Response to Referees' Comments
Response to Reviewer #1:
The authors have used 27 years of data collected by variety of instruments at the ARM SGP site to determine PBL height using machine learning. The method uses the PBL height derived by radiosondes, ceilometer, doppler lidar etc. at variety of temporal resolution to derive PBL height as hourly resolution. The results compare well with the evaluation data. The method is then applied to data collected during two field campaigns, CACTI and GoAmazon showcasing reasonable results. The authors argue that this demonstrates the utility of the deep learning models in predicting PBL height (Line 39). The article is well-written, and a lot of work has gone into it. However, I find some flaws with it and encourage the authors to revise it as it will make it better.
Response: We appreciate the reviewer's thoughtful feedback and recognition of the extensive work involved in our study. In response, we have addressed the concerns raised and have integrated more analyses to strengthen the manuscript. All of the comments and concerns raised by the referee have been carefully considered and incorporated into the revised manuscript. Detailed responses to the specific points are provided below.

Major Comments:

1. *It is unclear to me whether the article is about highlighting the uniqueness of deep learning model or it is about implementing the model to derive PBL height for atmospheric science. From the abstract and discussion, it seems that it is an article demonstrating the uniqueness of machine learning model, which is fine but might make it unsuitable for ACP. If it is for doing science from the derived high resolution PBL height values, then maybe some more analysis should be included in the paper.*

Response: We appreciate the reviewer's comments on the focus of our manuscript. The primary aim is to demonstrate the utility of the deep learning model for deriving PBLH and to highlight its implications for deriving reliable values under different scenarios. In response to this feedback, we have expanded our discussion to better elucidate the physical meaning and implications of the feature importance derived from our deep learning model as follows in Section 3.3:

"Figure 2 presents the importance scores to demonstrate each primary feature's relative influence on the model's performance. Prominently, features such as the BLH_{Parcel}, morning potential temperature profiles (θ profile), and surface relative humidity are identified as most important three features, with their substantial impact on the accuracy of PBLH estimation being highlighted. BLH_{Parcel} is defined as the height where the morning potential temperature first exceeds the current surface potential temperature by more than 1.5 K (Holzworth, 1964; Chu et al., 2019). Among these features, BLH_{Parcel} captures the response of the PBL to surface heating, which can drastically affect local convection and thus serves as one of the key parameters in the DNN model. Incorporating this parameter and its association with PBL development better simulates diurnal variations of PBLH in the DNN model. Meanwhile, the morning θ profile represents the vertical stratification of thermodynamics and is essential for understanding stability and mixing processes within the PBL. Thus, θ profile serves as the initial boundary condition for the PBLH estimation with a significant importance score. Surface relative humidity also emerges as a key influencer, affecting the model's performance significantly. Humidity levels influence the condensation and evaporation processes within the PBL, which are important in determining its vertical extent layer and structure. Fair-weather and dry

conditions are typically associated with a more turbulent and higher PBL. Conversely, high surface humidity often contributes to the formation of boundary layer clouds, which introduces complex interactions with PBL thermodynamics."

In addition, we have incorporated a new analysis that examines the performance of our DNN-derived PBL heights under shallow cumulus cloud conditions. This analysis provides further validation of the model's capabilities and offers the physical perspective of the PBL evolution, as well as its association with boundary layer clouds. The details of the analyses can be found in the response to comment #3. Thus, these revisions align our study more closely with the scientific objectives of ACP. These enhancements aim to clarify the scientific contributions of our work and its relevance to the application of deep learning in boundary layer processes.

2. *Table 3: The table lists feature importance of the input variables. Thereby it should highlight the variables that are most important for predicting the PBL height. The values are very small, and it is unclear why they don't add up to one. I highly encourage the authors to normalize the values before presenting them in the table. Please see the paper below for more information. Something like their Figure 7 would be great.*

Response: Thanks for the insightful comments regarding the presentation of feature importance values. We recognize the importance of normalizing these values to enhance their interpretability and to facilitate an intuitive comparison across different model inputs. Following the comment, we have normalized the importance scores so that they now sum to 100%. Now, these relative importance scores are expressed as percentages. Each score quantitatively represents how much the shuffling of a feature increases the MAE, indicating the relative significance of that feature in the model's predictive accuracy and facilitating a straightforward comparison of the influence of each feature within the model.

Following the style of Figure 7 in Gagne et al. (2019), which utilized the permutation feature importance method to rank input variables based on the impact of randomizing their values on prediction error, we have similarly revised Figure 2 (shown in Figure R1). This revision ensures consistency between the Figure 2 and Table 3. It's important to note that Gagne et al. (2019) employed AUC, the area under the ROC curve, as a measure of total prediction skill in a classification context (i.e., positive and negative events), while we use Mean Absolute Error (MAE) as the key metric to evaluate our models. The revised Figure 2 now effectively illustrates the relative importance of each input variable in a more visually accessible format, making it easier to discern which variables are most critical for estimating the PBLH.

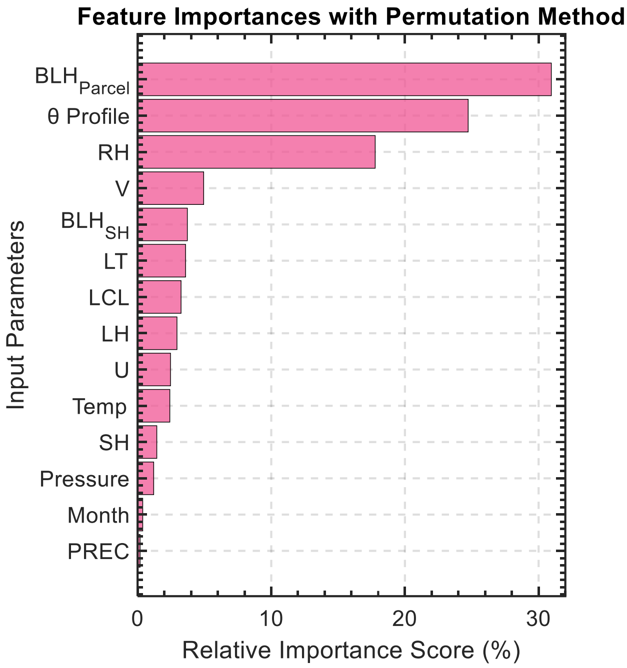

Figure R1 (the revised Figure 2). Feature importance with permutation method in the deep learning model. This table presents the importance scores of each input feature used in the deep learning model to estimate the PBLH. The features include local time (LT), month, relative humidity (RH), surface U and V wind components, pressure at the surface (Pressure), precipitation (PREC), surface temperature (Temp), sensible and latent heat (SH and LH), surface-derived lifting condensation level (LCL), boundary layer height derived from sensible heat and parcel methods (BLH_{Parcel} and BLH_{SH}), and morning profiles of potential temperature (θ Profile). The importance scores are presented as percentages, representing each feature's relative contribution to the model's predictive accuracy, normalized to sum to 100%.

3. *The second author Dr. Zhang has done a lot of work on the SGP site, especially on shallow cumulus clouds and their controls pertaining to land-atmosphere interactions. It will be great if the authors can use either the shallow cumulus case library made by Dr. Zhang, or the shallow cumulus cases simulated by LASSO activity to probe how the new DNN derived PBL heights compare with cloud boundaries. As of now it is hard to tell whether the DNN derived PBL heights are physically consistent.*

Response: We appreciate the valuable suggestion to compare our DNN-derived PBL heights with cloud boundaries. In response, we have incorporated an analysis using the shallow cumulus cases to verify the physical consistency of our DNN outputs with observed cloud boundary conditions. The results from this comparison are now included in Section 4.2 of the revised manuscript. They indicate a good alignment between the DNN-derived PBL heights and the cloud-base height, further validating the accuracy and reliability of the DNN model in capturing PBL evolutions. The detailed discussions are presented as follows.

"The evolution of the PBLH under shallow cumulus conditions offers insights into the interactions between clouds, PBL, and land surface (Zhang and Klein, 2010, 2013). Figure 10 (Figure R2) demonstrates the variations of PBLH measurements from different methods during conditions typical of shallow cumulus clouds. Shallow cumulus clouds were identified following Su et al. (2024).

Specifically, these coupled clouds form post-sunrise; and the sky must not be overcast, characterized by a cloud fraction less than 90%. This selection criterion ensures that the observed cloud formations are primarily driven by surface heating and local convection. The DNN model closely matches the SONDE-derived PBLH and the cloud-based height from ARSCL. This alignment underscores the physical validity of the DNN approach, confirming its capability to replicate traditional measurement techniques accurately. Meanwhile, Doppler lidar-derived PBLH retrievals also show high consistency with SONDE measurements, whereas ceilometer-derived PBLH generally underestimates values under shallow cumulus conditions.

Figure 10 (Figure R2) also demonstrates the general relationship between the development of shallow cumulus clouds and the PBL, which are driven by local convection and turbulence. The formation of these cumulus clouds is linked to rising thermals and an increase in surface heat fluxes, essential for driving vertical mixing within the sub-cloud layer. This relationship is evidenced by the increased occurrence of cumulus clouds along with an increase in DNN-derived PBLH from morning to late afternoon. Specifically, during periods with a high frequency of shallow cumulus, the DNN-derived PBLH often surpasses the cloud base height. This indicates that rising air parcels extend beyond the condensation level, facilitating the formation and development of coupled cumulus clouds. In this context, these analyses confirm the physical consistency of DNN-derived PBLH with traditional measurement techniques and highlight its physically reasonable variations during cloudy conditions.”

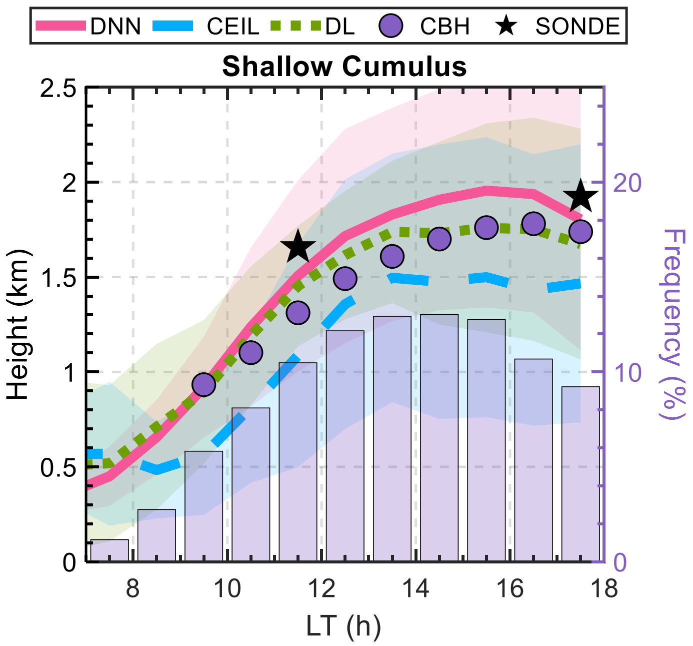

Figure R2 (the revised Figure 10). Daytime evolution of planetary boundary layer height (PBLH) derived from various methods under the shallow cumulus condition. PBLH values estimated by the deep neural network (DNN) are shown in red, ceilometer (CEIL) estimates in blue, Doppler lidar (DL) in green. Observed radiosonde (SONDE) data are represented by black stars. Purple bars show the relative frequency of shallow cumulus occurrences throughout the day, while purple dots mark the corresponding cloud-base heights (CBH). Shaded areas around each line reflect the standard deviations for each method.

4. *Line 240: you have used the lidar derived PBL height when the radiosonde data are not available, with a caveat that they agree within 200m. Can you please tell us how many of them did not agree within the 200m threshold and what was done for those periods? Thanks.*

Response: Thanks for pointing out the need for additional details regarding the agreement between lidar-derived and SONDE-derived PBLH. Specifically, out of the total comparisons during the study period, 40.2% of the lidar measurements do not agree within the 0.2 km threshold with the SONDE results. The cases with relatively larger inconsistencies stem from various factors, including instrumental errors, rainy conditions, stable PBL conditions, differing definitions, and lidar signal attenuation, as discussed in previous studies (Su et al., 2020; Kotthaus et al., 2023). These cases were excluded from the DNN model training to maintain the quality of the process. We have incorporated these discussions into Section 3.2 of the revised manuscript to clarify our methodology.

Minor Comments:

Line 117: add height above mean sea level.

Response: We added the elevation of SGP site for the clarity.

Line 119-120: Add references to the field campaigns.

Response: References to the CACTI and GoAmazon field campaigns have been included (Varble et al. 2021; Martin et al. 2016).

References

Martin, S. T., Artaxo, P., Machado, L. A. T., Manzi, A. O., Souza, R. A. F. D., Schumacher, C., ... & Wendisch, M. (2016). Introduction: observations and modeling of the Green Ocean Amazon (GoAmazon2014/5). Atmospheric Chemistry and Physics, 16(8), 4785-4797.

Varble, A. C., Nesbitt, S. W., Salio, P., Hardin, J. C., Bharadwaj, N., Borque, P., ... & Zipser, E. J. (2021). Utilizing a storm-generating hotspot to study convective cloud transitions: The CACTI experiment. Bulletin of the American Meteorological Society, 102(8), E1597-E1620.

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. and Mona, L., 2023. Atmospheric boundary layer height from ground-based remote sensing: a review of capabilities and limitations. Atmospheric Measurement Techniques, 16(2), pp.433-479.

Zhang, Y., & Klein, S. A. (2010). Mechanisms affecting the transition from shallow to deep convection over land: Inferences from observations of the diurnal cycle collected at the ARM Southern Great Plains site. Journal of the Atmospheric Sciences, 67(9), 2943–2959. https://doi.org/10.1175/2010jas3366.1

Su, T., Li, Z., Zhang, Y., Zheng, Y., & Zhang, H. (2024). Observation and Reanalysis Derived Relationships Between Cloud and Land Surface Fluxes Across Cumulus and Stratiform Coupling Over the Southern Great Plains. Geophysical Research Letters, 51(8).

Gagne II, D. J., Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. Monthly Weather Review, 147(8), 2827-2845.