Article AMT-2020-78 – Response of Referee 1's comments

Q1 "The abstract lacks of motivation. Currently, there are a number of different methods to retrieve the ABL from lidar observations, whose results have been widely tested. Why should we use KABL and ADAABL algorithms?"

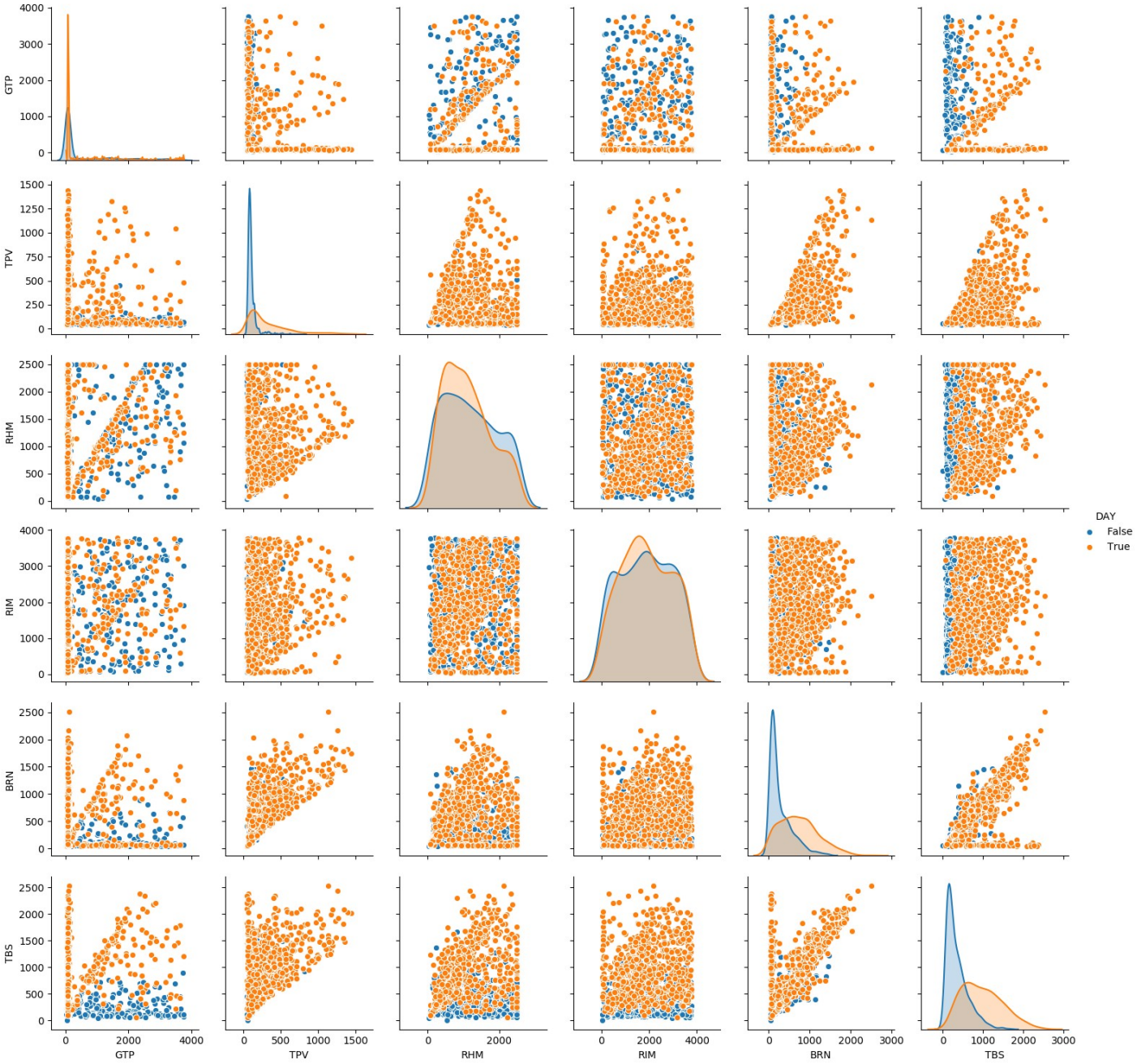
KABL and ADABL are open-source, usable by anyone, and not strongly tied to a particular instrument. Several methods exist to derive ABL from lidar observations, but it is also acknowledged that none of them are fully satisfying and the research on this topic is still active.
KABL is the reproduction of an existing method (scientifically interesting to reproduce previous results) with open-source libraries (technically interesting to give access to the source code and allow its reuse).
ADABL is a new algorithm (also open-source) that enables the reproduction of human expertise that we believe is valuable for deriving BLH. We develop this last point in our answer to the question 4, later in this document.

Q2 "The abstract states that "ADABL algorithm is performing better than KABL. . .". However, the authors do not indicate what they mean with 'performing better', or based on which results. In addition, the comparison uses radiosondes that always launched at the same time. How might this affect the study findings?"

The statement "ADABL algorithm is performing better than KABL. . ." refers to correlation and RMSE with RS, as both are better for ADABL.
The comparison with RS launched only twice a day is indeed a strong limitation of the evaluation. First, it reduces drastically the amount of KABL and ADABL estimations subject to evaluation (only 0.7 % are evaluated). Second, the RS are launched at 11:15 UTC and 23:15 UTC which are respectively at the end of the morning transition and at night. Such periods are known to be problematic for an accurate ABL estimation (for all methods). Third, the evolution of ABL throughout the day, which is important to assess the estimation quality, cannot be evaluated.

Q3 "Radiosonde data section, last sentence: 'After testing some of these methods on our dataset, we chose to derive boundary layer height with parcel method for the 11:15 sounding and bulk Richardson number for the 23:15 one.' Since the ABL height retrieved from radiosonde data is taken as reference in this study, the authors must explain why the parcel and Richardson methods were chosen. Also, they must explain how these different methods were tested."

The chosen rule "parcel method for 11:15 and bulk Richarson number at 23:15" follows the recommendations of Seibert et al. (2000), figure 10, assuming that morning launch is in unstable atmosphere and evening launch is in stable atmosphere.
However, our experience showed that estimation of BLH with the RS is not straightforward, even though they are considered as the reference. Several methods were implemented and applied on the 2-year data, but they hardly match, as it is shown in this figure :



One can see the pairwise comparison of each method with the others. It shows a matrix of subplots, each line and each columns corresponding to a method to derive BLH from RS. Off-diagonal subplots are the scatter-plot of one method against the other. Diagonal subplots are the smoothed histogram of the corresponding method. For all subplots, launches of 11:15 UTC (blue) and 23:15 UTC (orange) have been distinguished. Methods are identified by the following shortcuts:

- GTP : gradient of potential temperature (maximum of second derivative)
- TPV : virtual potential temperature profile (theta_v > moyenne niveaux inf +0.25K)
- GRM : minimum of mixing ratio gradient
- RHM : minimum of relative humidity gradient
- RIM : maximum of gradient Richardson number
- BRN : bulk Richardson number >0.25
- TBS : parcel method

This comparison shows how variable the estimation of BLH with RS can be. The two most agreeing methods are BRN and TBS, namely the bulk Richardson number and the parcel method. However, BRN gives a suspiciously large subset of estimations close to the ground during the day. Therefore our investigation is coherent with Seibert's recommendation.

Q4 "Section 3.1.2 and figure 4: The authors state "For few days where the boundary layer is easily visible for a human expert, the boundary layer top is drawn by hand: all points below this. . .". **I do not believe this is a criterion to estimate the ABL height. I suggest to use some of the other methods that have been previously tested and used in the literature."**

On this point, the authors disagree with the referee: although many automatic methods exist, they all have limitations (acknowledged in the literature, e.g. Seibert et. al., 2000, conclusions: "Since none of the methods and models are perfect, it is recommended to have results obtained in an operational context checked by a qualified scientist, considering the basic data").

MH retrieval methods consist in general of 2 steps: the first is the detection of the edges that separate two different layers, the second step is the attribution of the MH to one of those edge. Second step is critical in these methods and errors are often due to wrong attribution to the MH. In that regard, human expertise that uses physical interpretation and context awareness might provide a valuable input to edge attribution. Estimation with human expertise has limitations too: it can vary from one case to another or be different among experts, but the estimation is always based on physical interpretation and aware of context. Supervised algorithms like ADABL, are meant to reproduce the reference they are trained with. Adding existing proven methods is always interesting to interpret results and add elements for the comparison, however our focus was to evaluate ML methods as it was seldom investigated in the literature.

Q5 "Figure 9 shows the RMSE and correlation obtained from their comparison. However, nothing is said regarding RMSE. How did they calculate it? How is it defined?

The RMSE is defined at the section 3.4.1 by the equation 1, line 247. The method of computation is different for the sensitivity analysis (the reference is the handmade BLH, as explained at lines 280-283) and for the 2-year comparison (the reference is the RS estimation and only well-defined BL conditions are kept, as explained at lines 316-328). Figure 9 shows the results of 2-year comparison, thus its is explained at lines 316-328.

Q6 What are the errors of the radiosonde and lidar retrievals ?

The authors are not sure whether this question refers to the primary parameters measured by the sensors or to the BLH estimate derived from them.

Absolute uncertainty given by the manufacturer for the M10 radiosonde is 0.3°C for Temperature, < 1hPa for pressure, 3% for relative humidity and 0.15 m/s for horizontal wind velocity.
A quantitative analysis of the error for the lidar retrievals was not conducted. However the source of errors

and the correction procedures to minimize them are discussed in (Campbell et al., 2002) for the MPL systems and the uncertainties were further discussed in a quantitative fashion in (Welton et al., 2002)[1]. As the miniMPL is a compact version of the MPL, the same processing is conducted and the conclusions of these studies should apply at least qualitatively. To our understanding, the main source of error in the near range is the uncertainty of the overlap function. The first processed value is provided at 120 m by the manufacturer to limit this effect. At this range, the overlap function reaches a value over 10%.

Concerning the uncertainties on BLH retrieval, few values are available in the literature. For estimations from RS with bulk Richardson number, Seidel et al. (2012) evaluate the uncertainty at 50% for BLH<1km and 20% for BLH>1km. For parcel method, Seibert et al., (2000), emphasize the large uncertainty of BLH estimation when temperature inversion are weak but give no value of uncertainty. For BLH derived from backscatter profiles, Haeffelin et al. (2012) give in table 6 the standard deviation of the difference between STRAT-2D and RS (equivalent to our RMSE) for several instruments. For ALS450 (lidar in UV), it is of 345m during the day and 712m during the night.

However, it is also known in the literature that BLH uncertainties from lidar retrievals are hard to quantify. The value given above is a comparison with a reference, not the variability of the estimation. Many studies give quality flag instead uncertainty estimation (Haeffelin et al. 2012).

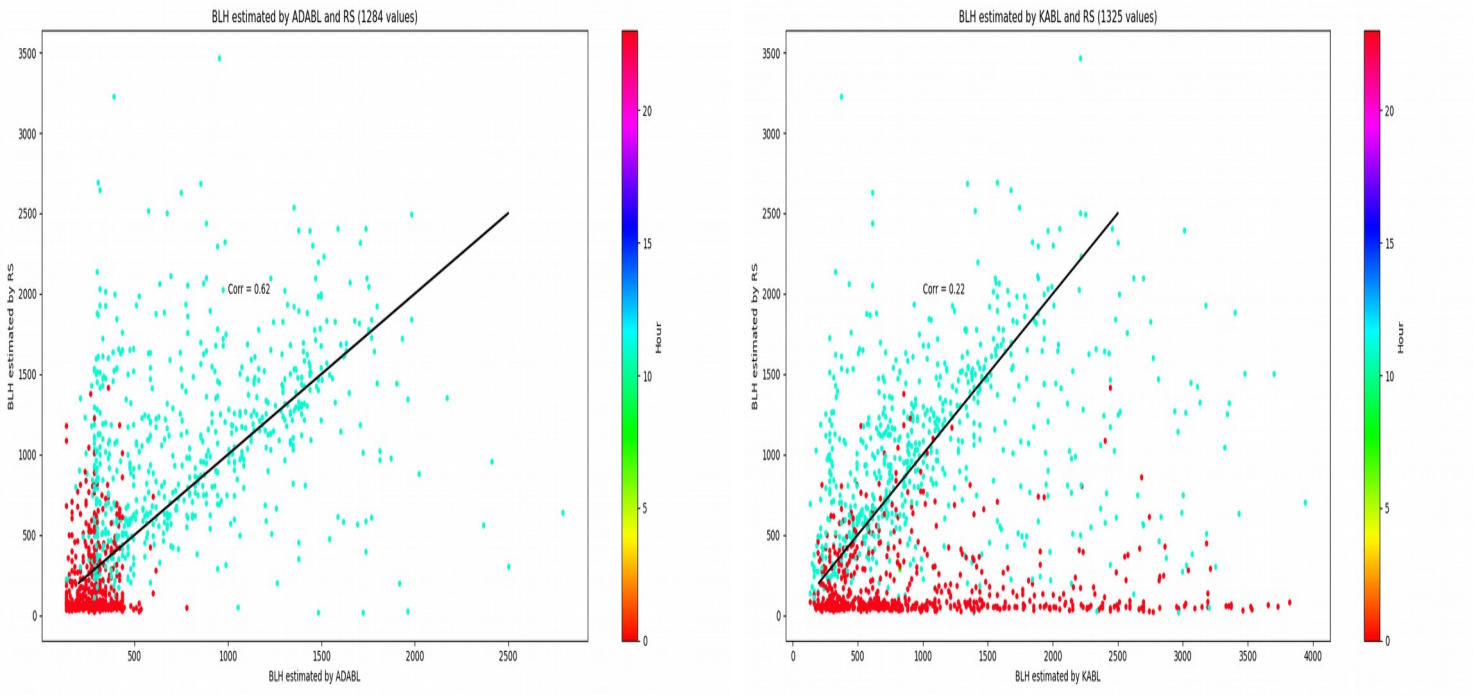
For KABL and ADABL, previously mentioned quality estimations are possible. The RMSE values given in figure 9 are comparable to the ones given in Haeffelin et al. (2012): we find for KABL 770m at Brest and 798m at Trappes, for ADABL 675m at Brest and 552m at Trappes. Our values are noticeably higher than in Haeffelin et al. (2012). We explain it by the length of our dataset (178 RS at Trappes, 101 at Brest, spanning over 2 years) and the little maturity of these algorithms. The referee's comment highlights that this comparison is missing in the manuscript. It will be added in the next version.

Moreover, for KABL and ADABL, it is possible to estimate the classification quality with internal scores, such as Davies-Bouldin, Calinski-Harabasz or the silhouette score. But the potential of turning internal scores into quality flags have not been made in this study. This would be a good prospect.

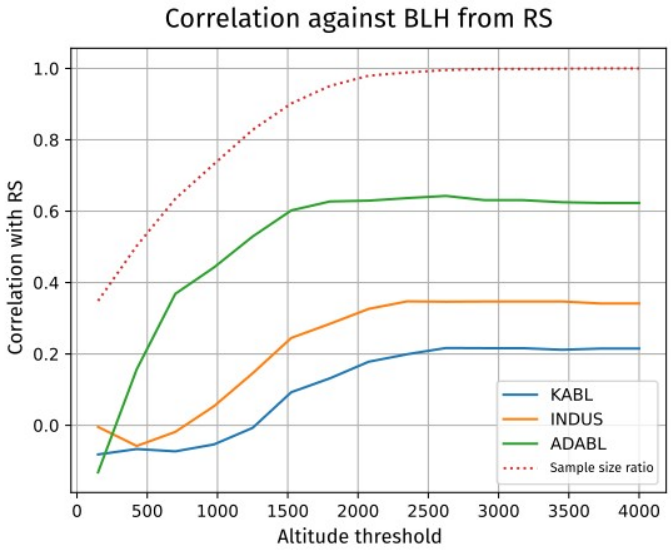
Q7 Does the correlation depend on the ABL height ?

These figures represent the RS estimation by against ADABL (left plot) and KABL (right plot) for the 2 years of data at Trappes. Coloration represents the hour in the day and the black solid line is the y=x correspondence.

1Welton, E. J., and J. R. Campbell, 2002: Micropulse Lidar Signals: Uncertainty Analysis. *J. Atmos. Oceanic Technol.*, 19, 2089–2094, https://doi.org/10.1175/1520-0426(2002)019<2089:MLSUA>2.0.CO;2.



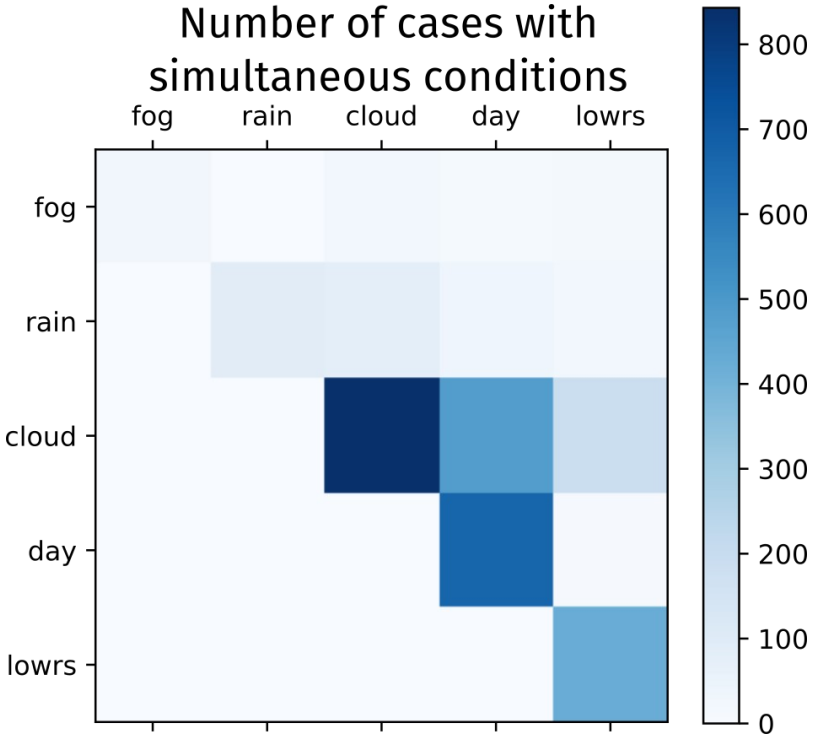
As we can see, most of the disagreement occur at low ABL height (according to RS estimation). Therefore, the correlation globally increases with height. It is confirmed by the following figure, representing the correlation with RS for all methods, considering only the points where the ABL estimation with RS is below an altitude threshold:



The correlation values given in these charts do not match with the ones in figure 9 because of the meteorological conditions filtering made before generating figure 9.

Q8 Also, it is said that a number of cases were not included in the analysis as a result of the meteorological conditions. How many cases?

As explained in the manuscript, cases with fog (2.2% of cases), with rain (7%), with cloud below 3000m (63%), nighttime (49.6%) and RS estimation below 120m were excluded from the comparison. As there are not simultaneous, the final proportion of excluded data is 84 % (Trappes). The statistics of the figure 9 are computed on 178 radiosoundings for Trappes, 101 for Brest.



Q9 Are the retrievals affected by the meteorological conditions, why?"

Yes. The lidar backscatter signal is rapidly attenuated by the presence of dense clouds, rain and fog to the point of being completely extinguished at a certain range. The boundary layer might be ill-defined in some meteorological conditions (e.g. in case of disturbance or storm).
In the study, three types of meteorological conditions were measured by ancillary probes: rain (rain gauge), fog (diffusometer) and cloud altitude (ceilometer). In case of rain, the boundary layer is ill-defined and the lidar is blind. In case of fog, the lidar is blind. In case of low cloud (below 3000m), the strongest gradient of backscatter will be at the cloud base, which is likely to fool the BLH estimation.

Therefore, all these meteorological conditions were excluded in order to ensure that the comparison is made on well-defined boundary layer.