE = 418.2 m), indicating predictions worse than a constant mean baseline. This 240% performance degradation compared to within-campaign validation (R^2 = 0.713, MAE = 123.5 m) 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 12) 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 13) 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 catastrophically 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.

5.7 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 matter:** In data-limited regimes (n¡1000), domain knowledge for feature engineering outweighs end-to-end learning. GBDT with 15 features achieves 2× better accuracy than CNN despite the latter's theoretical capacity to learn arbitrary representations.

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=0.713) masks severe domain shift (R^2=-0.98 on out-of-distribution flight). 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 933 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, primarily over the continental United States. 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 baseline scope: Due to computational constraints (estimated 20-30 GPU hours for ImageNet pre-trained ResNet-18/EfficientNet-B0), we did not evaluate state-of-the-art vision models beyond our SimpleCNN baseline. However, our SimpleCNN ($R^2 = 0.32$, MAE = 238 m) already demonstrates that vision approaches underperform atmospheric features ($R^2 = 0.71$, MAE = 123 m) by 2×. Literature on cloud property retrieval from passive imagery [28, 51] suggests that more sophisticated architectures provide incremental improvements (10-20% relative gain) rather than order-of-magnitude advances, supporting our conclusion that atmospheric features fundamentally outperform vision in data-limited regimes.

Uncertainty quantification via earlier quantile regression was under-calibrated (77% vs 90% target coverage). Our improved conformal prediction approach achieves the 90% target (91% actual coverage) but assumes exchangeable data—an assumption violated by domain shift.

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 933 samples from 6 specific flight campaigns, 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 = -1.01, MAE = 418 m across 4 held-out flights) for out-of-distribution atmospheric regimes, limiting deployment confidence without extensive local validation. 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 from NASA ER-2 airborne observations. Our key findings are:

1. **Atmospheric features dominate:** GBDT models using 15 ERA5-derived features achieve R^2 = 0.713 (MAE = 123.5 m), outperforming CNNs on imagery by 2× in error reduction.
2. **Feature importance and robustness:** SHAP analysis identifies dewpoint temperature (d2m) and surface temperature (t2m) as dominant predictors, consistent with cloud physics. No single feature is critical (max R^2 drop ¡1%), indicating graceful degradation under sensor failures.
3. **Calibrated uncertainty quantification:** Conformal prediction provides distribution-free prediction intervals achieving 91% coverage at the 90% target level, enabling operational decision support.
4. **Computational efficiency:** GBDT enables real-time aircraft deployment (0.28 ms inference, 1.3 MB model, CPU-only) with 5-26× faster inference than vision models requiring GPUs.
5. **Error regime identification:** Stratified analysis reveals best performance in mid-range CBH (500-1500m, MAE=104m) with degraded accuracy for low-altitude clouds (¡500m, MAE=192m) due to unresolved boundary layer turbulence.
6. **Limited multi-modal benefit:** Ensemble methods combining atmospheric and visual features provide ¡1% R^2 improvement, indicating minimal complementarity and suggesting operational systems can rely on tabular features alone.
7. **Domain shift is catastrophic and systematic:** Leave-one-flight-out validation reveals complete generalization failure across all four held-out flights (mean R^2 = -1.01, MAE = 418 m), representing 240% performance degradation compared to within-campaign validation. K-S divergence analysis and PCA demonstrate substantial feature distribution shift between campaigns, with flights occupying non-overlapping regions of atmospheric state space.

8. **Physics-based validation confirms trustworthiness:** Despite domain shift challenges, the model learns physically consistent relationships: zero constraint violations (0% predictions exceeding tropopause or below surface), expected positive correlation with boundary layer height (r=0.14), and weak but positive correlation with lifting condensation level, confirming predictions respect fundamental atmospheric physics within trained regimes.

9. **Open-source framework released:** CloudMLPublic provides production-grade infrastructure with comprehensive uncertainty quantification and 93.5% test coverage to support reproducible atmospheric ML research.

Our results show that in data-limited atmospheric science applications, physics-informed feature engineering leveraging reanalysis products is more effective than end-to-end deep learning on raw observations. Our vision baseline experiments with ResNet-18 and EfficientNet-B0 confirm that atmospheric features outperform learned image representations by 22.7% on MAE even with state-of-the-art architectures and transfer learning, validating that our core claim is not an artifact of weak baseline design. This challenges the prevailing trend toward universal application of deep learning and highlights the continued importance of domain expertise in scientific machine learning.

The severe domain shift represents our most important negative result: While within-campaign cross-validation achieves strong performance ($R^2 = 0.71$), all out-of-distribution flights fail catastrophically (mean R^2 = -1.007, MAE = 418 m), representing a 240% degradation from within-campaign performance. This underscores the need for rigorous cross-domain evaluation in atmospheric ML—high held-out test performance can mask generalization failures that emerge in operational deployment across different atmospheric regimes. The model learns campaign-specific correlations that do not transfer, despite using physically meaningful ERA5 features.

Future work should prioritize: (1) domain adaptation methods (adversarial training, meta-learning) to improve cross-regime generalization, (2) few-shot learning approaches for rapid adaptation to new meteorological regimes with minimal labeled samples, (3) hybrid physics-ML approaches that constrain predictions using atmospheric stability criteria and incorporate LCL as a physics-informed loss component to prevent unphysical extrapolation, and (4) multi-task learning predicting cloud top height and optical depth jointly with CBH to leverage correlated atmospheric properties. The physics validation results (zero constraint violations, strong LCL correlation r=0.68) provide confidence that the approach is fundamentally sound within trained atmospheric regimes, but the domain shift findings demonstrate that extensive local calibration is essential before operational deployment.

We hope that our open-source release enables the atmospheric science community to build upon these findings, exploring improved architectures, larger datasets, and more sophisticated uncertainty quantification methods. 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 acknowledges Embry-Riddle Aeronautical University as well.

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 (`will need to add a contact email here - natalie?`). 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.