Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., author comment AC2
https://doi.org/10.5194/amt-2021-383-AC2, 2022

Reply on RC2

Gina Jozef et al.

Author comment on "Testing the efficacy of atmospheric boundary layer height detection algorithms using uncrewed aircraft system data from MOSAiC" by Gina Jozef et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2021-383-AC2, 2022

The authors thank anonymous referee 2 for taking the time to review our manuscript and for their helpful comments, which have improved the manuscript. Each referee comment is given below in ***bold italics*** followed by our response to the comment. The line numbers provided in our responses refer to line numbers in the revised manuscript.

The manuscript "Testing the efficacy of atmospheric boundary layer height detection algorithms using uncrewed aircraft system data from MOSAiC", by Gina Jozef, John Cassano, Sandro Dahlke, and Gijs de Boer, evaluates different published methods for the determination of the ABL height for a unique data set sampled by an uncrewed aircraft system over the sea ice of the Arctic Ocean. These "objective" methods are verified against a "subjective" or visual method to evaluate the height of the ABL, and the robustness of this approach is shown by applying it also to radiosonde profiles sampled in close spatiotemporal proximity. The manuscript is well within the scope of AMT but requires major revisions before it can be accepted for publication.

Here my major comments:

The information given in the introduction on the concept of an atmospheric boundary layer height is very superficial. Here the authors should expand and clarify why this concept is important, for which applications it is used, and also in which way its definition is debatable. In my eyes, it is a diagnostic parameter used to quantify the altitude up to which the direct surface-atmosphere interaction can be considered relevant. It is some sort of simplification, which is helpful for many applications, but it is not a physical property of the atmosphere, e.g. it may not be continuous for example during regime transitions. A more critical reflection on this should be included. It should also be clear that either method for identifying the ABL height is only providing an estimate (one could even claim that there is no true ABL height only different methods to estimate or diagnose it).

Thank you for your comments. The authors have added a paragraph to the introduction which explains in more detail why ABL height identification, especially in the Arctic, is important (lines 74-84). It has also been clarified that ABL height is an approximation, which is why the work presented in the paper – to find the best approximation – is important and useful (lines 82-84). We have also added some text discussing why the definition of ABL height is debatable (lines 94-95 and 113-114).

This also has implications for your line of argumentation, which seems to be based on the assumption that your "subjective method" is giving the true ABL height. I agree that a visual evaluation of the ABL height by an expert may be generally better than any objective method, but also the expert can be wrong, e.g. due to misleading observations (DH2 and RS provide only a "quasi-snapshot").

The authors now clarify throughout the text that the 'subjective' ABL height is not necessarily the 'true' ABL height, but rather, the best estimate of ABL height given the available data (lines 113-114, lines 220-224).

More details on the calculation of the Rib number and the determination of underlying parameters (gradients) are required. In particular, I lack details on the determination of the wind speed from the helical flight patterns. There is a reference to unpublished work, but some important details should be mentioned here. I assume you use the 10Hz 3D wind data? How do potential time lags and/or inaccuracies in IMU or GNSS data influence instantaneous wind measurements?

The authors have added a section to the paper before the section on determining stability regime called 'Preparing the DataHawk2 data for analysis' (beginning on line 187), which gives more detail on the calculation of Rib and gradients. For the winds specifically, a more thorough description of the method used to diagnose winds from the UAS data is currently in review in de Boer et al. (submitted 2022) and we cite this manuscript (line 155), although a brief description of the wind estimation is now provided in the revised manuscript (lines 148-152). A summary of the wind processing can also be found in the metadata for the DH2 data used in this study, which is now cited (line 154). In addition, another paper describing the DataHawk2 platform and use of its data (this paper is more general, not specific to MOSAiC) will imminently be submitted for publication to AMT by Hamilton et al, so the authors now include the Hamilton et al. citation as well (line 155). Though it is unfortunate that neither of these papers are published yet, the authors are hopeful they may be at preprint stage once the final version of this manuscript is submitted.

I also noted that the DH2 temperature profiles start at lower altitudes than the corresponding wind speed profiles. Why is this?

This is because near-surface wind speed values from the DH2 are unreliable, due to the manual, rather than autopilot flight, during take-off and landing, which conflict with the measurements and calculations of wind speed. Therefore, we don't consider wind speeds below 30 m to be reliable. The authors have added text to the manuscript to specify this (lines 156-158).

In which coordinate system is the wind speed measured, relative to sea ice or in earth coordinates? I think this has only implications for the Richardson number at the first level, but this is still important to mention.

All variables are measured in the same coordinate system, which is Earth-relative (now written on line 144). We recognize thus that the wind speed in the Earth-relative coordinate system does not represent the actual wind shear between the atmosphere and sea ice that is also moving, and that the correct shear is that between the true wind and the speed and direction of the sea ice movement. However, we don't believe this to be an issue for the calculation of Rib at the lowest level because the ice was moving slowly relative to the wind speed. Krumpen et al. (2021) at https://doi.org/10.5194/tc-15-3897-2021 found that the average drift speed of the ice during MOSAiC was less than 0.1 m/s, which is very small compared to the wind speeds observed. Therefore, an assumption of 0 wind speed at the surface is sufficient for the calculation of Rib at the first level. Above this, we are comparing two wind speed measurements, so the drift of the ice is no longer relevant. The authors have added a sentence to the manuscript to explain that the drift of the ice is not crucial to consider (lines 215-217).

Furthermore, the response time of the temperature and humidity sensors and their implications for the computation of Ri and theta_v and determination of the ABL height should be discussed. For theta_v two very different response times are combined (T and RH sensor). What effect does bin-averaging, including ascent and descent data have?

Differences in response times of the RSS421 temperature and RH sensors has a negligible impact on the calculation of theta_v because the moisture content in the Arctic atmosphere is very low. In the coldest temperatures sampled, theta and theta_v values typically differed on the order of ~0.1 K, and in the warmest temperatures sampled, theta and theta_v values typically differed on the order of ~1 K. Regardless, the addition of humidity does not change the structure and location of features in the theta_v or Rib profiles, which is what is important for ABL height identification. We have included a figure in the supplement attached to these author comments (Figure 1 in author_comment_figs.pdf) which show an example of a cold and warm case, to demonstrate how little the potential temperate and Rib profiles differ when theta versus theta_v is used. We have also added text to the manuscript to discuss this (lines 189-193). To address your question about the effect of bin averaging through ascent and descent, we have added some text to the manuscript which describes how we average the theta_v, humidity, and wind speed variables over 1 m altitude bins throughout the entire flight to further eliminate the effects of differences in sensor response times during ascent and descent (lines 193-199).

How may non-stationarity, e.g. a substantial temperature change near the surface within the ~30m flight time affect the results, and are there any observations indicating that this may have been an issue (One could simply make use of surface-based observations to detect non-stationary conditions during the flight period)?

Visually comparing ascent and descent profiles of the UAS flights allowed us to see that there was never substantial change in the near-surface values of variables over the ~30 min flights. However, averaging in 1 m bins throughout the flight would mitigate any impacts this would have on the theta_v or Rib profiles. The authors have added a sentence on lines 195-197 which states this.

I am missing a general assessment of the different sampling methods (radiosonde vs DH2). There are some important differences like the radiosonde can only sample during ascent, differences in the vertical climb speeds, response time, wind measurements. It' is therefore well possible that one of the data sets is generally more smooth or has higher uncertainties in particular in a specific altitude region or under certain conditions.

The authors recognize that the sampling methods of the DH2 and radiosonde are different. However, both platforms still observed atmospheric features similarly, which can be seen in the comparisons of DH2 and radiosonde profiles in Figure 7 of the original manuscript (Figure 3 of the revised manuscript) and the Supplementary Figures showing every flight.

Some paired DH2 and radiosonde profiles agree better than others (they agree well in the revised Supplementary Figure S23 but poorly in S40), but the ones that don't agree as well are cases with greater time between DH2 and radiosonde launch in which atmospheric conditions may have changed. Nevertheless, the point of this paper is not to discuss the impacts of different sampling methods on the efficacy of the methods. The purpose of this work was to improve ABL height identification methods using the high resolution DH2 data, discover which one works best in the Arctic regime, and then to test whether these adapted methods work well on another platform (i.e., the radiosonde), despite differences in sampling methods. The results show that they do work with similar efficacy, which argues that one can take these methods and apply them to UAS or radiosonde data alike, without having to tweak them. The authors now try to make it clearer the purpose of including radiosonde profiles (lines 375-378, 623-625), and discuss some of the differences in sampling methods (lines 541-548), with a focus on how these differences don't significantly change the efficacy of the objective ABL height detection methods (lines 548-550, 632-635). The authors have also added a few sentences to explain that the radiosonde data have some smoothing applied to account for the swinging of the pendulum after launch, as well as how this smoothing might affect the choice of critical bulk Richardson number (lines 521-526).

The results/discussion section is kept on a rather general level and has little content compared to the methods chapter. Interesting research questions are not addressed (systematically), e.g.:

- ***How do different sampling and data processing methods affect the differences for radiosonde vs DH2 based ABL height estimates?***

As described in our response to the previous comment, the authors have added some text to describe the differences in sampling methods between the DH2 and radiosonde, but despite this, both datasets produce similar results (lines 541-550, 632-635), and thus the methods used in this paper are applicable to both platforms. Lastly, we also address how differences in processing methods (smoothing or vertical resolution) affect the efficacy of different objective methods (lines 521-526 and 551-561).

In Sect 3.2 you address the question of how stability or specific features in the ABL structure cause certain methods to perform better or worse than others (you also scratch on the surface of this in Sect. 2.2 ff) but this is done in a rather episodic manner. Table 3 would be a good starting point for expanding on this. In the corresponding section, you mention that one could list problematic features to be used in a pre-screening procedure. I think you should have the data and knowledge to propose such a list. I consider this as very relevant for the research community using similar systems to determine the ABL height.

Referee #3 suggested to use Figures 9 and 10 of the original manuscript (Figures 5 and 6 of the revised manuscript) to differentiate between the efficacy of the different methods based on stability regime. We have addressed this by creating identical figures which are shown in the Supplementary Figures document, which contain only SBL datapoints (S70 and S72) and only NBL datapoint (S71 and S73). We discuss the main takeaways of these figures in section 3.1 (lines 491-496 and 531-536) by highlighting which methods have the highest efficacy for each stability regime. In section 3.2, we recommend which objective method(s) to use if one were to choose a different method based on stability regime (lines 592-600). For listing problematic features to be used in a pre-screening procedure, we now list these in Table 3 of the original manuscript (Table 4 in the revised manuscript, beginning line 584) and have trimmed the text throughout section 3.2 to not be redundant. We hope that with presenting this information in a table in list form will make it easier for a reader to flag these features in a pre-screening procedure.

- ***How sensitive are the different methods to the choice data processing methods, or sensors used, e.g. vertical averaging procedure? This would also be of interest to the community, e.g., to make adjustments for different measurement systems.***

To address this, we have re-run our procedures with DH2 data averaged in 5 m, 10 m, and 20 m bins instead of the original 1 m bins, to determine if the methods and results are sensitive to the vertical averaging procedure. We have found that there is no significant difference in the efficacy of objective methods using the coarser data, aside from the Liu-Liang method, and the Heffter method at 10 and 20 m resolution. This suggests that the Liu-Liang and Heffter methods are more sensitive to vertical resolution, but the Rib and TGRDM methods are not. We have added a paragraph at the end of section 3.1 summarizing this (lines 551-561). We also include scatter plots identical to those presented in Figure 9 of the original manuscript (Figure 5 of the revised manuscript) in the supplement attached to the author comments (Figure 2 in author_comment_figs.pdf) which show the results of the 20 m resolution data.

Minor comments/suggestions

Since this is a technical paper small details may need some additional attention, e.g. the difference between level and layer (which has a thickness) and the corresponding indexing. Are meteorological parameters, gradients at level k averaged over k to k+1, and how exactly are averages and gradients determined?

The authors have taken care to make sure the uses of level and layer are correct. Values at altitude k are averaged between k-0.5 m to k+0.5 m. This information, along with more detail on how averaged and gradients are calculated was added to a new section titled "Preparing the DataHawk2 data for analysis" which appear before the determination of stability regime (lines 187-224).

I also regard the term "subjective" as misleading. In Section 2.3.1-3 you describe criteria that could also be automated (you may have done that). I would therefore claim it is a semi-objective method, where the final decision is made through a visual interpretation by an expert. "Visual" and "automatic" may be the better terms to distinguish the two types of methods. When reading the abstract my first impression was that it is strange to evaluate an "objective" method with a "subjective" one. Generally speaking, one may want to trust an objective method more than a subjective one. This confusion could be avoided by sticking to the terms "visual" and "automatic".

The authors prefer not to change the wording throughout the paper from 'subjective' and 'objective' to 'visual' and 'automatic', because the inherent dichotomy in the words 'subjective' and 'objective' give these two a clear distinction, which we would like to keep. However, we have adjusted the wording to make it clear throughout the paper, including in the abstract (lines 19-24), that the 'subjective' method refers to a visual/manual method, while the 'objective' methods refer to automated method (lines 111-112 and 220-224). The authors also disagree that the subjective methods can be automated. The strength of the subjective method is that an expert assesses all the features of the profile to identify the subjective ABL height. Here, the knowledge of the expert is critical and can't be automated. We have added some text to the beginning of section 2.4 which highlights this (lines 306-308).

In general, I also see the potential for condensing the content of in particular the introduction and methods sections.

We have drastically reduced the amount of content in the methods section by listing the subjective criteria in a table rather than paragraphs (Table 3 of the revised manuscript, beginning on line 291) and by removing a lot of the information provided about each objective method, including specific equations, and instead pointing the reader to the original citation for additional information. We also have moved Figures 3-6 of the original manuscript to the Supplementary Figures document. We have also done our best to condense the parts of the introduction that were present in the original manuscript, however comments from you and other referees have asked for additional information to be added to the introduction (mainly including more background information on the typical Arctic ABL structure, and discussing why knowing the ABL height is important). Therefore, the length of the introduction has not overall been reduced.

Readability could be increased by avoiding some passive phrases and using "we", as done elsewhere.

The authors have worked to change passive phrases to using "we" instead.

Specific comments, technical corrections, and suggestions for improvement:

L15: "fixed-wing uncrewed aircraft system"

This change has been made (lines 15-16 and 143).

L16: consider introducing an abbreviation for the ABL height, e.g. simply "Z" or "H" or with subscript "ABL"

The authors now use 'Z_{ABL}' throughout the paper instead of 'ABL height.' This is introduced on line 81.

L16: "the ABL structure".

This change has been made (line 17).

L18: "the ABL height". In general, I have the feeling that some articles are missing, in particular when abbreviations are used. Please add wherever this could increase the readability. Note that if you would simply use "H" instead of "ABL height" you would not need the article in this case.

We have changed 'ABL height' to 'Z_{ABL}' throughout the paper, so most problems like this are no longer an issue. However, we have also double checked that the articles are correct in each case.

L39: "pack ice" might be better.

This change has been made (line 40).

L49-60: This paragraph could be rewritten, giving a general description of the ABL and its typical structures of the ABL and above. This would then naturally lead to the concept of the ABL height.

An additional paragraph has been added (lines 61-73) which gives a general description of the Arctic ABL and its typical structure.

L51ff: "...the ABL is mostly impacted by interactions between the atmosphere and sea ice surface features, including the generation of turbulence through surface energy fluxes emitted from open water regions such as leads...". This is

hard to understand. I think you want to mention both, mechanical and buoyant production of turbulence, but it reads like buoyancy is part of the mechanical production (by interaction with surface features). It should also be made clear that buoyancy is mostly negative over sea ice.

We now specify that the ABL is impacted by interactions between the atmosphere and sea ice pack features, which includes the open water and lead components as well. Now, we first introduce that the Arctic ABL is impacted by both buoyant and mechanical turbulence. Then, we provide some examples of each. This can be found in lines 50-58. The authors disagree that buoyant turbulence is only present over open water features. For example, buoyant turbulence can also occur if colder air is advected over warmer sea ice.

L54: I consider mixing as a turbulent process so "radiative mixing" could be misleading. How about "turbulent mixing forced/triggered by radiation–cloud interaction"...

This part of the sentence has been changed to accommodate your comment as well as that of another referee to now specify that we refer to "turbulent mixing below cloud base due to cloud top radiative cooling" (line 55).

L55: Consider mentioning the effect of ice edges e.g., at leads. The roughness is increased due to the freeboard.

This is now mentioned (line 57).

L59: include "(LLJ)" here.

The authors prefer to introduce the abbreviation LLJ farther down on line 77, when we discuss the important features with which the ABL interacts. We fear that if we introduce the abbreviation LLJ where you suggest, it may get lost in all the citations, and it is more easy to notice when introduced on line 77.

L76: only one ")"

Two ")" are necessary since this is the end of a citation within a statement in parentheses.

L76: Here one could expand a bit on where the ABL height could be identified be when there is a capping LLJ? At the core, somewhere below or above? Are there different opinions about this?

When a capping LLJ is used to identify ABL height (as is done in one of the Liu-Liang SBL methods), the ABL height is identified to be the height of the LLJ core. The authors now specify this more clearly (line 103-104). The authors now also specify in lines 567-568 that according to our observations and to the literature, the ABL is typically at or below the LLJ core. We also cite literature which agrees that identifying the ABL height as the LLJ core height is often not accurate (lines 568-570).

L114: Is the SHT-85 really measuring at 100Hz. The response time of this sensor is rather slow, but of course, it is allowed to oversample.

Yes, the SHT is really sampling at 100 Hz. We include this information in section 2.1 for the sake of completeness in describing the instrumentation carried by the DH2. However, as you point out, the response time of the SHT is quite slow, so this is why we don't use it for the current analysis, but instead use data from the RSS421.

L117: Here more details on the 3D wind estimates could be given (see comment

above).

Per your above comments, more information has been added about 3D wind estimates (lines 148-152). We have also cited more sources where additional information can be found (lines 154-155).

L110-L120: Can you also provide similar information (doesn't have to be as detailed) for the radiosonde sensors (here or elsewhere). Does the radiosonde also contain a Vaisala sensor (I think the corresponding radiosonde sensor would be RS41)? Can the radiosonde sensor also be included in Table 2?

We have added a sentence when we introduce applying the objective methods to the radiosonde profile which indicates that the radiosonde sensor (RS41-SGP) pressure, temperature, and humidity variables have the same resolution, repeatability, and response time as those for the DH2's RSS421 variables, as listed in Table 2. We also add the uncertainty and resolution in the wind speed and direction of the RS41-SGP (lines 382-384).

L147ff: Were there any cases when a clear determination of the boundary layer height was not possible, even though the max altitude was sufficient. Potential reasons could be non-stationarity or internal boundary layers?

We were always able to determine the ABL height when the max altitude was sufficient. Sometimes this was difficult, but this is why we consider several variables (theta_v, humidity, Rib) when determining ABL height. Considering all this information, we could always determine the ABL height with reasonable confidence.

L152-155: Does this mean data from both, the ascent and descent were used? If so you may cause a kink at the level where you start using the first data after takeoff but have data from the descend before landing.

The authors are aware of this potential influence, but did not notice this to cause in issue in creating false kinks. Nonetheless, we were conscious of this when looking for kinks in the profiles for ABL height identification.

L174, 175, 177: Should be "an SBL/NBL" but "a CBL". Check the entire manuscript for this type of typo.

This typo has been fixed throughout the manuscript.

L176: You mean that the SBL "can range from …" but what is written refers to theta_v.

The structure of this sentence has been changed and moved to the introduction to accommodate other referee comments, so is no longer a problem (lines 61-63).

L183: It is not clear what "i" refers to.

The authors specify now that "i" refers to the initial altitude of the DH2 referenced in the below equations (lines 232-233).

L208-209: Your data may even suggest that there is a tendency to more stable conditions during the seasons you observed.

Since we had an almost equal number of SBL and NBL cases during the seasons we observed (actually 1 more NBL than SBL), this is not true (lines 248-249). However, we have removed this statement in the process of condensing the methods section.

L227: "The Bulk Richardson number"

We have changed 'bulk Richardson number' to 'Rib' in this case, so 'the' is no longer needed (line 272).

L227-228: Consider rephrasing the first sentence to allow for buoyant suppression. "buoyancy" may be better than "buoyant production"

We have changed this sentence to also mention buoyant suppression from static stability (lines 272-274).

L240ff: Here you should be more precise. See for example https://glossary.ametsoc.org/wiki/Bulk_richardson_number: "In the limit of layer thickness becoming small, the bulk Richardson number approaches the gradient Richardson number, for which a critical Richardson number is roughly Ric = 0.25 … Unfortunately, a critical value is not well defined for the bulk Richardson number, leading to uncertainty in turbulence likelihood for values near the critical value."

The authors have adjusted the wording in this paragraph to summarize the information provided by the AMS glossary (which we now cite), which highlights that a critical value of 0.25 is not necessarily always the exact number that corresponds to the transition between turbulent and laminar conditions, but low Rib is generally expected in the ABL, and high Rib is expected above the ABL (lines 274-283).

L247: Here more detailed information would be necessary. The raw data is bin averaged using 1-m bins,. then you use 30-m bins for Rib with 5-m resolution. Note that the choice of bin size may have implications for the choice of thresholds, e.g., for Ri_b. This could be discussed further.

The authors have added a new section titled "Preparing the DataHawk2 data for analysis" (beginning on line 187) before "Determining stability regime" in which we describe the processing methods (averaging raw data over 1 m bins, and then calculating Rib and dtheta_v/dz over an altitude range of 30 m with 5 m resolution) in greater detail to accommodate this and other referee comments. We discuss why we choose 1 m as the bin size for averaging the raw data (lines 197-199) as well as how the choice of this bin size affects the results (lines 551-561).

L250-255: Assuming that in the lowermost ~5m theta_v will have its strongest gradient this method has its weakness when no data from this level can be used, as the CBL would be found at higher theta_v, thus resulting in an underestimation. If only the ascent data below ~5m is ignored but descent data is used this should be clearly indicated. An alternative would be to use IR surface temperature estimates. Overall this appears like a very objective method, which can be fully automated.

Both ascent and descent data below ~5 m is ignored, since typically below 5 m is sampled with manual flight during takeoff and landing. We have added a sentence to explain this (lines 202-204). The IR surface temperature estimates from the DH2 are not accurate enough to aid in this because they are not thoroughly calibrated, so they are primarily used to qualitatively identify variations in surface temperature (e.g., resulting from leads or melt ponds). Now that we present the subjective criteria in a table, we have condensed the information and in doing so, have also removed the insinuation that CBL height is dependent on the exact 5 m theta_v value when it is subjectively identified. A similar fully automated approach is applied with the objective Liu-Liang method for a CBL (lines 317-319), and since this objective method uses exact thresholds, it is dependent on the lowest DH2 value and thus is more subject to underestimation (this is noted as a drawback of the objective methods in the conclusion on lines 647-650). This is why the subjective ABL height is taken to be the more accurate one.

L250. "theta_v" without "the"

Due to the presentation of the subjective methods in the form of a table, this sentence has been rephrased and this comment is no longer applicable.

L251: Replace "identify" e.g. with "determine" to avoid repetition.

Due to the presentation of the subjective methods in the form of a table, this sentence has been rephrased and this comment is no longer applicable.

L266: Consider changing to "slope of theta_v". At which heights are these multiple shifts?

These multiple shifts occur around 30 m, 100 m, and 130 m. However, to accommodate presentation this information in a table, rather than paragraphs, we had to removed excess information, including details about the location of specific kinks in the example profiles. We hope the difference between the theta_v profiles in Fig. 2b and 2c are evident enough that we don't need to explicitly state this information in the manuscript.

L271: The determination is not made entirely based on the humidity. In the last step, theta_v is used again.

The way this information is presented was changed to fit into the table format, so this sentence has been restructured. Thus, this comment is no longer applicable.

L278-280: If DH2 data is only considered from altitudes above a certain threshold this statement is not well supported unless surface temperature estimates from an onboard IR sensor are taken into consideration. Note that IR surface temperature estimates may be subject to uncertainties related to sensor temperature stability, the emissivity of the surface, radiation flux divergence, and sensor tilt.

This statement was meant to explain a characteristic of a SBL, by definition, rather than by what was observed with the DH2. However, due to restructuring this information into table format, this sentence has been removed and this comment is no longer applicable, though we do still note that the gradient of theta_v is positive in an SBL on lines 229-230. And again, as explained in response to one of your above comments, the onboard IR sensor measurements cannot be used to derive exact surface temperature measurements.

L284-285: This is often related to an inflection point in the wind profile or at least the layer where wind shear approaches zero. Showing both profiles of theta_v and wind speed would be illustrative, for the interpretation of Ri_b.

We specify now that Rib increases when wind shear decreases above the ABL (line 283). We already provide both profiles of theta_v and wind speed in the original manuscript's Figure 7 (Figure 3 of the revised manuscript), as well as for each flight in the Supplementary Figures (S5-69), so this information can be found there. We don't add this to the subjective methods figure, as the wind speed profile is not directly used.

L290-L291: This statement implies that the correct SBL height is known. One

could also choose to define the SBL height as the level where there is such a clear shift in Ri_b. One may then simply end up with a different height.

The way this information is presented was changed to fit into the table format, so this sentence has been removed. Thus, this comment is no longer applicable.

L305: See previous comment: Does this imply that the ABL height can be determined with a resolution of 5m or 30m?

Objective ABL height can be determined with a resolution of 5 m when the method ultimately relies on the dtheta_v/dz or Rib profiles. Otherwise, it can be determined with a 1 m resolution. The authors have added two sentences to specify this on lines 310-312.

L308-309: Do you use equations 1-3 for the determination of the regime? Please include a statement, making this crossreference since this (is as you indicate) slightly different from the Liu-Liang method.

The authors now specify on line 317 that the regime is determined using what are labelled as equations 2-4 in the revised manuscript.

L318: Note that the height of the lowest level is critical. If the levels close to the surface are not sampled this may become an issue. You may want to discuss this later on.

Since the DH2 samples down to ~5m (and sometimes lower), we can be fairly confident that CBL heights calculated using the DH2 data are accurate to a few meters. Issues do arise in the radiosonde data which does not sample as close to the surface (lowest 23m or higher), but the result of this is that a shallow convective layer is missed, and the stability regime is identified as an NBL or SBL. This uncertainty then is not due to the height of the CBL, but rather due to the fact that other methods were applied, as the CBL was not recognized. We have added a few sentences to the conclusion stating this drawback of the radiosonde data (lines 647-650).

L320: "the atmosphere"

In the revised manuscript this sentence was removed so this comment is no longer applicable.

L323: The notations for vertical gradients are not consistent, compare e.g., line 304. Please stick to one notation.

Due to shortening of the methods section, we removed this section, and this comment is no longer applicable.

L326, Eq 6: Gradients would have to be determined between two layers so from k to k+1 (or k-1 to k+1). It is hard to follow which of these two levels is chosen as the ABL height. Depending on your resolution this makes a difference for the ABL height. Since this is a reoccurring issue for all methods relying on vertical gradients a general statement at the beginning of section 2.2 would help.

The section we added before section 2.2 titled "Preparing the DataHawk2 data for analysis" (beginning on line 187) describes the resolution and how averages and gradients are calculated, which should clear this issue up.

L334-335: The chosen threshold is quite different from the originally proposed one. What is the reason for considering this as inappropriate?

We consider it inappropriate because the ABL heights found with the original threshold were far too low compared with the subjective ABL height and had no physical basis when analyzing the profiles; this may be due to differences in the vertical resolution or smoothing methods of our data versus that used by Liu and Liang (2010). The authors add some text on lines 321-323 to specify this.

L341-342: This statement appears very general. TKE is just one way to define the SBL height. This goes back to my general comment on a more critical reflection on ABL heights. It should also be moved to the introduction.

The authors have moved this statement to the introduction, and specify now that TKE is just one methods that often works well for SBL height identification, but is not used since we do not have this data (lines 106-108).

L343: Below the ABL there is only the surface, which is only buoyant in an oceanographic sense.

The authors meant to say "within the ABL." We have fixed this (line 326).

L344: simpler: "the SBL height" or "the height of an SBL".

This change has been made (line 327).

L356: Consider changing the subscript to account for different ranges of theta^dot_r for different regimes (compare Sections 2.4.1.1-3). BTW, what does the r stand for?

Due to shortening of the methods section, we removed this sentence, and this comment is no longer applicable.

L391: better "starts" - "extends" may be associated with an upward direction.

Due to shortening of the methods section, we removed this sentence, and this comment is no longer applicable.

L396: layer or level?

Layer is correct (line 336).

L402: "the stability regime"

Due to shortening of the methods section, we removed this sentence, and this comment is no longer applicable.

L413: "local maximum"

This change has been made (line 349).

L416: "local minimum"

This change has been made (line 351).

L420: Can you provide a brief interpretation of this figure as done for the previous methods/figures?

To shorten the methods section, we have removed interpretations of the figures for the other methods and moved Figs 3-6 to the Supplementary Figures. To remain consistent, we don't add interpretation of this figure.

L429: I suggest using the term "threshold value" to avoid a bit of the discussion on the critical value of Ri (see also comment above). One could interpret the following paragraphs as you were trying to find Ri_bc for the transition between turbulent and laminar flow based on your observations, but in fact, this is not the scope of this paper and you don't make use of any turbulence observations.

We no longer use the term 'Ri_bc', but rather use the phrase 'threshold values' (line 360) to indicate how we identify the ABL height. We don't believe that the text can be interpreted that we are trying to find the critical value for the transition between turbulent and laminar flow based on our observations (line 358-360 and 363-364). If you feel that this is still implied, please let us know what statements convey this idea and we will revise them.

L445-447: I get an idea of what is meant here, but I would suggest reformulating this sentence to make it more clear. Can this be broken up into two sentences?

Based on the suggestion of another referee, we have removed this sentence altogether.

L451-452: Can you summarize the main differences that may play a role?

We have added a brief summary of the main modifications to the original objective methods (lines 369-372) and a brief discussion of why they are necessary (lines 372-375): "These adaptations are necessary in part because previous implementations involved analysis of radiosonde profiles, which have a lower resolution than the DH2 profiles, and in mid-latitude locations, where the ABL structure is often quite different than that observed in the Arctic (due to the lack of daytime convection or a diurnal cycle in the Arctic most of the time)."

L453: "...applied to radiosonde data ..." is enough, "to identify ABL height" is implicit.

This paragraph has been restructured, so this comment is no longer applicable. However, we made sure in revising this paragraph to remove redundant or implicit information.

L454: It would be natural to state the number of the radiosonde profiles you used somewhere here.

This has been added (line 380-382).

L458: no "the" before "theta_v"

This paragraph has been restructured, so this comment is no longer applicable.

L459: Consider changing to e.g., "create profiles of the same parameters as for the DH2 data"

This paragraph has been restructured, so this comment is no longer applicable.

L462: can you give the reason for this inaccuracy?

This is due to the Polarstern acting as a heat source. Specification of this has been added to the text (lines 388-390).

L469: Stick to one common unit for vertical temperature gradients. I propose K/km.

The authors disagree in this case, and think that changing the units in this sentence would be confusing to a reader. This sentence is discussing how we calculate stability regime for the radiosondes over a 30 m range using an adaptation of the values given in equations 2-4 of the revised manuscript. Changing the units here would make it confusing where we got the numbers from, whereas now we believe it is clear that we are adapting the original threshold of δ_s = 0.2 K over 40 m to 0.15 K over 30 m (see lines 394-397).

L474-476: Do you mean: "Similar figures for all available DH2 and radiosonde profiles can be found in ...". Is it possible to use a hyperlink to get directed to this online supplementary material?

We have changed the wording of this sentence to match what you recommend (lines 401-402). It does not seem to be possible to hyperlink to the Supplementary Figures document, however this is provided on the AMT site at the same DOI as the paper.

L481: Consider reformulating, e.g.: "In general, the deviation between ABL heights from DH2 and the radiosonde increases with decreasing time proximity".

This change has been made (lines 414-415).

L507: This is one example where the subjective method appears as the "truth", although it is most likely not perfect, either. Consider adding "... compared to the subjective method".

This change has been made (line 445).

L510-511: Please note that there have been some high-level debate on the use of p-values and the 5% statistical significance, see e.g., https://www.nature.com/articles/d41586-019-00857-9. I am not an expert in statistics but I recommend at least using a somewhat "softer" formulation, like "can be considered statistically significant when the p-value is less than 5% (or 0.05)." If your p-value is 0.05, this means there is still a 5% chance that your result is completely random.

We have adjusted the text to be "softer" based on your recommendations (lines 445-447).

L513 and elsewhere: Consider using superscripts "Ri_b^0.5" and Ri_b^0.75.

Instead of using superscripts, the authors use Rib(0.5) and Rib(0.75). These are introduced on lines 364-366. We fear that the use of superscripts might make it appear to a reader like we are taking an exponent of the bulk Richardson number.

L519-520: See comment above

We have clarified that this is compared to the subjective ABL height (line 454-455).

L520: What is more complex, the method or the result from the method?'

We specify now that we refer to the results of the method, not the method itself (line 455-456).

L532: "R^2" without "number"

This change has been made (line 465).

L534: "0.5 to 1"

The authors prefer the original text, as the wording you suggest may be confusing to a reader (line 467).

L556-559: Very wordy sentence for saying that you assess the (cumulative) frequency distribution for the difference of the objective methods relative to the subjective one.

We have condensed this sentence using the wording you suggest (line 497-498).

L571: Consider changing to "number of cases within each (relative difference) category"

We have made this change (line 510).

L575: Consider replacing "predicts" since it's a diagnostic method.

We have replaced "predicts" with "results in" (line 514). We have also made this change wherever else the word "predicts" was used.

L581-584: Here, I would like to see some discussion on such aspects. Are there differences in the sampling or data processing methods that may lead to the fact that different threshold values work best? Such discussions may be very useful for the research community as they may have to adapt threshold values depending on their observational approach.

We predict that the better efficacy of the lower Rib critical value for the radiosonde data results from the fact that more smoothing procedures are applied to the radiosonde data when they are processed by Vaisala. Minimal smoothing is applied to the DH2 data – the only smoothing comes from the vertical averaging over 1 m bins. We believe this to be the cause, rather than the difference in vertical resolution, because we have run the routines on the DH2 with different vertical resolutions (we tried vertically averaging the data over 5 m, 10 m, and 20 m bins) and found no significant difference in the efficacy of the Rib method with either critical value. We have added some text on lines 521-526 (before Fig 6 of the revised manuscript) which states the hypothesis that smoothing is what makes the difference in Rib critical value, and some more text on lines 551-561 which discusses that Rib critical value is not sensitive to vertical resolution.

L588: "the ABL heights"

We now say "After comparing Z_{ABL} from the different objective methods…" as we have changed the syntax for ABL height (line 537).

L595: Consider replacing "it is not consistent enough to be reliable" with "it is not reliable".

This change has been made (line 563).

L598-599: Do you mean the LLJ core? This statement could need a reference.

We now clarify that we refer to the LLJ core (line 567). This was intended to mean that the LLJ core was observed to usually be above the subjective ABL height in the data used in the study, which also agrees with the literature. We have thus added citations, but also note that this is seen in our observations (lines 567-568).

L600-601: "throughout the whole profile"

This section has been restructured to list the reasons for objective method failure primarily in a table, rather than paragraphs. Thus, this sentence no longer exists as it originally did, but we use your suggestion when we phrase option 1 of Liu-Liang failures in the table (Table 4 of the revised manuscript, beginning on line 584).

L625: radiosondes were launched from the deck so "right at the surface" should be changed to "close to the surface"

Due to restructuring of this information in table form, some details were removed, including this one. Therefore, this comment is no longer applicable.

L628: I suggest changing to "during polar night".

Due to restructuring of this information in table form, some details were removed, including this one. Therefore, this comment is no longer applicable.

L630: There is at least a debate on whether the free atmosphere above the ABL is really laminar or rather weakly turbulent. The choice of a threshold value for Rib largely depends on the vertical resolution you use to compute Ri_b.

The authors have restructured this sentence, and now say: "Lastly, the failure of the Ri_b method occurs due to the difficulty of defining an accurate critical value which correctly captures the likelihood of turbulence for all cases" (lines 579-580). In attempting to determine if the threshold value for Rib depends on the vertical resolution of the data used to compute Rib, we found no sensitivity of the efficacy of the critical values of 0.5 and 0.75 when vertical resolution of 1 m, 5 m, 10 m, and 20 m, are tested (discussed in lines 551-561). Above this, the range over which Rib itself is calculated would have to be adjusted, and at this point the applicable threshold value for Rib would probably increase. We have added a sentence to the end of section 3.1 which recognizes this (lines 560-561).

L649: only "several", "different" is implicit

This change has been made (line 607).

L649: "methods (i.e., Liu-Liang ...)"

This change has been made (line 607).

L652: You could state the threshold values 0.5 and 0.75.

This change has been made (line 610)

L656: It is occasionally quite good, better to use"largely".

This change has been made (line 613).

L663: Is this also true for largely ice-free areas in the Arctic, which are likely underrepresented by a sea-ice-based campaign?

We now specify that this refers to sea ice regions, as the ABL structure over ice free regions in the Arctic can be quite different and we don't address this in the current study (line 617).

L671: This repetition should not be necessary.

This sentence has been removed.

L672: It should be safe to use active voice: "would change minimally"

This change has been made (626).

L679: "These similar conclusions" (plural)

This typo has been fixed (632).

L681: "no method" or "no single method"

We now say "no single method" (line 636).

L681-683: Again: The objective methods may be better than a visual inspection by a non-expert. A combination of both visual + objective may be better. For the semi-automatic approaches, the list of features that may cause certain methods to fail would be very useful. One different approach could be to use an ensemble of automatic methods and visually inspect only the profiles for which the resulting ABL heights diverge.

The authors now specify that the subjective methods are most accurate, but require much time and knowledge of ABL dynamics, and thus the objective methods may be a better choice for a non-expert (lines 636-641). We also now suggest the semi-automatic approach of using an ensemble of automatic methods and visually inspect only the profiles for which the resulting ABL heights diverge, as you suggest (lines 651-653). Additionally, the features that cause certain methods to fail is now listed in a table (Table 4 of the revised manuscript), so that they can be more easily found and applied by a reader. We now reference this table in the conclusion (lines 646-647) so that in case a reader is skimming the paper by reading the conclusion first, they will find that this information is included in the paper.

Figure 2: Caption: "each flight" is misleading when showing only selected flights

We now specify that we refer to each flight shown in the figure (line 294).

Figure 3: The caption is extremely long. The legends could be merged and plotted only once (applies to more figures)

We have moved Figures 3-6 of the original manuscript to the Supplementary Figures document and now provide one legend for each figure, instead of for each panel (see Supplementary Figures S1-S4). With this, we also have shortened the captions without removing key information.

Figure 7: This figure should be redone. Here are some suggestions: Use different color schemes for the figures, e.g., not two different shades of green in the same panel (Panel 5 also has several different shades of green for the horizontal lines to indicate ABL heights from the different methods). What are the dashed lines in Panel 5? Since the ABL heights shown as text in Panel 1 and 2 are also related to Panel 3-5 it would make more sense to put them in a small table (2 lines 6-7 rows), placed under the 5 panels. Use one common legend for all 5 panels. Consider using a logarithmic scale for Ri (only if all values are positive) or a

narrower range. Condense the caption and structure it better. These suggestions could partially also improve the other figures.

This figure has been redone according to your suggestions (see Fig. 3 of the revised manuscript). Now, we use black and grey lines for the vertical profiles of DH2 and radiosonde data respectively. Only horizontal lines are colored, and we have changed these colors to be more distinct. Additionally, all DH2 related lines are solid and all radiosonde related lines are dashed. We have moved the wind speed profile to be the 5th panel, instead of the 3rd. We now provide only one common legend which refers to the whole figure. We have decreased the range for Rib from 10 to 5. We have added a SBL example to this figure since we have moved Figures 3-6 of the original manuscript to the Supplementary Figures document. We leave the text of ABL heights on Panels 1 and 2 so that the information is all in one place and is easier to keep track of, now that we have added an additional set of panels for the SBL. With all of these changes, we were also able to shorten the caption without removing key information.

Figure 8: Consider, using a smaller range for the y-scale in the bottom panel and rather mention that a few outliers are not visible with this scaling. It may also be simpler to use "Relative difference" instead of "Absolute value of the percent difference".

These changes have been made (see Figure 4 of the revised manuscript).

Figure 10: This may be my personal preference, but it might be better to display this as a CDF plot (four lines) or a histogram (four bars for each bin) using bins with a constant width (e.g., ranging from 10% to 20%). It would also be possible to combine the CDF and histograms in one panel. Consider using a different y-label, e.g., "frequency of occurrence" and only one common x-label. For the "No ABL Height found" class you could simply add NaN or display them differently, e.g., plot them as shaded bars or horizontal dashed lines (sort of downward from 100%). I also note that the bars in the last two columns don't add up to 100%. Do the missing cases indicate a relative difference exceeding 100%?

Some of these changes have been applied, but we still use the same bar plot as before, as we believe it accomplishes what you want to see (see Fig. 6 of the revised manuscript). The changes made are 1) we use the y-label of "frequency of occurrence" that you recommend, and now list the platform type in a title for each panel, 2) we use one common x-label and just indicate the % below each set of plots, 3) for the "No ABL height found" class, we label this as "No Z subscript ABL", as we found NaN to not be fitting, and 4) we use diagonal lines in the bars for the "No Z subscript ABL" cases to differentiate them. Yes, the missing values that cause it not to add up to 100% indicate a relative difference exceeding 100%. We do not add comments on this to the text as we believe this is implied.

Please also note the supplement to this comment:
https://amt.copernicus.org/preprints/amt-2021-383/amt-2021-383-AC2-supplement.pdf