

Performance Analysis of CatBoost, XGBoost and Random Forest Models for Cloud Top Height Prediction Using Satellite Data

Performance Analysis of CatBoost, XGBoost and Random Forest Models for Cloud Top Height Prediction Using Satellite Data

Ade Sekarningsih^{a,1}, Choirul Basir^{b,2}, Sudarno^{a,3}

^a Master of Informatics Engineering Program, Universitas Pamulang, South Tangerang, 15417, Indonesia

^b Departement of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Pamulang, South Tangerang, 15417, Indonesia

^c Directorate of Aviation Meteorology, Indonesian Agency for Meteorology Climatology and Geophysics (BMKG), Jakarta, Indonesia

ade.sekarningsih@bmkgo.id¹, dosen02278@unpam.ac.id², dosen00119@unpam.ac.id³

ABSTRACT

Cloud Top Height (CTH) is a critical parameter for convective weather monitoring and aviation meteorology, particularly over the Indonesian maritime continent, where deep convection occurs frequently. This study evaluates the predictive performance and behavior of three tree-based ensemble models XGBoost, CatBoost, and Random Forest—for estimating CTH from Himawari multispectral observations using the Sentinel-5P/TROPOMI CTH product as the reference. The input variables comprise Himawari brightness-temperature channels, cloud fraction, latitude, and longitude. Evaluation on an independent spatial validation set of 3,444,543 observations demonstrates the model's success, with XGBoost achieving the best aggregate performance. It recorded a mean absolute error of 971.71 m, a root mean square error of 1297.57 m, and a coefficient of determination of 0.646. This performance successfully outperformed CatBoost and Random Forest, demonstrating an RMSE reduction of 3.2% and 5.3%, respectively. These quantitative indicators confirm XGBoost as a highly reliable baseline for CTH prediction. Multi-method interpretability analysis using built-in importance, permutation importance, group permutation, SHAP, partial dependence, and individual conditional expectation indicates that spectral channels and cloud fraction are the principal information sources, whereas coordinates provide weaker but non-zero spatial context. The error audit reveals relatively stable predictions for mid-level clouds but strong underprediction for extremely high cloud tops. Inter-model disagreement is positively but weakly related to actual error, indicating that it is a useful diagnostic signal rather than a calibrated uncertainty estimate. These results demonstrate that model selection for operational meteorology should consider not only aggregate accuracy but also error regimes, spatial dependence, and physical consistency.

Keywords : Cloud Top Height, Himawari, Sentinel-5P/TROPOMI, XGBoost; CatBoost, Random Forest

Submitted:

Article Info :
Reviewed:

Accepted :

Copyright © 2026 – CSRID Journal. All rights reserved.

1. INTRODUCTION

Cloud Top Height (CTH) describes the altitude of the uppermost radiatively emitting cloud layer and is a key variable for representing cloud vertical development. It contributes to the analysis of cloud-radiation interactions, convective intensity, precipitation-producing systems, and the assimilation of satellite observations into numerical weather prediction [1], [2]. In aviation meteorology, CTH is particularly relevant because high and rapidly growing cloud tops may accompany deep convection, severe turbulence, icing, lightning, and other weather hazards that affect flight-level selection and route

planning. An operational CTH product must therefore be spatially extensive, frequently updated, and sufficiently reliable across both common and extreme cloud conditions.

Direct vertical cloud profiling is available from active instruments such as spaceborne radar and lidar. These sensors provide highly informative vertical measurements but are generally carried on polar-orbiting platforms with narrow swaths and limited temporal sampling over a fixed location [2], [5]. Passive imagers on geostationary satellites offer a complementary capability. The Advanced Himawari Imager (AHI) provides frequent multispectral observations over the Asia-Pacific region, enabling continuous monitoring of cloud development over Indonesia. However, CTH retrieval from passive radiances is indirect. Thermal brightness temperature can approximate the temperature of an optically thick cloud top, but the conversion from temperature to altitude is affected by atmospheric temperature profiles, cloud emissivity, multilayer cloud structures, thin cirrus, mixed pixels, and sensor viewing geometry [1], [4], [5].

The use of multiple infrared and water-vapor channels helps reduce this ambiguity because different channels respond to different atmospheric layers and cloud properties. Nevertheless, physically based inversion methods commonly require radiative-transfer assumptions and auxiliary atmospheric profiles. Machine-learning methods provide an alternative by learning nonlinear mappings between multispectral observations and reference cloud products. Previous studies have shown that XGBoost, neural networks, and other data-driven approaches can improve satellite-based cloud-height retrieval under complex conditions [3], [11], [14], [18]. Tree-based ensembles are especially attractive for large tabular satellite datasets because they can represent nonlinear feature interactions, tolerate heterogeneous predictor scales, and provide efficient inference [7], [8], [21].

Despite these advantages, a model with a low global RMSE is not automatically suitable for scientific or operational use. Aggregate metrics do not reveal whether the model depends on physically meaningful spectral information, whether it exploits geographical coordinates as a spatial shortcut, or whether its errors are concentrated in safety-relevant regimes. They also do not indicate whether response curves follow plausible atmospheric behavior. This limitation is important over the Indonesian maritime continent, where land-ocean contrasts, strong moisture variability, and frequent deep convection produce a heterogeneous prediction domain [13]. A model may perform well for the densely represented middle of the CTH distribution while substantially underestimating the rare upper tail.

A review of the state-of-the-art literature reveals that most earlier CTH retrieval studies primarily reported accuracy improvements or comparisons with reference sensors [3], [10], [14], [19]. However, a critical research gap remains: less attention has been given to combining performance evaluation with a multi-method behavioral audit. Most studies treat the predictor merely as a numerical regression function without identifying its physical failure modes. Therefore, the novelty and scientific contribution of this study lie in establishing a comprehensive evaluation framework that not only compares the aggregate accuracy of XGBoost, CatBoost, and Random Forest under the same Himawari-TROPOMI configuration, but also systematically audits their physical consistency, spatial-coordinate dependence, and regime-specific errors.

This study addresses that need by comparing XGBoost, CatBoost, and Random Forest under the same Himawari-TROPOMI data configuration. The contributions are fourfold. First, the study provides a controlled comparison of aggregate predictive performance on a large spatial validation dataset. Second, it combines built-in importance, permutation importance, group permutation, and SHAP to assess whether the models rely primarily on spectral, geophysical, or spatial information. Third, partial dependence and individual conditional expectation are used to examine the physical regularity of learned responses. Fourth, the residuals are stratified by CTH and cloud fraction, and inter-model disagreement is evaluated as a diagnostic signal of difficult predictions.

2. METHOD

A. Study Area and Data Sources

The study domain covers the Indonesian region, a tropical maritime area characterized by strong diurnal convection and substantial spatial heterogeneity. The analysis uses observations collected during 2024 from two complementary satellite systems. Himawari AHI data provide the predictor variables, while the Sentinel-5P/TROPOMI Level-3 CTH product provides the supervised reference target. TROPOMI is treated as a reference retrieval rather than an error-free ground truth because satellite cloud products retain retrieval uncertainty, particularly for optically thin, multilayer, and very high clouds [10].

Table 1. Performance Metric Comparison

Sources	Product / Variable	Role in the model	Main information
Himawari AHI	Brightness temperature, channels 08-16	Predictors	Water-vapor and thermal cloud signals
Himawari / cloud product	Cloud fraction	Predictor	Cloud coverage and scene homogeneity
Geolocation grid	Latitude and longitude	Predictors	Regional spatial context
Sentinel-5P/TROPOMI	Level-3 Cloud Top Height	Reference target	Retrieved CTH in meters

B. Preprocessing and Spatio-Temporal Collocation

The two satellite systems observe the atmosphere using different orbits, spatial footprints, and acquisition schedules. A spatio-temporal collocation procedure was therefore required to establish paired predictor-target observations. For each TROPOMI overpass, the nearest available Himawari scan was selected to reduce temporal mismatch. Spatial harmonization then associated the corresponding Himawari spectral observations with each TROPOMI CTH retrieval. The resulting dataset represents simultaneous or near-simultaneous descriptions of the same atmospheric scenes.

Quality control was performed before model training. Observations with missing values, invalid coordinates, or physically unrealistic CTH and brightness-temperature values were removed. The units were standardized so that brightness temperature was represented consistently and CTH was expressed in meters. Quality information from the reference product was used to retain suitable retrievals. Because satellite data contain strong spatial autocorrelation, model evaluation was conducted on an independent spatial validation set rather than interpreting neighboring pixels as fully independent random observations. The final validation set contained 3,444,543 observations and was used identically for all three algorithms.

C. Feature Engineering and Model Development

The feature matrix included the original brightness-temperature channels, cloud_fraction, latitude, and longitude. The selected channels were retained because water-vapor bands provide information about moisture and upper-tropospheric emission, while thermal-window bands are sensitive to cloud-top temperature. The use of multiple channels enables the models to learn combinations that reduce the ambiguity of any single thermal observation. All models received the same predictors and target observations, ensuring that differences in performance were attributable to the learning algorithms rather than unequal data inputs.

XGBoost constructs an additive sequence of decision trees, with each new tree trained to reduce the residual errors of the current ensemble. Its objective function includes regularization that controls tree complexity and supports efficient learning on large tabular datasets [7]. CatBoost also uses gradient boosting but employs ordered boosting to reduce prediction shift during sequential learning [21]. Although the present features are numerical, CatBoost is included to evaluate whether its boosting procedure produces a competitive and stable CTH mapping. Random Forest serves as a bagging baseline; it builds independent trees from bootstrap samples and random feature subsets, and averages their predictions to reduce variance [8].

The computational workflow was implemented in Python using array- and table-oriented libraries for satellite processing and machine learning. Multidimensional satellite files were read and transformed using tools designed for NetCDF or HDF data, while the regression models were developed with the XGBoost, CatBoost, and scikit-learn ecosystems. The software workflow separated data preparation, model fitting, prediction, and interpretation into distinct stages so that the same collocated feature matrix could be reused across algorithms. This separation reduced the risk that model-specific preprocessing would create an unfair comparison.

The large number of collocated observations required memory-conscious processing. Rather than repeatedly loading the complete archive for every experiment, the prepared numerical arrays were stored in a consistent feature order and processed in reusable batches. Missing-value filtering and quality-control masks were applied before the modeling stage so that all algorithms received identical observations. Intermediate outputs, including predictions and residuals, were retained for the subsequent SHAP, permutation, PDP, ICE, and stratified-error analyses.

Comparability was treated as a central experimental requirement. The three algorithms were trained for the same regression target and evaluated against exactly the same validation records. Performance metrics were calculated from the unrounded predictions, and the reported values were rounded only for presentation. The interpretability analyses were conducted on the fitted models and validation distribution, rather than on a separate synthetic dataset. Consequently, differences in feature reliance reflect differences in model behavior under a shared data environment.

For PDP and ICE visualization, feature values are displayed in standardized units. The horizontal axes in those plots should therefore be interpreted as relative positions within the observed feature distribution rather than as raw Kelvin values or geographic degrees. This distinction is important when translating response curves into meteorological meaning. The statistical standardization changes the axis scale but does not change the ordering of observations or the qualitative direction of the learned response.

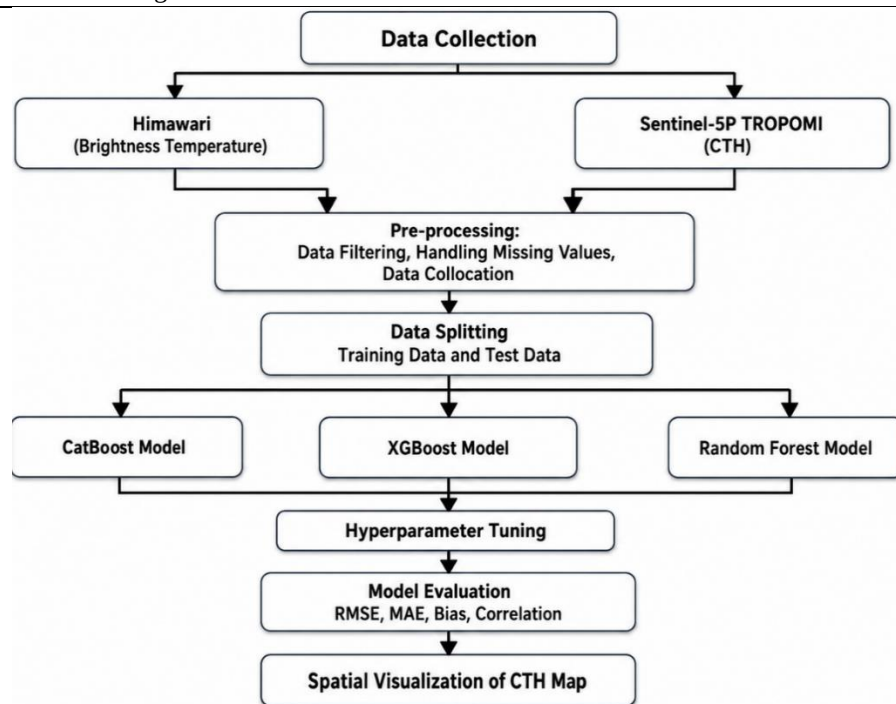


Figure 1. Flowchart Research Design

The workflow in Figure 1 begins with satellite data collection, quality filtering, and collocation. The integrated observations are then separated into model-development and evaluation subsets. XGBoost, CatBoost, and Random Forest are trained in parallel, and the resulting predictions are assessed using the same validation cases. The final stage extends beyond performance ranking by examining feature reliance, response behavior, residual structure, and model disagreement.

D. Feature Engineering and Model Development

Predictive performance was summarized using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), coefficient of determination (R^2), and mean bias. MAE is the average absolute difference between the predicted and reference CTH and therefore provides an intuitive estimate of typical vertical error. RMSE assigns a larger penalty to extreme residuals and is particularly relevant when large errors in deep-convective cloud tops have operational consequences. R^2 measures the fraction of target variance represented by the model, while mean bias indicates systematic overprediction or underprediction. Bias was defined as the average of predicted CTH minus reference CTH; consequently, negative values indicate underprediction.

Interpretability was assessed through complementary methods because no single importance measure provides a complete explanation. Built-in tree importance describes how strongly a feature participates in tree construction. Permutation importance quantifies the increase in validation RMSE after a feature is randomly shuffled. Group permutation applies the same procedure to physically related sets of predictors, including all spectral channels, water-vapor bands, thermal-window bands, cloud fraction, and coordinates. SHAP values estimate feature contributions to individual predictions and are summarized through mean absolute attribution [22].

Partial Dependence Plots (PDP) were used to describe average prediction responses, while Individual Conditional Expectation (ICE) curves were used to inspect heterogeneity among individual observations. A monotonicity audit calculated the fraction of adjacent PDP segments that violated a simple expected

negative relation between selected thermal brightness temperatures and CTH. This audit is diagnostic rather than a definitive physical test because channel responses can be modified by cloud emissivity, atmospheric profiles, multilayer structures, and correlations among channels.

Residual analysis was performed across CTH altitude intervals and cloud-fraction intervals. Finally, inter-model disagreement was calculated using the standard deviation and range of the three predictions for each observation. The relationship between disagreement and the absolute error of the ensemble-mean prediction was evaluated to determine whether divergence among models could identify difficult atmospheric scenes.

Table 2. Behavioral-audit components and interpretation objectives

Audit Componen	Primary Question	Output
Aggregate metrics	Which model is most accurate overall?	MAE, RMSE, R^2 , mean bias
Feature reliance	Which information sources support the prediction?	Built-in, permutation, group permutation, SHAP
Response consistency	How does prediction change across feature values?	PDP, ICE, monotonicity violations
Regime-specific error	Where are the largest residuals concentrated?	CTH-bin and cloud-fraction-bin errors

3. RESULT AND DISCUSSION

A. Statistical Performance Model

All three models were evaluated on the same 3,444,543 spatial validation observations. Table 3 shows that XGBoost produced the strongest aggregate performance, with MAE of 971.71 m, RMSE of 1297.57 m, R^2 of 0.646, and mean bias of -33.55 m. CatBoost ranked second, while Random Forest produced the largest overall errors. The consistent ranking across MAE, RMSE, and R^2 indicates that the advantage of XGBoost was not limited to a single evaluation measure.

Table 3. Perfomance Metric Comparison

Model Architecture	MAE (meters)	RMSE (meters)	Mean Bias (meters)	R^2 Score
XGBoost	971.71	1297.57	-33.55	0.646
CatBoost	1008.61	1340.95	-25.07	0.622
Random Forest	1030.44	1371.11	-38.72	0.605

The RMSE difference between XGBoost and CatBoost was 43.38 m, and the difference between XGBoost and Random Forest was 73.54 m. These margins are modest relative to the overall error magnitude, but they are consistent across millions of validation observations. The result suggests that sequential boosting more effectively captured the nonlinear interactions among the satellite predictors than bootstrap averaging. Nevertheless, the relatively small performance gap also indicates that model

selection cannot be justified by the leaderboard alone. The algorithms may achieve similar global scores while relying on different feature combinations or failing under different cloud regimes.

The mean biases of all models were small compared with their RMSE values. The negative signs indicate a slight overall tendency to underestimate CTH, but the magnitudes of 25-39 m are insufficient to explain RMSE values above 1.2 km. Therefore, the aggregate error is not primarily a uniform vertical offset. Instead, large positive and negative residuals in particular regimes largely cancel in the global mean. This result motivates the subsequent stratified analysis[14].

B. Feature Reliance

The feature-reliance analysis consistently identified spectral observations and cloud fraction as the principal sources of predictive information. In XGBoost, permuting `him_tbb_16` increased RMSE by 1123.61 m, followed by `him_tbb_10` with an increase of 819.84 m, `cloud_fraction` with 713.95 m, `him_tbb_11` with 560.28 m, and `him_tbb_09` with 297.08 m. CatBoost showed its largest deterioration after permutation of `him_tbb_16` (1146.12 m) and `cloud_fraction` (957.47 m). Random Forest relied most strongly on `cloud_fraction` (782.58 m), followed by `him_tbb_16` (483.54 m), `him_tbb_10` (425.60 m), `him_tbb_09` (334.14 m), and `him_tbb_08` (269.07 m).

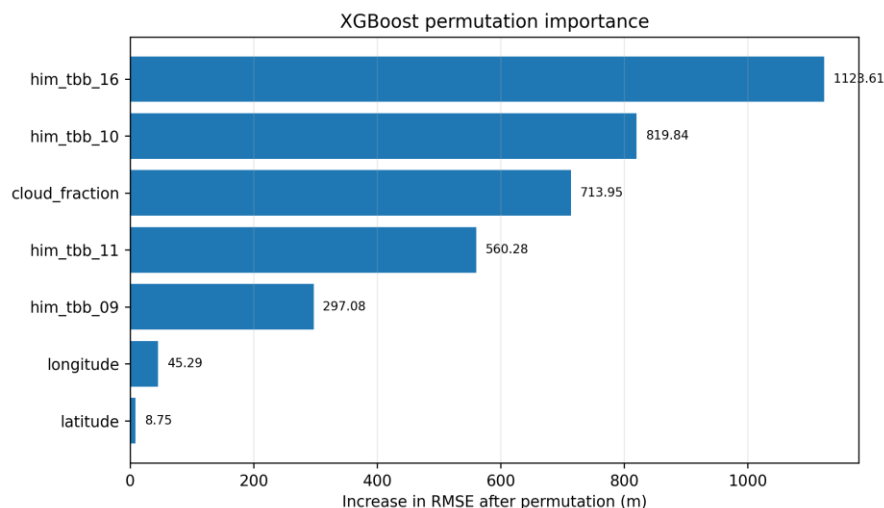


Figure 2. XGBoost permutation importance expressed as the increase in validation RMSE

Figure 2 emphasizes the substantial performance degradation caused by disrupting the leading XGBoost predictors. SHAP analysis provided a complementary local perspective. For XGBoost, `cloud_fraction` had the highest mean absolute SHAP value (approximately 657.61), followed closely by `him_tbb_16` (647.91), `him_tbb_10` (539.05), and `him_tbb_09` (478.58). Longitude ranked fifth with a value of 219.36, while latitude contributed 128.36. Thus, spatial coordinates influenced some individual predictions, but the dominant local contributions arose from cloud and spectral variables.

The importance rankings were broadly consistent but not identical. Spearman rank correlation was approximately 0.802 between SHAP and permutation importance, 0.753 between SHAP and built-in importance, and 0.758 between permutation and built-in importance. The agreement confirms that the highest-ranked predictors are robust to the interpretation method. The remaining disagreement is scientifically informative. For example, `him_tbb_09` ranked first in built-in importance but lower in permutation and SHAP rankings, suggesting that it was frequently used in tree splits yet was more replaceable by correlated spectral information. Conversely, `cloud_fraction` ranked highest in SHAP, indicating strong local contributions even when its structural tree ranking was not always first.

Table 4. Increase in Validation RMSE After Permutation of Feature Groups

Model Architecture	MAE (meters)	RMSE (meters)	Mean Bias (meters)
All spectral channels	1921.61	1603.08	1469.39
Thermal-window bands	1227.14	1018.41	526.33
Water-vapor bands	1120.84	597.79	1183.68
Cloud fraction	713.95	957.47	782.58

Group permutation provides the clearest evidence that the models were not driven primarily by location. Shuffling all spectral predictors increased XGBoost RMSE by 1921.61 m, CatBoost RMSE by 1603.08 m, and Random Forest RMSE by 1469.39 m. In contrast, jointly permuting latitude and longitude increased RMSE by only 48-55 m. The relative importance of water-vapor and thermal-window groups differed among algorithms. XGBoost and CatBoost were more sensitive to thermal-window bands, whereas Random Forest showed greater dependence on the water-vapor group. This difference indicates that the models organize the same atmospheric information through distinct tree structures.

C. Spatial Coordinate Dependence

The small group-permutation effect of the coordinates reduces, but does not eliminate, concern about spatial shortcuts. When assessed individually, permutation of latitude increased RMSE by only 8.75 m for XGBoost, 6.61 m for CatBoost, and 0.80 m for Random Forest. Longitude had a larger effect: 45.29 m for XGBoost, 41.39 m for CatBoost, and 53.67 m for Random Forest. This pattern indicates that the east-west spatial structure of the domain carried more predictive information than the north-south structure.

The apparently different conclusions from permutation importance and SHAP are not contradictory. Permutation asks how much global predictive performance deteriorates when a feature is destroyed; SHAP asks how much a feature changes individual predictions under the observed data distribution. Longitude can therefore contribute hundreds of meters to selected predictions even if other spectral variables compensate for its removal at the global level. In the Indonesian domain, longitude may represent legitimate regional differences in land-ocean distribution, climatological moisture, and cloud regimes. However, it may also serve as a proxy for patterns specific to the training domain.

The appropriate conclusion is therefore not that coordinates are either harmless or problematic in all circumstances. The present evidence shows that spectral and cloud information dominate, while coordinates provide secondary context. Before transfer to a new region or a substantially different season, the coordinate contribution should be tested through models trained without coordinates, leave-region-out validation, and external evaluation. A model that maintains accuracy without explicit coordinates may offer stronger geographical transferability, whereas a coordinate-aware model may remain advantageous for a fixed operational domain. Figure 3

D. Analysis Response Curve

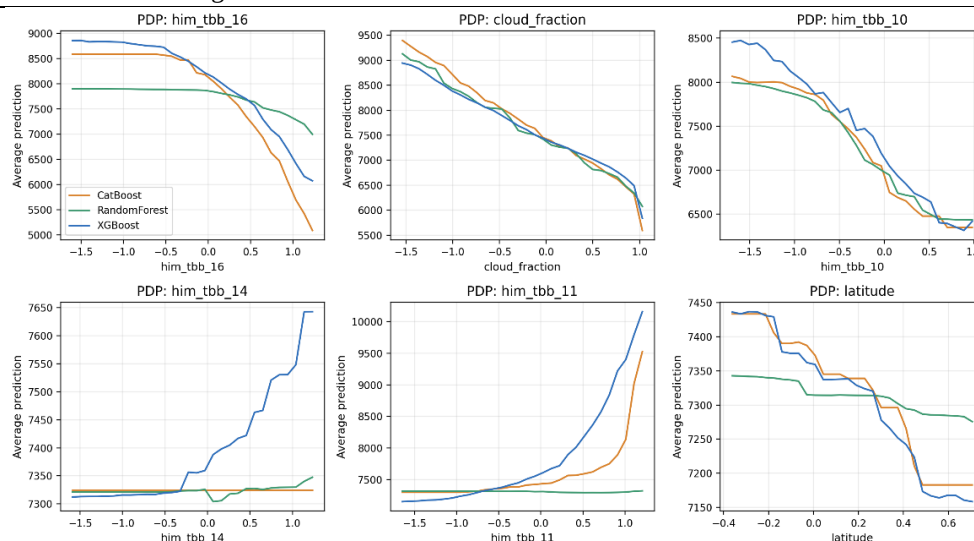


Figure 3. Partial dependence responses for selected spectral, geophysical, and spatial predictors

Figure 3 shows that the average response to him_tbb_16 follows the expected negative direction. As standardized him_tbb_16 increases, average predicted CTH decreases from approximately 8860 m to 6074 m for XGBoost, from 8590 m to 5089 m for CatBoost, and from 7902 m to 6998 m for Random Forest. This relation is physically plausible because warmer thermal-window brightness temperatures generally correspond to lower and warmer cloud tops, whereas colder observations are associated with higher cloud tops. The negative response to him_tbb_10 is also apparent for the three models.

Cloud_fraction produced a strong decreasing average response. Across the evaluated standardized range, XGBoost decreased from approximately 8942 m to 5840 m, CatBoost from 9396 m to 5596 m, and Random Forest from 9129 m to 6077 m. This empirical relation should not be interpreted as a universal physical law. Instead, it indicates that in the collocated dataset, high cloud-fraction scenes were associated with a different CTH distribution and more homogeneous radiometric conditions than partially cloudy scenes.

Table 5. Fraction of PDP Segments Violating The Expected Negative Thermal Response

Feature	XGBoost	CatBoost	Random Forest
him_tbb_16	0.069	0.000	0.000
him_tbb_11	1.000	0.724	0.414
him_tbb_14	0.897	0.000	0.655

The monotonicity audit in Table 5 demonstrates that predictive importance and physical regularity are not equivalent. Him_tbb_16 was both highly important and relatively consistent: CatBoost and Random Forest had no detected violations on the audited PDP grid, and XGBoost had a small violation fraction of 0.069. In contrast, XGBoost showed violation fractions of 1.000 for him_tbb_11 and 0.897 for him_tbb_14. CatBoost was more regular for him_tbb_14 but still showed substantial violations for him_tbb_11. Random Forest exhibited intermediate behavior.

These findings illustrate an accuracy-regularity trade-off. XGBoost achieved the best global metrics, yet it did not produce the most physically regular response for every thermal channel. The non-monotonic behavior may reflect legitimate multivariate interactions, correlated channels, mixed cloud conditions,

or model fragmentation rather than a simple physical error. ICE curves in the thesis analysis also showed that individual responses were heterogeneous, meaning that a smooth PDP can hide different sample-level behavior. For this reason, the violation fractions should be viewed as flags for further meteorological investigation, not as standalone proof that a model is physically invalid.

E. Error Structure Across CTH Regimes

The strongest limitation of the three models becomes visible after stratifying errors by reference CTH. Table 6 shows that the lowest RMSE occurred in the middle-altitude interval of 6910-8439 m. This interval also contained the largest number of observations, which gave the models extensive representation during learning. Errors increased substantially at both the low and high ends of the target distribution

Table 6. Fraction of PDP Segments Violating The Expected Negative Thermal Response

CTH interval (m)	Sample count	XGBoost RMSE (m)	CatBoost RMSE (m)	Random Forest RMSE (m)
793-2322	245,120	1964.39	1926.20	2056.40
6910-8439	901,018	990.27	1009.63	1069.85
13026-14556	4,543	4262.59	4411.67	4633.54
14556-16085	332	5152.52	4812.64	5878.33

At low CTH, all models tended to overpredict. In the 793-2322 m interval, the biases were +1608.92 m for XGBoost, +1590.08 m for CatBoost, and +1617.56 m for Random Forest. In the middle interval, the biases approached zero (-19.60 m, +17.33 m, and +4.25 m, respectively). At the extreme upper interval of 14556-16085 m, the direction reversed sharply: biases reached -5137.00 m for XGBoost, -4776.10 m for CatBoost, and -5853.13 m for Random Forest.

The sign reversal indicates regression toward the central region of the target distribution. Low clouds are pulled upward, while very high clouds are pulled downward. The upper-tail failure is especially important because the highest interval contained only 332 observations, compared with 901,018 observations in the most stable middle interval. A loss function optimized over the full dataset therefore receives far more information from common mid-level cases than from rare extreme cloud tops. The result is strong aggregate performance combined with poor preservation of the target extremes.

The difficulty is also physically plausible. Very high CTH may be associated with deep convective systems, overshooting tops, complex multilayer structures, or retrieval conditions in which passive brightness temperatures alone do not uniquely determine geometric height. Moreover, TROPOMI itself may underestimate or show increased uncertainty for some high-cloud situations [10]. Thus, the extreme residuals can reflect both model limitations and uncertainty in the reference product. CatBoost produced a lower RMSE than XGBoost in the highest interval, demonstrating that the globally best model is not necessarily the best model for every operationally important subset.

F. Error Validation Across Cloud-Fraction Regimes

Stratification by cloud fraction produced a different pattern. The largest errors did not occur in the most cloud-filled scenes. XGBoost and CatBoost reached their highest cloud-fraction-stratified RMSE in the interval 0.338-0.432, with values of 1729.61 m and 1796.83 m. Random Forest reached its largest value, 1925.14 m, in the lowest interval of 0.054-0.149. By contrast, all models achieved their lowest RMSE in the highest cloud-fraction interval of 0.905-1.000: 972.18 m for XGBoost, 1006.77 m for CatBoost, and 1031.04 m for Random Forest.

High cloud fraction likely produces a more homogeneous footprint in which cloud radiance dominates the observation. Low and intermediate cloud fractions may contain cloud edges, broken cloud fields, or mixtures of surface and cloud emission within one footprint. Such mixtures weaken the relation between observed brightness temperature and the retrieved cloud top. The biases in the highest cloud-fraction interval were also close to zero (+16.93 m, +31.81 m, and +38.24 m), whereas Random Forest produced a bias of -599.24 m in the lowest interval.

This result has two implications. First, `cloud_fraction` is not merely a statistical convenience; it identifies scene conditions that alter the reliability of the spectral-height relationship. Second, an operational system can use cloud fraction as a quality-context variable. Predictions from low or intermediate cloud-fraction scenes should receive greater caution, especially when accompanied by high model disagreement or unusual spectral combinations.

G. Granual Error Structure

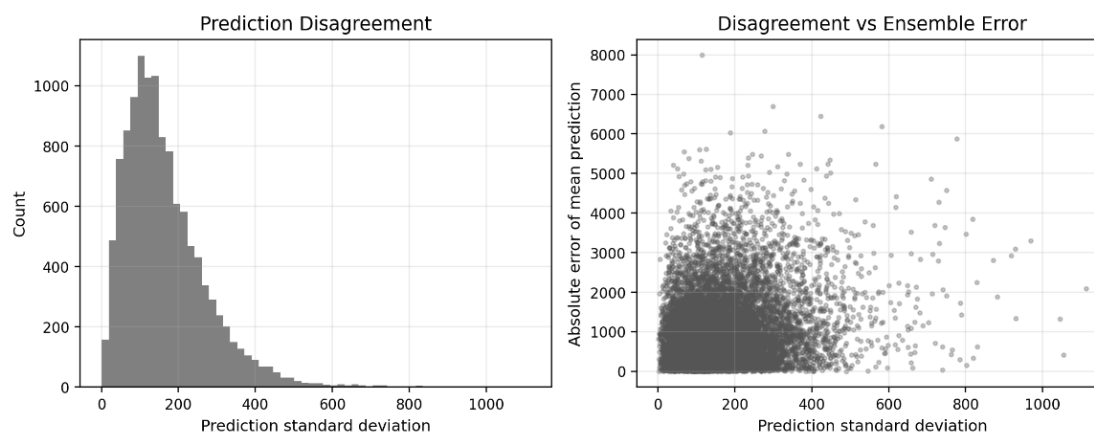


Figure 4. Distribution of inter-model disagreement and its relationship with ensemble absolute error.

The model predictions were strongly correlated. The correlation was approximately 0.987 between XGBoost and CatBoost, 0.976 between XGBoost and Random Forest, and 0.977 between CatBoost and Random Forest. XGBoost and CatBoost were the most similar pair, with a mean absolute prediction difference of 209.19 m and an inter-prediction RMSE of 279.43 m. The corresponding mean absolute differences were 292.65 m for XGBoost-Random Forest and 277.30 m for CatBoost-Random Forest.

Despite the high agreement, local divergence contained useful information. The correlation between prediction standard deviation and the absolute error of the ensemble mean was approximately 0.192. This is a positive but weak association. The 5% of observations with the lowest disagreement had a mean absolute error of approximately 786.34 m, whereas the 5% with the highest disagreement reached approximately 1465.16 m. High-disagreement scenes also tended to have lower cloud fraction; cloud fraction was negatively correlated with absolute error at approximately -0.268.

Figure 4 therefore supports the use of disagreement as a screening signal, but not as a calibrated probability of error. A low standard deviation cannot guarantee correctness because all three models may

share the same bias, particularly for extremely high clouds. Conversely, disagreement can increase because the algorithms generalize differently even when the ensemble mean remains reasonable. The signal is most useful for prioritizing cases for additional checks, human review, or comparison with other meteorological observations. Formal operational uncertainty would require calibration methods such as quantile models, conformal prediction, probabilistic ensembles, or reliability analysis on independent periods.

H. Comparison with Previous Studies

Critically analyzing these results against previous literature highlights the fundamental limitations of passive sensor retrievals. The XGBoost RMSE of approximately 1.30 km in this study demonstrates a significant improvement compared to the 1.74 km value reported for the Himawari-based XGBoost study by Dong et al. [3]. While this indicates algorithmic success, our regime-specific error analysis reveals a critical physical limitation: the severe underprediction at the extreme upper tail ($\text{CTH} > 14$ km) persists across all models.

The physical cause of this underprediction is deeply rooted in atmospheric theory. Optically thick deep-convective clouds saturate the thermal infrared signals, causing the brightness temperature to reach a minimum limit at the tropopause [1], [5]. Consequently, passive sensors struggle to distinguish the geometric height of overshooting tops, a challenge also documented in multi-sensor validation studies [10], [19]. The implication of this finding is profound: the error in predicting extreme high clouds is not merely a statistical flaw of the machine learning algorithms, but a physical saturation limit of the input features. This suggests that future advancements require integrating external physical constraints, rather than just changing the tabular algorithm.

I. Comparison with Previous Studies

Several limitations define the scope of the conclusions. First, the target is a TROPOMI retrieval and not a direct geometric measurement. Errors or biases in the reference product are inherited by the supervised models. This is especially relevant to thin, multilayer, and high-cloud conditions [10]. Validation with active sensors, ground-based cloud radar, lidar, all-sky imaging networks, or independent aircraft observations would provide a stronger assessment of absolute CTH accuracy [28], [29].

Second, the predictors are tabular summaries and do not fully represent the three-dimensional atmosphere. The models do not explicitly use vertical temperature and humidity profiles, cloud phase, optical thickness, emissivity, surface temperature, or satellite viewing angles. If two atmospheric scenes produce similar spectral observations but different geometric heights, the missing physical context imposes a limit that cannot be eliminated solely through algorithm tuning. Future models should incorporate atmospheric reanalysis or numerical-weather-prediction profiles and spatial texture features and multisource retrieval constraints [30].

Third, interpretation methods are affected by correlations among channels. Permuting one channel can create unrealistic feature combinations, while SHAP can distribute contributions among correlated predictors in method-dependent ways. PDP also averages across combinations that may be rare in the observed atmosphere. The use of multiple methods reduces reliance on any single explanation, but it does not establish causal relationships between a channel and CTH.

Fourth, the monotonicity audit uses a simplified physical expectation. A negative relation between thermal brightness temperature and CTH is reasonable for many optically thick single-layer clouds, but it is not universally valid across all channels and atmospheric states. Therefore, the reported violations should guide targeted investigation by cloud type, season, and scene condition rather than be interpreted as automatic model rejection.

Fifth, the evaluation is based on one spatial validation configuration and one observation year. Stronger evidence of generalization requires leave-region-out tests, seasonal holdouts, independent-year validation, and evaluation during extreme convective events. The contribution of latitude and longitude should also be tested by comparing otherwise identical models with and without explicit coordinates.

Sixth, disagreement among the three deterministic models is not a complete representation of uncertainty. It mainly reflects differences among the selected algorithms and cannot detect conditions in which all models fail in the same direction. The severe shared underprediction at extreme CTH is a clear example. Operational implementation should therefore combine model disagreement with regime-based warnings, cloud-fraction information, and independent meteorological indicators.

Within these limitations, XGBoost provides a strong baseline for high-frequency CTH estimation. Its aggregate performance, spectral reliance, and relatively small global bias support further development. Nevertheless, the model should not be used as the sole decision source for aviation hazards. An operational workflow should flag very high predicted or reference-like regimes, low and intermediate cloud-fraction scenes, and observations with high disagreement. Such flags can trigger additional validation with radar, lightning, numerical weather guidance, or forecaster assessment.

Improvement of the extreme upper tail should be prioritized. Potential strategies include sample reweighting, stratified training, specialized high-cloud regressors, two-stage classification-regression models, tail-sensitive loss functions, and the inclusion of vertical atmospheric predictors. Physical constraints may also be considered for channels with robust monotonic relationships, although constraints should be applied selectively to avoid suppressing legitimate nonlinear interactions.

J. Operational Integration for Aviation Meteorology

CTH should be interpreted as one component of an aviation-weather decision system rather than as a direct stand-alone hazard label. A high cloud top may indicate deep vertical development, but the operational significance of a cloud system also depends on its horizontal extent, growth rate, lightning activity, precipitation structure, wind environment, and position relative to an aircraft route. Consequently, the proposed model is most useful as a rapid satellite-derived layer that complements radar, numerical-weather-prediction guidance, lightning observations, and forecaster interpretation [2], [15].

A practical implementation can produce three outputs for every valid satellite pixel: the predicted CTH, a regime-based quality flag, and a diagnostic confidence indicator. The quality flag can incorporate cloud fraction, spectral validity, and whether the prediction lies near a poorly represented CTH range. The confidence indicator can incorporate inter-model disagreement, while retaining the warning that low disagreement does not guarantee correctness. This layered output is more informative than a single deterministic CTH map because it communicates both the estimated atmospheric structure and the conditions under which the estimate should be treated cautiously.

The extreme-cloud analysis suggests that conservative handling is required when the model estimates high CTH or when other observations indicate deep convection. Because the tested models compress very high reference values toward the distribution center, an operational threshold based only on the raw prediction could miss some severe cases. A safer approach is to apply an upper-tail correction, a dedicated high-cloud model, or an alert rule that combines the predicted CTH with cold brightness temperatures and rapid temporal cooling. Such a rule should be validated independently and should not be interpreted as an automatic flight-safety decision.

The high temporal frequency of Himawari creates an opportunity to monitor cloud-top evolution rather than isolated snapshots [3]. Consecutive CTH maps can be used to identify rapid vertical development, persistent high-cloud regions, or abrupt changes in prediction confidence. Temporal

monitoring may be especially valuable when a cloud system moves through areas with limited ground-based observations. However, frame-to-frame consistency must be evaluated because small radiometric changes or missing pixels can cause artificial jumps in a purely tabular model. Temporal smoothing or sequence-aware models may reduce this instability, provided that they do not suppress genuine rapid convective growth.

Before deployment, the system should undergo prospective monitoring for data drift and seasonal performance changes. Satellite calibration updates, changes in the reference product, unusual climate conditions, and differences between training and operational domains can alter the feature-target relationship. Performance dashboards should therefore report overall metrics together with errors by altitude, cloud fraction, region, and time period. Retraining should be triggered by documented degradation rather than performed without evaluation. A human-in-the-loop design remains essential: the model can prioritize scenes, summarize vertical cloud development, and highlight low-confidence areas, while qualified meteorological personnel retain responsibility for interpreting the complete operational context.

K. Operational Integration for Aviation Meteorology

The first priority for future work is to improve representation of the upper and lower tails of the CTH distribution. The present training distribution is dominated by mid-level cases, while the highest interval contains only a few hundred validation observations. Future experiments should compare sample reweighting, stratified mini-batches, tail-aware loss functions, and separate specialist models for low and very high clouds. A two-stage design may first classify the atmospheric regime and then apply a regime-specific regressor. Such a strategy could reduce regression toward the mean while preserving stable performance in the common middle range.

The second direction is feature enrichment. Brightness temperatures and cloud fraction carry substantial information, but they do not fully resolve atmospheric vertical structure. Temperature and humidity profiles from reanalysis or numerical models, cloud phase, optical thickness, surface characteristics, viewing geometry, and temporal change rates should be evaluated as additional predictors. Spatial texture around each pixel may also help distinguish homogeneous cloud decks from convective towers and broken cloud boundaries. Hybrid methods that combine radiative-transfer constraints with machine learning may provide stronger extrapolation than a purely data-driven tabular mapping [30].

The third direction is stronger external validation. Leave-region-out evaluation can determine whether the learned relationship transfers across different parts of Indonesia, while seasonal and independent-year holdouts can reveal temporal drift. Validation against active sensors, ground-based radar or lidar, and all-sky imaging systems can separate model error from uncertainty in the TROPOMI reference product [23], [28], [29]. Particular attention should be given to deep-convective events because average annual validation may contain too few extreme examples to support safety-oriented conclusions.

The fourth direction concerns uncertainty quantification. The present disagreement analysis is useful for screening but is not calibrated. Quantile regression, conformal prediction, bootstrap ensembles, and probabilistic boosting should be compared using coverage and sharpness metrics. Calibration should also be evaluated separately by CTH range, cloud fraction, region, and season. An uncertainty interval that performs well on average but fails for high clouds would not resolve the main operational weakness identified in this study.

Finally, future work should extend the audit from feature-level interpretation to spatial and representation-level analysis. Tree-leaf embeddings, neural-network hidden representations, and spatial deep-learning models can be examined to determine how cloud regimes are organized internally. Any increase in model complexity should be accompanied by equally strong interpretability, physical-

consistency, and error-regime evaluation. The objective is not merely to obtain a lower global RMSE, but to develop a CTH product whose limitations are transparent, whose uncertainty is measurable, and whose behavior remains reliable under the atmospheric conditions most relevant to aviation operations.

4. CONCLUSION

This study evaluated XGBoost, CatBoost, and Random Forest for predicting Cloud Top Height from Himawari spectral observations using Sentinel-5P/TROPOMI as the reference product. On 3,444,543 spatial validation observations, XGBoost achieved the best aggregate performance with MAE of 971.71 m, RMSE of 1297.57 m, R^2 of 0.646, and mean bias of -33.55 m. CatBoost and Random Forest followed with RMSE values of 1340.95 m and 1371.11 m.

The behavioral audit shows that spectral channels and cloud fraction are the main predictive information sources. Coordinates have much smaller global importance, although longitude contributes to selected local predictions. Him_tbb_16 combines high predictive importance with a physically plausible average response, while other channels show stronger monotonicity violations and therefore require more cautious interpretation.

The most important limitation is the concentration of errors at the target extremes. The models are relatively stable for mid-level CTH but overpredict low clouds and strongly underpredict extremely high cloud tops. This behavior is associated with severe sample imbalance and the complexity of passive high-cloud retrieval. High cloud fraction generally corresponds to lower errors, whereas partially cloudy scenes are more ambiguous.

Inter-model disagreement is positively related to actual error and can support diagnostic screening, but the association is too weak for disagreement to be interpreted as calibrated uncertainty. Overall, XGBoost is the strongest baseline among the tested models. To ensure reliable operational use, several specific directions for future research are highly recommended. Regarding methodological development, future studies should explore two-stage classification-regression models or tail-sensitive loss functions to explicitly mitigate the underprediction of extreme high-cloud regimes. In terms of feature addition, incorporating vertical atmospheric profiles (temperature and humidity from numerical models) alongside spatial texture features is critical to resolving thermal signal ambiguities. Furthermore, increasing the volume of data by extending the temporal coverage to multi-year datasets will significantly improve the statistical representation of rare, extreme convective events. Finally, regarding the expansion of research objects, the proposed models must be cross-validated across a broader range of geographical domains—such as comparing different tropical and subtropical maritime continents—to definitively prove their global spatial transferability.

ACKNOWLEDGEMENT

The authors express their gratitude to the relevant meteorological agencies for providing the Himawari-8 and Sentinel-5P/TROPOMI datasets, and the Master of Informatics Engineering Program at Universitas Pamulang for the academic support during this research.

REFERENCES

- [1] Chahine, M. T. (1974). Remote Sounding of Cloudy Atmospheres. I. The Single Cloud Layer. *Journal of the Atmospheric Sciences*, 31(1), 233–243.
- [2] Stephens, G. L., et al. (2008). CloudSat mission: Performance and early science after the first year of operation. *Journal of Geophysical Research: Atmospheres*, 113(D8), 2008JD009982.
- [3] Dong, Y., Sun, X., & Li, Q. (2022). A Method for Retrieving Cloud-Top Height Based on a Machine Learning Model Using the Himawari-8 Combined with Near Infrared Data. *Remote Sensing*, 14(24), 6367.

- [4] Fischer, J., Cordes, W., Schmitz-Peiffer, A., Renger, W., & Mörl, P. (1991). Detection of Cloud-Top Height from Backscattered Radiances within the Oxygen A Band. Part 2: Measurements. *Journal of Applied Meteorology*, 30, 1260-1267.
- [5] Kahn, B. H., Chahine, M. T., Stephens, G. L., Mace, G. G., Marchand, R., Wang, Z., Barnet, C., Eldering, A., Holz, R. E., Kuehn, R. E., & Vane, D. (2008). Cloud type comparisons of AIRS, CloudSat, and CALIPSO cloud height and amount. *Atmospheric Chemistry and Physics*, 8, 1231-1248.
- [6] Baran, A., Lerch, S., Ayari, M. E., & Baran, S. (2021). Machine learning for total cloud cover prediction. *Neural Computing & Applications (Print)*.
- [7] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- [8] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
- [9] Lenhardt, J., Quaas, J., & Sejdinović, D. (2024). Marine cloud base height retrieval from MODIS cloud properties using machine learning. *Atmospheric Measurement Techniques*, 17, 5655-5668.
- [10] Compernelle, S., Argyrouli, A., Lutz, R. B., Sneep, M., Lambert, J., Fjæraa, A. M., Hubert, D., Keppens, A., Loyola, D., O'connor, E., Romahn, F., Stammes, P., Verhoelst, T., & Wang, P. (2021). Validation of the Sentinel-5 Precursor TROPOMI cloud data with Cloudnet, Aura OMI O2–O2, MODIS, and Suomi-NPP VIIRS. *Atmospheric Measurement Techniques*, 14, 2451-2476.
- [11] Cheng, Y., He, H., Xue, Q., Yang, J., Zhong, W., Zhu, X., & Peng, X. (2024). Remote Sensing Retrieval of Cloud Top Height Using Neural Networks and Data from Cloud-Aerosol Lidar with Orthogonal Polarization. *Sensors*, 24(2), 541.
- [12] Hong, G., Yang, P., Huang, H.-L., Baum, B. A., Hu, Y., & Platnick, S. (2007). The Sensitivity of Ice Cloud Optical and Microphysical Passive Satellite Retrievals to Cloud Geometrical Thickness. *IEEE Transactions on Geoscience and Remote Sensing*.
- [13] King, M. D., Platnick, S., Menzel, W. P., Ackerman, S. A., & Hubanks, P. A. (2013). Spatial and Temporal Distribution of Clouds Observed by MODIS Onboard the Terra and Aqua Satellites. *IEEE Transactions on Geoscience and Remote Sensing*, 51, 3826-3852.
- [14] Håkansson, N., Adok, C., Thoss, A., Scheirer, R., & Hörnquist, S. (2018). Neural network cloud top pressure and height for MODIS. *Atmospheric Measurement Techniques*, 11, 3177-3118.
- [15] Biondi, R., Ho, S., Randel, W. J., Syndergaard, S., & Neubert, T. (2013). Tropical cyclone cloud-top height and vertical temperature structure detection using GPS radio occultation measurements. *Journal of Geophysical Research: Atmospheres*, 118, 5247-5259.
- [16] Wang, M., Min, M., Li, J., Lin, H., Liang, Y., Chen, B., Yao, Z., Xu, N., & Zhang, M. (2024). Technical note: Applicability of physics-based and machine-learning-based algorithms of a geostationary satellite in retrieving the diurnal cycle of cloud base height. *Atmospheric Chemistry and Physics*, 24(24), 14239–14256.
- [17] Li, J., Zhang, F., Li, W., Tong, X., Pan, B., Li, J., Han, L., Letu, H., & Mustafa, F. (2023). Transfer-Learning-Based Approach to Retrieve the Cloud Properties Using Diverse Remote Sensing Datasets. *IEEE Transactions on Geoscience and Remote Sensing*.
- [18] Li, T., Chen, N., Tao, F., Hu, S., Xue, J., Han, R., & Wu, D. (2024). Cloud Top Height Retrieval from FY-4A Data: A Residual Module and Genetic Algorithm Approach. *Atmosphere*, 15(6), 643.
- [19] Liu, B., Huo, J., Lyu, D., & Wang, X. (2021). Assessment of FY-4A and Himawari-8 Cloud Top Height Retrieval through Comparison with Ground-Based Millimeter Radar at Sites in Tibet and Beijing. *Advances in Atmospheric Sciences*, 38(8), 1334–1350.

-
- [20] Marchand, R., Ackerman, T. P., Smyth, M., & Rossow, W. B. (2010). A review of cloud top height and optical depth histograms from MISR, ISCCP, and MODIS. *Journal of Geophysical Research: Atmospheres*.