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 # 1*

We thank the reviewer for their time and comments on our paper. 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 blue, italicized text. Any line number references are to those in the originally submitted manuscript.

This paper presents the results of using a fuzzy logic algorithm with a combination of doppler lidar, radiance [interferometer] and microwave radiometer, with validation done by comparisons with radiosonde data. I feel the paper needs a major revision as the presentation is confusing and the results are not very clear. I realize that there are lots of different cases to present, and many options in the radiosonde comparisons, but in the end, it's not obvious as to how these retrievals might ultimately be used. But I think a rewrite could make it much easier to extract this information. My specific comments are:

1. *Both the and abstract should state more clearly what observations are being used for the retrieval and validation (see the first sentence above).*
   We modified language in both the abstract and the Summary and Outlook sections to be more explicit about the three instruments used to develop the algorithm, while remaining clear that the algorithm could be adapted to use other similar observation types.
   Abstract: In the second paragraph the first sentence and last sentences were modified:
   This study introduces a fuzzy-logic algorithm that leverages the synergy of multiple remote-sensing boundary-layer profiling instruments: Doppler lidar, infrared spectrometer, and microwave radiometer.
   ...
   While developed with the three instruments mentioned above, the fuzzy-logic boundary-layer top detection algorithm, called BLISS-FL, could be adapted for other wind and thermodynamic profilers. BLISS-FL is released publicly fostering collaboration and advancement within the research community.
   Summary and Outlook: In the opening paragraph of this section we added a parenthetical after platforms
   While this algorithm was developed for use with the CLAMPS platforms (i.e., Doppler lidar, infrared spectrometer, and microwave radiometer), it could be applied to or adapted for similarly instrumented facilities such as but not limited to the Department of Energy's Atmospheric Radiation Measurement profiling facilities.

2. *Line 110: do you mean the first generation step in B18, or is this changed in the current algorithm?*
   We mean there is an addition to the algorithm at a certain stage, and we will highlight it when we get there. In addition to the existing language "with any deviations from or expansions upon B18 highlighted", we modified the text to more clearly state "*At this stage the algorithm design deviates from the design of B18*" at the beginning of the paragraph which started on line 125 in the original manuscript. This language should make it clearer that the differences are now intentional and by design, not only differences in available variables. We hope that adding this language at the location when the addition to the algorithm is introduced makes it clear that prior to this the algorithm design is generally consistent.
3. *Line 126: Why are buoyant processes important during the night? This seems counter intuitive as buoyancy should grow during the daytime.*
   Buoyancy does grow during the day. Buoyant processes are not limited to the growth of buoyancy. Specifically in the scope of the algorithm we are discussing, the first generation step relies heavily on measures of mixing, or in other words observations showing that mixing is ongoing. In this case, that means mostly mechanical mixing observable by Doppler lidar. Of course, the buoyant processes related to daytime heating can lead to that mixing, but those buoyant processes are not directly observed by the Doppler lidar. Buoyant processes are important during the night when the daytime boundary layer decays and the mechanically driven processes may become less dominant, which in the scope of this algorithm can make the Doppler lidar observations of mixing less informative. As such information about the presence of the daytime capping inversion, thermodynamic structure in the residual layer, and the evolution of any surface inversion is critical at night. In non-canonical cases, the classical "Stull" evolution is not representative of boundary-layer evolution including mechanical or buoyant processes. Including both, specifically through the night, is useful during transitions and during non-canonical nocturnal boundary-layer cases.
4. *Line 131: So you mean that the B18 algorithm was not successful in estimating overnight BL height?*
   The reviewer brings up a fair point that the B18 algorithm is applied throughout the full diurnal cycle. In their paper, they do not offer any assessment or verification at night. In any case, the current phrasing implies that they only applied their method to daytime hours which is inaccurate. We will rephrase this for accuracy.
   In order to capitalize on the availability of thermodynamic profiles and to improve the capability and robustness of our algorithm during nocturnal hours when the mechanical mixing that backscatter-based instruments (e.g., Doppler lidar) observe can be absent or limited to the residual layer (Schween et al. 2014), the first-generation step also includes temperature inversion height as an input variable (Fig 2c).
   Schween, J. H., A. Hirsikko, U. Löhnert, and S. Crewell, 2014: Mixing-layer height

retrieval with ceilometer and Doppler lidar: From case studies to long-term assessment. Atmos. Meas. Tech., 7, 3685–3704, https://doi.org/10.5194/amt-7-3685-2014.

5. *Section 3: Do radiosonde observations measure thermodynamic BLH rather than the ML height? If so, is it really a good comparison?*

   In the scope of this algorithm development, we view mixing and thermodynamic contributions to boundary-layer height as coupled. The algorithm includes measures of mixing as it occurs and indicators mixing has occurred. Thermodynamic BLH parameters can be related to indicators (i.e., mixing has occurred) while MLH would be related to measures (i.e., mixing is occurring). Also, as mentioned in the paper, radiosonde observations are a commonly available and known source of data, which allows the comparison to be made against a dataset that is already integral and 'standard' in many atmospheric data communities.

6. *Line 402: It is fine not to recommend a preferred radiosonde PBLH algorithm, but it would be really helpful to see a plot like*

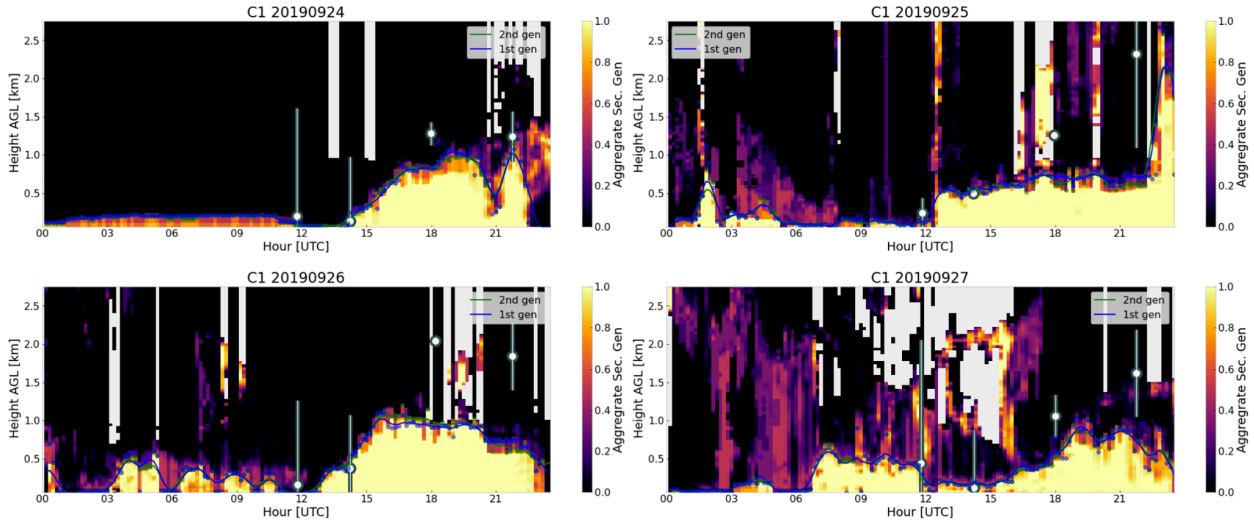
   In the absence of any additional context, we are treating this comment as if it is related to comment 7 as it seems incomplete on its own. We apologize if our assumption is incorrect.

7. *Figure 3 is really helpful to see the evolution of the PBLH estimate during the course of a day. It would be even better if you could plot the radiosonde estimates of PBLH on top of this. There is lots of discussion of how the algorithm does at different times of the day, but much of this could be clarified with a comparison like this.*

   The strength of an algorithm like the one we present here is we are able to capture or at least make estimates about BL evolution that is not able to be captured in typical radiosonde observations. In our case, we are fortunate to have some additional temporal resolution in the CHEESEHEAD radiosonde dataset (which featured up to 4 radiosonde launches/day) that makes this sort of comparison partially possible. However, the days and regimes in which comparison is available are not fully representative of the comparison we do with the full radiosonde dataset. CHEESEHEAD data are collected in a forest canopy, which perhaps has implications for BL height evolution and BL height detection which are not accounted for in this work. In any case, it is the dataset we have available, so we can include the comparison as requested. We now include a series of four days from the CHEESEHEAD project while the 4-times daily radiosonde releases were occurring. This comparison shows a spread of 'performance,' but generally shows what is expected based on the scatter plot comparisons. Differences are smaller during the early part of the daytime BL. The biggest differences typically occur latest in the afternoon.

   To include this figure in this text, we had to reorganize some of the text in Section 4. Previously the text read as introducing the median soundings scatter plot, describing what it shows, then comparing to the scatter plot of all the sounding methods to describe why the median is important. Now the text reads as introducing the median scatter plot,

moving straight to comparing to scatter plots of all methods to describe why all methods are important, then returning to the median scatter plot to describe what it shows. Then we follow that description with complementary information now available from showing the aggregates from CHEESEHEAD with soundings overlaid with respect to time.



8. *The summary section (5) doesn't really have any concrete conclusions. As you say, it's hard to make a fair comparison since a 100 m difference means something very different depending on the time of day. But plotting the height estimates in comparison to radiosondes as a function of time would really help with this.*

The radiosonde time series plot is now included in the manuscript in response to another comment. In a more direct response to this comment, the Summary and Outlook section, particularly the 4th paragraph, was revised to be more explicit about conclusions and outcomes. We discuss that while there is agreement between radiosondes and the algorithm, there is a pattern of underestimation. We speculate about the reasons behind this including the limitations of remote sensors. We also now more explicitly discuss that exploring errors through comparison with respect to time would require more data, specifically more data around the diurnal cycle to create discrete samples and or develop normalization techniques. The changes appear starting at approximately line 422 in the original manuscript.

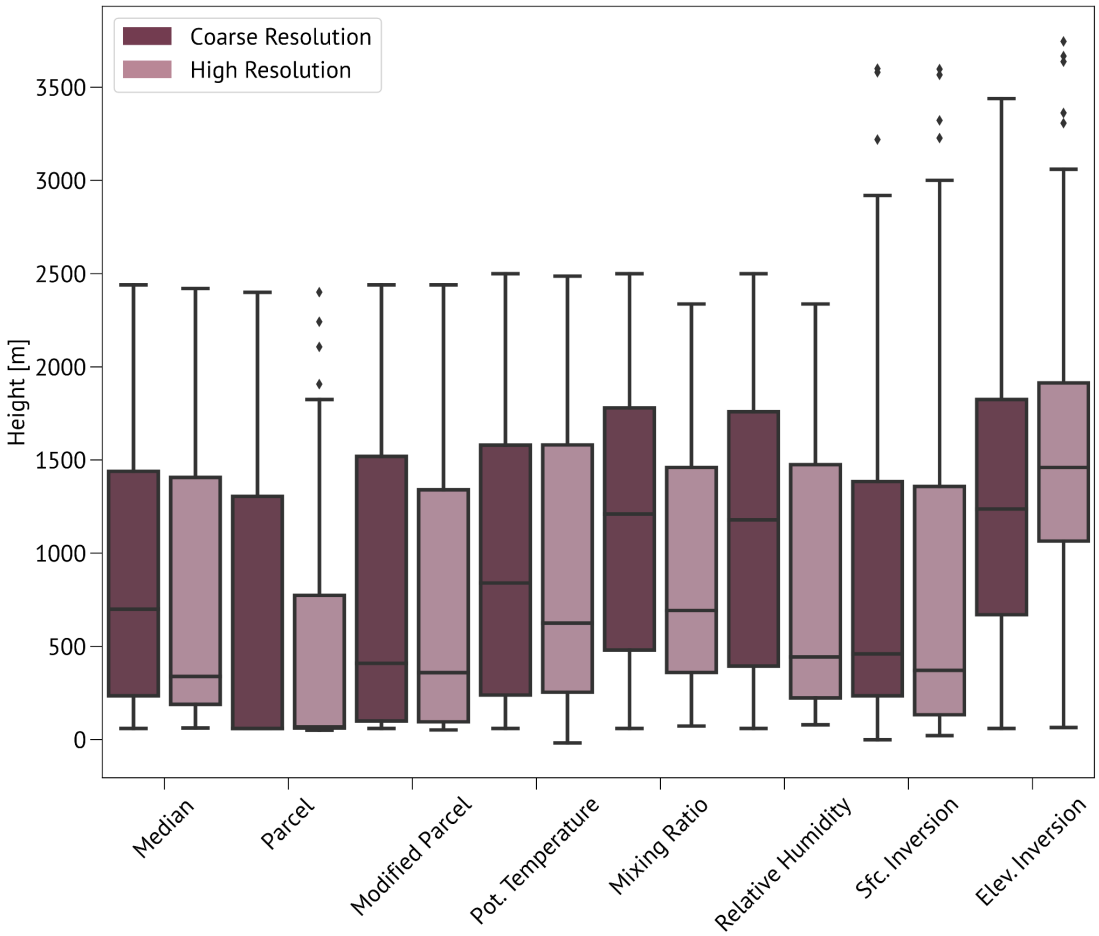
This comparison suggests fairly strong agreement between the techniques, but still with a maintained pattern of the fuzzy logic algorithm applied to CLAMPS observations underestimating BL height compared to the median of radiosonde-based methods. Given the instrumentation types included in the algorithm development and evaluation (Doppler lidar, infrared spectrometer, microwave radiometer), we speculate that even in the best-case comparisons, remote-sensing-based observations are simply more inclined to detect the lowest possible indication of BL height (especially in cases where the BL top itself may be a layer with depth). These types of instruments are also more likely to lose measurement capability and resolution with distance from the surface. These reasons

could combine to result in systematically lower estimates of BL height than those provided by in-situ platforms like radiosondes.

Unlike some similar algorithms, this approach has the capability to utilize thermodynamic observation information and kinematic observation information (when available) to provide BL height estimates throughout the diurnal cycle. Specific analysis is focused on the early morning and late overnight periods. While some statistics included suggest perhaps even minor improvement compared to the daytime group, this may be a misleading result. For example, a bias of 100 meters means something different when the BL height is O(100 m), which is common in the overnight and morning hours, compared to O(1000 m) BL height values in fully developed daytime and afternoon BL. More data is needed in this comparison to understand the role time of day plays in how the fuzzy logic algorithm behaves. More data could provide the opportunity to build time-dependent samples or to develop temporal normalization approaches to control for this important sensitivity.

9. *Figure 7: It's not clear what you mean by dark and light colors.*
   In the original manuscript, "dark" and "light" colors referred to the shading/brightness of boxplots for each set of colors (e.g., for the modified parcel method, dark green referred to the results using coarse radiosonde data and light green referred to the results using high-resolution data). We revisited this figure to determine how to present the information more clearly. We came to the conclusion that using different colors for each method and relying on the light/dark shading to show the differences between the box pairs for each method was too confusing. We updated the figure to use a single color scheme for all methods, with the labels along the x-axis describing each method. We also are more specific about using the language of 'pairs' to point out that two boxes go with each method. For all of the pairs the darker shade depicts the results for the given method (or median) using coarse resolution data and the lighter shade shows the results for the same method using high resolution data. We added a legend to the figure to reiterate what

the light and dark shades refer to. The updated figure is shown below.



10. *Table 3 appears to be the primary result of this work. Would it help to include a percent difference along with the absolute difference?*
    We have added percent differences at the suggestion of the reviewer. The values generally support the same outcomes. There is a roughly -33% bias overall, decreasing with the first two exclusions. When broken down by time of day the morning bias (~ -42.5%) is roughly twice the afternoon bias (~ -20%).

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*

We thank the reviewer for their thorough and helpful review of our paper. 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 blue, italicized text. Any line number references are to those in the originally submitted manuscript.

This study advances the fuzzy logic methodology initially introduced by Bonin et al. (2018) for the detection of boundary-layer top by integrating data from multiple remote-sensing instrument observations including both kinematic and thermodynamic observations. The research demonstrates that this novel approach yields reasonable estimations of boundary-layer height under both daytime and nocturnal conditions. The utilization of a synergy between multiple instrument measurements is critical for enhancing and ensuring the accuracy of boundary-layer top estimations. Consequently, this study [is] scientifically significant and aligns well with the journal's scope. However, the manuscript is now well organized. Noteworthy concerns are outlined in major comments provided below. Significant revisions are needed before the manuscript be accepted for publication.

Major comments:

1. *Given the new approach expand upon the work of Bonin et al. (2018), it is essential to include BL estimations by the Bonin et al. (2018) method in the comparisons with radiosonde BL estimations, to show and validate the improvement of the new approach.*
   First, we would like to clarify that our intention is not to *improve* the method that exists in Bonin et al. (2018). We did not mean to convey that we were trying to do so, and tried to avoid language that suggested otherwise. We have revised the text noted by the reviewer to attempt to make it more clear that our intent is to use the B18 framework as a starting point for the development of this algorithm.

   Bonin et al. (2018) proposed a fuzzy logic algorithm for determining BL height(footnote) from Doppler lidar data and found promising results using data from the Indianapolis Flux Experiment (INFLUX; Davis and Coauthors, 2017). This approach is advantageous in that it combines multiple estimates of BL height, provides a measure of uncertainty for each estimate, and is adaptable to users' individual needs and cases. However, no thermodynamic information is incorporated. In a review of BL height detection approaches, Kotthaus et. al (2023) specifically described the potential benefits of instrument synergy for a variety of applications. We hypothesize that, similar to the suggestions made in Kotthaus et al. (2023), the capability and applicability of BL a height estimation method (in the present case, a fuzzy logic algorithm) could be expanded

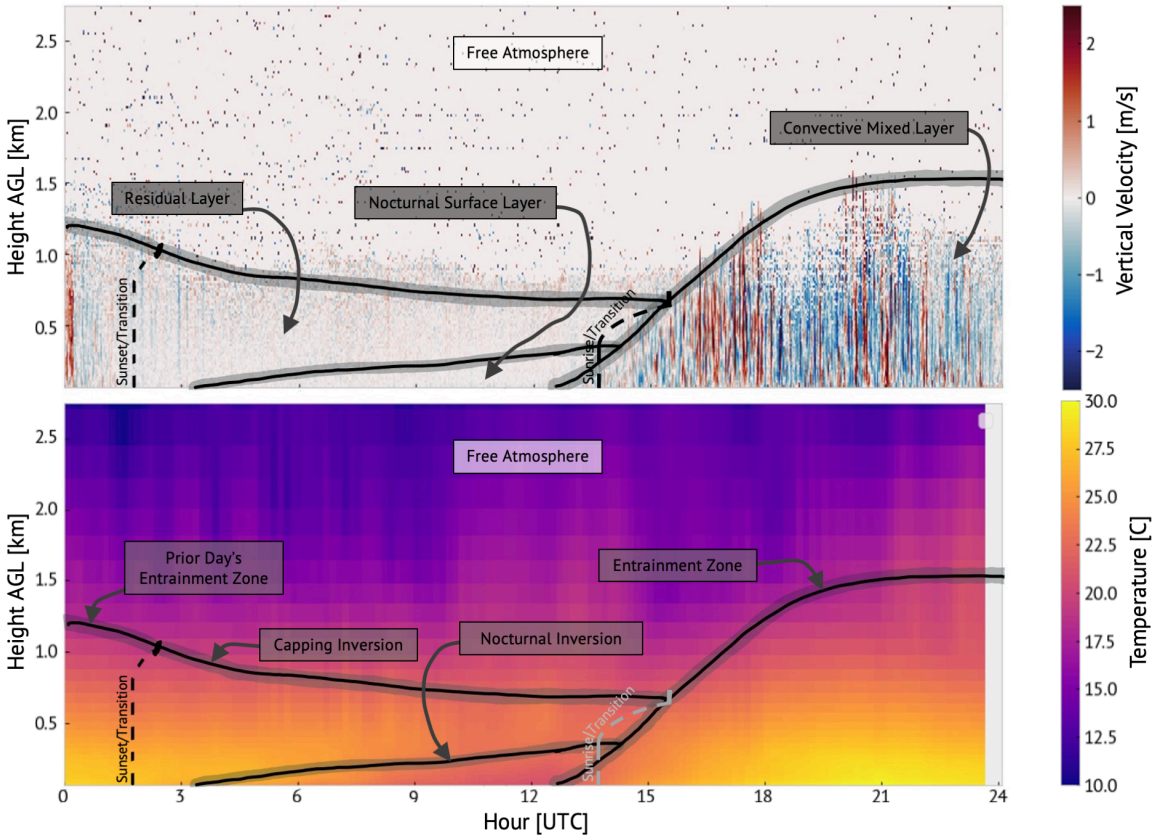
by incorporating multiple instrument datastreams.

The method presented in B18 is likely quite appropriate for the deployment in which it was developed, but that deployment isn't entirely representative of other boundary layer observation deployments. Often when deploying instruments to measure in the boundary layer, there can be multiple observation goals which can prevent or at least hinder developing a lidar scan pattern that can depict boundary-layer flow and mixing conditions in such detail (see Bonin et al. 2018; table 2). This is certainly the case for the CLAMPS platforms, which provide the data in this paper. Here we have a small subset of the lidar scan types compared to that of Bonin et al. (2018). For example, their scans include vertical stares, low elevation directional stares to the south and east, RHI scans, and PPI scans at multiple elevations and took about 20 minutes to complete one cycle. Our scan strategy is usually developed for more general observation of the mean wind with faster update times, so it includes a single PPI and a vertical stare. As such a direct comparison between our method and theirs is not possible within our dataset.

2. *Figure 1 lower panel: the Entrainment Zone is misleading in the plot. Is the entrainment zone the region between the capping inversion and free atmosphere, with values of ~ 1km?*

   We reviewed figure 1 carefully after receiving this reviewer comment. As mentioned in the caption, this figure is modeled after the well known Stull (1988) diagram of PBL diurnal evolution. In this case, we adapt that evolution diagram by overlaying it on real-world observations and pointing out features defined in the original Stull diagram. Since these are real observations, the features are not *exactly* canonical, but they illustrate the same diurnal cycle nonetheless. These terms and their placements are well established following Stull (1988). We think some of the clarity issues arose from our labeling, particularly in the thermodynamic panel. In the original version, the label for the entrainment zone, which is shown prior to the evening transition and again after the morning transition, was shared across both instances with two arrows. In the new version we split the labels to each instance. We are also explicit about the early evening hours just after 00 UTC including the prior day's entrainment zone. We also added a marker for the morning transition period to help separate the capping inversion and entrainment zone.

Finally, we made all arrow markers larger for clarity. The updated figure is shown below.



3. *Figure 2: Caption: what are the names and units of the variables in panels (a)-(j)? What's the definition of 'high-frequency' vertical velocity variance? Figure 2 comprises eight panels, but only panels 1-4 were discussed in the text. Why different color schemes are used for different panels. The plots and color schemes make it difficult to see boundary layer structures. What information should reader get from these panels?*
   It was an oversight to not include references to panels (e)-(j) in the text. They should have been included and have now been added in the description of the second generation step. We also included a more explicit definition for high frequency vertical velocity variance. The colors are different for different variables or variable types. The ranges for the color bars are standardized for comparison to each other (i.e., *u, v*) and to represent typical mid-latitude observed ranges. This multipanel figure, while dense, is intended to offer the reader the opportunity to see the data that is ingested into the algorithm to produce the 1st generation and 2nd generation aggregates and the BL height estimates used in the walk through of the algorithm.
4. *Page 5 line 154: How the sensitivity of the Haar wavelet dilation was examined and why it is low? From Sawyer and Li (2013), the dilation length is critical to find the signal peaks. References: Sawyer, V., & Li, Z. (2013). Detection, variations and intercomparison of the planetary boundary layer depth from radiosonde, lidar and infrared spectrometer. Atmospheric environment, 79, 518-528.*

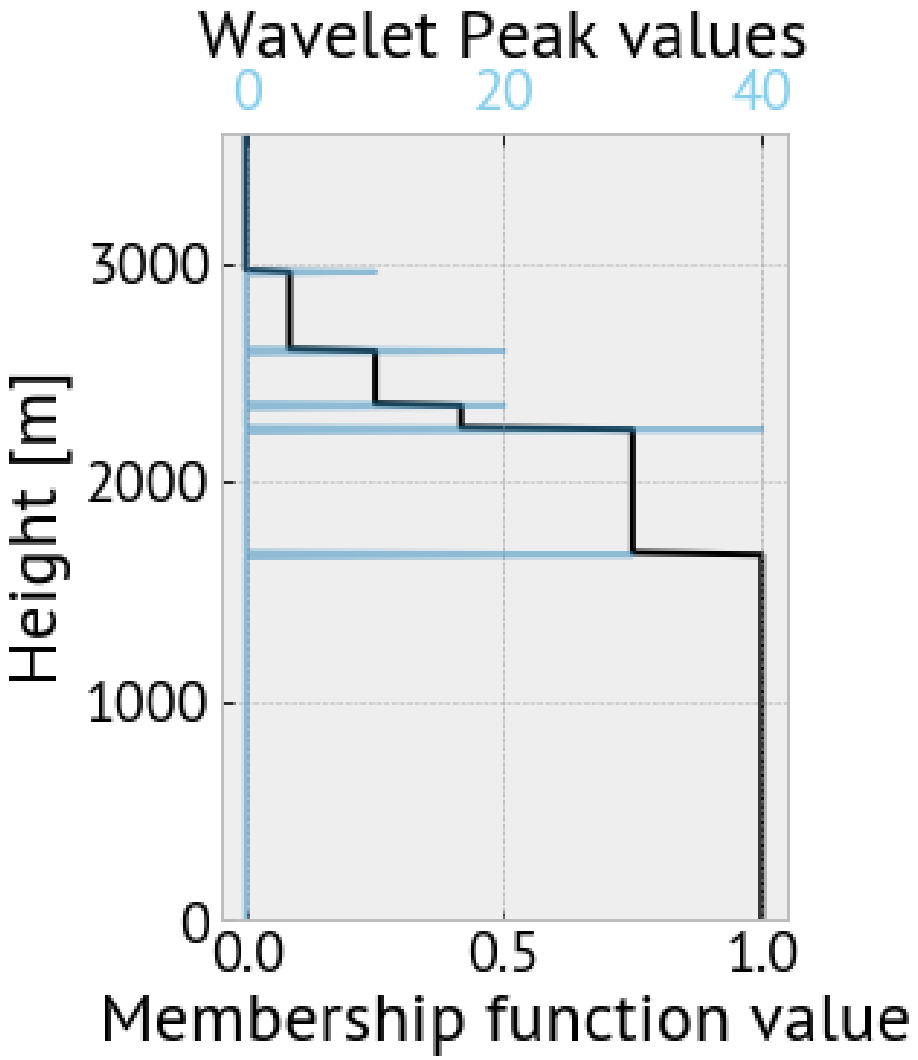
The sensitivity was examined by trialing different dilations, including the 100 m used in this application, 200 m as used in Bonin et al. 2018, and 300 m. In these trials, there were differences to the wavelet transform magnitude itself. However, applying the peak search and conditions to that wavelet transform led to results that were not sensitive to the dilation.

5. *Page 6 line 161: what physically is P? Is it the peak height or the peak magnitude?*
We revisited the expression of these conditions to make sure we had them in a clear form. We came to the conclusion that in addition to better defining P we could clarify the whole expression. The equation itself has been rewritten so the notation is improved and the middle condition is expressed more directly. We also revised the text following the expression to more clearly and accurately explain the terms and how/when they result in different values.

$$M(z) = \begin{cases} 1, & z \leq zmin \\ 1 - \frac{\int_{z_{min}}^{z} P(z)dz}{\int_{z_{min}}^{z_{max}} P(z)dz}, & z_{min} < z \leq z_{max} \\ 0, & z > z_{max} \end{cases}$$

where M is the membership function, P is the profile of peak magnitudes, and z is height above the surface. Since only select peaks are retained, P is nonzero only at the heights of retained peaks. In this notation, z_{min} and z_{max} are the heights where the lowest and highest peaks occur, respectively.

As is now written in the manuscript P is the vertical profile of peak magnitudes (shown below in blue in the figure below) and is zero everywhere except for at the height of the retained peaks. The new definition for M(z) shows that between the lowest and highest peaks (i.e., the middle condition), the membership function (shown in black below) starts at 1 and decreases at each peak proportional to that peak's magnitude relative to all peaks together. We hope this makes the membership function more clear.



6. *Page 8 line 233: Is 'high-resolution' radiosonde refer to vertical or temporal resolution? How frequent was radiosonde launched each day? Similarly, how coarse is the coarse radiosonde data, e.g., what's the resolution?*
   These paragraphs have been edited to include the requested information about resolution and timing.

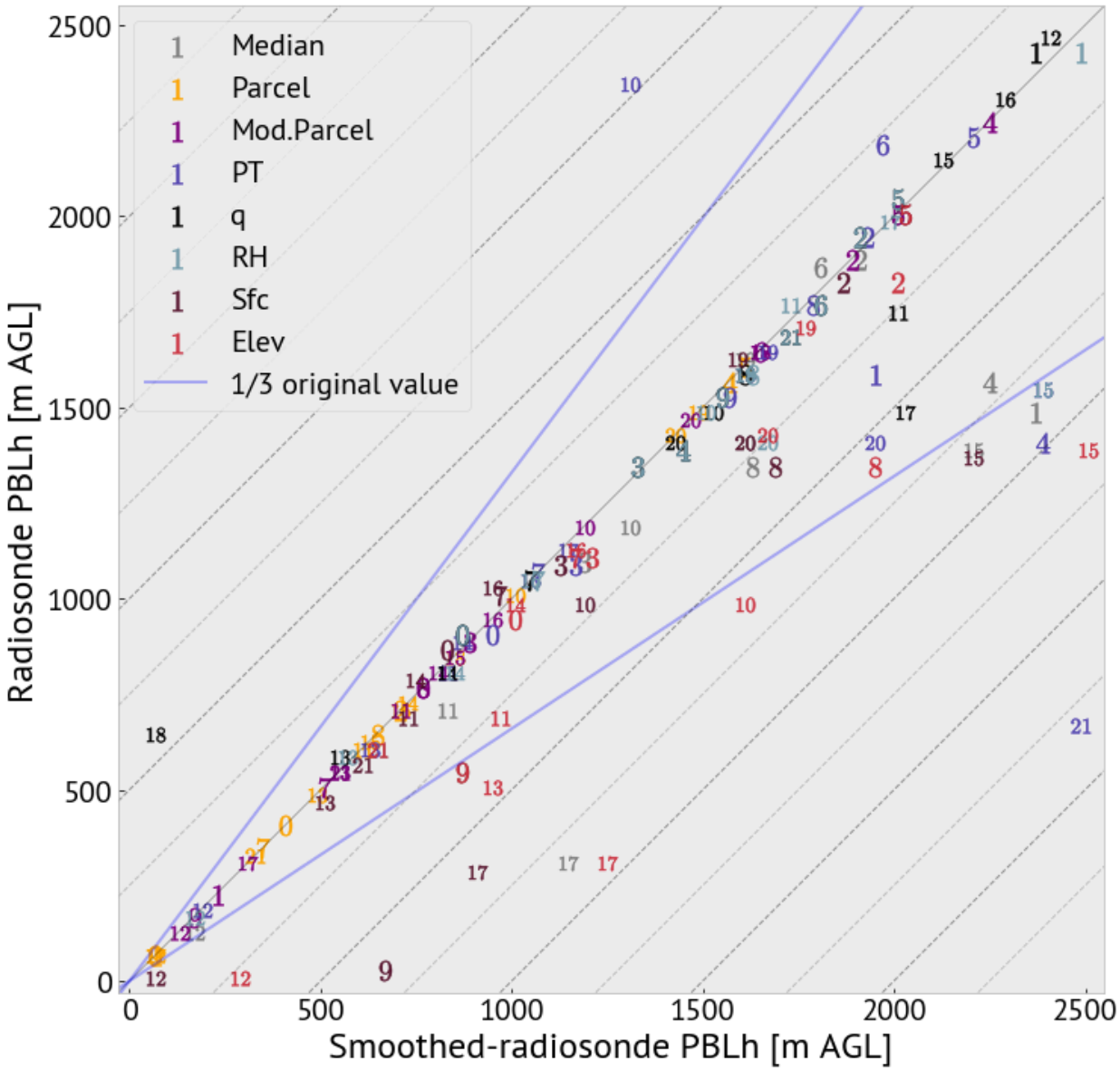
   Research radiosondes from the CHEESEHEAD campaign (Vaisala research grade radiosondes processed by the National Center for Atmospheric Research; NCAR/EOL In-Situ Sensing Facility and University of Wisconsin Space Science and Engineering Center (SSEC) 2019) and operational National Weather Service radiosondes at OUN from the NWC/RIL and PBLTops campaigns make up the high-resolution radiosonde dataset. These soundings have a mean vertical resolution of approximately 5 m in the lowest 3 km and the exact launch time recorded. Operational radiosondes are typically launched only twice per day for 0000 and 1200 UTC observations. In cases of forecast

need, special radiosondes are occasionally launched at other times, usually on a 6-hourly interval. During CHEESEHEAD, radiosondes were nominally launched once daily at 1800 UTC. During two intensive periods the launch frequency increased to four times daily. CLAMPS was only deployed during one of those periods which occurred from 23 September 2019 to 28 September 2019. To prevent erroneous…

…

At SHV during PBLTops and whenever high-resolution data were otherwise unavailable, coarse-vertical-resolution, publicly available radiosonde observations were used instead. These radiosondes are provided with data recorded at mandatory and significant pressure levels. According to Schwartz and Govett (1992) these levels were determined for a variety of reasons, including but not limited to consistency between existing database conventions at the time of alignment. The resulting vertical resolution varies with height, from O(10) m to O(100) m spacing; 300 m is the average.

7. *Figure 5: For cases when smoothed-radiosonde PBLH are deeper than radiosonde PBLHT, how to determine/make sure that the deeper values are better/more accurate?*
   There is no real measure by which we can say one radiosonde BL height estimate, smoothed or not, is more accurate. In our original approach, the authors visually examined the median and spread of the multiple methods on the suite of CHEESEHEAD soundings to verify that applying smoothing was not introducing unphysical or detrimental estimates. To examine the reviewer's question more carefully, the authors returned to this task. To identify the most impacted cases, for each CHEESEHEAD sounding available the BL heights estimated using unsmoothed data were directly compared to those estimated with the smoothed data. In any case where the difference between unsmoothed and smoothed estimates (from the same technique) was more than ⅓ of the BL height magnitude (represented on the 1-1 line), that sounding case was flagged for closer examination; 10 soundings were identified. Radiosonde BL height estimates were plotted in a similar manner as in figure 5, using color and index numbers to identify the BL height estimation method and sounding (by index number in our datafile), respectively. The blue lines show the PBL height estimate difference threshold boundaries and grey lines show intervals of 250 m.



The soundings that fall outside the boundaries are as follows (with the method resulting in the large difference listed alongside the sounding timestamp):

1: 2019-09-21 18:05:01 — Median
4: 2019-09-25 21:45:03 — PT
9: 2019-09-28 03:00:01 — Sfc, Elev
10: 2019-09-28 21:44:15 — PT, Elev
12: 2019-09-30 18:00:06 — Sfc, Elev
13: 2019-10-01 18:00:48 — Elev
15: 2019-10-03 18:17:36 — Median, Sfc, Elev, RH
17: 2019-10-05 18:01:29 — Median, Sfc, Elev
18: 2019-10-06 18:01:24 — q
21: 2019-10-10 18:00:03 — PT

These soundings were carefully examined manually to compare the unsmoothed profiles to the smoothed profiles and the resulting suite of BL height estimates from each. Most commonly, inversion based methods were noted (ten instances; multiple methods can be noted on a single sounding). Gradient based methods were noted five times, and large shifts in the median were only noted in three instances. In a number of these soundings there are saturated or near-saturated layers below ~2.5 km. In non-saturated cases, the moisture profile is usually still quite moist and characterized by many local maxima and minima. It is possible that these are real local pockets of moistening and drying in shallow layers in the mid boundary layer, but it is unlikely that such shallow layers (especially when there are many) are markers of the boundary layer top. While moisture was a recurring pattern in the soundings, the impacts on boundary-layer height detection methods were not limited to moisture variables. Similar minor, shallow pockets in temperature profiles impacted temperature gradient and inversion-based methods. In some of these cases, smoothing helps the BL height estimation methods which rely on gradients select what looks like a more "reasonable" BL height in the authors' visual assessment. In others, smoothing simply leads to these methods selecting the next most dramatic gradient available in the series of maxima and minima in a noisy profile. In saturated cases, smoothing may not lead to any change, or it may lead to some moisture based methods identifying the top of a cloud layer instead of the bottom (or vice versa). In short, some moisture profiles are difficult to use for characterization of BL height using moisture based methods. It is worth noting that this difficulty, particularly with gradient-based methods, is consistent with the results we note in our full vs coarse resolution comparisons at the end of section 3. Those results were different than the Siedel et al. 2010 findings, but consistent with the more detailed look we have described in this response. Overall, this points to the importance of using the median, which is rarely shifted by much with smoothing. When the median is shifted, it is likely due to outlier adjustment.

8. *Page 9 Line 272-286: This paragraph is confusing. Line 275 states that 'this comparison appears to support the Seidel et al. (2010) findings' that 'high-resolution data can change the estimate in statistically significant ways', while line 285 states that 'there is no indication…using high resolution data yields more accurate BL height values'. Aren't these two statements contradicting with each other?*

   The language in this paragraph was a bit confusing, we agree. We made some specific edits to the language around the comparison of Fig 6 and 7 and removed the word error in reference to profile resolution differences to hopefully lead the reader more directly through the subtle differences between what our data and the Seidel et al. (2010) data show. We do agree with Seidel et al (2010)'s finding that high-resolution data can change the estimate in statistically significant ways, but without any real measure of 'truth' we cannot say what is really more accurate. Using the word error when discussing profile resolution differences made that part particularly unclear and we are glad the reviewer

brought this forward for us to fix.

Seidel et al. (2010) found that while coarse data can be sufficient to detect BL height, the use of high-resolution data can change the estimate in statistically significant ways. A comparison of the medians of radiosonde-derived BL heights in cases where high-resolution and coarse data were available simultaneously is shown in Figure 6. This comparison initially appears to support the Seidel et al. (2010) findings. While many of the afternoon and early evening soundings are close to the one-to-one line, several of the morning soundings fall to the right and below the one-to-one line, suggesting the high-resolution based estimates are lower than the coarse estimates. The 'error' (more accurately, uncertainty) bars, which represent the spread (i.e., interquartile range) of BL heights detected by the seven methods, also appear to cover a wider range for the high-resolution soundings. However, if we further examine the comparison by looking at individual methods separately, we find slightly different results than those in Seidel et al. (2010). Their results showed methods that were computed from the surface up, such as the parcel- and inversion- based methods, were most sensitive to profile resolution, while Fig 7 suggests…

9. *Exclusion of cases were done manually for the four criteria. These criteria are subjective and not well defined. For example, how to determine cases that BLH height was ambiguous (criteria 1)? How to determine at what height CLAMPS observation not reliable to detect PBLH (criteria 2)? When radiosonde and CLAMPS methods identified different parts of the transition region, which one should be trusted (criteria 3)? How to define 'complex/non-canonical BL structures (criteria 4)? Practically, how do users know when to trust or not trust CLAMPS PBLH estimations? Suppose CLAMPS will be run automatically, how to qc label CLAMPS PBLH estimations from the 'bad atmospheric conditions?*

   We describe in the text that each matched pair was manually interrogated for inclusion or exclusion in a similar approach as Banghoff et al. (2018). We do not intend to suggest that these criteria are any sort of objective measures. As we describe in the text our initial comparisons (shown in Figure 8) are made without consideration of the atmospheric conditions at the time of the observations. In this set of comparisons with successive exclusions, we want only to better understand the behavior of the algorithm by eliminating potential 'outside' influences not related to the algorithm itself. The validation of new proposed algorithms using subjective human assessment is a common practice when a known truth is not available. For example, Hubbert et al. (2018) used human experts in convection to evaluate a proposed hydrometeor classification algorithm. Similar to this study, Bianco et al. (2008) used two human experts to evaluate a fuzzy-logic approach for estimating PBL depth. As such, there are not objective criteria to cite for each exclusion condition. For example, if the authors with experience identifying PBL height from radiosondes could not themselves confidently identify a

PBL height for Criterias 1 and 4, it was deemed ambiguous or complex and not a good case for evaluating algorithm performance. For Criteria 2, if the PBL depth increased steadily and then suddenly leveled off despite continued mixing coinciding with a threshold value of scatterers, it suggests a limitation in how far the lidar signal could reach. Finally, for Criteria 3, it is not our intent to say which estimate is more trustworthy or representative. In fact, the fact that we cannot is the reason we elect to remove those cases from the database in the first place. We aimed to be transparent in this successive elimination by including both the results with all data points and those with low-confidence cases (as judged by human experts) excluded, and by having each author perform an independent assessment of each of the four criteria. Despite these exclusions, we still retained nearly two-thirds of the cases originally included. To make the criteria themselves more clear, we have amended their descriptions some, providing a few examples in some cases. We have set them inline in a numbered list in the manuscript.

In order to better understand the potential causes of discrepancies between the radiosonde and CLAMPS-based BL height estimates and ensure a robust comparison, each matched pair was manually interrogated for inclusion or exclusion in a similar approach as Banghoff et al. (2018). In our case, criteria for exclusion were developed to describe instances in which discrepancies between CLAMPS and radiosonde BL heights may not necessarily reflect the intrinsic performance of the proposed algorithm. These criteria were as follows:
1) ambiguous cases where the BL height was unable to be confidently determined from the radiosonde data as multiple methods failed to identify BL height at all or identified levels that upon visual inspection were likely related to other structures (e.g., residual layer, clouds, etc.);
2) cases where CLAMPS observations were not able to be collected over a deep enough layer to capture the likely full depth of the BL (primarily DL observations due to a lack of scatterers);
3) cases where the BL top (e.g., entrainment layer or capping inversion) was deep and the radiosonde and CLAMPS methods identified different parts of this transition region;
4) cases with complex/non-canonical BL structures (e.g., multiple inversions, low-level cloud layers, etc.).
To identify cases for exclusion, both authors independently evaluated each time-matched CLAMPS and radiosonde profile to determine whether one or more of these criteria were met without respect to how their BL height estimates compared, then discussed and reconciled any differences. These criteria were used to successively exclude data pairs and get a better sense of algorithm performance. As each criteria was applied successively, more pairs were removed from the comparison dataset as summarized in Table 3.

Given this paper is introducing an algorithm and not a dataset from CLAMPS, the latter part of the comment is outside the scope of the present work. Any retrieval method or algorithm is only as good as the data it is provided. It is the responsibility of the user to know what they provide to an algorithm. Note that many of the standard accepted radiosonde-based BL height estimation methods (e.g., finding where the surface potential temperature intersects the profile) also rely on assumptions of a canonical BL (i.e., those with surface-up mixing processes) and could similarly fail in non-canonical BLs. In the case of providing it as a dataset, we already provide one quasi measure of confidence in the estimate in the form of the standard deviation over the moving windows, which makes the assumption that very variable BL height estimates are less confident or more uncertain. Regarding QC, that step should be done to data likely before it is provided to the algorithm. Theoretically flags could be passed through the algorithm from the datasets it ingests in a real-time setting, but this use case has not been developed at this time. The criteria discussed here would not be applied as QC criteria as they are not meant to be QC criteria. They are meant to be subjectively applied elimination criteria to provide us a best-case dataset for comparison, eliminating outside impacts on potential algorithm performance.

Bianco, L., J. M. Wilczak, and A. B. White, 2008: Convective boundary layer depth estimation from wind profilers: Statistical comparison between an automated algorithm and expert estimations. *J. Oceanic Atmos. Tech.,* **25**, 1397-1413. doi:10.1175/2008JTECHA981.1.

Hubbert, J. C., J. W. Wilson, T. M. Weckwerth, S. M. Ellis, M. Dixon, and E. Loew, 2018: S-Pol's Polarimetric Data Reveal Detailed Storm Features (And Insect Behavior). *Bull. Amer. Met. Soc.,* **99**, 2045-2060. doi:10.1175/BAMS-D-17-0317.1.

10. *Are all the comparisons for clear-sky conditions, e.g., no clouds and precipitation?*
All comparisons are in precipitation free periods (over the profilers). The algorithm requires Doppler lidar data and thermodynamic profiler data to provide a boundary layer height estimate. In our case, we are using the AERI instrument for thermodynamic profiles. This instrument does not collect data during precipitation. It detects precipitation and closes a protective hatch to keep its detector dry and safe. Cloudy conditions were not intentionally excluded. It is true that some clouds could have an impact on the algorithm (e.g., premature extinction of lidar signal, radiative impacts on the AERI measurement). However, there are reasons why we don't exclude clouds. First, deciding the threshold for 'what is a cloud' based on the types of observations we have available to us is non-trivial. Second, at times the formation of clouds and their positions can be related to the boundary layer depth itself. Without an independent measurement of cloud conditions, we did not feel confident in creating an objective exclusion criteria.

11. *Figure 10 and Figure 11 could be consolidated into a single figure for better clarity and coherence.*
    These panels have been combined into a single figure.
12. *Figure 11: Even after all exclusions, CLAMPS PBLH is generally still lower than radiosonde PBLH, any speculations of the causes?*
    In response to this comment and another reviewer comments, revisions were made to the Summary and Outlook section, particularly to the fourth paragraph, that address this comment. It is now more explicit about conclusions and outcomes. We discuss that while there is agreement between radiosondes and the algorithm, there is a pattern of underestimation. We speculate about the reasons behind this including the limitations of remote sensors (capabilities and resolution). We also now more explicitly discuss that exploring errors through comparison with respect to time would require more data, specifically more data around the diurnal cycle to create discrete samples and or develop normalization techniques.

Minor comments:

1. *Page 1 line 19: Given numerous past studies of atmospheric boundary layer and various of BL height estimation methods, it is unrealistic to claim that BL 'yet is also one of the least observed portions of the atmosphere'.*
   These authors argue that it is realistic. Many of these observation studies are just that: studies. In general the boundary-layer remains routinely under-observed due to a lack of widely available affordable, maintainable, deployable solutions capable of capturing the atmospheric conditions between the surface and hundreds to a few thousand meters above it, where remote systems such as satellites and radars can be more successful at monitoring. This is even more of an issue over Earth's oceans, though this is not discussed in the present study. See Bell et al, 2020, NASEM 2018a, NASEM 2018b, NRC 2010, NRC 2009 (all cited within this manuscript) for more discussion. To make the distinction more clear we have added the word routinely to the paper in the same fashion it is used in this response.
2. *Page 2 line 49: it is not clear what does the 'positive impacts' mean.*
   We rephrased this sentence for clarity.
   Recently Tangborn et al. (2021) found the accuracy of the temperature and wind fields in the simulated afternoon convective BL to be improved compared to radiosonde observations when assimilating observations of BL height.
3. *Page 3 line 72: what are buoyancy processes within nocturnal stable boundary layers?*
   Buoyant processes are not limited to the growth of buoyancy. Specifically in the scope of the algorithm we are discussing, the first generation step relies heavily on measures of mixing, or in other words observations showing that mixing is ongoing. In this case, that means mostly mechanical mixing observable by Doppler lidar. Of course, the buoyant processes related to daytime heating can lead to that mixing, but those buoyant processes

are not directly observed by the Doppler lidar. Buoyant processes are important during the night when the daytime boundary layer decays and the mechanically driven processes can become less dominant, which in the scope of this algorithm may make the Doppler lidar observations of mixing less informative. As such, information about the presence of the daytime capping inversion, thermodynamic structure in the residual layer, and the evolution of any surface inversion is critical at night. In non-canonical cases, the classical "Stull" evolution is often not representative of boundary-layer evolution including mechanical or buoyant processes. Including both is useful during transitions and during non-canonical nocturnal boundary-layer cases.

4. *Page 3 line 77: Repeating words 'such as the'.*
   We have made this correction.
5. *Page 5 line 129: what is a 'complete failure to detect a buoyancy-driven BL'? is the 'buoyancy-driven BL' similar as the buoyancy processes within nocturnal stable boundary layers? Under what conditions and how often does the 'complete failure' occur?*
   A failure to detect a buoyancy-driven BL would be the event in which there are no mechanical processes for the algorithm to identify so the algorithm fails, incorrectly, to identify the BL height while buoyant processes dominate. This then carries forward, as this initial estimate is used to constrain which peaks are retained for refining the estimate using the thermodynamic information. We did not examine the frequency, but this could occur anytime there is an absence of measurable mechanical mixing for the algorithm to detect as a measure of mixing ongoing. This is certainly possible at night. We modified the language in the paper to make this a bit clearer.

   Buoyant processes also play a role in BL development, and are often a dominant process at night when stable boundary layers are more common. In later steps the BL height estimate from the first-generation step is used as a type of constraint on the second-generation step. Thus if there is a complete failure to detect a BL height (presumably in situations in which the BL depth is driven by buoyant instead of mechanical processes) in the first-generation step, the second-generation step is unable to recover.
6. *Page 10 line 303-304: Description of Fig.8 is repeating as in the line 294-295.*
   This language was intended to draw a comparison between the two figures, not be a repetition of information. We think this helps the reader along in interpreting the meaning of the multiple panels on figure 9, since they are the same analysis as figure 8, just with different data subsets. However, after fulfilling some requests from other reviewer comments this particular section of writing has been reorganized. In the revision there is less repetition naturally, but we also took care to make the comparison between the figures without repeating the description of either.