Predicting Boundary Layer Height Using Neural Networks: A Practical Approach

Zhuorui He

June 19, 2024

Abstract

Determining the atmospheric boundary layer height (BLH) is essential for air quality evaluation and weather forecasting. This study analyzes data from the Kansas Field Station of the AmeriFlux network to forecast BLH using neural networks. Three models were assessed: simple linear regression, a recurrent neural network (RNN), and an improved RNN with hyperparameter tuning and k-fold cross-validation. The simple RNN outperformed the linear regression model in terms of Mean Absolute Error(MAE), but it had difficulties with variance and outliers. With reduced Mean Square Erro (MSE) and MAE, the improved RNN demonstrated notable gains. The high overall MSE, however, suggests that more advancements are required. To improve forecast accuracy, future research should address data quality and include more advanced modeling techniques.

Introduction

The Earth's climate system and weather patterns are significantly influenced by the atmospheric boundary layer (ABL). Being at the lowest point in the atmosphere, it is immediately impacted by the Earth's surface. A crucial factor in meteorology and climatology is the height of the boundary layer, also referred to as the boundary layer height (BLH). It affects climate models, weather forecasts, and evaluations of air quality by calculating the volume of air in which heat, momentum, and contaminants are combined. There are notable variations in air movement above and below the boundary layer's top, with the boundary layer being the primary location for turbulence and mixing. [1, 2].

Because of the intricate movements and interactions that occur inside the boundary layer, it is crucial to comprehend and forecast BLH. Turbulent fluxes that arise from the interaction between the surface and the atmosphere above are what define the

ABL. A highly dynamic environment where temperature, humidity, wind speed, and pressure can change quickly and unexpectedly is created by these turbulent motions. Accurately measuring and modeling the behavior of the boundary layer is difficult due to its complexity. The majority of measurements are made near the surface, and the few soundings and satellite observations that are accessible above make it more challenging to get a comprehensive picture of the dynamics of the boundary layer. [3, 4].

Accurately measuring the factors that characterize the state of the boundary layer is one of the major challenges in researching it. Commonly used tools for observing BLH include lidars and radiosondes, although the accuracy and precision of the measurements as well as the geographical and temporal resolution of these methods might be problematic. Furthermore, due to the boundary layer's extreme variability, even tiny measurement mistakes might result in considerable BLH estimation errors [5].

We choose to use neural networks due to their capability to predict the state without requiring known physical formulas. Machine learning, though it has obstacles similar to physical models, can predict target variables using other measurements not directly related to the target. This allows for the incorporation of a wide range of observational data, potentially improving the accuracy and reliability of BLH predictions even when direct measurements are sparse or noisy [6, 7].

In this work, we investigate the prediction of boundary layer height using neural networks. Because of their capacity to infer intricate patterns from data, neural networks present a viable way around some of the problems with conventional modeling and measurement methods. Our goal is to create a prediction model that can yield dependable and accurate estimations of BLH by utilizing historical data and advanced machine-learning methods. This model will ultimately help to enhance climate models and weather forecasts. [6, 8].

Methods

Data Source

The Kansas Field Station on the AmeriFlux network provided primary data for this investigation. AmeriFlux is an extensive network of research sites that monitors water, energy, and carbon dioxide fluxes in the American ecosystems. Since its founding in 1996, the AmeriFlux network has provided high-temporal resolution observations from a variety of ecosystems, including tundra, grasslands, savannas, and forests, to support a broad range of ecological and climate-related studies. The network seeks to offer vital information for comprehending the processes of terrestrial ecosystems and how they react to climate change. At the Kansas Field Station, data were gathered for this study half-hourly between January 1, 2017, and December 31, 2018. Megan Metz handled the data's initial processing, guaranteeing its quality.[15].

Preprocessing of Data

The atmospheric data was highly variable and complex, thus a great deal of preprocessing was required to get the information ready for analysis. Among the crucial actions in this process were:

1. Filtering Unrealistic Data: We plotted the PBLH data to find values that were out of the ordinary. We used common boundary layer heights and real-world physics to filter the data so that only values between 0 and 2500 meters were included.

2. Managing Missing Values: In order to prevent interference, we eliminated variables that had more than 50% missing values from the analysis. The missing data points for the other variables were filled in using the proper imputation techniques.

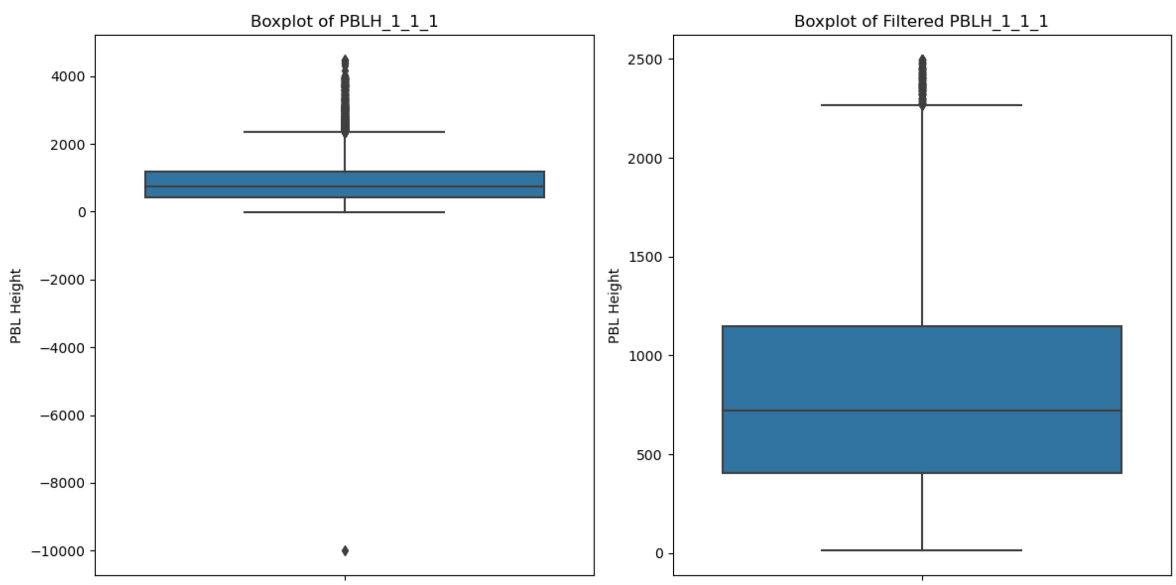

Figure 1: PBLH data comparison before and after filtering

Identification of Related Variables

To determine the impact of each variable on PBLH, we ran a correlation check against PBLH. Significant variables were those with a correlation coefficient greater than 0.2 in absolute value. It is assumed that there is a linear relationship between these factors and BLH. Among the variables that emerged were listed below in descending order:

- Friction Velocity (USTAR)

- Air Temperature (TA)
- Net Ecosystem Exchange (NEE)
- Sensible Heat Flux (H)
- Latent Heat Flux (LE)
- Net Radiation (NETRAD)
- Shortwave Radiation In (SW_IN)
- Shortwave Radiation Out (SW_OUT)
- Longwave Radiation Out (LW_OUT)
- Gross Primary Production (GPP)

Now we check the physical basis of the relationship between them and PBLH. Friction velocity (USTAR) is critical for understanding turbulent mixing in the boundary layer, which directly affects BLH by enhancing vertical mixing of heat and momentum. Studies have shown that higher friction velocities correlate with increased turbulent mixing and, consequently, a higher boundary layer height [9, 1].

Air temperature (TA) influences the thermal structure of the boundary layer, with higher temperatures generally promoting a deeper boundary layer due to increased buoyancy and turbulence. The correlation between air temperature and BLH is well-documented, with warmer surface temperatures leading to greater thermal instability and a higher boundary layer [4, 10].

Net ecosystem exchange (NEE) and gross primary production (GPP) are responses to changes in BLH rather than drivers. These variables reflect the ecosystem's carbon flux dynamics, which can be influenced by the boundary layer's height. While they may not drive BLH directly, their inclusion as features in the model can provide additional context and improve the accuracy of BLH predictions [11, 12].

Sensible heat flux (H) and latent heat flux (LE) are essential for determining the energy balance at the surface, which in turn drives the thermal dynamics of the boundary layer. Sensible heat flux (H) directly heats the air, while latent heat flux (LE) involves the phase change of water, both processes contributing to the vertical movement and mixing of air [2, 5].

Net radiation (NETRAD) represents the balance of incoming and outgoing radiation, impacting surface heating and subsequently boundary layer development. Thermodynamically, net radiation changes the amount of energy at the surface, heating the air and promoting convective processes that cause the boundary layer to expand.

The relationship between the longwave and shortwave radiation components (LW_OUT, SW_IN, and LW_OUT) provides further information about the radiative processes influencing the boundary layer height and, in turn, the surface energy budget [2, 13]. This mechanism has a significant impact on the height and dynamics of the boundary layer by affecting its stability and mixing.

These variables were selected as features for the neural network model, given their significant correlation with BLH and their relevance to boundary layer dynamics.

Recurrent Neural Network (RNN) Model

A simple Recurrent Neural Network (RNN) was employed to predict PBLH using the identified significant variables. RNNs are well-suited for time series data because they process a sequence step-by-step, maintaining an internal state that can capture temporal dependencies. The architecture of the RNN model includes:

- **Input Layer**: Accepts the values of the identified significant variables.
- **Hidden Layers**: The model included eight hidden layers, each employing a ReLU activation function to introduce non-linearity and enhance learning capabilities.
- **Output Layer**: A single neuron output layer was used to predict the PBLH.

The model was trained using the Adam optimizer, which is effective for training deep learning models due to its adaptive learning rate capabilities and robustness. The Mean Squared Error (MSE) loss function was used to quantify the difference between the predicted and actual PBLH values during training. The use of RNNs for time series prediction is supported by their ability to learn and remember long-term dependencies in the data, which is crucial for accurate forecasting in dynamic environments [14].

The model's performance was evaluated using a separate test set, and the MSE and Mean Absolute Error (MAE) were calculated to assess the accuracy of the predictions. Additionally, the true versus predicted PBLH values were plotted to visually inspect the model's performance.

Results

The performance of the boundary layer height (BLH) prediction was evaluated using three different models: a simple linear regression, a simple recurrent neural network (RNN), and an RNN with k-fold cross-validation and hyperparameter tuning.

In the first model, a basic linear regression was used to predict BLH. The performance metrics indicated poor results, with a MSE of 195320 and a Mean Absolute

Error (MAE) of 358.63. The linear regression model exhibited a significant limitation: the predictions clustered between 500 and 1500 meters, failing to accurately predict higher BLH values. This clustering suggests that the assumption of a linear relationship between the variables and BLH is insufficient for capturing the complexities of the atmospheric data.

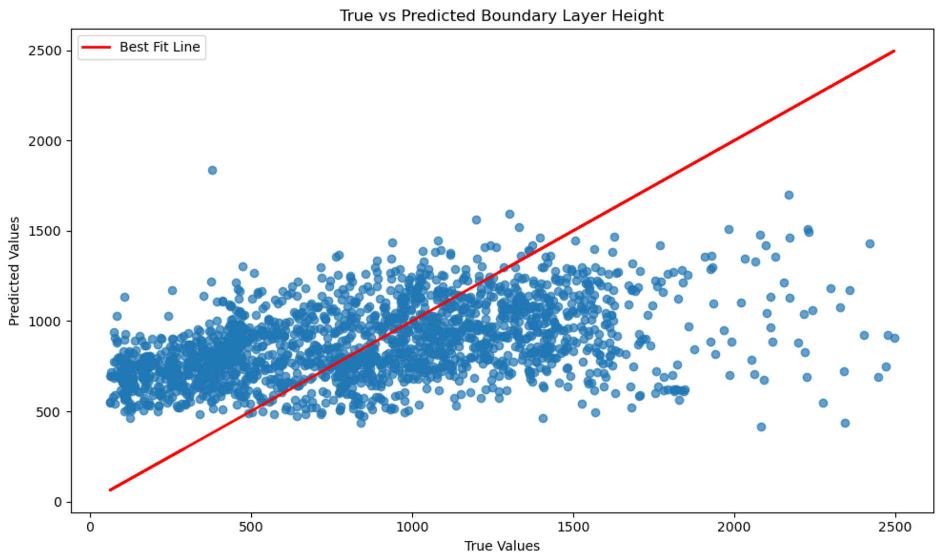

Figure 2: True vs Predicted Boundary Layer Height using Linear Regression.

The second model employed a simple RNN to capture the temporal dependencies in the data. The RNN model showed some improvement over the linear regression model, with a lower MSE of 152563 and a better MAE of 307.80. However, the variance in the predictions was high, and outliers were evident in the results. Additionally, the simple RNN tended to predict higher BLH values when they were supposed to be lower and lower BLH values when they were supposed to be higher. This discrepancy indicates that while the RNN could better handle the general patterns in the data, it struggled with accurately predicting the extremes [14].

The third model enhanced the RNN by incorporating k-fold cross-validation and hyperparameter tuning. K-fold cross-validation involves dividing the data into k subsets and training the model k times, each time using a different subset as the validation set and the remaining k-1 subsets as the training set. This technique helps assess the model's performance more reliably and reduces overfitting [16, 6]. Hyperparameter tuning involves optimizing the model parameters, such as the number of layers, learning rate, and batch size, to improve performance [17]. The enhanced RNN model with these techniques resulted in significant improvements, achieving an MSE of 96135 and an MAE of 244.01. The variance of the predictions decreased, suggesting that the model became more robust and reliable. This approach allowed the model to better

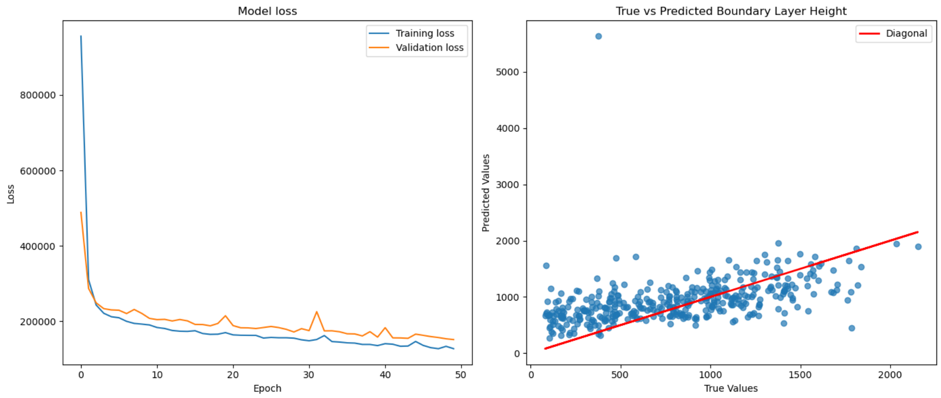

Figure 3: Model loss and True vs Predicted Boundary Layer Height using Simple RNN.

generalize from the training data to unseen data, addressing the issues seen in the previous models. The performance ranking from worst to best is: linear regression, simple RNN, and enhanced RNN with k-fold cross-validation.

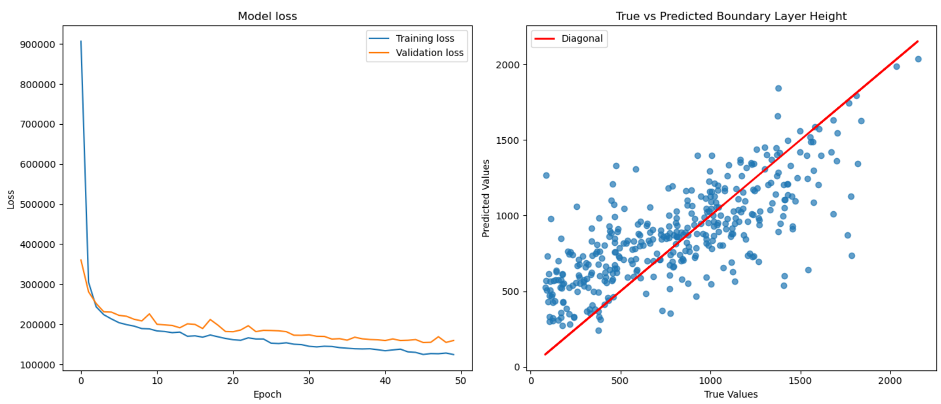

Figure 4: Model loss and True vs Predicted Boundary Layer Height using RNN with K-Fold Cross-Validation and Hyperparameter Tuning.

Overall, the high MSE across all models indicates that there is still significant room for improvement in predicting BLH. However, the progressive decrease in MSE and MAE from the linear regression model to the enhanced RNN model demonstrates the benefits of using more sophisticated models and techniques such as RNNs and cross-validation. The trend observed in the training and testing error graphs suggests that overfitting was not an issue in these models, further validating the effectiveness of the preprocessing and modeling approaches employed.

Discussion

The performance of the best model so far, though improved, is still poor. One major reason for this is the initial assumption that the relationships between the variables and the PBLH are linear. This assumption does not hold true for some of the variables, such as friction velocity. A more nuanced approach is needed to evaluate the degree of correlation for each variable and possibly adopt non-linear methods or transformations to better capture these relationships.

Another limitation arises from the constrained set of variables available in the raw data. For instance, a key variable such as momentum flux (TAU) was excluded due to the high number of missing values. Including TAU within the range where it has valid data significantly reduces the number of samples, which adversely impacts the performance of RNN models that rely on large sample sizes. Better strategies for data filtering and imputation could help mitigate these issues and enhance model performance.

The majority of the data being surface observations poses another challenge. The lack of aerial data complicates the modeling of chaotic atmospheric effects, limiting the model's ability to accurately predict BLH. Incorporating aerial data, even minimally, could add a crucial height dimension to the model, thereby improving its complexity and accuracy. This would enable the model to account for vertical profiles and interactions within the atmosphere more effectively.

Additionally, the neural network model employed in this study is relatively simplistic. More advanced models, such as physics-informed neural networks (PINNs), offer greater potential. However, PINNs require physical formulas to guide the model performance. The two formulas tested, the boundary layer height and the Monin-Obukhov length, did not yield satisfactory results in this context. Exploring other relevant physical models or improving the existing formulations could be a fruitful area for future research.

In summary, improving the model's performance will require addressing the linearity assumption, enhancing data quality and coverage, incorporating aerial data, and exploring more sophisticated modeling techniques. These steps are crucial for developing a robust model capable of accurately predicting boundary layer height in varied and complex atmospheric conditions.

References

[1] Stull, R. B. (1988). *An Introduction to Boundary Layer Meteorology*. Springer Science & Business Media.

[2] Garratt, J. R. (1994). *The Atmospheric Boundary Layer*. Cambridge University Press.

[3] Li, Y., & Tang, J. (2023). Atmospheric Boundary Layer Processes, Characteristics, and Parameterization. *Atmosphere*, 14(4), 691. doi:10.3390/atmos14040691.

[4] Seidel, D. J., Ao, C. O., & Li, K. (2010). Estimating climatological planetary boundary layer heights from radiosonde observations: Comparison of methods and uncertainty analysis. *Journal of Geophysical Research: Atmospheres*, 115(D16). doi:10.1029/2009JD013680.

[5] Holton, J. R. (2004). *An Introduction to Dynamic Meteorology*. Elsevier Academic Press.

[6] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[7] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536. doi:10.1038/323533a0.

[8] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. doi:10.1038/nature14539.

[9] Li, Y., & Tang, J. (2017). Atmospheric Boundary Layer Processes, Characteristics, and Parameterization. *Atmosphere*, 14(4), 691. doi:10.3390/atmos14040691.

[10] Guo, X., Fu, D., & Wang, N. (2016). Diurnal and seasonal variation of planetary boundary layer height over East Asia and its climatic change as seen in the ERA-5 reanalysis data. *Discover Applied Sciences*, Springer. doi:10.1007/s00703-016-0423-4.

[11] Gilmanov, T. G., et al. (2007). Partitioning European grassland net ecosystem CO2 exchange into gross primary production and ecosystem respiration using light response function analysis. *Agriculture, Ecosystems and Environment*, 121(1-2), 93-120.

[12] Flanagan, L. B., Wever, L. A., & Carlson, P. J. (2002). Seasonal and interannual variation in carbon dioxide exchange and carbon balance in a northern temperate grassland. *Global Change Biology*, 8(7), 599-615.

[13] Liu, J., & Pu, Z. (2019). Does soil moisture have an influence on near-surface temperature? *J Geophy Res Atmos*, 124, 6444–6466. doi:10.1029/2018JD029750.

[14] Brownlee, J. (2017). Time Series Prediction with LSTM Recurrent Neural Networks in Python. Machine Learning Mastery.

[15] AmeriFlux Network. About the AmeriFlux Network. Retrieved from https://ameriflux.lbl.gov/about/about-ameriflux/

[16] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, 14(2), 1137-1143.

[17] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2), 281-305.