Atmospheric Research 293 (2023) 106925

Contents lists available at ScienceDirect

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmosres

# Machine learning model to accurately estimate the planetary boundary layer height of Beijing urban area with ERA5 data

Kecheng Peng [a,b], Jinyuan Xin [b,c,*], Xiaoqian Zhu [a,*], Xiaoyuan Wang [d], Xiaoqun Cao [a], Yongjing Ma [b], Xinbing Ren [b,c], Dandan Zhao [b], Junji Cao [b], Zifa Wang [b]

[a] College of Computer, National University of Defense Technology, Changsha 410000, China
[b] State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry (LAPC), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China
[c] College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[d] Zhejiang Ecological and Environmental Monitoring Center, Zhejiang, Hangzhou 310012, China

ARTICLE INFO

*Keywords:*
Planetary Boundary Layer Height
ERA5 reanalysis data
Hybrid Machine Learning model
Shapley Additive Explanations
Beijing urban area

ABSTRACT

The planetary boundary layer height (PBLH) is one of the most important parameters in the environmental, weather, and climatic research. Therefore, it is of great significance and application value to accurately estimate the PBLH by using the available conventional reanalysis meteorological datasets. This study established a hybrid machine learning (HML) method by combining three different ensemble algorithms with parallel training to estimate the urban PBLH of Beijing with ERA5 reanalysis data. The Mean Absolute Percentage Error (MAPE) between the ERA5-BLH and the observed BLH is as high as 68%, whereas it decreased to 10% for the PBLH estimated by the HML model. The radiation-related thermal factors of surface temperature and dew point temperature in ERA5, play a critical role in summer in regulating the PBLH, while the dynamic factors of wind speed and pressure dominate in the other seasons. The MAPE for all the seasons decreases by 0.3–1.6% after introducing the measured temperature and humidity profiles. Shapley Additive Explanations (SHAP) method shows that higher RH contributes less for PBLH estimated in spring, while 2 m temperature is positively correlated with HML performance in summer. Finally, an optimal HML model of PBLH is synthesized as the contribution of meteorological elements. The MAPE drops to 5.2%–15.1% throughout the year. The Mean Absolute Error (MAE) is <50 m in autumn and winter, and the maximum MAE is up to 80 m in the summer afternoon, when the convection is intensely developed. Therefore, the HML is capable of accurately estimating the urban PBLH in high resolution, which provides great significances and references for the investigations regarding the atmospheric environment capacity, as well as for advancing weather forecasting.

## 1. Introduction

The PBLH is defined as the physical height of the mixed layer that forms in the lowest region of the troposphere, influenced by dynamic factors such as wind speed and thermal factors such as radiation variables (Stull, 1988). PBLH has a pivotal role in evaluating the near-surface atmospheric pollutant diffusion model (Geiss et al., 2017; Li et al., 2018; Chen et al., 2022), the heat and momentum exchange between the earth's surface and the troposphere (Palmén and Newton, 1969), greenhouse gas concentration budgets (Gerbig et al., 2008) and numerical weather forecast process (Chen et al., 2011; Illingworth et al., 2019). PBLH structure exhibits diurnal variation, mainly influenced by the turbulence process that increases with solar radiation after sunrise, peaks at noon, and weakens during sunset (Mahrt, 1999). The regional-scale variation of PBLH is significantly affected by the terrain and the drastic variation of near-surface wind speed, temperature, and relative humidity, causing uncertainty. Therefore, a long-term PBLH estimation with a high spatiotemporal resolution is a major public problem (Trentmann et al., 2009). To improve the accuracy of PBLH estimation and capture continuous and long-term variation patterns, previous studies have utilized modern remote sensing devices. Radio soundings are the most commonly used instrument for estimating PBLH,

* Corresponding author at: State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry (LAPC), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China.
*E-mail address:* xjy@mail.iap.ac.cn (J. Xin).

https://doi.org/10.1016/j.atmosres.2023.106925
Received 21 May 2023; Received in revised form 5 July 2023; Accepted 13 July 2023
Available online 15 July 2023

but their sparse observations and limited spatial coverage make it difficult to use them in urban and regional research (Su et al., 2017). In recent decades, other remote sensing techniques, such as ground-based lidar systems, including Raman lidars (Turner et al., 2014), ceilometers (Haeffelin et al., 2012; Caicedo et al., 2017; Lee et al., 2019; Jiang et al., 2021), Doppler lidar (Berg et al., 2017; Manninen et al., 2018; Marques et al., 2018; Banakh et al., 2021), microwave radiometers (e.g. Bravo-Aranda et al., 2017; Jiang et al., 2021; Ma et al., 2022) and atmospheric emitted radiation interferometer (Knuteson et al., 2004; Sawyer and Li, 2013), have been used to drive and enhance the spatiotemporal detection ability of PBLH. Compared with these remote sensing instruments, ceilometers have the advantages of low cost and high accuracy, and the ability to detect aerosols and clouds (e.g. Caicedo et al., 2017; Moreira et al., 2020; Moreira et al., 2022), although their performance may be reduced when monitoring in clouds, rain, and dust weather conditions or at high altitude (Kotthaus et al., 2023). However, thermal profile remote sensing instruments, such as microwave radiometers (MWR) (Cimini et al., 2006), is limited to low vertical resolution, while their vertical temperature gradient and water vapor detection are crucial for PBLH estimation (Saeed et al., 2016). Combining the MWR profile with the ceilometer data can improve the quality and accuracy of PBLH estimation and provide a more detailed understanding of the physical process.

The advent of the big data era and the abundance of computing resources have made Machine Learning (ML) models increasingly important, as traditional numerical methods require significant computational power to analyze massive meteorological data. ML algorithms have the advantage of combining contributions from multiple features (McGovern et al., 2017), and many of them have shown strong generalization ability and high predictive performance in the atmospheric field. ML models have been widely applied in atmospheric remote sensing (Cadeddu et al., 2009; Wei et al., 2019; Moreira et al., 2022; Muñoz-Esparza et al., 2022), atmospheric chemistry (Chen et al., 2018; Li et al., 2021a, 2021b; Wei et al., 2021; Chen et al., 2022), atmospheric phenomenon classification (McGovern et al., 2017; Gagne et al., 2017), and climate prediction (Ham et al., 2019). Notable research works include the spatiotemporal LightGBM model, which performs well in predicting $PM_{2.5}$ pollution and addresses the challenges of high complexity and large sample size (Wei et al., 2021). This ML model can also process large-scale data with high accuracy and low memory consumption. Liu et al. (2022) have also used ML methods to reconstruct a non-grid ground wind speed field based on meteorological background fields and geographic information, which can improve the historical ground wind speed field at any grid resolution. These ML methods have important applications in wind energy and aviation safety assessments. Additionally, traditional ML models such as random forest (RF) and neural networks have been used to estimate PBLH in recent years, achieving remarkable results (e.g. Krishnamurthy et al., 2021; Molero et al., 2022). Molero et al. (2022) employed ML techniques to estimate the mixing layer heights (MLH) and PBLH across various seasons and periods, introducing a methodology that enables direct estimation of these parameters from ground-level meteorological data. Moreira et al. (2022) utilized gradient boosting regression trees to achieve a precise estimation of the PBLH and discussed the impact of different weather conditions on the model estimation. However, these researches lack comparative analysis with publicly accessible reanalysis data, quantification of feature importance, and in-depth analysis of seasonal factors.

As a mega-city and central city of China, Beijing's air pollution incidents become more severe in recent years due to the city's rapid modernization process. However, there is still a lack of relevant research on estimating PBLH with high spatiotemporal resolution in the urban area of Beijing (Ma et al., 2022; Xin et al., 2023). PBLH of ERA5 reanalysis data has been proven to be the closest dataset to the observation profile (Guo et al., 2021), while it still suffers from low spatiotemporal resolution and poor accuracy. Considering the defects in the above dataset and the advantages of the model, this paper explores the potential of using the Hybrid Machine Learning (HML) method to compensate for the data deficiency. The combination of various meteorological datasets from ERA5 reanalysis data and MWR profiles allows us to estimate the local PBLH of Beijing with greater accuracy. And the estimates are comparable to the results detected by ceilometers. The purpose of this study is to demonstrate how HML methods can be used to rapidly, robustly, accurately, and automatically estimating local PBLH with high spatiotemporal resolution, overcoming the limitations of previous research.

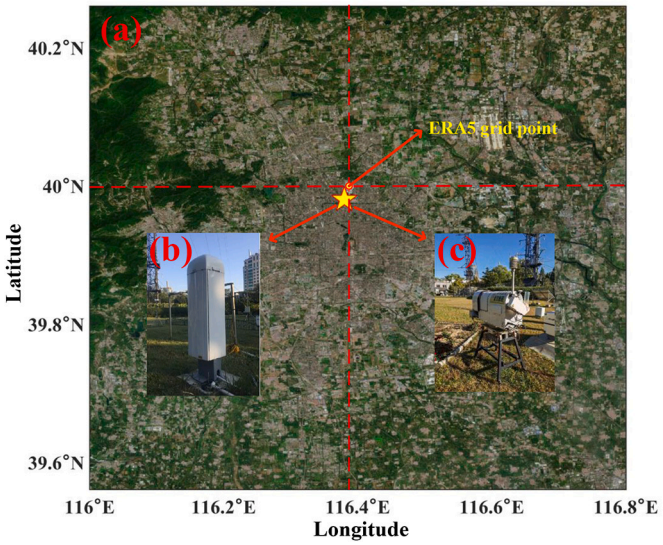


**Fig. 1.** Location of the observation station and key instruments used in this work. (a) represents the geographical location of Beijing urban area, the position marked by a five-pointed star represents the location of the observation station, where placed the main remote sensing instruments used in this study, (b) is ground-based multi-channel microwave radiometer (MWR) RPG-HATPR-G5 (temperature and humidity profiles), and (c) is Vaisala CL51 ceilometer. The position of the circular dots represents the grid point of the ERA5 reanalysis data selected in this study.

## 2. Data source

The observation site of this study is the comprehensive observation station of the Institute of Atmospheric Physics (IAP) of the Chinese Academy of Sciences in the Beijing urban area (40° N, 116.4° E), as indicated by the five-pointed star marked in Fig. 1 (a). The Vaisala CL51 ceilometer based on laser detection and ranging (Lidar) technology is used to retrieve PBLH as a label for supervised learning training. As shown in Fig. 1 (b), the ceilometer has a vertical resolution of 10 m, and a time resolution of 15–16 s (Zhu et al., 2018; Mues et al., 2017). We collected the PBLH driven from ceilometer in 2018 and 2019. Additionally, temperature and humidity profile data collected by the RPG-HATPR-G5 microwave radiometer (MWR) during the same period are utilized. The MWR, shown in Fig. 1(c), offers full atmospheric coverage of up to 93 layers in the vertical direction, with a temporal resolution of 1 s (Zhao et al., 2019). Since the PBLH is basically within 0–3 km, the temperature and RH profiles of 60 layers within 0–3 km from MWR are selected, with a total of 120 independent meteorological variables. Quality control measures are applied to both types of observation data to remove any abnormal or missing values.

The global atmospheric reanalysis data of the Fifth Generation European Medium-Range Weather Forecast Center (ERA5) provides the highest spatial resolution and most comprehensive reanalysis data. It includes atmospheric and terrestrial surface meteorological elements collected every hour since 1979, with 0.25° × 0.25° horizontal resolution. The land version of the ERA5 data (ERA5-land) has an even higher horizontal resolution of 0.1° × 0.1°. Both data have a time resolution of

**Table 1**
The instruments and ERA5 reanalysis data used in this study and their corresponding meteorological variables.

| Instrument/Data | Variable (Abbreviation) | Range |
|---|---|---|
| Ceilometer (Vaisala CL51) | Boundary Layer Height | 100–3000 m |
| Microwave Radiometer (RPG) | Temperature (0–3 km) | 259.9–306.6 K |
| | Humidity (0–3 km) | 0–100% |
| Time | Hour | Categorical Variable |
| | Day | Categorical Variable |
| | Month | Categorical Variable |
| | Season | Categorical Variable |
| ERA-5 Reanalysis Data | Boundary Layer Height (BLH) | 10.2–3678.0 m |
| | 2 m Temperature (T) | 264.7–300.2 K |
| | 2 m Dewpoint Temperature (DT_2m) | 248.5–292.4 K |
| | Relative Humidity (RH) | 0–100% |
| | 10 m U (10 U) | −3.6–4.1 m $s^{-1}$ |
| | 10 m V (10 V) | −5.5–3.9 m $s^{-1}$ |
| | Surface Pressure (SP) | 868.89–897.99 hPa |
| | Skin temperature (Skin_T) | 255.7–304.5 K |
| | Total Evaporation (TE) | 0–0.0049 m |
| | Total Precipitation (TP) | 0–0.0252 m |
| | Forecast Albedo (dimensionless) | 0.093–0.598 |
| | Surface Sensible Heat Flux (Surface_SF) | −308.1–4681.9 KJ $m^{-2}$ |
| | Surface Thermal Radiation Downwards (SRRD) | 683.4–34,633.8 KJ $m^{-2}$ |
| | Surface Latent Heat Flux (LF) | −2.1–12,166.1 KJ $m^{-2}$ |
| | Surface Solar Radiation Downwards (SSRD) | 52.9–24,671.1 KJ $m^{-2}$ |
| | Surface Net Solar Radiation (SNSR) | 46.9–21,800.7 KJ $m^{-2}$ |
| | Surface Net Thermal Radiation (SNTR) | 5–9484 KJ $m^{-2}$ |
| | Leaf area index, High vegetation (Lai_High) | 2.33–2.47 $m^2$ $m^{-2}$ |
| | Leaf area index, Low vegetation (Lai_Low) | 1.18–2.66 $m^2$ $m^{-2}$ |

1 h. Both versions of the data have a temporal resolution of 1 h. In this study, 18 meteorological variables including 2 m Temperature (Tem), Relative Humidity (RH), 10 m U and V wind components (U10 and V10), Surface Pressure (SP), Total Precipitation (TP), etc. are selected for PBLH estimation. Previous studies have demonstrated that the variation of these meteorological variables has a strong correlation with PBLH and significantly impacts its occurrence and development process (Haeffelin et al., 2012; Wang et al., 2016; Allabakash and Lim, 2020; Krishnamurthy et al., 2021; Rey-Sanchez et al., 2021). In addition, the PBLH dataset of ERA5-land (ERA5-BLH) is used to compare with model estimates for the same period as the observational data. Note that all variables and abbreviations are shown in Table 1.

## 3. Proposed Hybrid machine learning (HML) model of urban atmospheric boundary layer

This study focuses on estimating the PBLH in urban areas of Beijing, and an HML model is proposed by combining data pre-processing and multiple prediction methods. Firstly, due to the mismatch of temporal and spatial resolution among ceilometer, MWR, and ERA5 reanalysis data, and the relatively minor variation of PBLH within seconds, a time resolution of 10 min is selected to match the three data sources. For ceilometer and MWR data, a 10-min average is directly calculated during data processing, and ERA5 and ERA5-land reanalysis data are matched to the corresponding instrument observation time through bilinear interpolation. Since the observation station locates in the urban area of Beijing, the grid point closest to the observation station's longitude and latitude is selected to avoid the influence of terrain, i.e., the selected grid point for ERA5-land is 40°N and 116.4°E, and for ERA5 is 40°N and 116.5°E. Considering the differences in observation time and the impact of missing values, a total of 18,230 data samples are obtained. When constructing the model, an hourly variable (Hour) is added to analyze the outputs over different periods. Additionally, due to the strong daily and seasonal variations of PBLH in different regions (Gallée et al., 2015; Singh et al., 2016), previous studies have shown that dividing research data into different periods and seasons can enhance the interpretability of ML models (Krishnamurthy et al., 2021; Moreira et al., 2022; Molero et al., 2022). Therefore, we divide the dataset into four groups, each representing a season. Note that the day, month, and season variables are used as auxiliary variables, not as separate feature inputs of the model. In summary, we constructed the entire dataset using 138 meteorological variables and a time variable as feature variables. The input features for the HML model have a final dimension of (18,230, 139), and the prediction label is (18,230, 1).

The radiant flux values of ERA5 are often $10^3$–$10^4$ magnitude higher than that of precipitation and temperature in the constructed dataset, and their distribution is different between day and night. This difference can bias the number of decision trees in the ML model when the HML constructs the estimation process for different periods (Krishnamurthy et al., 2021). To reduce the impact of data magnitude and computational costs on the model, all feature variables are standardized before training and then denormalized after training. The dataset is then divided into an 80% training set and a 20% test set, with the validation set accounting for 10% of the training set data. The distribution of the divided data is shown in Table S1, where the number of samples is the largest in autumn, and the sum of training samples and test samples accounts for 41% of the total samples. Summer samples accounted for only 4% of the total sample since the overlap rate of the covering periods of MWR and ceilometer observations is low. However, we reduce the influence of sample unevenness by adjusting hyperparameter parameters and cross-validation.

To prevent model overfitting and data overload caused by inputting multi-layer MWR temperature and humidity data into the model, as well as to reduce the model complexity, this study adopts a feature selection method to select the most important layer data of MWR as input variables. The feature selection process is primarily based on the impact of each feature on the model accuracy, selecting the most relevant features from the initial data set (Guyon and Elisseeff, 2003). To reduce the number of the input feature variables from MWR, this study utilizes the Recursive Feature Elimination (RFE) method to analyze the importance of each feature. A predetermined model is trained on the original dataset, and the least important MWR temperature and humidity features are discarded, after which the model is adjusted with each iteration (Yu and Liu, 2003). In this study, the discarded features must ensure the decrease in model correlation is <2% and the increase in MAE is <5 m at each iteration. Finally, the temperature and humidity data at 100 m and 3 km from MWR are chosen as the input variables for HML. After feature selection, the input feature dimension is reduced to (18,744, 23).

To leverage the benefits of different ML methods, this study utilizes ensemble algorithms and parallel training as the foundation to combine three different ML methods for PBLH estimation. Specifically, the Random Forest (RF) algorithm utilizes the bootstrap sampling technique to randomly select n individual samples, each containing all the necessary meteorological variables. This process is repeated T times to generate T sample datasets, represented by different colored squares in Fig. 2. Each sample dataset is then trained on a basic learner (Rote learner), and the resulting basic learners are aggregated to obtain the average value of estimated PBLH. Although the RF algorithm is characterized by high classification and prediction accuracy, fast convergence, strong generalization ability, and robustness (Breiman, 2001), it is time-consuming and lacks iterative improvement. To enhance efficiency, this study also employs parallel training and incorporates the tree-based LightGBM regression algorithm (Ke et al., 2017). The

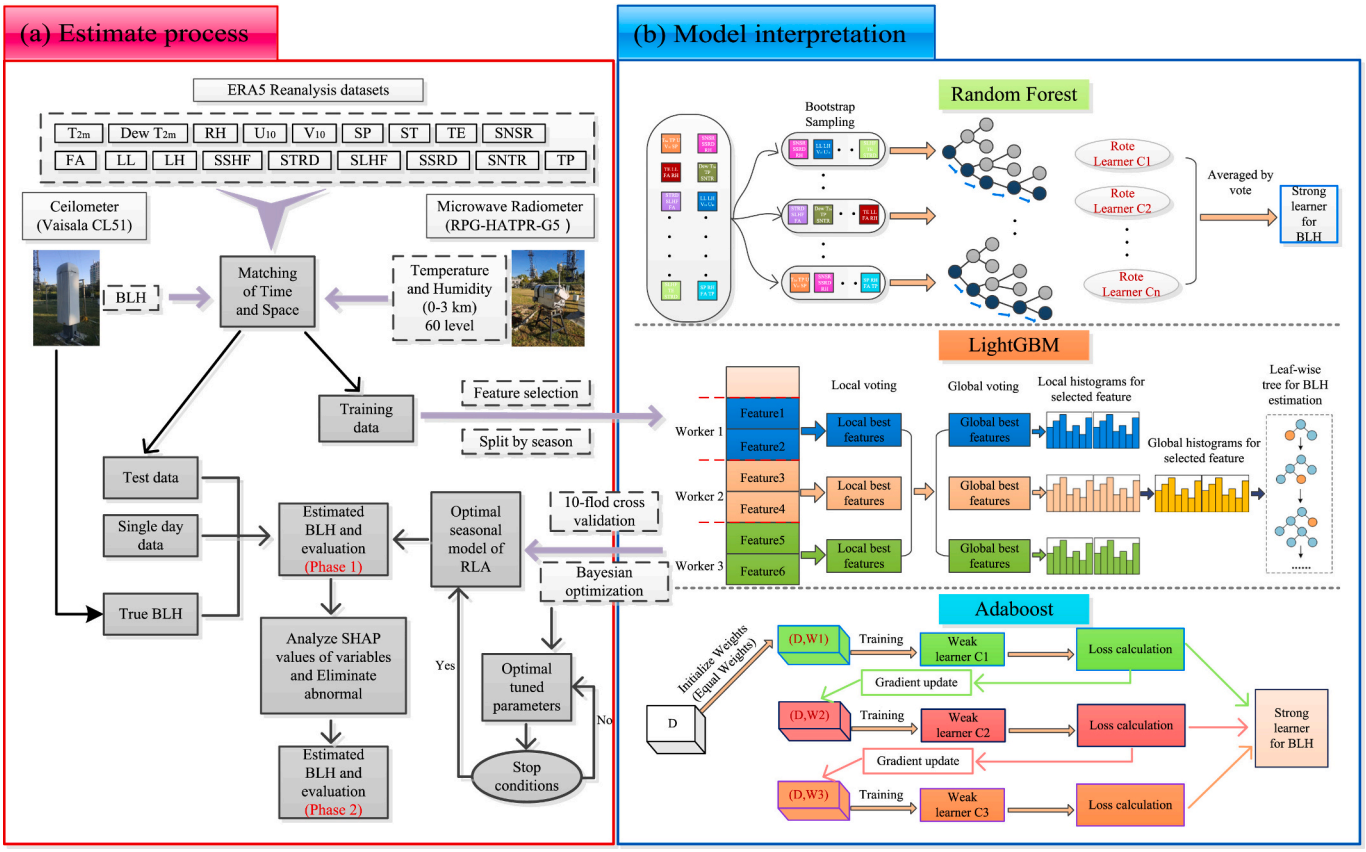


**Fig. 2.** Proposed Hybrid Machine Learning algorithm framework for PBLH estimation, including estimate process module (a) and model interpretation module (b).

histogram algorithm traverses the training data to find the optimal split point based on discrete values in LightGBM, significantly reducing the model time complexity from O(data*feature) to O(K*feature). During the PBLH estimation, the leaf-wise optimization strategy is adopted, as shown in the LightGBM module of Fig. 2. The algorithm selects the leaf with the highest split gain, which typically contains the largest amount of dataset, and then split it at each iteration. This process repeats until the desired accuracy is achieved. The leaf-wise optimization strategy can reduce errors and achieve better accuracy. To prevent overfitting and achieve better accuracy, we add a maximum depth constraint to the leaf-wise optimization strategy. However, since the above algorithms use a direct prediction process without iterative updates to the model learner, we introduce the Adaptive Boosting (AdaBoost) algorithm (Freund and Schapire, 1997; Rieutord et al., 2021) to correct deviations generated by RF and LightGBM in the HML model. As shown in Fig. 2 of the model explanation, the AdaBoost module initializes with a sample set of m samples, where each training sample point shares the same weight of 1/m. The initial sample set and weights are then used for the first weak learner training, and the AdaBoost gradient is adjusted after calculating the training loss, adaptively updating the weight of the training sample until the specified maximum iteration number is reached or the predetermined error rate is achieved. Finally, the various weak learners trained are integrated to form a powerful learner, where those with greater importance scores are assigned a greater weight in the final prediction outcomes.

To adjust hyperparameters and evaluate the stability of the HML model, we utilized Bayesian optimization and five-fold cross-validation methods. Since tree depth (max_depth) and the number of regression trees (n_estimators) are commonly used hyperparameters in ML modeling, we optimized both in this study, as suggested by Breiman (2001). Bayesian optimization experiments demonstrate the HML achieves optimal performance and computational efficiency when the max_depth and estimators are set to 25 and 500, respectively. Subsequently, we obtain the optimal solution through parallel training based on the parameter settings of the ensemble algorithm, and allocate it to different seasonal models.

After completing basic training, the interpretability of the HML is investigated using SHAP (Shapley Additive Explanations). The primary concept behind SHAP is to utilize the size of the Shapley value to quantitatively assess the various elements of the ML model. (Lundberg and Lee, 2017; Lundberg et al., 2018). The SHAP method evaluates the contribution of a feature to the estimated value of the HML model using coalitional game theory, thereby allocating a quantitative index of importance to the feature. The HML model developed in this study encompasses analyses of each feature variable as well as individual anomalous samples. Subsequently, the HML model will be optimized by eliminating the impact of outliers and readjusting the hyperparameters, and case studies are conducted based on continuous samples and a complete observation period in a single day.

Finally, the PBLH obtained from the ceilometer ($ABLH_{true}$) with the PBLH estimated by the HML model ($ABLH_{predict}$) and ERA5-BLH are compared to demonstrate the estimation performance of the HML model intuitively. Root Mean Square Error (RMSE) is used as one of the evaluation metrics.:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(ABLH_{estimate_i} - ABLH_{true_i}\right)^2}$$

Where $N$ represents the number of input samples, $ABLH_{predict_i}$ represents the estimated PBLH value of the $i$th sample, $ABLH_{true_i}$ represents the true PBLH value of the $i$th sample. The sensitivity of RMSE to outliers may cause non-robust in some cases. Hence, additional evaluation metrics such as mean absolute error (MAE) and mean absolute percentage error (MAPE) are introduced to quantify model errors:

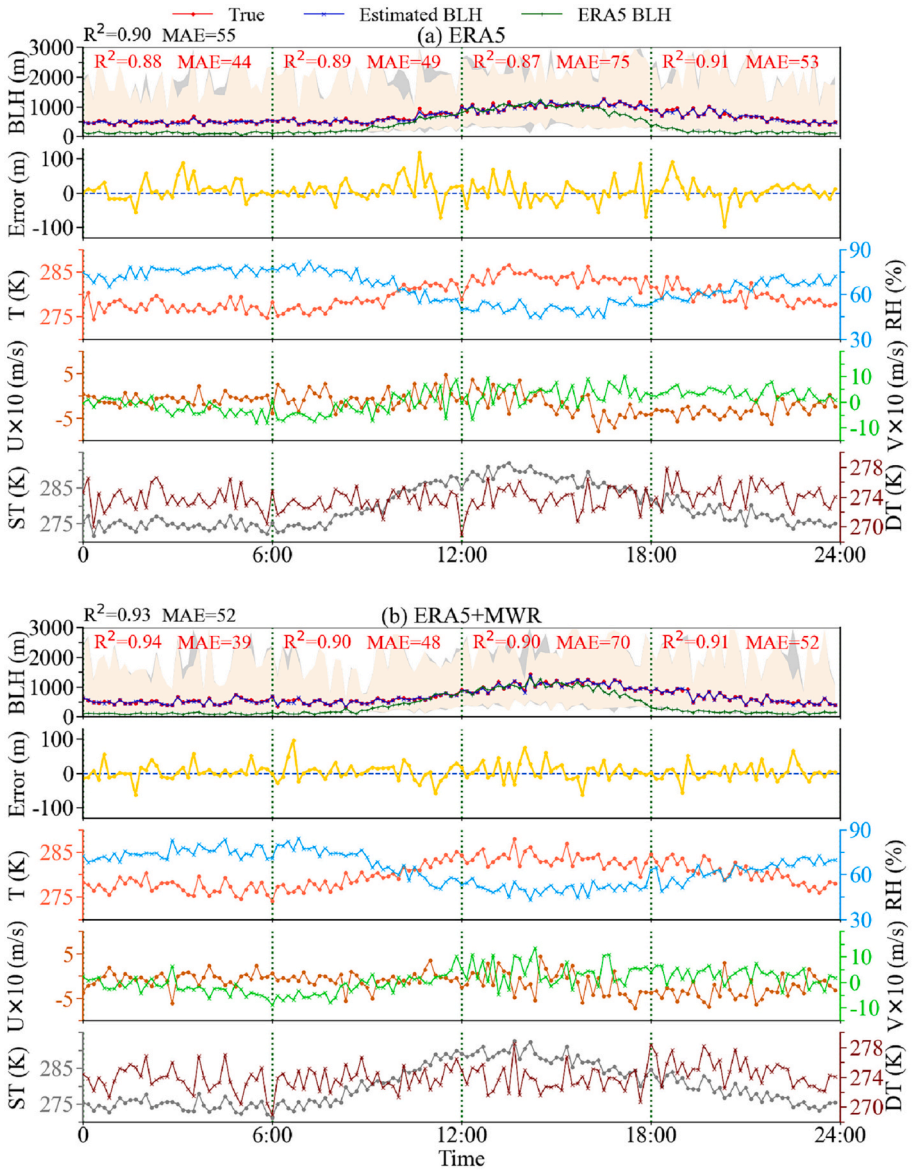


**Fig. 3.** Time series of observed and estimated Beijing urban PBLH, the error between them, ERA5-BLH, and corresponding meteorological variables including ERA5-temperature, ERA5-RH, ERA5-10 m U, ERA5-10 m V, ERA5-skin temperature, ERA5-2 m dewpoint temperature. (a) represents the estimation results using ERA5 reanalysis data, while (b) represents the estimation results after incorporating the MWR data.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|ABLH_{estimate_i} - ABLH_{true_i}|$$

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{ABLH_{estimate_i} - ABLH_{true_i}}{ABLH_{true_i}}\right|$$

In order to evaluate the fitting degree and generalization ability of the HML model, the coefficient of determination ($R^2$) is also calculated:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(ABLH_{true_i} - ABLH_{estimate_i})^2}{\sum_{i=1}^{N}(ABLH_{true_i} - \overline{ABLH_{estimate}})^2}$$

Where $\overline{ABLH_{predict}}$ represents the average value of PBLH estimated by the HML model.

## 4. Results and discussion

### 4.1. Comparison of ERA5-BLH and HML estimation

Figure 3(a) shows the daily variations of estimated PBLH, true values, and various variables using only ERA5 reanalysis data, while (b) shows the improved results after incorporating MWR data. The daily variation of the annual average PBLH in Beijing shows a gradually increase from sunrise (6:00–8:00), reaching the maximum in the afternoon (12:00–17:00) with significant fluctuations during this phase, followed by a decrease during the sunset period (17:00–18:00) and relatively stable and low PBLH development at night, which is consistent with the delayed pattern of solar radiation during the day (Moreira et al., 2020). Ground thermal radiation and turbulence mixing during the period of 12:00–18:00 CST significantly affect PBLH estimation using only ERA5 data, resulting in significant deviations at certain moments. The HML model cannot accurately capture the impact of feature variations on PBLH. After adding MWR observation data, the HML model takes the importance of features and the combined characteristics of various

**Table 2**
The quality indicators of the HML model with and without selected MWR profiles for PBLH estimation split by season.

| | Evaluation | All | Spring | Summer | Autumn | Winter |
|---|---|---|---|---|---|---|
| ERA5 | MAPE (%) | 67.8 | 77.6 | 60.4 | 60.2 | 74.8 |
| HML with | $R^2$ | 0.90 | 0.90 | 0.94 | 0.86 | 0.92 |
| ERA5 | MAE (m) | 55 | 77 | 65 | 56 | 53 |
| | RMSE (m) | 160 | 198 | 151 | 186 | 138 |
| | MAPE (%) | 9.6 | 8.7 | 13.3 | 10.3 | 9.2 |
| **HML with** | **$R^2$** | **0.93** | **0.96** | **0.96** | **0.89** | **0.92** |
| **ERA5 + MWR** | **MAE (m)** | **52** | **60** | **60** | **47** | **49** |
| | **RMSE (m)** | **136** | **141** | **105** | **153** | **128** |
| | **MAPE (%)** | **8.6** | **7.9** | **13.0** | **8.7** | **8.8** |

variables into account, which can reduce errors during abnormal periods. The findings presented in Table 2 demonstrate that the utilization of diverse meteorological variables and the HML model substantially improves PBLH estimation accuracy, resulting in an error rate as low as 10% compared to the error rate of about 70% observed when using ERA5-BLH. In addition, including MWR temperature and humidity data in the model significantly enhances PBLH estimation accuracy in all seasons, with an average increase of 0.03 in $R^2$, a reduction in MAPE of 0.3–1.6%, and a decrease in MAE by 4–17 m. The average RMSE in spring and autumn are reduced by 58 m and 32 m, respectively. These statistical results suggest that the proposed HML has the potential to enhance the precision of PBLH estimations.

### 4.2. Assessment of feature importance

Further investigation demonstrates that the effect of meteorological variables on PBLH estimation varies significantly among different time periods. As depicted in Fig. 3, the accuracy of the HML model declines when the temperature is relatively high, the humidity is relatively low, and the 10 m horizontal wind speed is relatively high along with drastic variations. The variation of dew point temperature shows a weak correlation with the model estimation error. Upon further analysis, it shows a relatively large difference between the PBLH data and the true PBLH data in the afternoon, while smaller errors occur during other periods. Consequently, the average deviation of PBLH estimation is <60 m at almost all times. The HML model solely relying on reanalysis exhibits notable deviations around 11:00 and 18:00 CST, which can be attributed to the intensified influence of ground heat radiation and turbulence mixing during the sunrise and sunset periods. This influence causes PBLH to undergo drastic variations, and the HML model fails to accurately capture the impact of feature variations on PBLH during this specific time frame.

Figure 4 presents scatterplots of PBLH estimates for all samples obtained using the LightGBM model in each season, providing a visual and impartial comparison of the results obtained from parallel training of the HML models. The PBLH estimates obtained from the LightGBM model exhibit a high level of consistency with the true values derived from the ceilometer, with $R^2$ values of 0.92, 0.96, 0.92, and 0.93 for the four seasons. Despite the relatively low density of summer samples, the correlation coefficient of 0.96 demonstrates the strong fitting ability of LightGBM, accurately fitting parameters with minimal training samples. The higher number of samples in autumn and winter enhances the generalization ability of LightGBM, resulting in significant improvements in MAE values of 65 m and 72 m, respectively, compared to 91 m in spring. Outliers are present in 13 cases in autumn and only 1 in summer, and the number of outliers will increase with the total samples of each season. These results indicate that the LightGBM model exhibits excellent estimation performance in all seasons and can utilize

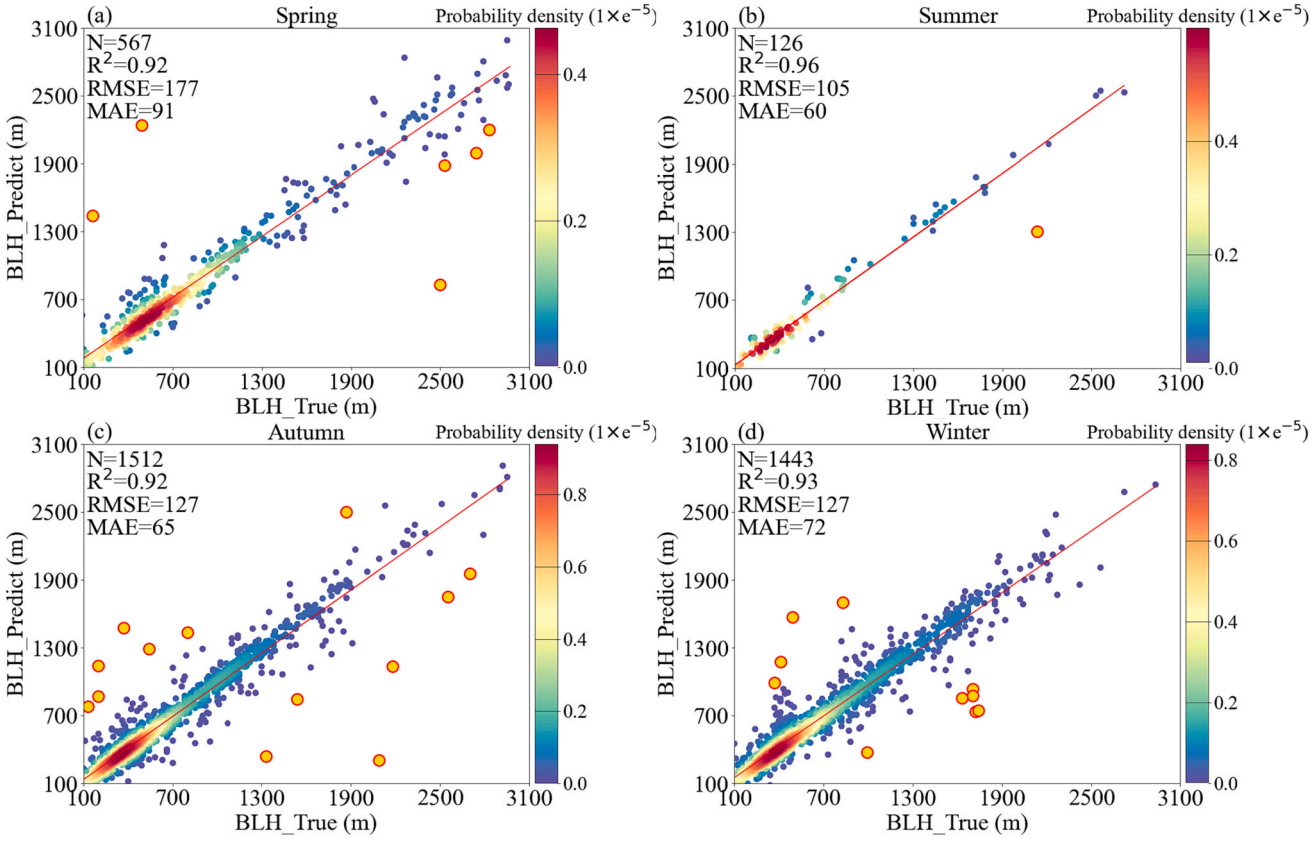


**Fig. 4.** Scatterplots of LightGBM estimated PBLH and corresponding histogram of feature importance in each season, (a) - (d) represent the results of spring, summer, autumn, and winter, respectively. The red solid line represents the fitted line, and the yellow dots represent the samples where the estimated PBLH differs from the observed value by >600 m. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

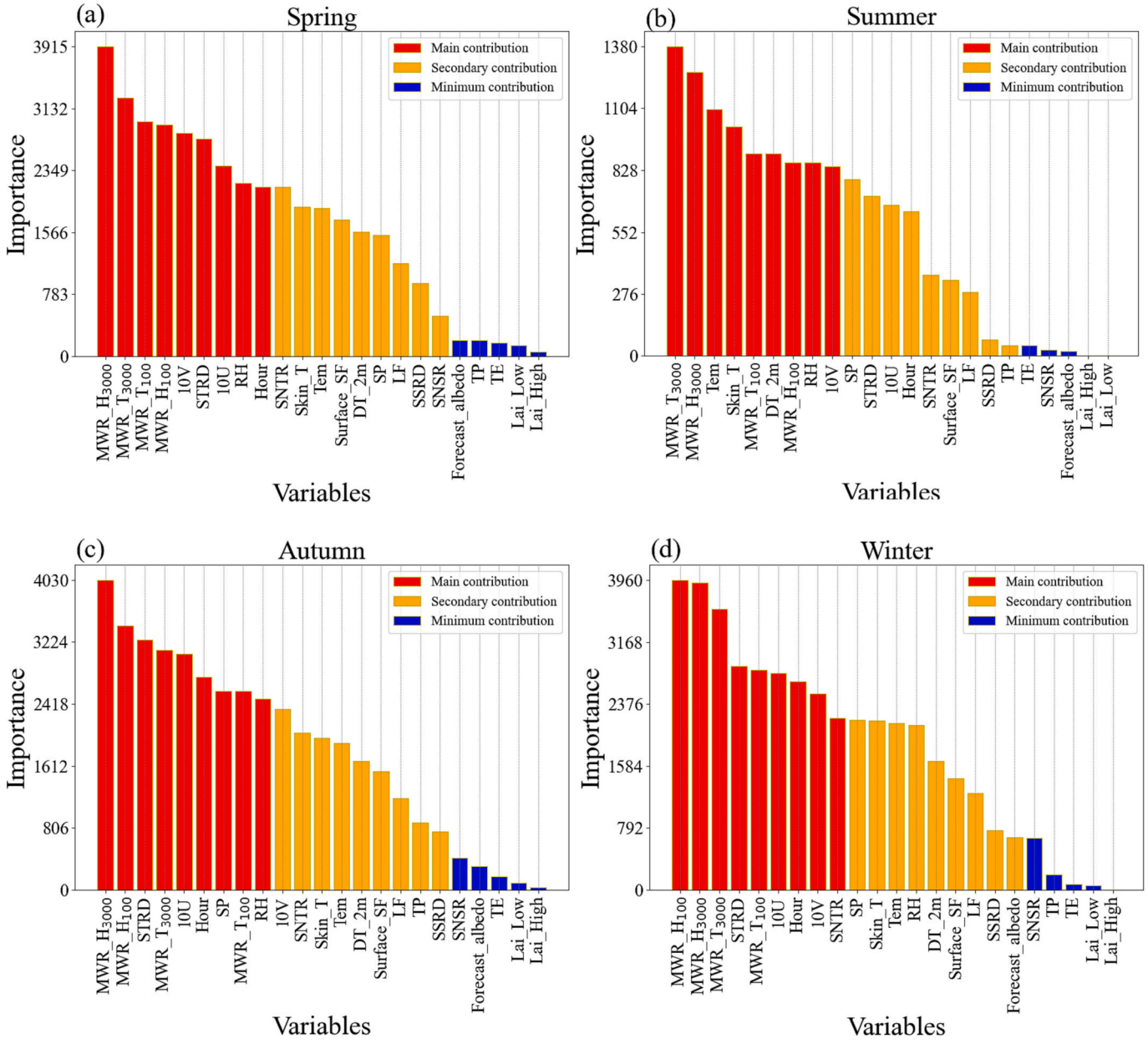


**Fig. 5.** Histogram of feature importance in each season, (a) - (d) represent the results of spring, summer, autumn, and winter, respectively.

meteorological variables to describe high spatiotemporal resolution PBLH.

The feature importance diagram in Fig. 5 indicates that the MWR temperature and RH at the height of 100 m and 3 km make significant contributions to the model throughout all seasons. However, the importance of ERA5 meteorological variables varies in different seasons. In summer, strong turbulent mixing and vigorous convection in the boundary layer make temperature a driving force resulting in more pronounced variations in PBLH (Ma et al., 2020; Jiang et al., 2021; Zhao et al., 2021). As a result, surface temperature, 2 m temperature, and dew point temperature, which are associated with ERA5 radiation variables, play a more crucial role in PBLH estimation during this season. In the other three seasons, relative humidity, 10 m wind speed, and surface pressure have greater significance due to the lower ground temperature and the weaker convective activity (Guo et al., 2016; Li et al., 2021a, 2021b; Jiang et al., 2021). During these seasons, thermal factors related to radiation occupy a less important position, and PBLH variations are more closely linked to dynamic factors.

### 4.3. Quantitative analysis of feature contribution

Most previous studies have explored the impact of ML models on PBLH estimation in different periods and seasons, as well as the overall importance of individual features on the estimation results (Moreira et al., 2022; Molero et al., 2022). However, the quantification of the contributions of various meteorological variables requires further investigation. In this study, SHAP values are employed to quantify the influence of meteorological variables on the PBLH estimation of the HML model.

The SHAP value feature density scatter plots for seasonal models are presented in Fig. 6. It is apparent that ERA5 RH plays a crucial role in precise PBLH estimation during spring, which is consistent with the conclusion drawn from Fig. 5. Additionally, a higher RH is associated with a more negative SHAP value, resulting in a decrease in accuracy. Similarly, elevated net surface radiation and decreased vegetation index positively impact the HML model estimation, and vice versa. In summer, dynamic factors, such as 10 m U and 10 m V, as well as thermal factors,

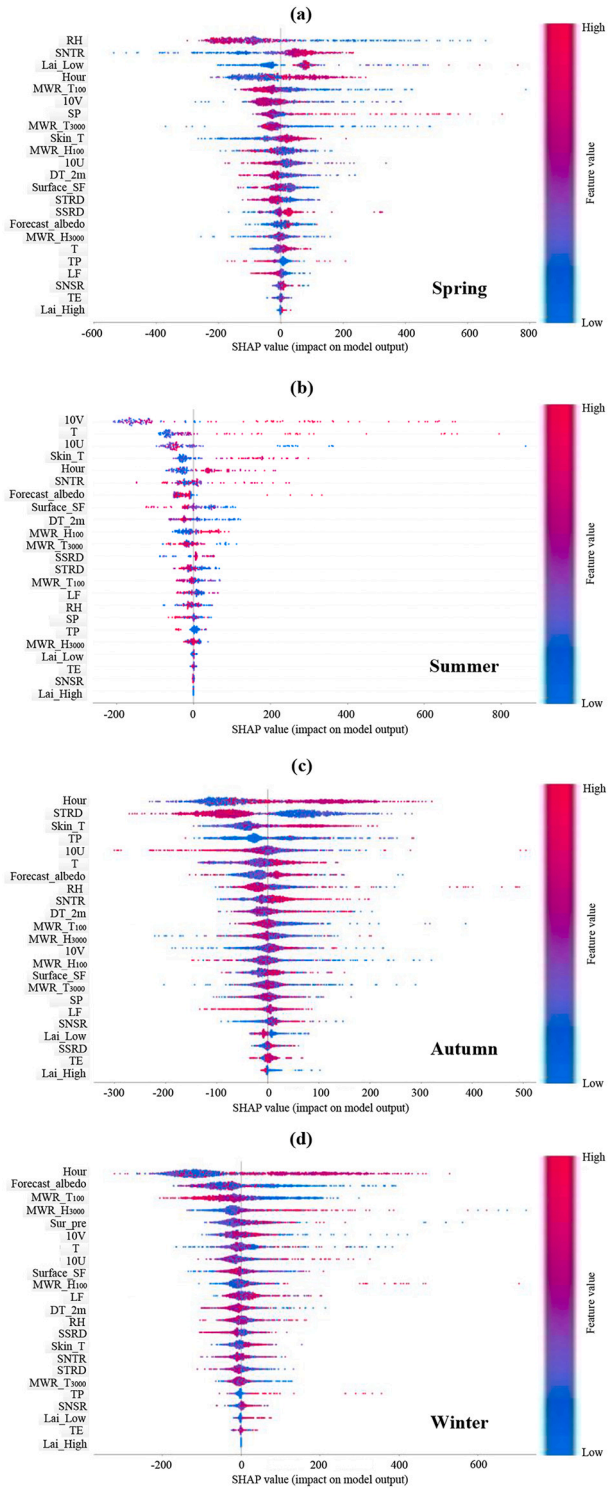


**Fig. 6.** The SHAP interpretable characteristic density scatterplots of each feature, from (a) to (d) represents the spring, summer, autumn, and winter samples, respectively. Each scatterplot represents the ranking of feature importance from top to bottom, and the width represents the concentration of the samples.

such as 2 m temperature and surface temperature, have a significant impact on PBLH estimation, with an elevated 2 m temperature leading to higher accuracy. However, the impact of RH is relatively insignificant. The hour variable, when treated as a categorical variable, exhibits positive gains in autumn and winter when the value is higher (i.e., during the afternoon to evening period), while negative gains are observed during the early morning to noon. It is worth noting that downward surface thermal radiation (STRD) exhibits notable differences in autumn, with lower STRD values contributing more to the model estimation. These quantified indicators confirm that the contributions of various meteorological variables to PBLH estimation vary significantly across different seasons, providing valuable guidance for the accurate estimation of PBLH in diverse weather processes and periods.

In addition, SHAP method are utilized in this study to quantify the respective contributions of various meteorological features to the estimated PBLH. As illustrated in Fig. 7, four outliers are selected during autumn to examine the impacts of anomalous estimation. The findings reveal that precipitation, radiation variables, and the associated temperature and humidity have a positive impact on the PBLH estimation in most cases. Conversely, other meteorological factors, such as wind speed and low vegetation index, are found to have negative feedbacks, leading to a deviation of the estimated value from the actual value.

### 4.4. Seasonal HML model performance and optimization

In addition to the variations in meteorological variables, errors in the model estimation may also arise from observation methods and extreme weather events. Through analysis of the periods of individual anomalous points, this study has found that these outliers often occur during sudden increases or decreases in PBLH, ranging from several hundred meters to over 2 km. Due to the backscattering coefficient of the aerosol detected by the ceilometer, the detected PBLH may undergo drastic variations when encountering sandstorms or strong winds that carry large amounts of aerosols or pollutants. Therefore, further optimization of the HML model method is necessary during such weather mutations or long-distance transmission of aerosols.

In this study, the periods during which outliers are identified as sudden variations will be excluded. Subsequently, the HML model will be retrained using the remaining data to optimize the hyperparameters. The evaluation results of ERA5-BLH, pre-optimization, and post-optimization are compared in Table 3. The results demonstrate that the HML model provides more precise and realistic estimates of PBLH than ERA5-BLH for different seasons and periods. Specifically, the overall errors of ERA5 in the four seasons are 77.6%, 60.4%, 74.8%, and 60.2%, respectively. This indicates that the HML model outperforms ERA5 in PBLH estimation. Moreover, consistent with the results depicted in Fig. 3, ERA5-BLH exhibits lower MAPE during the afternoon period in summer, autumn, and winter compared to the other three periods. In particular, the summer afternoon period shows a significantly reduced MAPE of 18.2%. This suggests that ERA5 reanalysis data can effectively capture the variations in PBLH during the afternoon period in urban areas of Beijing.

The findings of the study demonstrate that the optimized HML model significantly improves the metrics of $R^2$, MAE, MAPE, and RMSE across all seasons and all periods. The MAPE for the four seasons has been reduced by 0.5%, 1.2%, 0.7%, and 0.9%, respectively, with corresponding values of 7.4%, 11.8%, 8.0%, and 7.9%. Additionally, the MAE decreased by 10 m, 10 m, 12 m, and 10 m, respectively. Except for the 12:00–23:50 time period in summer and spring, the average PBLH error estimated in all other periods has dropped below 50 m. Notably, the HML model exhibited relatively strong model fitting abilities in autumn and winter, with corresponding MAE values of 35 m and 39 m, respectively. Furthermore, $R^2$ of the basic HML model reaches a maximum of 0.7–0.9, while it can be improved by 0.1–0.2 after optimization. Particularly in the summer periods of 6:00–11:50 and 12:00–17:50, $R^2$ has improved from 0.79 and 0.86 to 0.97 and 0.96, respectively. This indicates that the optimized HML model can accurately estimate PBLH during periods of vigorous convection and relatively drastic PBLH variations in summer. Therefore, the method of removing outliers can effectively improve the parameters of the model and enhance the estimation performance of the HML model, making it more adaptable to the

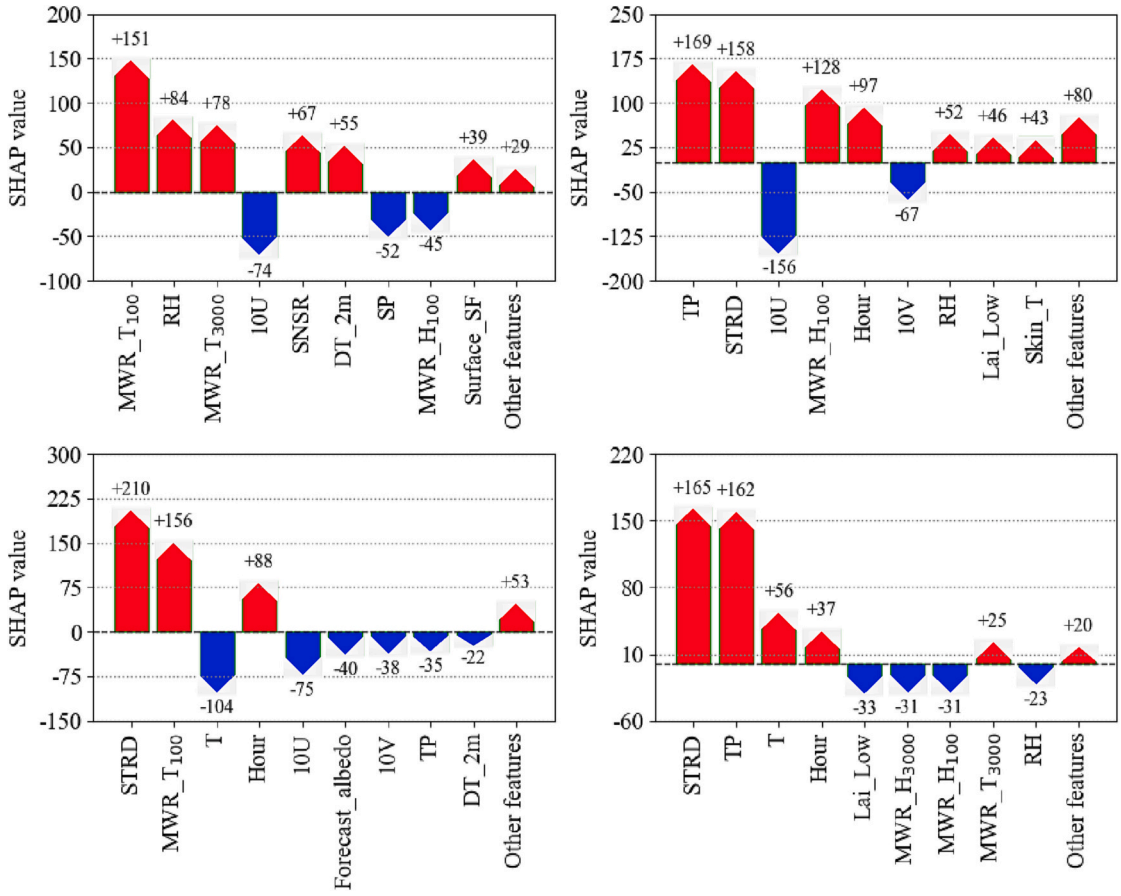


**Fig. 7.** Histogram of positive and negative contributions of each variable in four outliers.

**Table 3**
The quality evaluation results of the pre-optimization model and the post-optimization model for various seasons, periods, and case studies.

| | | | | Before optimization | | | | After optimization | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Season | Time | Averaged BLH (m) | ERA5 MAPE (%) | $R^2$ | MAE (m) | RMSE(m) | MAPE (%) | $R^2$ | MAE (m) | RMSE(m) | MAPE (%) |
| Spring | 0:00–5:50 | 591 | 89.3 | 0.94 | 46 | 103 | 8.5 | **0.96** | **41** | **84** | **7.6** |
| | 6:00–11:50 | 788 | 75.2 | 0.96 | 47 | 120 | 8.2 | **0.99** | **39** | **62** | **7.9** |
| | 12:00–17:50 | 1339 | 75.8 | 0.91 | 98 | 217 | 7.0 | **0.97** | **69** | **124** | **5.9** |
| | 18:00–23:50 | 914 | 69.9 | 0.97 | 58 | 108 | 7.7 | **0.97** | **58** | **108** | **7.7** |
| | Whole day | 886 | 77.6 | 0.96 | 60 | 141 | 7.9 | **0.98** | **50** | **94** | **7.4** |
| | Case study | 836 | 49.4 | 0.63 | 199 | 262 | 23.4 | **0.76** | **154** | **212** | **19.2** |
| Summer | 0:00–5:50 | 429 | 76.6 | 0.86 | 50 | 105 | 15.1 | **0.86** | **50** | **105** | **15.1** |
| | 6:00–11:50 | 428 | 47.2 | 0.79 | 58 | 147 | 12.1 | **0.97** | **33** | **46** | **10.9** |
| | 12:00–17:50 | 1683 | 18.2 | 0.86 | 113 | 200 | 7.0 | **0.96** | **78** | **108** | **5.4** |
| | 18:00–23:50 | 777 | 73.3 | 0.98 | 51 | 83 | 11.3 | **0.98** | **51** | **83** | **11.3** |
| | Whole day | 718 | 60.4 | 0.96 | 60 | 129 | 13.0 | **0.98** | **50** | **90** | **11.8** |
| | Case study | 970 | 47.0 | 0.25 | 246 | 289 | 30.4 | **0.49** | **191** | **239** | **22.1** |
| Autumn | 0:00–5:50 | 409 | 84.4 | 0.96 | 29 | 65 | 9.3 | **0.97** | **28** | **56** | **9.3** |
| | 6:00–11:50 | 501 | 71.8 | 0.88 | 44 | 114 | 8.4 | **0.95** | **33** | **65** | **7.6** |
| | 12:00–17:50 | 1002 | 63.1 | 0.83 | 64 | 229 | 6.4 | **0.98** | **42** | **68** | **5.2** |
| | 18:00–23:50 | 557 | 79.8 | 0.85 | 49 | 156 | 10.6 | **0.96** | **38** | **73** | **10.0** |
| | Whole day | 615 | 74.8 | 0.89 | 47 | 153 | 8.7 | **0.98** | **35** | **67** | **8.0** |
| | Case study | 612 | 32.4 | 0.39 | 186 | 271 | 27.4 | **0.55** | **142** | **232** | **19.8** |
| Winter | 0:00–5:50 | 511 | 71.2 | 0.96 | 36 | 81 | 9.4 | **0.98** | **33** | **62** | **9.2** |
| | 6:00–11:50 | 524 | 55.2 | 0.80 | 51 | 150 | 9.0 | **0.95** | **35** | **70** | **7.7** |
| | 12:00–17:50 | 981 | 43.1 | 0.90 | 61 | 145 | 7.6 | **0.98** | **48** | **70** | **6.1** |
| | 18:00–23:50 | 633 | 70.2 | 0.93 | 49 | 124 | 9.1 | **0.98** | **41** | **74** | **8.3** |
| | Whole day | 654 | 60.2 | 0.92 | 49 | 128 | 8.8 | **0.98** | **39** | **69** | **7.9** |
| | Case study | 729 | 49.1 | −0.57 | 216 | 271 | 31.6 | **0.06** | **171** | **209** | **25.2** |

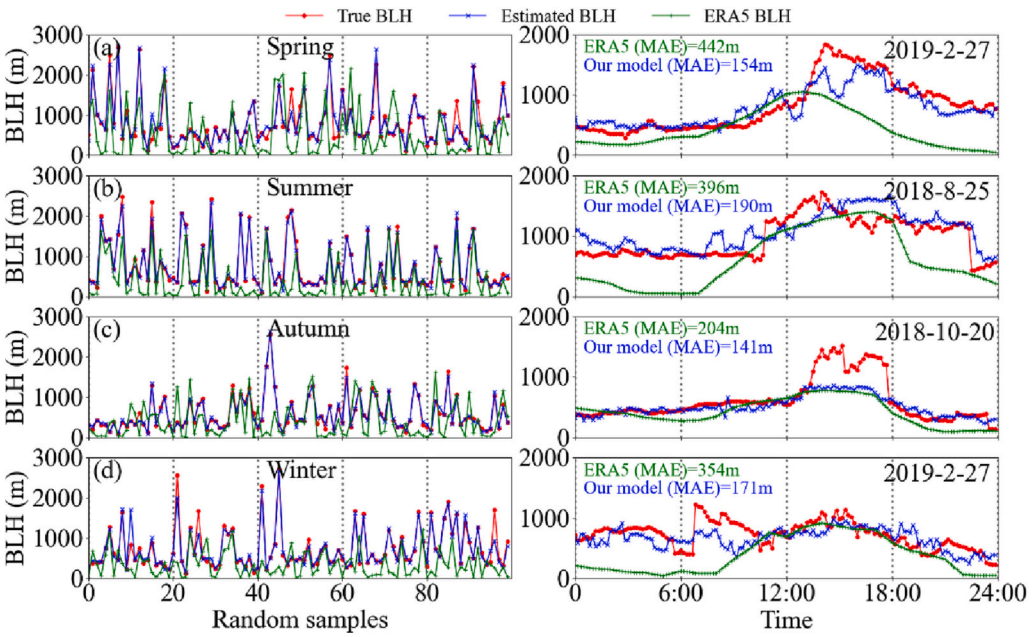


**Fig. 8.** Time series of ERA5-BLH, HML estimated PBLH, and observation results for each season in testing datasets four selected cases in each season. The date of the spring case selected in (a) is 2019-4-12, the date of the summer case selected in (b) is 2018-8-25, the date of the autumn case selected in (c) is 2018-10-20, and the date of the winter case selected in (d) is 2019-2-27.

variation of PBLH.

The winter PBLH estimation results of the HML model before (a) and after (b) optimization are illustrated in Fig. S1. The error curve depicts that the instances of significant deviations in the estimated PBLH have substantially reduced after optimization, resulting in a smoother curve. Although the MAE remains the highest in the afternoon period, it has improved by 13 m compared to the model before optimization. Notably, during the period of 12:00–23:50, characterized by high temperature, low humidity, and relatively drastic variations in horizontal and latitudinal winds, the estimation error of PBLH is more substantial than that during the period of 0:00–11:50. In contrast to the estimation results in Fig. 3, the variation of 2 m dew point temperature is more drastic during the daytime in winter, which could impact the accuracy of the HML model.

### 4.5. Case studies

This study also conducted predictive research on selected continuous samples from one day in each of the four seasons. Since the observation data are often discontinuous, this study selected cases with complete observation periods on a single day. As shown in Table 3, the MAE of the four selected cases is 199 m, 246 m, 186 m, and 216 m, with corresponding MAPE of 23.4%, 30.4%, 27.4%, and 31.6%, respectively. Our model outperforms the ERA5-BLH, which has MAPE results of 49.4%, 47.0%, 32.4%, and 49.1%. However, the estimation quality of the four examples is often relatively large compared to the results of the test set, especially in summer and winter, with an average error of >200 m, as shown in Fig. 8. This is primarily attributed to the learning method adopted by the ML models, which shuffle the dataset without considering the continuous variation pattern of PBLH. The model parameters are primarily tuned for irregular datasets based solely on variations in meteorological variables.

To improve the low accuracy of individual predictions, this study adopts the five-fold cross-validation method to remove 20% of the outliers in the test dataset each time. The optimized HML is used to improve case estimations after adjusting the model parameters using the complete training set samples. As shown in Table 3, the MAPE of the case estimation results for each season improved by 4.2%–8.3%, and the error decreased by 44–55 m after optimization. The winter case study result before (a) and after (b) optimization in Fig. S2 shows the daily PBLH estimation result has significantly enhanced, with the MAE increasing by nearly 50 m. Due to the limited number of training samples and the impact of long-distance transmission detected by the ceilometer, the accuracy of the HML model based solely on meteorological variables is poor, especially around 7:00 in 2019-2-27. Post-optimized HML model can improve the estimation results of such abrupt variations, but it still cannot entirely enhance the estimation results. Moreover, the stability and generalization of the HML model need further enhancement. In general, both pre-optimized and post-optimized model results are significantly better than the PBLH in ERA5 reanalysis data.

## 5. Conclusion

This paper proposes a high temporal resolution HML model to estimate the PBLH of the Beijing urban area using diverse meteorological data. The estimates are compared with ERA5-BLH, which is currently recognized as the highest resolution and accuracy dataset (Guo et al., 2021). The HML model demonstrates stronger generalization ability than the single ML algorithm (LightGBM) and higher precision than ERA5-BLH. Our analysis shows that MWR-specific height temperature and RH (100 m and 3 km) can be used as the major features to enhance the accuracy of the estimated PBLH when compared with various elements of ERA5 reanalysis data. The feature importance of different seasonal models shows that thermal factors contribute the most to the estimation of PBLH in summer, while dynamic factors dominate in other seasons. Quantitative analysis of SHAP values shows that selected meteorological elements have varying positive or negative effects in different seasons, and the outliers can affect the overall accuracy of the HML model. After removing some outliers and fine-tuning the parameters of the HML model, we achieve improved results for different phases of all seasons, with MAPE ranging between 5.2% and 15.1%, and the results of four case studies decreased by 4.2%–8.3%.

Therefore, differing from previous ML models for some specific regions based on real ground-based observations data, our model achieves high performance and shows great potentials by combining ERA5 reanalysis data with MWR profiles. This not only provides valuable

support for urban air quality forecasting and atmospheric environmental capacity, but also improves the PBLH accuracy from ERA5 reanalysis data. Nevertheless, the interference with turbulent kinetic energy can be further added in future research, designing a multi-parameter HML to improve the estimation results of PBLH in some typical scenarios such as strong convection and extreme weather. Moreover, aiming at strengthening the continuous long-term prediction ability of ML models, massive related datasets will be considered in the training process as well in future.

## Funding

This study was supported by the National Key Research and Development Program of China (2022YFF0802501), the CAS Strategic Priority Research Program (XDA23020301), the China Postdoctoral Foundation (2021 M700140; 2022TQ0332), the Key Laboratory of Ecological Environment Big Data Open Foundation of Zhejiang (EEBD-2022-01) and National Natural Science Foundation of China (Grant No. 42005003 and 41475094).

## Author statement

We thank the Editor-in-Chief and reviewers for the detailed and helpful comments to improve the manuscript. The manuscript with traces of revision has been submitted based on reviewers' comments. In general, Xin Jinyuan designed this research. Kecheng Peng and Xin Jinyuan analyzed the data, and wrote the manuscript. All authors provided experimental assistance and commented on the manuscript.

## CRediT authorship contribution statement

**Kecheng Peng:** Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jinyuan Xin:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Investigation, Writing – review & editing. **Xiaoqian Zhu:** Conceptualization, Methodology, Investigation, Validation, Writing – review & editing. **Xiaoyuan Wang:** Data curation, Validation. **Xiaoqun Cao:** Formal analysis, Writing – review & editing. **Yongjing Ma:** Methodology, Validation. **Xinbing Ren:** Data curation, Formal analysis. **Dandan Zhao:** Investigation, Validation. **Junji Cao:** Methodology, Funding acquisition. **Zifa Wang:** Investigation, Funding acquisition.

## Declaration of Competing Interest

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.

## Data availability

DData will be made available on request.

## Acknowledgments

We thank the following people or organizations for providing the data and package used in this paper and list the resources here. ECWMF data are available at https://cds.climate.copernicus.eu.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosres.2023.106925.

## References

Allabakash, S., Lim, S., 2020. Climatology of planetary boundary layer height-controlling meteorological parameters over the Korean Peninsula. Remote Sens. 12 (16), 2571. https://doi.org/10.3390/rs12162571.

Banakh, V.A., Smalikho, I.N., Falits, A.V., 2021. Estimation of the height of the turbulent mixing layer from data of Doppler lidar measurements using conical scanning by a probe beam. Atmos. Measure. Techn. 14 (2), 1511–1524. https://doi.org/10.5194/amt-14-1511-2021.

Berg, L.K., Newsom, R.K., Turner, D.D., 2017. Year-long vertical velocity statistics derived from Doppler lidar data for the continental convective boundary layer. J. Appl. Meteorol. Climatol. 56 (9), 2441–2454. https://doi.org/10.1175/JAMC-D-16-0359.1.

Bravo-Aranda, J.A., de Arruda Moreira, G., Navas-Guzmán, F., Granados-Muñoz, M.J., Guerrero-Rascado, J.L., Pozo-Vázquez, D., Arbizu-Barrena, C., Olmo Reyes, F.J., Mallet, M., Alados-Arboledas, L., 2017. A new methodology for PBL height estimations based on lidar depolarization measurements: analysis and comparison against MWR and WRF model-based results. Atmos. Chem. Phys. 17, 6839–6851.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cadeddu, M.P., Turner, D.D., Liljegren, J.C., 2009. A neural network for real-time retrievals of PWV and LWP from Arctic millimeter-wave ground-based observations. IEEE Trans. Geosci. Remote Sens. 47 (6), 1887–1900. https://doi.org/10.1109/TGRS.2008.2010459.

Caicedo, V., Rappenglück, B., Lefer, B., Morris, G., Toledo, D., Delgado, R., 2017. Comparison of aerosol lidar retrieval methods for boundary layer height detection using ceilometer aerosol backscatter data. Atmos. Measure. Techn. 10, 1609–1622. https://doi.org/10.5194/amt-10-1609-2017.

Chen, F., Kusaka, H., Bornstein, R., Ching, J., Grimmond, C., Grossman-Clarke, S., Loridan, T., Manning, K.W., Martilli, A., Miao, S., 2011. The integrated WRF/urban modelling system: development, evaluation, and applications to urban environmental problems. Int. J. Climatol. 31 (2), 273–288. https://doi.org/10.1002/joc.2158.

Chen, G., Li, S., Knibbs, L.D., Hamm, N.A.S., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M.J., Guo, Y., 2018. A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information. Sci. Total Environ. 636, 52–60. https://doi.org/10.1016/j.scitotenv.2018.04.366.

Chen, S., Tong, B., Russell, L.M., Wei, J., Guo, J., Mao, F., Liu, D., Huang, Z., Xie, Y., Qi, B., 2022. Lidar-based daytime boundary layer height variation and impact on the regional satellite-based $PM_{2.5}$ estimate. Remote Sens. Environ. 281, 113224 https://doi.org/10.1016/j.rse.2022.113224.

Cimini, T., Hewison, T., Martin, L., Güldner, J., Gaffard, C., Marzano, F.S., 2006. Temperature and humidity profile retrievals from ground-based microwave radiometers during TUC. Meteorol. Z. 15 (1), 45–56. https://doi.org/10.1127/0941-2948/2006/0093.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139. https://doi.org/10.1006/jcss.1997.1504.

Gagne, D.J., McGovern, A., Haupt, S.E., Williams, J.K., 2017. Evaluation of statistical learning configurations for gridded solar irradiance forecasting. Sol. Energy 150, 383–393. https://doi.org/10.1016/j.solener.2017.04.031.

Gallée, H., Preunkert, S., Argentini, S., Frey, M., Genthon, C., Jourdain, B., Pietroni, I., Casasanta, G., Barral, H., Vignon, E., 2015. Characterization of the boundary layer at Dome C (East Antarctica) during the OPALE summer campaign. Atmos. Chem. Phys. 15 (11), 6225–6236. https://doi.org/10.5194/acp-15-6225-2015.

Geiss, A., Wiegner, M., Bonn, B., Schaefer, K., Forkel, R., von Schneidemesser, E., Muenkel, C., Chan, K.L., Nothard, R., 2017. Mixing layer height as an indicator for urban air quality? Atmos. Measure. Techn. 10 (8), 2969–2988. https://doi.org/10.5194/amt-10-2969-2017.

Gerbig, C., Körner, S., Lin, J.C., 2008. Vertical mixing in atmospheric tracer transport models: error characterization and propagation. Atmos. Chem. Phys. 8 (3), 591–602. https://doi.org/10.5194/acp-8-591-2008.

Guo, J., Miao, Y., Zhang, Y., Liu, H., Li, Z., Zhang, W., He, J., Lou, M., Yan, Y., Bian, L., 2016. The climatology of planetary boundary layer height in China derived from radiosonde and reanalysis data. Atmos. Chem. Phys. 16 (20), 13309–13319. https://doi.org/10.5194/acp-16-13309-2016.

Guo, J., Zhang, J., Yang, K., Liao, H., Zhang, S., Huang, K., Lv, Y., Shao, J., Yu, T., Tong, B., Li, J., Su, T., Yim, S.H.L., Stoffelen, A., Zhai, P., Xu, X., 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 reanalysis. Atmos. Chem. Phys. 21, 17079–17097. https://doi.org/10.5194/acp-21-17079-2021.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3 (Mar), 1157–1182.

Haeffelin, M., Angelini, F., Morille, Y., Martucci, G., Frey, S., Gobbi, G.P., Lolli, S., O'Dowd, C.D., Sauvage, L., Xueref-R'emy, I., Wastine, B., Feist, D.G., 2012. Evaluation of mixing-height retrievals from automatic profiling lidars and ceilometers in view of future integrated networks in Europe. Bound.-Layer Meteorol. 143 (1), 49–75. https://doi.org/10.1007/s10546-011-9643-z.

Ham, Y.G., Kim, J.H., Luo, J.J., 2019. Deep learning for multi-year ENSO forecasts. Nature 573, 568–572. https://doi.org/10.1038/s41586-019-1559-7.

Illingworth, A.J., Cimini, D., Haefele, A., Haeffelin, M., Hervo, M., Kotthaus, S., Löhnert, U., Martinet, P., Mattis, I., O'connor, E., 2019. How can existing ground-based profiling instruments improve European weather forecasts? Bull. Am. Meteorol. Soc. 100 (4), 605–619. https://doi.org/10.1175/BAMS-D-17-0231.1.

Jiang, Y., Xin, J., Zhao, D., Jia, D., Tang, G., Quan, J., Wang, M., Dai, L., 2021. Analysis of differences between thermodynamic and material boundary layer structure: comparison of detection by ceilometer and microwave radiometer. Atmos. Res. 248, 105179 https://doi.org/10.1016/j.atmosres.2020.105179.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems ACM, Long Beach, CA, USA, 30, pp. 3149–3157. https://doi.org/10.5555/3294996.3295074.

Knuteson, R.O., Revercomb, H.E., Best, F.A., Ciganovich, N.C., Dedecker, R.G., Dirkx, T. P., Ellington, S.C., Feltz, W.F., Garcia, R.K., Howell, H.B., Smith, W.L., 2004. Atmospheric emitted radiance interferometer. Part II: Instrument performance. J. Atmos. Ocean. Technol. 21, 1777–1789. https://doi.org/10.1175/JTECH-1663.1.

Kotthaus, S., Bravo-Aranda, J.A., Collaud Coen, M., Guerrero-Rascado, J.L., Costa, M.J., Cimini, D., O'Connor, E.J., Hervo, M., Alados-Arboledas, L., Jiménez-Portaz, M., 2023. Atmospheric boundary layer height from ground-based remote sensing: a review of capabilities and limitations. Atmos. Measure. Techn. 16 (2), 433–479. https://doi.org/10.5194/AMT-16-433-2023.

Krishnamurthy, R., Newsom, R.K., Berg, L.K., Xiao, H., Ma, P.-L., Turner, D.D., 2021. On the estimation of boundary layer heights: a machine learning approach. Atmos. Measure. Techn. 14 (6), 4403–4424. https://doi.org/10.5194/AMT-14-4403-2021.

Lee, J., Hong, J.W., Lee, K., Hong, J., Velasco, E., Lim, Y.J., Lee, J.B., Nam, K., Park, J., 2019. Ceilometer monitoring of boundary-layer height and its application in evaluating the dilution effect on air pollution. Bound.-Layer Meteorol. 172, 435–455. https://doi.org/10.1007/s10546-019-00452-5.

Li, H., Yang, Y., Wang, H., Li, B., Wang, P., Li, J., Liao, H., 2021b. Constructing a spatiotemporally coherent long-term $PM_{2.5}$ concentration dataset over China during 1980–2019 using a machine learning approach. Sci. Total Environ. 765, 144263 https://doi.org/10.1016/J.SCITOTENV.2020.144263.

Li, X.-B., Wang, D.-S., Lu, Q.-C., Peng, Z.-R., Wang, Z.-Y., 2018. Investigating vertical distribution patterns of lower tropospheric $PM_{2.5}$ using unmanned aerial vehicle measurements. Atmos. Environ. 173, 62–71. https://doi.org/10.1016/j.atmosenv.2017.11.009.

Li, Y., Li, J., Zhao, Y., Lei, M., Zhao, Y., Jian, B., Zhang, M., Huang, J., 2021a. Long-term variation of boundary layer height and possible contribution factors: a global analysis. Sci. Total Environ. 796, 148950 https://doi.org/10.1016/j.scitotenv.2021.148950.

Liu, N., Yan, Z., Tong, X., Jiang, J., Li, H., Xia, J., Lou, X., Ren, R., Fang, Y., 2022. Meshless Surface Wind speed Field Reconstruction based on Machine Learning. Adv. Atmos. Sci. 39 (10), 1721–1733. https://doi.org/10.1007/s00376-022-1343-8.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Proces. Syst. 30.

Lundberg, S.M., Erion, G.G., Lee, S.-I., 2018. Consistent individualized feature attribution for tree ensembles. arXiv:1706.06060.v6.

Ma, Y., Ye, J., Xin, J., Zhang, W., Vilà-Guerau de Arellano, J., Wang, S., Zhao, D., Dai, L., Ma, Y., Wu, X., 2020. The stove, dome, and umbrella effects of atmospheric aerosol on the development of the planetary boundary layer in hazy regions. Geophys. Res. Lett. 47 (13) https://doi.org/10.1029/2020GL087373 e2020GL087373.

Ma, Y., Xin, J., Wang, Z., Tian, Y., Wu, L., Tang, G., Zhang, W., de Arellano, J.V.-G., Zhao, D., Jia, D., 2022. How do aerosols above the residual layer affect the planetary boundary layer height? Sci. Total Environ. 814, 151953 https://doi.org/10.1016/J.SCITOTENV.2021.151953.

Mahrt, L., 1999. Stratified atmospheric boundary layers. Bound.-Layer Meteorol. 90 (3), 375–396. https://doi.org/10.1023/A:1001765727956.

Manninen, A., Marke, T., Tuononen, M., O'Connor, E., 2018. Atmospheric boundary layer classification with Doppler lidar. J. Geophys. Res.-Atmos. 123 (15), 8172–8189. https://doi.org/10.1029/2017JD028169.

Marques, M.T., Moreira, G.D.A., Pinero, M., Oliveira, A.P., Landulfo, E., 2018. Estimating the Planetary Boundary Layer Height from Radiosonde and Doppler Lidar Measurements in the City of São Paulo-Brazil. EPJ Web of Conferences, EDP Sciences, p. 06015. https://doi.org/10.1051/epjconf/201817606015.

McGovern, A., Elmore, K.L., Gagne, D.J., Haupt, S.E., Karstens, C.D., Lagerquist, R., Smith, T., Williams, J.K., 2017. Using artificial intelligence to improve real-time decision-making for high-impact weather. Bull. Am. Meteorol. Soc. 98 (10), 2073–2090. https://doi.org/10.1175/BAMS-D-16-0123.1.

Molero, F., Barragán, R., Artíñano, B., 2022. Estimation of the atmospheric boundary layer height by means of machine learning techniques using ground-level meteorological data. Atmos. Res. 279, 106401 https://doi.org/10.1016/j.atmosres.2022.106401.

Moreira, G.A., Guerrero-Rascado, J.L., Bravo-Aranda, J.A., Foyo-Moreno, I., Cazorla, A., Alados, I., Lyamani, H., Landulfo, E., Alados-Arboledas, L., 2020. Study of the planetary boundary layer height in an urban environment using a combination of microwave radiometer and ceilometer. Atmos. Res. 240, 104932 https://doi.org/10.1016/j.atmosres.2020.104932.

Moreira, G.A., Sánchez-Hernández, G., Guerrero-Rascado, J.L., Cazorla, A., Alados-Arboledas, L., 2022. Estimating the urban atmospheric boundary layer height from remote sensing applying machine learning techniques. Atmos. Res. 266, 105962 https://doi.org/10.1016/j.atmosres.2021.105962.

Mues, A., Rupakheti, M., Münkel, C., Lauer, A., Bozem, H., Hoor, P., Butler, T., Lawrence, M.G., 2017. Investigation of the mixing layer height derived from ceilometer measurements in the Kathmandu Valley and implications for local air quality. Atmos. Chem. Phys. 17 (13), 8157–8176. https://doi.org/10.5194/acp-17-8157-2017.

Muñoz-Esparza, D., Becker, C., Sauer, J.A., Gagne, D.J., Schreck, J., Kosović, B., 2022. On the application of an observations-based machine learning parameterization of surface layer fluxes within an atmospheric large-eddy simulation model. J. Geophys. Res. Atmos. 127 (16) https://doi.org/10.1029/2021JD036214 e2021JD036214.

Palmén, E., Newton, C.W., 1969. Atmospheric Circulation Systems: Their Structure and Physical Interpretation, vol. 13. Academic press (ISBN 978-0-12-544550-4.).

Rey-Sanchez, C., Wharton, S., Vil'a-Guerau de Arellano, J., Paw, U., Hemes, K.S., Fuentes, J.D., Osuna, J., Szutu, D., Ribeiro, J.V., Verfaillie, J., Baldocchi, D., 2021. Evaluation of atmospheric boundary layer height from wind profiling radar and slab models and its responses to seasonality of land cover, subsidence, and advection. J. Geophys. Res.-Atmos. 126 (7) https://doi.org/10.1029/2020JD033775e2020JD033775.

Rieutord, T., Aubert, S., Machado, T., 2021. Deriving boundary layer height from aerosol lidar using machine learning: KABL and ADABL algorithms. Atmos. Measure. Techn. 14 (6), 4335–4353. https://doi.org/10.5194/amt-14-4335-2021.

Saeed, U., Rocadenbosch, F., Crewell, S., 2016. Adaptive estimation of the stable boundary layer height using combined lidar and microwave radiometer observations. IEEE Trans. Geosci. Remote Sens. 54 (12), 6895–6906. https://doi.org/10.1186/s12868-016-0283-6.

Sawyer, V., Li, Z., 2013. Detection, variations and intercomparison of the planetary boundary layer depth from radiosonde, lidar and infrared spectrometer. Atmos. Environ. 79, 518–528. https://doi.org/10.1016/j.atmosenv.2013.07.019.

Singh, N., Solanki, R., Ojha, N., Janssen, R.H.H., Pozzer, A., Dhaka, S.K., 2016. Boundary layer evolution over the Central Himalayas from radio wind profiler and model simulations. Atmos. Chem. Phys. 16, 10559–10572. https://doi.org/10.5194/acp-16-10559-2016.

Stull, R.B., 1988. An Introduction to Boundary Layer Meteorology. Kluwer Acad, Dordrecht, Netherlands, pp. 666–pp.

Su, T., Li, J., Li, C., Xiang, P., Lau, A.K.H., Guo, J., Yang, D., Miao, Y., 2017. An intercomparison of long-term planetary boundary layer heights retrieved from CALIPSO, ground-based lidar, and radiosonde measurements over Hong Kong. J. Geophys. Res.-Atmos. 122 (7), 3929–3943. https://doi.org/10.1002/2016JD025937.

Trentmann, J., Keil, C., Salzmann, M., Barthlott, C., Bauer, H.-S., Schwitalla, T., Lawrence, M.G., Leuenberger, D., Wulfmeyer, V., Corsmeier, U., Kottmeier, C., Wernli, H., 2009. Multi-model simulations of a convective situation in low-mountain terrain in Central Europe. Meteorog. Atmos. Phys. 103 (1), 95–103. https://doi.org/10.1007/s00703-008-0323-6.

Turner, D.D., Wulfmeyer, V., Berg, L.K., Schween, J.H., 2014. Water vapor turbulence profiles in stationary continental convective mixed layers. J. Geophys. Res. Atmos. 119, 11–151. https://doi.org/10.1002/2014JD022202.

Wang, W., Mao, F., Gong, W., Pan, Z., Du, L., 2016. Evaluating the governing factors of variability in nocturnal boundary layer height based on elastic lidar in Wuhan. Int. J. Environ. Res. Public Health 13 (11), 1071. https://doi.org/10.3390/ijerph13111071.

Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., Cribb, M., 2019. Estimating 1-km-resolution $PM_{2.5}$ concentrations across China using the space-time random forest approach. Remote Sens. Environ. 231, 111221. https://doi.org/10.1016/j.rse.2019.111221.

Wei, J., Li, Z., Pinker, R.T., Wang, J., Sun, L., Xue, W., Li, R., Cribb, M., 2021. Himawari-8-derived diurnal variations in ground-level $PM_{2.5}$ pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM). Atmos. Chem. Phys. 21, 7863–7880. https://doi.org/10.5194/acp-21-7863-2021.

Xin, J., Ma, Y., Zhao, D., Gong, C., Ren, X., Tang, G., Xia, X., Wang, Z., Cao, J., de Arellano, J.V.-G., 2023. The feedback effects of aerosols from different sources on the urban boundary layer in Beijing China. Environ. Pollut. 325, 121440 https://doi.org/10.1016/J.ENVPOL.2023.121440.

Yu, L., Liu, H., 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03), pp. 856-863.

Zhao, D., Xin, J., Gong, C., Quan, J., Liu, G., Zhao, W., Wang, Y., Liu, Z., Song, T., 2019. The formation mechanism of air pollution episodes in Beijing city: insights into the measured feedback between aerosol radiative forcing and the atmospheric boundary layer stability. Sci. Total Environ. 692, 371–381. https://doi.org/10.1016/j.scitotenv.2019.07.255.

Zhao, D., Xin, J., Gong, C., Quan, J., Wang, Y., Tang, G., Ma, Y., Dai, L., Wu, X., Liu, G., 2021. The impact threshold of the aerosol radiative forcing on the boundary layer structure in the pollution region. Atmos. Chem. Phys. 21 (7), 5739–5753. https://doi.org/10.5194/ACP-21-5739-2021.

Zhu, X., Tang, G., Lv, F., Hu, B., Cheng, M., Münkel, C., Schäfer, K., Xin, J., An, X., Wang, G., Li, X., Wang, Y., 2018. The spatial representativeness of mixing layer height observations in the North China Plain. Atmos. Res. 209, 204–211. https://doi.org/10.1016/j.atmosres.2018.03.019.