UNPAD

Vol. 10, No. 02 (2026) 199 – 215
JIIF (Jurnal Ilmu dan Inovasi Fisika)
https://jurnal.unpad.ac.id/jiif/

ESTIMATION OF PLANETARY BOUNDARY LAYER HEIGHT FROM SURFACE METEOROLOGICAL OBSERVATION USING RANDOM FOREST

ADILAKSA KHARISMA[1], YOSAFAT DONNI HARYANTO[1,*], PURWANTI LELLY SABRINA[1], MAMAN SUDARISMAN[2]

[1]Department of Meteorology, State College of Meteorology Climatology and Geophysics
[2]Department of Climatology, State College of Meteorology Climatology and Geophysics
Jl. Meteorologi No.5, Tanah Tinggi, Tangerang, Kota Tangerang, Banten, Indonesia 15119

**Corresponding author*
Email: yosafat.haryanto@stmkg.ac.id

Diserahkan: 10/07/2026
Diterima: 15/08/2026
Dipublikasikan: 19/08/2026

Abstrak. The Planetary Boundary Layer (PBL) is the lowest layer of the atmosphere that directly interacts with earth surface and play crucial role in momentum exchange and particle dispersion. Conventional observations of the PBL are limited by spatial coverage, operational costs, and temporal availability. Therefore, this study aims to develop a Random Forest model for estimating the PBL Height (PBLH) using surface meteorological observations. This research was conducted in Central Kalimantan during the dry season (June – November) from 2018 to 2024. Surface meteorological observation and PBLH ERA5 were used as predictor variables, while PBLH derived from radiosonde using the Bulk Richardson Number (BRN) method served as the target variable, with the complete dataset used for final validation. On the independent test set, the model achieved a Pearson correlation of 0.5020 ($p < 0.001$), R^2 of 0.2478, RMSE of 207.85 m, and MAE of 162.49 m. Feature importance analysis identified PBLH ERA5, sea level pressure, and wind speed as the most influential predictors, consistent with physical processes governing PBL stability and mixing. Validation across the complete dataset showed that the model substantially reduced the systematic negative bias of PBLH ERA5 relative to radiosonde, from −325.35 m to only −0.63 m ($r = 0.8238$, $R^2 = 0.5982$, RMSE = 155.41 m, MAE = 117.27 m). However, the model still showed limited skill in reproducing the most extreme PBLH values recorded by radiosonde. Overall, the proposed model provides a practical, low-cost approach for correcting reanalysis PBLH toward radiosonde estimates during the dry season.

Kata kunci: PBLH, radiosonde, surface meteorological observation, random forest, machine learning.

Abstract. Planetary Boundary Layer (PBL) merupakan lapisan terbawah atmosfer yang berinteraksi langsung dengan permukaan dan berperan penting dalam pertukaran momentum dan dispersi partikel. Pengamatan PBL secara konvensional terbatas dalam cakupan wilayah, biaya, dan waktu. Maka dari itu, penelitian ini bertujuan mengembangkan sistem estimasi ketinggian PBL (PBLH) menggunakan data permukaan sebagai data input untuk mengetahui nilai PBLH berbasis machine learning Random Forest . Penelitian ini berfokus pada Provinsi Kalimantan Tengah dengan rentang waktu 2018-2024 pada musim kemarau (Juni – November).

DOI: https://doi.org/10.24198/jiif.v10i2.72186 e-ISSN: 2549-7014

Data observasi permukaan dan PBLH ERA5 digunakan sebagai variabel prediktor, sedangkan PBLH hasil perhitungan radiosonde dengan metode Bulk Richardson Number (BRN) digunakan sebagai variabel target dan seluruh dataset selanjutnya digunakan untuk validasi akhir konsistensi ketiga sumber data PBLH. Pada data uji independen, model yang dikembangkan menghasilkan nilai koefisien korelasi 0,5020 ($p < 0,001$), nilai R^2 0,2478, RMSE 207,85 m, dan MAE 162,49 m. Proses bangun model menunjukkan bahwa PBLH ERA5, tekanan permukaan laut, dan kecepatan angin merupakan prediktor dengan kontribusi terbesar, sejalan dengan proses stabilitas termodinamika dan pencampuran dinamis pada PBL. Validasi pada keseluruhan dataset menunjukkan bahwa model Random Forest mampu menurunkan secara signifikan bias sistematik negatif PBLH ERA5 terhadap radiosonde, dari −325,35 m menjadi hanya −0,63 m (r = 0,8238, R^2 = 0,5982, RMSE = 155,41 m, MAE = 117,27 m). Meskipun demikian, model masih memiliki keterbatasan dalam mengestimasi nilai PBLH ekstrem yang tercatat oleh radiosonde. Secara keseluruhan, model yang dikembangkan memberikan pendekatan praktis dan berbiaya rendah untuk mengoreksi PBLH berbasis reanalysis agar lebih konsisten dengan observasi radiosonde pada musim kemarau.

Keywords: PBLH, radiosonde, pengamatan meteorologi permukaan, random forest, machine learning.

1. Pendahuluan

Planetary Boundary Layer (PBL) is the lowest part of troposphere and directly interacts with the earth surface through turbulent exchanges of momentum, heat, moisture, and pollutants [1], [2]. The evolution of Planetary Boundary Layer Height (PBLH) strongly influences weather, air quality, cloud formation, and atmospheric particle dispersion [3]. Especially during the dry season, enhanced solar radiation generally produces a deeper convective boundary layer, whereas variations in humidity, wind, and cloud cover may substantially modify PBL development [4]. Consequently, accurate PBLH estimation during the dry season is essential for meteorological and operational applications of weather forecasting. This is particularly crucial in Central Kalimantan, where vast peatland areas are susceptible to forest and land fires during the dry season and the PBLH contributes to controlling the vertical dispersion and near-surface accumulation of smoke pollutants [5], [6].

Conventional PBL observations still rely on observations of radiosonde, lidar, ceilometer, and numerical reanalysis product [7], [8]. Although these approaches generally provide reliable estimates, their operational use is often constrained by limited spatial and temporal coverage, high operational costs, or computational requirements [9]. However, continuous surface meteorological observations at weather stations can be an alternative for estimating PBLH through machine learning-based approach [10], [11]. Since these observations are routinely collected as part of operational weather monitoring, they provide a cost-effective data source with high temporal resolution for developing machine learning-based PBLH estimation models [1].

Recent advances in the application of machine learning have enabled the estimation of PBLH using surface meteorological variables with various algorithms capable of capturing the nonlinear relationship between surface atmospheric conditions and the development of PBL [10]. Several studies have shown that Random Forest successfully captures the nonlinear relationships between meteorological variables and PBLH, outperforming conventional regression methods [12]. This algorithm has been widely adopted due to its robustness against overfitting, ability to model nonlinear interactions, and relatively simple implementation [9], [13]. In addition, Random Forest provides feature importance measures that contribute to the interpretation of predictor, making it particularly suitable for atmospheric and environmental applications where understanding variable influence is essential [14].

Alongside these advancements, various prior studies have developed machine learning models using PBLH reanalysis data and radiosonde observations [11]. Notably, Xi et al. [11] and Guo et al. [13] demonstrated that machine learning can be used to learn and correct the systematic bias of PBLH ERA5 relative to radiosonde observations by treating ERA5 together with surface meteorological variables as predictors of the PBLH Radiosonde-based estimate. This correction is needed because PBLH ERA5 cannot simply be taken at face value, particularly over tropical regions, where its errors relative to radiosonde have been shown to be large and systematic. Virman et al. [15] found that ERA5 deviates more strongly from tropical radiosonde observations than its predecessor ERA-Interim, while Guo et al. [16] and Kumar et al. [17] reported boundary layer height biases of several hundred meters between ERA5 and radiosonde, comparable to or larger than the PBLH itself under the shallow boundary layers typical of the humid tropics. Given the presence of bias in reanalysis data for tropical regions, the direct use of ERA5 PBLH risks introducing significant errors in downstream applications, such as air quality forecasting and smoke dispersion modelling within tropical peatland areas. However, this bias-correction approach remains largely unevaluated under tropical dry-season conditions, such as in Indonesia [18].

Therefore, this study develops a Random Forest model to estimate PBLH from surface meteorological observations during the dry season (June – November) in Central Kalimantan. PBLH from radiosonde profile using the Bulk Richardson Number (BRN) method was used as the target variable, while surface meteorological observations and ERA5 PBLH used as predictor variables. The complete dataset, comprising both training and testing subsets is then used to validate the overall consistency among ERA5 PBLH, radiosonde PBLH, and Random Forest estimates. The proposed framework integrates feature importance analysis, statistical evaluation on an independent test set, and validation using the full dataset to assess the capability of surface observations and ERA5 PBLH to estimate and correct PBLH values so they align with radiosonde measurements. The study findings are expected to provide a sustainable and operationally viable approach for estimating PBLH, addressing the limitations of radiosonde observations, particularly during the dry season [9], [10], [11].

2. Research Methods

2.1 Data

The datasets used in this study were obtained from Iskandar Meteorological Station in Central Kalimantan with the WMO Station No. 96645 and coordinates 2.73° S and 111.66° E. This study period covered the dry season (June – November) from 2018 to 2024 using daily observations at 00:00 UTC. Three datasets were used in this study, specifically surface meteorological observations, radiosonde observations, and ERA5 boundary layer height reanalysis data. These datasets were integrated to develop a Random Forest model for estimating PBLH Radiosonde from surface meteorological observations and PBLH ERA5, and to subsequently validate the consistency between the three PBLH sources using the complete dataset.

Surface meteorological observations were obtained from Ogimet database [19]. Surface meteorological parameters included air temperature (T_avg, T_min, T_max, and AppT_avg), dew point temperature (Td_avg), relative humidity (Hr_avg), sea level pressure (Press_sl), wind speed (Wind_speed), wind direction (Wind_dir), visibility (Vis), sunshine duration (Sun), precipitation (Prec), total cloud cover (Tot_cl), low cloud cover (Low_cl), and present weather observations (W1 – W8). The weather condition variables (W1 – W8) represent weather observations recorded over successive 3-hour periods. For example, W1 corresponds to observations from 00:00 to 03:00 UTC. These weather codes were converted into numerical values before being used as predictor variable.

Radiosonde observations were obtained from the University of Wyoming Atmospheric Science Radiosonde Archive [20]. The radiosonde profiles were processed using Bulk Richardson Number (BRN) method to estimate the Planetary Boundary Layer Height (PBLH) [8], [21]. In this study, PBLH derived from this BRN calculation (PBLH Radiosonde) was used as the target variable of the Random Forest model. The BRN was calculated using Equation (1):

$$Ri_B = \frac{g}{\theta_v}\frac{\Delta\theta_v \Delta z}{(\Delta u)^2 + (\Delta v)^2} \qquad (1)$$

As shown in Equation (1), Ri_B denotes the Bulk Richardson Number, g is gravitational acceleration (9.81 ms^{-2}), θ_v is the virtual potential temperature, u and v denote the zonal and meridional wind components, and z represents the observation height above the ground. In this study, PBLH was identified as the first altitude where the BRN exceeded the critical value of 0.25 [21].

ERA5 derived PBLH (PBLH ERA5) data were obtained from the ERA5 Hourly Data on Single Levels dataset provided by the Copernicus Climate Data Store (CDS). The PBLH ERA5 is diagnosed within the Integrated Forecasting System (IFS) using the BRN method, in which the boundary layer height is identified as the first level where the BRN reach es a critical value of 0.25, makes it physically consistent with the PBLH Radiosonde calculated using the same criterion [21], [22]. The downloaded NetCDF files were spatially interpolated using bilinear interpolation method to produce a spatial resolution of 1 km x 1 km to improve the spatial representativeness of the extracted grid relative to the observation site while preserving continuity of the data [11], [23]. Subsequently, the grid cell corresponding to the study location was extracted and used as the target variable for model development throughout the study period. Conversely, the ERA5 PBLH is included as a predictor variable alongside surface meteorological observations so that its systematic bias relative to PBLH Radiosonde-based estimates can be studied and reduced by the model.

2.2 Random Forest Model Development

Prior to model development, all datasets were integrated into single Microsoft Excel (.xlsx) file and subjected to a preprocessing procedure. The preprocessing included data quality control by identifying missing values, duplicate data, and outliers [24]. No duplicate records were found in the 1281-day dataset. Because PBLH Radiosonde was used as the target variable, the 174 daily records (13.6%) without a corresponding radiosonde observation were removed, leaving 1107 records for model development.

Missing values in the remaining predictor variables, including PBLH ERA5, were then imputed using the median, while outliers in the continuous variables (excluding the discrete present weather codes W1–W8) were capped at the 1.5× interquartile range (IQR) boundary rather than removed, so as not to distort the shape of the data distribution. The final dataset consisted of 23 predictor variables, twenty-two surface meteorological parameters together with PBLH ERA5, and PBLH Radiosonde as the target variable. Subsequently, the processed dataset (1107 records) was randomly split into training and testing subsets using an 80:20 ratio, following the data partitioning adopted in a previous study [9].

Random forest is an ensemble learning algorithm that constructs multiple decision trees using bootstrap sampling and random feature selection during the training process [14]. The final prediction is obtained by averaging the predictions from all decision trees, thereby reducing prediction variance and improving model generalization as represented in Figure 1. Owing to its ability to capture complex nonlinear relationships and interactions among predictor variables, Random Forest has been applied in estimating PBLH [9], [10], [13].

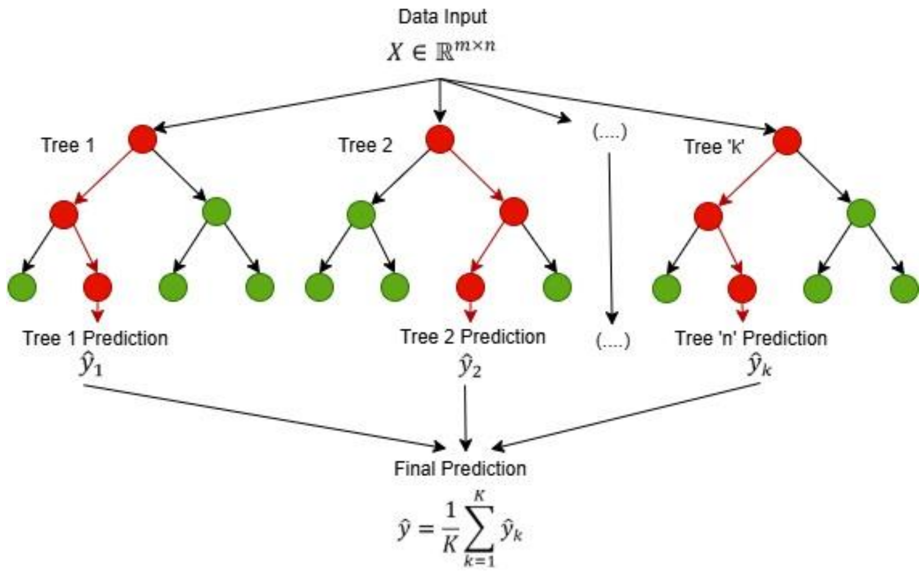


Figure 1. Random forest regressor algorithm

Figure 1. illustrates the input data for the Random Forest Regressor model are represented as $X \in R^{n\times m}$, where n denoted the number of samples and m represents the number of predictor variables. The input data are then supplied to all decision trees, although each tree is trained on a different dataset through bootstrap sampling and the random feature selection. Each regression tree is expressed as $h(x, \theta_k)$, where h denotes the regression tree function, x represents the input predictor vector, and θ_k represents the random component generated from the combination of a bootstrap sample and randomly selected subset of features. The red arrows from X to each tree in Figure 1 indicate that every tree is trained independently using different bootstrap samples and tree structures. Through this process, each tree produces a numerical prediction denoted by $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_k$. The final prediction is denoted by $\hat{y}$, which the final estimated value is obtained by averaging the predictions from all K regression trees rather than relying on the output of a single tree [14].

In this study, a random Forest Regressor model was developed in the Python programming environment using Jupyter Notebook. Random Forest was selected because of its capability to model complex nonlinear relationships and its robustness against overfitting in atmospheric and environmental prediction problems [9], [25], [26]. After constructing the model pipeline, hyperparameter optimization was performed using the Optuna framework [27]. Six hyperparameters were optimized, including *max_depth* to control tree depth, *min_samples_leaf* to set the minimum samples in a leaf node, *min_samples_split* to determine the minimum samples for node splitting, *n_estimators* to specify the number of decision trees, *random_state* to ensure reproducibility, and *n_jobs* to control parallel computation. After hyperparameter optimization, the model was employed to estimate PBLH from surface meteorological observations and PBLH ERA5.

2.3 Model Evaluation

The performance of the developed Random Forest model was evaluated using the Pearson correlation coefficient (r), coefficient of determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and *p*-value. These metrics were used to assess the agreement between the estimated PBLH and the actual PBLH in terms of correlation, explained variance, predicted error, systematic deviation, and statistical significance.

The Pearson correlation coefficient measures the strength and direction of the linear relationship between the estimated and actual PBLH as shown in Equation 2.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (2)$$

The statistical significance of the Pearson correlation coefficient was evaluated using the *p*-value. A correlation was considered statistically significant when p-value was less than 0.05 corresponding to a 95% confidence level [28]. Equation (3) presents the coefficient of determination (R^2) that quantifies the proportion of variance in the actual PBLH explained by the predicted PBLH [10]. Higher R^2 value approaching to 1, indicate better model performance and greater explanatory power [29].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (3)$$

RMSE measures the average magnitude of estimation errors, with greater emphasis on larger errors described in Equation (4) [30].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \qquad (4)$$

Moreover, MAE represents the average absolute difference between the estimated and actual PBLH as explained in the Equation (5) [30].

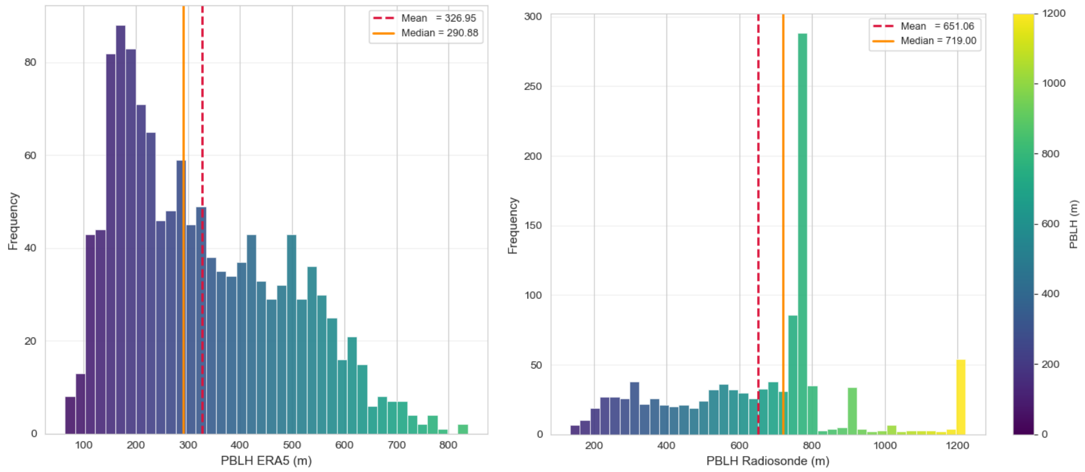
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \qquad (5)$$

Where x_i denotes the actual PBLH (PBLH Radiosonde), y_i denotes the estimated PBLH (PBLH RF), $\bar{x}$ is the mean actual PBLH, $\bar{y}$ is the mean estimated PBLH, and n is the total number of observations.

3. Results and Discussion

3.1 Characteristics of PBLH and Meteorological Variables

Before developing a Random Forest model, it is essential to examine the statistical characteristics of the PBLH datasets in this study. Understanding the distribution of the target variable provides insight into its variability, central tendency, and dispersion of the data which can influence the ability of a machine learning model to learn the relationship between surface meteorological parameters and PBLH [24]. The comparison of PBLH ERA5 and PBLH Radiosonde offers an initial assessment of the consistency between the two datasets prior to model development and validation. Figure 2. illustrates the distributions of PBLH derived from ERA5 (Figure 2.a) and the BRN method using radiosonde observations (Figure 2.b).



(a) (b)

Figure 1. (a) PBLH ERA5 and (b) PBLH Radiosonde distribution

Figure 2.a shows that the ERA5 derived PBLH exhibits a relatively continuous distribution, ranging from approximately 120 to 780 m, with a mean of 326.95 m and a median of 290.88 m. The higher mean than median indicates a slightly right skewed distribution, suggesting that several high PBLH events shift the average toward larger values. In contrast, from Figure 2.b the PBLH Radiosonde presents a markedly different distribution. Most observations are concentrated around 600-750 m, with a smaller cluster occurring at approximately 300-500 m and above 1000 m. The PBLH Radiosonde has a mean of 663.69 m and median of 719.00 m, indicating a slightly left-skewed distribution. Across nearly the entire range, PBLH ERA5 values are considerably lower than PBLH Radiosonde values, providing an initial visual indication of a systematic negative bias of the reanalysis product relative to the radiosonde-based

estimate in this region, a discrepancy that is examined quantitatively through the full-dataset validation presented in Section 3.3.

In addition, to investigate the relationship among meteorological variables and their association with PBLH, a Pearson correlation (*r*) analysis was performed using all variables. The correlation matrix provides an initial assessment of the linear relationship between predictor variables and the target variables prior to developing the Random Forest model. As shown in Figure 3., several surface meteorological variables exhibit moderate to strong correlations that are physically consistent with atmospheric processes. Air temperature variables are positively correlated with the apparent temperature (*r* = 0.76) and sunshine duration (*r* = 0.63), while relative humidity is negatively correlated with air temperature (*r* = -0.73) and sunshine duration (*r* = -0.71). Likewise, total cloud cover and low cloud cover show a strong positive correlation (*r* = 0.83), whereas both variables are negatively correlated with sunshine duration (*r* = -0.56 and -0.47). These relationships indicate that the observational dataset preserves the expected thermodynamic interactions between radiation, cloud cover, temperature, and humidity, suggesting good physical consistency among the measured surface variables [1], [3].

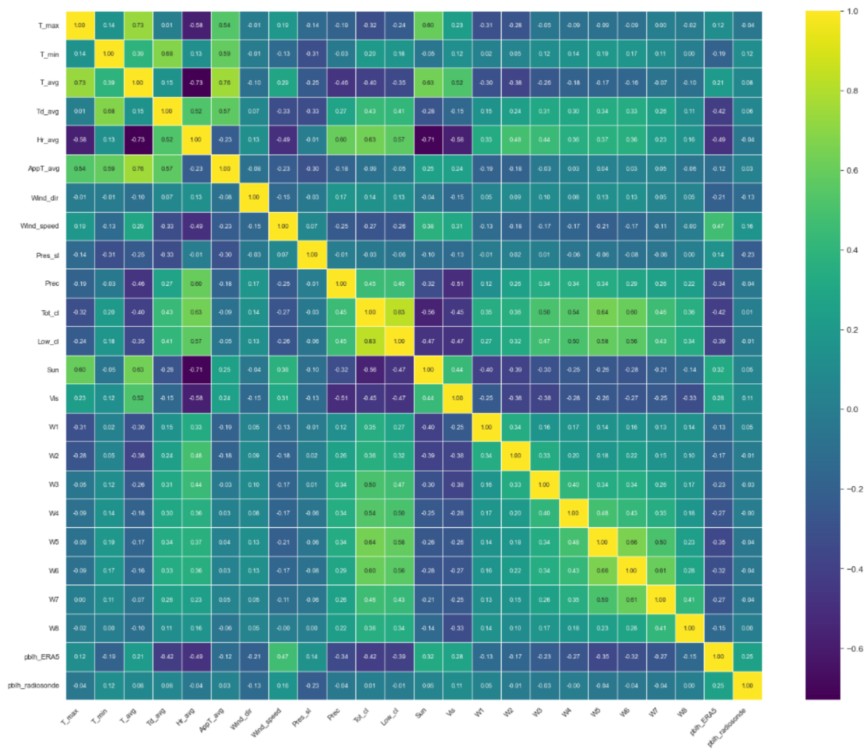


Figure 3. Correlation heatmap of all variables

Figure 3 shows that PBLH Radiosonde as the target variable, indicate only weak linear correlations with the surface meteorological parameters. Among all predictors, PBLH ERA5 exhibits the strongest correlation with PBLH Radiosonde (*r* = 0.25), followed by wind speed (*r* = 0.16), minimum temperature (*r* = 0.12), and visibility (*r* = 0.11), while sea level pressure shows the strongest negative correlation (*r* = −0.23), followed by wind direction (*r* = −0.13). The remaining surface variables, including relative humidity (*r* = −0.04), show correlation coefficients close to zero. This indicates that surface meteorological observations alone carry limited linear information about the PBLH

Radiosonde, consistent with the more complex, vertically-integrated thermodynamic and dynamic structure represented by the BRN method [8], [19]. Nevertheless, PBLH ERA5 stands out as the single most informative predictor among the available variables, which supports its inclusion as a predictor rather than as the target in the modelling framework adopted in this study, allowing the Random Forest model to combine the coarse but physically consistent signal of ERA5 with surface observations to approximate the radiosonde-based estimate.

This weak linear relationship between PBLH Radiosonde and surface observations does not necessarily imply that the surface variables are uninformative. Rather, it reflects the fact that boundary layer height derived from the Bulk Richardson Number method depends on the full vertical thermodynamic and dynamic profile of the atmosphere, a structure not fully captured by instantaneous surface measurements alone [8]. At the same time, PBLH ERA5 itself carries a systematic bias relative to independent radiosonde observations, particularly over tropical regions during the dry season, arising from limitations in the boundary layer parameterization scheme used in the Integrated Forecasting System [22]. Consequently, the Random Forest model in this study was designed to learn a nonlinear mapping from surface meteorological observations and PBLH ERA5 toward the radiosonde based target, effectively functioning as a data driven bias correction of PBLH ERA5 rather than as an independent PBLH estimator built purely from surface observations.

This difference between PBLH ERA5 and PBLH Radiosonde is consistent with previous studies on PBLH ERA5 compared to independent observations. Guo et al. [13] reported a systematic overestimation of PBLH ERA5 relative to radiosonde-based estimates by approximately 90 m when compared across multiple stations, while Xi et al. [11] found that PBLH ERA5 bias against MOSAiC radiosonde observations varied significantly depending on PBL fluctuation, cloud cover, and season. These findings support the interpretation that the weak correlation between PBLH Radiosonde and surface meteorological variables observed in this study does not solely originate from limitations of surface-based predictors, but also reflects an inherent structural difference between the PBLH Radiosonde estimation and the ERA5 reanalysis diagnosis of PBLH.

3.2 Random Forest Model Development and Performance Evaluation

The Random Forest model was developed using a machine learning pipeline implemented in Python. Prior to model training, hyperparameter optimization was performed using the Optuna framework to identify the optimal combination of model parameters and improve prediction performance [9], [27]. The optimization process resulted in a maximum tree depth (*max_depth*) of 31, a minimum number of samples required at a leaf node (*min_samples_leaf*) of 4, and a minimum number of samples required to split an internal node (*min_samples_split*) of 10. In addition, the optimal model consisted of 134 decision trees (*n_estimators* = 134) with a fixed random seed (*random_state* = 18) to ensure reproducibility, while the number of parallel processing jobs (*n_jobs*) was set to -1. The optimized hyperparameters were subsequently used to train the final Random Forest model to estimate PBLH from surface meteorological observations. The model performance was then evaluated through feature importance analysis (Figure 4.), prediction accuracy assessment (Figure 5.), residual analysis (Figure 6.), and statistical evaluation metrics (Table 1.).

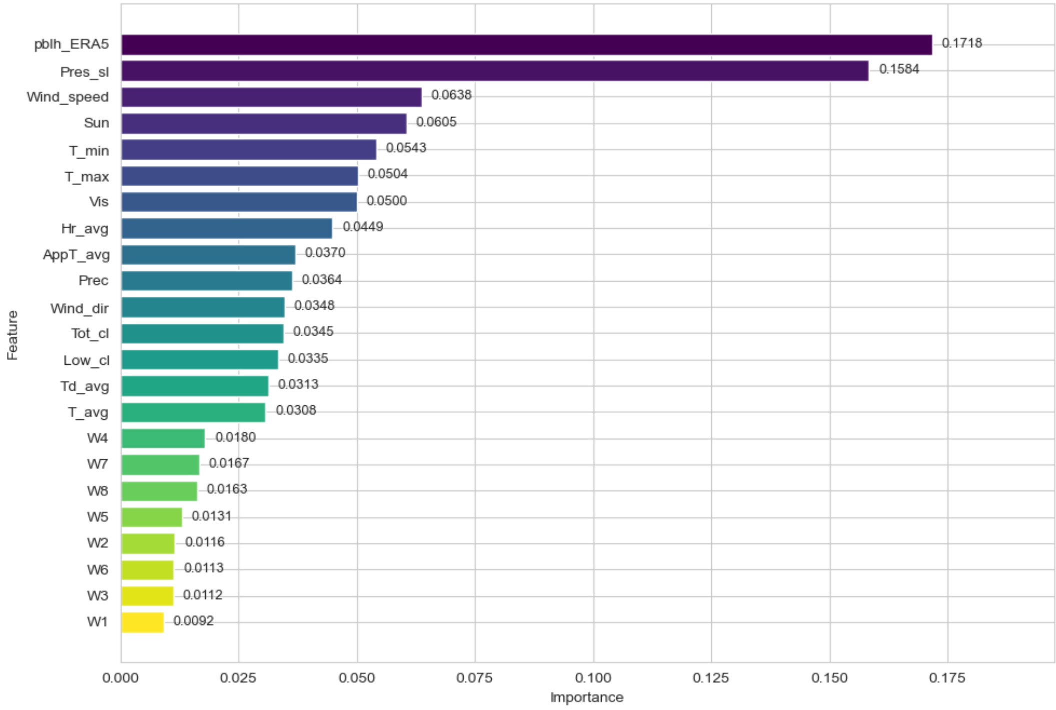


Figure 4. Feature importance of random forest model

Figure 4. presents the relative importance of each predictor variable in estimating PBLH Radiosonde using the optimized Random Forest model. PBLH ERA5 was identified as the most influential predictor, contributing 17.18% of the total feature importance, followed by sea level pressure (15.84%), wind speed (6.38%), sunshine duration (6.05%), and minimum temperature (5.43%). Maximum temperature (5.04%), visibility (5.00%), relative humidity (4.49%), apparent temperature (3.70%), precipitation (3.64%), wind direction (3.48%), total and low cloud cover (3.45% and 3.35%), dew point temperature (3.13%), and average temperature (3.08%) provided moderate contributions, while the present-weather condition variables (W1–W8) each contributed less than 2%.

The dominance of PBLH ERA5 as the leading predictor confirms that, despite its systematic bias, the reanalysis product retains meaningful information regarding boundary layer evolution that the Random Forest model exploits as an anchor before adjusting it using surface observations. The high importance of sea level pressure and wind speed indicates that synoptic-scale pressure variation and dynamical mixing play an important role in explaining the discrepancy between PBLH ERA5 and the radiosonde-based estimate, while the moderate contribution of sunshine duration and temperature variables reflects the influence of surface heating on boundary layer development [4], [8]. Compared with a surface-only configuration, the substantial share of importance attributed to PBLH ERA5 illustrates that the model developed in this study behaves as a hybrid reanalysis correction estimator rather than a purely surface-based PBLH estimator.

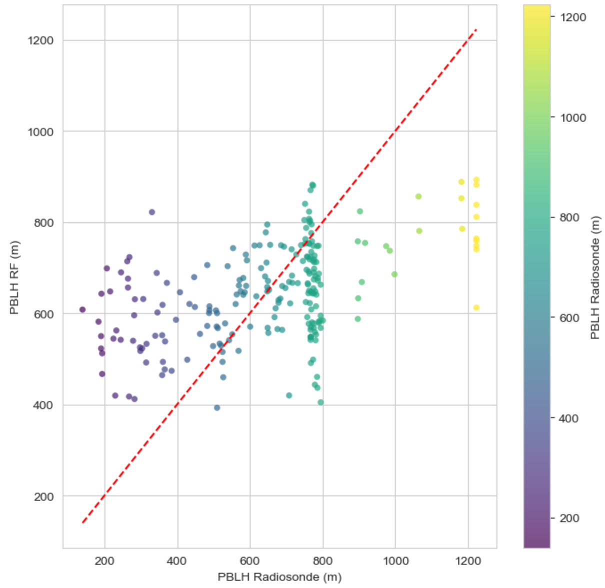


Figure 5. Scatter plot of PBLH Radiosonde (actual) vs PBLH RF (prediction)

Figure 5. compares the Random Forest estimated PBLH (PBLH RF) with the PBLH Radiosonde on the independent test set. The predicted values show a positive association with the 1:1 line (red line), but with considerable scatter. For lower radiosonde values (below approximately 400 m) the model tends to overestimate PBLH, predicting values clustered around 500–700 m, whereas for higher radiosonde values (above approximately 900 m) the model systematically underestimates PBLH, rarely predicting above approximately 900 m even when the observed value exceeds 1200 m. This compression pattern, in which predictions regress toward the mean of the training target, is characteristic of ensemble tree-based regressors and has been reported in comparable studies [14], [29]. This behavior is consistent with the moderate R^2 (0.2478) and RMSE (207.85 m) obtained on the test set (Table 1) and indicates that, although the Random Forest model captures the general tendency of PBLH Radiosonde, its skill in reproducing the most extreme boundary layer heights remains limited when only routine surface meteorological observations and PBLH ERA5 are used as predictors. Furthermore, the comparison of residual RF prediction and ERA5 values is presented in Figure 6.

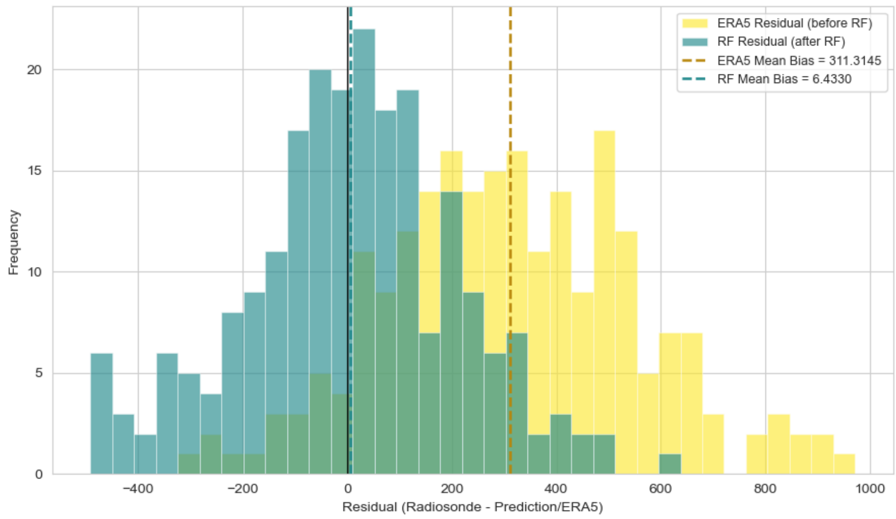


Figure 6. Residual of PBLH RF and PBLH ERA5 against PBLH Radiosonde

Figure 6. compares the residuals of the Random Forest prediction (Radiosonde − RF) with the residuals of the raw PBLH ERA5 (Radiosonde − ERA5) on the test set. The residuals of raw PBLH ERA5 are almost entirely positive, with a mean bias of 311.31 m, confirming a strong systematic underestimation of PBLH Radiosonde by the reanalysis product. In contrast, the residuals of the Random Forest prediction are centered close to zero, with a mean bias of only 6.43 m, a reduction of 304.88 m, equivalent to a 97.9% decrease in systematic bias relative to raw PBLH ERA5 on the test set. The narrower and more symmetric distribution of the Random Forest residuals in Figure 6. indicates that the model largely succeeds in removing the systematic component of the ERA5 bias, although a considerable random error component (RMSE = 207.85 m) remains, consistent with the underestimation of extreme values observed in Figure 5.

The model performance evaluation on the independent test set is summarized in Table 1. The model achieved a Pearson correlation coefficient of 0.5020 ($p < 0.001$), indicating a moderate but statistically significant positive relationship between the Random Forest prediction and PBLH Radiosonde. The coefficient of determination ($R^2 = 0.2478$) indicates that approximately 24.8% of the variability in PBLH Radiosonde can be explained by the model when only routine surface meteorological observations and PBLH ERA5 are used as predictors. The model produced an RMSE of 207.85 m and an MAE of 162.49 m, considerably lower than the corresponding errors of raw PBLH ERA5 against radiosonde on the same test set (RMSE = 393.70 m, MAE = 331.43 m; see Section 3.3), indicating that although the model ability to independently explain radiosonde variability is moderate, it still provides a substantial reduction in error relative to using PBLH ERA5 without correction.

Table 1. PBLH Random Forest Evaluation

Model Evaluation	Score
RMSE	207.85
MAE	162.49
R^2	0.2478
Pearson (r)	0.5020
p-value	1.4205e-15

These results should be interpreted differently from previous studies that trained Random Forest models directly against reanalysis- or remote-sensing-derived PBLH. Molero et al. [10], for instance, reported a considerably higher R^2 of 0.66 on the test set when Random Forest was trained against ceilometer-derived boundary layer height using ground-level meteorological predictors. Silva et al. [9] similarly reported stronger explanatory power when the target was a remote-sensing-derived PBLH rather than an independently observed radiosonde profile. This is consistent with the stronger correlation between ERA5 and surface meteorological variables identified in Section 3.1. In this study, because the target was instead the independently observed PBLH Radiosonde, which is only weakly correlated with surface variables, a more moderate R^2 (0,2478) was expected, and a more meaningful benchmark than a surface-only model is the raw PBLH ERA5 baseline itself. Evaluated against this benchmark, the substantial bias reduction achieved by the model (Figure 6.) is comparable to the improvement reported by Xi et al. [11] and Guo et al. [13], both of which used machine learning to correct PBLH ERA5 toward radiosonde observations using near-surface and land-surface predictors. This supports the interpretation that the model developed in this study functions primarily as a bias-correction tool for PBLH ERA5 rather than as a standalone surface-based PBLH estimator, and that relative humidity, wind, and pressure remain physically relevant predictors of PBL development even when their direct correlation with the radiosonde target is weak, consistent with Silva et al. [9] and Peng et al. [25].

3.3 Temporal Validation of the Estimated PBLH

To further evaluate the developed Random Forest model, temporal validation was performed by comparing the predicted PBLH RF, PBLH ERA5, and PBLH Radiosonde. This validation utilizes the entire dataset rather than relying solely on a test set to assess the overall consistency among the three PBLH sources and to measure the extent to which the Random Forest model corrects ERA5 PBLH bias relative to independent radiosonde observations. It should be noted that, unlike Table 1, this validation includes the training records seen by the model, reflecting the model's overall bias-correction capability rather than independent generalization performance. Daily time series (Figure 7) and monthly averages (Figure 8) were analyzed to assess the ability of each PBLH source to reproduce the temporal variation of PBLH Radiosonde during the study period.

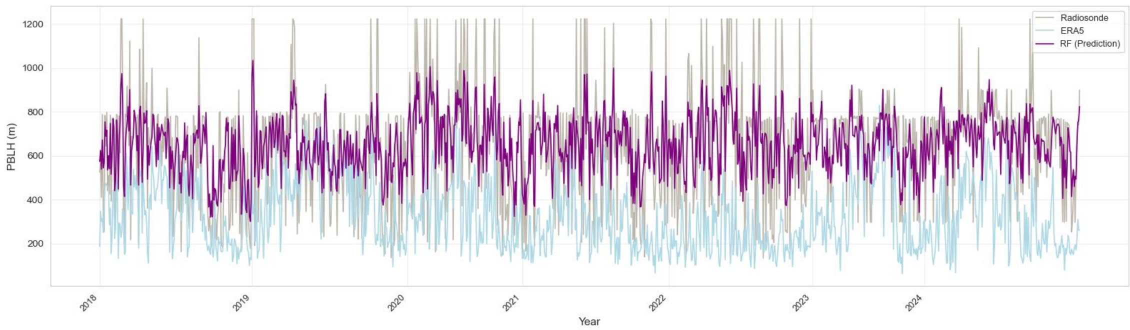


Figure 7. Timeseries comparison of PBLH ERA5, RF, and Radisonde (June - November)

Figure 7. illustrates the daily timeseries of PBLH ERA5, RF and Radiosonde during the dry season (June – November) from 2018 to 2024. Across the full dataset, PBLH RF closely tracks the temporal evolution of PBLH Radiosonde, including many of the short-duration peaks exceeding 1000 m, while PBLH ERA5 remains persistently and

substantially lower throughout nearly the entire period. This close agreement between PBLH RF and PBLH Radiosonde is expected given that the majority of these records (885 of 1107, or 80%) were used during model training. Nonetheless, the figure clearly demonstrates the magnitude of the correction applied by the Random Forest model relative to the raw ERA5 signal, which otherwise underestimates PBLH throughout almost the entire study period.

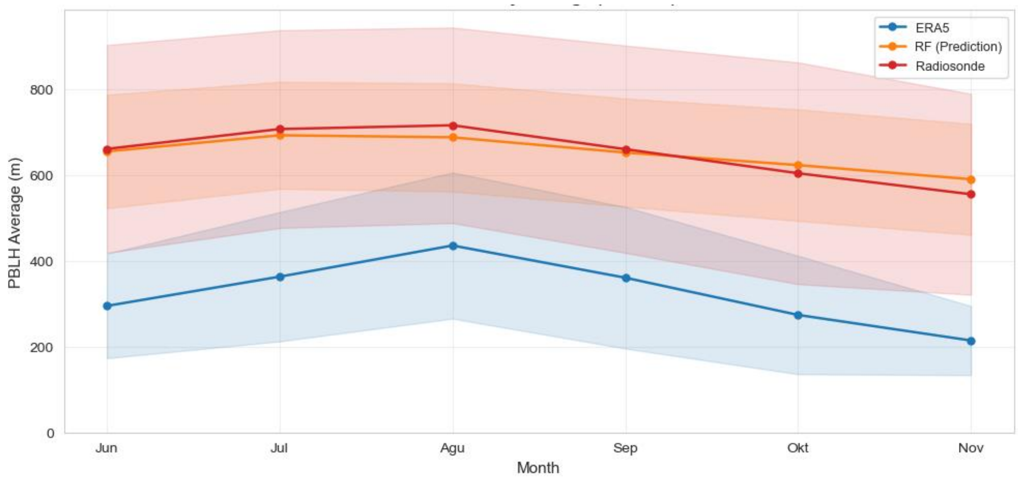


Figure 8. PBLH monthly average (June – November)

Figure 8. presents the monthly average of PBLH ERA5, PBLH RF, and PBLH Radiosonde during the dry season. PBLH Radiosonde and PBLH RF show a consistent seasonal cycle, both peaking in August (mean 715.5 m and 687.4 m) and gradually decreasing toward November (589.8 m and 554.8 m). Meanwhile, PBLH ERA5 follows a broadly similar seasonal shape but at a substantially lower level throughout the season, ranging from approximately 220 to 440 m. The close agreement between the monthly mean RF PBLH and radiosonde PBLH contrasts with the consistent discrepancy between ERA5 PBLH and radiosonde PBLH, demonstrating the Random Forest model ability to reproduce radiosonde-based seasonal mean PBLH values, in addition to the daily variability shown in Figure 7.

Table 2. PBLH Validation

PBLH	Pearson (r)	p-value	R^2	RMSE	MAE	Bias
ERA5 vs Radiosonde	0.2507	2.4994e-17	-1.8558	414.34	346.36	-325.35
RF vs ERA5	0.3909	9.8853e-42	-4.2207	363.01	328.29	324.72
RF vs Radiosonde	0.8238	1.2963e-274	0.5982	155.41	117.27	-0.63

The statistical evaluation presented in Table 2. confirms the temporal comparison shown in Figures 7. and 8. PBLH ERA5 shows a very weak and effectively non-predictive relationship with PBLH Radiosonde across the full dataset together with a large negative mean bias of −325.35 m, RMSE of 414.34 m, and MAE of 346.36 m, confirming a strong and systematic underestimation of the radiosonde-based PBLH. The comparison between PBLH RF and PBLH ERA5 (r = 0.3909, R^2 = −4.2207) further illustrates that the Random Forest prediction has substantially diverged from the original ERA5 signal that partly informed it, in the direction of the radiosonde observations, consistent with the intended bias-correction behavior of the model. In marked contrast, PBLH RF shows a strong and highly significant agreement with PBLH Radiosonde (r = 0.8238, p < 0.001, R^2 = 0.5982) with mean bias of only −0.63 m, a reduction of 324.72 m (99.8%) relative to the raw ERA5 bias. As noted above, this strong full-dataset agreement partly reflects the fact that most of the records evaluated in Table 2. were also used during model training, so Table 2. should be interpreted as a measure of overall bias-correction capability, while the independent test-set metrics in Table 1. remain the more conservative estimate of the model's true predictive skill.

The magnitude of the −325.35 m ERA5 bias found for Central Kalimantan is broadly consistent with previous radiosonde-based evaluations, although its direction varies by region. The bias in this study is consistent in sign with Guo et al. [16] and Kumar et al. [17] that underestimations of approximately 130 m and 200–650 m and falls toward the upper end of these reported magnitudes, suggesting a comparatively large ERA5 underestimation over this tropical peatland region. This contrast with the Arctic overestimation reported by Xi et al. [11], together with the systematic regional differences noted by Guo et al. [13], indicates that the ERA5 PBLH bias is not a fixed offset but depends strongly on regional surface, land-cover, and boundary-layer characteristics.

The substantial bias reduction achieved by the Random Forest model in this study is broadly consistent with improvements reported when machine learning is used to correct ERA5 PBLH toward radiosonde or other independent references, including the index of agreement of up to 0.71 achieved by Xi et al. [11] and the reduced mean deviation reported by Guo et al. [13]. However, the model's test-set R^2 (0.2478) is more moderate than the R^2 reported when Random Forest is trained directly against ERA5-derived PBLH [9], [10]. This shows that correcting reanalysis PBLH toward independent radiosonde observations is harder than merely reproducing ERA5 itself, especially for extreme boundary layer heights, which were also underestimated by RF models over Central Amazonia [9].

4. Conclusions

This study demonstrates that a Random Forest model combining surface meteorological observations with PBLH ERA5 can estimate and, more importantly, correct the systematic bias of PBLH ERA5 against independently observed PBLH Radiosonde in Central Kalimantan during the dry season. On the independent test set, the model achieved r = 0.5020, R^2 = 0.2478, RMSE = 207.85 m, and MAE = 162.49 m, with a 97.9% reduction in ERA5's systematic bias. Across the full dataset, agreement was stronger (r = 0.8238, R^2 = 0.5982), reducing the mean bias from −325.35 m to just −0.63 m (99.8% reduction), though this evaluation is not fully independent of the training data. Feature importance analysis identifies PBLH ERA5, sea level pressure, and wind speed as the dominant predictors, indicating the model functions primarily as a data-driven bias-correction of the reanalysis product rather than a purely surface-based estimator. Although still limited in reproducing extreme PBLH Radiosonde values, the model offers a practical, cost-effective way to improve reanalysis-based PBLH consistency with observations in data-sparse regions like Central Kalimantan. Future work should apply independent, temporally or station-blocked validation, add further atmospheric predictors, and test transferability across seasons and locations

ACKNOWLEDGMENT

The authors gratefully acknowledge the Department of Meteorology, State College of Meteorology, Climatology, and Geophysics for its academic support and research facilities throughout this study. The authors also thank the Iskandar Meteorological Station, Kotawaringin Barat for providing the surface meteorological and radiosonde observation data that were indispensable to this research.

REFERENCES

1. R. B. Stull, An Introduction to Boundary Layer Meteorology. Dordrecht: Kluwer Academic Publishers, 1988.
2. S. Liu and X.-Z. Liang, “Observed Diurnal Cycle Climatology of Planetary Boundary Layer Height,” J. Clim., vol. 23, no. 21, pp. 5790–5809, 2010.
3. J. M. Wallace and P. V Hobbs, Atmospheric Science: an Introductory Survey, vol. 92. Elsevier, 2006.
4. J. A. Mantovani Júnior, J. A. Aravéquia, R. G. Carneiro, and G. Fisch, “Evaluation of PBL Parameterization Schemes in WRF Model Predictions During the Dry Season of the Central Amazon Basin,” Atmosphere (Basel)., vol. 14, no. 5, p. 850, 2023.
5. A. Sullivan, “Inside the Inferno: Fundamental Processes of Wildland Fire Behaviour: Part 2: Heat Transfer and Interactions,” Current Forestry Reports, vol. 3, pp. 150–171, Jun. 2017, doi: 10.1007/s40725-017-0058-z.
6. M. A. Rahman et al., “Weather Radar Detection of Planetary Boundary Layer and Smoke Layer Top of Peatland Fire in Central Kalimantan, Indonesia,” Sci. Rep., vol. 11, no. 1, p. 367, 2021.
7. J. Gu, Y. Zhang, N. Yang, and R. Wang, “Diurnal Variability of The Planetary Boundary Layer Height Estimated from Radiosonde Data,” Earth and Planetary Physics, vol. 4, no. 5, pp. 479–492, 2020.
8. D. J. Seidel, C. O. Ao, and K. Li, “Estimating Climatological Planetary Boundary Layer Heights from Radiosonde Observations: Comparison of Methods and Uncertainty Analysis,” Journal of Geophysical Research: Atmospheres, vol. 115, no. D16, 2010.
9. P. R. P. Silva, R. G. Carneiro, A. O. Moraes, C. Q. Dias-Junior, and G. Fisch, “Estimating Planetary Boundary Layer Height over Central Amazonia Using Random Forest,” Atmosphere (Basel)., vol. 16, no. 8, p. 941, 2025, doi: 10.3390/atmos16080941.
10. F. Molero, R. Barragán, and B. Artíñano, “Estimation of The Atmospheric Boundary Layer Height by Means of Machine Learning Techniques Using Ground-Level Meteorological Data,” Atmos. Res., vol. 279, p. 106401, 2022.
11. X. Xi et al., “Evaluation of the Planetary Boundary Layer Height from ERA5 Reanalysis with MOSAiC Observations over the Arctic Ocean,” Journal of Geophysical Research: Atmospheres, vol. 129, no. 12, p. e2024JD040779, 2024.
12. R. Krishnamurthy, R. K. Newsom, L. K. Berg, H. Xiao, P.-L. Ma, and D. D. Turner, “On The Estimation of Boundary Layer Heights: A Machine Learning Approach,” Atmospheric Measurement Techniques Discussions, vol. 2020, pp. 1–34, 2020, doi: 10.5194/amt-14-4403-2021.
13. J. Guo et al., “A Merged Continental Planetary Boundary Layer Height Dataset Based on High-Resolution Radiosonde Measurements, ERA5 Reanalysis, and GLDAS,” Earth Syst. Sci. Data, vol. 16, no. 1, pp. 1–14, 2024.
14. L. Breiman, “Random Forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

15. M. Virman, M. Bister, J. Räisänen, V. A. Sinclair, and H. Järvinen, "Radiosonde comparison of ERA5 and ERA-Interim reanalysis datasets over tropical oceans," Tellus A: Dynamic Meteorology and Oceanography, vol. 73, no. 1, pp. 1–7, 2021.
16. J. Guo et al., "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," Atmos. Chem. Phys., vol. 21, no. 22, pp. 17079–17097, Nov. 2021, doi: 10.5194/acp-21-17079-2021.
17. A. H. Kumar, A. Sandeep, P. Prasad, and B. Bv, "Evaluation of ERA5 boundary layer height against radiosonde measurements: Regional drivers, spatiotemporal variability, and trends over the Indian Monsoon Region," Atmos. Res., p. 108511, 2025.
18. J. T. Lestari, E. H. Sujiono, and M. Arsyad, "Analisis Hubungan Ketinggian Planetary Boundary Layer (PBL) dengan Surface Flux di Makassar dan Sekitarnya," Jurnal Fisika Unand, vol. 13, no. 4, pp. 459–468, 2024.
19. Ogimet, "Ogimet: Professional Information about Meteorological Conditions in the World," Ogimet. Accessed: Jul. 09, 2026. [Online]. Available: https://www.ogimet.com/
20. University of Wyoming, "University of Wyoming Atmospheric Science Radiosonde Archive," Wyoming Weather Web. Accessed: Jul. 09, 2026. [Online]. Available: https://weather.uwyo.edu/upperair/sounding.shtml
21. D. H. P. Vogelezang and A. A. M. Holtslag, "Evaluation and Model Impacts of Alternative Boundary-Layer Height Formulations," Boundary. Layer. Meteorol., vol. 81, no. 3, pp. 245–269, 1996.
22. European Centre for Medium-Range Weather Forecasts, "IFS Documentation CY49R1: Part IV – Physical Processes," ECMWF, 2024, doi: https://doi.org/10.21957/c731ee1102.
23. Z. Li et al., "A Temperature Refinement Method Using the ERA5 Reanalysis Data," Atmosphere (Basel)., vol. 13, no. 10, p. 1622, 2022.
24. M. Komorowski, D. Marshall, J. Salciccioli, and Y. Crutain, "Exploratory Data Analysis," in Secondary Analysis of Electronic Health Records, 2016, pp. 185–203. doi: 10.1007/978-3-319-43742-2_15.
25. K. Peng et al., "Machine Learning Model to Accurately Estimate The Planetary Boundary Layer Height of Beijing Urban Area with ERA5 Data," Atmos. Res., vol. 293, p. 106925, 2023.
26. D. Zhang et al., "Best Estimate of The Planetary Boundary Layer Height from Multiple Remote Sensing Measurements," Atmos. Meas. Tech., vol. 18, no. 14, pp. 3453–3475, 2025.
27. T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-Generation Hyperparameter Optimization Framework," in Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.
28. R. L. Wasserstein and N. A. Lazar, "The ASA Statement on p-values: Context, Process, and Purpose," 2016, Taylor & Francis.
29. G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, vol. 103. Springer, 2013.
30. D. Wilks, Statistical Methods in the Atmospheric Sciences Second Edition. Oxford: Elsevier, 2006.