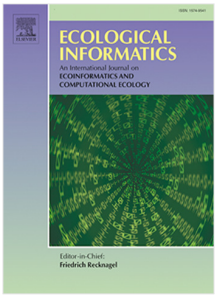
Journal Pre-proof



Retrieval of planetary boundary layer height from CALIPSO satellite observations using a machine learning approach

A. Salcedo-Bosch, F. Rocadenbosch, C.I. Argañaraz, G. Curci, S. Lolli

PII: S1574-9541(25)00440-6
DOI: https://doi.org/10.1016/j.ecoinf.2025.103431
Reference: ECOINF 103431

To appear in: *Ecological Informatics*

Received date : 15 May 2025
Revised date : 11 September 2025
Accepted date : 13 September 2025

Please cite this article as: A. Salcedo-Bosch, F. Rocadenbosch, C.I. Argañaraz et al., Retrieval of planetary boundary layer height from CALIPSO satellite observations using a machine learning approach. *Ecological Informatics* (2025), doi: https://doi.org/10.1016/j.ecoinf.2025.103431.

Retrieval of Planetary Boundary Layer Height from CALIPSO Satellite Observations Using a Machine Learning Approach

A. Salcedo-Bosch[1], F. Rocadenbosch[2], C.I. Argañaraz[3], G. Curci[3] and S. Lolli[1]

[1] CNR-IMAA, Contrada S. Loja snc, 85050 Tito Scalo (PZ), Italy
[3] Physics Dept., Univ. degli Studi dell'Aquila, Via Vetoio, 67100 L'Aquila, Italy
[2] CommSensLab, Universitat Politècnica de Catalunya (UPC), Jordi Girona, 31, Barcelona, 08034, Spain

E-mail: `simone.lolli@cnr.it`

Abstract. The planetary boundary layer height (PBLH) is a key variable in air quality, climate modeling, and weather prediction. Traditional retrieval methods, such as radiosondes, provide high accuracy but lack spatial coverage. This study presents a Random Forest (RF) model based on Machine Learning (ML) to estimate PBLH from ten years of Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) on board the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO), using radiosonde measurements as a reference. The model achieves an R^2 of 0.67 and an RMSE of 278.02 m with a spatial resolution of $\approx$ 20x20 km^2 in a test set that covers mainly Europe and North America. Unlike previous methods, our approach does not require atmospheric typing and uses minimal data filtering, demonstrating robustness under diverse aerosol and cloud conditions. Although validation is currently limited to mid-latitude regions, the method offers a scalable approach to global monitoring and supports the management of climate and air quality. Future work will extend the validation to other geographic zones and explore deep learning models for further improvements.

Submitted to:

1. Introduction

The Planetary Boundary Layer (PBL), in direct contact with the Earth's surface [1], significantly influences air quality by regulating the dispersion of pollutants [2]. In fact, the depth of the boundary layer determines the spatial volume available for the emission of surface pollutants, having a direct influence on their concentration levels [3], and monitoring of these pollutants, such as $PM_{2.5}$, is a critical aspect of urban air quality management that integrates ground-based, satellite, and machine learning approaches [4]. In extratropical regions, especially during winter high-pressure anticyclones, a shallow PBL contributes to severe pollution events and haze formation, because of insufficient solar radiation to drive convection and mixing [5]. In contrast, in tropical areas, the circulation of the monsoon primarily controls the depth of the PBL [6]. Accurate estimation of PBL height (PBLH) is essential to improve air quality forecasts and reanalysis, particularly in urban and metropolitan settings where pollutant accumulation is a major concern [7]. However, PBLH measurements remain sparse due to the limited number of observational platforms, with global data primarily obtained from a combination of independent national radiosonde networks. These networks, although extensive, do not cover every geographic region uniformly, resulting in data gaps. The World Meteorological Organization (WMO) radiosounding network plays an essential role in filling these gaps by providing standardized and coordinated observations in different countries. Nevertheless, the current density of stations in this network remains inadequate for full global coverage, requiring PBLH measurements to rely significantly on regional projects and cooperative efforts to enhance the availability and reliability of data. The entrainment zone at the top of the PBL plays a significant role in air pollution. Its complexity due to dynamics such as turbulence and convection allows the transfer of thermal energy, moisture, and atmospheric gases [1], making its vertical structure critical for multiple scientific applications. These include photovoltaic energy optimization [8], numerical weather prediction [9], and climate modeling [10], among others, improving their representation of atmospheric mixing processes.

Common methodologies for retrieving PBLH using thermodynamical variables include the parcel method [11], the gradient method [12], and the Richardson number method [13]. These techniques incorporate temperature, humidity, and wind atmospheric profiles, which are derived from rawinsonde measurements [13]. Since most modern radiosonde (RS) launches are in fact rawinsondes, we refer to them simply as RS throughout this manuscript. Ground-based remote sensing techniques, such as lidars [14], microwave radiometers [15], ceilometers [16], and radar wind profilers [17], offer additional means of estimation of PBLH, often using a proxy, i.e. aerosols as tracers for lidars and ceilometers, including synergistic techniques for its disambiguation. Although widely studied and proved accurate, ground-based estimations offer limited spatial coverage and require the deployment of permanent observational sites, which, at remote locations, results in too expensive or even unfeasible results. In order to obtain global retrievals of the PBLH in a cost-effective manner, space-based remote sensing

instruments have been used, including GPS radio occultation [18], the Atmospheric Infrared Sounder (AIRS) [19], and the Moderate Resolution Imaging Spectroradiometer (MODIS) [20]. However, they offer coarse spatial resolution and limited estimation accuracy.

One of the most widely used space-born instruments for PBLH estimation is the Cloud-Aerosol lidar with Orthogonal Polarization (CALIOP; [21]) onboard the Cloud-Aerosol lidar and Infrared Pathfinder Satellite Observation (CALIPSO) [22, 23, 24], despite missing an official mission product. CALIOP level 1 (L1) data products encompass the calibrated and geolocated lidar backscatter profiles [25], from which level 2 (L2) products such as the cloud base and highest altitudes [26], and the aerosol layer typing [27] are derived. Although providing vertically attenuated high-resolution backscatter profiles, the retrieval of PBLH from CALIOP presents significant challenges due to the low signal-to-noise ratio (SNR) [21], surface backscatter contamination [28], signal attenuation near the surface [21], and dust and cloud screening [22]. Numerous methodologies have been developed in the literature to address these challenges. For example, Zhang et al. [23] applied a maximum variance algorithm (MV) [29] combined with the Haar-wavelet covariance transform (WCT) [30] to estimate daytime PBLH in China, achieving moderate correlations with radiosonde data ($R = 0.65$ in rural areas, $R = 0.59$ in urban areas), but their approach relied on cloud layer products of CALIPSO Level 2 and required a coarse resolution of 0.2^{o} x 0.2^{o} ($\sim$ 20km x 20km). Similarly, Leventidou et al. [22] used CALIPSO level 2 aerosol classification products to obtain PBLH, and, by dust removal scenarios, produced more accurate PBLH retrievals in Thessaloniki, Greece. Although promising, the results achieved were limited to a single location and were based on scenarios in which the advected dust layers were manually removed. On the other hand, Kim et al. [24] applied the WCT technique to CALIOP data, validating the results against a ground-based lidar in Korean Peninsula. Although their method improved instrument agreement ($R = 0.81$, Root Mean Squared Difference ($RMSD$) = 250 m during the day; $R = 0.51$, $RMSD$ = 560 m at night), the results found were again limited to a single location and required *manual removal of multilayer aerosol profiles*. Similarly, Su et al. [31] applied the WCT over CALIOP measurements to obtain PBLH estimates over Hong Kong and validated the results with ground-based lidars and radiosonde launches. By applying a *cloud screening* method and a horizontal smoothing window ($\sim$ 7 km), they were able to obtain acceptable PBLH estimations from CALIOP with respect to radiosondes ($R = 0.52$ and root mean squared error ($RMSE$) = 277 m) and ground-based lidars ($R = 0.65$ and $RMSE = 237$ m). Liu et al. [32] presented an image processing algorithm that grouped the CALIOP-measured backscatter signal into two categories using a k-means approach, which were then identified as the clusters above and below the PBLH. The method was able to achieve high PBLH measurement accuracy at high spatial resolution ($\sim$ 5 km) with respect to a ground-based lidar in Wuhan, China ($R^2 = 0.7$ and $RMSE = 390$ m). Unfortunately, the algorithm was only tested at night and proved to be ineffective in cloudy or dusty scenarios.

The studies mentioned above highlight limitations in CALIOP-derived PBLH estimates. On the one hand, there is an overestimation of PBLH in presence of multiple aerosol layers or under high aerosol loading, leading to retrieval biases relative to ground-based lidar and RS measurements [24]. However, there are discrepancies in the nighttime PBLH retrievals between CALIOP, ground-based lidar, and radiosonde data, particularly in stable boundary layers [31]. Finally, there is a limited range of seasons, locations, measurement times, and atmospheric scenarios in which the methods were validated.

To address these challenges and fill an important research gap, we propose a novel data fusion method based on decision tree machine learning (ML) techniques to estimate daytime and nighttime PBLH from CALIOP Level 1 data measurements combined with historical radiosonde data. Using a comprehensive dataset of radiosonde launches and intersecting CALIPSO overpasses spanning 10 years, our approach retrieves PBLH from CALIOP total attenuated backscatter (TAB) profiles without requiring data filtering, atmospheric typing, or reliance on CALIOP Level 2 data, thus utilizing historical data from the radiosonde dataset. ML methods consider CALIOP measurements and contextual data (e.g. latitude, longitude, and altitude) as input features and are trained to predict the PBLH as measured by radiosondes. Additionally, the method is validated using radiosonde launches across extensive regions of the world with a high density of radiosonde stations, specifically in North America and Europe, where data quality and availability are ensured. In comparison to alternative techniques such as ground-based lidars, RS can provide PBLH estimates in a wide range of atmospheric conditions, including cloudy skies or high aerosol loads. This robustness allows for a standardized assessment of method performance under diverse environmental scenarios, allowing more comprehensive intercomparisons.

Accurate and large-scale estimation of PBLH is critical for air quality management, climate adaptation, and environmental policy [33, 34, 35, 36]. Real-time monitoring can improve early warning systems for pollution events, enabling targeted actions such as traffic restrictions [37] and emission controls, as well as health alerts for vulnerable populations [38]. Integrating satellite-based PBLH data into forecasting models improves pollution dispersion predictions [39], supporting evidence-based regulations, especially in urban and industrial areas. Therefore, this study offers a scalable approach that empowers policymakers, researchers, and planners with better tools to manage air quality, urban heat, and renewable energy deployment.

The paper is organized as follows. Section 2.1 describes the radiosonde datasets and the CALIPSO mission. Sections 2.3 to 3.1 outline the methodologies for estimating PBLH from radiosonde data and introduce the ML framework. Section 3 presents and discusses the PBLH estimation results. Section 4 concludes with final remarks.

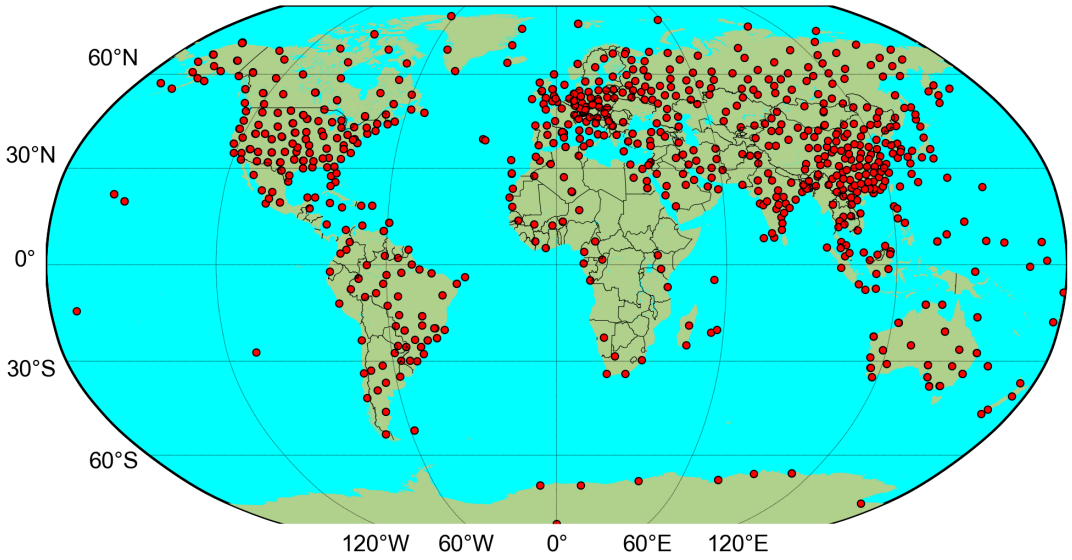


Figure 1. World map showing the location of the 693 radiosonde launching stations (red dots) considered in this work as a reference for CALIPSO retrievals of the PBLH.

2. Materials and Methods

2.1. Instrumentation

Satellite lidar observations The CALIPSO mission [21], launched on April 28, 2006, is part of NASA's A-Train satellite constellation and aims to improve understanding of the atmospheric system by providing detailed observations of clouds and aerosols in the troposphere and low stratosphere. It follows a sun-synchronous polar orbit (705 km), with a 16-day revisit cycle and an equator-crossing time of 13:30 local solar time. After 2018, CALIPSO joined the C-Train satellite constellation at a lower orbit (688 km) [40], essentially mantaining the revisit cycle and equator crossing time [41]. CALIPSO's core instrument, the CALIOP, is a solid-state laser that emits simultaneous co-aligned pulses at 1064 and 532 nm wavelengths in a near-nadir configuration [42]. CALIOP measures total-attenuated-backscatter (TAB) vertical profiles at a spatial resolution of 30 m vertically and 335 m horizontally, being able to measure the vertical distribution, altitude, thickness, and optical properties of clouds and aerosols, critical for climate modeling and weather forecasting. CALIPSO data are publicly available through NASA's Earthdata portal [43], enabling exploration of how clouds and aerosols influence Earth's energy balance, atmospheric chemistry, and meteorological patterns. The mission produces different data products [44], including detailed profiles

and classifications of clouds and aerosols, aiding research in climate dynamics and atmospheric interaction. CALIPSO's advanced technology and methodologies provide essential insights on atmospheric phenomena, supporting climate change research, air quality assessment, and predictions in meteorological science. Although the mission ended after 17 years of operation (1st August 2023), CALIPSO remains a key contributor to the scientific understanding necessary to manage global environmental challenges. Along with few system malfunctioning cases, in the last 7 years of CALIPSO mission, CALIOP emitted intermittent low energy laser pulses which degraded those measurements quality [45].

Radiosonde measurements In this work, we combined radiosonde data from multiple sources, covering worldwide launches conducted between 2010 and 2020. The datasets included:

- The German Weather Service (DWD),
- the Global Climate Observing System Reference Upper-Air Network (GRUAN website),
- the National Oceanic and Atmospheric Administration (NOAA),
- and the University of Wyoming upper-air sounding data-set (UWYO).

Combining these four datasets, a total of 693 launch stations were considered worldwide, as shown in Fig. 1. A total of 4,309,109 radiosonde launches were processed. Each processed launch included measurements of the vertical profile of wind (U), pressure (P), temperature (T), relative humidity (RH) and dew point temperature (T_d), which are essential for the estimation of PBLH from radiosondes using the bulk Richardson number method (see Sect. 2.3).

Region	# Stations	# Valid Launches
World	693	4,309,109
N. America	140	1,413,692
S. America	55	395,310
Europe	123	774,624
Africa	36	131,203
Asia	251	1,301,059
Oceania	43	121,825
Antarctica	7	21,288

Table 1. Description of the RS launches distribution by World region.

2.2. High Performance Computing Environment

Due to the large amount of data involved in the study, the data pre-processing and ML models training were carried out in the High Performance Computing environment

CALCULA at CommSensLab. CALCULA uses the Slurm Workload Manager as a resource management system and the GlusterFS system for data management. Details on the environment can be found in [46]. The CPU architecture used for the computations was an Intel Xeon Gold 5218 2.30GHz, and 64 GB of RAM were required.

2.3. Bulk-Richardson Number Method

The bulk Richardson number (Ri_b) method is a widely used approach to estimate PBLH [13]. Ri_b quantifies the ratio of buoyancy-induced turbulence to mechanical shear and is formulated as

$$Ri_b(z) = \frac{gz}{\theta_v(s)} \frac{\theta_v(z) - \theta_v(s)}{(U(z))^2}, \tag{1}$$

where g is the gravitational acceleration, z is the measurement altitude, θ_v is the virtual temperature, and U is the horizontal wind speed. s refers to the surface altitude. In other words, R_i expresses the ratio of buoyant forces to shear forces. Details on the computation of θ_v from RS data can be found in Eqs. 9-12 from [13].

A critical R_i threshold is used to identify the transition from a mixed turbulence-dominated layer to a more stable free atmosphere above the PBL, thus determining PBLH [47]. Threshold values typically range between 0.25 and 0.3, with minimal impact on PBLH estimates [48]. Following the study by Guo et al. [49], PBLH was derived from RS measurements as the altitude where R_i first exceeds a threshold of 0.25.

This method is a common and accepted approach in boundary layer meteorology [50]. It works effectively under different conditions as it can be applied to both convective and stable boundary layers. Additionally, it does not require additional data, as it can be computed directly from radiosonde profiles.

2.4. Decision-tree-based Machine Learning Methods Review

The objective of this study is to obtain a *model* able to represent the relationship between input data (CALIOP-measured TAB measurements) and a desired variable (RS-measured $PBLH$). This study utilizes ensemble ML methods based on decision trees (DTs) to estimate this *model*, due to their ability to handle non-linear problems while performing inherent scenario classification [51]. The two *methods* evaluated were Random Forest (RF) and Gradient Boosting (GB). The learning process and *model* structure obtained by RF and GB *methods* are controlled by configuration variables, the so-called hyperparameters.

Decision trees (DT) are a simple ML *method* that recursively partitions the input feature space X (here, the CALIOP-measured TAB profiles) into K regions $R = R_0, \ldots, R_{K-1}$. Each region is assigned a *prediction value* y_k (here, *the PBLH measured with RS*). The partitioning criteria of the feature space are defined as the nodes, the links between nodes are the branches, and the leaves represent the output values. DTs are built top-down, recursively splitting the feature space X into regions with the goal

of improving the homogeneity of the target values y_k within each split region R_k. DTs are defined by several *hyperparameters,* such as splitting criteria, maximum tree depth, and minimum samples per leaf. Although simple and interpretable, DTs often overfit training data and exhibit high variance.

Random Forests (RF) reduce the prediction variance of DTs by bagging, that is, using an ensemble of B independently trained DTs, each using a bootstrap sample of the training dataset [52], and thus reducing the variance of estimations by a factor B. RFs excel at handling high-dimensional data and outliers while minimizing the estimation variance. Their built-in randomness helps mitigate overfitting, particularly in datasets with noisy features, therefore enhancing generalization. Although they lack the straightforward interpretability of DTs, RFs can still reveal meaningful patterns and feature significance within a dataset. In this work, we resort to the *RandomForestRegressor* class of the Scikit-learn Python library for the implementation of RF. More details on the RF class used can be found in [53].

Gradient Boosting (GB) models are an ensemble predictor method similar to RFs in the sense that they combine multiple DTs. However, they build DTs sequentially, where each aims to minimize the error of its predecessor [54]. GB models often achieve higher predictive accuracy than RFs, but are more prone to overfitting.

2.5. Machine Learning Workflow

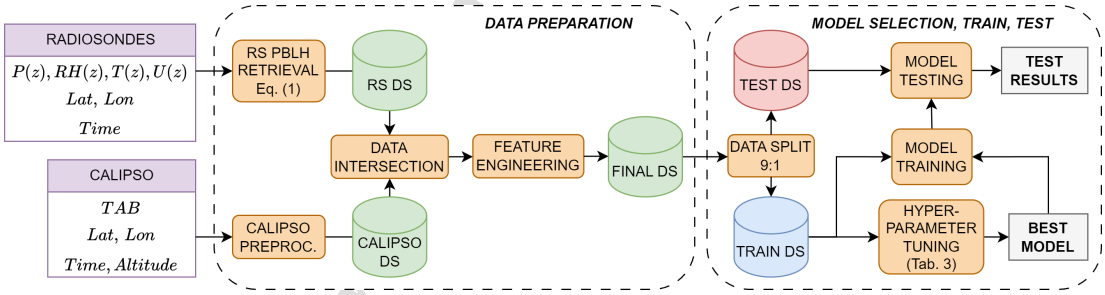


Figure 2. ML workflow. The workflow integrates data from radiosondes and CALIPSO satellite mission, involving data preparation steps like retrieval, intersection, and feature engineering to form the final dataset (DS). The dataset is then split into training and testing subsets for model selection, training, hyperparameter tuning, and evaluation to identify the best predictive model.

This section describes the workflow considered to obtain the ML PBLH estimation model. The diagram in Fig. 2 summarizes the procedure followed, which consists mainly of the preparation of the dataset followed by the selection, training, and testing of the model.

2.5.1. Data Preparation.- The raw measurement data for this study consisted of: (i) vertical profiles measured by RS of pressure ($P(z)$), relative humidity ($RH(z)$),

temperature ($T(z)$) and horizontal wind ($U(z)$) as a function of the measurement altitude (z); and (ii) CALIOP-measured TAB profiles. The data preparation steps were as follows:

(i) *RS PBLH Retrieval.-* PBLH values were calculated from RS measurements using the bulk-Richardson number method (see Sect. 2.3). The vertical profiles of the input variables were interpolated at a vertical resolution of 30 m, to match that of CALIOP. RS-derived PBLH values that exceeded 5000 m or equal to the launch altitude were excluded. The 5000 m threshold was applied as a quality control criterion, as higher values are typically the result of algorithmic misidentification of elevated inversions or free-tropospheric features rather than the true PBLH. RS launches with measurement gaps greater than 500 m were also removed, resulting in 2,162,107 PBLH ground-truth records, which make up the RS dataset.

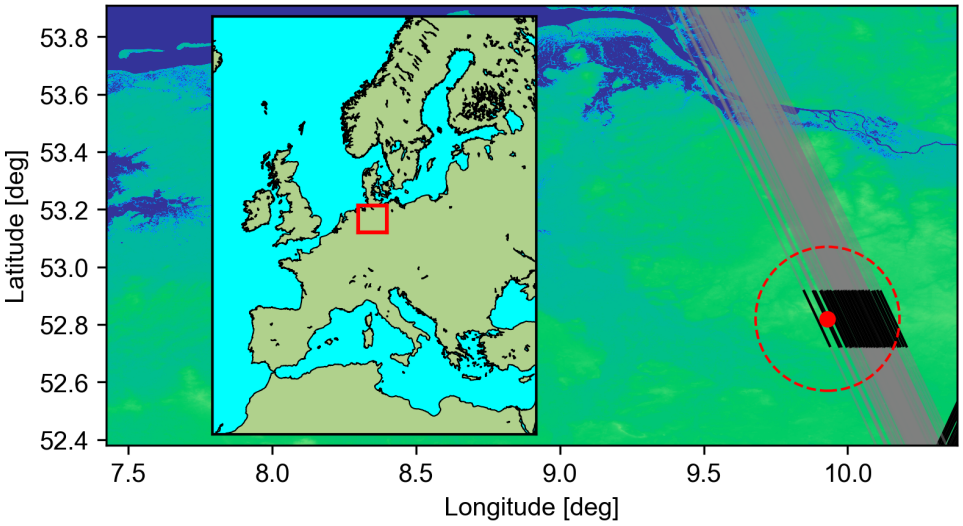


Figure 3. CALIPSO overpasses (gray traces) and Rotenburg RS launch station (red dot, N 52º49'12", E 9º55'48"). The red dashed circle indicates the maximum distance (0.25º) from the RS launching station to consider as a coincidental CALIPSO overpass. Black traces indicate CALIPSO TAB measurement segments (≈ 20 km long) used for this particular station.

(ii) *CALIPSO TAB profiles filtering.-* About 30% of the of all CALIOP profiles contain opaque layers, i.e., the backscatter signal is completely attenuated before reaching the surface, and thus, are not able to provide information for PBLH estimation. Therefore, profiles which do not have a surface reflection (maximum TAB below

$10^{-3} km^{-1} \cdot sr^{-1}$ at altitudes lower than 3000 m [55]) are filtered out. In order to eliminate the low-energy laser profiles and malfunctioning cases, laser profiles with pulse energy lower than 10 mJ [56] or QC flag different than "0" are removed as well.

(iii) *CALIPSO pre-processing.*- CALIPSO overpasses coinciding with RS launch station coordinates with a maximum distance of 0.25° (≈ 25 km) were selected (red dashed circle in Fig. 3). CALIOP level-1 TAB profiles were segmented into 2D windows (black traces in Fig. 3) consisting of 66 laser pulses, spanning approximately 20 km, and 180 vertical samples - covering around 5 km in height, including negative altitude values. Each resulting segment is a 66×180 matrix of TAB measurements, corresponding to an input dimensionality of 66×180 (Fig. 4a).

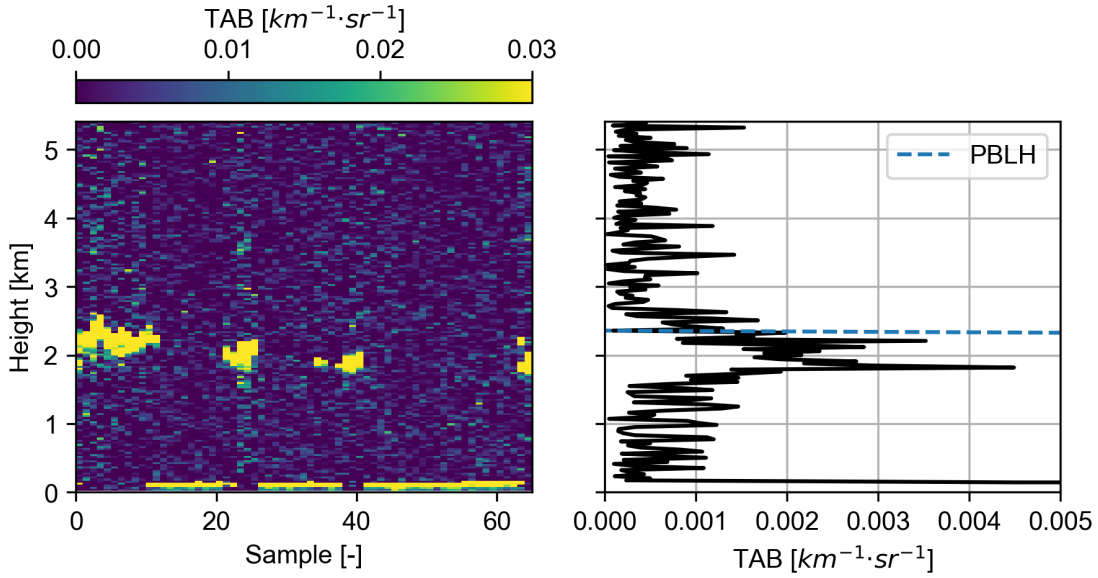


Figure 4. Pre-processing of CALIOP TAB measurements: (a) TAB profiles with 0.1° horizontal resolution, and (b) vertical TAB profile derived from horizontal median values.

(iv) *Data intersection.*- The RS and CALIPSO datasets were also intersected in time, considering a time difference of up to 3 hours as a coincidence overpass.

(v) *Feature extraction and reduction of dimensionality.*- First, to reduce the dimensionality of the input data of the CALIPSO (66×180) and the measurement noise, a median filter [57] was applied. The median filter computes the horizontal median TAB at each measurement altitude, reducing the dimensionality to 1×180 continuous values (Fig. 4b). The following statistical features were then extracted from the 1D TAB profiles to further reduce dimensionality and improve training performance: (i) signal energy (E_{TAB}), (ii) skewness ($skew$), (iii) kurtosis ($kurt$) and (iv) vertical median every 300 m (med_z). The 1D-TAB signal energy can be

obtained as

$$E_{TAB} = \sum_{n=0}^{N-1} {TAB_n}^2 \Delta z, \tag{2}$$

where TAB_n is the nth sample of the 1D-TAB vertical profile, $N = 180$ is the number of samples of the profile and $\Delta z = 30$ m is the height increment.
Furthermore, latitude (lat), longitude (lon), ground altitude (alt), month ($month$), solar zenith angle (θ_s, [58]) and distance to the coast ($dist$) were considered input features to provide information about geographical, temporal and climatological context. Tab. 2 summarizes the features considered for the PBLH estimation model.

Feature	Symbol	Units	# of Features
Latitude	Lat	deg	1
Longitude	Lon	deg	1
Month	$month$	month	1
Distance to coast	$dist$	km	1
Ground Altitude	alt	km	1
Solar Zenith Angle	θ_s	deg	1
TAB Signal Energy	E_{TAB}	$[km^{-1}sr^{-2}]$	1
TAB Signal Kurtosis	$kurt$	$[-]$	1
TAB Signal Skewness	$skew$	$[-]$	1
TAB Vertical Median	med_z	$[km^{-1}sr^{-1}]$	16

Table 2. Summary of the features used for the estimation model.

2.5.2. Model Selection and hyperparameter Tuning.- The dataset resulting from the preparation pipeline (Sect. 2.5.1) is split into training and test subsets using an approximately 9:1 ratio. The split is performed based on locations rather than individual samples to ensure that no RS launch station appears in both the training and the test sets. The test locations are randomly selected from Europe and North America to take advantage of the higher density of RS stations in these regions. The training set is used to identify and train the best performing model, which is then evaluated in the test set.

The performance of the RF and GB models (see Sect. 2.4) is highly sensitive to their hyperparameters, which must be optimized for each specific application through a process known as hyperparameter tuning. In this study, the adjustment was carried out using a k-fold cross-validation on the training dataset. Specifically, 5-fold cross-validation (k = 5) was used, where data was divided into five subsets. In each iteration, one subset was used for validation, while the remaining four were used for training, ensuring that each portion of the data served as both training and validation data across the folds. To prevent data leakage and ensure reliable evaluation, samples from the same RS launching station were not allowed to appear in the training and validation sets. The

performance of the model for each hyperparameter configuration was evaluated using two statistical metrics: R^2 and $RMSE$ between the estimated and observed values of $PBLH$.

The hyperparameters tuned for the RF model included the number of estimators, maximum tree depth, minimum number of samples required for a leaf node, minimum number of samples required to split a node, and maximum number of features considered by each tree. In addition to these, the learning rate was also tuned for the GB model. The complete list of hyperparameters and their corresponding value ranges is provided in Table 3.

Next, from the hyperparameter tuning process, the best-performing model was selected for PBLH estimation, trained with the entire training dataset, and evaluated with the test dataset, which is data collected in unseen locations by the model.

3. Results

3.1. Data-set analysis

The data set resulting from the preparation pipeline is presented in Sect. 2.5.1 is comprised of 22,928 records, divided into training (20,636 records) and test (2,292 records) subsets using an approximate 9:1 ratio, while ensuring that the locations of the test subset are not used for training.

Fig. 5 shows the coincident points between CALIPSO overpasses and the RS launches. It can be observed that mainly Europe and North America contain a dense cloud of coincident points, whereas other locations such as South America, East Asia, and Oceania have very few or no coincident points. This is because CALIPSO follows a near-polar orbit, meaning that it overpasses approximately the same location at the same time, whereas RSs are launched globally at synoptic times of 00:00UTC and 12:00UTC, and at 06:00UTC and 18:00UTC at some stations. Consequently, in specific regions of the globe, there exists a temporal discrepancy exceeding three hours between the overpassing schedule of CALIPSO and the launching times of RS, rendering them unsuitable for comparative analysis.

We selected a random set of locations within Europe and North America (red dots in Fig. 5) as the *test set* to validate the PBLH estimation model. This is due to the fact that the model is better suited for the estimation of PBLH in locations with climatological and geographical conditions similar to the training ones (yellow dots in Fig. 5). Consequently, extreme or isolated coincidental locations are exclusively utilized for model training. The datasets include day- and night-time measurements, encompass stations operational throughout the entire year, and encompass a comprehensive range of meteorological conditions(e.g. cloud or dust scenarios), and measurement locations are in both rural and urban areas.

Fig. 6 shows the PBLH distributions observed in the training and test datasets. On initial examination, the training and test datasets exhibit comparable exponential-like

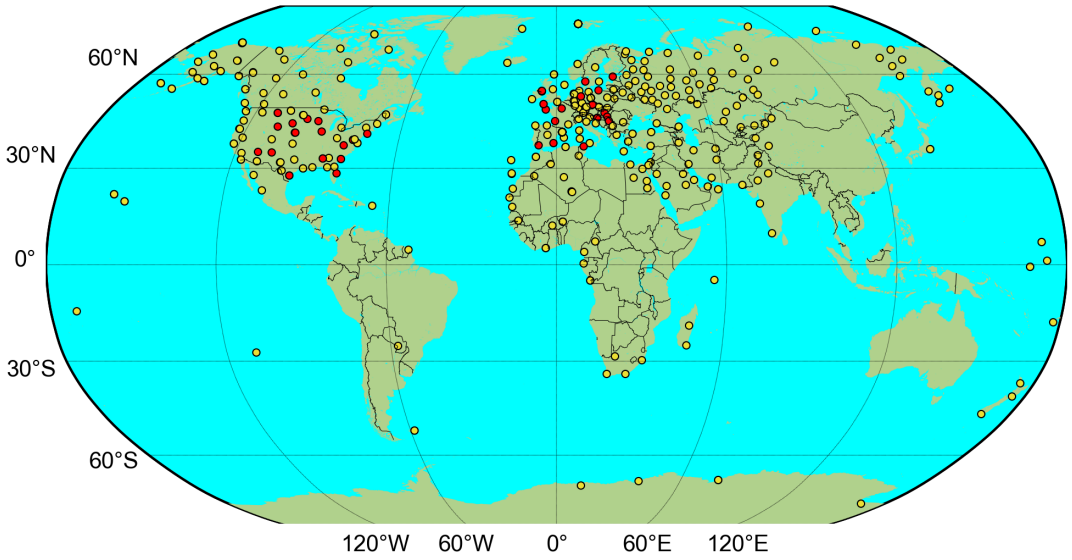


Figure 5. Collocated CALIPSO overpasses and radiosonde launch points used for PBLH estimation model training and testing. Yellow dots represent training data locations, while red dots indicate test data locations.

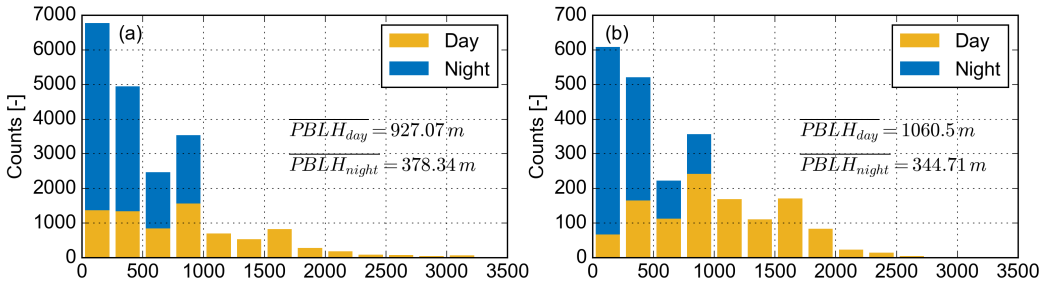


Figure 6. Histogram distribution of the RS-measured PBLH in the training (panel (a)) and test (panel (b)) datasets.

distributions of PBLH, with an underrepresentation of higher PBLH (above 1700 m). The mean values of PBLH for both daytime ($\overline{PBLH_{day}}$) and nighttime ($\overline{PBLH_{night}}$) measurements are observed to align closely between the training and test datasets. The fact that these two sets are not identical is due to the fact that the RS training set geographically includes not only the test set but also remote locations around the world. As expected, night-time PBLHs (blue bars) are lower than daytime PBLHs (yellow bars) and do not exceed 1000 m height.

3.2. hyperparameter tuning results

Table 3 shows the results of the hyperparameter adjustment and the optimal performance of the ML models considered: RF and GB. The number of tests required to evaluate all the different combinations of hyperparameters is very large (more than 10^4 tests). Therefore, we first applied a randomized search approach considering 100 random sets of hyperparameters to narrow down the best candidates. The randomized search results with all the 100 hyper-parameter sets and corresponding cross-validation indicators are given as supplementary materials. Then, we applied a grid search around the preliminary best candidates to fine-tune the hyperparameters.

Model	hyperparameter	Values Range	Best	Best Indicators ($RMSE$ in $[m]$)
RF	n. estimators	50 - 1000, steps of 10	950	$R^2 = 0.51$ $RMSE = 375.33$
	max. depth	1 - 50, steps of 2	19	
	min. leaf	1 - 20, steps of 1	6	
	min. split	1 - 20, steps of 1	2	
	max. features	0.1 - 1, steps of 0.1	0.4	
GB	num. estimators	50 - 1000, steps of 10	390	$R^2 = 0.49$ $RMSE = 378.89$
	max. depth	1 - 50, steps of 2	16	
	min. leaf	1 - 16, steps of 1	6	
	min. split	1 - 16, steps of 1	2	
	learning rate	0.01 - 0.1, steps of 0.01	0.03	
	max. features	0.1 - 1, steps of 0.1	0.4	

Table 3. ML models hyperparameter tuning results. In bold are highlighted the best indicators encountered. The number of tests corresponds to the total combinations of hyperparameters multiplied by the number of cross-validation folds per test (5 folds).

The tuned RF and GB models provide similar performance, with virtually identical indicators $R^2 \approx 0.50$ and $RMSE \approx 375$ m. However, the RF model was preferred over the GB model because of the faster training process (parallelization) and higher explainability. The chosen RF model that provided the most accurate estimation results had the following hyperparameters: 950 tree estimators, a maximum tree depth of 19, 6 minimum samples per leaf, 2 minimum samples to split a node and 40% maximum number of features used to obtain each node split (that is, 0.4 in Tab. 3).

3.3. Global Results

We trained the RF model with the optimal hyperparameters (see Sect. 3.2) using the training dataset (28,021 records) and validated on the test dataset (3,113 records). The PBLH estimations of the RF model derived from the CALIPSO data ($PBLH_{RF}$) were compared against the RS PBLH measurements ($PBLH_{RS}$) as a reference. Note that the test set locations are in regions of North America and Europe (see Sect. 3.1), and

thus, the validity of the results is limited to those. To quantitatively assess the accuracy of the model prediction, we considered the statistical indicators R^2, $RMSE$, normalized $RMSE$ ($nRMSE$) and the mean signed deviation (MSD). The $nRMSE$ allows for the comparison between datasets with different ranges and can be formulated as

$$nRMSE = \frac{RMSE}{max(PBLH_{RS}) - min(PBLH_{RS})}. \tag{3}$$

The MSD is indicative of the mean bias and is defined as

$$MSD = \frac{\sum_{n=0}^{N-1} PBLH_{RS,n} - PBLH_{RF,n}}{N}, \tag{4}$$

where $PBLH_{RF,n}$ and $PBLH_{RS,n}$ are the nth sample of the test set, and N the number of samples.

The scatter plot in Fig. 7 (panel (a)) illustrates the comparison between $PBLH_{RF}$ and $PBLH_{RS}$ for the whole test dataset. It can be observed that data samples fall along the ideal 1:1 correlation line, and, except for particular outliers, the estimates show an acceptable correspondence with the reference measurements, as evidenced by the statistical indicators $RMSE = 278.02$ m, $nRMSE = 0.11$ and $R^2 = 0.67$. It is important to note that these results were obtained with minimal data screening, i.e., only excluding opaque layer measurements and cases of technical malfunction. They therefore also include multilayer clouds and scenarios with advected aerosol layers (e.g. dust), showing the robustness of the presented methodology. Furthermore, the RF exhibits a low estimation bias, as shown by a MSD of -1.71 m (0.41% of the mean $PBLH$ measured by RS), with the mean PBLH estimated by the RF being $\overline{PBLH}_{RF} = 420.14$ m, and the mean PBLH estimated by the RS being $\overline{PBLH}_{RS} = 418.43$ m.

Regardless of its overall efficacy, the RF demonstrated specific limitations. Firstly, it exhibited an underestimation of elevated PBLH values, as indicated by the linear regression slope and intercept values of 0.65 and 147.28, respectively. This is due to the fact that PBLH values greater than 2000 m are underrepresented in the dataset (see panel (b)) [23], which challenges generalization. Secondly, although the RF is able to estimate the PBLH values at lower altitudes, high dispersion is observed, probably due to the presence of the residual layer in night-time regimes.

Regarding the estimation of outliers, a significant overestimation is visible at lower values of $PBLH_{RS}$ (the red ellipse in Fig. 7). These points are mainly made up of observations taken at noon in the summer months. Two potential causes for these discrepancies can be hypothesized. Firstly, the retrieval of $PBLH$ from CALIOP is based on the vertical distribution of aerosols, which serve as a proxy for the PBLH, indicating a possible bias relative to the retrievals from RS [59, 60, 61], which is based on a thermodynamic definition obtained by meteorological profiles. However, more significant discrepancies arise at elevated PBLH values, since the RF model relies more on contextual features (see Tab. 4) rather than on CALIOP TAB measurements to estimate the $PBLH$, indirectly using information from historical RS measurements, which results in overestimations. In other words, the model assigns

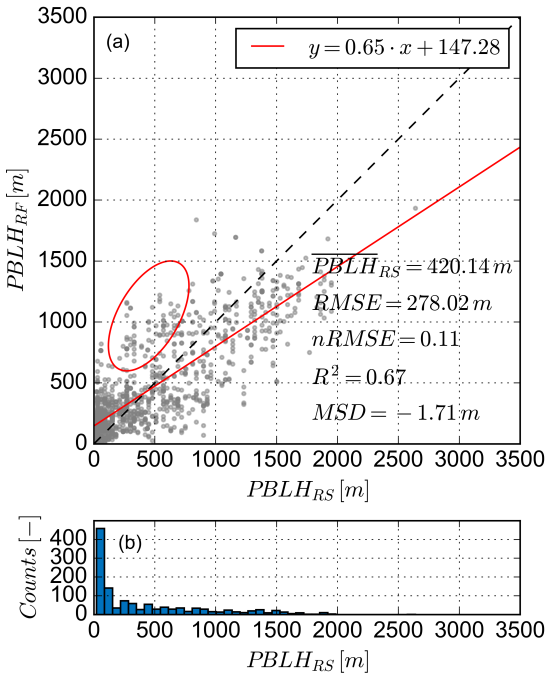


Figure 7. Global RF $PBLH$ estimation results. (a) Scatter plot comparing the PBLH retrieved from CALIPSO using RFs ($PBLH_{CALIPSO}$, y-axis) against RS retrievals of the PBLH via the bulk-Richardson number method ($PBLH_{RS}$) as a reference. (b) Histogram of the RS-measured $PBLH$ in panel (a).

a greater weight to temporal characteristics such as latitude, longitude, hour, month, and day/night indicators than to direct measurements of CALIOP TAB profiles. This misinterpretation introduces a bias, causing these contextual features to outweigh the actual profile data. However, at the same time, incorporating contextual data in the RF model allows for robust PBLH estimation in multi-layer or cloudy atmospheric conditions. Thus, there is a trade-off between feature bias and the model's ability to handle complex atmospheric scenarios.

3.4. Feature Importance and Ablation Study

We conducted an ablation study to assess the impact of contextual information on the precision of RF for the estimation $PBLH$. To this end, we trained and tested the best performing RF while selectively excluding input features from three categories: (i) geographical (latitude, longitude, distance to the coast and ground altitude), (ii) temporal (month and solar zenith angle) and (iii) statistical descriptors of TAB (energy, kurtosis and skewness).

Tab. 4 presents the results. The study shows that both geographical and temporal contexts are crucial for accurate RF performance. Excluding them significantly degrades the accuracy of the model, with R^2 dropping from 0.61 to 0.20 and $RMSE$ increasing

from 278.02 m to 431.56 m. Temporal characteristics proved to be the most critical: their exclusion resulted in the lowest R^2 (0.22) and the highest $RMSE$ (427.50 m), highlighting their role in distinguishing between day and night, as well as seasonal variations in $PBLH$ (see Sections 3.6.1 and 3.6.3). Geographical and Statistical features had similar impact on the RF performance, improving the test set indicators to $R^2 \approx 0.48$ and $RMSE \approx 345$ m.

Excluded Information	Excluded Features	Test Set Indicators
Geographical	Latitude, Longitude Distance to Coast, Ground Altitude	$R^2 = 0.48$ $RMSE = 349.21$ m
Temporal	Month, Solar Zenith Angle	$R^2 = 0.22$ $RMSE = 427.50$ m
Geographical and Temporal	Latitude, Longitude Distance to Coast, Ground Altitude Month, Solar Zenith Angle	$R^2 = 0.20$ $RMSE = 431.56$ m
Statistical	Energy, Kurtosis, Skewness	$R^2 = 0.49$ $RMSE = 343.01$ m
None	None	$R^2 = 0.67$ $RMSE = 278.02$ m

Table 4. Ablation study results. Impact of excluding feature categories on RF performance for $PBLH$ estimation.

3.5. Method Comparison with the State of the Art

Table 5 compares the performance and validation conditions of the presented method with the state of the art. The RF method of this study achieves an R^2 value of 0.67 and a $RMSE$ of 278.02 m with reference to RS in the test set, providing coverage in Europe and North America in day / night regimes and in all types of atmospheric scenarios, that is, without any type of data detection. In contrast, the traditional WCT method and its variations show R^2 values ranging from 0.26 to 0.66, and $RMSE$ values from 250 m up to 560 m, by applying multilayer and cloud scenario screening. For example, Kim et al. [24] obtained R^2 values of 0.26 (night) and 0.66 (day), with corresponding $RMSE$ values of 250 m (night) and 560 m (day), eliminating multilayer cases. Without data screening, the results are markedly worse, with $R^2 = 0.15$ (night) and 0.17 (day), and $RMSE = 740$ m (night) and 610 m (day). Similar conclusions can be drawn from the retrieval of PBLH from CALIPSO L-2 aerosol layer product by Leventidou et al. [22], which consider the base height of the L-2 aerosol layer product as the $PBLH$. Meanwhile, the novel graphic algorithm presented by Liu et al. [32] is able to obtain more accurate PBLH estimations, with $R^2 = 0.72$ and $RMSE = 340$ m for the boundary layer at night. However, it also requires filtering out multilayer and cloud scenarios.

With respect to the number of locations used for validation of the methods, note that the current studies do not utilize more than two locations for a 1-to-1 comparison

of PBLH retrievals with their corresponding reference. Moreover, most of them are limited to day- or night-time regimes. Therefore, the validity of the results achieved may be limited to those locations, providing worse estimation results under different atmospheric or climatologic regimes, for example. Although the results obtained by the RF in this study are also limited to the locations of reference RS launch stations, they span a much wider range of atmospheric and climate regimes, ensuring a much higher generalization of the results obtained.

Finally, note that most of the methods use ground-based elastic lidars as a reference. Although being a widely accepted method for PBLH retrieval, their estimations have been reported to differ from RS retrievals under certain conditions, for example, measuring the residual layer during night periods [11]. Therefore, PBLH estimations may correspond to micropulse lidar (MPL) retrievals, but not to real PBLH [11, 62, 63].

Method	R^2	RMSE [m]	Locations	Times	Reference	Screening
RF (this study)	0.67	278.02	North America and Europe	Day/Night	RS	No
WCT (Kim et al.[24])	0.26 - 0.66	250 - 560	Single Location	Day/Night	MPL	Multilayer and Cloud
WCT (Su et al. [31])	0.31	277	Single Location	Day	MPL, RS	Multilayer and Cloud
MV + WCT (Zhang et al.[23])	0.35 - 0.42	-	Two Locations	Day	MPL	Multilayer and Cloud
Graphics Algorithm (Liu et al. [32])	0.72	340 m	Single Location	Night	MPL	Multilayer and Cloud
Base height of CALIPSO L-2 aerosol layer product [22]	0.28	-	Single Location	Day	MPL	Dust Scenarios

Table 5. Comparison of the state-of-the-art methods for PBLH estimation from CALIOP measurements.

A notable limitation of our approach is the underrepresentation of extreme PBLH values, particularly those above 2000 m, which affects the model accuracy in these regimes. This is reflected in the slope and intercept of the regression in Figure 7. Mitigation strategies include applying stratified sampling, quantile-aware loss functions, or upweighting extreme cases during training. Incorporating such methods could enhance the capacity of the RF model to resolve edge-case atmospheric scenarios.

3.6. Cluster Results

In this section, we evaluate the performance metrics and the complete effectiveness of the RF model in retrieving the height of the PBL under different conditions. The analysis focuses on evaluating the effectiveness and reliability of the model in different seasons, regimes, and geographical characteristics. The investigation rigorously assesses the complex interactions between seasonal variations in atmospheric conditions and distinct regional differences, including land use practices and topographical features, which can profoundly affect the precision and reliability of the model's outputs and predictions. Through a comprehensive analysis of these multifaceted factors, we aim

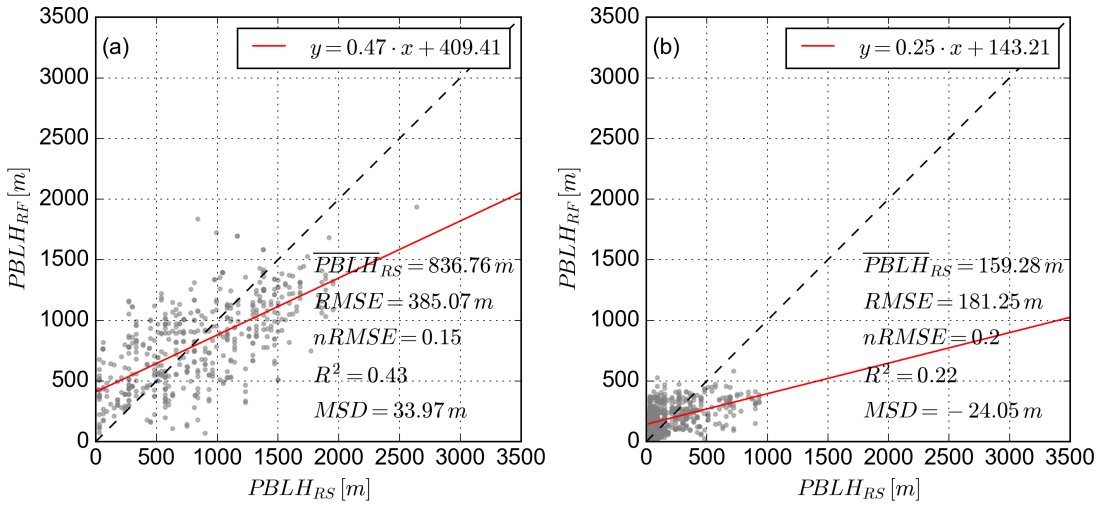


Figure 8. The same as Fig. 7 but clustered as a function of daytime (a) and nighttime (b) regimes.

to elucidate meaningful insights into the model's inherent strengths and potential limitations, ultimately contributing to the advancement of methodologies and techniques for retrieving PBL height across diverse environmental contexts and conditions. Toward this purpose, the prediction results in the test set have been clustered as a function of the following binning variables: *(i) day/night regimes, (ii) Season, and (iii) Rural/Urban location.*

3.6.1. Day/Night Clustering Results Fig. 8 compares the performance of RF PBLH estimation in the day and night regimes. At first glance, the daytime convective PBLH (Fig. 8, panel (a)) exhibits significantly higher values and a larger dispersion compared to the nighttime stable PBLH (panel (b)). The mean PBLH observed by RS ($\overline{PBLH_{RS}}$) further evidences this difference, with daytime $\overline{PBLH_{RS}}$ being 836.76 m and nighttime 159.28 m. On the one hand, the PBLH measured in the daytime RS ranges from virtually 0 m to altitudes higher than 2500 m. This is due to the fact that a wide range of atmospheric conditions (e.g. clouds, rain), latitudes (from 30° up to 60°), seasons and land uses (urban, rural, industrial, etc.) that influence convective PBLH are included in this cluster. However, night-time PBLH is less affected by these characteristics and therefore ranges between 0 and 1000 m.

Regarding the performance of the RF estimation, better results are observed for daytime conditions, mainly due to the stronger aerosol concentration gradient at the $PBLH$, which the RF is able to track. This is reflected in the statistical indicators: during the day, $R^2 = 0.43$ and $RMSE = 385.07$ m ($nRMSE = 0.15$); at night, $R^2 = 0.22$ and $RMSE = 181.25$ m ($nRMSE = 0.20$). The poorer performance during night regimes is probably due to the fact that the residual layer creates ambiguities when estimating the shallower stable PBL. In terms of estimation bias, thanks to its

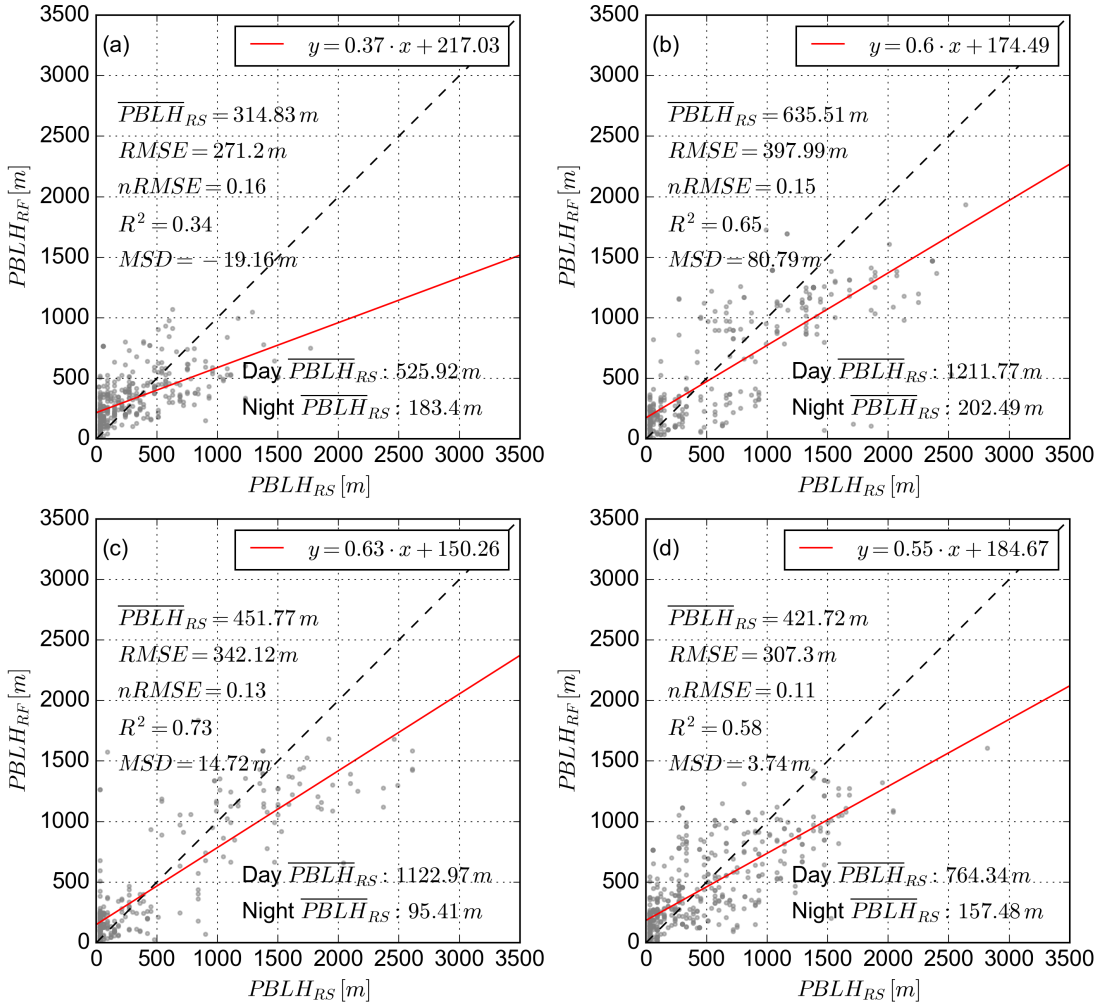


Figure 9. Same as Fig. 7 but clustered as a function of the season: (a) winter, (b) spring, (c) summer, and (d) autumn.

inherent structure, the RF is able to provide low MSD figures, that is, $MSD = 33.97$ m and $MSD = -24.05$ m in the daytime and nighttime regimes, respectively.

3.6.2. Seasonal Clustering Results As shown in Fig. 9, seasonality has a clear influence on PBLH variability. Although site-specific characteristics can be observed, particularly in Mediterranean climates due to the diversity of atmospheric flow types [64, 65], overall seasonal patterns are consistent with previous studies in Europe and the US using the Richardson-number approach [66]. During the warmer seasons, spring and summer (panels (b) and (c), respectively), the daytime PBLH values exhibit greater variability, with mean heights of 1211.77 m and 1122.97 m and peak values reaching 2640 m and 2610 m, respectively. This increase is primarily driven by enhanced solar radiation and stronger convective activity. In contrast, during the colder seasons, winter and autumn

(panels (a) and (d)), the dynamic range of daytime PBLH is considerably reduced. The mean heights at the day drop to 525.92 m and 764.34 m, with maximum values of 1770 m and 2070 m, respectively, reflecting weaker surface heating and more limited convection. At night, however, this trend reverses: winter exhibits the second highest nighttime PBLH values after spring, with a mean of 183.40 m, probably due to stronger climatological wind speeds typical of the colder months [66].

Seasonal variations also play a role in the performance of RF, with a larger relative dispersion observed during winter (panel (a)). This is possibly related to the lower vertical mixing of aerosols in the colder seasons, producing a weaker TAB gradient at the PBLH which the RF struggles to identify. In terms of estimation bias, the RF exhibits consistent performance in all seasonal groups, with $|MSD|$ lower than 20 m in all seasons except spring (panel (b)). These findings underscore the need for careful consideration of seasonal factors when using CALIOP-derived PBLH and highlight the importance of further refinement in retrieval algorithms to minimize biases and improve accuracy.

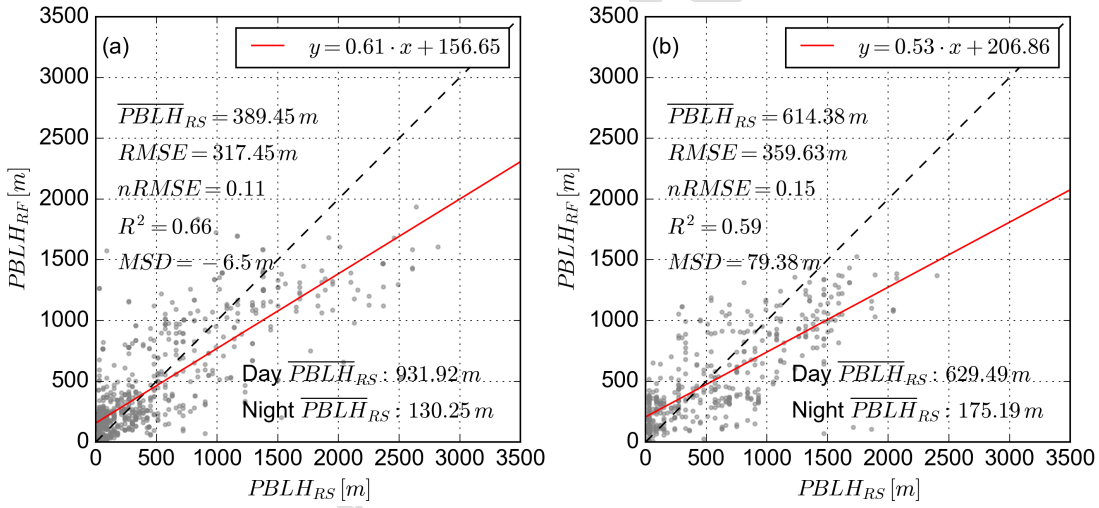


Figure 10. Same as Fig. 7 but clustered as a function of the population density: (a) urban scenario (density higher than 100 people/km^2), (b) non-urban scenario (density lower than 100 people/km^2).

3.6.3. Land Use Clustering Results Finally, we clustered the test dataset as a function of the population density in order to observe how the land use influences the PBLH and the RF performance. Towards this purpose, we used the *Gridded Population of the World* dataset [67], providing world-wide population density values for the year 2020 at a resolution of 0.1°×0.1° (≈10km×10km). We considered as an urban area coordinates with population density higher than 100 people/km^2, and rural otherwise.

Fig. 10 shows the results of the land use clustering. Surprisingly, higher mean PBLH values are observed in nonurban areas. Highly dense urban environments

typically contribute to higher atmospheric convection due to surface radiating properties, buildings' effects on atmospheric dynamics, and human activity, and thus leading to higher PBLH, whereas non-urban environments mitigate this effect thanks to vegetation, higher albedo, and soil moisture, among others. In the test dataset, 79. 62% of the nonurban measurements were collected during the day, compared to only 24.43% in urban areas. This temporal sampling bias contributes to the elevated mean PBLH observed in nonurban clusters. When restricting the analysis to daytime measurements only, the mean PBLH in urban areas ($\overline{PBLH}_{RS} = 931.92\,m$) is 302.43 m higher than in rural areas ($\overline{PBLH}_{RS} = 629.49\,m$).

Regarding RF, its performance is better in urban environments, with $R^2 = 0.66$ and $RMSE = 317.45$ m ($nRMSE = 0.11$), compared to $R^2 = 0.59$ and $RMSE = 359.63$ m ($nRMSE = 0.15$) in non-urban areas. Although further research for improvement in the latter is required, an adequate estimate of PBLH in urban areas is of crucial importance in terms of pollution management and early warning, highlighting the potential impact of the RF estimation method.

3.7. Temporal Series

Fig. 11 shows an example time series of the PBLH estimated by the RF overlaid on the CALIOP-measured TAB. In this case, CALIOP follows a north-to-south orbit, overpassing Europe around local noon in summer. At first glance, the TAB structures are well captured by the RF, which reproduces a consistent and smooth transition from low $PBLH$ values at high latitudes to steadily increasing values toward the south. This is particularly clear between 43.2° and 33.5° N, where the RF successfully tracks the aerosol layer at altitudes between 1500 m and 2500 m—values consistent with typical summer midday $PBLH$ in southern Europe, when the daytime boundary layer is still developing at night.

Between 54.2° and 47.2° N (covering Germany and the eastern edge of the Alps), the RF provides meaningful $PBLH$ estimates even under cloudy conditions, benefiting from contextual predictors such as latitude, longitude, surface elevation, and solar zenith angle. At latitudes between 63° and 58° N, the RF estimates a $PBLH$ of approximately 500 m, which aligns with expected values for Scandinavian coastal climates in mid-June [68]. Finally, at latitudes above 63° N (that is, over the Arctic Ocean), RF performance degrades, reflecting the scarcity of training data available for those regions.

4. Discussion and Conclusions

This study introduces a novel approach to the estimation of PBLH by integrating a decade-long, globally sourced RS-derived dataset with raw CALIOP Level 1 profiles, eliminating dependence on higher-level products, aerosol classification or atmospheric typing. The RF model was trained on 28,021 global records and validated on 3,113 records from Europe and North America, achieving high precision ($R^2 = 0.67$, $RMSE =$

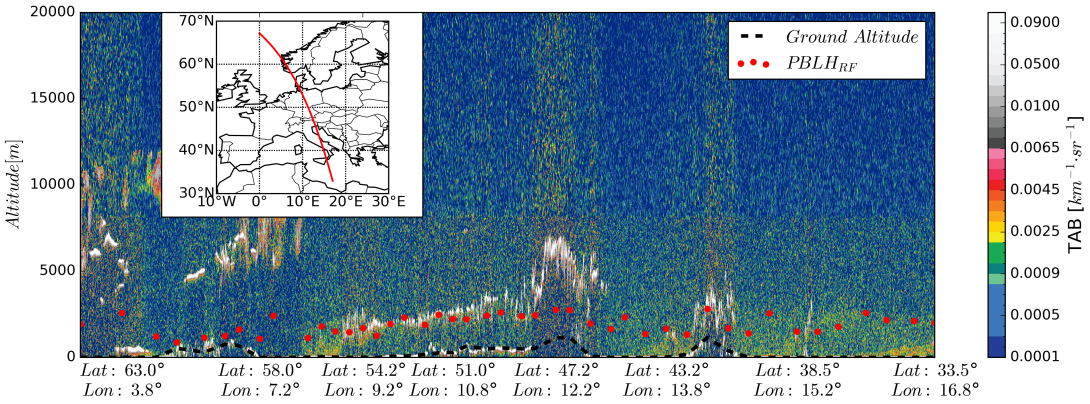


Figure 11. Temporal series of RF-estimated PBLH ($PBLH_{RF}$) overlaid on time–altitude CALIOP-measured TAB profiles across Europe on 27 June 2018 at 11:48:11 UTC. The inset map shows the CALIPSO orbit track during the measurements.

278.02 m) and demonstrating resilience in urban and industrial environments, as well as in both day and night conditions. This performance exceeds that of previous techniques (see Tab. 5) while supporting broad spatial coverage with minimal data filtering or atmospheric screening, i.e., only technical malfunctioning or opaque layer cases, thus enhancing operational feasibility for large-scale applications such as monitoring air quality, adaptation to urban climate, and prediction of pollution. However, systematic estimation errors were observed—specifically, underestimating high PBLH values and miss-estimating low ones—due to four main factors: differences between aerosol-based CALIOP measurements and thermodynamically-based RS measurements, underrepresentation of extreme PBLH cases in the training set, residual layer effect on night-time TAB measurements, and biases linked to features such as latitude, altitude, and solar zenith angle. These limitations also affected the predictions under specific conditions, such as noon in the summer, emphasizing the need for feature refinement and expanded data coverage. The applicability of the model remains geographically limited to regions with adequate RS–CALIOP overlap, and further validation is necessary in tropical, arid, and polar zones. Future work will explore deep learning architectures to better model spatial dependencies in CALIOP profiles, address feature-related biases, and integrate hybrid physical–data-driven methods to improve robustness. Furthermore, ground-based lidar-derived PBLH estimates from large-scale networks such as EARLINET in Europe warrant comparison with the retrieval method proposed in this study. In general, this research advances the state-of-the-art in satellite-based PBLH retrieval and offers a scalable and accurate alternative to traditional methods, with important implications for environmental monitoring, climate resilience planning, and evidence-based policy design.

Data and Code Availability Statement

The RS, CALIOP, training, and test datasets are publicly available at `https://zenodo.org/records/15639750` [69]. The code repository can be accessed at `https://github.com/andreukalu/pblh_ML.git` [70].

Funding

The authors acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 1409 published on 14.9.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU – Project Title PBLhsat CUP P20224AT3W Grant Assignment Decree No. 965 adopted on 30 June 2023 by the Italian Ministry of Ministry of University and Research (MUR).

[1] R. B. Stull, *An introduction to boundary layer meteorology*. The Netherlands: Kluwer Academic Publishers, 1988.

[2] S. K. Nair, M. Madhusoodanan, and R. Mehajan, "The role of boundary layer height (blh) variations on pollution dispersion over a coastal station in the southwest peninsular india," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 179, pp. 273–280, 2018. [Online]. Available: https://doi.org/10.1016/j.jastp.2018.07.011

[3] T. Su, Z. Li, and R. Kahn, "Relationships between the planetary boundary layer height and surface pollutants derived from lidar observations over china: regional pattern and influencing factors," *Atmospheric Chemistry and Physics*, vol. 18, no. 21, pp. 15 921–15 935, 2018. [Online]. Available: https://doi.org/10.5194/acp-18-15921-2018

[4] S. Lolli, "Urban pm2.5 concentration monitoring: A review of recent advances in ground-based, satellite, model, and machine learning integration," *Urban Climate*, vol. 63, p. 102566, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212095525002822

[5] Y. Miao, J. Guo, S. Liu, H. Liu, Z. Li, W. Zhang, and P. Zhai, "Classification of summertime synoptic patterns in beijing and their associations with boundary layer structure affecting aerosol pollution," *Atmospheric Chemistry and Physics*, vol. 17, no. 4, pp. 3097–3110, 2017. [Online]. Available: https://doi.org/10.5194/acp-17-3097-2017

[6] S. Lolli, W. Y. Khor, M. M. Z. Matjafri, and H. S. Lim, "Monsoon season quantitative assessment of biomass burning clear-sky aerosol radiative effect at surface by ground-based lidar observations in pulau pinang, malaysia in 2014," *Remote Sensing*, vol. 11, no. 22, 2019. [Online]. Available: https://doi.org/10.3390/rs11222660

[7] W. Jia and X. Zhang, "The role of the planetary boundary layer parameterization schemes on the meteorological and aerosol pollution simulations: A review," *Atmospheric Research*, vol. 239, p. 104890, 2020. [Online]. Available: https://doi.org/10.1016/j.atmosres.2020.104890

[8] S. Lolli, "Is the air too polluted for outdoor activities? check by using your photovoltaic system as an air-quality monitoring device," *Sensors*, vol. 21, no. 19, 2021. [Online]. Available: https://doi.org/10.3390/s21196342

[9] J. M. Edwards, A. C. M. Beljaars, A. A. M. Holtslag, and A. P. Lock, "Representation of boundary-layer processes in numerical weather prediction and climate models," *Boundary-Layer Meteorology*, vol. 177, no. 2, pp. 511–539, Dec 2020. [Online]. Available: https://doi.org/10.1007/s10546-020-00530-z

[10] A. von Engeln and J. Teixeira, "A planetary boundary layer height climatology derived from ecmwf reanalysis data," *Journal of Climate*, vol. 26, no. 17, pp. 6575 – 6590, 2013. [Online]. Available: https://doi.org/10.1175/JCLI-D-12-00385.1

[11] M. P. Araújo da Silva, F. Rocadenbosch, R. L. Tanamachi, and U. Saeed, "Motivating a synergistic

mixing-layer height retrieval method using backscatter lidar returns and microwave-radiometer temperature observations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022. [Online]. Available: https://doi.org/10.1109/TGRS.2022.3158401

[12] R. H. Johnson, P. E. Ciesielski, and J. A. Cotturone, "Multiscale variability of the atmospheric mixed layer over the western pacific warm pool," *Journal of the Atmospheric Sciences*, vol. 58, no. 18, pp. 2729 – 2750, 2001. [Online]. Available: https://doi.org/10.1175/1520-0469(2001)058⟨2729:MVOTAM⟩2.0.CO;2

[13] C. Sivaraman, S. McFarlane, E. Chapman, M. Jensen, Toto, S. Liu, and M. Fischer, "Planetary boundary layer height (pbl) value added product (vap): Radiosonde retrievals," DOE Office of Science Atmospheric Radiation Measurement (ARM) Program (United States), Tech. Rep., 08 2013. [Online]. Available: https://www.osti.gov/biblio/1808688

[14] M. Haeffelin, F. Angelini, Y. Morille, G. Martucci, S. Frey, G. Gobbi, S. Lolli, C. O'Dowd, L. Sauvage, I. Xueref-Rémy *et al.*, "Evaluation of mixing-height retrievals from automatic profiling lidars and ceilometers in view of future integrated networks in europe," *Boundary-Layer Meteorology*, vol. 143, no. 1, pp. 49–75, 2012. [Online]. Available: https://doi.org/10.1007/s10546-011-9643-z

[15] G. de Arruda Moreira, J. L. Guerrero-Rascado, J. A. Bravo-Aranda, I. Foyo-Moreno, A. Cazorla, I. Alados, H. Lyamani, E. Landulfo, and L. Alados-Arboledas, "Study of the planetary boundary layer height in an urban environment using a combination of microwave radiometer and ceilometer," *Atmospheric Research*, vol. 240, p. 104932, 2020. [Online]. Available: https://doi.org/10.1016/j.atmosres.2020.104932

[16] J. Lee, J.-W. Hong, K. Lee, J. Hong, E. Velasco, Y. J. Lim, J. B. Lee, K. Nam, and J. Park, "Ceilometer monitoring of boundary-layer height and its application in evaluating the dilution effect on air pollution," *Boundary-Layer Meteorology*, vol. 172, no. 3, pp. 435–455, Sep 2019. [Online]. Available: https://doi.org/10.1007/s10546-019-00452-5

[17] D. Lange, F. Rocadenbosch, J. Tiana-Alsina, and S. Frasier, "Atmospheric boundary layer height estimation using a kalman filter and a frequency-modulated continuous-wave radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3338–3349, 2015. [Online]. Available: https://doi.org/10.1109/TGRS.2014.2374233

[18] G. Basha and M. V. Ratnam, "Identification of atmospheric boundary layer height over a tropical station using high-resolution radiosonde refractivity profiles: Comparison with gps radio occultation measurements," *Journal of Geophysical Research: Atmospheres*, vol. 114, no. D16, 2009. [Online]. Available: https://doi.org/10.1029/2008JD011692

[19] J. P. A. Martins, J. Teixeira, P. M. M. Soares, P. M. A. Miranda, B. H. Kahn, V. T. Dang, F. W. Irion, E. J. Fetzer, and E. Fishbein, "Infrared sounding of the trade-wind boundary layer: Airs and the rico experiment," *Geophysical Research Letters*, vol. 37, no. 24, 2010. [Online]. Available: https://doi.org/10.1029/2010GL045902

[20] R. Wood and C. S. Bretherton, "Boundary layer depth, entrainment, and decoupling in the cloud-capped subtropical and tropical marine boundary layer," *Journal of Climate*, vol. 17, no. 18, pp. 3576 – 3588, 2004. [Online]. Available: https://doi.org/10.1175/1520-0442(2004)017⟨3576:BLDEAD⟩2.0.CO;2

[21] D. M. Winker, M. A. Vaughan, A. Omar, Y. Hu, K. A. Powell, Z. Liu, W. H. Hunt, and S. A. Young, "Overview of the calipso mission and caliop data processing algorithms," *Journal of Atmospheric and Oceanic Technology*, vol. 26, no. 11, pp. 2310–2323, 2009. [Online]. Available: https://doi.org/10.1175/2009JTECHA1281.1.

[22] E. Leventidou, P. Zanis, D. Balis, E. Giannakaki, I. Pytharoulis, and V. Amiridis, "Factors affecting the comparisons of planetary boundary layer height retrievals from calipso, ecmwf and radiosondes over thessaloniki, greece," *Atmospheric Environment*, vol. 74, pp. 360–366, 2013. [Online]. Available: https://doi.org/10.1016/j.atmosenv.2013.04.007

[23] W. Zhang, J. Guo, Y. Miao, H. Liu, Y. Zhang, Z. Li, and P. Zhai, "Planetary boundary layer height from caliop compared to radiosonde over china," *Atmospheric Chemistry and Physics*, vol. 16,

no. 15, pp. 9951–9963, 2016. [Online]. Available: https://doi.org/10.5194/acp-16-9951-2016

[24] M.-H. Kim, H. Yeo, S. Park, D.-H. Park, A. Omar, T. Nishizawa, A. Shimizu, and S.-W. Kim, "Assessing caliop-derived planetary boundary layer height using ground-based lidar," *Remote Sensing*, vol. 13, no. 8, 2021. [Online]. Available: https://doi.org/10.3390/rs13081496

[25] NASA/LARC/SD/ASDC, "Calipso lidar level 1b profile data, v4-51." [Online]. Available: https://doi.org/10.5067/CALIOP/CALIPSO/CAL_LID_L1-Standard-V4-51

[26] ——, "Calipso lidar level 2 5 km cloud layer, v4-51." [Online]. Available: https://doi.org/10.5067/CALIOP/CALIPSO/CAL_LID_L2_05kmCLay-Standard-V4-51

[27] ——, "Calipso lidar level 2 aerosol profile, v4-51." [Online]. Available: https://doi.org/10.5067/CALIOP/CALIPSO/CAL_LID_L2_05kmAPro-Standard-V4-51

[28] X. Lu, Y. Hu, M. Vaughan, S. Rodier, C. Trepte, P. Lucker, and A. Omar, "New attenuated backscatter profile by removing the caliop receiver's transient response," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 255, p. 107244, 2020. [Online]. Available: https://doi.org/10.1016/j.jqsrt.2020.107244

[29] W. P. Hooper and E. W. Eloranta, "Lidar measurements of wind in the planetary boundary layer: the method, accuracy and results from joint measurements with radiosonde and kytoon," *Journal of climate and applied meteorology*, pp. 990–1001, 1986. [Online]. Available: https://doi.org/10.1175/1520-0450(1986)025⟨0990:LMOWIT⟩2.0.CO;2

[30] N. Gamage and C. Hagelberg, "Detection and analysis of microfronts and associated coherent events using localized transforms," *Journal of Atmospheric Sciences*, vol. 50, no. 5, pp. 750–756, 1993. [Online]. Available: https://doi.org/10.1175/1520-0469(1993)050⟨0750:DAAOMA⟩2.0.CO;2

[31] T. Su, J. Li, C. LI, P. Xiang, A. Lau, J. Guo, D. Yang, and Y. Miao, "An intercomparison of long-term planetary boundary layer heights retrieved from calipso, ground-based lidar and radiosonde measurements over hong kong," *Journal of Geophysical Research: Atmospheres*, vol. 122, 03 2017. [Online]. Available: https://doi.org/10.1002/2016JD025937

[32] B. Liu, Y. Ma, J. Liu, W. Gong, W. Wang, and M. Zhang, "Graphics algorithm for deriving atmospheric boundary layer heights from calipso data," *Atmospheric Measurement Techniques*, vol. 11, no. 9, pp. 5075–5085, 2018. [Online]. Available: https://doi.org/10.5194/amt-11-5075-2018

[33] Y. Miao, S. Liu, J. Guo, S. Huang, Y. Yan, and M. Lou, "Unraveling the relationships between boundary layer height and pm2. 5 pollution in china based on four-year radiosonde measurements," *Environmental pollution*, vol. 243, pp. 1186–1195, 2018. [Online]. Available: https://doi.org/10.1016/j.envpol.2018.09.070

[34] M. Wang, G. Tang, Y. Liu, M. Ma, M. Yu, B. Hu, Y. Zhang, Y. Wang, and Y. Wang, "The difference in the boundary layer height between urban and suburban areas in beijing and its implications for air pollution," *Atmospheric Environment*, vol. 260, p. 118552, 2021. [Online]. Available: https://doi.org/10.1016/j.atmosenv.2021.118552

[35] A. Salcedo-Bosch, L. Zong, J. B. Cohen, Y. Yang, and S. Lolli, "Comparative analysis of precipitation scavenging effects on aerosol loading in Shanghai: seasonal variability and implications for air quality management," in *Remote Sensing of Clouds and the Atmosphere XXIX*, E. I. Kassianov and S. Lolli, Eds., vol. 13193, International Society for Optics and Photonics. SPIE, 2024, p. 1319304. [Online]. Available: https://doi.org/10.1117/12.3031200

[36] A. Salcedo-Bosch, L. Zong, Y. Yang, J. B. Cohen, and S. Lolli, "Forecasting particulate matter concentration in shanghai using a small-scale long-term dataset," *Environmental Sciences Europe*, vol. 37, no. 1, p. 47, 2025. [Online]. Available: https://doi.org/10.1186/s12302-025-01068-y

[37] M. Gonçalves, P. Jiménez-Guerrero, E. López, and J. M. Baldasano, "Air quality models sensitivity to on-road traffic speed representation: effects on air quality of 80 km h- 1 speed limit in the barcelona metropolitan area," *Atmospheric Environment*, vol. 42, no. 36, pp. 8389–8402, 2008. [Online]. Available: https://doi.org/10.1016/j.atmosenv.2008.08.022

[38] F. J. Kelly, G. W. Fuller, H. A. Walton, and J. C. Fussell, "Monitoring air pollution: Use of early warning systems for public health," *Respirology*, vol. 17, no. 1, pp. 7–19, 2012. [Online]. Available: https://doi.org/10.1111/j.1440-1843.2011.02065.x

[39] Y. Huang, B. Guo, H. Sun, H. Liu, and S. X. Chen, "Relative importance of meteorological variables on air quality and role of boundary layer height," *Atmospheric Environment*, vol. 267, p. 118737, 2021. [Online]. Available: https://doi.org/10.1016/j.atmosenv.2021.118737

[40] B. M. Braun, T. H. Sweetser, C. Graham, and J. Bartsch, "Cloudsat's a-train exit and the formation of the c-train: An orbital dynamics perspective," in *2019 IEEE Aerospace Conference*, 2019, pp. 1–10. [Online]. Available: https://doi.org/10.1109/AERO.2019.8741958

[41] A. Z. Kotarba, "Impact of the revisit frequency on cloud climatology for calipso, earthcare, aeolus, and icesat-2 satellite lidar missions," *Atmospheric Measurement Techniques*, vol. 15, no. 14, pp. 4307–4322, 2022. [Online]. Available: https://doi.org/10.5194/amt-15-4307-2022

[42] W. H. Hunt, D. M. Winker, M. A. Vaughan, K. A. Powell, P. L. Lucker, and C. Weimer, "Calipso lidar description and performance assessment," *Journal of Atmospheric and Oceanic Technology*, vol. 26, no. 7, pp. 1214–1228, 2009. [Online]. Available: https://doi.org/10.1175/2009JTECHA1223.1

[43] N. Aeronautics and S. Administration, "Your gateway to nasa earth observation data — nasa earthdata," 1 2025, [Online; accessed 2025-02-06]. [Online]. Available: https://www.earthdata.nasa.gov/

[44] "Calipso - lidar level 2 blowing snow - antarctica v1.00," [Online; accessed 2025-02-06]. [Online]. Available: https://www-calipso.larc.nasa.gov/resources/calipso_users_guide/data_desc/cal_lid_l2_blowingsnow_antarctica_v2-00_desc.php

[45] J. L. Tackett, R. A. Ryan, A. E. Garnier, J. Kar, B. J. Getzewich, X. Cai, M. A. Vaughan, C. R. Trepte, R. C. Verhappen, D. M. Winker, and K.-P. A. Lee, "Mitigating impacts of low energy laser pulses on caliop data products," *EGUsphere*, vol. 2025, pp. 1–36, 2025. [Online]. Available: https://doi.org/10.5194/egusphere-2025-2376

[46] D. of Signal Theory and U. P. d. C. U. Communications (TSC), "Calcula computing services — signal theory and communications department — upc. universitat politècnica de catalunya," [Online; accessed 2025-02-06]. [Online]. Available: https://tsc.upc.edu/en/it-services/computing-services/

[47] P. Seibert, F. Beyrich, S.-E. Gryning, S. Joffre, A. Rasmussen, and P. Tercier, "Review and intercomparison of operational methods for the determination of the mixing height," *Atmospheric Environment*, vol. 34, no. 7, pp. 1001–1027, 2000. [Online]. Available: https://doi.org/10.1016/S1352-2310(99)00349-0

[48] J. Guo, Y. Miao, Y. Zhang, H. Liu, Z. Li, W. Zhang, J. He, M. Lou, Y. Yan, L. Bian, and P. Zhai, "The climatology of planetary boundary layer height in china derived from radiosonde and reanalysis data," *Atmospheric Chemistry and Physics*, vol. 16, no. 20, pp. 13 309–13 319, 2016. [Online]. Available: https://doi.org/10.5194/acp-16-13309-2016

[49] J. Guo, J. Zhang, J. Shao, T. Chen, K. Bai, Y. Sun, N. Li, J. Wu, R. Li, J. Li, Q. Guo, J. B. Cohen, P. Zhai, X. Xu, and F. Hu, "A merged continental planetary boundary layer height dataset based on high-resolution radiosonde measurements, era5 reanalysis, and gldas," *Earth System Science Data*, vol. 16, no. 1, pp. 1–14, 2024. [Online]. Available: https://doi.org/10.5194/essd-16-1-2024

[50] Y. Zhang, Z. Gao, D. Li, Y. Li, N. Zhang, X. Zhao, and J. Chen, "On the computation of planetary boundary-layer height using the bulk richardson number method," *Geoscientific Model Development*, vol. 7, no. 6, pp. 2599–2611, 2014. [Online]. Available: https://doi.org/10.5194/gmd-7-2599-2014

[51] S. Lolli, "Machine learning techniques for vertical lidar-based detection, characterization, and classification of aerosols and clouds: A comprehensive survey," *Remote Sensing*, vol. 15, no. 17, p. 4318, 2023. [Online]. Available: https://doi.org/10.3390/rs15174318

[52] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*,

1st ed. Chapman and Hall/CRC, 1984.

[53] "Randomforestregressor — scikit-learn 1.6.1 documentation," [Online; accessed 2025-02-06]. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[54] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, 2013. [Online]. Available: https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2013.00021

[55] R. Guzman, H. Chepfer, V. Noel, T. Vaillant de Guélis, J. E. Kay, P. Raberanto, G. Cesana, M. A. Vaughan, and D. M. Winker, "Direct atmosphere opacity observations from calipso provide new constraints on cloud-radiation interactions," *Journal of Geophysical Research: Atmospheres*, vol. 122, no. 2, pp. 1066–1085, 2017. [Online]. Available: https://doi.org/10.1002/2016JD025946

[56] J. Tackett, R. Ryan, M. Vaughan, A. Garnier, B. Getzewich, D. Winker, and C. Trepte, *Mitigation Strategy for the Impact of Low Energy Laser Pulses in CALIOP Calibration and Level 2 Retrievals*, 12 2023, pp. 617–623. [Online]. Available: https://doi.org/10.1007/978-3-031-37818-8_79

[57] S. Dhanani and M. Parker, "4 - introduction to digital filtering," in *Digital Video Processing for Engineers*, S. Dhanani and M. Parker, Eds. Oxford: Newnes, 2013, pp. 19–28. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780124157606000040

[58] M. Z. Jacobson, *Fundamentals of Atmospheric Modeling*, 2nd ed. Cambridge University Press, 2005.

[59] G. Vivone, G. D'Amico, D. Summa, S. Lolli, A. Amodeo, D. Bortoli, and G. Pappalardo, "Atmospheric boundary layer height estimation from aerosol lidar: a new approach based on morphological image processing techniques," *Atmospheric Chemistry and Physics*, vol. 21, no. 6, pp. 4249–4265, 2021. [Online]. Available: https://doi.org/10.5194/acp-21-4249-2021

[60] N. Roldán-Henao, T. Su, and Z. Li, "Refining planetary boundary layer height retrievals from micropulse-lidar at multiple arm sites around the world," *Journal of Geophysical Research: Atmospheres*, vol. 129, no. 13, p. e2023JD040207, 2024. [Online]. Available: https://doi.org/10.1029/2023JD040207

[61] T. Zhong, N. Wang, X. Shen, D. Xiao, Z. Xiang, and D. Liu, "Determination of planetary boundary layer height with lidar signals using maximum limited height initialization and range restriction (mlhi-rr)," *Remote Sensing*, vol. 12, no. 14, 2020. [Online]. Available: https://doi.org/10.3390/rs12142272

[62] U. Saeed, J. Tiana, S. Crewell, and F. Rocadenbosch, *Atmospheric boundary layer height estimation using combined microwave radiometer and lidar data*, 2014, p. 799. [Online]. Available: http://doi.org/10.1007/s10546-020-00514-z

[63] P. Kokkalis, D. Alexiou, A. Papayannis, F. Rocadenbosch, O. Soupiona, P. Raptis, M. Mylonaki, C. Tzanis, and J. Christodoulakis, "Application and testing of the extended-kalman-filtering technique for determining the planetary boundary-layer height over athens, greece," *Boundary-layer meteorology*, vol. 176, p. 125–147, Jan 2020. [Online]. Available: https://doi.org/10.1007/s10546-020-00514-z

[64] R. F. Banks, J. Tiana-Alsina, F. Rocadenbosch, and J. M. Baldasano, "Performance evaluation of the boundary-layer height from lidar and the weather research and forecasting model at an urban coastal site in the north-east iberian peninsula," *Boundary-Layer Meteorology*, vol. 157, no. 2, pp. 265–292, Nov 2015. [Online]. Available: https://doi.org/10.1007/s10546-015-0056-2

[65] R. F. Banks, J. Tiana-Alsina, J. M. Baldasano, F. Rocadenbosch, A. Papayannis, S. Solomos, and C. G. Tzanis, "Sensitivity of boundary-layer variables to pbl schemes in the wrf model based on surface meteorological observations, lidar, and radiosondes during the hygra-cd campaign," *Atmospheric Research*, vol. 176-177, pp. 185–201, 2016. [Online]. Available: https://doi.org/10.1016/j.atmosres.2016.02.024

[66] D. J. Seidel, Y. Zhang, A. Beljaars, J.-C. Golaz, A. R. Jacobson, and B. Medeiros, "Climatology of the planetary boundary layer over the continental united states and europe," *Journal of Geophysical Research: Atmospheres*, vol. 117, no. D17, 2012. [Online]. Available:

https://doi.org/10.1029/2012JD018143

[67] C. for International Earth Science Information Network-CIESIN-Columbia University, "Gridded population of the world, version 4 (gpwv4): Administrative unit center points with population estimates, revision 11 (version 4.11)," Palisades, NY, 2018, [Data set]. [Online]. Available: https://doi.org/10.7927/H4BC3WMT

[68] V. A. Sinclair, J. Ritvanen, G. Urbancic, I. Erner, Y. Batrak, D. Moisseev, and M. Kurppa, "Boundary-layer height and surface stability at hyytiälä, finland, in era5 and observations," *Atmospheric Measurement Techniques*, vol. 15, no. 10, pp. 3075–3103, 2022. [Online]. Available: https://doi.org/10.5194/amt-15-3075-2022

[69] A. Salcedo-Bosch, "Planetary boundary layer height 10-years long (2010–2020) global dataset from radiosonde," https://doi.org/10.5281/zenodo.15639749, 2025, data set.

[70] ——, "pblh_ML: ML-Based PBLH Retrieval from CALIOP TAB Measurements," https://github.com/andreukalu/pblh_ML, 2024, gitHub repository.

Retrieval of Planetary Boundary Layer Height from CALIPSO Satellite Observations Using a Machine Learning Approach

By A. Salcedo-Bosch et al.

Highlights

- **Accurate PBLH Estimation**: Random Forest model achieves R^2 = 0.71 and RMSE = 308 m using CALIPSO and radiosonde data.
- **Global Dataset**: Trained on 2.16 million radiosonde launches across 693 stations worldwide.
- **No Data Filtering Needed**: Model works in all atmospheric conditions, including multilayer clouds and dust.
- **Context Matters**: Temporal and geographic features are key to performance; removing them halves accuracy.
- **Urban-Ready and Scalable**: Performs well in urban areas (R^2 = 0.68), suitable for air quality and climate monitoring.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: