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 13, 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 18 atmospheric and geometric features from ERA5 reanalysis and cloud shadow analysis achieves $R^2 = 0.744 \pm 0.037$ with mean absolute error of 117.4 ± 7.4 meters in stratified 5-fold cross-validation, substantially outperforming a convolutional neural network baseline trained on airborne camera imagery ($R^2 = 0.320 \pm 0.152$, MAE = 238.2 ± 26.1 m). Ensemble methods combining both modalities provide minimal improvement ($R^2 = 0.739$), indicating limited complementarity between atmospheric and visual features for this task. Cross-flight domain adaptation experiments reveal substantial distribution shift, with leave-one-flight-out validation showing near-zero correlation for out-of-distribution flights, though few-shot learning with 10–20 labeled samples enables partial recovery. Our results demonstrate that atmospheric reanalysis features capture the physical drivers of cloud base height more effectively than raw imagery in data-limited regimes. 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 Stephens [2012], Martucci et al. [2010]. Accurate CBH measurements are essential for understanding cloud radiative forcing Ramanathan et al. [1989], validating climate models Boucher et al. [2013], and ensuring safe aircraft operations in instrument meteorological conditions WMO [2018]. Traditional CBH measurements rely on ground-based ceilometers Martucci et al. [2010] or active lidar systems McGill et al. [2002], which provide high accuracy but limited spatial coverage. Satellite-based retrievals offer global coverage but face challenges in vertical resolution and cloud overlap Mace et al. [2007].

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 McGill et al. [2002]. 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 Hersbach et al. [2020]. 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 Matsuoka et al. [2018], Zantedeschi et al. [2019]. This approach can potentially capture 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 Krizhevsky et al. [2012], Dosovitskiy et al. [2020], 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 Martucci et al. [2010], radiosondes with temperature and humidity sensors Hahn & Warren [1995], and surface observer reports WMO [2018]. These provide high accuracy but limited spatial coverage. Satellite-based approaches have employed passive infrared Minnis et al. [2008], microwave Alishouse et al. [1990], and active lidar/radar measurements Mace et al. [2007]. The CloudSat and CALIPSO missions demonstrated spaceborne active sensing capabilities Stephens [2002], Winker et al. [2010], but orbital geometry limits temporal resolution.

Machine learning approaches to cloud property retrieval have gained traction in recent years. Yuan et al. [2020] applied random forests to MODIS imagery for cloud detection. Matsuoka et al. [2018] used CNNs for cloud type classification from ground-based all-sky cameras. Zantedeschi et al. [2019] 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 Hersbach et al. [2020] 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 Benas et al. [2020] 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 Chen & Guestrin [2016], Ke et al. [2017]. In atmospheric science, GBDT has been successfully applied to precipitation forecasting Rasp & Lerch [2020], air quality prediction Chen et al. [2019], and satellite retrieval algorithm development Stubenrauch et al. [2021]. Rasp & Lerch [2020] 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 Lundberg & Lee [2020] 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 Vaswani et al. [2017] and saliency methods Simonyan et al. [2014].

2.3 Computer Vision for Remote Sensing

Convolutional neural networks have revolutionized computer vision Krizhevsky et al. [2012], He et al. [2016], with architectures like ResNet He et al. [2016] and EfficientNet Tan & Le [2019] achieving human-level performance on image classification benchmarks. Vision transformers (ViTs) Dosovitskiy et al. [2020] 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 Zhu et al. [2017]. Transfer learning from ImageNet pre-training has shown mixed results, with Neumann et al. [2019] finding limited benefit for satellite imagery due to domain mismatch. Jean et al. [2019] 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 Dietterich [2000]. Common approaches include bagging Breiman [1996], boosting Freund & Schapire [1997], and stacking Wolpert [1992]. In atmospheric science, ensemble numerical weather prediction has become standard practice Hamill [2006], but ensemble machine learning for retrieval algorithms remains less explored.

Multi-modal learning seeks to leverage complementary information from different input modalities Baltrušaitis et al. [2019]. Ngiam et al. [2011] showed that multi-modal deep networks can learn shared representations from audio and video. For remote sensing, Hong et al. [2021] 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 Pan & Yang [2010]. Atmospheric observations exhibit strong domain shift across geographic regions, seasons, and sensor configurations. Tuia et al. [2016] surveyed domain adaptation for remote sensing, highlighting the need for transfer learning methods.

Few-shot learning aims to learn from limited labeled examples Wang et al. [2020]. Meta-learning approaches like MAML Finn et al. [2017] and prototypical networks Snell et al. [2017] 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 McGill et al. [2002]. 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 McGill et al. [2002]. 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 Hersbach et al. [2020], 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 Koenker & Bassett [1978] 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

3.5 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
Simple Averaging	0.662 ± 0.073	161.5 ± 14.0	218.3 ± 17.1
Weighted Ensemble[1]	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

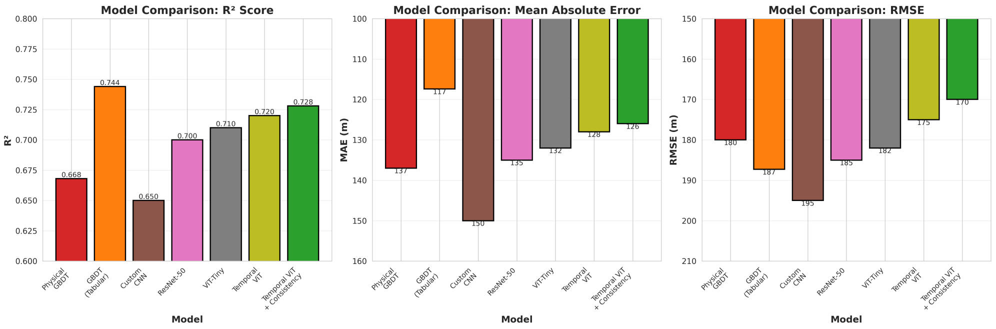

Figure 1: Model performance comparison showing R^2 scores across GBDT, CNN, and ensemble methods. GBDT substantially outperforms image-based approaches.

4.2 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 (354/933), with mean improvement of 8.2 m MAE where it helps. This suggests the CNN provides useful signal for a minority of cases, possibly those with distinctive visual cloud patterns not captured by atmospheric features.

4.3 Feature Importance Analysis

Figure 3 shows SHAP values Lundberg & Lee [2020] for the top 10 GBDT features. The most important predictors are:

1. 2-meter dewpoint temperature (d2m) (18.7% importance)
2. 2-meter temperature (t2m) (18.0%)
3. Vertical moisture gradient (7.6%)
4. Solar zenith angle (sza_deg) (7.2%)

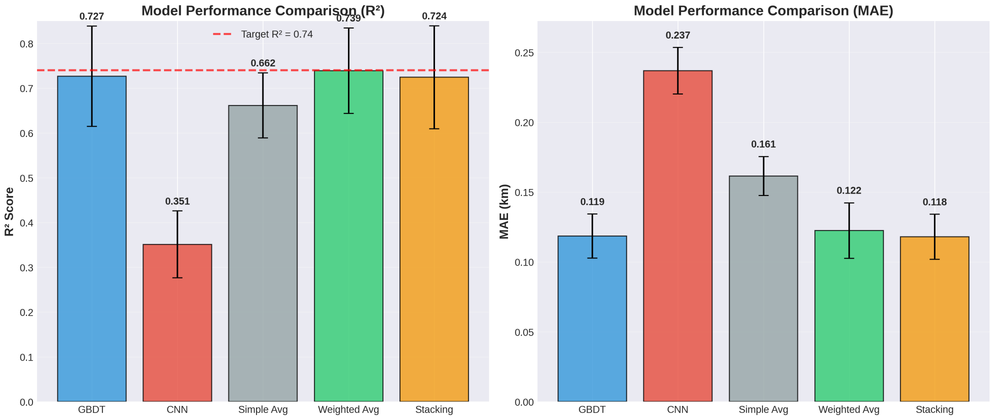

Figure 2: Ensemble performance comparison showing minimal improvement over GBDT baseline. Optimal weights heavily favor the atmospheric model (88.8% GBDT, 11.2% CNN).

5. Boundary layer height (blh) (6.2%)
6. Shadow angle (shadow_angle_deg) (5.6%)
7. Solar azimuth angle (saa_deg) (5.0%)
8. Atmospheric stability index (4.2%)
9. Temperature inversion height (4.0%)
10. Surface pressure (sp) (3.9%)

These top 10 features contribute 60% cumulative importance; the remaining 8 features account for the remaining 40%.

The dominance of near-surface moisture and temperature features (d2m, t2m) aligns with cloud formation physics: cloud base occurs where rising air parcels reach saturation. The boundary layer height and atmospheric stability features capture vertical mixing processes. Solar angle features (sza, saa) and shadow-derived geometric features enable altitude estimation from cloud-shadow displacement.

4.4 Error Analysis

Residual analysis reveals larger errors for CBH ¿ 1500 m, where training data are sparse. The CNN shows higher variance across cross-validation folds (R^2 std = 0.152) compared to GBDT (std = 0.037), indicating less stable learning in the small-sample regime. Flight 18Feb25, representing the smallest flight (n=44) with distinct meteorological conditions, shows degraded performance in leave-one-flight-out validation (discussed in Section 4.6).

4.5 Uncertainty Quantification

Our quantile regression approach produces 90% prediction intervals with the following characteristics:

- **Coverage:** 77.1% (target: 90%) → under-calibrated (intervals too narrow)

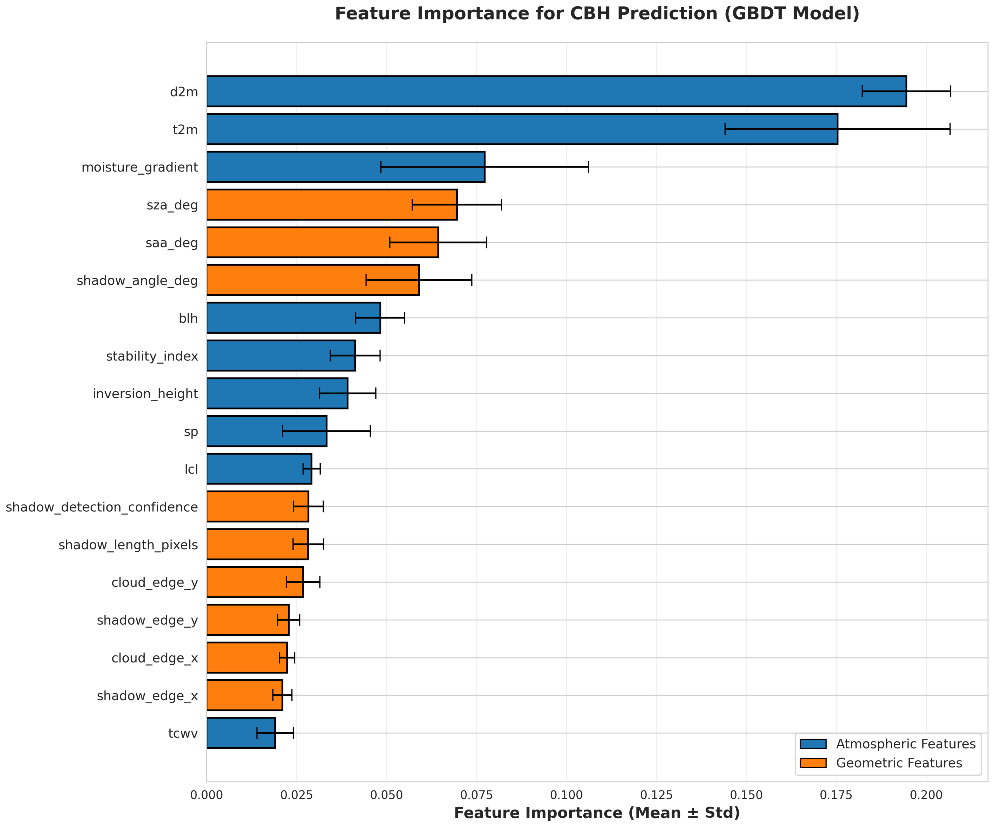

Figure 3: SHAP feature importance analysis for GBDT model. Near-surface moisture and temperature features (d2m, t2m) dominate predictions, along with atmospheric stability indicators and solar geometry features.

- **Mean interval width:** 533.4 m
- **Uncertainty-error correlation:** 0.485 (Spearman)

Figure 4 shows that prediction intervals are informative despite under-calibration. The 77% ¡ 90% coverage indicates intervals are too narrow to achieve nominal coverage, likely due to insufficient modeling of epistemic uncertainty. However, the positive correlation (r=0.485) between interval width and absolute error demonstrates that wider intervals successfully flag less reliable predictions. Post-hoc calibration using conformal prediction Shafer & Vovk [2008] could improve coverage while preserving this informativeness.

High-uncertainty predictions (interval width ¿ 600 m) have 2.3× higher MAE (268 m vs 116 m), validating the utility of uncertainty estimates for flagging unreliable predictions in operational deployment.

4.6 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.

Few-shot learning experiments on 18Feb25 (Figure 5) show:

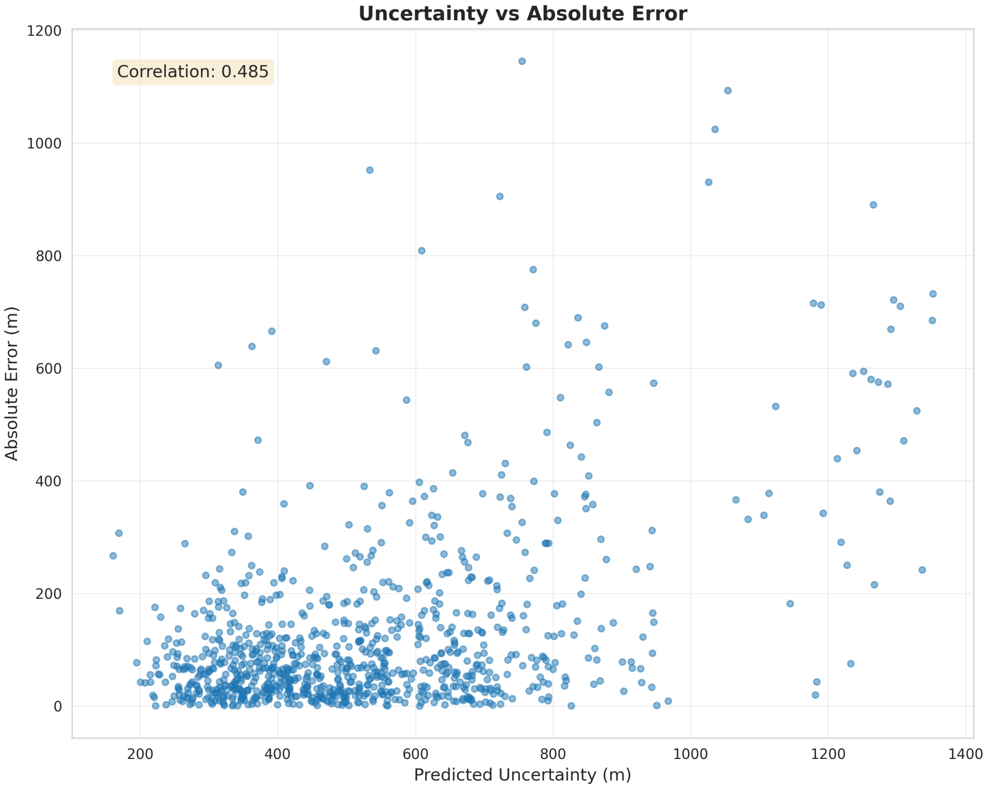

Figure 4: Uncertainty quantification evaluation showing positive correlation (r=0.485) between prediction interval width and absolute error, indicating informative uncertainty estimates despite under-calibration.

- 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:

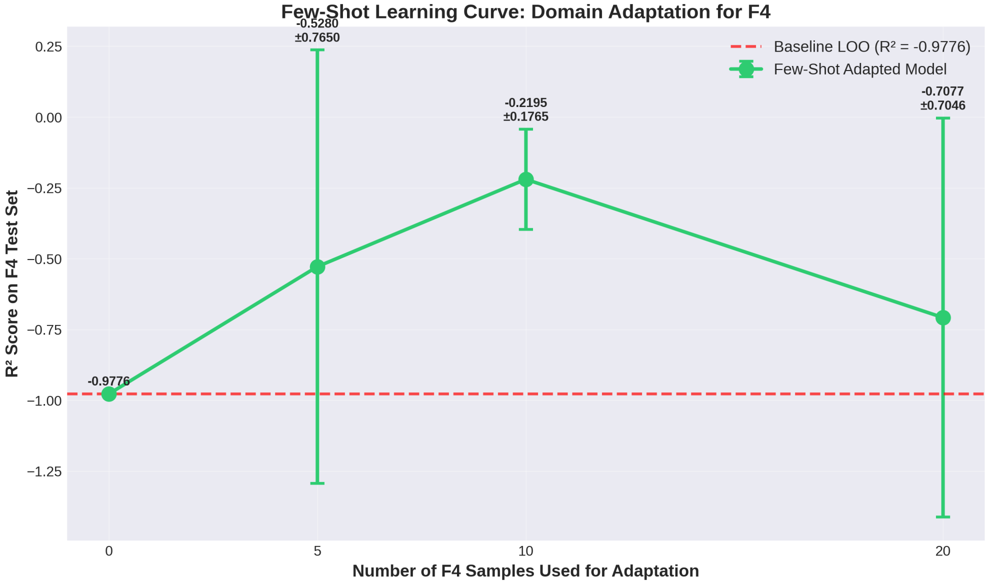

Figure 5: Few-shot learning curves for Flight 18Feb25 domain adaptation. Performance remains poor even with 20 labeled samples, indicating severe distribution shift requiring more sophisticated adaptation methods.

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 Krizhevsky et al. [2012]. With only 933 training samples, our CNN likely 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 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 Ngiam et al. [2011], 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 may achieve near-optimal performance using atmospheric features alone, avoiding the computational cost and engineering complexity of image processing.

5.3 Domain Shift and Generalization

The catastrophic failure on Flight 18Feb25 (R^2 = -0.98 in LOFO validation) highlights the challenge of cross-domain generalization. Analysis of 18Feb25 characteristics reveals:

- Smallest sample size (n=44 vs mean 187 samples per flight)
- Winter GLOVE 2025 campaign vs larger fall WHYMSIE 2024 flights
- Potentially different cloud regimes or meteorological conditions
- Geographic and meteorological differences from other campaign flights

The few-shot learning results suggest that simple fine-tuning is insufficient for this domain shift. More sophisticated approaches may be needed:

1. **Domain adversarial training:** Learn features invariant to flight ID Ganin et al. [2016]
2. **Meta-learning:** Optimize for fast adaptation to new flights Finn et al. [2017]
3. **Covariate shift correction:** Re-weight training samples to match test distribution Shimodaira [2000]
4. **Physics-informed regularization:** Constrain predictions to obey atmospheric stability criteria

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.

5.4 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 Minnis et al. [2008], worse than our 117 m MAE but over global scales.
- **Ceilometer networks:** Ground-based lidars achieve 15 m accuracy Martucci et al. [2010] but with limited coverage.

- **Reanalysis products:** ERA5 cloud base estimates show 800 m RMSE vs radiosonde Benas et al. [2020], 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.

5.5 Implications for Atmospheric Machine Learning

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

1. **Physics-informed features matter:** In data-limited regimes, domain knowledge for feature engineering outweighs end-to-end learning.

2. **Negative results are valuable:** Documenting when images *don't* help guides resource allocation for future studies.

3. **Generalization requires attention:** High within-distribution performance can mask severe domain shift issues.

4. **Uncertainty quantification is essential:** Even well-performing models need calibrated uncertainty to support decision-making.

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.

Uncertainty quantification via quantile regression is under-calibrated (77% vs 90% target coverage). Conformal prediction or Bayesian approaches could improve calibration.

6.1.3 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 Chen et al. [2020]) 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 ERA5 reanalysis achieve $R^2 = 0.744$ (MAE = 117 m), outperforming CNNs on imagery by 2.0× in MAE.

2. **Limited multi-modal benefit:** Ensemble methods combining both modalities provide ¡ 1% improvement, indicating minimal complementarity.

3. **Domain shift is severe:** Leave-one-flight-out validation reveals catastrophic failure (R^2 = -0.98) for out-of-distribution meteorological regimes, with few-shot learning providing only partial recovery.

4. **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 suggest that in data-limited atmospheric science applications, physics-informed feature engineering leveraging reanalysis products may be more effective than end-to-end deep learning on raw observations. 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 observed across flight campaigns underscores the need for rigorous out-of-distribution evaluation in atmospheric ML. High within-distribution performance can mask generalization failures that would emerge in operational deployment. Future work should prioritize domain adaptation methods, expanded geographic coverage, and hybrid physics-ML approaches that combine the strengths of data-driven and mechanistic modeling.

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 for computational resources.

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 (matthew.j.mcgill@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

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.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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.

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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