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.

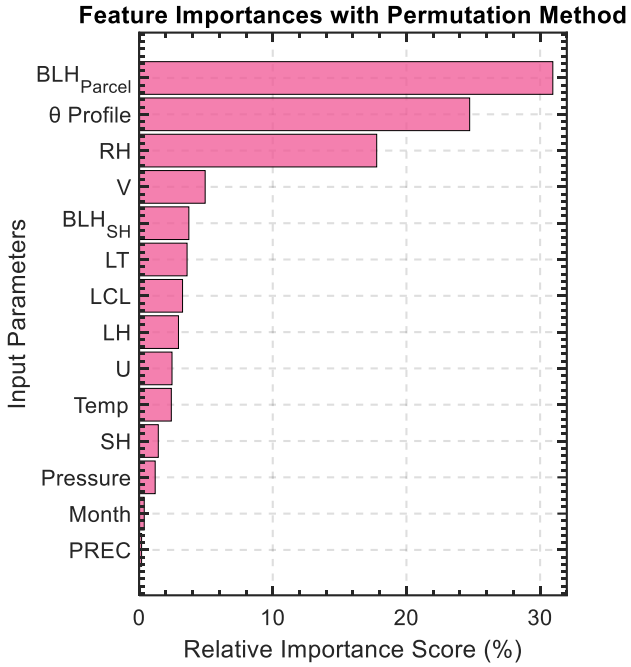

Figure R1 (the revised Figure 2). Feature importance with permutation method in the deep learning model. This table presents the importance scores of each input feature used in the deep learning model to estimate the PBLH. The features include local time (LT), month, relative humidity (RH), surface U 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."

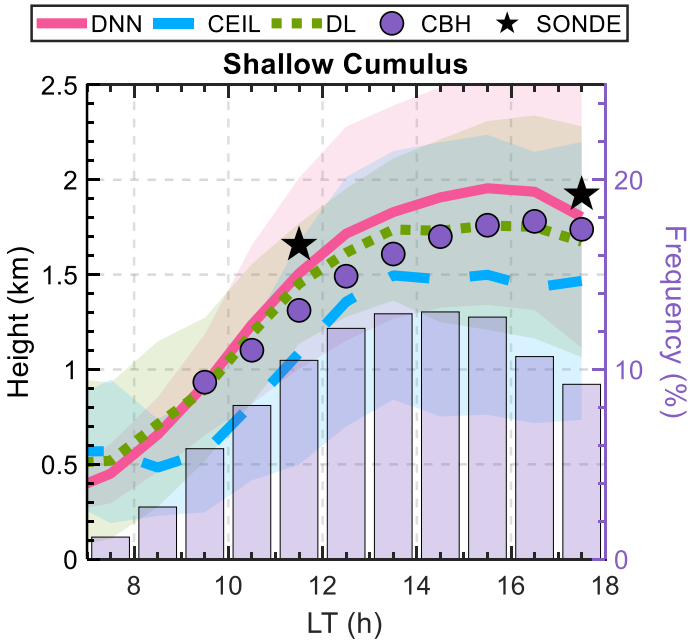

Figure R2 (the revised Figure 10). Daytime evolution of planetary boundary layer height (PBLH) derived from various methods under the shallow cumulus condition. PBLH values estimated by the deep neural network (DNN) are shown in red, ceilometer (CEIL) estimates in blue, Doppler lidar (DL) in green. Observed 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.

Response to Reviewer #2:

While traditional machine learning methodologies (e.g., Random Forest) have been widely used to estimate PBLH, most studies heavily rely on specific remote sensing instruments or focuses on limited time-period or specific region of interest. More importantly, lack of enough physical explanation is another concern. To address this issue, this manuscript introduces a multi-structure deep neural network (DNN) model that is used to generate yield a robust 27-year PBLH dataset over the Southern Great Plains from 1994 to 2020. Through leveraging a variety of meteorological data, independent of remote sensing instruments, this model yielded an PBLH dataset over the SGP with robust accuracy, consistently yielding lower bias values across various conditions and datasets. Besides, the generalizability of this model to different geographic regions and climate zones are explored, exhibiting high potential and less uncertainties in terms of seasonal, diurnal variability. Overall, this manuscript is well organized with clear enough logic, I would like to offer the following suggestions for further improvement:

Response: We appreciate the reviewer's positive and comprehensive comments on our work. Following these insights, we have refined our manuscript to enhance its clarity. We have carefully considered and addressed all comments and concerns raised by the reviewer in this revision. Our detailed responses to each point are provided below.

Major Comments:

Introduction: Except for the lidar systems, the authors seem to ignore the radar wind profiler, which provides the direct measurements of turbulence in the atmosphere and thus affords the retrievals of PBLH. A variety of algorithms or methods in the literature have been proposed to accomplish this task. Therefore, the authors can argue the current literature review in this regard.

Response: Thanks for the helpful suggestion. We acknowledge the importance of radar wind profilers in measuring atmospheric turbulence and their utility in PBLH retrieval and acknowledge the relevant studies in the introduction as follows:

"In addition, wind profilers can estimate the PBLH using algorithms that analyze the signal-to-noise ratio from wind profiler data (Molod et al. 2015; Solanki et al. 2022; Liu et al. 2019; Salmun et al. 2023; Bianco and Wilczak 2002; Bianco et al. 2008; Tao et al. 2021)."

References:

Solanki, R., Guo, J., Lv, Y., Zhang, J., Wu, J., Tong, B., & Li, J. (2022). Elucidating the atmospheric boundary layer turbulence by combining UHF radar wind profiler and radiosonde measurements over urban area of Beijing. Urban Climate, 43, 101151.

Liu, B., Ma, Y., Guo, J., Gong, W., Zhang, Y., Mao, F., ... & Shi, Y. (2019). Boundary layer heights as derived from ground-based Radar wind profiler in Beijing. IEEE Transactions on Geoscience and Remote Sensing, 57(10), 8095-8104.

Molod, A., Salmun, H., and Dempsey, M., 2015: Estimating Planetary Boundary Layer Heights from NOAA Profiler Network Wind Profiler Data, J. Atmos. Ocean. Tech., 32, 1545–1561, https://doi.org/10.1175/JTECH-D-14-00155.1.

Salmun, H., Josephs, H., & Molod, A. (2023). GRWP-PBLH: Global Radar Wind Profiler Planetary Boundary Layer Height Data. Bulletin of the American Meteorological Society, 104(5), E1044-E1057.

Bianco, L., Wilczak, J. M., & White, A. B. (2008). Convective boundary layer depth estimation from wind profilers: Statistical comparison between an automated algorithm

and expert estimations. Journal of Atmospheric and Oceanic Technology, 25(8), 1397-1413.
Bianco, L., and J. M. Wilczak, 2002: Convective boundary layer depth: Improved measurements by Doppler radar wind profiler using fuzzy logic methods. J. Atmos. Oceanic Technol., 19, 1745–1758, https://doi.org/10.1175/1520-0426(2002)019,1745: CBLDIM.2.0.CO;2.
Tao, C., Y. Zhang, Q. Tang, H. Ma, V. P. Ghate, S. Tang, S. Xie, and J. A. Santanello, 2021: Land–Atmosphere Coupling at the U.S. Southern Great Plains: A Comparison on Local Convective Regimes between ARM Observations, Reanalysis, and Climate Model Simulations. J. Hydrometeor., 22, 463–481, https://doi.org/10.1175/JHM-D-20-0078.1.

Line 89-102: The reason for the selection of multi-structure deep neural network (DNN) in the retrieval of PBLH lacks necessary literature support. Are there similar models constructed based on DNN? If any, how is the performance compared with other models or methods? This should be clarified and some necessary references are required to be cited here.

Response: Response: We appreciate the comment regarding the need for a clearer explanation for the choice of the deep learning model. In response, we have added a detailed discussion on the introduction as follows:

"We aim to leverage and integrate the comprehensive field observations (i.e., radiosonde and remote sensing techniques) to develop a deep learning model for direct PBLH estimation from conventional meteorological data. This strategy circumvents the limitations of relying on particular remote sensing technologies. Furthermore, our model employs an advanced deep neural network (DNN) approach (Sze et al. 2017; Schmidhuber, 2015; Nielsen, 2015; Pang et al. 2020), diverging from traditional ML methods like random forest. This deep learning model utilizes ensemble techniques, constructing arrays of various structures and using their average for the final estimation. This approach method provides particular advantages in the context of complex and nonlinear processes (Ganaie et al. 2022; Mohammed and Kora. 2023). Ensemble DNN with multi-structure designs shows very strong flexibility and robustness, so it relatively performs better and has high stability across a wide range of conditions (Xue et al. 2020; Dong et al. 2020). This facilitates the adaptability of DNN as a tool for PBLH estimation, which can be utilized under different scenarios and locations."

References:

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University-Computer and Information Sciences, 35(2), 757-774.
Xue, W., Dai, X., & Liu, L. (2020). Remote sensing scene classification based on multi-structure deep features fusion. IEEE Access, 8, 28746-28755.
Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. Frontiers of Computer Science, 14, 241-258.
Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115, 105151.
Sze, V., Chen, Y.H., Yang, T.J. and Emer, J.S., 2017. Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), pp.2295-2329.
Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural networks, 61, pp.85-117.
Nielsen, M.A., 2015. Neural networks and deep learning (Vol. 25, pp. 15-24). San Francisco, CA, USA: Determination press.

Pang, B., Nijkamp, E. and Wu, Y.N., 2020. Deep learning with tensorflow: A review. Journal of Educational and Behavioral Statistics, 45(2), pp.227-248.

Specific comments:

1. *Line 50: "it" is redundant and can be removed.*

Response: Thanks for catching this typo. We have revised it.

2. *Line 54: "climate models" -> "climate projections"*

Response: Revised as suggested.

3. *Line 83: "PBL heights using thermodynamic profiles or backscatter profiles from Lidar or Atmospheric Emitted Radiance Interferometer (AERI)" -> "PBLH using thermodynamic profiles Atmospheric Emitted Radiance Interferometer (AERI) or using backscatter profiles from Lidar".*

Response: We have rephrased the sentence as suggested.

4. *Line 87: "Moreover," -> "For example, "*

Response: The suggestion has been incorporated.

5. *Line 89: "marked progress" -> "made great progress"*

Response: Revised as suggested.

6. *Line 136: Some words are missing between "latent heat fluxes" and "the surface instruments"*

Response: We revised it as "latent heat fluxes from the surface instruments".

7. *Line 365-367: is there any supporting material for the threshold used to define low cloud (maximum cloud fraction between 0-4 km exceeding 1%)?*

Response: The ECWMF also use 1% as the threshold to identify the cloud base height. Specifically, cloud base is calculated by searching from the second lowest model level upwards, to the height of the level where cloud fraction becomes greater than 1% and condensate content greater than $1.E^{-6}$ kg kg^{-1} (Hersbach et al. 2023). We include this reference in the manuscript.

References:

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2023): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: https://doi.org/10.24381/cds.adbb2d47