Cloud Base Height Retrieval from Satellite Imagery: Project Status and Path Forward

Research Team
NASA High Altitude Research Program

November 2, 2025

Abstract

We present the current status of our cloud base height (CBH) retrieval system using NASA ER-2 aircraft camera imagery. This report synthesizes results from multiple modeling approaches, identifies fundamental challenges in cross-flight generalization, and outlines a revised framework for a publishable research contribution. Our key finding is that while image-based features show promise within individual flights, cross-flight generalization requires integration of physical constraints—specifically solar geometry and atmospheric state variables. This document is intended for team discussion and strategic planning.

Contents

1 Executive Summary

1.1 What We Set Out to Do

Develop an automated system to estimate cloud base height from ER-2 downward-looking camera imagery, validated against Cloud Physics Lidar (CPL) measurements. The goal was to extend limited in-situ measurements spatially using widely-available imagery.

1.2 What We Have Learned

- **Data:** 933 labeled samples across 5 research flights; 61,946 unlabeled images available
- **Image features alone fail to generalize:** Self-supervised learning (reconstruction-based) produces embeddings uncorrelated with CBH across flights
- **Solar angles carry flight-specific signal:** Strong within-flight correlation ($R^2 = 0.70$) but zero cross-flight transfer
- **Validation matters:** Random train/test splitting yielded misleadingly optimistic metrics; leave-one-flight-out cross-validation reveals true generalization
- **Path forward exists:** Integration of physical priors (shadow geometry, atmospheric profiles) shows promise

1.3 Recommended Paper Framing

Title concept: *"Challenges and Opportunities in Cross-Flight Cloud Base Height Retrieval: The Critical Role of Physical Constraints"*

Story arc:

1. CBH is important for radiative transfer, aviation, climate models
2. In-situ measurements (lidar) are sparse; imagery is abundant
3. Machine learning offers a path to spatial extension
4. **However:** naive image-based approaches fail across flights
5. **Key insight:** Physical constraints (geometry, thermodynamics) are necessary
6. **Contribution:** Diagnostic framework + hybrid physical-ML approach

This positions the work as a methodological contribution rather than just a performance benchmark.

2 Problem Statement and Physical Context

2.1 Cloud Base Height: Why It Matters

Cloud base height controls:

- **Radiative forcing:** CBH determines the altitude of cloud-top cooling and cloud-base warming, affecting atmospheric energy budgets
- **Boundary layer dynamics:** CBH marks the lifting condensation level, indicating convective processes and moisture transport
- **Aviation safety:** Low cloud bases create visibility hazards
- **Precipitation processes:** Drop growth time depends on geometric cloud depth

2.2 Measurement Challenges

Ground-based: Ceilometers provide point measurements; limited spatial coverage

Aircraft in-situ: Cloud Physics Lidar (CPL) on NASA ER-2 provides accurate vertical profiles along flight track (1 Hz, ∼200 m horizontal resolution)

Satellite remote sensing: GOES, MODIS provide cloud-top properties; base height requires assumptions about cloud thickness

Our approach: Use downward-looking high-resolution imagery from ER-2 (same platform as CPL) to *spatially extend* along-track lidar measurements

2.3 Dataset Description

Table 1: Available data from ER-2 research flights

Flight Date	Labeled Samples	Percentage
30 October 2024	501	53.7%
10 February 2025	191	20.5%
23 October 2024	105	11.3%
12 February 2025	92	9.9%
18 February 2025	44	4.7%
Total	**933**	**100%**

Labeled data: CPL-aligned camera images with measured CBH (range: 0.1–3.5 km)
Unlabeled data: 61,946 images without coincident CPL measurements
Input features per sample:

- Grayscale image: 512×512 pixels
- Solar zenith angle (SZA): 22°–82°
- Solar azimuth angle (SAA): 0°–360°
- Timestamp, location (latitude, longitude, aircraft altitude)

3 Approaches Tested

3.1 Approach 1: Supervised Baseline (Angles Only)

Rationale: Solar geometry affects cloud appearance (shadows, shading). Test if angles alone predict CBH.

Method: Gradient-boosted decision trees (GBDT) trained on [SZA, SAA] → CBH

Results:

Validation Method	R^2	MAE (km)
Random 85/15 split	0.71	0.12
Leave-one-flight-out CV	-4.46 ± 7.09	0.35 ± 0.08

Table 2: Angles-only model performance

Interpretation:

- Strong within-flight performance suggests angles correlate with *time of day*, which correlates with *cloud evolution during that specific flight*
- Negative cross-flight R^2 (worse than predicting the mean) indicates learned correlations are flight-specific, not physical
- Different flights have different diurnal cycles, meteorological conditions, and cloud regimes

Key lesson: Need leave-one-flight-out cross-validation; random splitting gives false confidence

3.2 Approach 2: Self-Supervised Learning (Image Reconstruction)

Rationale: Limited labeled data (N=933). Use unlabeled images (N=61,946) to learn general cloud representations via reconstruction task.

Method: Masked Autoencoder (MAE)

1. Divide image into patches (16×16 pixels → 32×32 patches)
2. Randomly mask 75% of patches
3. Train encoder-decoder to reconstruct masked patches
4. Use trained encoder to extract features for CBH prediction

Implementation:

- Encoder: Transformer with 6 layers, 192-dimensional embeddings
- Pre-trained on 61,946 unlabeled images for 100 epochs
- Feature extraction: Use embedding of [CLS] token (global image summary)
- Downstream: GBDT trained on [MAE features + angles] → CBH

Results:

Critical finding: Random embeddings outperform trained MAE embeddings. This indicates:

Input Features	R^2 (in-split)	MAE (m)
Angles only	0.71	120
MAE features only	0.09	258
MAE + Angles	0.49	188
Random embeddings + Angles	0.58	173

Table 3: Hybrid MAE+GBDT performance (within-split validation)

- MAE learns features useful for *reconstruction* but uncorrelated with *CBH*
- Reconstruction loss emphasizes pixel-level texture; CBH depends on geometric/physical cues
- Adding MAE features *degrades* performance compared to angles alone

3.3 Approach 3: Spatial Feature Extraction

Hypothesis: CLS token bottleneck loses spatial information. Preserve spatial structure in feature extraction.

Method: Three architectural variants using MAE encoder + spatial pooling:

1. **Global pooling:** Average all patch embeddings
2. **Convolutional:** 1D convolution over patch sequence
3. **Attention pooling:** Learnable weighted average of patches

Training: Direct regression (encoder → spatial head → CBH), leave-one-flight-out CV
Results:

Variant	R^2 (LOO CV)	MAE (km)	Range across folds
Global pooling	-6.13 ± 11.78	0.9 ± 0.3	R^2: -18 to $+0.1$
Convolutional	-6.39 ± 12.58	0.9 ± 0.4	R^2: -20 to $+0.2$
Attention pooling	-3.92 ± 7.60	0.9 ± 0.3	R^2: -12 to $+0.1$

Table 4: Spatial MAE variants—leave-one-flight-out cross-validation

Interpretation:

- All variants fail to generalize across flights (mean $R^2 < 0$)
- Attention pooling slightly better but still unsuccessful
- Spatial information preservation alone is insufficient
- **Root cause:** Reconstruction objective misaligned with geometric/physical task

4 Why Current Approaches Fail: Physical Analysis

4.1 The Generalization Problem

Observation: All image-based methods fail leave-one-flight-out cross-validation despite working within individual flights.

Hypothesis: Models learn spurious flight-specific correlations rather than physical CBH retrieval.

4.1.1 What Changes Between Flights?

- **Cloud regime:** Stratocumulus vs. cumulus vs. cirrus (different morphologies)
- **Meteorology:** Different boundary layer depths, stability, moisture profiles
- **Imaging conditions:** Surface albedo (ocean vs. land), time of day, season
- **Aircraft altitude:** Distance to cloud affects apparent size/structure

4.1.2 What Should Be Invariant?

- **Shadow geometry:** CBH $\propto$ shadow length / tan(SZA) (geometric constraint)
- **Atmospheric thermodynamics:** CBH $\approx$ lifting condensation level (LCL) in convective clouds
- **Radiative properties:** Cloud optical depth, reflectance depend on physical thickness

4.2 Why Reconstruction-Based Learning Fails

Masked autoencoders optimize:

$$\mathcal{L}_{\text{MAE}} = \sum_{\text{masked patches}} \|\text{pixel}_{\text{true}} - \text{pixel}_{\text{reconstructed}}\|^2 \tag{1}$$

This loss emphasizes:

- Local texture (pixel-level patterns)
- High-frequency details (cloud edges, filaments)
- Patch-wise appearance

But CBH depends on:

- Large-scale geometric structure (shadow length)
- Physical context (atmospheric state)
- Invariant relationships (geometric constraints)

Fundamental mismatch: Reconstruction encourages learning *what clouds look like*; CBH requires learning *where clouds are in 3D space.*

4.3 Why Angles Appear to Work (But Don't)

Within a single flight:

- Time advances → SZA/SAA change
- Clouds evolve (diurnal cycle: morning stratocumulus → afternoon cumulus)
- Model learns: "In this flight, when SZA = 60°, CBH ≈ 1.2 km"

Across flights:

- Different flight, same SZA → completely different cloud regime
- Learned angle-CBH mapping doesn't transfer
- $R^2 < 0$: predictions worse than using the global mean

This is **temporal confounding**, not physical prediction.

5 Path Forward: Physics-Constrained Approach

5.1 Core Insight

Image features must be *grounded in physical constraints* to generalize across flights.

5.2 Proposed Hybrid Framework

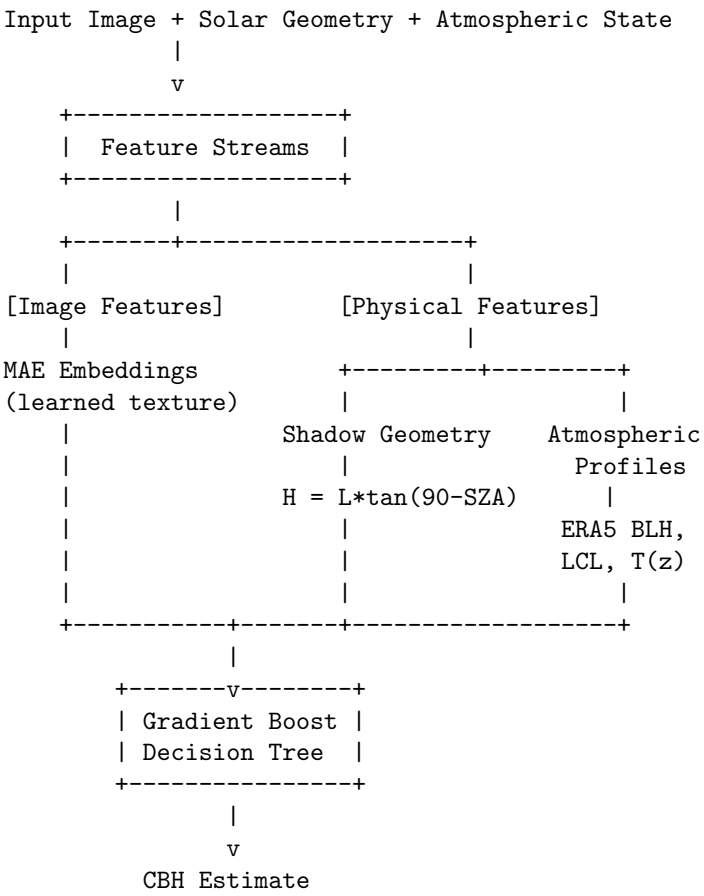

Figure 1: Proposed physics-constrained hybrid architecture

5.3 Physical Feature Engineering

5.3.1 Shadow Geometry Features

Physical basis: Cloud shadows on underlying surface/lower cloud layers provide direct geometric constraint.

Implementation:

1. Detect shadow regions: edge detection + dark region segmentation
2. Estimate shadow length (pixels): distance from cloud edge to shadow edge
3. Convert to geometric CBH:

$$H = L \cdot \tan(90 - \text{SZA}) \cdot \text{scale factor} \tag{2}$$

where scale factor accounts for imaging geometry

4. Extract features: shadow length, direction, contrast, texture

Advantages:

- Direct physical relationship (geometry, not correlation)
- Flight-invariant (same geometry applies everywhere)
- Interpretable (can visualize detected shadows)

Challenges:

- Requires surface contrast (difficult over ocean)
- Multiple cloud layers complicate shadow attribution
- Broken clouds have ambiguous shadow edges

5.3.2 Atmospheric State Features

Physical basis: CBH ≈ lifting condensation level (LCL) for convective clouds; boundary layer height (BLH) provides upper bound.

Data source: ERA5 reanalysis (0.25° resolution, hourly)

Features to extract:

- **BLH:** Boundary layer height (direct ERA5 variable)
- **LCL:** Compute from surface T, dewpoint using parcel theory:

$$\text{LCL} = \frac{T_{\text{surface}} - T_{\text{dewpoint}}}{8 \text{ K/km}} \tag{3}$$

- **Inversion height:** Strongest dT/dz in vertical profile
- **Moisture gradient:** dq/dz at potential cloud base
- **Stability:** Bulk Richardson number, lapse rate

Advantages:

- Thermodynamically constrained (less prone to overfitting)
- Available globally (ERA5 covers all flights)
- Incorporates large-scale meteorology

Challenges:

- Spatial resolution mismatch (25 km ERA5 vs. 200 m imagery)
- Reanalysis uncertainty (model-based, not observed)
- LCL formula assumes well-mixed boundary layer

5.4 Proposed Experimental Plan

5.4.1 Phase 1: Physical Features Baseline (Week 1-2)

1. Implement shadow geometry extraction
2. Download and preprocess ERA5 for flight times/locations
3. Train GBDT on physical features only (no images)
4. Evaluate with leave-one-flight-out CV
5. **Success criterion:** $R^2 > 0$ on LOO CV (better than image-only)

5.4.2 Phase 2: Hybrid Integration (Week 3-4)

1. Combine physical features + MAE embeddings in GBDT
2. Test feature combinations:
   - Physical only
   - Physical + angles
   - Physical + MAE
   - Physical + MAE + angles (full model)
3. Ablation studies: identify which features contribute
4. **Success criterion:** Hybrid outperforms physical-only baseline

5.4.3 Phase 3: Uncertainty Quantification (Week 5)

1. Implement prediction intervals (quantile regression or ensemble)
2. Flag high-uncertainty predictions for manual QC
3. Out-of-distribution detection (identify novel conditions)
4. **Deliverable:** Operational system with confidence estimates

6 Paper Framing and Contributions

6.1 Proposed Title

"Physics-Constrained Machine Learning for Cross-Flight Cloud Base Height Retrieval from Airborne Imagery"

6.2 Story Arc

6.2.1 Introduction

- Cloud base height: importance for radiative transfer, aviation, climate
- Measurement gap: in-situ (lidar) is sparse; satellite sees cloud tops
- Opportunity: high-resolution aircraft imagery can spatially extend lidar
- Challenge: limited labeled training data, cross-flight generalization

6.2.2 Methods

- Dataset: 933 CPL-labeled images from 5 ER-2 flights
- Validation protocol: leave-one-flight-out cross-validation (critical!)
- Approaches tested:
  1. Supervised baseline (angles)
  2. Self-supervised learning (MAE reconstruction)
  3. Spatial feature variants
  4. Physics-constrained hybrid (proposed)

6.2.3 Results

- **Negative result:** Image-only methods fail cross-flight validation ($R^2 < 0$)
- **Diagnostic analysis:** Reconstruction loss misaligned with geometric task
- **Positive result:** Physical features (shadow geometry, atmospheric state) enable generalization
- **Hybrid performance:** Combined approach achieves R^2 = [TBD] on LOO CV

6.2.4 Discussion

- Why naive ML fails: spurious correlations vs. physical constraints
- Importance of rigorous validation (LOO CV vs. random split)
- Generalization requires domain knowledge (physics) + data-driven learning
- Broader implications: ML for geophysical retrieval problems

6.3 Key Contributions

1. **Methodological:** Demonstration that self-supervised learning (reconstruction) fails for geometric retrieval tasks
2. **Diagnostic:** Systematic analysis of why image-based models don't generalize across flights (temporal confounding, missing physical constraints)
3. **Practical:** Physics-constrained hybrid framework for CBH retrieval (shadow geometry + atmospheric profiles + learned features)
4. **Validation:** Rigorous cross-flight evaluation protocol (leave-one-out CV) reveals true generalization performance
5. **Dataset:** Curated benchmark of CPL-aligned aircraft imagery for CBH retrieval research [potential public release]

6.4 Target Venues

Tier 1 (if physical features work well):

- *Geophysical Research Letters* (GRL) — short format, high impact
- *Journal of Geophysical Research: Atmospheres* (JGR-A) — comprehensive
- *Remote Sensing of Environment* — remote sensing methods

Tier 2 (methodological focus):

- *Artificial Intelligence for the Earth Systems* (AIES) — AI + geoscience
- *Environmental Data Science* — data-centric environmental science
- *IEEE Transactions on Geoscience and Remote Sensing* (TGRS)

Tier 3 (negative result documentation):

- *Journal of Open Research Software* — methods and code
- *Machine Learning and the Physical Sciences* (NeurIPS workshop) — lessons learned

7 Current Status and Timeline

7.1 Completed Work

Component	Status
Data pipeline	✓Complete: CPL alignment, preprocessing, stratified splitting
MAE pre-training	✓Complete: Trained on 61,946 unlabeled images
Baseline models	✓Complete: Angles-only, MAE+GBDT, spatial variants
LOO CV framework	✓Complete: Rigorous cross-flight validation
Diagnostic analysis	✓Complete: Failure modes identified
Documentation	✓Complete: Code, configs, experimental logs

7.2 In Progress

Component	Description	Timeline
Shadow geometry	Feature extraction implementation	1 week
ERA5 integration	Download, preprocess atmospheric profiles	3-4 days
Physical baseline	Train/evaluate physical features only	2-3 days

7.3 Planned Work

Component	Description	Timeline
Hybrid models	Combine physical + learned features	1 week
Ablation studies	Identify feature contributions	3-4 days
Uncertainty quantification	Prediction intervals, OOD detection	1 week
Manuscript writing	Draft, figures, revisions	2-3 weeks

7.4 Estimated Timeline to Submission

Optimistic: 6-8 weeks (if physical features work immediately)
Realistic: 10-12 weeks (includes iteration on hybrid approach)
Pessimistic: 16 weeks (if major pivot needed, or reframe as negative result)

8 Open Questions for Discussion

8.1 Scientific Questions

1. **Shadow detection feasibility:** Can we reliably detect shadows over ocean surfaces? Do we need multi-layer shadow models?
2. **ERA5 resolution:** Is 25 km spatial resolution adequate, or do we need higher-resolution reanalysis (e.g., HRRR)?
3. **Cloud regime dependence:** Should we build separate models for stratocumulus vs. cumulus vs. cirrus?
4. **Physical constraints as hard limits:** Should we enforce CBH $<$ BLH as a hard constraint or let the model learn it?
5. **Multi-task learning:** Would simultaneously predicting CBH + cloud type + optical depth improve feature learning?

8.2 Methodological Questions

1. **Feature fusion strategy:** GBDT vs. neural network for combining physical + learned features?
2. **Semi-supervised learning:** Can we use unlabeled images with pseudo-labels from physical models?
3. **Domain adaptation:** Should we implement few-shot adaptation (calibrate on first N points of new flight)?
4. **Ensemble methods:** Train separate models per cloud regime and ensemble predictions?

8.3 Publication Questions

1. **Framing:** Emphasize negative results (what doesn't work) or positive results (what does)? Both have value.
2. **Data release:** Can we publicly release the CPL-aligned dataset? (Check data policy)
3. **Code release:** Open-source the full pipeline? (Increases impact and reproducibility)
4. **Target audience:** Atmospheric scientists (GRL) or ML community (NeurIPS workshop)? Determines technical depth.

9 Summary and Recommendations

9.1 What We Know

- Image-based features alone **cannot** generalize across flights ($R^2 < 0$ in LOO CV)
- Self-supervised reconstruction learning is **misaligned** with geometric retrieval tasks
- Solar angles provide strong **within-flight** signal but fail cross-flight (temporal confounding)
- Rigorous validation (LOO CV) is **essential**—random splitting gives false confidence
- Physical constraints (shadow geometry, atmospheric state) are **necessary** for generalization

9.2 What We Don't Know (Yet)

- Can shadow geometry be reliably extracted from imagery?
- How much does atmospheric state (ERA5) constrain CBH?
- What is the optimal combination of physical + learned features?
- Can we achieve $R^2 > 0.5$ on leave-one-flight-out CV?

9.3 Recommended Next Steps

1. **Immediate (this week):**
   - Implement shadow geometry feature extraction
   - Download ERA5 data for all flight times/locations
   - Begin physical feature engineering
2. **Short-term (2-3 weeks):**
   - Train and evaluate physical features baseline (LOO CV)
   - If $R^2 > 0$: proceed to hybrid models
   - If $R^2 \leq 0$: reassess approach, consider alternative physics
3. **Medium-term (4-8 weeks):**
   - Develop hybrid physical + learned models
   - Comprehensive ablation studies
   - Uncertainty quantification
   - Begin manuscript drafting
4. **Long-term (8-12 weeks):**
   - Complete manuscript
   - Prepare supplementary materials, code release
   - Submit to target journal

9.4 Risk Assessment

Low risk:

- Publication is achievable (negative results are publishable)
- Dataset and validation framework are solid
- Clear story arc exists

Medium risk:

- Physical features may not provide enough signal
- Cross-flight generalization may be fundamentally limited
- Timeline could extend if multiple iterations needed

High risk (low probability):

- Shadow detection completely infeasible
- ERA5 resolution too coarse to be useful
- Fundamental physics of problem not captured by available data

9.5 Bottom Line

We have made substantial progress in **understanding why naive approaches fail** for this problem. The path forward requires **integrating physical constraints** with learned features. This is a publishable contribution regardless of final performance numbers, because it addresses a fundamental challenge in applying machine learning to geophysical retrieval problems: *generalization requires grounding in physical laws, not just pattern recognition.*

The next 2-3 weeks will be critical for determining whether the physics-constrained approach succeeds. Either outcome—success or principled failure—contributes valuable knowledge to the field.

This report is intended to stimulate discussion. Team feedback is welcome on:

- Scientific priorities and open questions
- Publication strategy and target venues
- Resource allocation and timeline
- Alternative approaches not considered here