Response to anonymous referee comments

The authors again thank the three anonymous referees for taking the time to review the resubmitted draft of our manuscript and for their helpful comments, which have improved the manuscript further. 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.

Anonymous referee #1

Second review of « Testing the efficacy of atmospheric boundary layer height detection algorithms using uncrewed aircraft system data from MOSAiC" by Jozef et al.

The manuscript was largely improved, the description of the methods is better structured and shortened and the results are more clearly presented. There is however still some points that have to be further clarified:

Major comment:
1. The calculation of the Richardson number (L213-215) has to be clarified: I don't understand why Rib is computed over 30 m range at 5 m resolution. If I understand the authors well, the Δθv and Δz of equation 1 are built over 30 m, i.e. the "ground reference" is always 30 m below the altitude at which Rib is measured ? I checked in Seibert et al 2000 and in Zhang et al., 2014, the Rib is always computed with the reference level at the ground. Stull mentioned (p. 177) that the critical value of 0.25 applies only for local gradients, not for finite differences across thick layers. No size is however given to estimate what is meant by "thick". Up to now, I have never seen such a calculation of Rib (if I understand well what the authors have done).

Additional information has been added to the manuscript to specify why we calculate bulk Richardson number over a 30 m altitude range, rather than using the ground as the reference level. First, on lines 222-225 we specify that we use this 30 m range rather than ground level in order to isolate local likelihood of turbulence rather than calculating that over the full depth from the surface. Next, on lines 289-299 we discuss why using this running 30 m range, rather than the ground as the reference level, is actually beneficial for having a critical value that is consistent throughout the profile, and for that critical value being somewhat close to 0.25. This is because, according to Stull 1988 (page 177) as you mentioned, the critical value of 0.25 applies only for local gradients, and with a thicker layer, we are more likely to average out large local gradients. Well, when we use a consistently 30 m thick layer, we are essentially finding local gradients and are a lot less likely to average out large local gradients which may be averaged out when calculating Rib over an ever-increasing depth using the ground as the reference level. Thus, a critical value of 0.25 is more likely to apply when using the methods of Rib calculation that we use. Lastly, such a calculation of Rib has been used in previous literature, so we are not the first to introduce it. While many papers do use the method in which the reference level is always the ground, Georgoulias et al., 2009 (now cited in the paper) and Dai et al., 2014 (already cited in the paper) both specify that they calculate the Rib profile using gradients over distinct layers in the atmosphere, not using the ground as a reference. And they then use these profiles of

Rib to identify ABL height. Therefore, I am confident that we are justified in our method for calculating and our application of Rib profiles, and in fact our method makes more sense for detecting the likelihood of turbulence than the method of always using the ground as a reference level.

Minor comments:
• Table 1: my comments on Table 1 remain valuable. E.g. how can we understand the differences between "virtual potential temperature" and "vertical gradient of virtual potential temperature" ? What is used in the first case ? Similarly, the gradient (or kink that is a difference in the gradient) are used for liquid water content and absolute humidity. The description could be much more precise.

We have added a column to Table 1 called "Application of Quantity" in which we describe how the variables in the "Quantity Used" column are applied to find ABL height (Table 1 begins on line 97). Within each "Quantity Used" there is one or more subsections in "Application of Quantity" since some variables are used in multiple ways depending on the method. We hope this addresses your comment as we now describe the difference in the application of simply "virtual potential temperature" versus "vertical gradient of virtual potential temperature." We also now specify how the moisture profiles are used to identify ABL height.

• L110: "manual visual analysis" or human visual analysis ?

I don't understand the difference. How can manual visual analysis refer to anything but that done by a human? If there is no difference, the authors prefer the word "manual" to "human" and we leave it as is (line 116).

• L114: can you explain the mention of "the definition of this quantity is not constant over time"?

The authors apologize for this being poorly worded. We did not mean to say that the definition of the quantity is not constant over time, we meant to say that ABL height is not constant over time. We have changed the wording to reflect this line 119-120.

• You should clearly mention in page 4 that "manual visual analysis" is refered as "subjective" later on. (this is done at line 221, but it is somewhat late)

We had already included on line 117-118 that manually determined ABL height is referred to as 'subjectively' determined ABL height. Therefore, we do already mention this on page 4 and no additional change has been made.

• Table 3 and Fig. 2:
o As stated in table 3 for the CBL cases " ZABL is the altitude at which the vertical gradient of θv is positive and is the bottom of a layer of enhanced stability ". Such a description remains quite evasive and the definition of the Parcel method suits probably much better ZABL: the altitude at which the virtual potential temperature becomes equal to the one at the ground (i.e. at 2 m). We can quite easily identify visually (and objectively) this altitude. If the increase in

Rib is found at the identified ZABL, the kink in humidity profile is not unique and could have been put at ~250 m

The reason we do not include the criteria that θ_v reaches that at the surface (as is the case in the Parcel method) is because the surface value of θ_v may not be known in instances when these subjective criteria are applied. For example, usually the lowest DH2 measurement in a given profile is at 5 m, and the lowest good radiosonde measurement in a given profile is around 23 m (at least for the radiosondes from MOSAiC). Thus, such subjective criteria that rely on a surface measurement cannot be applied in many cases. Instead, indicating that the gradient of θ_v is positive and is the bottom of a layer of enhanced stability at the top of the ABL does not require the use of a surface measurement, and essentially results in the same conclusion of what the ABL height is in the central Arctic where the unstable layer of the ABL, if present, will be very shallow. With regard to your comment that a kink in the humidity profile can be also found at 250 m, this is true of the mixing ratio profile, but a clear kink at 136 m can be seen in the RH profile, and a faint kink here can also be seen in the mixing ratio profile. Since this altitude has kinks in several of the shown profiles (θ_v, RH, r, and Rib), this is the altitude chosen for ABL height. We have thus changed the wording in this section of the table to say "…corresponding to a kink in the relative and/or absolute humidity profiles…" (CBL section of Table 3, beginning on line 311).

o Fig. 2c: once again the increase in Rib is well marked but the kink in humidity seems much more obvious at ~150 m.

The authors disagree with you in this case. While there is a change in the humidity profile at around 150 m, it is certainly not as pronounced of a kink as that which occurs at 100 m. In this case, we believe the increase in humidity above 100 m to be a humidity inversion, and when there is a humidity inversion, this usually occurs above the ABL.

o Fig. 2d, 2e: similarly to Fig 2c, I would estimate the kink in humidity some 30-50 m higher. (I deduce that the unique description of Table would not allow several independent persons to deduce the same ZABL and I do understand much better the significance of "subjective". Moreover Rib seems to be a more determinant parameter than what is explained in the paper.

We have laid out in our subjective criteria that we primarily look for a kink in the θ_v profile and secondarily in the humidity profiles (line 307-309); thus, if there are multiple clear humidity kinks, we look for the humidity kink which corresponds to a θ_v kink. Given that, in 2d and 2e, the relevant humidity kinks were found to be those that also corresponded to the most clear kink in the θ_v profile and an increase in Rib. You have brought up a good point that Rib seems to be a more determinant parameter than what is explained in the paper, so we have changed the wording to indicate that Rib is used more heavily for CBL and NBL cases, but for many SBL cases, the humidity profiles are used more heavily (lines 307-310) as they often provide more useful information especially in cases of very weak turbulence Lastly, to clear up some confusion as to which humidity kink or θ_v kink to consider as ZABL, we have added a sentence to specify that, if there are multiple clear kinks in different profiles at different altitudes, then preferential treatment is given to the kink that also corresponds to an increase in Ri_b. We also clarify that if kinks in the relative humidity and mixing ratio profiles occur at different altitudes, preferential treatment is given to the kink which occurs at the same altitude as that in the θ_v and/or Rib profiles (lines 323-325). We hope this added specification will better result in independent persons finding the same ZABL from these criteria. But in the end, these are "subjective" not "objective" criteria for a reason – they cannot be automated and are subject to uncertainty, but we have addressed this uncertainty (lines 318-327), and have concluded that the uncertainty is relatively small.

• Fig. S4a: I do not understand why the ZABL identified with the Rib method is not set at the first altitude where Rib is greater than the threshold at about 100 m? This is also the case for e.g. Fig. S10, S13, S18, S19, S20, S21, S23

The reason we do not simply take the first altitude at which Rib exceeds the threshold as ZABL is because, since we are calculating Rib over a 30 m range with ascending altitude, it is possible for there to be small local non-turbulent layer within the turbulent ABL. By requiring Rib to exceed the threshold for at least 20 m, we are looking for where the likelihood for turbulence has truly ceased above the ABL. We have added some discussion to section 2.4.4 to cover this (lines 389-393). Looking at Figure S4a, based on the θ_v profile, and the other subjective criteria, the subjective top of the ABL is just below 200 m, which is also the level above which Rib is consistently greater than the threshold. So this is actually a good example as why we do not simply take the first altitude where Rib exceeds the threshold as ZABL, because if we took it at the level of first exceedance of the threshold, the ABL height found would be too low. For the other supplementary figures you reference, if we take the first altitude where Rib exceeds the threshold as ZABL, this will also give a ZABL much lower than the subjective ZABL. We settled on a exceedance of the threshold by a consecutive 20 m (4 datapoints) as the criteria for identifying ZABL after trying many options and seeing which one performed best. Some examples of other options we tried are 2 of 3 consecutive datapoint, 3 of 3, 3 of 4, 3 of 5, 4 of 5, 5 of 5, etc.

• Fig 4: since the ABL is rather shallow in Arctic and the uncertainties remains of about some 10 m (e.g. an uncertainty of 30 m is given for the subjective method (p. 11)). An absolute difference would probably give a better view. From Fig. 5f we can see that the relative difference between the subjective method applied to DH2 and RS can be as high as 200%.

We have changed this plot to show absolute difference, rather than relative difference.

• Fig. 5a: the red slope (correlation between the subjective method and RS) is mostly determined by the isolated points> 500m and does not seems me a relevant value to characterize the difference.

This is a good point, and additionally, the very low R2 value for the Liu-Liang method indicates that there is not much correlation between the objective and subjective ZABL for this method, so analysis of the slope does not provide reliable information. We have added a sentence to the manuscript to address this (line 510-512).

• L507-508: could you give a tentative explanation for the high number of cases with no Liu-Liang ZABL identification and with a relative difference > 100% ?

The reasons for failure of the Liu-Liang method are already listed in Table 4 and discussed in Section 3.4, but we add a preview sentence at the point in the paper to which your comment refers, that the primary reason for the failure of the Liu-Liang method is the high prevalence of a weak θv inversion that persists throughout the entire lower atmosphere in the Arctic (lines 566-568).

<u>Anonymous referee #2</u>

Minor comments:

L68: The entire free atmosphere and in cases of an SBL even the entire troposphere can be characterized by a potential temperature inversion. Do you mean "capping temperature inversion"? In an SBL the inversion is often surface-based and transitions directly into the free atmosphere without a capping inversion.

Thank you for this comment – we agree that the way it was written before does not properly get the point across. We have changed the sentence to now say: "While the various forms that the Arctic ABL may take are complex, most of the time, the Arctic ABL is capped by a temperature inversion (which may extend to the surface for a stable ABL) and local maximum in potential temperature gradient, marking the entrainment zone…" (lines 71-73). We hope this clarifies what we mean.

Table 1: It is not clear to me in which way the "Quantities used" are evaluated, e.g. 1) "virtual potential temperature" vs. 2) "vertical gradient of virtual potential temperature". Is this supposed to mean that in 1) you look at a bulk temperature gradient or difference between two layers? If taking the difference of potT between two layers this is essentially also a gradient. Or are you referring to a 2nd order derivative (curvature)? I suggest trying to make this clear by always using terms like: "gradient", "(bulk) difference", etc. Note that "bulk Richardson number" can stay as it is since the gradients/differences are implicit and since you use a simple threshold (same for TKE).

We have added a column to Table 1 called "Application of Quantity" in which we describe how the variables in the "Quantity Used" column are applied to find ABL height (see updated Table 1, beginning on line 97). Within each "Quantity Used" there is one or more subsections in "Application of Quantity" since some variables are used in multiple ways depending on the method. We hope this addresses your comment as we now describe the difference in the application of simply "virtual potential temperature" versus "vertical gradient of virtual potential temperature."

L104-106: The statement here contradicts what is stated in L68-69. I agree with the statement made here.

We have changed the wording on the original L68-69 (now lines 73-74) so that it does not seem we are implying that the temperature inversion ceases at the top of the ABL: "…marking the entrainment zone, which is a stable layer that makes the transition from the ABL to the free atmosphere (Stull, 1988).” Now, it should read that the temperature inversion just makes the transition between the ABL and the free atmosphere, but that temperature inversion can still extend (though weaker) throughout the free atmosphere. We hope you no longer find these two statements contradictory.

Section 2.1/2.1.1: How is the altitude used for all of the following analyses determined? I suppose the DH2 provides different altitude estimates from different sources, e.g. GNSS, barometer, extended Kalman-Filter using both, etc. Each method has its uncertainty. Unless RTK is used GNSS uncertainty is in the order of 5m and often behaves “jumpy” when satellites are disappearing or popping up; barometers are subject to drift due to non-sufficient temperature stabilization of the autopilot (on older models) or background pressure changes (not always negligible over a 30min period). The bigger problem with pressure-based altitude estimation is that most autopilots do this internally based on the assumption of a standard atmosphere, which is, however, way too warm for Arctic conditions and thus causing errors in the order of ~10% or worse in very in cold conditions. Recomputing altitude based on pressure detrending and the hypsometric equation using ambient (measured) temperature is a simple way to reduce this uncertainty (see Barbieri et al., 2018 (https://www.mdpi.com/1424-8220/19/9/2179); Greene et al. 2022 (https://link.springer.com/article/10.1007/s10546-022-00693-x)). Radiosonde data may be subject to similar uncertainties.

The altitude estimates are obtained using a GNSS receiver and barometer onboard. The barometer derived altitude (pressure altitude) is typically accurate to within a couple of meters but due to long observation periods (on the order of 30-40 minutes) tends to drift because the barometer is calibrated only once before each flight. Your concern with GNSS derived altitude uncertainty is justified. Although GNSS altitudes are ‘jumpy’ locally, they provide drift-free altitude estimates over the duration of the flight. The altitude used in DH2 analysis, termed 'GPS calibrated pressure altitude', corrects for the drift in high-resolution barometric pressure altitude with the GPS/GNSS obtained over the duration of the flight. We have added two sentences to summarize how the altitudes used for this study were obtained (line 154-157). As far as the radiosonde data goes, altitude is calculated using pressure measurements which are compared to the initial pressure at 10 m to determine altitude via the hydrostatic equation (line 418). Thus, one source of uncertainty is due to that of the pressure measurement, which is 1 hPa. Additionally, the pendulum swing and other motion under the balloon after launch could influence the altitude calculation, but these effects are generally smoothed during the data processing by Vaisala.

L148 ff: Although this has been improved I am still not sure how the wind speed is computed: Can you cite one published paper describing at least a similar method. Mayer et. al (2012, https://doi.org/10.1260/1756-8293.4.1.15) is to my knowledge the first publication introducing the no-flow sensor method for determining horizontal wind speed based on GPS speed along circular flight paths. I can only guess that your method is similar, possibly including a correction for variations in true airspeed. Another useful reference might be Rautenberg et al. (2018, https://doi.org/10.3390/atmos9110422).

We have restructured this paragraph so that the information for how the winds are calculated can primarily be found in the references cited. Winds from the DH2 have been calculated by two methods, which are both included in the B1 netcdf files, via a “standard” approach, and a “hybrid” approach. We now briefly introduce each of these approaches and point the reader to two citations for each approach where they can read about the methods in depth. We also now specify that the winds used in this study are those from the “hybrid” approach (line 158-166). We hope that this now satisfies your comment, and a reader should easily be able to find the information on how the DH2 winds are calculated by primarily referring to the references, now that the paper describing the technical information regarding the DH2, which includes detailed information about how the winds are calculated, is in preprint (Hamilton et al., 2022: https://doi.org/10.5194/amt-2022-96).

L154: There is some additional information given here, however, I don't regard this as very clear. Can you add some references here? It is not really subject to this review, but I expected the information to be also found in the metadata of the *.nc files. For the sake of creating self-explaining files, can you add this?

With the restructuring of this paragraph, we have added more references which describe how the DH2 winds are calculated. Now, at the end of the paragraph where we describe the wind estimation, we provide a sentence to inform the reader that more information about the processing of all of the variables that the DH2 measures can be found in the metadata for the netcdf files (line 168-170). This information is, and has been, provided in the metadata for the netcdf files already, so we are not sure what you are missing.

L168-169: What is meant by “filtered to remove the impact of angle and ground speed”. Do you mean a coordinate transformation from true airspeed to wind speed or filtering to remove high-frequency variations in the angle (which angle? Yaw or all Eularian attitude angles) and ground speed?

This information is actually no longer true for the current way the DH2 winds are calculated. In previous iterations of the wind processing, we had applied more filtering routines, but in the final iteration, we do not. Thank you for bringing our attention to this sentence, as it was not our intention to include it. Thus, we have removed this sentence from the manuscript.

L452: “0.3 to 1.2” not “1” also in line 467 an upper limit of 1 (which may be understood as 1.0) is not appropriate. I also suggest giving the same amount of digits for all values.

The authors apologize for not being clear with what we meant here. In this section, when we discuss the slope values, we are comparing the slope to an ideal value of 1.00, which is what the slope would be if objective ZABL = subjective ZABL in every case (if the intercept is also 0). We have added some text on line 480-481 to introduce this concept before we get into the discussion. Then, when we said all the slopes are “within 0.3 of 1” for the DH2 and “within 0.5 of 1” for the radiosonde, what we really meant was that all slopes fall within 1.00 +/- 0.30 and 1.00 +/- 0.50 respectively. We have changed the text to reflect this (line 490 and line 505). We have also made sure to give the same number of digits for all slope values and R2 values in our discussion.

L492-496 and S70-71: It is great that you included these additional analyses. I consider these

results very interesting and would encourage you to give higher attention to this. At first sight, I noted some additional important results that should, in my eyes, be highlighted, e.g.: the high discrepancy in the performance of the Heffter and TGRDM methods for the two different regimes. The Ri methods show less dependency on stability (for DH2) however they work better with a higher threshold value for NBL and lower SBL cases. In addition, the discrepancy between radiosonde and DH profiles becomes larger for some methods when dividing between regimes (Rib for NBL and RS is performing very poorly – any idea why?). This may suggest that some of the methods are not as robust across platforms as originally thought. These points should also be reflected in the discussion. Liu-Liang does apart from the 4-5 extreme outliers do very well in the NBL regime (see also S73). Consider moving these two figures to the main manuscript or maybe better, mark the different regimes in Fig5 with different symbols and add a table with the results now in the different figure panels to not overload the plots.

We have moved supplementary figures S70 and S71 from the previous draft to the main text, and have combined them as a two-part figure that is now Figure 6 in the revised manuscript (line 544). We already tried portraying this information as one figure with different symbols for the different stability regimes, but found it to be much too busy to try to decipher the information from – it is much easier to visualize the main takeaways when you can see the datapoints from the different stability regimes separately. We have also added more discussion along with these figures to address the points you mention in your comment. Specifically, we address the difference in the efficacy of the Heffter and TGRDM methods between different stability regimes, the stability dependence of the Rib method, how the Rib method especially with threshold value of 0.75 does not perform as well for NBL cases, and how the Liu-Liang is actually pretty good for NBL cases. This discussion can be found in lines 532-543.

Section 3.2 and Table 4. I wonder whether the question should rather be "Why does the subjective method result in different results" rather than "why does it fail?". E.g., 1a. " LLJ core altitude is well
above the ABL top" is not very helpful since we don't know the true ABL height. It would be more helpful to focus on the feature that is causing the subjective method to be different, e.g., wind shear or humidity profile which is not taken into consideration in the objective method. I am not 100% sure whether an approach like this is feasible for pre-screening, but I am not convinced that some of the currently listed features are either. For example, how should one use feature Hefter 1 ("SBL height is not the altitude at which θv is 2 K warmer than θv at the surface") if the SBL height is unknown? I recommend at least checking all the listed features for their applicability in a pre-screening routine.

We have specified now that when we refer to "failure" of a method, we really mean that the objective ZABL is much different than the subjective ZABL (line 627-628). Since we have already concluded that the subjective ZABL is the best estimate of ZABL that we have, given the available data, if the objective ZABL is much different, we believe this constitutes a failure, i.e. the objective ZABL is most likely incorrect. The authors would like to keep the information provided in the table, but we have also added some discussion to point out that some objective methods might produce different results than the subjective method due to the lack of inclusion of humidity and/or wind shear features into the calculation (line 655-659). We have also added some text which specifies that not all features in Table 4 may be conducive for pre-screening, but at least it is a starting point to identify some cases in which it is likely that certain objective methods may fail (lines 660-665).

Technical corrections and suggestions:

Abstract: Include a statement on the stability dependency?

We have added a statement which mentions we also discuss in the paper how the success of the methods differs based on stability regime (line 24).

L52 "sea ice pack features": Remove "pack" and mention also open water. Maybe it's best to use "underlying surface" and then specify the different surfaces and their features that can be relevant. Here a distinction between thick and thin ice (allowing for a substantial heat flux) may also be relevant and can be combined with the cold air advection example given in the following sentence.

The two sentences have been adjusted to read: "In the central Arctic, the ABL is impacted by interactions between the atmosphere and underlying surface, including both sea ice and open water portions, which can cause either buoyantly or mechanically produced turbulence. The generation of buoyant turbulence can occur through surface energy fluxes emitted from open water regions such as leads (Lüpkes et al., 2008), cold air advection, especially over thin ice (Vihma et al., 2005), or turbulent mixing below cloud base due to cloud top radiative cooling (Tjernström et al., 2004)." (line 52-57).

L55-58: It would be more natural to start with mechanical production due to surface roughness and larger features such as ridges (ocean waves are also roughness features only moving) and ice edges and then continue with LLJs.

The sentence has been adjusted to read: "Mechanical generation, which is the dominant driver of turbulence in the central Arctic (Brooks et al., 2017), can occur due to the interaction between the atmosphere and surface roughness features such as ridges and ice edges (Andreas et al., 2010) or oceanic waves (Jenkins et al., 2012), or due to the presence of a low-level jet (Brooks et al., 2017; Banta, 2003)." (line 57-60).

L59: change to "plays a minor role". The entire sentence is a bit misleading since solar heating still can play a role even though it may not cause buoyant thermals but can still decrease stability substantially.

We have changed the wording to "plays only a minor role" (line 62), as you are correct that, even if solar radiation isn't creating an unstable situation with buoyant thermals, the addition of solar radiation to the surface energy budget will lessen the radiative cooling and thus lead to less stable conditions then if no solar radiation were present.

L61 ff: Add the information on the frequency of occurrence of CBL to the beginning of the

paragraph. Although rather rare CBL is still important in the central Arctic and it is often linked to openings in the sea ice (leads and polynyas).

We now mention that a CBL is rare in the first sentence of the paragraph (line 63). We also mention later on, when we describe how a convective ABL occurs, that when this does occur in the Arctic, it is likely due the presence of leads or polynyas (line 70-71).

L 74: Consider mentioning the relevance for NWP or atmospheric modeling in terms of enabling boundary layer parameterization schemes.

We now include mention of ABL parameterization in NWP models as a reason why it is important to know the ABL height from observations (line 86-87).

L 84: add a statement like "based on thermodynamic and kinematic profile data from UAS"

The sentence now reads: "The goal of the current work is to determine which methods, based on thermodynamic and kinematic UAS profile data, can best accomplish this" (line 88-90)

L85-87: some redundancies

We have combined the information on these lines into one sentence, which reads: "The depth of the ABL has been previously defined using a variety of approaches that involve visualizing the profiles of different thermodynamic and kinematic variables, which are listed in Table 1, along with some examples of associated literature that references use of that variable" (line 91-93). This should reduce the redundancy.

L88: "vertical structure"

This change has been made (line 94).

L122: Which method is adapted to best suit DH2 data?

All objective methods, aside from the Heffter method, needed some level of adaptation to best suit the DH2 data. This is now stated (line 128-129).

L173: "near-surface wind speeds"?

We now clarify that it is indeed the near-surface wind speeds (line 183).

L186: "unfavorable environmental conditions"?

This change has been made (line 196).

L193-197: I suggest reformulating these two sentences. You give three reasons for bin averaging the entire flight, the first at the beginning and the second at the end of the first sentence. Reason 2 links well to the sentence before so why not start with this.

We have restructured these sentences to read: “To further eliminate the effects of differences in sensor response times during ascent and descent, and for ease of visualization, we average the θv, humidity, and wind speed variables over 1 m altitude bins throughout the entire flight (e.g., values at 10.5 m are averaged from 10 to 11 m). This also mitigates the effect of changes in atmospheric conditions near the surface throughout the span of a flight, though the near-surface observations largely remained constant during a given flight” (line 203-207).

L213: should be “(delta z)"

This change has been made (line 223).

L215: consider adding that this results in Ri at z=15m, 45m, 50m, 55m, etc.

This has been added (line 227).

L216: the average is only an argument that it doesn't matter in most cases, but there might still be extremes that cause an error. Can you mention the maximum drift speed during your flights and include a statement that the error is in any case limited to the first level where Rib is determined?

We include now the maximum drift speed during the DH2 flights, which was 0.3 m/s (we also determined the average drift speed during DH2 flights was 0.09 m/s, which is approx. the same as the stated average drift throughout mosaic), and have added a sentence stating that any error that due to the drift speed is limited to the first level where Rib is determined (line 229-231).

L219: replace “dthetav/dz” with “it”

This change has been made (line 232).

L224: “physical processes” doesn’t fit in this context.

What we meant with this statement is that the underlying physical processes that dictate the ABL structure and height are considered similarly in both the subjective and objective methods. We have changed the wording to clarify this (lines 237-238).

L225: “the stabiltiy regime”

This change has been made (line 239).

L238: replace “gradient” with “difference”

This change has been made (line 252).

L243: replace “number” with “threshold”

This change has been made (line 257).

L272-273: More precisely, Rib is an approximation of this ratio.

We specify now that Rib is an approximation of this ratio (line 286).

L286: “stability regimes”

The authors disagree that “regimes” should be plural here. We therefore leave it as is (line 305).

L287 ff: Can these kinks be found more reliably from the second derivative (or its bulk approximation)?

While these kinks can also be found using a second derivative, this removes the benefit of analyzing the direct profiles of the variables, which allows an expert to better visualize how different physical processes which impact the ABL may be at play. With the visualization of the direct profiles, one can understand better the ABL structure, so we believe that searching for the kinks this way, rather than using a second derivative profile, yields a better estimate for the ZABL.

L290: Better something like: “Only in few especially difficult cases the Rib profiles were used heavily” or even “Only in few especially difficult cases we largely relied on the Rib profiles”?

Based on another reviewer’s comments, we have restructured/changed the end of this paragraph, but to also accommodate your comment, this now reads: “The primary methods applied to determine ZABL are those in which there are either one or two θv kinks, where we rely most heavily on the θv profile, and secondarily on the humidity and Rib profiles. For SBL cases, the humidity profiles often provide more insight than the Rib profile in identifying ZABL. In only a few especially difficult cases, we relied primarily on the Rib profiles” (lines 307-310).

Fig 2 and similar figures: Adding y-tick marks (I don't mean labels) to all subplots would make it much easier to read the heights from e.g. the Rib profiles. In addition, you could increase the width of each subplot by decreasing the horizontal spacing between them (keep a larger space between the right and left groups (a)-(b), etc.

I have added y-tick marks to all subplots in Figure 2, Figure 3, and Supplementary Figures S1-S69. I have also increased the width of subplots and increased the horizontal spacing between the groups (a), (b), (c), etc. for these same figures.

L299-301: Can you refer to the corresponding examples in the supplementary figures?

We have added a list of the corresponding examples in the supplementary figures, which were cases in which the ABL height was more ambiguous (line 319-320).

L303: Can you make use of this 30m max. uncertainty, e.g. in the analyses related to Fig. 5 and 6?

While we state an uncertainty in the subjective ZABL to be less than 30 m, this is only applicable to a handful of DH2 flights (~15%), whereas the majority have an uncertainty on the order of only ~1 m, due to the vertical averaging procedure and sensor response time. Therefore, we do not expect this uncertainty to make any significant effect on the results. We have added some text to the manuscript later on to mention this (line 622-625), but since we expect no significant effect on the results, we do not discuss it further.

L306: I would argue it can be automated e.g. through machine learning, but it may be complex and time-consuming.

We have added a note about the possibility of automation with machine learning, with the caveat that this may still not be fully reliable (lines 330).

L314: Considering that there were only two CBL cases it might make sense to exclude CBLs entirely from the analyses carried out in this manuscript. This would make it possible to slightly shorten the methodology section, e.g. L317-319, first row in Table 3, Fig 2a, etc.

The authors wish to include the discussion of the CBLs in the manuscript. While rare, as you yourself mentioned in a previous comments, they can be important still. More importantly, we want this paper to be a one-stop-shop for how to determine ABL height in the Arctic, so the exclusion of any discussion about a CBL would not accomplish this.

L321-323: Could this also be related to the fact that Liu and Liang use ~40m and 160m to determine the stability regime? Can you give more details on their vertical resolution and smoothing? I also wonder whether the method would work better with more smoothing and original thresholds. This can be discussed a bit later on.

We have added some text which discusses in more detail the difference in vertical resolution between ours and that used in Liu and Liang (2010). The vertical resolution of their data was ~40-50 m, which is much coarser than ours. This is likely the reason they were able to use a lower threshold. However, it does not make sense to interpolate the DH2 data to their resolution and use original thresholds because this would eliminate the ability the identify key features in the often shallow Arctic ABL (line 347-351). We don’t believe that the need for a higher threshold has anything to do with the fact that Liu and Liang use a different altitude range to determine stability, as the altitude range they use is simply not appliable to the central Arctic environment, but it does relate back to the core issue of their having much lower resolution data.

L356: It would be good to avoid the use of the term “critical value” for Rib here and in other occurrences. Further down you use “threshold value”, but I get that you first want to indicate that this threshold value is sort of related to the critical value.

We now use the word “threshold” rather than “critical” in this section (line 382-398), and throughout the paper when we refer to the threshold values which are applied to the DH2 and RS data. We only use the word “critical” when we explicitly refer to the standard critical Richardson number of 0.25, to which the threshold values we use are related.

L373: “vertical resolution”

This change has been made (line 405).

L374-375: “(due to the lack of daytime convection or a diurnal cycle in the Arctic most of the time)” is mentioned before.

We have removed this statement in parentheses here, as it is already discussed in the introduction.

L382-384: I consider the sampling rate, climb rate, response time (mentioned in Table 2), and resulting vertical resolution as equally important as the accuracy specs.

We have added information on the sampling rate, climb rate, and resulting vertical resolution for the radiosonde (line 416-417). We have also added information on the uncertainty in the pressure, temperature, and humidity measurements (line 414-415). We do not provide uncertainty from the RSS421 on the DH2 since this information is not provided in the data sheet.

L393: “the determination of the stability regime”

This change has been made (line 427).

L394-395: Do these adaptations in some cases result in different stability regimes compared to the corresponding DH2 profiles?

We have added the following sentence to address this: “These adaptations in themselves do not result in the identification of a different stability regime than is found in the DH2 profiles; instead, differences in stability regime between the two platforms may result from the lack of near-surface observations from the radiosonde, or a change in atmospheric structure between the two corresponding launches” (line 431-434).

L396: avoid this hybrid between text and mathematical expression or at least replace “x” with “times”. Better to use “\delta z”.

These changes have been made (line 430).

L415 and 416: “relative difference”

Another reviewer has asked we change this plot to absolute difference rather than relative difference, so this comment is no longer relevant.

L434 and 435: You use only one “subjective method”.

We have changed the wording to reflect one “subjective method” (line 471).

L435: Consider including "arguably" or similar. You don't have any proof supporting this statement.

We have added "arguably" to the sentence (line 472).

L448 ff: I suggest writing "(R2 = 0.653)" and "(slope = …", consider also replacing "slope" with e.g. "m". You can also simply use "R2" instead of "R2 value".

These changes have all been made (line 479-522).

L459 ff: I suggest consequently sticking to terminology like "can be considered as statistically significant". Although highly unlikely it can all be a coincidence.

This wording has been changed as you suggest (line 496-497, 515-516).

L478: I suggest changing to "rather strong correlation". For a strong correlation, I would expect R2>0.90 or even 0.95

This change has been made (line 518).

L545: "very close to the surface (~5 m)"

This change has been made (line 604).

L547: Consider specifying. Higher time resolution and slower climb rate result in higher vertical resolution.

We now say: "Additionally, the DH2 samples with higher vertical resolution (due to higher time resolution of instrumentation and slower climb rate)…" (line 606-607).

L549: What is meant by "improved"?

What we really meant is "adjusted." This refers to the fact that some threshold values were changed from the original published methods to best work for the DH2 data, so we mean to say that the objective methods with adjusted threshold values also work well for the radiosonde data. We have changed the wording accordingly (line 609).

L556: "10 m or 20 m bin size" or rather "10-m or 20-m bin size"

We say "10 m or 20 m bin size" (line 616). The AMT journal specifically says they do not want authors to use a dash between numbers and units.

L561: Why a "higher critical Rib value" and not a lower one. For the smoother RS data 0.5 works better. Consider using "threshold".

You make a good point. We have changed this sentence to say: “For vertical resolution of 30 m or coarser, the altitude range over which Rib is calculated would have to be increased, and at this point a lower threshold Rib value may be more applicable” (line 620-621).

L563 ff: Consider starting more general with all methods. It reads like Table 4 lists only the reasons for Liu-Liang.

We have restructured this section to first introduce what is shown in Table 4 for all objective methods. We then go on to discuss the specific failures for each method. This should now read like Table 4 lists reasons for all methods, and no one method is singled out (lines 627-629).

L570: A sentence on why it works relatively well for NBL cases would fit in here. May it have something to do that there is often no LLJ when the ABL is neutral?

We have added a sentence here which reads: “The Liu-Liang method likely performs better for NBL cases (as is evident in Fig. 6 and Supplementary Figure S71) than SBL cases because the Liu-Liang method for an NBL is not dependent on the sufficient diminishment of the θv inversion, nor the presence or altitude of a LLJ” (line 635-637).

L614: What is meant by “qualifying values”?

We have changed this to simply say “threshold values” (689).

L620-621: Consider using “Arctic pack ice to near the Arctic ice edge” or “marginal ice zone” if this fits better.

We now say: “…deep in the Arctic pack ice to near the marginal ice zone…” (line 695-696).

L631: Consider changing “percent difference” to “relative difference” throughout the manuscript.

We have made this change throughout the manuscript.

Anonymous referee #3

None