Anonymous Referee #2

Review of the article titled "On the estimation of boundary layer heights: a machine learning approach" by Krishnamurthy and coauthors for publication in the Atmospheric Measurement Technique.

The authors have used a machine learning (ML) approach to improve the retrieval of boundary layer depth from the data collected by the Doppler Lidar. They first develop a ML model to calibrate the DL retrieved PBL depth with that derived from the radiosondes. As the radiosonde measurements are temporally sparse, they use the higher resolution PBL depth retrieved from the Doppler Lidar to understand boundary layer parameters affecting it. In the end they also evaluate two days of output from two different models. The article is overall well-written and is easy to follow. However, the article can be further improved by addressing the following concerns. These can be regarded as minor revisions.

We thank the reviewer for carefully reading the article and providing constructive feedback. We believe the quality of the article has improved by addressing the comments and hope our revisions are acceptable to the reviewer. Below the reviewer comments are in black and the authors responses are in blue.

Major Concerns:
It will be good to add some discussion in the last section on the use of machine learning in deducing PBL depth, and understanding its controls. The authors have mentioned and acknowledged several things in the text, i) like the training could have been performed by using a different estimate of PBL depth from the radiosonde, and ii) how the authors are only demonstrating the use of ML for deriving the PBL in the nighttime, but refrain to call it the "true" nighttime Zi (Line 313-315). This is simple the limitation of the use of ML in deriving physical understanding. This should be discussed in the text in detail. If the authors truly believe (#2 above) to be the case, then can you trust the numbers reported in Table 4 and 5? Maybe the Tucker method is correct and just the training needs to be done on a different dataset. This concern does not mean that the article is not valuable, however this needs to be addressed in the text. Thank you.

These are very good points raised by the reviewer and we have tried to address these in the updated manuscript and provide further explanations below.

The sensitivity of the RF model indeed depends on the reference data used for training, but the choice was made by comparing the standard Tucker method to all radiosonde z_i estimates generally used by the community. We found that the correlations were higher based on the Tucker method, which we know is generally biased low in many conditions but is precise. So, we believe the choice of the reference data (Liu & Liang based approach) is better suited for the ARM SGP C1 site. For another site, it would be prudent to conduct a preliminary comparison against lidar derived estimates and evaluate which the radiosonde algorithm correlates well with the remote sensing observations. For example, in arctic climate (say Alaska) we would expect higher percentage of stable boundary layers and therefore the z_i estimates based on Richardson number might be well suited for training. We have clarified that the choice of reference data is indeed site specific in the updated manuscript (Page 8 Line 189 to 198).

With regards to the nighttime z_i estimates, this is a part of ongoing research by the larger community to define the true nocturnal boundary-layer height (Zilitinkevich and Baklanov 2002, Vickers and Mahrt 2004, Steeneveld et al. 2006, Richardson et al. 2013). During stable boundary layers, the determination of the PBL height is very uncertain. Turbulence in the stable boundary layer, can result from either buoyancy forcing or wind shear. At SGP C1, the nose of the low-level jet can also be used to define the height of the boundary layer (Sivaraman et al. 2013). Therefore, the subject of this article is not delve onto which nocturnal z_i estimates from the radiosondes are more accurate but rather to develop a methodology/framework to provide continuous nocturnal z_i estimates. We have made the point clearer in the updated manuscript that we are not doubting the results from the RF model but we are showing a methodology that can be adapted to future research needs associated with the depth of stable boundary layers (Page 8 Line 190 to 198). Moreover, in many stable boundary layer conditions, the true boundary layer height could be lower than the height of the first range gate of the lidar, so such a technique could provide accurate z_i estimates for such conditions.

During convective boundary layers, the definition is clearer and can more easily be discerned from radiosonde datasets. In Section 4, we provided a thorough evaluation during both convective and stable boundary layers compared to radiosonde estimates, which are assumed to be the true boundary layer height estimate at SGP C1. We hope from the results shown in Section 4 (Figure 5 and Table 3), readers can conclude that the Tucker method is not as accurate as the RF model z_i estimates.

References:

Zilitinkevich, S., & Baklanov, A. (2002). Calculation of the height of the stable boundary layer in practical applications. Boundary-Layer Meteorology, 105(3), 389-409.

Steeneveld, G. J., Van de Wiel, B. J. H., & Holtslag, A. A. M. (2007). Diagnostic equations for the stable boundary layer height: Evaluation and dimensional analysis. Journal of applied meteorology and climatology, 46(2), 212-225.

Vickers, D., & Mahrt, L. (2004). Evaluating formulations of stable boundary layer height. Journal of applied meteorology, 43(11), 1736-1749.

Richardson, H., Basu, S., & Holtslag, A. A. M. (2013). Improving stable boundary-layer height estimation using a stability-dependent critical bulk Richardson number. Boundary-layer meteorology, 148(1), 93-109.

Figure 10 and associated text: it is a bit confusing as to the whole purpose of this exercise. Just because the variance is being scaled by a higher PBL depth, the profile will look different. So not sure how it speaks to the Random Forest (RF) PBL depth being better than that derived by the Tucker method. Also, the variability of variance is probably huge, so the differences wouldn't be statistically significant anyways. This needs to be clarified in the text, or else removed from the manuscript. Thanks.

The last two sections in the manuscript are case studies showing the importance of an accurate boundary-layer height estimation. We agree that a higher boundary-layer height will change the profile, as we are scaling it with a different PBL height (Page 24 line 528). Since we have already shown that the RF model z_i is more accurate than the Tucker method, this section is showing the amount of uncertainty in using one method over the other at SGP C1. As shown in Page 24 Line 525, the average uncertainty in using the RF model z_i compared to the Tucker method z_i can result in approximately 10% uncertainty in convective velocity estimates and almost 15% in vertical velocity variance profiles. Normalized profiles are generally used in boundary-studies and atmospheric models (Lenschow et al. 1980), and understanding the importance of accurate z_i estimates is highlighted. We have provided additional motivation for this section in the updated manuscript (page 23 line 502 -507).

Minor Concerns:

Line 14: Might be better to say four years rather than multi-year. Thanks.

Agreed. Corrected.

Line 41: MISR is mis-spelled.

Since the abbreviation was only used once in the article, this has been removed in the updated manuscript.

Line 42:43: The satellites measure cloud top temperature from which the cloud top heights are calculated. During cloudy conditions, it is assumed that the PBL top corresponds to cloud top heights. This statement states that there has not been any validation of the satellite derived cloud top heights. Please add reference to support this, or else remove. Thanks.

For brevity, we have removed that statement from the paper. Although, we did some further research, and found only one recent paper (Böhm et al., 2019) which does a statistical evaluation of the MISR cloud base heights and ground-based Ceilometer observations.

Böhm, C., Sourdeval, O., Mülmenstädt, J., Quaas, J., & Crewell, S. (2019). Cloud base height retrieval from multi-angle satellite data. Atmospheric Measurement Techniques, 12(3), 1841-1860.

Line 58: you mean rely and not relay.

Typo corrected. Thanks.

Line 64: better word would be "lowest gate" rather than minimum range.

Agreed. Corrected in the updated manuscript.

Table 1: It will be good if you add units to the measurement features. Thanks.

Agreed, we currently have added units to the measurement features in Table 1.

Line 145-146: please revise this sentence. Thanks.

Agreed. We have added some additional text to this section, based on some of the earlier comments and comments from the other Reviewer. The sentence was rephrased as well. Please see Page 7 line 154 to line 164 for the changes.

Line 165: The numbers do not add up. Four years of data should equal 1460 days, not sure how you got 1785 days.

The ARM site has at least 2 radiosondes during daytime conditions and a higher frequency during some field campaigns, so for 4 years there are approximately 2920 radiosonde launches. We used 4 years of data and found approximately 1785 cases (not days) in our analysis. This typo was fixed and the statement has been rephrased to "…1785 cases with radiosonde data and daytime clear (identified as periods when surface heat flux is positive from sunrise to sunset and cloud base height is zero) or shallow cumulus conditions (identified as cloud base height less than 5 km from Doppler lidar and cloud fraction less than 0.1) for the years 2016 through 2019."

Line 285: you mean "hourly" cloud fraction greater than 0.1?

Yes, we use a rolling hourly average estimate of the cloud fraction to differentiate the measurements. The Doppler lidar, although, provides a cloud fraction estimate every 15 minutes. We have included this in the manuscript.

Figure 8: Please describe the vertical bars in the caption.

Yes, we have mentioned the vertical bars represent one standard deviation in the caption.

Figure 11: Looks like the LASSO simulations are able to accurately capture the development of the daytime PBL. I assume that the E3SM values are within the model range resolution as well. So, this is very good news for the modelling community and should be highlighted.

We thank the reviewer for the comment. We agree and have highlighted the motivation in the updated manuscript.