Reply to Reviewer #2's comments

This is a well-written paper describing the use of remote sensing measurements to derive estimates of PBL height. The authors describe the use of various measurements and algorithms to derive the PBL height and how the remote sensing measurements can be used in machine learning algorithms to improve retrievals of PBL height. Recommend publications after the authors address the minor comments below (pay particular attention to item 21).

Answer: We appreciate the reviewer's valuable suggestions and comments. We have carefully revised the manuscript in accordance with the feedback.

1. *Lines 51-55. The references provided in lines 51-55 refer to ground-based remote sensing measurements. The authors should also include references to airborne remote sensing measurements.*

Answer: We appreciate the reviewer's constructive suggestion and have carefully considered it. However, since the leading sentence highlights the high temporal resolution of PBLHT estimates from remote sensing, an area where airborne remote sensing measurements are less strong, we decided not to include airborne remote sensing studies in that section. Instead, we have added references to airborne remote sensing observations at the end of the paragraph when discussing remote sensing instruments across various platforms in lines 60-62.

2. *Line 60. This should also mention aircraft as well as ground stations and satellites.*

Answer: We added references to aircraft remote sensing observations as suggested in lines 60-62.

3. *Line 66. More reliable than what? Radiosonde profiles also generally have higher vertical resolution than many remote sensing techniques.*

Answer: We agree with the reviewer's comment and have revised the text for clarity. We added "and high vertical resolution" in line 65 and "than that from remote sensing observations" in line 66 to make the point clearer.

4. *Line 81. This statement says that "…combining PBLHT estimates from multiple remote sensing techniques can lead to more accurate estimates under various conditions." If multiple remote sensing techniques each provide a different estimate of the PBLHT, how does one combine these, especially if each technique may have some advantages and disadvantages? Are these combined using different weights for the different techniques?*

Answer: We replaced the word "combining" with "integrating" to more accurately describe the process. As demonstrated in the manuscript, these different techniques were integrated using ML methods.

5. *Line 101. What is the true PBL height if (as mentioned in line 40) there are several well-accepted definitions of the PBL? The authors should indicate the PBL height varies depending on the method used to define the PBL so there is no "true" PBL height.*

Answer: We agree with the reviewer that there is no single "true" PBL height. Therefore, we replaced "the true" with "improving" to better reflect this concept.

6. *Line 115. Should read "…most extensive ground-based climate research facility".*

Answer: Thanks. We have updated the sentence as suggested.

7. *Line 132. The Heffter technique requires a critical potential temperature lapse rate as well as a potential temperature difference between the top and bottom of a layer. What was the critical lapse rate used in this study and was this the 0.001 K/m recommended by DelleMonache et al. (2004) for the SGP site instead of the more typical 0.005 K/m?*

Answer: In this study, we directly used PBLHT-Sonde data from the DOE ARM PBLHT-SONDE Value-Added Product (VAP). As stated in the PBLHT-SONDE VAP technique report (https://doi.org/10.2172/1808688), a critical lapse rate of 0.005 K/m was applied. Rather than including these details in the text, we refer readers to the technical report for further information in line 137.

8. *Figure 1d. The green triangle seems to correspond to a critical Richardson number of 0.15 rather than 0.25.*

Answer: We appreciate the reviewer for identifying the issue and have updated the figure accordingly.

9. *Figure 2 shows PBLHT estimates compared with their median. The shaded regions correspond to the kernel distribution estimates. It would be helpful if the paper provided a short description of the kernel distribution estimate (KDE), how this is computed, and how to interpret the values.*

Answer: As suggested, we added several sentences between lines 186 and 189 to explain KDE and how to interpret its values.

10. *Line 215. There is likely also some attenuation by water vapor at 910 nm, which may also cause the ceilometer to have a lower S/N than the MPL.*

Answer: We appreciate the reviewer's suggestion. We add 'and stronger water vapor absorption' in line 233.

11. *Line 224. Since there are often Cumulus clouds at/near the top of the daytime BL, why would a higher quality index be assigned when there were no clouds detected near the boundary layer? Is this because clouds interfere with the remote sensing measurements? See also item 24.*

Answer: That's correct. Münkel and Roininen (2010) state that the enhanced gradient method incorporates a cloud and precipitation filter during averaging to prevent false layer identification. To ensure accuracy, high backscatter from clouds and precipitation is excluded from the averaging process.

Reference: Münkel, C., Roininen, R.: Automatic Monitoring of Boundary Layer Structures with Ceilometer. vol. 184 Vaisala News., 2010.

12. *Line 224. How do the various techniques (MPL, Ceil, radiosonde, thermos, DL, etc.) deal with the presence of clouds?*

Answer: The PBLHT-SONDE VAP does not explicitly account for the influence of clouds (Sivaraman et al., 2013). Similarly, PBLHT estimates derived from remote sensing also do not specifically consider the presence of clouds. To minimize the impact of mid- and high-level clouds, PBLHT detection algorithms are limited to within 4 km of the surface. Since low-level cloud bases generally occur at or near the PBLHT, detection algorithms often identify the cloud base as the PBLHT. This does not introduce significant errors when compared with PBLHT-SONDE (Sawyer and Li, 2013; Zhang et al., 2022). We added this paragraph in the text between line 354 and 358.

References:

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

Sivaraman, C., McFarlane, S., Chapman, E., Jensen, M., Toto, T., Liu, S., and Fischer, M.: Planetary boundary layer (PBL) height value added product (VAP): Radiosonde retrievals, U.S. Department of Energy Rep. DOE/SC-ARM-TR-132, 36 pp., https://www.arm.gov/publications/tech_reports/doe-sc-arm-tr-132.pdf (last access: 9 August 2022), 2013.

Zhang, D., Comstock, J., and Morris, V.: Comparison of planetary boundary layer height from ceilometer with ARM radiosonde data, Atmos. Meas. Tech., 15, 4735–4749, https://doi.org/10.5194/amt-15-4735-2022, 2022.

13. *Line 244. Should read "σ_w^2 remains low at night and in the early morning."*

Answer: We updated the sentence as suggested.

14. *Line 264. What is (are) the uncertainties in the RL temperature retrievals? How do these uncertainties compare with the temperature uncertainty in the radiosonde temperature measurement? How do the RL temperature uncertainties vary with altitude? With daytime vs. nighttime operations?*

Answer: As noted in line 284, “The uncertainty in the RL’s temperature retrievals is calculated using standard error analysis, and by default, temperature retrievals with relative uncertainties greater than 0.05 are excluded.” Figure r2 illustrates the probability distribution function (PDF) of RL temperature uncertainties, along with their variations with altitude and between daytime and nighttime operations. In the atmospheric boundary layer, RL temperature retrieval uncertainties are generally within 2 K, which is larger than the typical SONDE temperature uncertainty of 0.3 K (Holdridge, 2020). These uncertainties tend to increase with altitude and show slight differences between day and night retrievals. However, since a detailed discussion of RL temperature uncertainty characteristics is beyond the scope of this manuscript, we have chosen not to include Figure r2 in the main text.

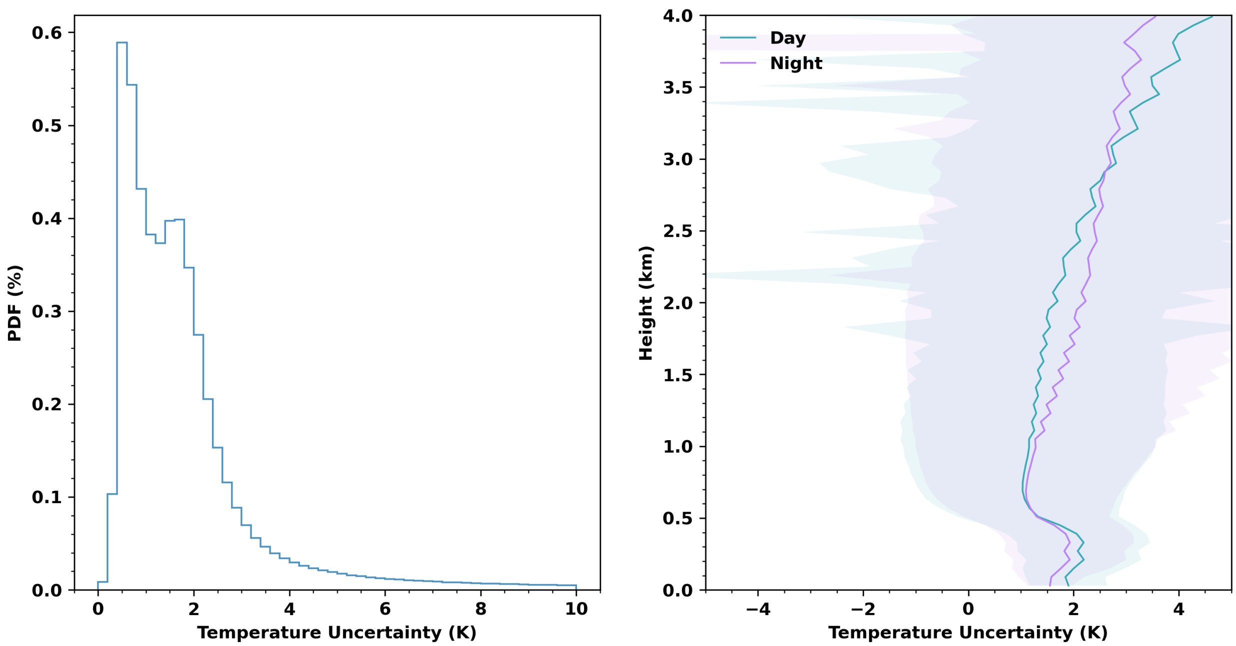

Figure r2. Raman Lidar (RL) temperature uncertainty: (a) probability distribution function (PDF) of RL temperature uncertainty; (b) variation of RL temperature uncertainty with altitude during daytime and nighttime operations.

Reference:

Holdridge, D. "Balloon-Borne Sounding System (SONDE) Instrument Handbook." , Nov. 2020. https://doi.org/10.2172/1020712

15. Line 284. Does PBLHT-THERMO use only the Heffter method? If so, why not also compute PBLHT-THERMO1 (using Heffter) and PBLHT-THERMO2 (using Liuliang) and use the mean of these? This could increase the agreement between PBLHT-THERMO (mean) and PBLHT-SONDE.

Answer: It is correct that PBLHT-THERMO currently uses only the Heffter method. We appreciate the reviewer’s suggestion to incorporate the Liuliang method. The Heffter method

identifies the PBLHT primarily by detecting potential temperature inversion layers, making it more robust to random temperature uncertainties. In contrast, the Liuliang method relies on calculating the temperature gradient at each level, which requires higher-quality temperature data. Additionally, the Liuliang method depends on low-level horizontal wind information to estimate PBLHT under stable boundary layer conditions. However, such wind data are not available from AERI or RL measurements. Although the ARM Doppler Lidar (DL) does provide horizontal wind retrievals, these measurements are affected by the lidar overlap effect at low altitudes and are limited by coarse spatial and temporal resolution.

16. *Line 372. In the case of missing values, why not also just leave out those cases where one or more pieces of data were missing, rather than trying to fill in these missing data values, especially for those cases that were used to train the ML model?*

Answer: We thank the reviewer for the suggestion. Our goal is to use data with all required inputs available for training the ML models while also maximizing the amount of usable data. To achieve this, we applied a data-filling method, but limited the filling to gaps of no more than one hour. The potential impact of missing data is further explored through the feature importance analysis presented in Section 3.2.2.

17. *Line 400. Hopefully the results shown in Figure 7 correspond to data that were NOT included in the training set. Is this true?*

Answer: That is correct. The evaluations at SGP are conducted solely using the testing dataset. We have added the phrase 'using the testing dataset' in line 436.

18. *Line 432. Rather than using local time, could you have used instead solar zenith angle to avoid issues with other locations?*

Answer: We thank the reviewer for the suggestion. Both local time and solar zenith angle are indirectly related to atmospheric radiative fluxes, which influence the depth of the PBL. However, since our input variables already include both shortwave and longwave radiation, adding local time or solar zenith angle is not essential. That said, one motivation for including local time is its practicality—it does not require advanced instrumentation and can serve as a convenient input for researchers without access to observational data.

19. *Figure 8. Based on these results, is it safe to say that satisfactory results can be obtained solely using the PBLHT-THERMO, PBLHT-DL, and PBLHT-MPL results without the use of the other measurements?*

Answer: We would like to point out that although PBLHT-THERMO, PBLHT-DL, and PBLHT_MPL are the top three most important features, SRAD and the PBL regime also contribute to the ML model predictions. Accordingly, we have added the phrase 'followed by SRAD and PBL regime' in line 489.

20. *Line 446. Isn't it likely that PBLHY-THERMO emerges as the most significant feature because two of the three items used to compute PBLHT-SONDE use potential temperature profiles as the metric rather than aerosol or water vapor profiles?*

Answer: We agree with the reviewer and have added the phrase 'uses potential temperature profiles to derive PBLHT' to the text.

21. *Line 454 and Table 1. This sentence implies that the higher S/N of the MPL led to a better performance when measuring aerosol gradients, which in turn, led to a better performance in determining the PBL height. If this is the case, then why were Raman lidar measurements of aerosol profiles (e.g. aerosol attenuated backscatter, aerosol unattenuated backscatter, extinction, and/or depolarization) not used in any of these analyses? The Raman lidar S/N should be much greater than the MPL S/N and hence be more likely to see weaker aerosol gradients than may correspond to PBL HT, especially at night. (If for some reason this was not true the paper should explain what was wrong with the RL.) These measurements have inherently high vertical and temporal resolution that could be very useful for such analyses. Furthermore, the RL provides (or is at least supposed to provide) high temporal and vertical resolution water vapor profiles during both daytime and nighttime operations. It would be interesting to see if the gradients in water vapor would be more or less useful than aerosol gradients to determine PBL HT, especially at night. This reviewer was surprised (and disappointed) that such RL measurements were not included in these analyses. This omission is even more surprising when considering that RL measurements of temperature were included. Consequently, the paper should address why such RL measurements of aerosols and water vapor were not included when RL measurements of temperature were included.*

Answer: We appreciate the reviewer's valuable suggestion. The PBLHT-THERMO approach follows the methodology outlined by Ferrare et al. (2012), and we have now added a reference to Ferrare's presentation. As noted in their work, "ML heights from Raman lidar water vapor have large high bias as compared to BL heights from radiosonde potential temperature." Based on this finding, we chose not to include PBLHT estimates derived from Raman lidar (RL) water vapor profiles.

Regarding RL aerosol measurements, the reviewer is correct that the RL system, operating at 355 nm, is more sensitive to smaller aerosol particles. However, it is important to note that molecular scattering also contributes significantly to the total backscatter at this wavelength. Figure r3 presents comparisons of the RL scattering ratio, RL particulate backscatter, and MPL backscatter. These comparisons show that MPL backscatter provides a clearer contrast between the planetary boundary layer and the free atmosphere. Therefore, we use MPL backscatter for estimating PBLHT.

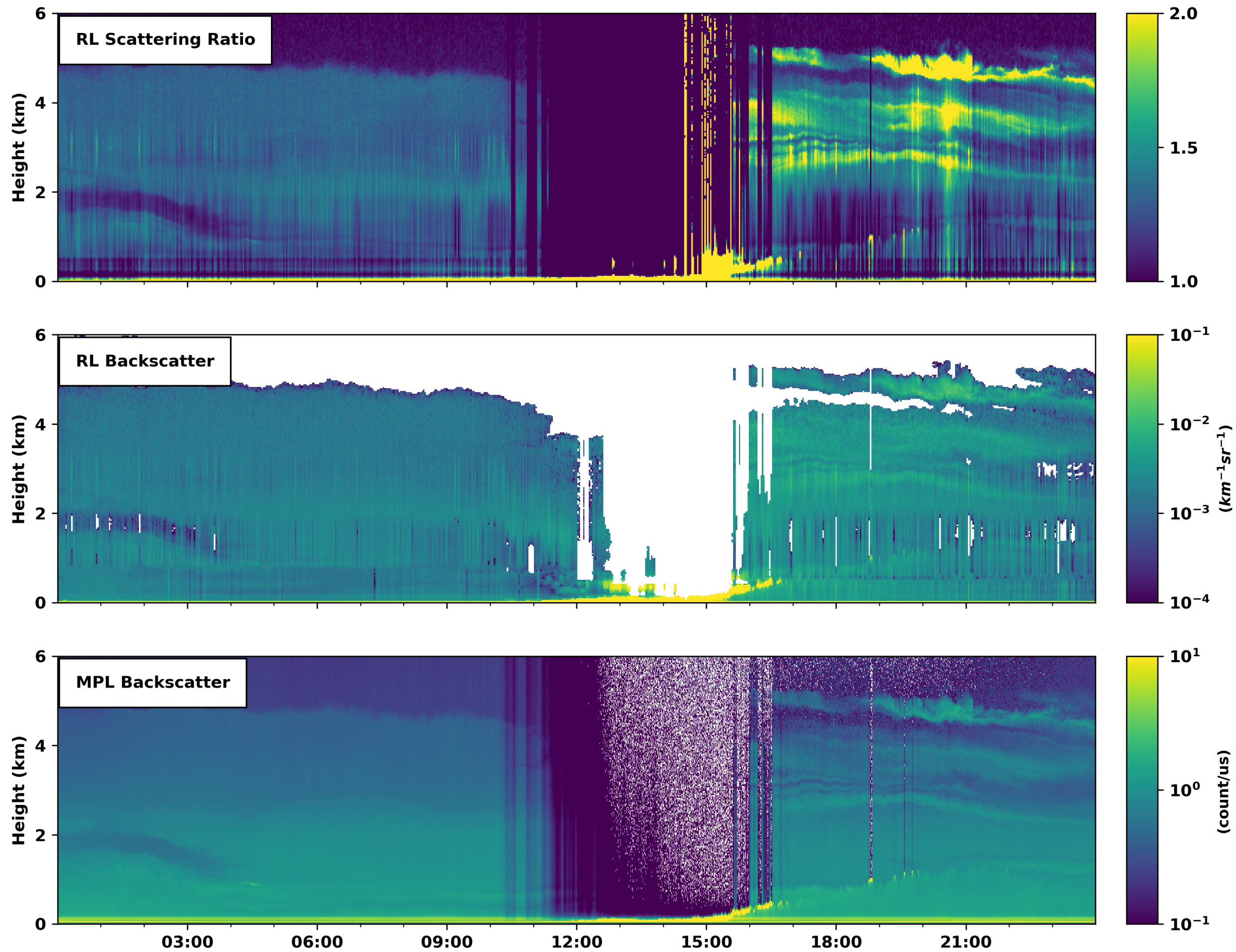

Figure r3. Time-height plots of Raman Lidar (RL) and Micropulse Lidar (MPL) measurements on October 10, 2020 at the ARM SGP site. (a) RL scattering ratio; (b) RL particulate backscatter; (c) MPL attenuated backscatter.

Reference:

Ferrare, R., Clayton, M., Turner, D., Newsom, R., Scarino, A.J., Burton, S., Hostetler, C., Hair, J., Obland, M., and Rogers, R.: Raman Lidar Retrievals of Mixed Layer Heights. DOE ASR Science Team Meeting, Virginia, USA, March 12-16, 2012.

22. Line 504. Tis line seems to indicate that the RL temperature retrievals have large uncertainties (at least for the altitudes associated with the strong convective periods). The paper should indicate what these uncertainties actually are and perhaps show a comparison of potential temperatures from THERMO as compared to those derived from the radiosondes. How small should these temperature uncertainties be to derive accurate PBLHT from the THERMO method?

Answer: We excluded PBLHT-THERMO from the PBLHT-Lidar ensemble because it clearly overestimates PBLHT under convective boundary layer conditions, as shown in Figure 4d. This overestimation is likely due to the use of the Heffter method in PBLHT-THERMO, which tends to produce higher values compared to PBLHT-Median, as illustrated in Figure 2a. Furthermore, larger uncertainties in temperature retrievals and weaker potential temperature gradients at higher altitudes—as shown in Figure r2—may further contribute to this bias. We have added several sentences between lines 337 and 340 to explain these causes.

It is challenging to determine a specific threshold for temperature uncertainty below which accurate PBLHT estimates can be ensured. Since potential temperature profiles are used to identify key features such as inversion layer bases and tops, any uncertainty in temperature measurements can directly or indirectly affect the identification of these features. Quantifying these impacts remains difficult.

23. *Line 519. This line suggests that the performance of the ML model may decrease for periods different from those included in the training sets. Perhaps more useful would be to compare the values of the various variables with those used in the actual cases. For example, comparisons of the BL heights, potential temperatures, potential temperature gradients, aerosol graidients, etc. for the data included in the training sets vs. the data that are in the data to be analyzed. This would one to see where the ML model must interpolate vs. extrapolate results. One would expect poorer performance for extrapolation vs. interpolation.*

Answer: The reviewer is correct that our results from Figure 11 and the related discussion in Section 3.2.3 indicate that "the performance of the ML model may decrease for periods different from those included in the training sets." We appreciate this valuable observation. Due to the lack of validation data outside of the radiosonde launch times, it is challenging to accurately assess the ML model's performance based solely on a comparison of input variables. Therefore, we did not include a discussion of this particular experiment.

24. *Lines 532 and 542. If PBLHT-THERMO is derived solely from the AERI data using the TROPoe algorithm, how well can this be expected to work with ubiquitous stratocumulus clouds in coastal California, and is this a reason why the PBLHT-THERMO overestimated PBLHT there?*

Answer: We agree with the reviewer that the overestimation of PBLHT from PBLHT-THERMO is likely due to large temperature retrieval uncertainties from TROPoe under opaque stratocumulus clouds, as discussed in Turner and Löhnert (2021). To reflect this, we have added a sentence in line 584 to illustrate this possible cause.

Reference:

Turner, D. D. and Löhnert, U.: Ground-based temperature and humidity profiling: combining active and passive remote sensors, Atmos. Meas. Tech., 14, 3033–3048, https://doi.org/10.5194/amt-14-3033-2021, 2021.

25. Line 574. I don't recall the use of water vapor profiles in PBLHT-THERMO.

Answer: We removed the phrase "and water vapor" from the text.

26. OK, so given these results, is there a plan to use this PBLHT-BE-ML algorithm operationally at the SGP site, or at other sites like ENA and/or BNF? If not, why not?

Answer: We appreciate the reviewer's suggestion. The PBLHT-BE-ML method is intended to be implemented as a Value-Added Product (VAP) to provide improved PBLHT estimates at ARM sites. Currently, the ML models have been trained using data from the SGP site, and we anticipate that the approach can be directly extended to other land-based sites such as BNF. While we have demonstrated that PBLHT-BE-ML significantly improves PBLHT estimates at ENA, we aim to further enhance the model by incorporating training data from a broader range of surface types, including ocean, ice, and snow-covered regions. This expansion will help improve performance over these more complex surface environments. We added a sentence about this plan in lines 664-665.