



Improved estimation of diurnal variations in near-global PBLH through a hybrid WCT and transfer learning approach

Yarong Li^{1,2,#}, Zeyang Liu^{3,#}, Jianjun He^{2,*}

5

¹College of Earth and Environmental Sciences, Lanzhou University, Lanzhou, China

²State Key Laboratory of Severe Weather Meteorological Science and Technology, Chinese Academy of Meteorological Sciences, Beijing, China

³College of Electronic and Information Engineering, West Anhui University, Lu'an, China

10 [#]These authors contributed equally to this work

^{*}Corresponding author: Jianjun He (Email: hejianjun@cma.gov.cn)

Abstract

15 Diurnal variations in planetary boundary layer height (PBLH) is highly linked to weather, climate, and environmental processes. However, remaining challenges persist in estimating its diurnal behavior at a large scale due to insufficient observations and limitations of operational retrieval algorithms. This study proposed a deep learning framework based on an attention-augmented residual neural network to estimate diurnal variations in near-global
20 PBLH, incorporating profiles from a non-sun-synchronous lidar (Cloud-Aerosol Transport System: CATS) and meteorological fields. The framework can largely address the issue of multi-layer structures in space-borne lidar signals, significantly improving the accuracy of PBLH retrieval during morning and evening (with accuracy improvement approach 40% compared to traditional algorithms). Due to insufficient observations aligned with CATS
25 orbits, a pre-trained model was firstly trained using pseudo-labels from reanalysis, and then was transferred to observation-based target labels. The transfer model demonstrated superior performance in most regions and periods, outperforming conventional algorithms in capturing PBLH magnitude and its diurnal variations, though under-performing over complex terrains. Further assessments over different land covers shown that the transfer-trained model



30 estimated PBLH and diurnal patterns were highly consistent with those from radiosondes,
surpassing reanalysis outputs. For model capability, wavelet covariance transformation
derived potential PBLH and temperature profiles emerged as dominant factors, with
contributions exhibiting diurnal patterns. Overall, this work proposes a novel framework for
large-scale PBLH estimation and provides insights for improving conventional algorithms,
35 particularly through integrating remote sensing and machine learning.

1. Introduction

The planetary boundary layer height (PBLH) plays key roles in land-air exchanges and
lower tropospheric processes (LeMone et al., 2019; Medeiros et al., 2005), such as the
40 transfer and exchange of heat, momentum, humidity, and materials (Garratt, 1994; Holtslag et
al., 2013; Stull, 1988). As an interface between the turbulent boundary layer and the free
atmosphere, PBLH acts as a significant barrier and represents the degree of turbulent
diffusion, determining the upper limit of boundary layer processes and playing vital roles in
weather, climate, and environmental studies (Che et al., 2019; Davy and Esau, 2016; Guo et
45 al., 2021; Li et al., 2017). Particularly, weather and pollution conditions are largely dependent
on the diurnal behaviors of PBLH, which dominates the atmospheric dispersion and vertical
mixing of pollutants (Ding et al., 2013; Huang et al., 2023; Li et al., 2025).

Despite the crucial importance, accurately measuring diurnal variation of PBLH across
large scaled areas remains challenging due to spatio-temporal limitations of current detection
50 instruments. Radiosonde and lidar measurements allow precise representation of vertical
atmospheric structure (Seidel et al., 2010; Seidel et al., 2012). The radiosonde derived PBLH
generally served as a benchmark for validating simulations, reanalysis, and remote sensing
(Guo et al., 2021; Li et al., 2023; Yue et al., 2021). However, global radiosondes are typically
launched two or four times per day, and its coverage is much sparse in less-developed regions
55 (like Africa and South America). Lidar systems serve as a promising tool for continuous
PBLH monitoring (Chen et al., 2022; Liu et al., 2021), benefiting from their operation at
sub-minute temporal resolution. While ground-based lidar has limited spatial representation,
space-borne lidar enables large-scale PBLH detection across diverse regions (Jordan et al.,



2010; McGrath-Spangler and Denning, 2012). Li et al. (2023) demonstrated diurnal
60 variations in large-scale PBLH from a non-sun-synchronous satellite. However, they
retrieved PBLH exhibited large deviations in accuracy and diurnal patterns due to
uncertainties of retrieval and signal noises such as multi-layer structures.

Conventional algorithms for retrieving PBLH from satellite signals are typically developed
either to detect abrupt jumps in backscatter profiles (Kumar et al., 2018; Liu et al., 2015) or
65 to identify the first exceeding of an empirical threshold (Palm et al., 2021). These algorithms
suffered from significant accuracy challenges, due to at least three limitations: (i) the
presence of elevated residual layers prevent downward staring lidar from detecting the true
PBLH; (ii) cloud contamination or advected aerosols induce noises into lidar echos; and (iii)
parameter selection of algorithm affect its sensitivity to diverse profile structures. The
70 primary challenge for retrieving the diurnal variation of PBLH perhaps lies in minimizing the
influences of residual layers or multi-layer structures during its morning and evening
transition periods (Su et al., 2020; Li et al., 2023). Numerous efforts have been taken to
enhance the algorithm performance in operating multi-layer profile structures of space-borne
lidar, such as utilizing graphic clustering (Liu et al., 2018) or implementing additional
75 physical constraints (Kim et al., 2021; Su et al., 2017). However, to date, current algorithms
have not yet achieved optimal performance, primarily due to their inability to effectively
resolve ambiguity signal structures through automated detection.

In recent years, machine learning has been increasingly integrated into PBLH estimation,
achieving evidenced improvements. Several studies have employed deep neural network
80 frameworks to estimate PBLH using near-surface and vertical atmospheric variables (Nguyen
et al., 2021; Su and Zhang, 2024), constructing non-linear mapping from meteorologies to
PBLH. Based on parameters acquired from surface observations, remoter sensors, reanalyses,
and simulations, several random forest models were developed to predict PBLH (Guo et al.,
2024; Krishnamurthy et al., 2021), the results demonstrated greater consistency with
85 radiosondes and effectively corrected some inherent biases. There are gradient boosting
learning models been proposed (de Arruda Moreira et al., 2022; Peng et al., 2023), which
sequentially fits multiple weak learners, allowing the model to learn iteratively and improve



prediction accuracy progressively. These methodologies essentially addressed the regression relationship between PBLH and associated meteorological variables. There are also machine learning models were employed to refine retrieval technique from only remote sensing data. Rieutord et al. (2021) compared an unsupervised (AdaBoost) and a supervised (K-means) learning, to judge whether the lidar signals originate from the boundary layer or the above free atmosphere. Mei et al. (2022) proposed a VGG16-based convolutional neural network for PBLH detection using wavelet covariance transformation (WCT) images of ground-lidar backscatter, which can effectively suppresses contamination from clouds and residual layers. Sleeman et al. (2020) improved PBLH measurement from backscatter profiles under cloud condition through convolutional network.

Existing machine learning methodologies exhibit significant advantages in capture PBLH and its diurnal variations from noisy lidar signals. However, these studies have almost been limited to ground-based sites, and either require additional meteorological variables affecting PBLH evolution to be provided or necessitate human intervention to process remote sensing signals. These site-scaled models may not be generalizable on larger regions or global scale. Few studies have focused on improving PBLH estimation from space-borne lidar through machine learning approaches. This is primarily due to training a robust model requires substantial feature samples been provided, yet ground-based observations aligned with space-borne lidar overpass orbits are extremely scarce, making it difficult to obtain sufficient target labels; while those unsupervised learning methods often fail to achieve the desired performance (Rieutord et al., 2021).

Given the aforementioned considerations, this study proposes to construct a temporally and spatially adaptive deep learning model to estimate PBLH and its diurnal variations on a near-global scale using space-borne Cloud-Aerosol Transport System (CATS). As the satellite operates on a non-sun-synchronous orbit, it can capture a complete diurnal cycle (Yorks et al., 2016). To address the issue of insufficient matching samples with satellite orbits, this paper employs a transfer learning strategy. The approach involves first establishing a pre-trained model using large-scale samples matched by reanalysis data. And then, the feature extraction capabilities of the pre-trained model are transferred to small samples matched with ground



truth values. By fine-tuning the model weights, its representation for real-world targets and generalization are enhanced, thereby constructing more accuracy measurements of diurnal variations in large-scaled PBLH. Overall, this work presents the first attempt to integrate
120 attention mechanisms and transfer learning for diurnal PBLH estimation at near-global scale, overcoming the limitations of conventional algorithms in handling multi-layer atmospheric structures.

2. Dataset

125 2.1 satellite-based lidar profiles

This study aims to develop a robust and generalizable deep learning framework for PBLH estimation from space-borne CATS lidar. CATS was initiated to monitor atmospheric clouds and aerosols using advanced lidar technology and was mounted on the International Space Station's (ISS) Japanese Experiment Module. Launched on Jan. 10, 2015, the ISS
130 operated in 51.6° inclined orbits at an altitude of ~405 km, covering tropical and mid-latitude regions. Unlike sun-synchronous satellite, CATS exhibited a repeat cycle of approximately three days and operated at non-fixed overpass times. These characteristics allow CATS to capture large-scale diurnal variations in aerosols (Yu et al., 2021) and clouds (Zhao et al., 2023), as well as in PBLH (Li et al., 2023) after approximately 16 days of running. Due to
135 instrument malfunctions, available CATS backscatters for PBLH retrieval only limited from Mar. 2015 to Oct. 2017, exclusively at the 1064 nm. Such wavelength is more sensitive to aerosol structure and variations compared to 532 nm (Winker et al., 2007), but with a lower signal-to-noise ratio (SNR); such that CATS signals necessitate more rigorous denoising processes. Herein, the study acquired 1064 nm 'Total_Attenuated_Backscatter' profiles (TAB)
140 from the CATS V3.00 Level 1B product and 'Feature_Type' data from the Level 2 product. The collected L1B and L2 products have horizontal resolutions of 350 m and 5 km, respectively, while both maintain a vertical resolution of 60 m. Several additional CATS products: 'Profile_UTC_Time', 'DEM_Mean_Elevation', 'Bin_Altitude_Array', 'Opacity', 'Layer_Top_Bin', 'Layer_Base_Bin', 'Surface_Type', 'Sky_Condition' were collected to refine



145 the input features when training model. Only the daytime products were acquired and the determination for nocturnal PBLH falls outside scope of this paper.

2.2 Radiosondes and reanalyses derived PBLH

Given that radiosonde derived PBLH is typically recognized as ground truth, this study employed sounding profiles from Integrated Global Radiosonde Archive (IGRA) V2.0, which serves to generate benchmark PBLH and to assess performances of our deep learning model. IGRA offers exceptional temporal and spatial coverage, with current 466 radiosondes sites (Fig. S1) available in CATS overpassing areas. We acquired IGRA data temporally aligned with the CATS orbits. Sounding profiles employed for PBLH determination encompass geo-potential height, temperature, dew point depression, wind speed and direction. The bulk Richardson number method (Vogelezang and Holtslag, 1996) was adopted to calculate the PBLH, which can even maintain enough effectiveness under stable atmosphere regimes and coarse sounding resolutions. Nevertheless, procedures were still conducted to eliminate soundings with coarse vertical resolution: within 5 km from the surface, the profiles must include at least seven vertical levels of temperature and humidity records; along with at least four levels of wind records. If valid wind observations are fewer than seven levels, a cubic spline interpolation was employed to fill missing values (Zhang et al., 2013). However, we should aware that radiosondes have standard launch schedule (fixed at two UTC), only a few soundings coincide with CATS orbits, spatio-temporal overlaps between the two datasets are quite scarce. Fig. S1 gives their match-up information, where relatively rough matching rules (with distance limited to 150 km and time to 1.5 hour) were performed to enlarge the number of samples. As a result, we obtained totally 5368 valid matching samples, which cover the majority of the Earth's land, and larger sampling numbers observed in mid-latitude regions.

While the robustness and reliability of radiosonde-based PBLH, using only 5,368 matched samples to train a model is far from sufficient, especially considering these samples fall throughout diverse periods and regions. Therefore, two reanalyses outputted PBLH, ERA5 (the fifth generation European Centre for Medium-Range Weather Forecasts atmospheric reanalysis) and MERRA2 (the Modern-Era Retrospective Analysis for Research and Applications Version 2), were further acquired in this study. Two sets of PBLH share the



same temporal resolution (1-hour) but with discrepant spatial grids: $0.25^\circ \times 0.25^\circ$ (ERA5)
175 and $0.625^\circ \times 0.5^\circ$ (MERRA2). The grid-based reanalyses were interpolated to the orbit-based
CATS data using inverse distance weighting to ensure they are spatially aligned. In this study,
the MERRA2 PBLH was employed to construct one of training sets for the model, partly
because it assimilates aerosol information compared to ERA5 (Gelaro et al., 2017), making it
more approach to the intrinsic nature of CATS retrievals. Our prior study also reported that
180 using conventional algorithm retrieved PBLH from CATS was more consistent with that from
MERRA2 (Li et al., 2023). Moreover, we acquired vertical profiles of temperature, humidity,
and wind speed from MERRA2 as meteorological features for training the model. These
variables represent 3-hourly averaged values and were matched with CATS orbit based on the
closest temporal and spatial proximity.

185

3. Methodologies

3.1 Generate training data

WCT is one of typical PBLH retrieval techniques from satellite-based backscatters. This
study employed the Haar wavelet transform (Gamage and Hagelberg, 1993):

190
$$W_f(a, b) = \frac{1}{a} \int_{z_b}^{z_t} B(z) h\left(\frac{z-b}{a}\right) dz \quad (1)$$

where, $W_f(a, b)$ is the WCT coefficient, a is dilation factor, b denotes the central location of
vertical translation, $B(z)$ is backscatters, z_b and z_t represent the bottom and top limits when
integrating the Haar function, respectively. The Haar wavelet function is:

$$h\left(\frac{z-b}{a}\right) = \begin{cases} 1, & b - a/2 \leq z \leq b \\ -1, & b \leq z \leq b + a/2 \\ 0, & \text{elsewhere} \end{cases} \quad (2)$$

195 inherently, the WCT is designed to check the similarity between the lidar profile and wavelet
stepping function, its maximum peak represents the sharpest signal gradient, and thereby
considered as PBLH. However, selecting a proper dilation factor is crucial, diverse dilation
values exhibit significant impacts on step WCT signals. Particularly, a smaller dilation cause
WCT being sensitive to smaller fluctuations in backscatter profile and is susceptible to noise



200 interference, whereas a larger dilation may smooth out thin aerosol layers.

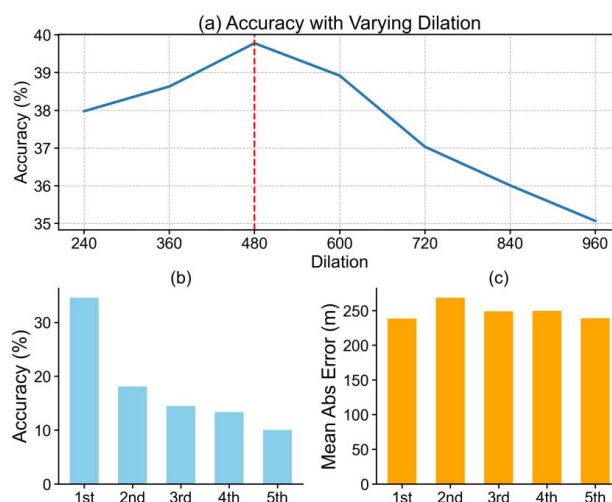


Fig. 1. Assessment for the WCT algorithm under different dilations (a); and accuracy (b) and MAE (c) compared against radiosonde derived PBLH when assuming one the first five peaks in WCT profiles (dilation=480) as PBLH.

205 Since varying sensitivity of different dilations to backscatter structures, we evaluated the retrieval accuracy of seven dilation values ranging 240–960 m (with an interval of 120 m) in Fig. 1a. Note that a tolerate PBLH bias of 500 m between WCT and radiosonde was utilized when calculating the retrieval accuracy, accounting for spatio-temporal matching errors and inherent algorithm differences between them. When compared against radiosonde derived
 210 PBLH, a dilation of 480 m yielded the optimal results. Therefore, a dilation of 480 m is taken as a benchmark for WCT in this work. However, its maximum accuracy of 39.7% does not meet reasonable desire, such uncertainty is mainly induced by multi-layer structures such as residual layer and advected aerosols, and inability of WCT algorithm (Li et al., 2023).

The WCT can, to some extent, be considered as a gradient-based algorithm, local peaks in
 215 WCT profile denote sharp changes in signal structure. A previous study adopted dynamic noise thresholds of ground-based lidar to identify the multiple layers (Li et al., 2023), but is not applicative to space-borne lidar profiles. Due to the magnitude of WCT represents the intensity of local changes in backscatter profile, we hypothesize in this study that the local



peaks in WCT profiles correspond to the top position of multi-layer aerosols; these peaks
220 were then compared against the radiosonde derived PBLH (Fig. 1b). The results show that the
first five peaks in WCT profiles aligned well with the truth PBLH, with their overall accuracy
exceeding 90% when we assumed one of the peaks to be PBLH. These peaks may not
necessarily originate from the PBLH and may be induced by other interfering signals,
whereas the first peak (i.e., benchmark for WCT algorithm), only capture few portion of truth
225 PBLH. In other words, the WCT can effectively detect complex signal structures, while its
maximum peak does not fully denote the PBLH. Therefore, the performance of WCT are
largely biased, particularly when it was utilized to CATS backscatters with strong temporal
variability. Fig. 1c further examine mean absolute errors (MAE) when assuming one of the
first five WCT peaks as PBLH, the values (~240 m) are much lower than that using WCT
230 algorithm (~1 km, not show here).

Consequently, this study proposes to develop a deep learning framework to identify the
optimal peak from the first several peaks of WCT profiles that aligned with the truth PBLH.
Three types of feature data: remotely sensed profiles, meteorological profiles, and auxiliary
parameters served as model inputs. We used the raw CATS backscatter profile as one of the
235 remotely sensed features. Due to the lower SNR, a series of pre-processing procedures were
implemented. First, we utilized the 'Opacity' parameter to remove opaque profiles, ensuring
downward-scanning CATS lidar can detect entire atmosphere columns. According to previous
retrieval practices (Li et al., 2023), profiles containing cloud layers were filtered using the
'Feature_Type' and corresponding 'Layer_Top_Bin', 'Layer_Base_Bin' from CATS L2 product.
240 Since CATS L1B and L2 products have diverse horizontal resolutions (a single L2 profile
involves 14 L1B profiles), all of 14 L1B profiles would be eliminated if any cloud layer was
detected in the L2 profile. Noting that cloud screening only applied below 5 km, profiles
remained available when the lowest cloud base exceeded this altitude. Prior studies have
suggested that long-distance horizontal smoothing can enhance SNR of daytime CATS
245 profiles (Nowottnick et al., 2022; Palm et al., 2021). Accordingly, the L1B profiles were then
horizontally averaged across 60 km, meaning each training unit aggregated 60 km of raw
CATS profiles. However, the solid ground generally return stronger signal echoes than the



above aerosols, which could potentially distort the long-distance smoothing. To address this, we re-aligned the CATS profiles according to their elevations, ensuring consistent bin for ground layers for a single training unit. Moreover, elevations of CATS profile extracted from the 'DEM_Mean_Elevation' may slightly bias from the true ground level, we thereby followed the same approach as Li et al. (2023) to re-calibrate the ground bin. Finally, to prevent the model from learning unforeseeable signal noises, we adopted a vertical smoothing window spanning three vertical bins into the profiles.

Based on the above cloud-screened, re-aligned, and horizontally averaged CATS profiles, we calculated the corresponding WCT profiles based on a dilation of 480 m. This study limits the PBLH estimation to height below 5 km (corresponding to 84 CATS bins), which covers the vast majority of global cases. Additionally, the two lowest bins (nearest the surface) were excluded to minimize ground noise contamination. Consequently, the derived PBLH values range from 360 m ($120 + a/2$) to 4800 m ($5040 - a/2$). From each WCT profile, we acquired an additional profile involves the candidate PBLH, with the same dimensions as the backscatter and WCT profiles. In candidate profile, most bins were assigned as '0', while the bins corresponding to local WCT peaks were marked as '1'. The WCT peaks were selected based on their sorted magnitudes, with a maximum of five peaks retained per profile. Overall, three remote sensing based profiles, encompassing averaged TAB, WCT, and candidate PBLH, each with dimensions of 84×1 , were incorporated as model inputs.

The meteorological profiles include temperature, relative humidity, and wind speeds obtained from MERRA2 3-D meteorological fields, which were first spatio-temporally matched with CATS orbits and then vertically interpolated to 84 corresponding CATS bins. In addition, the model inputs incorporated several non-profile parameters extracted from CATS auxiliary products, including geography information (latitude, longitude, altitude), local standard time (LST; converted from UTC of each profile), surface type, and sky conditions. These non-profile parameters were subsequently resampled to match the dimensions of the profile features, and finally forming a standardized input array ($84 \text{ bins} \times 12 \text{ features}$) for training the model, as the input layer shown in Fig. 2.

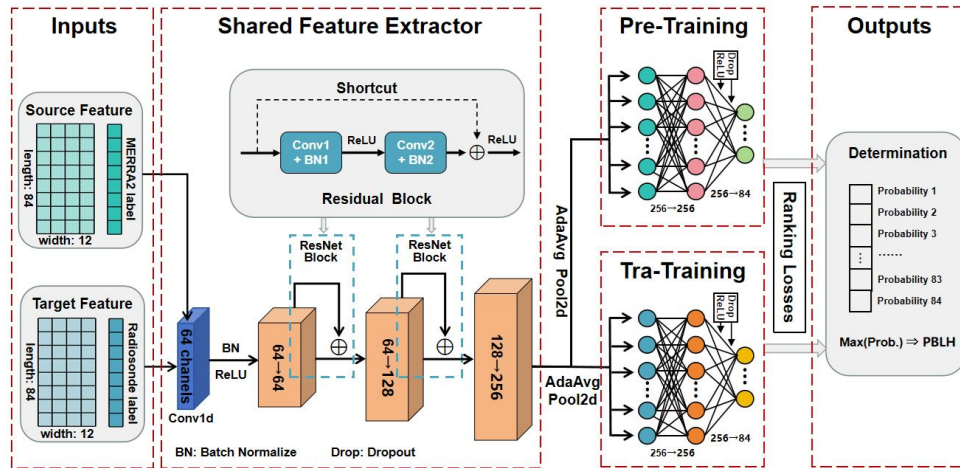


Fig. 2. Visualization for the model architecture, encompassing input layer (pre-training set and transfer training set), shared feature extractor (involving two attention augmented residual blocks), prediction heads (two discrepant fully connected layers for pre-training and transfer-training), and output layer.

In principle, the target labels for model training should be generated based on radiosonde derived PBLH . However, this study obtained only 5,368 matched samples between CATS and radiosonde data, which is far too limited to train a model capable of capturing both temporal and spatial PBLH variations. To address this challenge, a transfer learning strategy was adopted. Specifically, a baseline model was pre-trained across a large feature set using pseudo-labels constrained by MERRA2 PBLH, after which the pre-trained model was fine-tuned on a smaller dataset with target labels constrained by radiosonde derived PBLH. During the pre-training phase, the training target was defined as the single peak in the WCT profile closest to the MERRA2 PBLH, allowing a maximum deviation of 480 m that equals to one-fold dilation value. This approach enabled the model to learn vast feature information and substantially expanding the training sample size. For pre-training, a feature dataset of 2016 covering a completed calender year was employed, comprising 113,488 samples in total, and were split into training (80%) and validation (20%) subsets. In the transfer-learning stage, the target labels were constrained by radiosonde derived PBLH. There are 4,662 feature samples were extracted from the matched CATS-radiosonde samples. Of these, 4,000 samples were used for transfer training, while the remaining 662 samples served as a common testing



set to assess model performances for both pre-training and transfer-learning stages.

3.2 Model architecture

A residual neural network (ResNet) attempts to learn the residual mapping between input
300 features and outputs, effectively alleviating the vanishing gradient problems in a deep neural
network. This study constructs a ResNet based transfer learning framework for target location
detection, aiming to identify the only bin representing the PBLH. The approach reshapes
inputted feature array and employs a deep neural network to estimate the probability of each
bin approximating the truth PBLH. As illustrated in Fig. 2, the model adopts a modified
305 ResNet-18 architecture (He et al., 2016), consisting of four main components. (i) Input layer:
the model receives 2-D feature vectors (84×12) without spatial reshape, maintaining the
original temporal structure. (ii) Initial feature extracting: a 1D convolutional layer with 64
channels ($\text{kernel_size}=7$) processes the input sequence, followed by batch normalization and
ReLU activation. This maintains the original sequence length while expands the channel
310 dimension. (iii) Attention augmented residual blocks: three groups of down-sampling
networks ($64 \rightarrow 128 \rightarrow 256$ channels) process the extracted features, containing two residual
blocks. Notably, all convolutions use $\text{kernel_size}=3$ with to preserve sequence length. Each
residual block incorporates a parallel attention mechanism, where the positions of candidate
PBLH are transformed through a 1D convolution to weight the feature maps. Skip
315 connections are implemented through 1×1 convolutions when channel dimensions change.
(iv) Prediction heads: our model architecture includes a global average pooling across the
temporal dimension to aggregate sequence information, and two fully connected layers ($256 \rightarrow 256 \rightarrow 84$) with ReLU activation and dropout. Sigmoid activation producing probability
scores for each bin, the losses during training process were ranked to ensure that the score of
320 target bin is higher than that of non-target bins.

The architecture involves an end-to-end supervised learning approach to train an enhanced
attention-based ResNet based on PyTorch framework, where candidate PBLH with single
channel was mapped to 64 channels via 1D convolution to align with the main ResNet



networks, transforming position information into attention weights that explicitly leverage
325 prior knowledge for improved PBLH prediction. For the hyper-parameter tuning, the model
was trained using the Adam optimizer with an initial learning rate of 0.001, and was
optimized via binary cross-entropy loss. To prevent over-fitting, a dropout regularization with
a rate of 0.3 was implemented in the last fully connected layers, and an early stopping
mechanism is enabled (patience=10). Training process will be terminated when the validating
330 accuracy does not improved for 10 epochs.

Transfer learning is an efficient deep learning strategy that leverages prior knowledge from
pre-trained models to address new tasks (Pan and Yang, 2010). In this study, we first
pre-trained a ResNet model as our base network on a largely sampled dataset with target
labels constrained by MERRA2 PBLH. By virtue of the strong feature extraction capability
335 of the pre-trained model to learn common hierarchical features from the input data, we then
transfer it to a new task, establishing the optimal prediction model. For this new task, the
classification head at the end of the pre-trained model was removed and replaced with new
fully connected layers, which were re-trained on the smaller transfer-training dataset.
Meanwhile, the weights of the original convolutional layers were kept frozen to preserve the
340 learned feature representations. During transfer training, we employed a fine-tuning strategy
with a lower learning rate (0.0001), reduced training epochs and early stopping tolerance (=8)
to prevent overfitting.

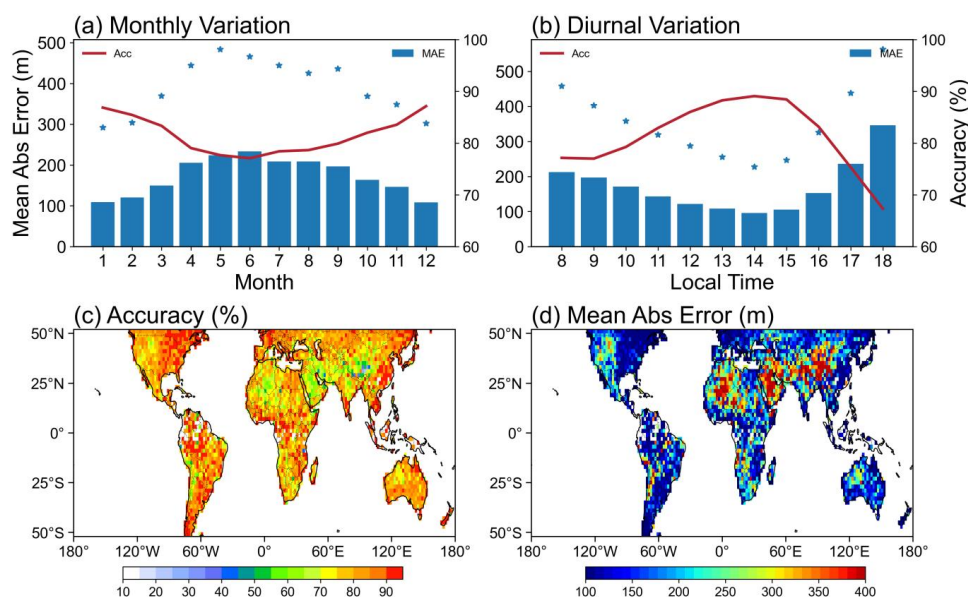


Fig. 3. Assessment of the pre-trained model. (a-b) give the accuracy (column), MAE (solid line) and standard deviation of MAE (asterisk) at monthly and hourly scale, respectively; (c) and (d) denote the spatial distribution of accuracy and MAE.

Given that the majority of matched CATS-radiosonde samples fall in land (Fig. S1), features over oceans were filtered out when training the pre-training model. As training curves shown in Fig. S2, the pre-training model achieved its optimal validation accuracy at 38th epoch, and training was stopped at 48th epoch due to early stopping. The optimal model demonstrated an accuracy of 80.24% on the training set and 81.18% on the validation set, with corresponding losses of 0.0209 and 0.0204. Over the common testing set, the pre-trained model achieved an accuracy of 65.87%. Such test accuracy surpasses a non-transfer trained model (training the model only across the radiosonde-constrained training set, with testing accuracy of 63.2%), see training curves in Fig. S3. However, our transfer training achieved a superior performances than both the pre-training and non-transfer models. The transfer model early stopped at 22th epoch (Fig. S4), reaching accuracies of 75.38% and 72.36% over training and validating sets, and a testing accuracy of 69.03%. This indicates that employing a transfer learning strategy can effectively enhance the model's learning capabilities and increase its generalization.



Fig. 3 preliminarily evaluates the temporal (monthly and hourly) and spatial differences in accuracy and MAE of the pre-trained model. The results indicate the pre-trained model performed well over most land areas. However, the model's representation in high-altitude regions (Tibetan Plateau, Rocky Mountains) and desert areas (Sahara, Arabian Peninsula) were somewhat weak, where the accuracy dropped below 70% and the MAE exceeded 400 m. These inabilities can be attributed to three main causes. First, the long-distance signal smoothing in processing raw CATS profiles may cause uncertainties over complex terrain. Second, grid-based MERRA2 data represents average state within a grid-cell, potentially leading to matching errors with orbital CATS observations in high-altitude areas. Finally, meteorological profiles and PBLH from MERRA2 may contain larger errors in these regions due to sparse observations available for assimilation. The capabilities of pre-trained model also exhibited seasonal and diurnal discrepancies. Particularly, the model demonstrate poorer performance from April to September compared to other months. As the poorer performances were primarily sourced from the Northern Hemisphere, it can be concluded that the model's representation in spring and summer seasons were somewhat weaker than that in autumn and winter. From a perspective of diurnal variation, the pre-trained model performed less effectively during morning and later afternoon hours compared to around midday, with particularly poor performance observed in the later afternoon.

3.3 Feature importance permutation

Based on the transfer-trained model, we examined the importance score of each input feature using permutation importance technique (Altmann et al., 2010; Breiman, 2001). By randomly shuffling individual feature and measuring decreases in model performance, this method directly quantifies feature importance and can capture the non-linear dependencies among different features. Since the proposed ResNet model is essentially a classification task, we quantify the feature importance scores by calculating the increase in MAE induced by feature shuffling. Specifically, permutation importance estimations were implemented based on radiosonde constrained dataset (4662 samples), and the baseline MAE over original test dataset was firstly derived. And then, we randomly shuffled the target feature across all samples, ensuring that 84 bins of target feature move synchronously from every input sample,



390 while keeping other features unchanged. This will break the association between the target feature and predict label and is much applicable for our position sensitive predict task. The importance score is determined by the increased magnitude of MAE after permutation, a larger MAE increase indicates an higher feature importance. To enhance the robustness of feature permutation, each feature undergoes 30 independent iterations with different random sequences, noting that the input features were shuffled using a common random seed in each iteration. The ultimate importance scores were represented as mean value across 30 iterations.

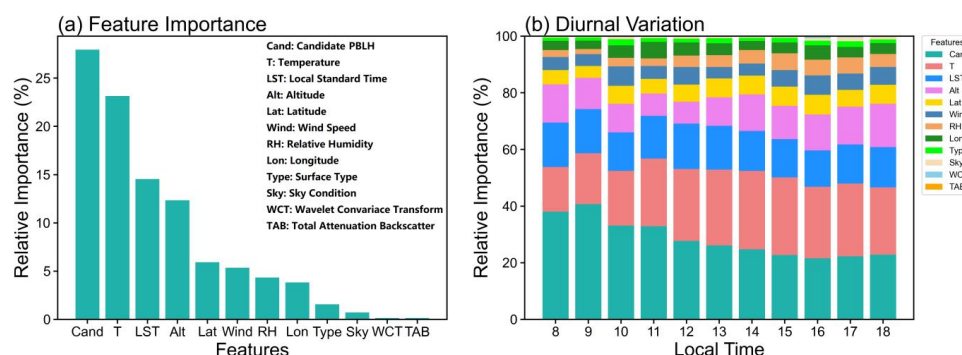


Fig. 4. The permutation importance of input features is measured by the increase in MAE when each individual feature is randomly shuffled. These importance scores are then normalized to represent their relative contributions (a), with the total importance summing to 100%. (b) illustrates their relative importance scores at each hour.

The obtained importance score of each input feature was recalculated to derive its relative contribution rate. As shown in Fig. 4a, two profile features (candidate PBLH, temperature), along with two non-profile features (LST, altitude) emerged as the most important features, each with relative importance exceeding 10%. Geographic associated variables (latitude, longitude) and two meteorological (humidity, wind speed) profiles contributed the secondary importance, collectively contributing over 20% to the total importance, whereas surface type and sky conditions contributed marginally. Among the three remotely sensed profiles, importance score of TAB and WCT were negligible, despite candidate PBLH playing the dominant role in the model. This implies that local peak/valley locations in backscatter profiles are more important than other shape features when estimating PBLH from CATS



profiles. This may also suggest potential direction for improving conventional retrieval algorithms, that is, the shape and structure of remotely sensed profiles provide limited information about the PBLH, efforts should be taken to incorporate other diagnostic data, as suggested by (Su et al., 2020).

We further extracted the permutation importance of input feature at each hour, and present their diurnal variations (Fig. 3b). The hourly importance scores of the two dominant contributors (candidate PBLH and temperature) varied evidently, while the diurnal variation of other importance scores were relatively slight. The combined importance of the two dominant contributors exceeded 50%, and their diurnal variation exhibited an alternating dominance pattern. Specifically, candidate PBLH dominates the model's capability during the morning periods with a gradually decreasing tendency, while the temperature emerges as the primary factor in the afternoon, with its importance scores essentially surpassing those of candidate PBLH. The diurnal variations in these importance scores might lead to diurnal behaviors of model performance (Fig. 3b), which will be discussed in the next section.

4. Results and Discussions

4.1 Assessing the model

Herein, we evaluated the performance of the transfer-trained model by checking whether the model effectively captured the target labels constrained by radiosonde derived PBLH. The assessment was categorized into training and testing subsets, Fig. 5a-d illustrate the diurnal variations in accuracy and MAE and for WCT, pre-trained and transfer-trained models. Notably, the calculated accuracy for WCT was slightly higher than that in Fig. 1a because the current assessment is carried out only on the feature samples (4662) rather than all of the matched CATS-radiosonde samples (5368). Both the pre-trained and transfer-trained models demonstrated different degree of enhancements related to the conventional WCT algorithm, and the performance of transfer-trained model was reasonably better than the pre-trained model. Quantitatively, the transfer-trained model achieved an increase of 26.1% (24.9%) in accuracy and a reduction of 537.2 (517.2) m in MAE compared to the WCT on training



(testing) subset, demonstrating the substantial advantage of transfer-training in refining PBLH measurements from CATS data.

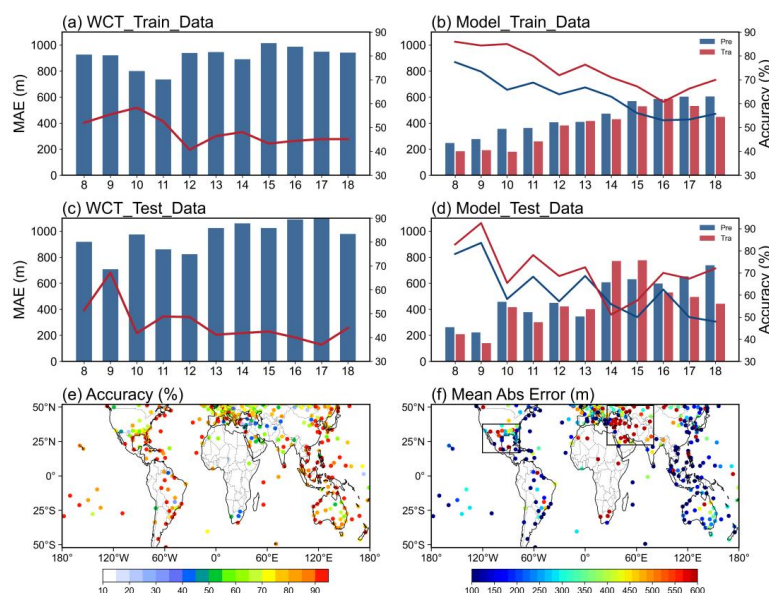


Fig. 5. Performance comparison of the WCT, pre-trained model, and transfer-trained model against radiosonde-constrained target labels. (a-b) show results for the training set, while (c-d) correspond to the testing set. (e-f) display the accuracy and MAE of the transfer-trained model on the combined training and testing dataset.

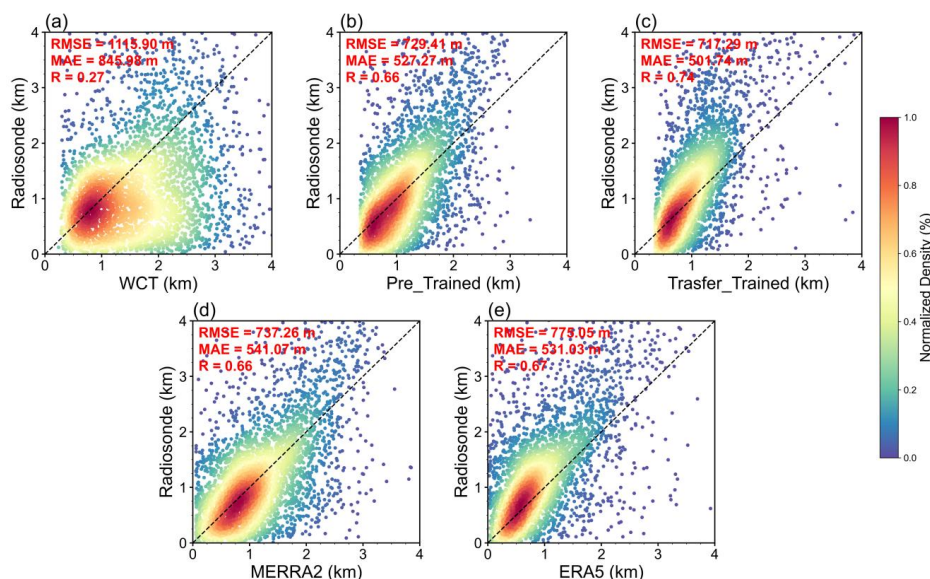
For the diurnal variations, transfer-trained model performed better during the morning and midday periods but poorly in the afternoon. In other words, its performance deteriorated over daytime hours. It is interesting to note that the diurnal variation of the model performance aligned closely with the importance scores of candidate PBLH in Fig. 4b, while exhibited an inverse tendency with that of temperature. This further underscores the dominant role these two factors in regulating the model's capability. These diurnal variations may be largely regulated by the spatial distribution of training samples. Since radiosondes are only launched at two fixed times (00:00 and 12:00 UTC), each sits can provide samples at different local time. The poorest accuracy and largest MAE typically occurred between 14:00 and 16:00 LST, with most samples originating from western North America and the Middle East (see rectangular boxes in Fig. 5f), regions characterized by high-altitude terrain or desert areas. As



illustrated in Fig. 3, the pre-trained model performed poorly in these regions. Even so, the transfer-trained model achieved significant improvements in the later afternoon. Additionally, the pre-trained model exhibited generally weak performance during morning and evening periods (Fig. 3b), whereas the transfer-trained model performed better in the morning than at other daily times. This may attribute to the fact that morning samples were predominantly collected from regions around 120° E and 60° W, where the pre-trained model performed stronger feature extraction capabilities in these low-altitude areas compared to others (Fig. 3c-d). Fig. S5 further provides the accuracy and MAE at site scale for WCT, pre-trained, and transfer-trained models, the transfer-trained model achieved enhancement at nearly all sites.

4.2 Inter-comparison of multi-sourced PBLH

The above analyses primarily involve to validate the model's capability in capturing target labels, where the positions typically correspond to the WCT peak closest to either MERRA2 or radiosonde derived PBLH. In fact, the core function of the model is selecting, from multiple WCT peaks, the one that most accurately represents the PBLH based on provided meteorological and physical conditions. It is crucial to aware that the model output remains a remotely sensed product, while radiosonde derived PBLH is regarded as closest to ground truth and generally serves as benchmarks for validating other measurements. Accordingly, Fig. 6 presents scatter plots comparing PBLH estimations from WCT, pre-trained model, transfer-trained model, MERRA2, and ERA5 against those from radiosondes. To enable systematic comparisons, their outputs within 150 km of sounding sites were averaged to derive statistical metrics, including correlation coefficients (R), MAE, and root mean square error ($RMSE$).



480

Fig. 6. Scatter plots comparing PBLH estimations from (a) WCT, (b) pre-trained model, (c) transfer-trained model, (d) MERRA2, and (e) ERA5 against radiosonde observations. Unlike Fig. 5, these comparisons employ direct radiosonde-derived PBLH rather than the radiosonde-constrained target labels. Statistical metrics, correlation (R), MAE, and RMSE are present in red at the upper left corner of each subplot.

485 As results, the conventional WCT algorithm had the poorest consistence with radiosondes, with the weakest correlation (0.27) and largest MAE (845.98 m) among all these comparisons. Although we have previously obtained reasonable consistency between them by filtering PBLH under stable regime and separately comparing samples under cloudy and clear-sky conditions (Li et al., 2023), those statistical metrics remained inferior to the comparisons
 490 between radiosonde and MERRA2 and ERA5. However, our ResNet model significantly enhanced the representation for the ground truth PBLH. The transfer-trained model demonstrated marked improvements in predictive capability, exhibiting higher consistency with radiosondes than both the pre-trained model and two reanalysis datasets, with the strongest correlation (0.74) and the lowest MAE (501.74 m). Since the pre-trained model
 495 using pseudo-labels constrained by MERRA2 PBLH, its statistical metrics with radiosonde closely consistent with those between MERRA2 and radiosonde. In addition, Table 1 gives the comparisons between the WCT, pre-trained model, and transfer-trained model and



radiosonde at each daytime hour. The transfer-trained model is also superior than others at hourly scale, indicating its ability to capture more accurate diurnal variations of PBLH. However, its correlations in the morning were somewhat weaker, This may seem to contradict the accuracies in Fig. 4. In fact, the correlation represents the linear relationship between model predicts and radiosonde calculation, accuracy is statistically based on deviations between them. The transfer-trained model predicted PBLH exhibited the lowest biases from ground truth during the morning period, which accordingly leads to its higher accuracy.

Table 1. Statistics metrics for comparing PBLH from WCT, pre-trained model, and transfer-trained model to radiosonde derived PBLH at each daytime hours.

Local Time		8	9	10	11	12	13	14	15	16	17	18
WCT	Correlation	0.18	0.1	0.12	0.24	0.31	0.3	0.25	0.28	0.32	0.21	0.27
	MAE (km)	0.87	0.93	0.8	0.64	0.74	0.84	0.88	0.96	1.01	0.97	0.92
	RMSE (km)	1.11	1.2	1.04	0.86	1	1.15	1.24	1.29	1.36	1.38	1.26
Pre-trained	Correlation	0.24	0.37	0.38	0.41	0.67	0.72	0.66	0.69	0.69	0.65	0.59
	MAE (km)	0.38	0.39	0.44	0.44	0.45	0.54	0.61	0.68	0.77	0.79	0.72
	RMSE (km)	0.5	0.5	0.56	0.58	0.65	0.79	0.88	0.96	1.07	1.11	1.08
Tra-trained	Correlation	0.44	0.34	0.49	0.48	0.71	0.72	0.71	0.7	0.76	0.72	0.68
	MAE (km)	0.33	0.34	0.33	0.39	0.46	0.56	0.63	0.72	0.73	0.7	0.6
	RMSE (km)	0.45	0.46	0.43	0.54	0.67	0.84	0.9	0.98	1.01	0.97	0.89

Although radiosondes are considered as ground truth and are often served as benchmarks for evaluating models and reanalysis outputs, complete quantitative consistency cannot be pursued due to mis-matches in both space and time with other datasets and discrepancies in retrieval algorithms. Instead, we can only ensure certain consistency in their spatio-temporal characteristics. The prominent consistency between the transfer-trained model and radiosonde demonstrates the superiority of deep learning approaches and validates the rationality of our experimental design. However, it should be noted that the matchup between orbital CATS data and radiosonde sites remain relatively crude, exhibiting considerable mismatches in



temporal, horizontal distance, and altitudes. Moreover, different vertical resolution of
 radiosondes induce uncertainties in deriving PBLH. Therefore, the PBLH differences
 between the transfer-trained models and the WCT compared to the sounding derived PBLH
 were examined in Fig. 7 and Fig. S6, with respect to their sensitivity to the matching
 differences in distance, time, and altitude, as well as vertical resolution of radiosonde. It can
 be observed that although the PBLH deviations exhibited slight dependence on time
 difference, distance difference, and vertical resolution, this dependence is not obvious. This
 implies that the matching criteria between the radiosonde sites and CATS orbits cannot cause
 substantial uncertainties in this study. However, significant PBLH differences emerged as the
 altitude difference increasing. This is related to the poor model performance over rugged
 terrain, and it also highlights the heterogeneity of PBLH over complex terrains.

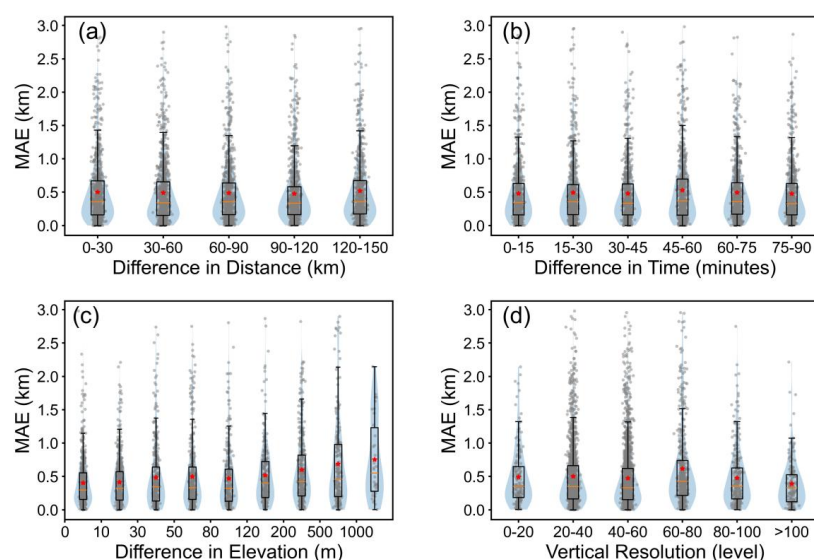


Fig. 7. Sensitivity analysis of PBLH differences between transfer-trained model predictions and radiosonde
 measurements with respect to: (a) matching distance, (b) matching time, (c) matching elevation, and (d)
 vertical resolution of sounding profiles.

4.3 Diurnal variations in near-global PBLH

Benefiting from the unique operational characteristics of the CATS, the near-global diurnal
 variations in PBLH can be obtained after approximately 16 days of operation. However, due



535 to interference from multi-layer structures and noises in backscatter signals, diurnal variations
derived by conventional WCT algorithm often present non-physical fluctuations (Li et al.,
2023). This study aims to extract more physically reasonable diurnal PBLH variations from
CATS data using a deep learning approach. Based on theory by Stull (1988), we assumed that
daytime PBLH evolution undergoes four distinct phases: morning transition (08:00–09:00),
540 rapid growth (10:00–14:00), maintenance (14:00–16:00), and decay in the late afternoon
(17:00–18:00). Fig. 8 presents spatial distributions of PBLH for the four evolution periods
derived from WCT, pre-trained model, transfer-trained model, MERRA2, and ERA5.
Additionally, Fig. S7 provides further details by highlighting the specific daytime PBLH at
each hour. These results demonstrate reasonable diurnal PBLH behaviors, and they shown
545 evident differences among different datasets or methodologies.

Similar as previous results by Li et al. (2023), the diurnal variation amplitudes derived
from the WCT algorithm were severely weakened, showing no significant difference between
the morning transition period and the afternoon maintenance period. In contrast, our ResNet
model captured clearer diurnal patterns: lower PBLH was observed in the morning transition
550 period, gradually increased at the growth period, reached its daily maximum in the
maintenance stage, and then began to decline during the decay period. However, such typical
diurnal variation pattern appeared more pronounced for the pre-trained model, while became
relatively indistinct for the transfer-trained model. It is evident that the pre-trained model
predicted PBLH exhibited significant spatial consistency with MERRA2 and ERA5 outputs
555 compared to the transfer-trained model. This is mainly because the transfer-trained model
predicted higher PBLH in high-altitude regions and in decay phase in the later afternoon.
However, higher consistency between the pre-trained model and reanalysis did not indicate
its predictions are closer to true values, as reanalysis outputs themselves still require careful
assessment.

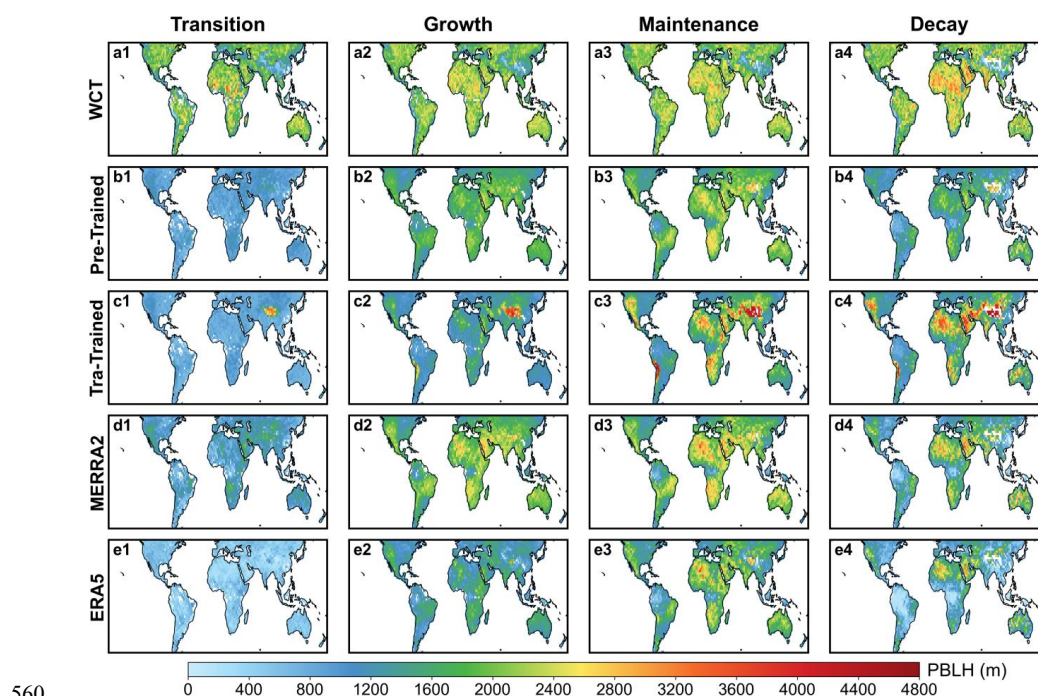


Fig. 8. Spatial distributions of PBLH derived from (a1-a4) WCT, (b1-b4) pre-trained model, (c1-c4) transfer-trained model, (d1-d4) MERRA2, and (e1-e4) ERA5 during four diurnal evolution phases: morning transition, rapid growth, maintenance, and afternoon decay.

However, we must acknowledge the transfer-trained model exhibited some anomalous performance. Such as its abnormally higher PBLH over the Tibetan Plateau during the morning transition and afternoon decay periods, which severely deviated from actual situations. This partly stems from the inherent limitation in feature extraction capability of the pre-trained model over high-altitude regions (Fig. 3). Furthermore, the scarcity of available training samples in high-altitude regions for the transfer-trained model can also cause substantial uncertainties in its performances. Therefore, it can be concluded that the transfer-trained model predicted PBLH in high-altitude regions being unreliable in this study. Additionally, the transfer-trained model predicted PBLH in the later afternoon did not significant decay and remained notably higher than those derived from other methods or datasets. Fig. S8 illustrates the diurnal variations of PBLH derived from the transfer model at four seasons. There were almost no discernible decays in PBLH during summer (JJA in the



Northern Hemisphere and SON in the Southern Hemisphere); instead, it even maintained an increasing trend. In contrast, only slight PBLH decays were observed in other seasons.

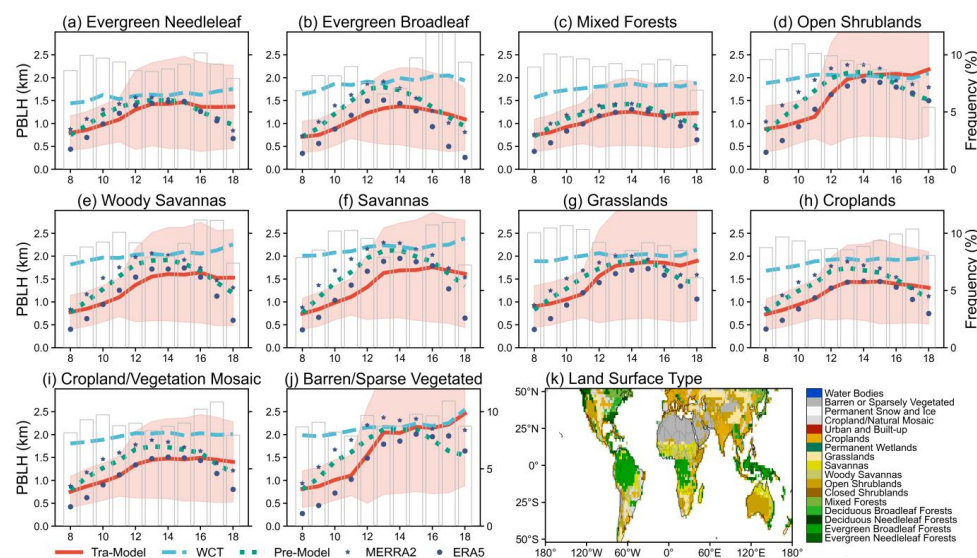
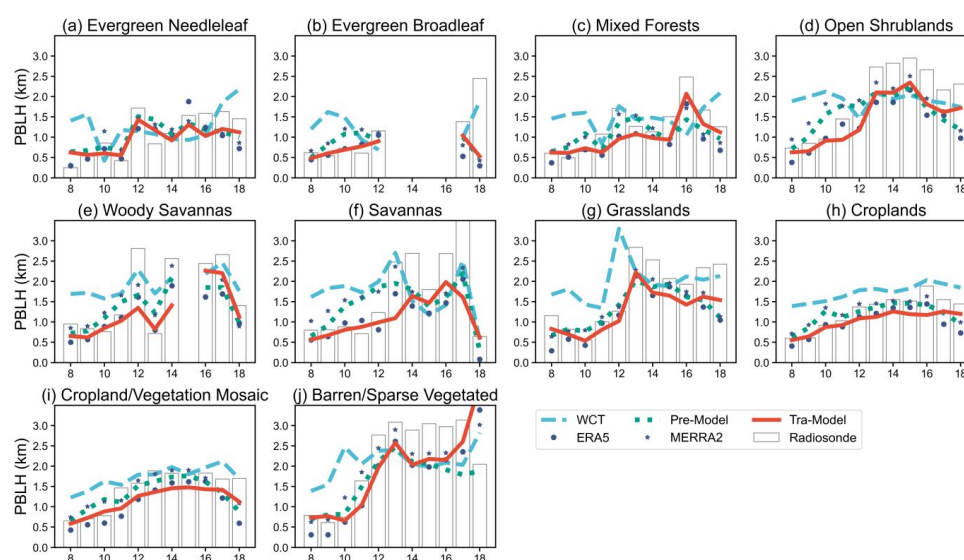


Fig. 9. Hourly PBLH from WCT, pre-trained model, transfer-trained model, MERRA2, and ERA5 over the
 580 major ten land cover types. The bar plots denote sampling frequency for a specified land cover at each
 daytime hours. (k) reveals the land cover distributions across $2^\circ \times 2^\circ$ grids.

Evolution of PBLH is mainly governed by surface conditions and is highly dependent on
 land surface types (Li et al., 2021). To better illustrate its diurnal variation, Fig. 9 presents the
 hourly PBLH across ten major land surface types (derived from the three approaches and two
 585 reanalyses). The transfer-trained model demonstrated significant improvements in capturing
 diurnal variations compared to WCT at most land covers, exhibiting more realistic diurnal
 patterns in terms of amplitude, growth duration, and peak timing. Particularly, the model
 present clearer morning growth phase and more accurate peak attainment timing. Additionally,
 the model predicted PBLH exhibited a more pronounced dependence on land cover, with
 590 higher PBLH and greater diurnal amplitude observed over bare soil and shrublands compared
 to forests, croplands, and grassland areas. These findings are consistent with our previous
 observation based report (Li et al., 2021), whereas the WCT predicted PBLH exhibited much
 smaller deviations across different land surface types. In addition, the diurnal PBLH variation



patterns (amplitude, peak timing) derived from our models aligned closely with those from
 595 the two reanalyses. Specifically, the pre-trained model displayed nearly identical diurnal
 patterns to MERRA2, while the transfer-trained model performed more closely with ERA5
 during the growth and maintenance period. However, the transfer-trained model predicted
 much higher PBLH than ERA5 during transition and decay phases.



600 Fig. 10. Comparing the diurnal variations of PBLH estimated from WCT, pre-trained model,
 transfer-trained model, MERRA2, and ERA5 to that derived from radiosondes (bar plots).

As mentioned above, our transfer-trained model derived PBLH decay in the later afternoon
 is not pronounced in most land covers, with PBLH magnitudes during this period being
 notably higher than those from the reanalyses and pre-trained model. This is primarily due to
 605 reanalysis outputted PBLH is highly dependent on thermodynamic conditions and begins to
 decay after surface thermal flux reaches its afternoon maximum. In contrast, the model
 predicted PBLH is associated with backscatter of aerosols, which does not diminish
 synchronously with thermodynamic weakening (Wang et al., 2012). Likewise, Pearson et al.
 (2010) synthesized numerous studies regarding the diurnal variations of PBLH, obtaining
 610 diurnal variation curves that resembled the results from our transfer-trained model, which
 suggested the credibility of the diurnal patterns predicted by our transfer-trained model. To



further support this conclusion, we qualitatively compared the diurnal PBLH patterns from WCT, model predictions, and reanalysis outputs, to the radiosonde observations across the ten major land types (Fig. 10). Although these diurnal variations were composited from
615 radiosonde sites at different longitudes (potentially inducing perturbations in diurnal curves), their one-to-one matchup with other PBLH can still provide certain effective evidences. The results demonstrated that over most land covers, PBLH from the two reanalyses shown more pronounced decays and lower magnitudes in the later afternoon than the radiosonde derived PBLH. Additionally, ERA5 exhibited lower PBLH than radiosonde observations during the
620 morning transition period. These findings indicate that our transfer-trained model captured more reasonable diurnal patterns, and the ResNet based transfer learning approach can effectively estimate near-global PBLH from CATS data.

5. Conclusions

This study developed a spatially and temporally applicative ResNet learning framework to
625 estimate near-global diurnal variations in PBLH from approximately three years of CATS lidar profiles. The proposed model demonstrates significant enhancement in estimating large-scale PBLH compared to conventional algorithm. The framework is designed based on the concept that the first few peaks in WCT profiles typically capture the true PBLH, and the model is inherently proposed to identify the peak with the highest probability of representing
630 the actual PBLH. Given the radiosonde measured PBLH samples for training a robust deep learning model are insufficient, this study adopted a transfer learning strategy. We first trained a base model using pseudo-target constrained by MERRA2 PBLH and then fine-tuned the base model on a smaller sampling dataset to generate the optimal model. This transfer-trained model retained the strong feature extraction capabilities of the pre-trained
635 model and demonstrated considerable improvement in performance when evaluated on unseen data.

The input features for the model include remotely sensed and meteorological profiles, geographic and temporal information, as well as surface/sky conditions. Among these, candidate PBLH derived from CATS backscatters and temperature profiles are the two
640 dominant factors influencing model performance, collectively accounting for more than 50%



of the importance scores. Their importance exhibits a distinct diurnal variation with alternating dominance: candidate PBLH primarily influences morning periods while temperature dominates the afternoon. This alternating dominance pattern further explains the diurnal variation in model performance, with higher accuracy and lower MAE observed
645 during morning hours and the opposite tendencies occurred in the late afternoon. Despite these temporal fluctuations, the transfer-trained model demonstrates overall superior performance metrics when compared against radiosondes, outperforming the results obtained from WCT, pre-trained model, MERRA2, and ERA5.

Regarding diurnal variation, the transfer-trained model predicted PBLH exhibited clear
650 diurnal patterns, demonstrating more reasonable diurnal amplitude, growth duration, and peak timing compared to the conventional WCT algorithm. Although the model struggled to capture PBLH over high-altitude regions like the Tibetan Plateau due to insufficient training samples, its performances in other regions were significantly better. Particularly, the model derived diurnal PBLH variations were sensitive to land covers. PBLH over bare and shrub
655 lands had higher magnitude and larger diurnal amplitudes than that over forests, croplands, and other vegetated areas. Furthermore, the model maintained high PBLH magnitudes in the late afternoon and shown only slight decay, differing from the pronounced decay phases of the two reanalyses derived PBLH. However, this non-prominent afternoon decay aligned well with radiosonde measurements, indicating its superior capability in capturing diurnal PBLH.

660 This study involves an initial attempt of using a deep neural network to address complex signal structure in CATS backscatter, and then to re-fine its measurement for PBLH on a near-global scale. Although utilizing attention augmented ResNet and transfer learning strategy can effectively improve the model's capability, its performances in high-altitude regions and in the morning and later afternoon periods remain poor. Future efforts would be
665 prospected to refine the model's applicability in rugged topography or in certain local time, and fine-resolution meteorological data and accurate target label are crucial for improving the model performances.



Author contributions

YL: Conceptualization, Methodology, Formal analysis, Investigation, Writing-original draft, Software, Validation. ZL: Software, Validation, Formal analysis, Writing-review and editing. JH: Conceptualization, Formal analysis, Methodology, Supervision, Writing-review and editing.

Code and data availability

Data and software used in this study are available as follows. Relevant datasets and scripts necessary to understand, evaluate, and extend the research findings reported in this paper were archived in the Zenodo repository, accessible under the DOI: 10.5281/zenodo.16907935 (<https://zenodo.org/records/16907935>). The IGRA V2 radiosonde data (Durre and Yin, 2008) is available at <https://www.ncei.noaa.gov/products/weather-balloon>. CATS (Yourks, et al., 2016) data can be acquired from <https://cats.gsfc.nasa.gov/>. ERA5 (Hersbach et al., 2023) data obtained from Copernicus Climate Change Service (C3S) Climate Data Store accessible at <https://cds.climate.copernicus.eu/>. MERRA2 data (Gelaro et al., 2017) is archived at <https://disc.gsfc.nasa.gov/datasets?keywords=MERRA-2&page=1>.

Acknowledgements

We acknowledge the provision of radiosonde by the National Oceanic and Atmospheric Administration (NOAA), space-borne CATS lidar by the National Aeronautics and Space Administration (NASA), ERA5 and MERRA2 reanalyses by Copernicus Climate Change Service and NASA Global Modeling and Assimilation Office.

Financial support

This study was supported by the National Key Research and Development Program of China (2023YFC3706304).

Competing interests

The contact author has declared that none of the authors has any competing interests.



References

- Altmann, A., Tološi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10), 1340-1347.
- 705 Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
<https://doi.org/10.1023/A:1010933404324>
- Che, H., Gui, K., Xia, X., Wang, Y., Holben, B. N., Goloub, P., et al. (2019). Large contribution of meteorological factors to inter-decadal changes in regional aerosol optical depth. *Atmospheric Chemistry and Physics*, 19(16), 10497-10523.
- 710 Chen, S., Tong, B., Russell, L. M., Wei, J., Guo, J., Mao, F., et al. (2022). Lidar-based daytime boundary layer height variation and impact on the regional satellite-based PM_{2.5} estimate. *Remote Sensing of Environment*, 281.
- Davy, R., and Esau, I. (2016). Differences in the efficacy of climate forcings explained by variations in atmospheric boundary layer depth. *Nature Communications*, 7(1), 11690.
715 <https://doi.org/10.1038/ncomms11690>
- de Arruda Moreira, G., Sánchez-Hernández, G., Guerrero-Rascado, J. L., Cazorla, A., and Alados-Arboledas, L. (2022). Estimating the urban atmospheric boundary layer height from remote sensing applying machine learning techniques. *Atmospheric Research*, 266.
- 720 Durre, I., and X. Yin, 2008: Enhanced radiosonde data for studies of vertical structure. Bulletin of the American Meteorological Society, 89, 1257-1262.
- Ding, A. J., Fu, C. B., Yang, X. Q., Sun, J. N., Petäjä, T., Kerminen, V. M., et al. (2013). Intense atmospheric pollution modifies weather: a case of mixed biomass burning with fossil fuel combustion pollution in eastern China. *Atmospheric Chemistry and Physics*, 13(20), 10545-10554.
- 725 Gamage, N., and Hagelberg, C. (1993). Detection and Analysis of Microfronts and Associated Coherent Events Using Localized Transforms. *Journal of Atmospheric Sciences*, 50(5), 750-756.
https://journals.ametsoc.org/view/journals/atsc/50/5/1520-0469_1993_050_0750_daaoma_2_0_co_2.xml
- 730 Garratt, J. R. (1994). Review: the atmospheric boundary layer. *Earth-Science Reviews*, 37(1), 89-134. <https://www.sciencedirect.com/science/article/pii/0012825294900264>
- Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., et al. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, 30(14), 5419-5454.
735 <https://journals.ametsoc.org/view/journals/clim/30/14/jcli-d-16-0758.1.xml>
- Guo, J., Zhang, J., Shao, J., Chen, T., Bai, K., Sun, Y., et al. (2024). A merged continental planetary boundary layer height dataset based on high-resolution radiosonde measurements, ERA5 reanalysis, and GLDAS. *Earth System Science Data*, 16(1), 1-14.
- 740 Guo, J., Zhang, J., Yang, K., Liao, H., Zhang, S., Huang, K., et al. (2021). Investigation of near-global daytime boundary layer height using high-resolution radiosondes: first



- results and comparison with ERA5, MERRA-2, JRA-55, and NCEP-2 reanalyses. *Atmospheric Chemistry and Physics*, 21(22), 17079-17097.
- 745 He, K., Zhang, X., Ren, S., and Sun, J. (2016). *Deep Residual Learning for Image Recognition*. Paper presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J.-N. (2023): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.adbb2d47
- 750 Holtzlag, A. A. M., Svensson, G., Baas, P., Basu, S., Beare, B., Beljaars, A. C. M., et al. (2013). Stable Atmospheric Boundary Layers and Diurnal Cycles: Challenges for Weather and Climate Models. *Bulletin of the American Meteorological Society*, 94(11), 1691-1706.
- 755 Huang, X., Wang, Y., Shang, Y., Song, X., Zhang, R., Wang, Y., et al. (2023). Contrasting the effect of aerosol properties on the planetary boundary layer height in Beijing and Nanjing. *Atmospheric Environment*, 308, 119861. <https://www.sciencedirect.com/science/article/pii/S135223102300287X>
- 760 Jordan, N. S., Hoff, R. M., and Bacmeister, J. T. (2010). Validation of Goddard Earth Observing System-version 5 MERRA planetary boundary layer heights using CALIPSO. *Journal of Geophysical Research: Atmospheres*, 115(D24).
- 765 Kim, M.-H., Yeo, H., Park, S., Park, D.-H., Omar, A., Nishizawa, T., et al. (2021). Assessing CALIOP-Derived Planetary Boundary Layer Height Using Ground-Based Lidar. *Remote Sensing*, 13(8). Retrieved from doi:10.3390/rs13081496
- Krishnamurthy, R., Newsom, R. K., Berg, L. K., Xiao, H., Ma, P.-L., and Turner, D. D. (2021). On the estimation of boundary layer heights: a machine learning approach. *Atmospheric Measurement Techniques*, 14(6), 4403-4424.
- 770 Kumar, A., Singh, N., Anshumali, and Solanki, R. (2018). Evaluation and utilization of MODIS and CALIPSO aerosol retrievals over a complex terrain in Himalaya. *Remote Sensing of Environment*, 206, 139-155.
- LeMone, M. A., Angevine, W. M., Bretherton, C. S., Chen, F., Dudhia, J., Fedorovich, E., et al. (2019). 100 Years of Progress in Boundary Layer Meteorology. *Meteorological Monographs*, 59, 9.1-9.85.
- 775 Li, Y., He, J., Ren, Y., and Wang, H. (2025). Aerosol-PBL relationship under diverse meteorological conditions: Insights from satellite/radiosonde measurements in North China. *Atmospheric Research*, 321, 108125. <https://www.sciencedirect.com/science/article/pii/S0169809525002170>
- 780 Li, Y., Li, J., Xu, S., Li, J., He, J., and Huang, J. (2023). Diurnal variation in the near-global planetary boundary layer height from satellite-based CATS lidar: Retrieval, evaluation, and influencing factors. *Remote Sensing of Environment*, 299.
- 785 Li, Y., Li, J., Zhao, Y., Lei, M., Zhao, Y., Jian, B., et al. (2021). Long-term variation of boundary layer height and possible contribution factors: A global analysis. *Science of The Total Environment*, 796.



- Li, Z., Guo, J., Ding, A., Liao, H., Liu, J., Sun, Y., et al. (2017). Aerosol and boundary-layer interactions and impact on air quality. *National Science Review*, 4(6), 810-833.
- Liu, B., Ma, Y., Liu, J., Gong, W., Wang, W., and Zhang, M. (2018). Graphics algorithm for deriving atmospheric boundary layer heights from CALIPSO data. *Atmospheric Measurement Techniques*, 11(9), 5075-5085.
- 790 Liu, F., Yi, F., Yin, Z., Zhang, Y., He, Y., and Yi, Y. (2021). Measurement report: characteristics of clear-day convective boundary layer and associated entrainment zone as observed by a ground-based polarization lidar over Wuhan (30.5° N, 114.4° E). *Atmospheric Chemistry and Physics*, 21(4), 2981-2998.
- 795 Liu, J., Huang, J., Chen, B., Zhou, T., Yan, H., Jin, H., et al. (2015). Comparisons of PBL heights derived from CALIPSO and ECMWF reanalysis data over China. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 153, 102-112.
- McGrath-Spangler, E. L., and Denning, A. S. (2012). Estimates of North American summertime planetary boundary layer depths derived from space-borne lidar. *Journal of Geophysical Research: Atmospheres*, 117(D15).
- 800 Medeiros, B., Hall, A., and Stevens, B. (2005). What Controls the Mean Depth of the PBL? *Journal of Climate*, 18(16), 3157-3172.
<https://journals.ametsoc.org/view/journals/clim/18/16/jcli3417.1.xml>
- Mei, L., Wang, X., Gong, Z., Liu, K., Hua, D., and Wang, X. (2022). Retrieval of the planetary boundary layer height from lidar measurements by a deep-learning method based on the wavelet covariance transform. *Optics Express*, 30(10).
- 805 Nguyen, P., Gite, R., Yang, Z., and Halem, M. (2021). *Deep Neural Network Architecture Search for Emulating Physical Parameterization of Planetary Boundary Layer Height*. Paper presented at the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS.
- 810 Nowottnick, E. P., Christian, K. E., Yorks, J. E., McGill, M. J., Midzak, N., Selmer, P. A., et al. (2022). Aerosol Detection from the Cloud–Aerosol Transport System on the International Space Station: Algorithm Overview and Implications for Diurnal Sampling. *Atmosphere*, 13(9).
- 815 Palm, S. P., Selmer, P., Yorks, J., Nicholls, S., and Nowottnick, E. (2021). Planetary Boundary Layer Height Estimates From ICESat-2 and CATS Backscatter Measurements. *Frontiers in Remote Sensing*, 2.
- Pan, S. J., and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345-1359.
- 820 Pearson, G., Davies, F., and Collier, C. (2010). Remote sensing of the tropical rain forest boundary layer using pulsed Doppler lidar. *Atmos. Chem. Phys.*, 10(13), 5891-5901.
<https://acp.copernicus.org/articles/10/5891/2010/>
- Peng, K., Xin, J., Zhu, X., Wang, X., Cao, X., Ma, Y., et al. (2023). Machine learning model to accurately estimate the planetary boundary layer height of Beijing urban area with ERA5 data. *Atmospheric Research*, 293.
- 825 Rieutord, T., Aubert, S., and Machado, T. (2021). Deriving boundary layer height from aerosol lidar using machine learning: KABL and ADABL algorithms. *Atmospheric Measurement Techniques*, 14(6), 4335-4353.



- Seidel, D. J., Ao, C. O., and Li, K. (2010). Estimating climatological planetary boundary
830 layer heights from radiosonde observations: Comparison of methods and uncertainty
analysis. *Journal of Geophysical Research: Atmospheres*, 115(D16).
- Seidel, D. J., Zhang, Y., Beljaars, A., Golaz, J. C., Jacobson, A. R., and Medeiros, B. (2012).
Climatology of the planetary boundary layer over the continental United States and
Europe. *Journal of Geophysical Research: Atmospheres*, 117(D17).
- 835 Sleeman, J., Halem, M., Yang, Z., Caicedo, V., Demoz, B., and Delgado, R. (2020). *A Deep
Machine Learning Approach for Lidar Based Boundary Layer Height Detection*.
Paper presented at the IGARSS 2020 - 2020 IEEE International Geoscience and
Remote Sensing Symposium.
- Stull, R. B. (1988). *An Introduction to Boundary Layer Meteorology*: Kluwer Academic
840 Publishers.
- Su, T., Li, J., Li, C., Xiang, P., Lau, A. K. H., Guo, J., et al. (2017). 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*, 122(7), 3929-3943.
- 845 Su, T., Li, Z., and Kahn, R. (2020). A new method to retrieve the diurnal variability of
planetary boundary layer height from lidar under different thermodynamic stability
conditions. *Remote Sensing of Environment*, 237, 111519.
<https://www.sciencedirect.com/science/article/pii/S0034425719305383>
- Su, T., and Zhang, Y. (2024). Deep-learning-derived planetary boundary layer height from
850 conventional meteorological measurements. *Atmospheric Chemistry and Physics*,
24(11), 6477-6493.
- Vogelezang, D. H. P., and Holtslag, A. A. M. (1996). Evaluation and model impacts of
alternative boundary-layer height formulations. *Boundary-Layer Meteorology*, 81(3),
245-269. <https://doi.org/10.1007/BF02430331>
- 855 Wang, Z., Cao, X., Zhang, L., Notholt, J., Zhou, B., Liu, R., and Zhang, B. (2012). Lidar
measurement of planetary boundary layer height and comparison with microwave
profiling radiometer observation. *Atmospheric Measurement Techniques*, 5(8),
1965-1972.
- Winker, D. M., Hunt, W. H., and McGill, M. J. (2007). Initial performance assessment of
860 CALIOP. *Geophysical Research Letters*, 34(19).
<https://doi.org/10.1029/2007GL030135>
- Yorks, J. E., McGill, M. J., Palm, S. P., Hlavka, D. L., Selmer, P. A., Nowottnick, E. P., et al.
(2016). An overview of the CATS level 1 processing algorithms and data products.
Geophysical Research Letters, 43(9), 4632-4639.
865 <https://doi.org/10.1002/2016GL068006>
- Yu, Y., Kalashnikova, O. V., Garay, M. J., Lee, H., Choi, M., Okin, G. S., et al. (2021). A
global analysis of diurnal variability in dust and dust mixture using CATS
observations. *Atmospheric Chemistry and Physics*, 21(3), 1427-1447.
- Yue, M., Wang, M., Guo, J., Zhang, H., Dong, X., and Liu, Y. (2021). Long-Term Trend
870 Comparison of Planetary Boundary Layer Height in Observations and CMIP6 Models
over China. *Journal of Climate*, 34(20), 8237-8256.



- Zhang, Y., Seidel, D. J., and Zhang, S. (2013). Trends in Planetary Boundary Layer Height over Europe. *Journal of Climate*, 26(24), 10071-10076. <https://journals.ametsoc.org/view/journals/clim/26/24/jcli-d-13-00108.1.xml>
- 875 Zhao, Y., Li, J., Zhang, L., Deng, C., Li, Y., Jian, B., and Huang, J. (2023). Diurnal cycles of cloud cover and its vertical distribution over the Tibetan Plateau revealed by satellite observations, reanalysis datasets, and CMIP6 outputs. *Atmos. Chem. Phys.*, 23(1), 743-769. <https://acp.copernicus.org/articles/23/743/2023/>