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 continued 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 a darker blue, italicized text. At times we include additional text from the manuscript in italicized grey text [the marker (TR) refers to tracked revised manuscript].

The revised manuscript improved a lot. However, there are still several noteworthy concerns as outlined in comments below. All line and page numbers are based on the tracked revised manuscript. Major comments:

Major comments:

1. *How do the authors know the new algorithm, i.e., the second-generation step, improves boundary layer height estimation, compared with the first-generation step? It should be noted that including low-quality data or including non-critical data could jeopardize BL estimates. For example, in Fig. 3, the second-generation algorithm estimated BL height is almost the same as the first-generation ones during the nighttime, only show some differences during the daytime. However, from the description between line 135-139, it indicates that the first-generation estimation is expected to work well during the daytime, but has issues during the nighttime, which is contradictory to what Fig. 4 shows. The new added figure, Fig. 10, shows the same that the second-generation BL height is almost as the first-generation BL height during the nighttime. Fig. 10 b) (top right) even shows that the first-generation BL heights are closer to sonde-estimated BL height at ~18 UTC and 22 UTC. For Fig. 8, it might be better to include the comparison between the first generation BL heights and sonde-estimated BL heights.*

   In order to best respond to this comment, we will break down the concepts addressed in it into two main parts. First there is the general question: How do we know that the new algorithm (i.e., the second generation step) improves boundary-layer height estimation compared with the first generation step? In some ways we could consider some of the reviewer's question (specifically about including low-quality data or non-critical data) to imply that if the second generation step could make the estimate worse, we should not embark on it, and then it follows that only lidar measured vertical velocity matters to this problem and adding any other information cannot help the process. Given the premise of our paper and the content of our literature review, we do not agree with this framing and instead we align with the notion that muli-instrument approaches offer the opportunity to add benefits to one another especially when some of the included instruments experience observation weaknesses from time-to-time. In this framing, we could say that lidar-only

methods of boundary-layer detection may be limited---for example during the nighttime or very stable periods when mechanically-driven mixing processes are unlikely to present much if any target by which the lidar could detect boundary-layer height, or during periods of fog or low cloud when the lidar signal is partially or completely extinguished by water droplets. Including additional instrumentation that can gather other information about the structure of the boundary layer which can be used to estimate the boundary-layer height is obviously beneficial. In a fuzzy-logic approach, which is what we use, it is important to understand the data sources/types (and the quality assurance measures that the user has hopefully applied or understands have been applied to them) that contribute to the result and assign membership functions, weights, and other parameters or controls accordingly to control for any 'issues' with 'low-quality' or 'non-critical' data. Those functions, weights, etc. should be appropriately assigned to allow the most impactful or direct datasets to contribute more. This is what our paper describes. Still, to make the point abundantly clear in the paper, we have added language to the end of the fuzzy-logic algorithm section, immediately prior to the data section.

In the case shown in Fig. 2 (and in the CLAMPS datasets described next in Section 3), care was taken to understand how the instruments operated, any quality assurance applied to the data during their initial collection and production, and to determine if any additional quality assurance was needed before providing the data to the fuzzy logic algorithm. The CLAMPS platform happens to be a highly automated system which applies a high level of quality assurance automatically, reducing (but not removing) user needs to further quality assure CLAMPS datasets. This may not always be the case, and users must understand the potential impacts data quality may or may not have when providing inputs to any algorithm. In the case of our fuzzy logic approach, the physically-based definition of membership functions limits in the first generation step and the BL height-range constraint in the second generation step both provide some safeguard against entirely artificial or problematic data, but ultimately users must consider the data quality provided to the algorithm when weighing the quality of BL height estimates.

The second part of this comment includes specific discussion of figures. In the comment, the reviewer mentions figure 4, which for us is the flowchart figure. We assume the reviewer meant figure 3 and are responding as if that is the case. Here the reviewer notes that the first and second generation estimates of boundary-layer height are similar/the same overnight and only different during the day. This is framed as being inconsistent with text on, e.g., lines 135-139 where we state that first-generation estimation is expected to work well during the daytime, but has issues during the nighttime. For completeness the text from those lines are reproduced here.
(TR) *Lines 135-139: "Since the first-generation estimate depends only on measures of*

mixing, it relies only on mechanically induced turbulence and mixing to determine BL height. 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.

Here we explain that the first generation step, as it exists at this point in the paper (and in line with the B18 approach) relies only on mechanically induced processes. We have not introduced any change to the B18 algorithm other than the statement "At this stage the algorithm design deviates from the design of B18." We bring up the potential problem of only relying on lidar data for boundary layer identification through the full diurnal cycle (e.g., buoyant processes) and how such a problem may have effects later in the algorithm (because the 1st gen estimate serves as a constraint in the 2nd gen). We then expand on this explanation, which also explains the behavior seen in the figure.

(*TR) Lines 139-150 Thus if there is a complete failure to detect a buoyancy-driven BL height in the 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. In order to capitalize on the availability of thermodynamic profiles and extend the capability 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). as an input variable (N.B.: this is not a measure of mixing). To limit the effect of this additional input beyond periods where buoyancy is more likely to dominate mechanical generation of mixing (e.g., nocturnal stable periods), a time-dependent weighting function is defined based on the local sunrise and sunset time (Fig 2d). This weighting function allows the inversion height to have an effect on the algorithm during the overnight hours with sloped increasing and decreasing weights during the evening and morning transitions, respectively. The membership function in this case is step wise, and thus not authentically a fuzzified field. All levels above and below the inversion height are assigned membership values of zero and one, respectively."*

In other words, inversion height is added as an input during the first generation step. This is the deviation from the B18 approach. Without it, for the case in which no boundary-layer height was detected in the first-generation, the second-generation would have no basis from which to start (since it is required to work within a constrained distance of the first generation estimate). We know that the inversion height is really most important when mechanically-driven mixing processes are minimized/buoyant processes are dominant and this is the time the lidar detection methods are most likely to fail. We expect this to most often occur at night, so we assign a weighting function that only allows this first generation inversion height input to have an impact during the nighttime

hours. During the second generation step, thermodynamic profiles are used at all times of the day. Given that overnight first generation estimates are often dominated by inversion height (i.e., based on thermodynamic profiles), it makes sense that adding more information from the thermodynamic profilers in particular at night would not lead to much change to the estimate in the second generation during that period. This same reasoning can be applied in the figure 10 panel the reviewer cites. In regard to the upper right hand panel of figure 10: this is a single case that we added on reviewer request as one of a set of examples. On any one day at any one observation time, we could choose to show a 'better' or 'worse' comparison to either generation estimate. We also point out that for this particular 22Z sounding there is a large degree of spread among sounding estimation methods. 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 first generation estimate 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.

2. *The caption of Fig. 2 is still not clear. The caption should be able to roughly explain the figure so that readers do not need to read carefully in the text to understand the figure. For example, what are first- and second- generation variables? Those need to be clear in the caption. Fig. 2c and line 144, how is the temperature inversion height derived from the first-generation step as it only calculates the vertical wind variance? Fig. 2d, is it the inversion height or weighting function? 'inversion weight' is not defined in the text.*
To completely define what first and second generation steps are would be extensive for a figure caption that is only referring to the variables provided to the algorithm, regardless of the order of their use in the algorithm. We have never stated that inversion height is derived from the first generation step; it is used in the fuzzy logic algorithm as part of the first generation step. Please refer to the text in lines 139-150. In panel d, the figure shows the weight (value of the weighting function with respect to time) that the inversion height estimates gets in the fuzzy logic algorithm. To be more specific, some minor changes have been made to the figure label. Bolded text shows new additions.

CLAMPS1 observations (collected 25 June 2020 in Norman, Oklahoma) used in the fuzzy logic algorithm. ***Variables in*** *panels (a)–(d) are* ***used as*** *first generation variables, while* ***variables in*** *panels (e)–(j) are* ***used as*** *second generation variables. Each observed field is labeled with units* ***(if applicable) on its respective panel color bar.*** *Note that panel (a1) is a subset time window of panel (a) with a more narrow colorbar to highlight overnight vertical velocity variance values. Panel (b) is similar to panel (a) but considers only the high-frequency fluctuations in vertical velocity, as described in the text. Panels (c) and*

(d) have local sunrise and sunset times marked for reference. Panel (d) is not an observed field, ***but shows the magnitude of the*** *weighting* ***function*** *applied to inversion height, shown in panel (c),* ***which itself is computed from temperature profile data shown in (i)****.*

3. *(TR) Line 139-141: the logic is confusing here. It is indicated that a complete failure to detect BL height in the first-generation step occurs when the 'buoyant processes' dominates. If the first-generation step fails, the second-generation step is 'unable to recover'. However, according to the manuscript, the main advancement of the second-generation step is to improve 'buoyant processes' dominated (nighttime) BL height estimates. The question is that if the first-generation step failed and the second-generation is 'unable to recover', how does the second-generation step improve BL height estimates?*

   As described in this part of the paper (and now also in our reply to the reviewer's comment 1), the first generation step in the overnight period includes inversion height, which is a thermodynamically derived property. The logic the reviewer describes as confusing is actually stating that it is important for us to include this capability and not only depend on mechanical processes overnight. Without including the inversion height, we would only allow the first generation estimate to be defined by mechanical processes detected by the Doppler lidar. At night in particular this is limiting. In the case no boundary-layer height was detected in the first-generation, the second-generation would have no basis from which to start (since it is required to work within a constrained distance of the first generation estimate). We know that the inversion height is really most important when mechanically-driven mixing processes are minimized/buoyant processes are dominant and this is the time the lidar detection methods are most likely to fail. We expect this to most often occur at night, so we assign a weighting function that only allows this first generation inversion height input to have an impact during the nighttime hours. During the second generation step, thermodynamic profiles are used at all times of the day, allowing buoyant processes (and other indicators that mixing has occurred, including well mixed wind profiles) to factor into boundary layer height estimation, within a constrained height range of the first generation estimate.

4. *The terminologies 'buoyant processes' and 'mechanical processes' are confusing. What exactly are 'buoyant processes' and what exactly are 'mechanical processes' in BL? Any references using these terminologies? In BL, we often use 'shear- or buoyant- driven' turbulence.*

   In our view, the terms buoyant processes and mechanical processes are more encompassing. In other words, "processes" in this case would include (but not be limited to) turbulence. It would also include the effects of turbulence on atmospheric properties and structures that are measurable by our systems, which is the motivation behind our choice of language versus the perhaps more common "-driven" turbulence language often

found in theoretical, analytical, and numerical studies. We don't feel as comfortable using this terminology or feel it is quite fitting in our application as we are not truly measuring the turbulence, driven by either mechanical mixing or buoyancy (or its decay). We are able to detect properties or fields used as measures of turbulent processes like vertical velocity variance, which we use to infer the presence of turbulence or that turbulence is occurring, or to find indicators like well mixed layers, which we use to tell us turbulent mixing has likely occurred. To be consistent with what we can detect, we use the term processes -- which again could include turbulence and the larger scale mixing, warming, dying, cooling, moistening processes (which are more likely to be detectable by our sensors) that may occur as a result of turbulence. *To come closer to the terminology the reviewer prefers and finds more common, we have added the word 'driven' where applicable throughout the paper, but retained processes in most cases.*

5. *The manuscript states that 'the Harr wavelet transform is not sensitive to the dilation', which contradict with previous studies (e.g., Sawyer and Li, 2013). Why physically 'the Harr wavelet transform is not sensitive to the dilation' in this case? If it is not sensitive to the dilation length, why choose to use 100 m instead of, for example, 50 m or 300 m?*
   In a previous reply to the reviewer we discussed our findings regarding differences that we did find. There were some differences in the wavelet transform magnitude itself between dilation trials. However, in our application we subsequently apply search functions and conditions to the wavelet transform to identify the most prominent peaks. The outcomes of that search step—i.e., the identified peaks—produced results that were not sensitive to the dilation. Regarding the choice of 100m: as described in the manuscript dilation of 100m, 200m (as in B2018) and 300m were examined. Seeing no sensitivity to the outcome of the application, we saw no reason to increase (or decrease) the dilation size.

6. *The author's reply to my comment 6: 'The resulting vertical resolution varies with height, from O(10) m to O(100) m spacing; 300 m is the average.' It is not clear what does 'O(10) m' and 'O (100) m' mean. How come the average is 300 m?*
   The notation used in the paper is a style often called 'Big-O Notation,' where the O stands for 'Order of," meaning the whatever is being discussed can be expected to have magnitude of the order x, which is referenced in the O(x) usually with units; this can be a variety of scale types for a variety of use cases. Calling this style 'Big-O' notation is a naming convention most commonly referenced in data science and algorithm development fields, but it is seen in many applications outside of these communities. *We expected this notation style to be ubiquitous, but finding this may not be the case based on this reviewer's comment, we can simply extend the language to use the full 'order of' language.*

The dataset discussed here is the publicly available radiosonde data. To be clear, this is not a dataset we generated from our own observations. These data are collected and archived based on defined data requirements and structures as described in Schwartz and Govett (1992), which is referenced directly in the text. This reference and related summary in our text mentions that there were several reasons for the archive of data ending up with the format it has today when at one point in the past, multiple data sources were merged to create a unified archive. The average is 300 m because that is the average computed from this dataset which has variable vertical spacing, as described in the manuscript. In other words, 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.

7. *Line 315: From Fig. 6 and Fig. 7, it is not sufficient get the assertion that 'there is no indication from this analysis that using high-resolution data necessarily yields more accurate BL height values'. 1) Fig. 6 shows that the differences between high- and low resolution mainly occurs at small BL height values., e.g., during time. While Fig.7 shows the absolute BL height ranges for all cases. Small BL heights generally have smaller absolute BL height estimation ranges (across different methods), which large BL heights generally have large absolute BL height estimation ranges. Therefore, Fig. 7 is biased by the large BL heights. 2) there is a clear trend that High-res BL height estimations are much smaller than coarse-res BL height estimations for majority of the methods, as show in Fig.7. Therefore, it is hard to believe that 'coarse-resolution data' has neglected impacts on BL estimations.*

   We find there is some potential confusion between the concepts referenced in this comment from the reviewer and what we include in our text and analysis. In this study, we have access to two sets of radiosonde data. One is higher resolution than the other, so for simplicity and clarity, that is the language we use throughout the paper. We do not intend to assign any specific connotation to a dataset by using these relative (high/coarse) terms. We must determine how to use the datasets, what data to use, etc. These choices are also important to consider in the context of comparing results from prior literature to our work (and vice versa in future works). How may the resolution of radiosonde data impact the analysis we do and what does that mean for the results and subsequent comparisons to other works? As we state in the paper, this is our motivation for these comparisons. As the reviewer points out, neither figure 6 or 7 in isolation is sufficient to provide all information relevant to the comparison between high and coarse resolution data. The reviewer states that 'it is not sufficient [to assert] that "there is no indication from this analysis that using high-resolution data necessarily yields more accurate BL height values"'. We disagree with this stance. We never state that there is no difference

between coarse and high-resolution sounding data or the BL depth estimates obtained from them; Figure 7 clearly shows otherwise for individual BL depth estimation methods. We do state that it is not obvious that the high-resolution estimates, being different from the estimates from coarse-resolution data, are intrinsically more correct, especially since small gradients can force detection events that appear unrelated to the BL-top. The manuscript also includes more information than what the reviewer referenced. This text is included below and the additional comments we make in the paper regarding this topic are bolded, including even a direct statement about our specific intention -- which nowhere states that coarse resolution data has neglected impacts on BL estimations. ***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, 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.*

8. *For the answers to my comment #9: These conditions are also the major reason for many other BL estimates fail too. How does the first-generation BL estimates compare with sonde BL estimates if these conditions were excluded? The core question is still that: how does the author prove that including other data/measurements improves the BL estimation?*
   Generally, we refer back to the first half of our response to comment 1 for a response to the question regarding the improvement of BL estimation derived from adding measurements. In the context of this comment, exclusion criteria are not linked to the first/second generation estimation steps. The exclusions were developed and applied based on conditions that were difficult for the instrumentation platforms considered to observe---including the reference radiosondes, thermodynamic instruments, *or* Doppler lidar. Any excluded condition would not represent the 'best case' comparison of a fully observed boundary-layer by a reference radiosonde and by *all* CLAMPS instruments, so it would be excluded from that set of comparisons.

9. *For the answers to my comment #10: 'deciding the threshold for 'what is a cloud' … is non-trivial'. Can't the DL backscatter intensity be used for cloud detection?*
   Yes, the lidar can detect cloud bases, especially directly overhead. However, this would limit the definition of 'cloud' to optically-opaque clouds in the visible (technically near-infrared) portion of the EM spectrum. We know that cloud-like/near-cloud features

that are not quite opaque to the human eye are optically thick or even opaque in other portions of the EM spectrum and have important impacts on infrared and microwave radiation (thus potentially making impacts on the passive radiative sensor observations; we do include a ‘cloud’ flag from the lidar-based estimation in the thermodynamic retrieval, simply to aid the retrieval’s optimal estimation in ‘knowing’ a cloud is there, but this would not be a robust criterion). With that knowledge, deciding when to or when not to classify periods as cloudy is non-trivial. It is also unclear how to handle mixed scenes when classifying cloud vs no-cloud in a binary way for this sort of exclusion task.