Reply to Reviewer #1's comments

General comment:
The new version of the manuscript looks better; however, I still see that the quality of the study needs to be improved. I have a serious problem with understanding your definition and use of the Richardson number. If your explanation is true (which I believe you confirmed in the responses), then I think your methodology is used incorrectly, which may invalid your investigation.

I strongly recommend that ALL the authors read the paper very carefully again. Please be very critical about every sentence you wrote, and ask yourselves if each of those sentence makes sense. It would also be good to check the final version of the text for errors and confusing formulations. Please see my specific comments below showing how difficult it was for me to understand some parts of the text.

Response: We thank the reviewer for these constructive suggestions and comments. We carefully revised the manuscript according to the reviewer's comments.

Specific comments:
Title: "Comparisons" - I would suggest to change it to a more standard 'Comparison' as your study is a comparison of different approaches to estimating PBLH.

Response: We changed the title using 'Comparison' as suggested.

Abstract:
There is no need to use 'stable atmospheric boundary layer conditions' or 'unstable atmospheric boundary layer conditions' so many times in one paragraph. Once you introduced it, a shorter version ('stable conditions' or 'unstable conditions') can be used.

Response: We thank the reviewer for the helpful suggestion on improving the readability of the manuscript. We used a shorter version after 'stable atmospheric boundary layer conditions' or 'unstable atmospheric boundary layer conditions' is introduced in the abstract.

L29: Still confusing because PBLHT as a parameter doesn't have any impact on air quality or climate. No parameter has any impact on those elements of the Earth system. I am not sure what the authors are trying to say here. Please consider reformulating it and think about the reader that may not understand everything.

Response: We thank the reviewer for pointing out the confusing sentence. We reformulated the sentence in the manuscript as following:

The structure and the depth of the planetary boundary layer play a critical role in near surface air quality, land-atmosphere interactions, and a wide range of atmospheric processes such as cloud formation and evolution, aerosol mixing and transport, and aerosol-cloud interactions (Seinfeld et al., 2006; Konor et al., 2009; Lemone et al., 2018). The height of the planetary boundary layer (PBLHT) is a key parameter that characterizes the structure of the lowest few kilometers of the atmosphere.

Tropics -> tropics

Response: We corrected it as suggested.

L54: 'Under the SBL condition, turbulence tends to be suppressed by the statically stable air above it and occurs only sporadically' - What do you mean by 'above it'? Above the boundary layer? But then it doesn't make much sense. This sentence is confusing and needs some clarification. Do you want to involve everything that happens above the PBL?

Response: We thank the reviewer for pointing the confusing sentence. We rewrote this sentence in the manuscript in lines 60-62 as following:

"The SBL is commonly formed during nighttime by surface radiative cooling or when warm air is advected over a cool surface. Under the SBL condition, virtual potential temperature increases with altitude in the boundary layer. Turbulence tends to be suppressed and occurs sporadically."

L62: *'The PBLHT in atmospheric models is usually calculated by using either diagnostic equations that take surface fluxes and the initial temperature profile as inputs ' – This is a very confusing sentence. It wrongly suggests that most of the models (I'm sure you know that atmospheric models include large-eddy simulation, regional, and global circulation models) calculate PBLHT this way, but this is not true. Why only initial conditions? How can PBLHT vary if you use initial conditions? PBLHT may be calculated in many ways, not necessarily by using surface fluxes. In one of the next sentences, you suggest that those methods are only applicable to observational data, which is questionable (once you have some data, you can apply any method you want). Please reformulate this paragraph. You may need to do some research on those methods before you update it.*

Response: We thank the reviewer for pointing out the confusing sentence. We rewrote the sentence in the manuscript in lines 70 -73 as following:

"The PBLHT in numerical weather prediction and climate models is usually calculated by using the Richardson number profile to find the first level where the Richardson number exceeds a critical value and in large-eddy simulation models by using turbulence kinetic energy or eddy diffusivity thresholds (Seibert et al., 2000; Noh et al., 2002; Seidel et al., 2012)."

As for the confusing sentences about PBLHT estimation methods for observations, we did not mean that these methods can't be applied to model simulations. To avoid confusion, we deleted the sentence 'On the other hand' in the manuscript.

L73: 'Observing atmospheric boundary layer transitions with high temporal-spatial resolutions is required to investigate atmospheric thermodynamic processes (Fritz et al., 2021).' – Another confusing sentence. What thermodynamic processes do you have in mind? That sentence suggests all of them. I think the message from the Fritz et al paper is different as they show that there are many important details that cannot be measured when using instruments with coarser resolution than their DTS, in addition to the fact that their method of determining PBLH based on the temperature gradient is quite reliable.

Response: We thank the reviewer for pointing out the confusing sentence. Since this paragraph focuses on PBLHT, we rewrote the sentence to reflect the need for studying PBLHT evolutions using high temporal-spatial observations in lines 80-82 as following:

"Observing atmospheric boundary layer transitions with high temporal-spatial resolutions is required to investigate the evolution of PBLHT, which will help to improve its representations in models (Su et al., 2020, Fritz et al., 2021)."

We agree with the reviewer that 'Fritz et al paper is different as they show that there are many important details that cannot be measured when using instruments with coarser resolution than their DTS'. In the paragraph, we wanted to emphasize that observations with higher temporal resolutions than ARM radiosonde launching are needed to study the evolution of PBLHT. To make this more clear, we added another reference to Su et al. (2020).

L115: 'limited cases' suggests a problem with those cases. Did you mean limited number of cases?

Response: We changed it to 'limited number of cases' in the manuscript.

L155: found from -> found in (or found from … to…)

Response: We thank the reviewer for pointing out the grammar issue. We changed it to 'found in' in the manuscript.

L162: Each ARM observatory generally launches four times a day (except twice daily at NSA and OLI) the balloon-borne sounding system (SONDE) at roughly 5:30, 11:30, 17:30, and 23:30 Universal Time Coordinate (UTC) (5:30 and 17:30 UTC at NSA 165 and 17:30 and 23:30 UTC at OLI). – This sentence is confusing. First you say 'each', then you say not each. Please be as precise as possible. You can say something like 'The balloon-borne sounding system is launched four times a day (at 5:30, 11:30, 17:30, and 23:30 UTC) at most of the sites, except twice a day at NSA (5:30 and 17:30 UTC) and OLI (17:30 and 23:30 UTC).

Are different UTC times chosen to match similar local times and facilitate morning and afternoon observations of PBL? If yes, please mention that.

Response: We thank the reviewer for pointing out the sentence structure issue and suggesting a better sentence structure. We rewrote the sentence based on the reviewer's suggestion.

At most ARM sites, SONDE launches are picked at different synoptic time (00Z, 06Z, 12Z, and 18Z) depending on the scientific justification, staffing, etc. For OLI, the SONDE launch times were chosen because it was when the observers were on site -7:00 am to 5:00 pm.

L167: "atmospheric dynamic environment" – justify why dynamic or simplify to atmospheric environment

Response: We deleted 'dynamic' in the manuscript.

L191: "boundary layer structure regimes " – boundary layer regimes (or structures);

Response: We deleted 'structure' in the manuscript.

L209: I raised that point before as to me you are calculating Ri in the layer from 0 to z. In your response, you said that you calculate Ri in a thin layer only, although in the text you now say that Ri is calculated at given altitude (not in a layer?). Then you show Eq. 1 and say that theta_vz and theta_v0 are the virtual potential temperatures at the surface and height z, which means that you actually calculate the temperature gradient in the entire layer between 0 and z, contrary to what you said. The same applies to wind: you can only use the values of wind at z knowing that they diminish at 0, otherwise their difference should be used.
I strongly suggest that the authors review their approach carefully. The gradient Richardson number formulation includes dz, which in your case is replaced by z, meaning that your layer always has the thickness of z: https://glossary.ametsoc.org/wiki/Gradient_richardson_number

Response: We thank the reviewer for pointing out the difference between gradient Richardson number and bulk Richardson number (Zoumakis 1992, Basu et al., 2014). The reviewer is correct that the bulk Richard number is calculated in the entire layer between 0 and Z. From literature, the bulk Richardson number is commonly used for radiosonde-based PLBHT estimations (Vogelezang and Holtslag 1996; Von Engeln and Teixeira, 2013; Chandra et al., 2014; Zhang e t al., 2014). To emphasize that bulk Richardson number is used in this study, we changed *Ri* to Ri_b and Ri_c to Ri_{bc} in the manuscript.

Reference:

Basu, S., Holtslag, A.A.M., Caporaso, L. *et al.* Observational Support for the Stability Dependence of the Bulk Richardson Number Across the Stable Boundary Layer. *Boundary-Layer Meteorol* 150, 515–523 (2014). https://doi.org/10.1007/s10546-013-9878-y

Chandra, S., Dwivedi, A.K. & Kumar, M. Characterization of the atmospheric boundary layer from radiosonde observations along eastern end of monsoon trough of India. *J Earth Syst Sci* 123, 1233–1240 (2014). https://doi.org/10.1007/s12040-014-0458-4

von Engeln, A., & Teixeira, J. (2013). A Planetary Boundary Layer Height Climatology Derived from ECMWF Reanalysis Data, *Journal of Climate*, *26*(17), 6575-6590.

Zoumakis, N.M. On the relationship between the gradient and the bulk Richardson number for the atmospheric surface layer. *Il Nuovo Cimento C* 15, 111–114 (1992). https://doi.org/10.1007/BF02507777

Vogelezang, D.H.P., Holtslag, A.A.M. Evaluation and model impacts of alternative boundary-layer height formulations. *Boundary-Layer Meteorol* 81, 245–269 (1996). https://doi.org/10.1007/BF02430331

Zhang, Y., Gao, Z., Li, D., Li, Y., Zhang, N., Zhao, X., and Chen, J.: On the computation of planetary boundary-layer height using the bulk Richardson number method, Geosci. Model Dev., 7, 2599–2611, https://doi.org/10.5194/gmd-7-2599-2014, 2014.

L420: If you hypothesize that PBLHT CEIL is greater than PBLHT Liu-Liang because of aerosols, explain what exactly you mean. Are you suggesting that free-tropospheric aerosol may cause some overestimation? If yes then be very specific in your explanation. What do you mean by 'transported aerosol layers'?

Response: We rewrite the sentence in lines 410-413 as following:

"probably because free-tropospheric aerosol layers transported from low latitudes have larger CEIL backscatter gradients than boundary layer aerosols and the top of the elevated aerosol layer is misidentified as the PBLHT by CEIL."

L423: "As being pointed by many previous studies" – this formulation seems grammatically incorrect

Response: We thank the reviewer for pointing out the grammar issue. We changed it to 'as was pointed by previous studies' in the manuscript.

L427: "The diurnal evolution of PBLHT is important to better understand boundary layer processes." – This sentence is confusing. Diurnal evolution of PBLH is the result of boundary layer processes, but I don't quite see what you want to say here. In the current form, this sentence says that without the diurnal evolution of PBLHT (i.e., for non-evolving PBL) you cannot understand PBL processes, which probably is not the authors intention.

Response: We agree with the reviewer. We deleted this sentence because the next sentence is enough to illustrate the point.

L483: "Ceilometer provides continuous measurements of aerosol backscatter profiles, which have been widely used to estimate the planetary boundary layer height (PBLHT). Good agreements between ceilometer- and radiosonde-estimated PBLHTs have been reported using

limited data from a single location or a short-term campaign. To test the robustness of ceilometer-estimated PBLHT under different atmospheric conditions, we compared ceilometer- and radiosonde-estimated PBLHTs using multiple years of U.S. DOE ARM measurements at six ARM observatories located around the world." – This paragraph needs some polishing (as many others), and I want to show the authors what I mean by that. Please see if my version makes more sense:
Ceilometer observations facilitate continuous measurements of aerosol backscatter profiles, which have been widely used to estimate the planetary boundary layer height (PBLHT). Good agreements between the ceilometer and radiosonde estimations have previously been reported for short-term campaigns at single locations. In this study, we extend that comparison to multi-year time series for nine different DOE ARM sites located over land and ocean in different climate zones.

Response: We thank the reviewer for showing an example of how to polish the paragraph, which greatly improved the readability of the paragraph. We changed this paragraph using the reviewer's version. We also carefully revised the manuscript with our best effort to improve the readability of the manuscript.

L487: "at six ARM observatories located around the world" – This is confusing. Once you say nine (Tab. 1, Fig. 1), then you say six. Again, I strongly encourage the authors to read their paper very carefully and check every single sentence for consistency.

Response: We thank the reviewer for pointing out the typo. The reviewer is correct that it should be nine.

Figs. 6, 8, 10 – what are the values corresponding to the colors (from red to blue)? Is that scale linear?

Response: We thank the reviewer for the comment. The KDE represents the continuous probability density function of observations in datasets and is derived with a bin size of 0.1 km in this study. We now use the same value range of from 0 to 1 in this revision and add a color bar for each figure. The scale is linear.

Reply to Reviewer #2's comments

Line 187: What is the lowest possible PBLHT from ceilometer? And is an event excluded from the statistics, when an altitude is found below this lowest altitude, also in any of the other methods? Can you include some info on the different vertical resolutions and their impact on the PBLHT (e.g. what happens if all data is smoothed to the same vertical resolution)?

Response: Therefore, the CL31 can provide PBLHT estimations from the ground up to 4 km (Münkel et al., 2010). However, since the vertical resolution for PBLHT-SONDE is 30 to 60 m, the minimum PBLHT from PBLHT-SONDE is usually higher than 90 m above the surface. Therefore, we only compare PBLHT higher than 90 m from both PBLHT-SONDE and PBLHT CEIL. We pointed out this in line 363-364.

We thank the review for pointing out the impact of different vertical solutions. To reduce the identification of spurious layers due to noisy data, the radiosonde data is subsampled at a 5 mb resolution, corresponding to vertical height resolutions of 30 to 60 meters depending on the atmospheric environment. We added the sentence in the manuscript in lines 166-168. On the other hand, the CL31 applies 30 minutes temporal and 360 m vertical sliding average and provides PBLHT estimations with a vertical resolution of 10 m. Compared with vertical solutions of 30-60 m for PBLHT-SONDE, the CL31 has a higher vertical resolution. We added these discussions in the manuscript in lines 280-281.

If all data are smoothed to the same coarser vertical resolution, it is expected the comparisons between PBLTH-SONDE and PBLHT CEIL will be better (in general, the lower the vertical resolution, the better the comparisons). However, this does not mean improved PBLHT estimations from either PBLHT-SONDE or PBLHT CEIL. In addition, it is also technically more complicated to smooth CEIL data to PBLHT-SONDE vertical resolutions as they are dependent on the atmospheric environment.

Line 196: Can you include some general info on the percentage within each quality index? How much was removed/used? Any statistical significant removal due to QC? This comment applies to all data sets where QC was used to focus on high quality data.

Response: The BL-VIEW algorithm assigns a quality index with a value from 1 to 3 to each boundary layer height candidate. The percentage of each quality index depends on boundary layer aerosol structures. We select the boundary layer height candidate with the highest quality index as the ceilometer estimated PBLHT for each ceilometer profile. Therefore, the quality index is not used to remove PBLHT estimations from the statistics.

We do use QC in the PBLHT-SONDE VAP to remove suspicious PBLHT estimations due to bad input data or unreasonable retrievals (e.g., estimated PBLHT > 4 km AGL). Unreasonable PBLHT estimations removed by QC flags are less than 10% of the total data. We added this

sentence in the manuscript in line 258.

Line 220: Richardson profile in lower left plot of Figure 3 starts only at about 400m?

Response: From Figure 3 c), potential temperatures decrease with height below ~400m, indicating that the atmosphere is superadiabatic near the surface. Therefore, bulk Richardson numbers are negative below ~400m and are not shown in the plot.

Line 341: Did you actually check if other bulk Richardson numbers than 0.25 and 0.5 perform better?

Response: We thank the reviewer for this suggestion. The magnitude of Ri_{bc} employed in previous studies ranged from 0.25 to 0.5. Seibert et al. (2000) suggests an optimal Ri_{bc} value of 0.25 for the PBLHT estimation when applied to radiosonde data. Therefore, the ARM PBLHT-SONDE Value-Added Product (VAP) provides estimated PBLHTs using two Ri_{bc} values of 0.25 and 0.5. We did not do extra testing of using other bulk Richardson numbers to derive PBLHTs as finding an optimal Ri_c for the PBLHT estimation is out of the scope of this study.

Reference:

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

Seibert, P., Beyrich, F., Gryning, S. E., Joffre, S., Rasmussen, A., and Tercier, P.: Review and intercomparison of operational methods for the determination of the mixing height, *Atmos. Environ.,* 34, 1001–1027, https://doi.org/10.1016/S1352-2310(99)00349-0, 2000.