Response to reviewers for the paper *A multi-instrument fuzzy-logic boundary-layer top detection algorithm* by Elizabeth N Smith and Jacob T Carlin - *Anonymous Reviewer # 2*

To guide the review process we have copied the reviewer comments in black, italicized text. Our responses are in regular blue font. We have responded to all the referee comments and, when appropriate to do so, shared the resulting alterations to our paper in a darker blue, italicized text.

The authors' reply clarifies some of my comments. However, there still some confusing statements/definitions/discussions for me as listed below.

Comments:

1. *While I concur that leveraging the synergy of multi-remote sensing boundary layer profiling measurements could enhance PBLH estimations by providing more constraining information, I find the authors' claim that 'the proposed method enables accurate boundary-layer height estimation both during daytime and nocturnal conditions' unconvincing based on the comparisons in Figures 8-11. The PBLHs estimated by remote-sensing (Fuzzy-logic) are consistently lower than those derived from radiosondes for PBLHs > 500 m or during afternoon convective boundary layer conditions (Figure 10). Why or what causes the underestimations from Fussy-logic PBLHs?*
   We make a statement that very closely matches the reviewer's observation lines 343-348. *From the bulk comparison shown in Fig 8, a few things become immediately apparent. The fuzzy logic method applied to CLAMPS observations is more likely to estimate a lower BL height than radiosondes. This is especially true in the afternoon hours, which correspond to 1800 and 2100 UTC for the area where these observations were collected. No 2100 UTC BL height comparisons (shown in yellow in Fig. 8) fall on or below the one-to-one line. The majority of 1800 UTC and most 0000 UTC BL height comparisons also trend above the one-to-one line, suggesting the fuzzy logic method applied to CLAMPS observations estimates a shallower BL. Earlier hours (0600 and 1200 UTC) do not demonstrate as much of a signal.* We then take a look at examples from CHEESEHEAD with more frequent soundings available to get a temporal comparison with radiosonde-based estimates in the following lines 348-355, confirming the prior statement. Then, to examine this further (i.e., answer the question posed in this comment) we do the exclusion analysis. This exclusion process was motivated specifically to better understand the potential causes of discrepancies between the radiosonde and CLAMPS-based BL height estimates. The exclusion criteria are described in the paper, and were used to successively exclude data pairs and get a better sense of algorithm performance. As each criteria was applied successively, pairs were removed from the comparison dataset and analysis was performed at each step, including statistics. This is all described in the paper in the final part of section 4. We can briefly review it here for more direct clarity and added commentary in the context of this review:

The first exclusion step eliminated cases that were deemed ambiguous without confident PBLH detection from radiosondes and/or when the detected PBLH appeared to be related to a different feature altogether (clouds, residual layers, etc). While this may appear 'subjective,' the dataset shown in 11a (after this exclusion) is similar to the dataset other studies might *start with*, never including such ambiguous cases (see, e.g., Banghoff et. al 2018-JTECH, Heinselman et al. 2009-JTECH). Removing these cases from the comparison did not present any pattern in the dataset, as expected.

As shown in Fig 11b, many of the pairs that indicated the fuzzy logic underestimated the PBLH relative to sondes were related to cases where CLAMPS observation platforms simply did not collect observations over a deep enough layer of the atmosphere (e.g., environments with relatively few available scatterers for the Doppler lidar, the deepest PBL cases resulting in important phenomena occurring in the upper, more coarse-resolution observing zone of the thermodynamic profiler, etc.). Users should be aware of the limitations of any instrument or dataset provided to any algorithm they use, and such limitations would likely apply to any remote-sensor, including lidar, based method. There are specialized lidar scanning strategies that could have been deployed (as in Bonin's work) and lidar post-processing algorithms, and special options in thermodynamic retrievals algorithms that could be applied by a user to attempt to increase the depth of the CLAMPS observed fields, but this is application specific and beyond the scope of understanding this algorithm.

Next, excluding cases where the radiosonde and CLAMPS-based methods identified different parts of the BL-top removed some more pairs mostly in the area just above the one-to-one line: instances when the fuzzy-logic method underestimated the PBLH relative to the sonde methods, but not by as large of a margin as in the previous set of excluded cases. Detecting different parts of PBLH can be connected to 1) ambiguity in defining what the PBLH is when the chosen defining parameters exist more as a *layer* (i.e., a relatively deep entrainment zone) than as a discrete level and 2) the methods by which different instrumentation senses the atmosphere and the PBLH estimation methods applied to those data. As discussed in the paper, the CLAMPS-based method is more likely to find a BL height closer to the surface since the fuzzy logic algorithm is a bottom-up algorithm that searches for the first point above the surface where the membership value crosses a threshold. Many points above this level could also be equivalent to this threshold value, but the algorithm will select the first one by design to make sure any PBLH detected is surface-based. The methods applied to radiosondes are, as discussed, variable. With regard to the instrumentation, thermodynamic retrievals will not detect shallow, minor gradients that radiosondes might. Similarly, sondes may also be more apt than the retrieval to capture thermodynamic structures associated with deep PBL top zones (e.g., double inversions, shallow elevated isothermal layers, etc.). Accordingly, CLAMPS is most likely to capture the *bottom* of these zones, while various methods applied to sondes may result in different representations.

The final exclusions are cases with complex structures in the PBL as determined by examining the radiosonde profiles and CLAMPS observations---not the PBLH estimates. As described in the case of detecting different parts of the PBL top, CLAMPS thermodynamic retrievals are not likely to fully represent shallow gradients due to the optimal estimation approach. In the event the complex structure includes saturated layers near the surface, the thermodynamic profiler data may not include as much independent information for the retrieval process, adding even more uncertainty. This is well known and documented in the literature concerning the retrieval and in dataset documentation. Additionally, the various PBLH techniques applied to radiosondes are not necessarily well suited for profiles with complex temperature and moisture gradients. For these reasons, we excluded these pairs in the final round of exclusions---seeing no clear pattern in the removed data which was expected---leaving a dataset that we found most representative of cases where the observing conditions were generally ideal for examining the behavior of the *algorithm* without conflating it with variability of observation quality across cases. That dataset is shown in panel 11d. While many, if not most, of the instances of overestimation have been excluded, some cases where the fuzzy logic algorithm applied to CLAMPS observations leads to an underestimation of BL height relative to sondes persist. There is not an immediate clear pattern to these pairs. The majority of the pairs are in the later half of the diurnal cycle, but they span all three considered locations. An obvious reason for this persistence is not yet clear. As stated in the paper, more frequent and regularly available radiosondes are required to extend this analysis in order to deepen understanding about the impact time of day has on the performance of the fuzzy logic algorithm itself. It would also be interesting to add more locations with profiler-sonde comparisons to evaluate the impacts of moisture in these comparisons.

2. *Based on Figure 2, it appears that the second-generation step yields even lower PBLHs than the first-generation step. This suggests that the second-generation step further underestimates PBLHs under conditions of a convective boundary layer. Is that correct?*
As stated in the prior response, on any one day at any one observation time, we could choose to show a 'better' or 'worse' comparison to either generation estimate. Forming arguments or opinions about data from a single comparison point is not a particularly informative method and not the method we choose to determine the 'performance' of any technique. Comparing with the estimates from the first and second generation is not the intent of this work or question we are seeking to address; instead we are looking to combine independent observations from a multi-instrument platform and explore the synergistic PBL information this approach can provide for boundary-layer height estimation, particularly via this fuzzy logic technique. Further, decreasing the height of the PBLH estimate is not necessarily a degradation---there are scenarios in which lidar-based estimates could be too high (e.g., elevated cloud bases). Setting that aside for now and following the reviewer down this path for a moment, taking a closer at

20200625 (Fig 3) we can see that the 2nd generation estimate does not always produce lower estimates. See the period between 16 and approximately 18 UTC, from 2030 to 21 UTC and several of the individual estimates without the ten-min triangle smoothing applied (dots), which are sometimes deeper estimates than their first-generation counterparts. Similar behavior is apparent on 20190926 (Fig 10), where the period from 15 to 19 UTC and from 10-11:30 UTC. At all times of the day, the second generation step adds synergistic observations. This can include more information from the lidar and information beyond what a lidar can provide. This manuscript fundamentally intends to introduce an algorithm with this capability. If users are interested in lidar-only methods, those are already available---eliminating the second generation would effectively turn this algorithm into a repackage of those techniques.

3. *The comparisons between Fuzzy-logic and radiosonde PBLHs do not seem as robust as those shown in previous studies that exclusively used Doppler lidar measurements, such as those in Tucker et al. (2009) and Krishnamurthy et al. (2021). Could you provide the correlation coefficients between Fuzzy-logic PBLHs and Radiosonde PBLHs as depicted in Figures 8, 9, and 11?*

   References:

   Tucker, S. C., C. J. Senff, A. M. Weickmann, W. A. Brewer, R. M. Banta, S. P. Sandberg, D. C. Law, and R. M. Hardesty, 2009: Doppler Lidar Estimation of Mixing Height Using Turbulence, Shear, and Aerosol Profiles. J. Atmos. Oceanic Technol., 26, 673–688, https://doi.org/10.1175/2008JTECHA1157.1.

   Krishnamurthy, R., Newsom, R. K., Berg, L. K., Xiao, H., Ma, P.-L., and Turner, D. D.: On the estimation of boundary layer heights: a machine learning approach, Atmos. Meas. Tech., 14, 4403–4424, https://doi.org/10.5194/amt-14-4403-2021

   The requested analysis was part of the original manuscript and has never been removed. Please see table 3, values reported in Fig. 9 and discussion of values throughout section 4. Re: Tucker et al 2009, we would expect those comparisons to match more closely than ours for a couple of reasons. 1) the lidar system used (HRDL) operates at a different wavelength and generally a bit differently than the system we used, so it often collects stronger returns from aerosol backscatter and can profile slightly deeper into the atmosphere 2) the approach described defaults to defining the mixing layer height using empirically derived thresholds of vertical velocity variance, but in the event an estimate cannot be made this way at a given time, the next best option is selected based on the environmental conditions, then finally a manual check to ‘eliminate problematic data and/or unexplained discontinuities.’ Re: Krishnamurthy et al. 2021: This is a random forest approach using a large number of inputs that is trained and tested at one location (without considering missing data, which is a common problem and focus of research in machine learning). If you examine their Figure 3, you will see some different perspectives on Tucker-based comparisons. Even with the largely different approaches and the more varied regimes of the data we include, the correlations we report from the

exclusion analysis in our Table 3 are generally comparable to the correlations shown in their Figure 5.

4. *The authors have noted that in the first-generation step, the temperature inversion height was used as an input. Could you clarify how the temperature inversion height is derived? What degree of temperature difference is considered to constitute an inversion? The term 'temperature inversion height' is introduced in line 139 without an explanation of how it is obtained. In the response, the authors claim, 'We have never stated that inversion height is derived from the first-generation step.' If so, could you specify when the 'inversion height' is derived?*

   The temperature inversion has been included in the algorithm and in the description of the algorithm in all versions of the manuscript. CLAMPS provides temperature profiles in the form of thermodynamic retrievals from infrared spectrometers and/or microwave radiometers (depending on instrument availability). The retrieval using an optimal estimation approach, which is an ill posed problem. This algorithm and its outcomes have been well documented and published over many years. We know that retrieved profiles are unlikely to represent shallow or minor gradients. As such, we compute the gradient directly as dT/dZ and find the inversion height as the first level where dT/dz reverses sign (e.g., becomes positive) above the surface and below 1.5 km, since elevated inversions are of interest for this application. Since the algorithm output is not expected to depict minor fluctuations in the thermodynamic profiles, there is no required ‘threshold’ for dT/dz to qualify as an inversion.

5. *Typically, a figure caption should provide a brief explanation of what the figure represents, such as the observations or variables depicted in Figure 2. This would provide readers with basic information, and I don't believe it would be as 'extensive' as the authors suggested in their response. They could relocate the detailed discussions or illustrations of the figure to the main body of the text.*

   The caption has been revised again. *Time-height cross-sections showing the diurnal evolution of CLAMPS1 observations collected on 25 June 2020 in Norman, Oklahoma which are used in the fuzzy logic algorithm to produce the aggregate fields and BL height estimates shown in Figure 3. Panels (a-c) show the variables used in the first-generation step of the fuzzy-logic algorithm [(a) vertical velocity variance, (b) high-frequency vertical velocity variance, (c) temperature inversion height], with the weighting given to panel (c) shown in (d) based on sunrise/sunset time, while panels (e-j) show the variables used in the second-generation step of the fuzzy-logic algorithm [(e) backscatter intensity, (f) backscatter intensity variance, (g) u-wind, (h) v-wind, (i) temperature, and (j) water vapor mixing ratio]. Panel (a) includes a subsetted window (a1) between 00—12 UTC with a finer colorscale to show overnight values of vertical velocity variance. Each observed field is labeled with units (if applicable) on its respective panel color bar.*

6. *Despite several rounds of discussion, the precise definitions of 'buoyancy-driven processes' and 'mechanically-driven processes' remain unclear*

The Tucker et al. 2009 paper the reviewer referenced uses yet another version of terminology: "shear-driven boundary-layer." At this stage, after many rounds of discussion, the authors respectfully propose we agree to disagree in the absence of obvious community defined norms.

7. *The authors replied that 'if dz = the spacing between any two levels in a profile, and we know that dz varies throughout the profile, 300 is the average of all dz values in the profile. The minimum dz that occurs in the lowest levels of the atmosphere is on the order of 10s of meters. The maximum dz that occurs higher up in the atmosphere is on the order of 100s of meters.' If a group of numbers have a minimum of 10 and maximum of 100, how can the average to be 300?*
   We clearly state **on the order of 100s of meters**, not that the maximum is 100 m. The "order" of 100 meters can include values as low as 100 m and up to and including 999 m. Written in mathematical interval notation, this would be expressed as [100,999]. The average value of dz can therefore be 300 m since 100 m < 300 m < 999 m.
8. *As for the comparison of high- vs. low-resolution sounding data, the authors state, 'Analyzing the methods separately highlights that using high-resolution sounding data tends to lead to lower median BL height, but the ranges and interquartile range values do not change much between coarse and high resolution datasets. There is no indication from this analysis that using high resolution data necessarily yields more accurate BL height values'. It appears that the authors are suggesting that because 'the coarse and high resolution datasets have similar PBLH ranges and interquartile range values', therefore, 'There is no indication from this analysis that using high resolution data necessarily yields more accurate BL height values'. I don't believe this is a solid conclusion. For instance, under an extreme condition where the coarse sounding data produce identical PBLHs, resulting in ranges and interquartile range values of zero, it certainly doesn't imply that coarse sounding data provides accurate PBLH estimations.*
   The reviewer provides a hypothetical condition "where the coarse sounding data produce identical PBLHs, resulting in ranges and interquartile range values of zero." We are interpreting this to mean the coarse data lead all applied methods for estimating PBLH to output the same result. It is unclear why this hypothetical "extreme condition" would only impact the coarse resolution data. Further, we state directly that "there is no indication *from this analysis*" meaning in the data presented in the paper. Those data do not include such a condition. Most fundamentally, we aren't sure exactly how this would be connected to our assertion that, based on our analysis, we cannot make any clear statement about one dataset being more accurate than the other. That assertion has less to do with the IQR of each method and is more about the fact that, absent any other independent source of data besides the radiosondes, it is not clear that the pros/cons of both the hi-res and coarse soundings (e.g., methods applied to the coarse data may not be misled by small fluctuations, but important small details may similarly be missed in other cases) necessarily mean one is better than the other overall. In any case, to hopefully

provide the assurance that is perhaps needed to resolve this concern we have performed additional testing on these comparisons. We used bootstrap hypothesis testing to estimate the confidence interval of test statistics by repeatedly calculating the test statistics on bootstrapped samples. If the test statistic falls outside of the resulting confidence interval, then we reject our hypothesis. Our hypothesis in this case is that the coarse-resolution and high-resolution datasets are similar for any given radiosonde-based PBLH estimate method. For each method, we performed a 10,000 bootstrap resampling analysis twice to compare the median and IQR as the test statistic for which we defined confidence intervals with 95% confidence. In all instances (that is for all methods in terms of the median and the IQR) the test statistic fell within the confidence interval, indicating that, regardless of the technique, differences between PBLH estimates based on coarse- and high-resolution are not statistically significant. The following was added to the text: *Analyzing the methods separately highlights that using high-resolution sounding data tends to lead to lower median BL height, but the ranges and interquartile range values do not change much between coarse and high resolution datasets. To be certain, we conducted 10,000 bootstrap resampling analyses for each method to evaluate if the median and interquartile range values were significantly different between high- and coarse-resolution datasets. We found with 95% confidence that there were no statistically significant differences. There is no indication from this analysis that using high-resolution data necessarily yields more accurate BL height values, and often users have access only to data that we have classified here as coarse resolution (i.e., publicly available and accessible National Weather Service radiosonde datasets). Our intention with this analysis is to understand and consider possible implications of including different resolution sounding datasets. Moving forward, we use high resolution sounding data when it is available. If it is not available a coarse resolution radiosonde profile is used instead.*