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

Note: every response to all referee are followed by corresponding changes (CC) in the manuscript. Changes are referenced with lines numbers from the marked-up version (later in this document) to make them easier to identify. Changes visible in marked-up version that are not referenced as referee's response are for English language improvement.

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.

Corresponding Changes (CC): *mainly at lines 12 to 19 (marked-up version), at lines 66 to 69, and at lines 72 to 77.*

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.

CC: *no change.*

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.

CC: *The sentence at line 148 was changed for "Following recommendations from figure 10 of Seibert et al. (2000)".*

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.

CC: *motivation to exploit human expertise is added at lines 72 to 77 in the introduction.*

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 344 (marked-up version). The method of computation is different for the sensitivity analysis (Figures 7 and 8) and for the two-year comparison (Figure 9). For the sensitivity analysis, the reference is the handmade BLH, as explained at lines 383-384; while for the two-year comparison, the reference is the RS estimate and only well-defined BL conditions are kept, as explained at lines 433-434.

CC: *no specific change. English corrections might make it clearer.*

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

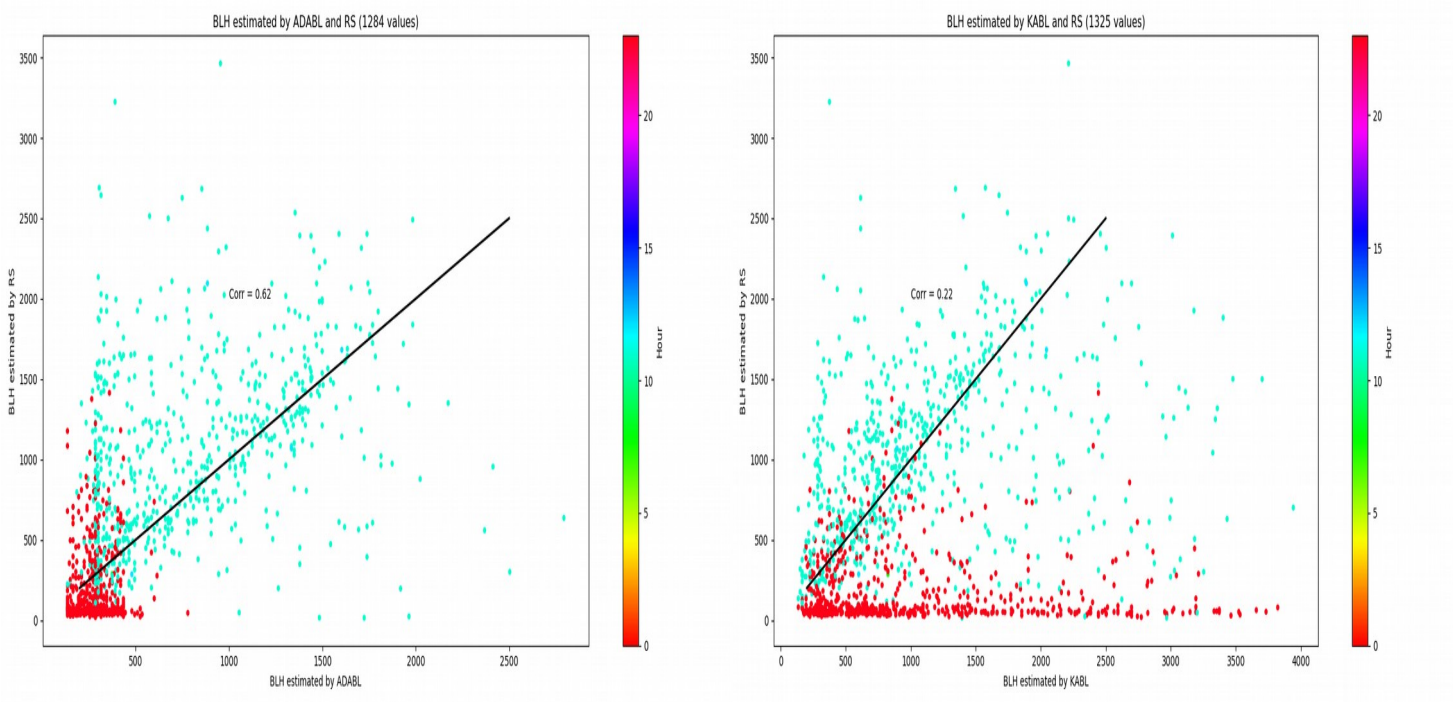
1Welton, E. J., and J. R. Campbell, 2002: Micropulse Lidar Signals: Uncertainty Analysis. *J. Atmos. Oceanic Technol.*, 19, 2089–2094, https://doi.org/10.1175/1520-0426(2002)019<2089:MLSUA>2.0.CO;2.

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.

CC: *add comparison with errors found in the literature at lines 459 to 470 and a new paragraph "5.8 quality flags" is added as a future prospect.*

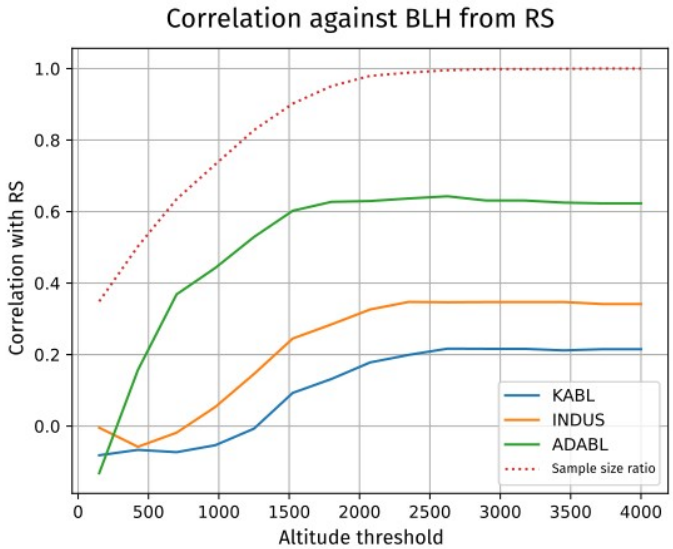
Q7 Does the correlation depend on the ABL height ?



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.

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:

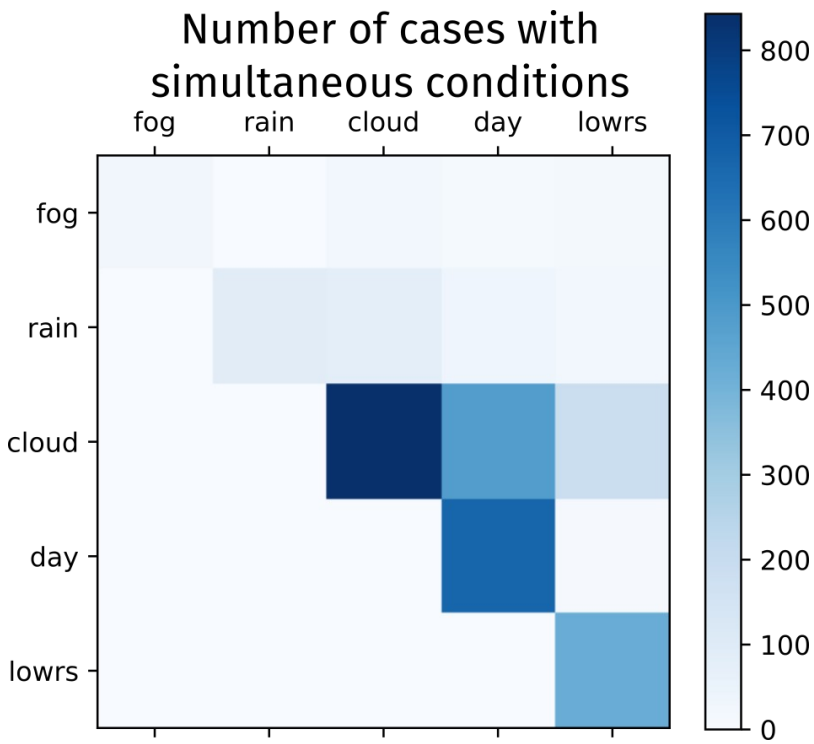
The correlation values given in these charts do not match with the ones in Figure 9 because of the meteorological conditions filtering made before generating Figure 9.



CC: *no change.*

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 Figure 9 are computed on 178 radiosoundings for Trappes, 101 for Brest.



CC: *the total cases used for the overall comparison are added (lines 446-447).*

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.

CC: *lines 573 to 579.*

Article AMT-2020-78 – Response to Referee 2's comments

Note: every response to all referee are followed by corresponding changes (CC) in the manuscript. Changes are referenced with lines numbers from the marked-up version (later in this document) to make them easier to identify. Changes visible in marked-up version that are not referenced as referee's response are for English language improvement.

General (Major) Comments

1. When training the supervised ML algorithm, the estimation of the accuracy of ADABL on the validation ensemble (by the cross-validation technique) is presented in line 171: 99.5%. I think it is important to present also the accuracy on some testing ensemble, at least on the case study of April 19, 2017. It will justify the generalization ability of the applied AdaBoost algorithm showing the algorithm performance in an independent dataset. In my opinion, the training dataset could be insufficient.

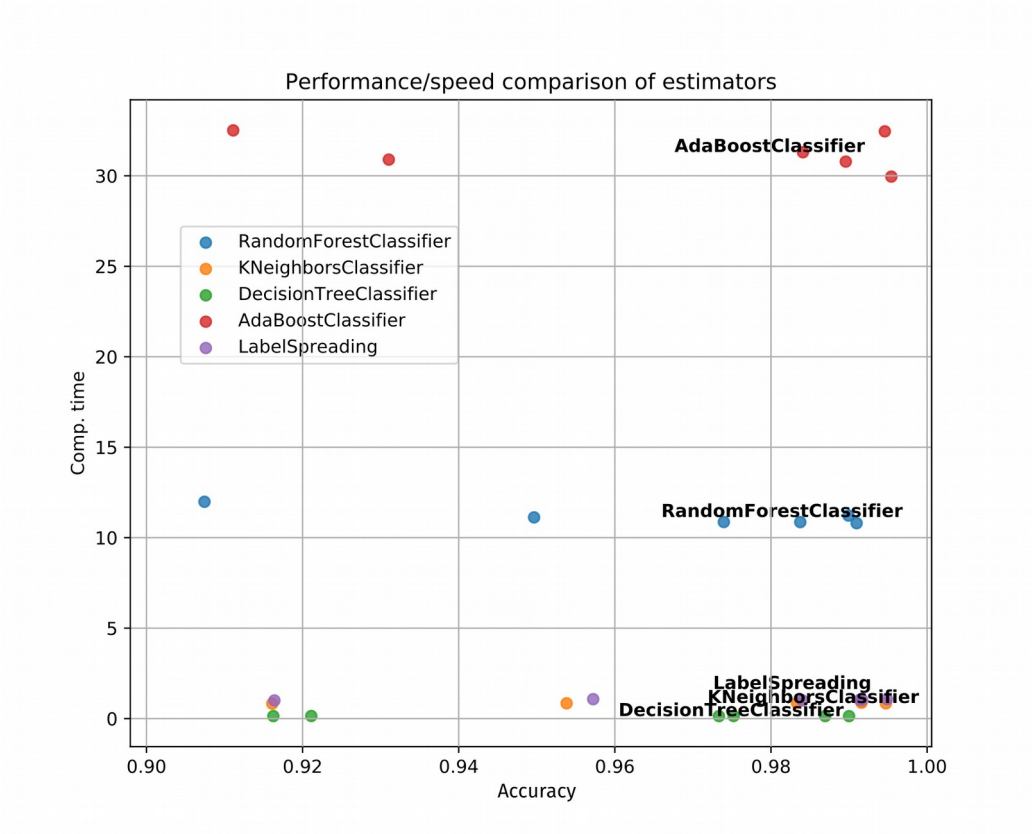
We acknowledge that the training dataset is insufficient, as it is commented in the results section. The point of this study is to show that, despite few training, ADABL has reasonably good results. As data labeling and algorithm training are time-consuming tasks, we would like to demonstrate first doing such tasks is of interest.
Time-consumption was also the reason why the case study of April 19 was not labeled, but the referee makes a very good point arguing that this would enable us to make a more accurate evaluation of ADABL (and KABL as well) on an brand new day. It will be added in the short term prospects.

In order to give a lower bound for the accuracy, we trained ADABL on one day and used the second day as validation set. The resulting accuracy is 87.6%. It is noticeably lower than the previous figure, but it suffers from a half-less rich training set.
In any case, the accuracy value is not meant to be a finding of this study. We extent this opinion in the next comment which is closely related.

Corresponding Changes (CC): *this comment is addressed from line 239 to 246 (marked-up version).*

2. Another reason why the accuracy of ADABL on validation ensemble is unrealistically high is the application of the random cross-validation split for time-correlated data (line 171). Using random selection from correlated datasets can lead to loss of generalization ability of the algorithm. The proposed ML method should be applied to new (independent) measurements. It means that AdaBoost should be trained-validated on uncorrelated parts of the dataset. If the data points were selected randomly from the whole dataset by the cross-validation procedure, it is highly probable that the similar neighbouring time points would be placed in both training and validation ensembles, which gives unrealistically good accuracy estimate on training-calibration datasets, but the worse result on another independent (test) dataset. I suggest the application of the block cross-validation.

We agree with the reviewer that the accuracy as it is estimated is too optimistic. Our effort were not directed towards having the best estimation of accuracy because it was only used to discriminate different supervised algorithms. The result of the comparison is not shown in the manuscript but can be found here: https://presentations.copernicus.org/EGU2020/EGU2020-19807_presentation.pdf
To meet with the referee's concern, the same study has been repeated with a block cross-validation. It was performed with group K-fold where groups were 4 hours chunks. The code to generate the figure is online on the Github repository (examples/perform_block_cv.py). The results are in the following figure. We can conclude that AdaBoost is still the most accurate among the one tested and its accuracy is rather 0.96 than 0.99.



To conclude, the accuracy value will be changed in the manuscript. The value of 0.96, obtained with block cross-validation is the most relevant in our opinion because it is more realistic than 0.99 and it was compared to the accuracy obtained by other algorithms with the same calculation. The lack of validation set will be highlighted as a limitation of the study.

CC: *this comment is addressed from line 237 to 246.*

3. If I understand right, the final configuration of the unsupervised ML algorithm (KABL) produces the classification using just one parameter - RCS0. In this case, the phrase in the conclusion (line 435) is misleading – "Both take the same input: one day of data generated by raw2l1 routine; . . ."

The final configuration of KABL does use only one parameter. But both algorithms still take the same input: a daily file generated by raw2l1 that contains all information needed for both algorithms, each algorithm extracts only what it needs.

CC: *line 624: "same input" changed for "same input file"*

4. Line 272: "number of invalid values (NaN or Inf) are recorded." - Please explain why algorithms return these kinds of values. Another question is how algorithms deal with undefined values in Lidar measurements.

Algorithms return NaNs when all the points of the profile are assigned to the same cluster. For ADABL, it happens when the profile is very different than the ones in the training set (not that rare). For KABL, it happens when we specified the initial centroids (it the case in the retained configuration) and only one of these points gather a cluster around it (very rare). When lidar has few undefined values in the profile, they are just ignored and the estimation is made on the available points.

CC: *Precision on BLH estimations at Nan or Inf is added at line 371.*

Specific (Minor) comments

5. Line 15: ". . . boundary layer height (BLH). . ." – please give somewhere a definition of the BHL.

BHL is probably a typo error for BLH, as a research for "BHL" inside the manuscript returned no results. The BLH definition we used here was in the next sentence (lines 15-16 of the submitted version) as "the depth of atmosphere where all pollutants emitted from the ground will remain". It was rephrased it place the definition before the acronym and improve the English.

CC: *lines 20 to 26 (marked-up version)*

6. Line 105: ". . . we chose to derive boundary layer height with parcel method for the 11:15 sounding and bulk Richardson number for the 23:15 one." Please justify why two different methods were used for morning and evening radiosounding.

We followed the recommendations in Seibert et al. (2000), figure 10, assuming that morning launch is in unstable atmosphere and evening launch is in stable atmosphere.

As the other referee had a major comment about the method used for RS, a more complete answer was given. It might be of interest here if any additional question arise.

CC: *The sentence at line 148 was changed for "Following recommendations from figure 10 of Seibert et al. (2000)".*

7. Line 113: Does false positives on cloud detection perturb a BLH detection? Please explain.

Cloud screening with the collocated CL31 was only used to exclude cases for the comparison with the radiosoundings. Therefore, false positives in cloud detection would have for only effect to improperly reduce the comparison sample. The MiniMPL detection of clouds was found to reduce too much the comparison sample, while CL31 detection of clouds looked more reliable.

CC: *use of cloud detection with the CL31 is clarified (lines 157-160).*

8. Line 113: In the following text, some basic ML concepts are introduced for readers, who are not familiar with the scope of ML. In this case, the "false positives" should also be explained or referenced.

The false positives refer here to the detection of clouds, not to the BLH detection.

To avoid confusion, the following sentences "Although the MiniMPL (...) to make some false positives." will be replaced by "Although the MiniMPL is perfectly capable of detecting clouds, we relied more on the cloud detection with the CL31, because MiniMPL's cloud detection was detecting cloud where there was not."

After the correction of the misleading use of words "false positives" and "algorithms" line 158 and 159, we do not think it is necessary to introduce false positive in section 3.

CC: *mention to false positive is removed (lines 157-160).*

9. Line 127: As the number of seconds, since midnight is a periodical function, the 'classical' distance could not take it in consideration correctly this variable. It means that the classical distance between one 00:01 and 23:59 will be nearly 24 hours. Please make sure, that ADABL algorithm works as expected in this case.

This remark is very important for the next stages of the algorithms development. However, at the moment, both algorithms have been used only on 24 hours chunks, so that the periodicity was not an issue.

CC: *precision added at lines 565-566.*

10. Line 142: I do not see any subsampling in figure 3. Is it a five-forks weak learner creation part? Please specify.

It is correct to see no subsampling in the figure 3: AdaBoost do not perform subsampling, as the referee points out in the next comment.

Figure 3 is only meant to illustrate the boosting algorithm (as figure 2 is only meant to illustrate decision tree). They do not reflect the actual settings of the algorithms, which would needlessly blur the pedagogical effort. The decision trees used as weak learner are five-forks trees or less.

CC: *captions of Figures 3 and 4 (line 199 and 198): "fake 2-dimensional data" was changed for "two-dimensional artificial data"*

11. Line 142: How these shallow decision trees are fitted? I have never heard about resampling in the classical AdaBoost. Is it Bagging? Please give a reference or explain the algorithm in detail.

The referee raises an error here: indeed the weak learners are not trained on a subsample but on the whole dataset. Although, the weight put on each sample changes from one weak learner to another. Weak learners are trained with CART algorithm. A very good description of the algorithm can be found in Hastie et al. (2009) page 307.

CC: *corrections are made at lines 196-198: "First a shallow decision tree is fitted on a random subsample of the dataset" was changed for "First a shallow decision tree is fitted to the entire dataset, using the*

Classification and Regression Tree (CART) algorithm (Hastie et al., 2009)"

12. Line 143: ". . .the error of the classifier is the number of misclassified points." - I am not sure that error is defined like that. Please explain or give a reference.

We acknowledge this sentence is incorrect: the error is the weighted average of misclassified points. For more details, the formula is given in Hastie et al. (2009) page 339, algorithm 10.1. The actual implementation can be checked in the Scikit-learn source code: sklearn.ensemble._weight_boosting.py: line 528.

CC: *line 198, add "All points have the same weight of this first step." and line 199 "number of misclassified points" changed for "weighted average of misclassified points"*

13. Line 146: The explanation is not sufficient. I propose to present here a reference to any popular textbook on AdaBoost or carefully introduce the algorithm. For example, in the expression, the performance was not introduced, the upper limit in the sum should be capitalized, etc.

The reference to Hastie et al. (2009), which is a popular textbook explaining the algorithm with many details and well written, was made at the section 3.1. It is completed by the reference to Freund & Schapire (1997), which is the original publication of AdaBoost and contains many theoretical results. The authors will complete theses references with Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference* (pp. 37-52). Springer, Berlin, Heidelberg.
This paragraph was kept short to avoid overwhelming the reader with technical details. We will reformulate in order to keep the main idea and refer to the literature for the details.

CC: *line 202 to 204: the text is shortened, unexploited formulae are removed and the references to Schapire (2013) and Hastie et al. (2009) are added.*

14. Line 170: "trade-off between accuracy and computing time" - I do not think that the limiting factor for this problem is the computing time. Normally this kind of problem could be sufficiently well resolved by parallel computing.

It is true that computing time is not critical here because it is always low, even without parallel computing. However, we do not want a classifier needlessly complex. For example, fully-grow trees would be long to train and test for very little extra-performance.

As computing time is not a major factor, we will change the sentence line 170 (submitted version)

CC: *lines 186-189 (marked-up version): "It was chosen by a trade-off between accuracy and computing time" was changed to "This configuration was chosen because more complex classifiers do not necessarily improve the performance."*

15. Line 169: "RCSco, RCScr" – please make sure that these names for copolarized and crosspolarized range-corrected backscatter signals persist in the following text (notably in tables 2 and 3).

RCSco and RCScr are named respectively RCS1 and RCS2 in the source code. We chose to use RCSco and

RCScr in the text because it is less ambiguous and to use RCS1 and RCS2 in the table so that readers going through the code would not be confused.
As this choice seems to be confusing for reader, we will keep only RCSco/cr in the paper and make the correspondence within the code.

CC: *all instances of "RCS1", "RCS0" or "RCS2" were removed (line 422 and in tables 2 and 3).*

16. Line 175: "It is possible to quantify the relative importance of the predictors (Breiman et al., 1984; Hastie et al., 2009). After the training, the time accounts for 30.3%, RCSco for 28.4%, RCScr for 26.5% and the altitude for 14.8%." – I have not found this information; could you please specify the corresponding page numbers?

In Hastie et al. (2009), it can be found page 368, equation 10.43.
In Breiman et al. (1984), it can be found page 147, definition 5.9.

CC: *no change.*

17. Line 184: "distances from all points to all centroids" – Are these the Euclidian distances?

Yes, K-means usually implies Euclidean distance. Although it is technically possible to use any distance, most of implementations do not provide this option.

CC: *no change.*

18. Line 196: "If we assume all Gaussian have the same fixed variance and that this variance tends to zero, EM and K-means algorithms are the same." – Could you provide a reference or explain the statement?

In Hastie et al. (2009), exercice 14.2, page 580.
Let's consider a sample generated by a mixture of two Gaussian with the same fixed variance. When the variance tends to zero, the "expectation" step is the same as the attribution to the closest centroid, and the "maximisation" step is the same as updating the centroids.

CC: *no change.*

19. Line 203: "Then the data they contain are normalized. . ." – if time and height are used in the KABL algorithm, are these variables also normalized?

Time and height are not used in KABL algorithm. However, all the data are normalized in order to avoid any unit comparison problem. The normalization consists in removing the mean and divide by the standard deviation. It is done in kabl.core.py: line 152 (release 1.0.0. Current commit: line 167).

CC: *no change*

20. Line 205: What values are included in predictors? If X matrix contains only signals RCSco and RCScr, it should be stated somewhere.

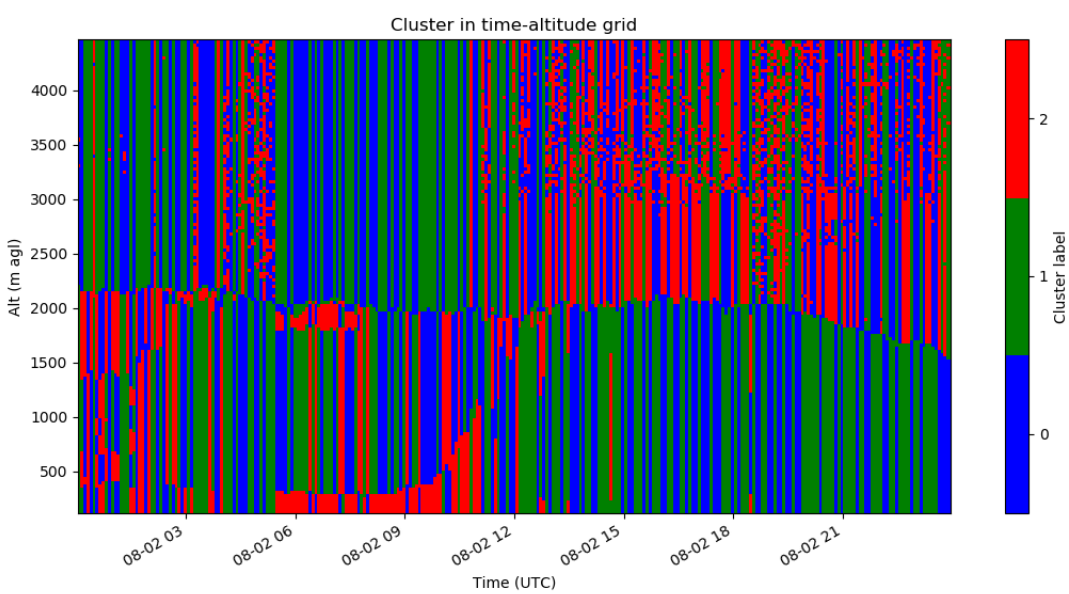
Predictors can be either RCSco all day, either RCSco and RCScr all day or RCSco for daytime and both RCSco and RCScr for nighttime, as precised in Table 2.
This question arises because of the ill-suited organisation of the text, as pointed out by comment 22, but this will be corrected.

CC: *paragraph 3.3 is reorganized (lines 275-325).*

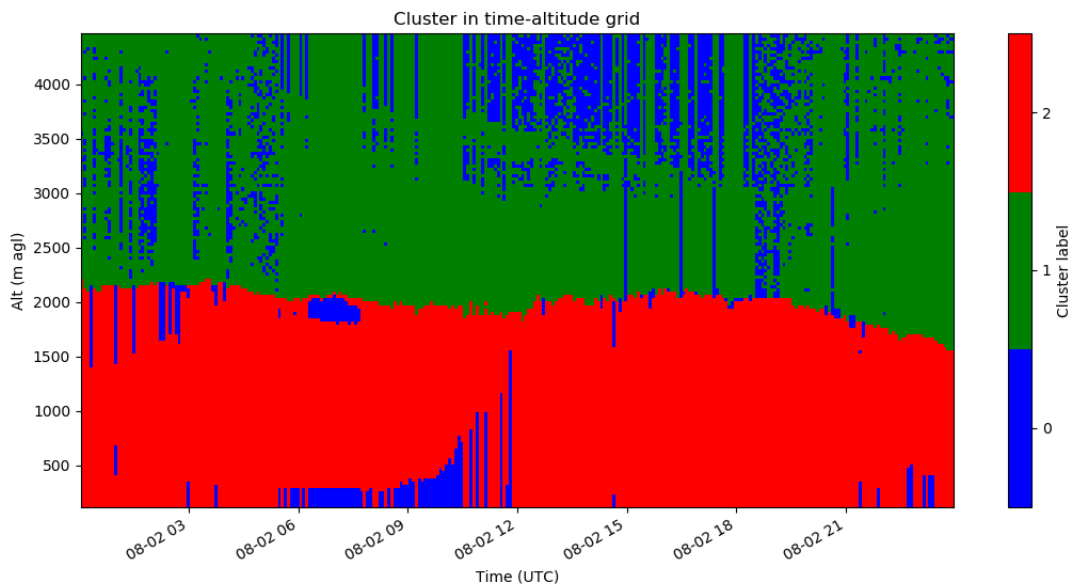
21. Line 209: "Finally, we look for the first change in clusters attribution, starting from the ground. This gives us the BLH for this profile." I am not sure that this algorithm is optimal, as it could lead to oscillations of BLH. To understand how it could be improved I suggest presenting and analyzing the altitude-time plot with pixels representing the results of the classification (like Fig 11 but with classes). Probably it is better not to take the first change, but a height above which the class is not changing, e.g. for three levels. Alternatively, a value of height could be selected that persist in time. These kinds of parameters could be optimally selected by the scores optimization. Another option is to modify the 'distance' definition.

We do observe some oscillations as the referee describes (e.g. see figure 11). The proposition made by the referee, to enforce vertical persistence of the clusters, is very relevant and will be added as a prospect. The time continuity is also very relevant, and existing methods already use such criteria, thus are a source of inspiration to implement it here. The distance definition is probably the "smartest" way, as it can help us learn about what really matters to distinguish boundary layer from the rest, but it is also the less intuitive. The optimization of parameters was done here thanks to global sensitivity analysis, but it would greatly benefit from being repeated after such new features are added.

The following figure shows the altitude-time plot of the classes attributed by KABL, with random initialization (K-means "vanilla"). It makes very visible the random attribution of the classes numbers in unsupervised classification: only borders matter.



A way to avoid such random attribution is to specify the initial centroids: the resulting plot is next. Initial centroids were put at typical backscatter values (modes of the histogram). The blue cluster has a very high backscatter (it detects cloud and shallow morning BL), the red cluster has high backscatter (it detects mixed layer or residual layer), the green cluster has low backscatter (it detects free atmosphere). We can see patches of blue in the free atmosphere: they are not realistic. They occur in profiles were there is no strong backscatter corresponding to this cluster. As we still ask for 3 clusters, the blue cluster sticks to noisy points.



The BL top defined as the height where cluster stop changing, as suggested by the referee, would be affected by such noisy points.

CC: *suggestions of the referees are added as prospects in the section 5, mainly in the new paragraph "5.2 Time and altitude continuity" (lines 550-559)*

22. Line 213: "The parameters of this computer code. . ." these parameters should be introduced at the beginning of the section 3.3, before they are referenced.

The organization of this paragraph will be changed accordingly.

CC: *paragraph 3.3 is reorganized (lines 275-325).*

23. Line 298: ". . . figure 8 the distribution (violin plots) of the relevant output conditionally to the parameter value." – I suggest adding here a reference on the construction of this kind of plot.

Here is a reference describing the plot and its use:
Hintze, J. L., & Nelson, R. D. (1998). *Violin Plots: A Box Plot-Density Trace Synergism. The American Statistician, 52(2), 181–184.*

CC: *reference to Hintze and Nelson (1998) is given at line 407.*

24. Line 304: "Parameters values are chosen to give the most optimal value for the metrics they have influence on." - The selection of locally optimal combination of parameters does not provide the globally optimal solution. How can you be sure that this combination gives the best precision?

This is a good remark: we cannot be sure. However, a sensitivity analysis on the 2-year long dataset would be computationally too expensive. The sensitivity analysis on a single day has the advantage to be more thorough than anything we could have done on the whole dataset. It helped us to selected few configurations to test on the whole dataset, including the configuration described in Table 3, which was elected because it had the best results. A sentence will be added to make clear that the sensitivity analysis on one day was used to select few configurations to be tried on two years.

CC: *clarifications added (lines 423-426).*

25. Line 316: "As the average gap E1 and the RMSE E2 are very similar. . ." – I suggest excluding the average gap E1 from the article for the sake of simplification.

Yes, this is something we will do in the next version of the manuscript.

CC: *all references to the average gap E1 are removed (lines 341, 344, 432, Table 1 and Figure 7).*

26. Line 326: "Nighttime (launch of 23:15 UTC)" - If nighttime radiosounding was not used, why to present this dataset in "2.2 Radiosonde data"? Probably it was used in supervised ML? Please specify.

Nighttime launches were not used in the supervised algorithm.
They were only used for the comparison to the diurnal BL cycles presented in Fig. 10.

CC: *no change.*

27. Figure 9: Adding the confidence intervals for RMSE and Correlation in Fig. 9 could be quite useful.

Yes it would. We will add bootstrap estimations of confidence intervals in the next version of the manuscript.

CC: *bootstrap estimations of confidence intervals were added in Figure 9 and its caption. The code to calculate the confidence intervals was also added to the public repository.*

28. Line 335 and Line 438: I think it would be advantages to understand how works the lidar manufacturer's software and to give some interpretations.

We asked the manufacturer for more details about their algorithm: a modified wavelet transform method described in Brooks, 2003 is used. In that regard, we will expand on the interpretations of the results for the revised manuscript.

CC: *manufacturer algorithm is precised and related comments are given (lines 467-470).*

29. Line 399: “A method to filter these oscillations will be needed, but it can also divest the "real-time" property.” – Instead of filtering, the criteria of the lowest transition of the class for KABL could be somehow modified, as I proposed in my comment for line 209.
The filtering could be of the "real-time" if it is done relatively the past classifications.

Yes, the comment 21 was full of good ideas to be put in the prospects. We will add the use of past classification to our answer to this comment.
As a matter of fact, the sensitivity analysis revealed that concatenating previous profiles do not solve these oscillations. Therefore, it is really the previous output of the classification that should be used in future filtering.

CC: *the question of oscillations is more thoroughly commented in the new paragraph “5.2 Time and altitude continuity” (line 550-559)*

30. Line 422: “5.6 KABL is "trainingless"” – I suggest that KABL could be used also by an expert to simplify the learning stage of supervised ML.

As we understand this comment, the referee suggests to make first an unsupervised classification (with KABL), correct it manually and then use it as a reference to train a supervised classifier (as ADABL). This a very interesting strategy to reduce the burden of supervision in ML methods, even beyond the only question of boundary layer height estimation. For example (still close to the topic), this method is currently under experimentation to make boundary layer classification: https://github.com/ThomasRieutord/bl-classification
However we would like to emphasis that the manual correction between unsupervised and supervised classification can hardly be by-passed. First, if the result of unsupervised learning can be used as a reference, why use a supervised model after? Second, unsupervised learning tells which classes are different but not which class is what. The identification of the classes must be done by a human expert or a reliable (physically-based) strategy.

CC: *no change.*

31. Line 417: “. . .strategies . . . for the training of ADABL. . .” – To decrease the sensitivity to “idealized” diurnal cycle of the BLH, I suggest trying to exclude the time predictor in ADABL.

Yes, this is an idea. Furthermore, a sensitivity analysis, as was done for KABL, would be very helpful to know better how to correctly set ADABL.

CC: *more general opening is added at line 577 “Further studies to determine how ADABL behaves without training and how it could be appropriately trained would be interesting.”*

Technical corrections

32. Line 13: “Atmospheric boundary layer concentrates many scientific challenges (small scale flows, turbulence...) and with high impacts due to its position of the interface between ground and atmosphere.” - awkward English, please correct.

CC: *beginning of introduction is rephrased (lines 20-26)*

33. Line 27: “(clouds, residual layers..)” -> “(clouds, residual layers. . .)”.

CC: *corrected*

34. Line 88: “SIRTA” - Please decrypt the abbreviation.

CC: *a footnote has been added.*

35. Line 92: “at 11:15 AM and PM” – Please utilize the same notation for the time here and further. I suggest the UTC format.

CC: *all time reference have been verified and are in UTC format.*

36. Line 94: Please explain what the theta is. Is it the potential temperature?

CC: *corrected (line 136)*

37. Line 127: I suggest inserting a comma after “height above ground”.

CC: *corrected*

38. Line 129: “[[1,N]]”- What does double brackets means? Please explain.

CC: *explained at line 181.*

39. Line 145: “tree” -> “trees”.

CC: *corrected at line 202.*

40. Line 146: “m=200” -> “M=200” + upper case in sum limit

CC: *corrected at line 202.*

41. Line 154, 157, fig. 4: “top”->”left”, “bottom”→”right”.

CC: *corrected*

42. Line 208: Init parameter is not defined.

CC: *paragraph 3.3 is reorganized (lines 223-272)*

43. Line 208: “specified in algo” -> “specified by algo parameter”

CC: *paragraph 3.3 is reorganized (lines 275-325)*

44. Line 270: “(0.20)” Is it the software version? Please specify.

CC: *line 368: “(>=0.20)” changed to “(version >=20)”*

45. Figure 10: Please introduce the INDUS abbreviation.

CC: *INDUS abbreviation is explained in the caption of Figure 10*

Mixing height derivation from aerosol lidar using machine learning: KABL and ADABL algorithms

Thomas Rieutord[1], Sylvain Aubert[2], and Tiago Machado[1,2]

[1]Centre National de Recherches Meteorologiques, Université de Toulouse, Météo-France, CNRS, Toulouse, France
[2]Direction des Systèmes d'Observation, Météo-France, Toulouse, France

Correspondence: Thomas RIEUTORD (thomas.rieutord@meteo.fr)

Abstract. ~~Atmospheric~~ The atmospheric boundary layer height (BLH) is a key parameter ~~for~~ in several meteorological applications ~~, for example air quality forecast. To measure it, a common practice is to use aerosol lidars:~~ such as air quality forecasts. A common method to measure BLH is via aerosol lidar, where a strong decrease in the backscatter signal indicates the top of the boundary layer. This paper describes and compares two ~~methods of machine learning to derive the BLH from backscatter profiles: the~~ machine learning methods, the K-means algorithm and the AdaBoost algorithm ~~. Their codes are available under a fully open access, with the name KABL (~~to derive the BLH from backscatter profiles The K-means for ~~Atmospheric Boundary Layer) and ADABL (~~Amtospheric Boundary Layer (KABL) and AdaBoost for Atmospheric Boundary Layer ~~)~~(ADABL) algorithm codes used in this study are free and open source. Both methods are compared to the lidar manufacturer's ~~software~~ algorithm and to reference ~~BLH derived from collocated radiosondes. The radiosondes are taken~~ BLHs derived from colocated radiosonde data. The radiosonde data were used as the reference for all ~~other methods. The comparison is carried out on~~ methods. A comparison was carried for a two-year ~~long~~ period (2017-2018) ~~on 2 Meteo-France's~~ for two Météo-France operational network sites (Trappes and Brest). ~~Results show that, although its training is limited, ADABL is performing better than KABL and can easily be improved by enhancing its training set. However, KABL can be easily adapted for other instrumental device and used to make instrument synergy, while ADABL must be fully re-trained at each change in the instrument settings~~A large discrepancy in the results was observed between the two sites. At the Trappes site, KABL and ADABL outperformed the manufacturer's algorithm, while the performance was clearly reversed at the Brest site. We conclude that ADABL is a promising algorithm but has training issues that need to be resolved, KABL has a lower performance than ADABL but is much more versatile, and the manufacturer algorithm is performing well with little tuning but is not open-source.

1 Introduction

~~Atmospheric boundary layer concentrates many scientific challenges (small scale flows, turbulence. ..) and with high impacts due to its position of interface between ground and atmosphere.For example, air quality forecasts rely on many meteorological parameters, and among them~~ The atmospheric boundary layer is the lowest part of the troposphere that is directly influenced by surface forcings. As a high-impact area where most human activities take place, all pollutants emitted from the ground are diluted in this layer. The key parameter used to model this dilution is the depth of this layer, i.e., the boundary layer height

(BLH)~~is of first importance. Indeed, the BLH is the depth of atmosphere where all pollutants emitted from the ground will remain. As it varies from a few tenth~~ . Because BLH can vary from a few tens meters to about 2 km within a single day, the ~~dilution/concentration can be very important and responsible for~~ amount of dilution can vary considerably and result in air quality warnings (Stull, 1988; Dupont et al., 2016). ~~Beside, it~~ In adddition, BLH is one of the largest ~~source~~ sources of uncertainty in air quality ~~model~~ models (Mohan et al., 2011) and there is a need ~~of better evaluation of~~ to better evaluate this parameter (Arciszewska and McClatchey, 2001). In numerical weather prediction models, physical processes ~~are not the same inside the~~ change within the boundary layer (Seity et al., 2011). Therefore it is ~~worth~~ important to compare BLH ~~from models and BLH~~ calculated in models with that derived from measurements.

However, measuring ~~the boundary layer height~~ BLH is not straightforward. As stated in Seibert et al. (2000), there ~~is no system matching~~ are no systems that match all the requirements to make ~~a reliable estimation of BLH . Best~~ reliable BLH estimations. The best BLH estimation can be achieved ~~through instrument synergy~~via the synergetic use of multiple instruments. However, adding instruments limits the ~~eligible sites where the estimation~~ number of sites where estimations can be made. In this paper we ~~choose to~~ focus on a single instrument, ~~the~~ aerosol lidar (see ~~section~~ Sect. 2.1.1 for more information), already widely used (Haeffelin et al., 2012). The boundary layer is detected by ~~a~~ the decrease of the lidar signal at its top. ~~But~~

However this decrease can be blurred ~~,~~ or perturbed by other strong signals (e.g., clouds, residual layers~~.) and numerical artifacts can occur~~, and small scale structures) and instrumental noise. For these reasons, ~~there exists numerous studies on BLH derivation from aerosol lidars. Melfi et al. (1985) make~~ numerous studies exist concerning the derivation of BLH from aerosol lidar. Melfi et al. (1985) use a simple thresholding of the signal. Others methods are based on derivatives~~:~~ . For example, Hayden et al. (1997) take the minimum of the gradient~~.~~ , Menut et al. (1999) use the height ~~of zeroing~~ where the second derivative ~~(~~is zero (the inflection point) ~~and also~~ as well as the variance of the signal~~.~~ , Senff et al. (1996) use the derivative of the logarithm of the ~~back-scattered~~ backscattered intensity along the height. One of the most used ~~method~~ methods is the wavelet covariance transform ~~(WCT): it looks~~ which searches for the maximum in the convolution between the signal profile and a Haar wavelet (Gamage and Hagelberg, 1993; Cohn and Angevine, 2000; Brooks, 2003). More recent studies ~~are~~ have been based on backscatter signal analysis ~~, like STRAT (Morille et al., 2007) or~~ such as STRAT (Morille et al., 2007) and CABAM (Kotthaus and Grimmond, 2018). ~~Other studies use graph theory~~ Graph theory has also been used to impose continuity constraints (both vertically and in time) in ~~the BLH estimation:~~ BLH estimations, e.g., Pathfinder (De Bruine et al., 2017). Inspired by image processing, some methods use Canny edge detection in addition to backscatter signal analysis (Morille et al., 2007; Haeffelin et al., 2012). STRAT and Pathfinder have ~~been merge~~ also been merged into PathfinderTURB by Poltera et al. (2017). ~~All this literature shows that finding~~ These studies demonstrate that deriving BLH from aerosol lidar is still an open ~~question~~area of research.

~~Furthermore~~In addition, artificial intelligence (AI) has ~~reborn~~ reemerged in the last decade ~~due to the concomitant increase~~ because of the simultaneous increase in the amount of available data and computational power. Both ~~reached levels allowing applications that were not possible before. AI has shown some ability to tackle~~ have reached levels that enable previously impossible applications. AI is capable of tackling complex classification problems, especially in image classification (Krizhevsky et al., 2012). Such breakthroughs were ~~done thanks to~~ made possible by deep convolutional neural networks (LeCun et al.,

2015)~~, but AI encompasses much more~~ ; however, AI encompasses many other techniques that also benefit from larger data and increased computational power (Besse et al., 2018). In this paper~~we will see how the BLH derivation from backscatter profile~~ , we explore how the derivation of BLH from backscatter profiles can be formulated as a classification problem and ~~we will apply appropriate algorithms to solve it. Toledo et al. (2014)~~ how appropriate algorithms can be applied to solve this problem. Toledo et al. (2014) have already described a method that falls into the ~~AI scope . It uses~~ scope of AI. They used unsupervised learning to classify whether the measurement points were within the ~~measure points in or out of the~~ boundary layer. This ~~idea~~ method has yielded convincing results in ~~past studies (Toledo et al., 2017; Rieutord et al., 2014). It has been pursued here (under the name of KABL) , and explored more carefully~~previous studies (Toledo et al., 2017; Caicedo et al., 2017; Rieutord et al., 2014) and is pursued here with the K-means for Atmospheric Boundary Layer (KABL) algorithm. KABL was extensively tested and is shared via an open-source code. In addition, ~~an alternative~~ we test an alternative adaptive boosting (AdaBoost) machine learning algorithm~~(named ADABL) was tested. As KABL, it classifies measure points in or out~~ , the AdaBoost for Atmospheric Boundary Layer (ADABL) algorithm. Both algorithms classify whether the measurement points are inside or outside of the boundary layer~~, but it~~ ; however, ADABL learns the characteristics of both groups from a training set. The training set consists of atmospheric boundary layer identifications made by human experts, which is acknowledged as being more reliable than available automatic methods (Seibert et al., 2000). Such supervised algorithms make it possible to automatically reproduce human expertise in boundary layer identification. To our knowledge, this is the first time that ~~boosting algorithms are~~ a supervised algorithm has been applied to this problem. This study is of practical interest because it includes the publication of the source code, which only uses free software.

~~First, in section~~ In Sect. 2, we ~~state which data have been~~ describe the data used in this study~~:~~ , i.e., the lidar data in ~~input of algorithms, radiosounding data as reference , ancillary data to sort~~ the algorithm inputs, reference radiosonde data, and ancillary data used to sort the meteorological conditions. ~~Next, in section~~ In Sect. 3, we ~~described~~ describe the two machine learning algorithms (KABL and ADABL) and the ~~procedure~~ procedures used to evaluate them. ~~Then, in section~~ In Sect. 4, we present the results of our study~~:~~ , which consists of a sensitivity analysis of the KABL algorithm, a comparison of ~~methods against radiosondes for a 2 years periodand case studies. Finally, in section 5, a discussion about~~ the methods with the radiosonde data over a two-year period, and a case study. In Sect. 5, we discuss the results, ~~the limitations and the prospects of this studyis proposed. Last~~ limitations and prospects of our study. The final section is dedicated to ~~take-home conclusions~~ the conclusions that can be drawn from our study.

2 Material

~~In this study , the data comes from Meteo-France's~~ Our study used data from Météo-France operational network. We used ~~collocated radiosoundings~~ colocated radiosonde and aerosol lidar data over two sites: Brest (a coastal city in ~~extreme West~~ an extreme Western region of France) and Trappes (a sub-urban area of Paris ~~, inland~~ in an inland region of Frane). The dataset ~~has a span of~~ spanned two years: 2017 and 2018. A case study ~~is taken on the 2nd of August~~ was taken conducted on August 2, 2018~~in Trappes~~ , for the Trappes site.

Figure 1. Typical MiniMPL unit from ~~Meteo-France~~ Météo-France network

2.1 Lidar data

2.1.1 Lidar network

~~Starting in~~ Since 2016, ~~Meteo-France~~ Météo-France has deployed a network of ~~6~~ six automatic backscatter lidars to help the Volcanic Ash Advisory Center ~~(VAAC)~~ of Toulouse characterize volcanic ash and ~~aerosols~~ aerosol layers. One of the six sensor can be quickly redeployed ~~on~~ at a more suitable geographic location depending on the transport event ~~to follow~~being tracked. The network, fully operational since April 2017, is ~~functioning continuously and has been able to detect aerosol events up to an altitude of~~ continuously functioning and has detected aerosol events at an altitude up to 17 km. It is part of the wider ~~Automatic Lidars and Ceilometers (ALC)~~ automatic lidar and ceilometers network of the E-PROFILE program described in Haefele et al. (2016).

Two sampling sites ~~of this network have been~~ in this network were selected: Brest (48.444° N, 4.412° W, 94 m ~~a.s.l~~above sea level) and Trappes (48.773 N, 2.0124 E, 166 m ~~a.s.l). Each site is~~ above sea level). Both sites are equipped with a Mini Micro Pulse LiDAR (MiniMPL), built by Sigma Space Corporation ~~; the exterior casing was~~ with an exterior casing provided by Envicontrol. A typical MiniMPL unit from ~~Meteo-France~~ Météo-France network is shown in figure 1. ~~The~~ MiniMPL is a compact version of the standard ~~MPL~~ micro pulse lidar systems deployed in the NASA global lidar network ~~(MPLNET)~~MPLNET. A comprehensive description of ~~the~~ MiniMPL can be found in Ware et al. (2016).

2.1.2 Data processing

~~The miniMPL~~ MiniMPL acquires profiles of atmospheric backscattering at high frequency (2500 Hz) using a low energy pulse (3.5 μJ) emitted by a Nd:YAG laser at 532 nm. The profiles are acquired in photon-counting mode and, in our present configuration, averaged over 5 ~~minutes~~ min and 30 ~~meters~~ m vertical resolution bins. The instrument uses a monostatic coaxial design ~~:~~ where the laser beam and the receiver optics share the same axis. ~~Due to~~ Because of geometrical limitations, only a fraction of the signal can be recovered in the near field. Therefore, in our system, the first usable data are ~~provided~~ available at 120 ~~meters~~ m above ground level~~on our system~~.

The instrument has polarization capabilities with the collection of photons on two different channels (~~more details in Flynn et al. (2007)):~~ for more details see Flynn et al. (2007)); the measured raw signals on the "copolarized" and "crosspolarized" ~~channel, respectively suffixed~~ channels are respectively co and cr suffixed. These raw signals are ~~then~~ processed to obtain the ~~quantities of interest:~~ quantity of interest, i.e., the range corrected signal RCS, also called the normalized relative backscatter~~, NRB~~. This industrial processing ~~comprises several procedures such as~~ consists of several procedures including background, overlap, afterpulse, and dead-time corrections. A comprehensive description of ~~this~~ the processing is given in Campbell et al. (2002). ~~Finally the~~ The "copolarized" and "crosspolarized" range corrected ~~signal~~ signals, RCS_{co} and RCS_{cr}, respectively, as delivered by the industrial software, ~~respectively RCS_{co} and RCS_{cr} will be~~ are used as predictors for the machine learning algorithms described in ~~part~~ Sect. 3.

~~Raw~~ The raw data type and format depends on the instrumental device used. To make the algorithms usable on other devices, we ~~used as input of the algorithms the files processed to a normalised format by~~ converted the files to a normalized format using the *raw2l1* routine and then used these files as the algorithm input. *raw2l1* ~~is~~ was developed by the ~~SIRTA and publicly available here:~~ Site Instrumental de Recherche par Télédétection Atmosphérique and is publicly available[1]

2.2 Radiosonde data

The algorithms ~~are evaluated against radiosounding estimations. Meteo-France operates several radiosounding~~ were evaluated with respect to radiosonde (RS) estimations. Météo-France operates several RS sites for the WMO Global Observing System. Two ~~radiosoundings sites are collocated~~ RS sites are colocated with the lidars of Brest and Trappes. They are equipped with a Meteomodem robotsonde and typically launch a Meteomodem M10 sonde at 11:15 ~~AM and PM~~ UTC and 23:15 UTC every day.

Many methods ~~exists~~ exist to derive BLH from ~~radiosondes, and have been variously~~ RS data, and several of which have been used in the literature. Some of these methods are listed below.

- Parcel method: BLH is the height at which the ~~θ profile~~ profile of the potential temperature θ reaches its ground value.
- Humidity gradient method: BLH is the height at which the gradient of the relative humidity is strongly negative.
- Bulk Richardson number method: BLH is the height at which the bulk Richardson number exceeds 0.25 (~~threshold depending on authors~~this threshold varies among studies).

[1] https://gitlab.in2p3.fr/ipsl/sirta/raw2l1

- Surface-based inversion: BLH is the height at which the gradient temperature profile reaches zero.

- Stable layer inversion: BLH is the height at which the gradient of the potential temperature profile reaches zero.

Hennemuth and Lammert (2006) ~~use the parcel method and the~~ used the parcel and humidity gradient methods. Collaud Coen et al. (2014) ~~use~~ uses all the techniques mentioned above and recommend the ~~Bulk~~ bulk Richardson number method for all cases. Guo et al. (2016) ~~use~~ used the bulk Richardson number for a ~~2 years~~ two-year climatology. Seidel et al. (2010) ~~compare the parcelmethod~~compared the parcel, the humidity gradient~~method~~, the surface-based inversion methods and other methods over a period of 10 years ~~and~~ at 505 sites ~~over the world. Seidel et al. (2012) compare many methods and recommend~~ worldwide. Seidel et al. (2012) compared several methods and recommended the bulk Richardson number method.

~~After testing some of these methods on our dataset~~Following the recommendations in Figure 10 of Seibert et al. (2000), we chose to derive ~~boundary layer height with~~ BLH using the parcel method for the 11:15 UTC sounding and bulk Richardson number for the 23:15 ~~one~~UTC sounding.

2.3 Ancillary data

Ancillary data ~~have been used for the description of the meteorological situation. They are~~ were used to decribe the meteorological situations at the observation sites. These data were not used by the machine learning algorithms. All the instruments ~~are collocated with lidar and radiosoundings~~were colocated with the lidar and radiosonde launchings.

- Rain gauges were used to detect rain events.

- Vaisala Ceilometer CL31 ~~ceilometers~~ instruments were used to detect the cloud base height ~~,~~ and distinguish cases with ~~cloud~~ clouds on top or inside the boundary layer. ~~Although~~ Even though the MiniMPL is ~~perfectly~~ capable of detecting clouds, we ~~chose not to use the industrial algorithm . The algorithm has shown in our experience to make some false positives~~relied on the cloud detection with the CL31, because MiniMPL algorithm was found to report non-existent clouds.

- Scatterometers were used to estimate the visibility and detect ~~fog cases~~the occurrence of fog.

3 Machine learning methods

~~Machine learning techniques are separated in two wide~~ Machine-learning techniques are categorized into two broad families: supervised learning (~~mimic~~ mimicking a reliable reference) and unsupervised learning (~~learn without~~ learning without a reference) (Hastie et al., 2009). First, we present the supervised algorithm leading to ADABL. ~~Second we present~~ Then, we present the unsupervised learning leading to KABL.

3.1 Supervised learning method

Supervised methods learn from a reference. ~~They are divided in~~ Such methods are divided into two families: classification~~(~~, which aims to find the frontiers between groups~~) and regression(~~, and regression, which aims to approximate a function~~)~~. In this ~~work, we consider the boundary layer height~~ study, we treat the BLH derivation as a classification problem. ~~From all points measured by the lidar , which are in the boundary layerand which are not?~~ We need to classify the measurement points of the lidar into two classes: 'boundary layer' or 'free atmosphere'. Then, the highest point of the ~~boundary layer class is the BLH~~ 'boundary layer' class indicates the BLH estimate. Boosting algorithms are a very powerful family of algorithms, ~~initially made~~ that were developed for classification but ~~they~~ can also be used for regression (Hastie et al., 2009). ~~AdaBoost (Adaptive Boosting) algorithm is well~~ In particular, the AdaBoost algorithm is designed for binary classification (Freund and Schapire, 1997), ~~thus it is the one we used~~and is therefore well suited to our problem.

3.1.1 AdaBoost algorithm

Let us consider the following problem~~: we~~. We have N vectors $x_i \in \mathbb{R}^p$ (here $p = 4$: seconds since midnight, height above ground~~copolarized channel, crosspolarized channel)~~, copolarized channel and cross-polarized channel), and for each vector, we have a binary indicator $y_i \in \{-1, 1\}$ (~~-1 for boundary layer, 1 for free atmosphere~~-1 for 'boundary layer', 1 for 'free atmosphere'). From the sample $(x_i, y_i)_{i \in [\![1,N]\!]}$, where $[\![1, N]\!]$ is the ensemble of integers from 1 to N, we want to predict the output indicator y_{new} of any new vector x_{new}. To do so, we must find a rule based on the x_{new} ~~coordinates~~ coordinate values (the features) to cast it into the appropriate class. Decision tree classifiers (Breiman et al., 1984) ~~do~~ perform this casting one feature at a time. For example, in ~~the figure~~ Figure 2, there are black ~~points~~ and white points in a ~~2-dimensional space. Black~~ two-dimensional space. The black points are mostly located where X_1 is low, hence the rule "if $X_1 < t_1$, then ~~it is black".But~~ the point is black." However, in the other region, where $X_1 > t_1$, there are still some black points, all with low X_2. Therefore, ~~at the output of the following rule,~~ we add the rule "if $X_2 < t_2$, then ~~it~~ the point is black, else it is white." Decision trees are classifiers made up of such "if" statements ~~,~~ with various depths and thresholds. The deeper the tree, the more accurate the border ~~,~~ but the more complex the decision and the longer it ~~is~~ takes to train. ~~As a matter of fact, deep~~ Deep trees are strongly subject to overfitting and ~~they~~ are less efficient than other methods. However, shallow decision trees are valuable because of their simplicity and their speed, even though their ~~performance~~ performances are quite limited (Hastie et al., 2009). They are often used as ~~a~~ *weak* ~~*learner*~~*learners*~~: that isto say~~, that is, classifiers with poor ~~performances (although still~~ (but better than random) ~~but~~ performances and are very simple (Freund and Schapire, 1997). In this study, weak learners in AdaBoost are trees with a maximum depth of ~~5 (maximum 5 forks between root an~~ five (a maximum of five forks between the root and leaves).

AdaBoost is based on decision tree classifiers. It aggregates ~~them in order to have~~ these classifiers to determine the most accurate border. The ~~idea of~~ concept behind AdaBoost is illustrated ~~by figure~~ in Figure 3. First, a shallow decision tree is fitted ~~on a random subsample of the dataset. Some points of~~ to the entire dataset using the Classification and Regression Tree (CART) algorithm (Hastie et al., 2009). All points have the same weight in this first step. Some points in the dataset are misclassified~~:~~ , and the error of the classifier is the ~~number of~~ weighted average of the misclassified points. Another shallow decision tree is

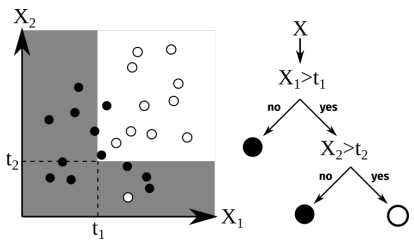


Figure 2. Illustration of binary classification with decision trees on ~~fake 2-dimensional~~ two-dimensional artificial data.

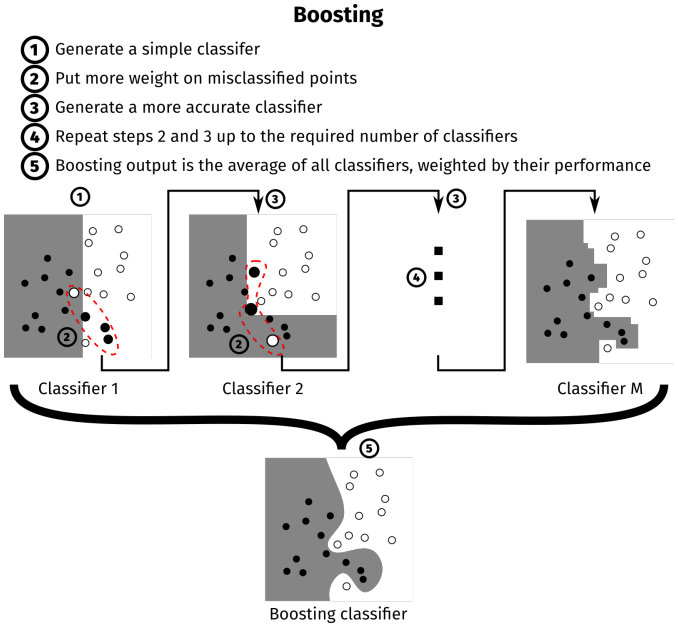


Figure 3. Illustration of boosting on ~~fake 2-dimensional~~ two-dimensional artificial data ~~and 2~~ with two classes.

then fitted on a ~~subsample of the~~ resampled dataset where the previously misclassified points are over-represented. This new tree has new misclassified points, that will be over-represented in the training of the next tree, and so on, up to the specified number of ~~tree ($m = 200$~~ trees ($M = 200$ in our case). The ~~classification given in output of AdaBoost is the average $\hat{y}$ of all the predicted class by the trees $\{\hat{y}_m\}_{m\in[\![1,M]\!]}$, weighted by their performance $\{\hat{\alpha}_m\}_{m\in[\![1,M]\!]}$: $\hat{y} = sign(\sum_{m=1}^{m} \hat{\alpha}_m \hat{y}_m)$~~ detailed algorithm is described in Hastie et al. (2009), algorithm 10.1, and in Schapire (2013).

3.1.2 Training of the algorithm

Such ~~algorithm must be trained from~~ an algorithm needs to be trained using a trustworthy reference. ~~For few~~ On days where the boundary layer is easily visible for a human expert, the ~~boundary layer top is~~ top of the boundary layer can be drawn by

hand~~:~~ ; all points below this limit are in the class ~~"boundary layer",~~ ‘boundary layer’ and all points above this limit are in the class ~~"free atmosphere"~~‘free atmosphere’.

~~Two~~ In this study, two days were labelled by hand. These two days where chosen because the boundary ~~layer is quite visibleand they are in different site~~ layers on these days were easily visible; the two labelled days were at different sites at different seasons. The first labelled day ~~is a clear day of summer at~~ was a clear summer day in Trappes, shown ~~on figure 4 (top):~~ in Figure 4 (left); a stable boundary layer is present near the ground ~~at~~ during the night, topped by a residual layer and few clouds between 02:00 UTC and 04:00 UTC. ~~The mixed layer starts developing at 9~~A mixed layer started to develop at 09:00 UTC and ~~stays around~~ remained approximately 2000 ~~meters~~ m for the rest of the day. ~~Around~~ At approximately 22:00 UTC, a new stable layer ~~seems~~ appeared to develop near the ground~~but~~ ; however, it is not very clear where ~~it starts and what is its extent~~ this layer started and what its extent was. The second labelled day ~~is a clear day of winter at~~ was a clear winter day in Brest, shown ~~on figure 4 (bottom~~in Figure 4 (right): a stable boundary layer ~~is~~ was present near the ground ~~at~~ during the night, topped by a residual layer, ~~shallower than at Trappes~~which was shallower than what was observed at the Trappes site. The mixed layer ~~starts developing at 9~~started to develop at 09:00 UTC and ~~stays around~~ remained at approximately 1000 ~~meters and decreasing along~~ m with the height of the layer gradually decreasing throughout the day. ~~Around~~ At approximately 17:00 UTC, aerosols ~~seems~~ appeared to accumulate in a thin layer close to the ground, therefore we ~~choose to drop the BLH at~~ chose to drop BLH to that level.

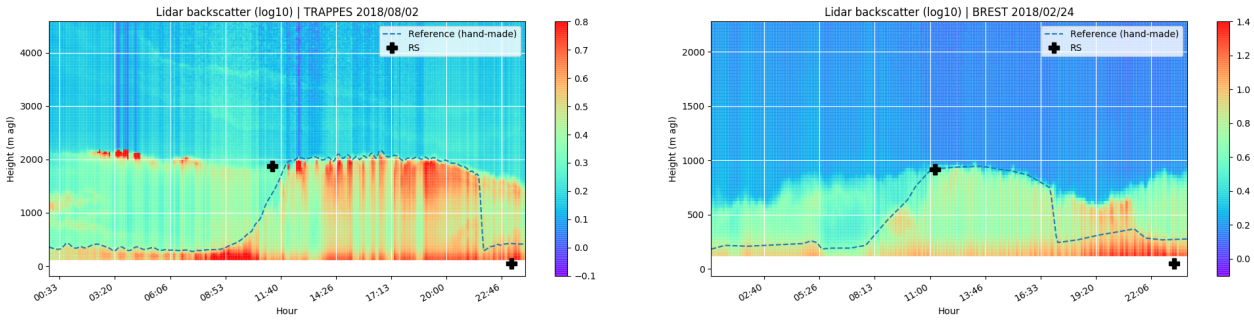


Figure 4. ~~Hand-made reference~~ Hand-drawn references and ~~RS estimation over~~ radiosonde estimates overlaying the lidar range-corrected intensity signal for two days: ~~2nd of~~ August 2, 2018, at the Trappes site (~~top~~left) and ~~24th of~~ February 24, 2018, at the Brest site (~~bottom~~right).

The ~~line of BLH made by hand is then loaded thanks to VGG Image Annotator software~~[2] ~~to draw the BLH by hand and get the coordinates of the curve's points~~coordinates of the points on the curves of the hand-drawn BLHs were obtained using the VGG Image Annotator software[2]. Then, the output ~~curve is interpolated with~~ curves were interpolated with a cubic spline to ~~meet~~ match the lidar temporal resolution. Given the resolution of the lidar, this ~~way~~ method of labelling the data ~~gives~~ results in $N = 86400$ individuals in total.

[2]~~Publicly available online following this URL:~~

[2]Publicly available online at https://www.robots.ox.ac.uk/~vgg/software/via/via-1.0.6.html.

3.1.3 Retained configuration

Four predictors ~~are~~ were used: the two lidar channels, ~~the~~ time (number of seconds since midnight) and ~~the~~ altitude (meters above ground level). The ~~current configuration of ADABL is thus the following:~~ ADABL configuration used was

- Weak learner: decision tree of depth ~~5~~ five;
- Number of weak learners: 200; and
- Predictors: time, altitude, RCS_{co}, and RCS_{cr}.

~~It was chosen by a trade-off between accuracy and computing time. The accuracy is~~ This configuration was chosen because more complex classifiers do not necessarily improve the performance. The computation time of the algorithm was still reasonable: training took 23s on the full dataset and predicting BLH for a full day took 3.7s. AdaBoost was chosen after a comparison of multiple classification algorithms, i.e., random forests, nearest neighbors, decision trees, and label spreading (study not shown here). The benchmark score was the accuracy as measured by the percentage of individuals that were well classified. ~~It is estimated by cross-validation (random split, 80% training set, 20% testing set) and reaches 99.5% of the testing test. The computing is still reasonable: it takes 23 seconds to train on the full dataset and 3.7 seconds to predict the BLH for a full day.~~ The accuracy was estimated by group K-fold, where labelled data-sets are grouped into chunks of four consecutive hours, one group was used as a testing set and all the rest as a training set. This operation was repeated until each group was used as the testing set. The resulting accuracy was 96%. However, this figure overestimates the generalization ability of AdaBoost. A more correct estimation would be obtained with an independent validation set (e.g. a new labelled day). An independent validation set was not used here because the accuracy was only used to discriminate between the classification algorithms.

It is possible to quantify the relative importance of the predictors (Breiman et al., 1984; Hastie et al., 2009). After ~~the~~ training, the ~~time accounts for~~ relative importance of the time, RCS_{co}, RCS_{cr}, and altitude predictors was 30.3%, ~~RCS_{co} for~~ 28.4%, ~~RCS_{cr} for~~ 26.5%~~and the altitude for~~ , and 14.8%, respectively.

3.2 Unsupervised learning methods

Unsupervised methods aim to ~~find groups in~~ identify groups in the data. In our case, we want to identify the group ~~"boundary layer". The boundary layer height~~ 'boundary layer'. The BLH estimate is then the ~~border~~ upper boundary of this group. Two unsupervised learning algorithms ~~have been~~ were tested: K-means and ~~Expectation-Maximisation~~ expectation-maximisation (EM).

3.2.1 K-means algorithm

The K-means algorithm is a well proven and commonly used algorithm ~~to make~~ for data segmentation (Jain et al., 1999; Pollard et al., 1981) ~~. The algorithm has 3 steps:~~ and consists of three steps.

1. ~~Initialisation~~Initialization: K centroids $m_1, ..., m_K$ are initialized at random ~~places~~ locations inside the feature space.

2. Attribution: The distances from all points to all centroids $(d(x_i, m_k))_{k\in[\![1,K]\!],i\in[\![1,N]\!]}$ are computed, and points are attributed to the closest centroid:
   $C(i) = arg\min_k \{d(x_i, m_k)\}$.
3. Update: The centroids are re-defined as the average point of the cluster: $m_k = \frac{\sum_{i=1}^{N} x_i \mathbf{1}_{C(i)=k}}{\sum_{i=1}^{N} \mathbf{1}_{C(i)=k}}$

Steps 2 and 3 are repeated until the centroids stop moving. It has been shown that this algorithm converges to a local minimum of the intra-cluster variance Selim and Ismail (1984). ~~The figure~~ Figure 5 (left) illustrates ~~it~~this algorithm.

3.2.2 ~~Expectation-Maximisation~~ EM algorithm

The ~~Expectation-Maximisation algorithm addresses the classification when the groups are Gaussian. It assumes~~ EM algorithm assumes that each group $k \in [\![1, K]\!]$ is generated by a Gaussian distribution (μ_k, Σ_k). The algorithm ~~estimates iteratively~~ iteratively estimates the parameters $\hat{\mu}_k$, $\hat{\Sigma}_k$ and the *responsibility* for each Gaussian $\hat{\gamma}_k^i$~~(~~, where the *responsibility* is the probability ~~for~~ of the point x^i ~~to be~~ being generated by the k-th Gaussian~~)~~. Points are then attributed to the group with the highest responsibility: $C(i) = arg\max_k(\hat{\gamma}_1^i, ..., \hat{\gamma}_K^i)$ ~~The figure~~ Figure 5 (right) illustrates ~~it~~this algorithm.

The K-means and EM ~~algorithm are quite similar: if we assume all Gaussian~~ algorithms are very similar. If we assume that all Gaussians have the same fixed variance and that this variance tends to zero, the EM and K-means algorithms are the same. However, K-means does not rely on ~~the Gaussian assumptionof the groups~~a Gaussian assumption.

3.3 KABL flowchart

~~The simplified flowchart of KABL is shown in figure 6. A netCDF file generated by the *raw2l1* software must be provided as input data of the KABL code. The data, namely: the vector of altitude z (size N_z), the vector of time t (size N_t), the range-corrected signals RCS_{co} and RCS_{cr} (matrices of shape $N_t \times N_z$), are extracted from this file. Such data are prepared to fulfil machine learning algorithms requirements. For each time, **n_profiles** last profiles are extracted. Then the data they contain are normalised (remove mean and divide by standard deviation) and this provides a matrix X (shape $N \times p$) with $N =$**n_profiles**$\times N_z$ and $p = |$**predictors**$|$ (number of elements in the list). The matrix X is the usual input for machine learning algorithm: it has one line per individual observation and one column for each variable (or predictor) observed. For the need of BLH retrieval, the preparation provides also a vector Z (size N) with the altitude of each individual observation. The algorithm (either K-means or EM, the one specified in **algo**) is applied on the matrix X, with the parameters **n_clusters**, **init** and **n_inits**. It provides a vector of *labels* (size N) which contains the cluster attribution of each individual. Finally, we look for the first change in cluster attribution, starting from the ground. This gives us the BLH for this profile. These operations are repeated up to the end of the netCDF file.~~

~~Simplified flowchart of KABL computer code~~

~~The parameters of this computer code were in bold font in the text and they are~~ The parameters of the KABL software are detailed here:

- **algo:** ~~the~~ The applied machine learning algorithm~~that will be applied~~. Possible values are

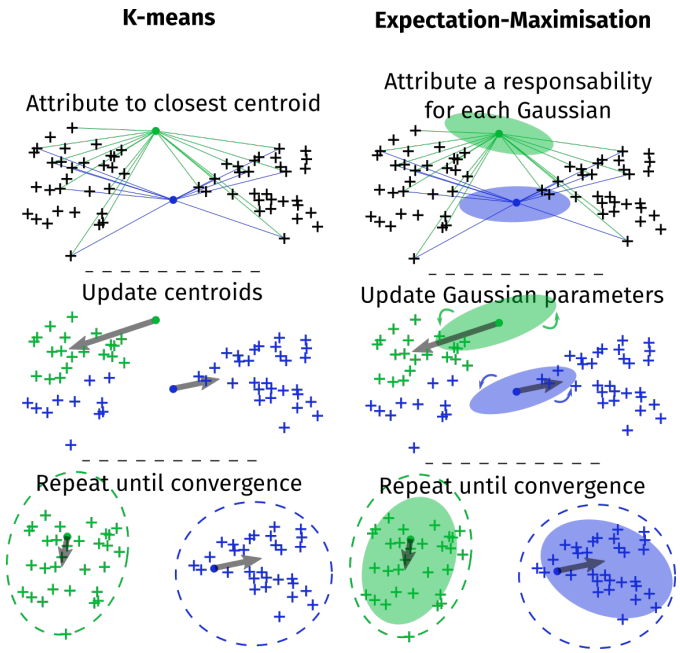


Figure 5. Illustration of the K-means and ~~EM~~ expectation-maximisation algorithm on ~~fake 2-dimensional~~ two-dimensional artificial data and ~~2~~ two clusters.

  - ~~'gmm',~~ 'gmm' for the EM algorithm (Gaussian mixture ~~)~~ model) and
  - ~~'kmeans',~~ 'kmeans' for the K-means algorithm

- **classif_score:** The internal score used to automatically choose the number of clusters (only used when n_clusters=~~'~~'auto'). See ~~section~~ Sect. 3.4 for a description of ~~these~~ the internal scores.

- **init:** initialisation strategy for both algorithms. Three choices are available:
  - ~~'~~'random': ~~pick randomly~~ randomly pick an individual as starting point (both Kmeans and ~~GMM)~~EM);
  - ~~'~~'advanced': ~~more sophisticated way to initialize~~ use a more sophisticated initialization (kmeans++ for Kmeans (Arthur and Vassilvitskii, 2007) ~~,~~ and the output a ~~Kmeans pass for GMM)~~K-means pass for EM); and
  - ~~'~~'given': start at explicitly passed point coordinates

- **max_height:** ~~height (meter~~ The height (meters above ground level) at which the profiles are cut.

- **n_clusters:** ~~the~~ The number of clusters to be formed (between ~~2 and 6). Either explicitly given , either~~ two and six). This is either explicitly given or determined automatically to ~~optimise~~ optimize the score given in classif_score.

- **n_inits:** ~~number~~ The number of repetitions of ~~repetition of~~ the algorithm. ~~The more it is , the more likely is~~ When this number is larger, the algorithm is more likely to find the global optimum ~~, but also the more timeit takes~~but requires more time.

- **n_profiles:** The number of profiles concatenated ~~before~~ prior to the application of the algorithm. For example, if n_profiles=~~1~~1, only the current profile is used. If n_profiles=~~3~~3, the current profile and the two previous profiles are concatenated and ~~put in input of~~ input to the algorithm.

- **predictors:** The list of variables used in the classification. ~~They~~ These variables can be different at night and ~~at~~ during the day. For both ~~, it~~ time periods, the variables can be chosen ~~among~~ from

  - RCS_{co}: the copolarized range-corrected backscatter signal; and
  - RCS_{cr}: ~~crosspolarized~~ the cross-polarized range-corrected backscatter signal

A simplified flowchart of KABL is shown in Figure 6 and the parameters of the KABL software are highlighted in bold in the following explanation of the KABL algorithm. A netCDF file generated by the *raw2l1* software needs to be provided as input data to KABL. The data, namely, the altitude vector z (size N_z), the time vector t (size N_t), the range-corrected signals RCS_{co} and RCS_{cr} ($N_t \times N_z$ matrices), are extracted from this file. Such data are prepared to fulfil the machine-learning algorithms requirements. For each time, the **n_profiles** last profiles are extracted. Then, the data they contain are normalized (by removing the mean and dividing by the standard deviation); this provides a matrix X ($N \times p$, where $N =$**n_profiles**$\cdot N_z$ and $p = |$**predictors**$|$ is the number of elements in the list). The matrix X is the usual input for a machine-learning algorithm; it has one line for each individual observation and one column for each variable (or predictor) observed. For the BLH retrieval, the preparation also provides a vector Z (size N) containing the altitude of each individual observation. The algorithm (either K-means or EM, as specified by **algo**) is applied to the matrix X, with the parameters **n_clusters**, **init** and **n_inits**. This results in a vector of *labels* (size N) that contains the cluster attribution of each individual. Finally, we look for the first change in the cluster attribution, starting from the ground level. This gives us the value of BLH for this profile. These operations are repeated until reaching the end of the netCDF file.

3.4 Performance metrics

Two ~~kind of metrics have been used:~~ types of metrics were used.

- External scores: ~~they~~ These metrics compare the result to a trustworthy reference. They have the advantage ~~to give~~ of providing a meaningful evaluation of the performance, but ~~they depend widely~~ depend strongly on the quality of the reference (i.e., its accuracy and availability).

- Internal scores: ~~they tell~~ These metrics rate how well the classification ~~is done,~~ performs based only on the distances between points. They have the advantage ~~to be always computable , but they~~ of being always computable but are not linked to any physical property ~~, hence~~ and therefore are not always meaningful.

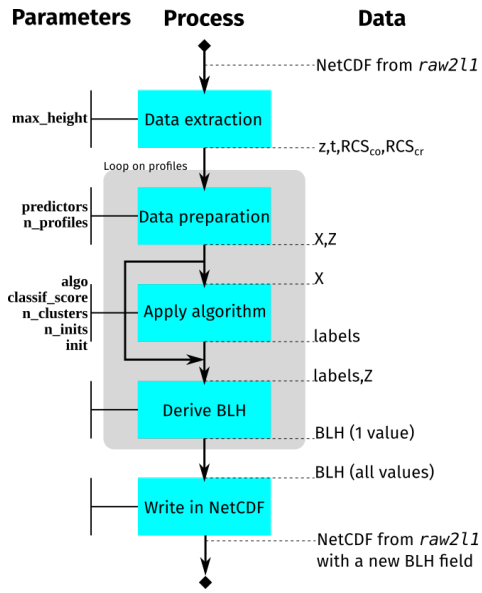


Figure 6. Simplified flowchart of the KABL algorithm.

~~As none of them is perfect~~None of these metrics are perfect; however, the information ~~brought by all give~~ they provide allows a broader understanding of the ~~algorithms~~ algorithm performance.

3.4.1 External scores

External scores use a reference to assess the quality of the result. In our case, the reference is the ~~BLH estimated from RS~~ RS-estimated BLH and, when available, the ~~BLH estimated by a human expert~~human expert-estimated BLH. If we denote ~~by~~ $\hat{Z}$ as the estimated BLH (by any of the previously introduced ~~algorithm) and by~~ algorithms) and Z_{ref} as the reference, the external scores used in this study are ~~denoted as follow: the root-mean-squared error (~~the root mean square error (RMSE) (E_2, equation 1) ~~, the average gap (E_1, equation ??), the Pearson 's~~ and the Pearson correlation (ρ, equation 2).

$$E_2 = \sqrt{\mathbb{E}\left[(\hat{Z} - Z_{ref})^2\right]} \tag{1}$$

$$E_1 = \mathbb{E}\left[|\hat{Z} - Z_{ref}|\right]$$

$$\rho = \frac{cov(\hat{Z}, Z_{ref})}{\sigma(\hat{Z})\sigma(Z_{ref})} \tag{2}$$

~~In these formulae~~Here, $\hat{Z}$ and Z_{ref} are random variables. ~~In the estimation of these scores , they~~ When these scores are estimated, the random variables are replaced by a ~~vector of sample ,~~ sample vector and the expectation and standard deviation are replaced by their usual estimators. ~~For unsupervised algorithms, such errors are calculated on all external information. For~~

~~supervised algorithms, such errors are calculated only on the external information that was not used to train the algorithm (the test set) which is about 20% of the total.~~

3.4.2 Internal scores

The quality of a classification can be quantified ~~by some scores only based~~ using scores that are based only on the labels and the distances between points. ~~It gives an estimation of~~ Such scores estimate how trustworthy an estimation is, without any external input. Many ~~scores exist,~~ such scores exist with different formulation ~~,~~ and different strengths and weaknesses (Desgraupes, 2013). In this study, three internal scores ~~are~~ were used:

- ~~Silhouette~~ the silhouette score (Rousseeuw, 1987).
  ~~Compares~~ which compares the average distance to its own group (a) to the average distance to the ~~neighbouring~~ neighboring group (b): $S_{sil} = \frac{b-a}{\max(a,b)}$~~.~~,
  ~~1~~ where 1 is the best classification, ~~0 is neutral. -1 the worst .~~ 0 is neutral, and -1 is the worst classification;

- the Calinski-Harabasz index (Caliński and Harabasz, 1974)~~.~~,
  ~~Compares~~ which compares the between-cluster dispersion (B) to the within-cluster dispersion (W_k): $S_{ch} = \frac{(N-K)B}{(K-1)\sum_{k=1}^{K} W_k}$~~.~~,
  where $+\infty$ is the best classification ~~, 0 the worst .~~ and 0 is the worst classification; and

- the Davies-Bouldin index (Davies and Bouldin, 1979).
  ~~Compares~~ which compares the average distance to its group center ($\bar{\delta}_k$) to the distance between the group centers ($d(\mu_k, \mu_{k'})$): $S_{db} = \max_{k' \neq k} \left(\frac{\bar{\delta}_k + \bar{\delta}_{k'}}{d(\mu_k, \mu_{k'})} \right)$,
  ~~0~~ where 0 is the best classification ~~,~~ and $+\infty$ ~~the worst~~is the worst classification.

These three scores ~~have been~~ were chosen to diversify the metrics and ~~they~~ are all implemented in Scikit-learn (version $\geq$0.20).

3.4.3 Other metrics

In addition to the internal and external scores, the computation time and the number of invalid values (NaN or Inf) ~~are recorded. Even though they don't~~ were recorded. BLH estimates of NaN or Inf can occur when all the points of the profile are assigned to the same cluster; this reflects a faulty configuration of the algorithm. Even though these metrics do not measure how well ~~the program is doing~~a program is performing, they are useful ~~for~~ to the user.

All the metrics used to measure the performance of KABL are summarized in ~~the table~~ Table 1.

4 Results

4.1 Sensitivity analysis of the KABL algorithm

A sensitivity analysis was ~~carried out~~ performed on KABL code ~~in order to find~~ to identify the "best" configuration. ~~Various configuration of KABL have been tested extensively~~ Various KABL configuration were extensively tested on a single day: ~~the~~

Table 1. Table of metrics used to measure the performance of KABL algorithm

Metric	Type	Description	Best/worst value
corr	External	Pearson correlation coefficient	~~1~~1/~~0~~ 0
~~**errl1**~~ ~~External~~ ~~Average~~ ~~absolute~~ ~~gap with~~ ~~reference~~ ~~0/+∞~~ **errl2**	External	~~Root-mean-squared gap with reference~~ Root mean square error (RMSE)	~~0~~0/$+\infty$
s_score	Internal	Silhouette ~~coefficient~~ score	~~1~~1/~~-1~~ -1
db_score	Internal	Davies-Bouldin index	~~0~~0/$+\infty$
ch_score	Internal	Calinski-Harabasz index	$+\infty$/~~0~~ 0
chrono	Other	Time to perform 24 ~~hours of BLH estimation~~ h of BLH estimations	~~0~~0/$+\infty$
n_invalid	Other	Number of invalid values (NaN or Inf) in 24 ~~hours~~ h of BLH estimations	~~0~~0/$+\infty$

~~Table of metrics used to measure the performance of KABL algorithm~~

~~2nd of August~~ August 2, 2018~~at Trappes~~, at the Trappes site, for which we have a hand-made reference (~~see figure 4 -top). The more relevant configuration is then~~ Figure 4 (top)). The most relevant configurations were retained and tested on the ~~two years~~ two-year dataset.

There are ~~8~~ height parameters in the KABL code (see ~~section~~ Sect. 3.3 for their ~~description~~descriptions). To assess the sensitivity of KABL to these parameters, the performance metrics (given in ~~section~~ Sect. 3.4) were estimated with the hand-made BLH as Z_{ref} and with ~~KABL's output~~ the output of KABL as $\hat{Z}$ for different ~~combination~~ combinations of input parameters. The tested values for the input parameters are given in ~~table~~ Table 2 and the output metrics are given in ~~table~~ Table 1. We ~~call a *configuration* a~~ refer to a set of values for ~~KABL parameters~~the KABL parameters as a *configuration* . Screening all the possible values listed in ~~table 2 would take~~ Table 2 would require 3240 different configurations.

To ~~look into~~ obtain an overview of these 3240 configurations~~at a glance~~, we started by estimating the influence of the ~~code~~ parameters (listed in ~~table 2) onto~~ Table 2) on the different metrics (listed in ~~table~~ Table 1). ~~Their influence is quantified by first order~~ The influence of the parameters was quantified using first-order Sobol indices (Sobol, 2001; Iooss and Lemaître, 2015; Rieutord, 2017), that is~~to say~~ , the ratio of the variance of the metric when the parameter ~~is~~ was fixed over the total variance of the metric. If we denote ~~by~~ Y as the metric and ~~by~~ X as the vector of parameters, all considered as random ~~, the~~

Table 2. Possible values for the parameters of the KABL code. The dependencies between parameters result in 3240 different configurations.

Parameter	Possible values	Meaning
algo	~~'~~'kmeans'	The K-means algorithm is used
	~~'~~'gmm'	The EM algorithm is used (Gaussian mixture model)
classif_score	~~'~~'silh'	~~Silhouette~~ The silhouette score is used
	~~'~~'db'	The Davies-Bouldin index is used
	~~'~~'ch'	The Calinski-Harabasz index is used
init	'random'	Starting ~~point~~ points are chosen randomly
	~~'~~'advanced'	Starting points are chosen with a smarter strategy
	~~'~~'given'	Starting points are explicitly given
max_height	3500	The height (meters above ground level) at which
	4500	profiles are cut
n_clusters	2	
	3	The number of clusters to be formed is explicitly
	4	passed and is always the same
	5	
	~~'~~'auto'	~~Automatically chosen to optimise~~ The number of clusters is automatically chosen to optimize **classif_score**
n_inits	10	The number of times the algorithm is repeated with
	80	different initializations (when **init** is not 'given')
n_profiles	1	Only the current profile is used
	2	The current profile and the previous profiles are used
	3	The current profile and the ~~2 previous~~ two previous profiles are used
	4	The current profile and the ~~3 previous~~ three previous profiles are used
predictors	~~RCS0~~ 'co'	~~Copolarized~~ The copolarized range-corrected signal is used at all times
	~~dRCS0nRCS12~~ 'co/co+cr'	~~Copolarized~~ The copolarized range-corrected signal is used during ~~daytime,~~ the daytime, and both polarization channels are used separately during nighttime
	~~RCS12~~ 'co+cr'	Both polarization ~~channel~~ channels are used separately at all times

~~Table of possible values for the parameters of KABL computer code. With the dependencies between parameters, it gives 3240 different configurations.~~

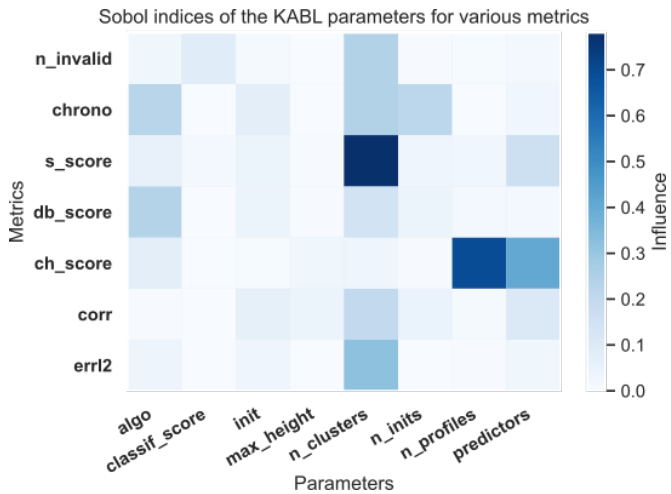


Figure 7. Relative influence ~~on~~ of parameters ~~over~~ on the different metrics~~: on~~. The x-axis ~~are~~ indicates the ~~code's~~ parameters of the code, ~~on~~ and the y-axis ~~are~~ indicates the metrics~~, the color shade~~. The shading represents the influence of the parameter ~~over~~ on the metric with darker shading indicating larger influence.

~~first order~~ variables, the first-order Sobol index of the i-th parameter is defined ~~by~~ as $S_i = V(\mathbb{E}[Y|X_i])/V(Y)$~~(with~~, where $V(\cdot)$ ~~denoting~~ denotes the variance and $\mathbb{E}[\cdot]$ ~~expectation). The higher the Sobol index , the higher the influence.~~ denotes the expectation. A higher Sobol index indicates a larger influence.

Figure 7 shows the Sobol indices obtained on the KABL computer code. ~~Reading~~ Examining the matrix line by line, one can see that the metrics are sensitive to different parameters~~:~~. For example, the silhouette score is very sensitive to **n_clusters** ~~, for example, while~~ while the Calinski-Harabasz index is sensitive to **n_profiles** and **predictors**. ~~Reading~~ Examining the matrix column by column, one can see that some parameters are more influential than others (e.g. **classif_score** is much less influential than **n_clusters**~~, for example). It also highlights what are~~). This matrix highlights the main effects of changing ~~this parameter , andhence,~~ a parameter and, therefore, how to set ~~it well~~each parameter appropriately. For each parameter, we ~~will look at the metrics it has an influence on and decide which configurationis better~~examined the metrics that it influences and determine the preferred configuration.

Critical parameters are ~~thus indicated in the figure~~ indicated in Figure 7 by the ~~deepest~~ darkest blue columns, namely~~:~~ , **n_clusters**, **algo**, **predictors** and **init**[3]. For each ~~of them, we have drawn in figure 8 the distribution (violin plots)~~ parameter, Figure 8 shows the distribution of the relevant output ~~conditionally to~~ given the parameter value (violin plots are explained in Hintze and Nelson (1998)). For example, Figure 8a has the ~~sub-figure 8-a has for abscissa the~~ value of **algo** ~~and for ordinate~~ on the x-axis and the computing time on the y-axis. The 3240 different configurations ~~are divided in two parts: the ones~~ were divided into two groups; those with **algo=**~~'~~'kmeans' and ~~the ones~~ those with **algo=**~~'~~'gmm'. ~~What we see in sub-figure 8-a is~~

[3] ~~Although~~ Even though **n_profiles** has a large Sobol index for the Calinski-Harabasz index, this influence ~~is~~ was not explored because it is ~~only~~ due to the increase of ~~**ch_score**~~ this index with the number of points.

Table 3. Retained values for the parameters of the KABL code after the sensitivity analysis

Parameter	Retained values
algo	~~'~~'kmeans'
classif_score	~~'~~'db'
init	~~'~~'given'
max_height	4500
n_clusters	3
n_inits	10
n_profiles	1
predictors	~~RCS0~~ 'co'

~~Retained values for KABL computer code's parameters after sensitivity analysis~~

~~the~~ Figure 8a shows a smoothed histogram of the computing time for the divided populations. The other ~~sub-figures are~~ panels in Figure 8 were constructed in the same way. Each line ~~correspond~~ corresponds to a critical parameter, and we represent the two most influenced outputs according to ~~figure~~ Figure 7.

~~Parameters values are~~ The parameters values were chosen to give the most optimal ~~value~~ values for the metrics they ~~have influenceon. Optimal~~ influence. The optimal values are indicated by a yellow star ~~on~~ in each plot. To set **algo**, we ~~better look at~~ examined the computing time (~~8-a~~Figure 8a) and the Davies-Bouldin ~~score (8-b): it results that '~~index (Figure 8b). These figures indicate that 'kmeans' is the best choice for both ~~(~~metrics (resulting in a lower computing time and a lower Davies-Bouldin index). To set **init**, we ~~better look at~~ examined the correlation (Figure 8-c) and the computing time (~~8-d): 'given' appear~~ Figure 8d). In this case, 'given' appears to be the best choice. To set **n_clusters**, we ~~better look at~~ examined the RMSE (8~~-e) : 3 clusters is the best;~~ e) and the silhouette score (8~~-f): 'auto' is the best~~ f). They indicate the best number of cluster is respectively three and 'auto'. We chose to give ~~the priority to the~~ priority to RMSE because silhouette score has also very high values for ~~2~~ two clusters, which is suspicious given the presence of a cloud and a residual layer this day. To set **predictors**, we ~~better look at~~ examined the silhouette score (8-g) and the Calinski-Harabasz ~~score (8-h): ' RCS0' appear~~ index (8h); here, 'co' appears to be the best choice. ~~The same methodologywas applied to the remaining parameters and the resulting configuration is in table~~ Following this methodology, we can identify a few configurations worth trying. These configuration were tested on the two-year dataset. The configuration used to generate the results in Sect. 4.2.1 is given in Table 3. It ~~will be used to generate the results in the next section~~was chosen to maximize correlation between KABL and RS at the Trappes site.

4.2 ~~Two years~~ Two-year comparison

~~All~~ The three methods (KABL, ADABL and the manufacturer's ~~) have been compared to radiosounding estimation for~~ algorithm) were compared to RS estimates over a two-year period.

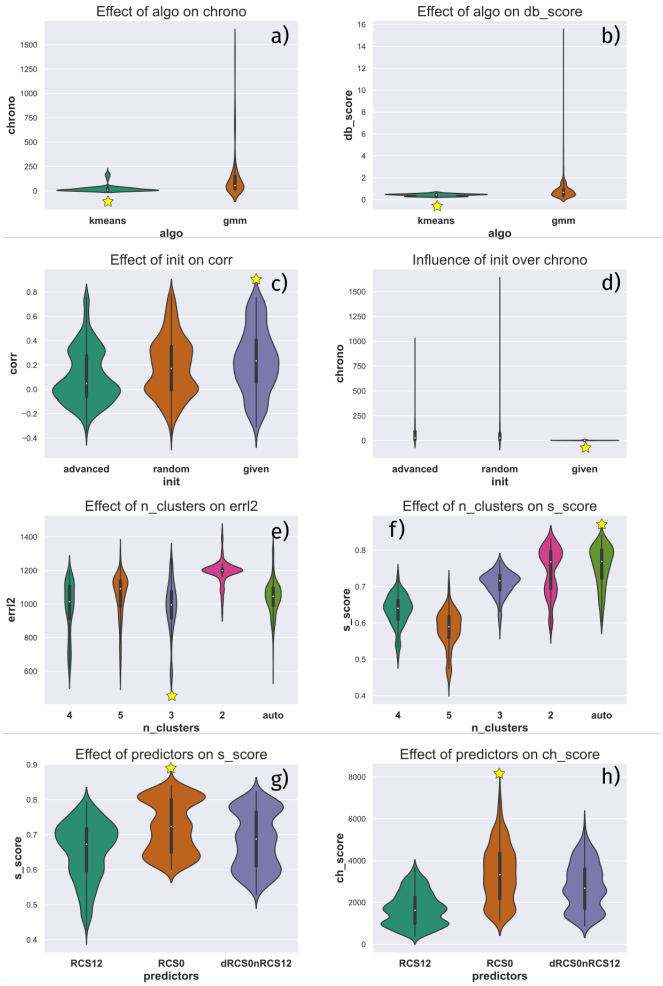


Figure 8. Distribution of the relevant outputs ~~of~~ for the critical inputs~~: a) and b) show the~~ . The effect of **algo** on (a) the computing time and ~~Davies-Bouldin index~~ (~~respectively~~b) ~~. c) and d) show~~ the Davies–Bouldin index. The effect of **init** on (c) the correlation and (d) the computing time. ~~e) and f) show the~~ The effect of **n_clusters** on (e) the root mean square error (RMSE~~and silhouette score. g~~) and ~~h~~(f) ~~show~~ the silhouette score. The effect of the predictors ~~over~~ on (g) the silhouette score and ~~Calinski-Harabasz score~~(h) the Calinski–Harabasz index. For each ~~sub-graph~~panel, the best parameter value is highlighted by ~~an~~ a yellow star.

4.2.1 Overall comparison

As explained in ~~the section 3.4, three external scoresare~~ Sect. 3.4, two external scores, RMSE and the correlation, were used to assess the quality of the ~~estimations: the RMSE, the average gap and the correlation. As the average gap E_1 and the RMSE E_2 are very similar, we will show only the RMSE E_2 and the correlation ρ. In formulae~~ estimates. In equations 1 and 2, the reference BLH Z_{ref} ~~is now the RS estimation~~was set to the RS estimate, as described in ~~section~~ Sect. 2.2. To ~~be able to compute such score, BLH estimation from lidar and from RS must be collocated~~compute the scores, the BLH estimates from the lidar and radiosonde must be co-located. At the time of each RS estimation, the corresponding lidar ~~estimation~~ estimate is the average of all ~~available within the next~~ the available estimates within the 10 ~~minutes after release (it means 1 or 2 lidar estimations~~min following the release of the radiosonde (this translates to one or two lidar estimates). The following meteorological conditions ~~have been~~ were discarded:

- ~~Rain~~ rain (rain gauge measures ~~RR~~the rainfall as >0 ~~)~~ mm);
- ~~Fog~~ fog (scatterometer measures ~~visibility~~ the visibility as <~~1000m)~~1000 m);
- ~~Low~~ low level cloud (ceilometer measures cloud base height <3000m);
- RS estimation below 120m (blind zone for lidar); and
- ~~Nighttime (launch of~~ nighttime (RS launched at 23:15 UTC)

This selection rejects a large part of the dataset ~~, but it ensures to keep~~ but ensures that only well-defined ~~boundary layer. Meteorological conditions are measured by~~ cases are retained for the comparison. In total, 178 RS measurements from Trappes and 101 RS measurements from Brest were used for the overall comparison. The meteorological conditions were measured using the ancillary instruments presented in ~~section~~ Sect. 2.3. The results of the comparison are shown in ~~figure~~ Figure 9.

In ~~figure 9~~Figure 9, we can see the results of the comparison between the KABL and RS estimates (blue bars), between the ADABL and RS estimates (grey bars), and between the manufacturer's ~~algorithm and RS~~ and RS estimates (orange bar). The first column represents ~~the~~ RMSE E_2 (~~the lower the better). The~~ lower is better), and the second column represents the correlation ρ (~~the higher the~~ higher is better). The ~~first line~~ upper row shows the results for the Brest site, the ~~second line for~~ lower row show the results for the Trappes site. ~~KABL has the largest error for both sites, the lowest correlation for Brest and the second lowest for Trappes. ADABL has the best performance (highest correlation , lowest error) for Trappes and medium performances for Brest. Manufacturer~~One can see very different results depending on the site. While both KABL and ADABL outperform the manufacturer's algorithm at the Trappes site, neither algorithm does at the Brest site. The correlation is strongly affected by the site. While the correlation for both KABL and ADABL is higher than that for the manufacturer's algorithm ~~has the best performance for Brest~~at the Trappes site, it collapses to close to zero for KABL at the Brest site (0.07 for ADABL). The RMSE values can be compared to the values given in Haeffelin et al. (2012). For KABL, we find 770 m at the Brest site and 798 m at the Trappes site, while for ADABL, ~~medium error and lowest correlation for Trappes . Overall, ADABL has better performances than KABL, and so has the manufacturer's. However, it is not clear which one between the~~

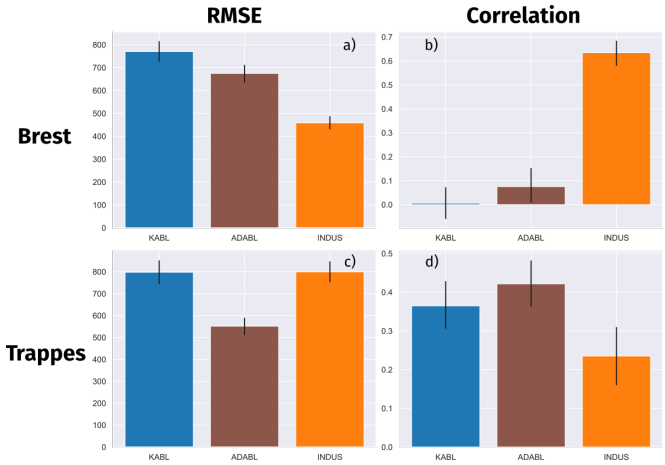


Figure 9. Results of ~~the~~ a two-year comparison with ~~radiosounding for 2 years, on~~ the radiosonde (RS) estimates at both sites ~~,~~ for two metrics: RMSE and correlation. Cases at night or with rain, fog, an RS-estimated BLH ~~estimated by RS~~ of under ~~120m~~ 120 m, or ~~cloud~~ clouds under ~~3000m have been~~ 3000 m were removed. The 95% confidence intervals were estimated using percentile bootstrapping (Davison and Hinkley, 1997)

~~manufacturer and ADABL is the best. Machine learning algorithmshave better performances at Trappes. It may be because the boundary layer is better defined in Trappes than in Brest (generally higher, larger aerosol load and less often perturbed by synoptic disturbances).~~ we find 675 m at the Brest site and 552 m at the Trappes site. Our values are notably higher than those in Haeffelin et al. (2012). This is likely due to the larger extent of our dataset (178 RS at Trappes, 101 at Brest, spanning over 2 years) and the low maturity of the algorithms. Between ADABL and KABL, ADABL has better correlation and RMSE values than KABL at both sites. The manufacturer's algorithm performs well without any specific tuning on our part. It uses a wavelet covariance transform as described in Brooks (2003). This result is not surprising because the wavelet method has been shown to be robust in numerous studies, especially in Caicedo et al. (2017), who included a cluster analysis method and concluded that the wavelet method should be preferred.

4.2.2 Seasonal and diurnal cycles

~~In order to qualify~~ To quantify the ability of the ~~algorithm to give a consistent estimation of BLH , we have drawn in figure 10~~ algorithms to provide a consistent BLH estimation, Figure 10 shows the seasonal cycle (monthly average) and the diurnal cycle (six-minute average) ~~on~~ at both sites. For each estimator, the thick line ~~is~~ represents the average BLH ~~estimation~~ estimate and the shaded area ~~in the inter-quartiles~~ represents the inter-quartile gap. Rain, fog~~and low clouds conditions have also been~~ , and low-cloud conditions were discarded. For the monthly average, the night values ~~are removedtoo~~were also removed. If we ~~take only night estimation~~include only night estimates, the seasonal cycle is reversed for RS ~~estimations: they~~ estimates, that is, the BLH estimates are lower in summer. For other estimators, we do not see such a difference between day and night seasonal ~~cycle~~ cycles (not shown).

~~In Brest (fig. 10-a), estimations from the manufacturer~~At the Brest site (Figure 10a), estimates made by the manufacturer's algorithm are lower than ~~estimations from~~ those made by KABL and ADABL ~~. Estimations from~~ and estimates made by ADABL are usually higher than ~~estimations from KABL (excepted~~ those made by KABL (except in July). ~~Radiosoundings give BLH estimate rather~~ The RS estimation gave BLH values that were low in summer (~~June to October), rather high~~ June–October), high in February and March (higher than ~~KABL's)in February and March and between~~ the KABL estimates), and between the manufacturer's and KABL ~~'s~~ estimates during the rest of the year. Overall, the manufacturer's ~~has the cycle the closest to RS~~algorithm displays the seasonal cycle that is closest to that of the RS estimates, while KABL ~~overestimates and ADABL overestimates even more. Inter-quartiles~~ and ADABL both overestimate BLH. The inter-quartile distances (shaded areas) are large for all estimations, reflecting the wide range of ~~values BLH estimations can take~~the BLH estimates.

~~In Trappes (fig. 10-c)also~~At the Trappes site (Figure 10c), KABL and ADABL also overestimate BLH in comparison to ~~RS, while~~ the RS estimates, while the manufacturer's ~~estimation~~ estimate is close. The seasonal cycle is more visible in Trappes than in Brest~~: all BLH estimations~~ , and all BLH estimates are higher in summer than in winter. The ~~one with the more marked cycle is~~ most pronounced cycle is given by KABL, while the ~~one with the less marked cycle is RS. Inter-quartiles~~ least pronounced cycle is given by the RS estimation. The inter-quartile distances are also very large, especially in summer, because the difference between ~~BLH at day and BLH~~ the BLH estimates during the day and at night is larger.

Figures 10~~-b and~~ b and 10d show the diurnal cycle~~: all values with the same 0.1 hour~~ , where all values within the same six-minute period in the day ~~(6 minutes)~~ were averaged. ~~The diurnal cycle of RS cannot be drawn because they~~ Because the radiosondes are only launched twice a day~~:~~, at 11:15 UTC and 23:15 UTC, an equivalent RS-estimated diurnal cycle cannot be drawn. However, we ~~have drawn~~ used the average and quartile values at these ~~hours, as a checkpoint~~ times as checkpoints for the other ~~estimations. Manufacturer~~estimates. The manufacturer's and KABL ~~estimations have both a quite smooth diurnal cycle~~estimates both have very smooth diurnal cycles, with lower BLH at night and ~~a maximum~~ maximum BLH around 15:00 UTC at ~~Trappes and~~ the Trappes site and around 13:00 ~~at Brest. KABL's~~ UTC at the Brest site. The KABL average is always higher than that calculated by the manufacturer's ~~one. ADABL's estimation show a really~~ algorithm. The ADABL estimation has a very different diurnal cycle, ~~quite~~ similar to the conceptual image we have of the boundary layer. Indeed, ~~it has been trained on~~ ADABL was trained using hand-made BLHs ~~which reflects~~ that reflect this conceptual image. ~~Thus it is no surprise ADABL reproduce it well, however~~Therefore, it is not surprising that ADABL reproduces this image well; however, it may fail to adapt to special cases. It ~~seems~~ appears that the "time" predictor (~~number of second~~ the number of seconds since midnight) has a large influence ~~and~~ that is not balanced by the other predictors. This is ~~probably due to the fact it has been~~ likely because ADABL was trained on only two days~~:~~ , resulting in an unbalanced importance for sunrise and sunset ~~this~~ on these particular days and ~~locations acquire an unbalanced importance~~at these locations. To balance this importance, the AdaBoost algorithm ~~should~~ needs to be trained on more days and ~~more sites,~~ at more sites with a representative selection of cases.

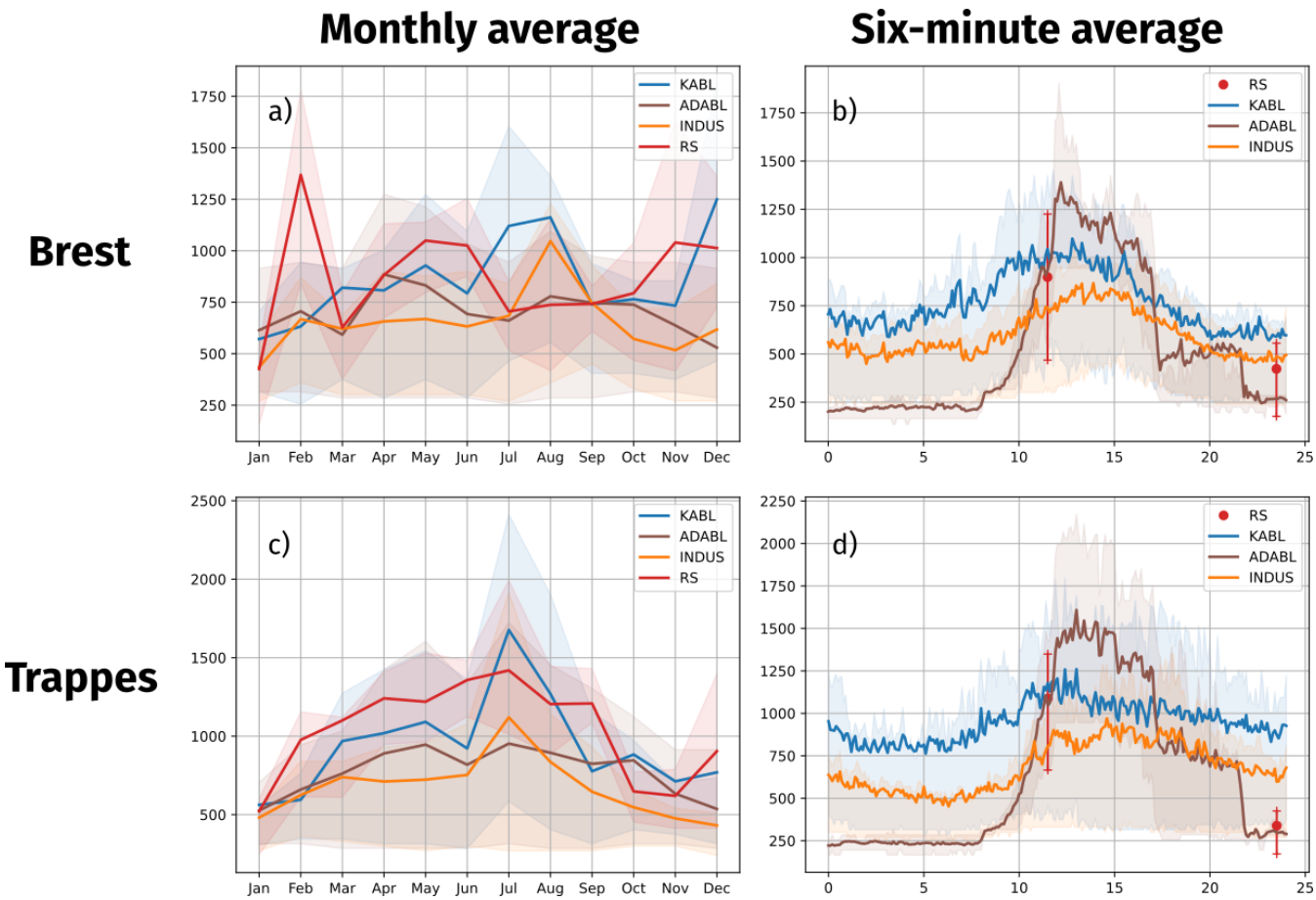


Figure 10. ~~Seasonal~~ (a~~and~~ , c) Seasonal and ~~diurnal~~ (b~~and~~ , d) diurnal cycles of all BLH ~~estimations on~~ estimates at both sites. INDUS indicates the manufacturer's algorithm. Thick lines represent the average, shaded area the quartiles.

4.3 Case study

The chosen case study ~~is the 19th of April~~ was for April 19, 2017~~at Trappes~~, at the Trappes site. The boundary layer ~~is~~ was clearly visible and ~~has almost~~ had nearly all the features of the conceptual image. ~~It~~ The case study day must be different than the days used for the training of ADABL, otherwise the comparison would be biased in favor of ADABL.

~~In figure 11 is represented the co-polarized~~ Figure 11 represents the copolarized backscatter intensity (RCS_{co}) in ~~shade of colors. Abscissa is~~ shaded colors. The x-axis indicates the hour of the day (UTC)~~. Ordinate is the height (meter~~ , and the y-axis indicates the height (meters above ground level). The different BLH estimates are ~~superimposed in~~ represented by dotted lines: blue ~~is~~ indicates KABL, orange ~~is~~ indicates the manufacturer's algorithm, ~~green is~~ and green indicates ADABL. At the beginning of the day, there is a thick residual layer ~~with some plumesinside~~containing some plumes. Both KABL and the manufacturer's ~~include these plume into~~ algorithm include these plumes in the boundary layer. Conversely, ADABL ~~give~~

gives a very low ~~estimation~~ estimate where there is no visible frontier. In the morning (from ~~8~~ 08:00 UTC to 12:00 UTC), ~~they all catch the~~ all the algorithms capture the transition reasonably well. However~~KABL has more irrelevant estimations~~ , KABL includes more irrelevant estimates (hitting what remains of the surface layer) than ~~others, and ADABL goes too high with~~ the other methods and ADABL gives an estimate that is too high for no apparent reason around 12:~~00.~~ 00 UTC. During the day, ADABL sticks to the ~~boundary layer top~~top of the boundary layer, the manufacturer's algorithm sticks to the surface layer (~~quite visible)~~which is very visible), and KABL oscillates between ~~both~~the two. The evening transition is blurry~~:~~ ; the surface layer slowly sends back ~~more and more signalto finally turn~~ increasing amounts of signal, finally turning the mixed layer into a residual layer. KABL locates ~~the~~ this transition very early (around 17:00 UTC), when it stops oscillating and sticks to the surface layer. ADABL makes the transition more smoothly, from 19:00 UTC to 22:~~00.~~ 00 UTC. The manufacturer's algorithm is the ~~latest~~ last to make the transition, at around 23:00 ~~, quite~~ UTC, and the transition then occurs very sharply. We can conclude from this case study that ~~no algorithm perfectly catches~~ none of the algorithms perfectly capture the boundary layer. Some of the limitations are physical~~:~~ , e.g., the evening transition is ~~ill-defined, therefore algorithmsdisagree~~ill defined, resulting in disagreement between the algorithms. The RS ~~of~~ estimate at 23:15 ~~gives an estimation which~~ UTC is close to the lower boundary of the lidar range. ~~It~~ This highlights the fact that ~~BLH below 120m are not that~~ BLHs below 120 m are not rare and cannot be detected with ~~the~~ lidar alone, whatever the method. Some ~~others are algorithmic:~~ of the other limitations are algorithmic; KABL has an unfortunate ~~trend~~ tendency to oscillate between several candidates for ~~boundary layer top~~ the top of the boundary layer (surface layer or clouds), ~~ADABL reproduces too much~~ and ADABL too closely reproduces the features of the days it has been trained on (~~night estimation and morning transition~~e.g., night estimates and morning transitions).

5 Discussion

This ~~sections~~ section discusses various aspects of the results and the methodology. For the sake of readability, it has been split ~~in many short paragraphs~~into several short subsections.

5.1 Algorithms maturity

Both of the examined algorithms are very recent. K-means algorithms have already been used to detect ~~BLH~~ BLHs in previous studies ~~(Toledo et al., 2014, 2017; Ricutord et al., 2014). Therefore it is the most mature methodand this~~ (Toledo et al., 2014; Caicedo et al. ; therefore, it is a more mature method. This is visible in this paper ~~by~~ via the level of investigation which was much higher for KABL than for ADABL. Concerning boosting, ~~it is~~ this the first time, to our knowledge, that such ~~algorithm is~~ an algorithm has been tested on this ~~problem. Thus~~ type of problem; therefore ADABL is a completely new algorithm~~and yet~~ . Yet, it outperforms KABL and competes favorably with the manufacturer's ~~. However, it raises training issues.~~ algorithm despite raising training issues.

5.2 Time and altitude continuity

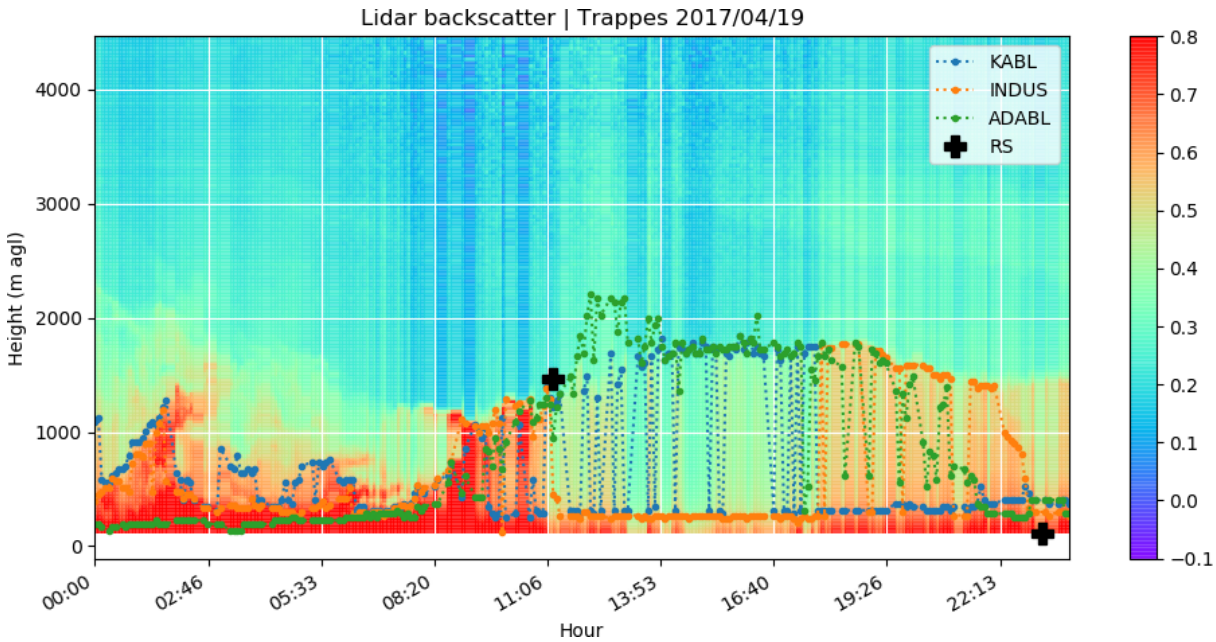


Figure 11. Case study: BLH ~~estimations by~~ estimates using different methods ~~the 19th of~~ on April 19, 2017, at the Trappes site.

The oscillations observed in the figure 11 are unrealistic and need to be avoided. They occur with KABL because clusters do not always have vertical persistence (some points are identified as free atmosphere in the middle of the boundary layer). Vertical persistence needs to be enforced, for example, by adding altitude to the KABL predictors. Another way to filter out these oscillations is to increase the time continuity in the post-processing, for example, with a moving average, or by imposing a maximum BLH growth rate (Poltera et al., 2017). The distance used in K-means could also be modified to incorporate these constraints, for example, by adding penalty terms. In ADABL, time and altitude continuity are ensured because they are within the predictors. However, ADABL yields BLH estimates that are too similar to the BLHs in the training set. Removing time and/or altitude from the predictors should be considered to force the algorithm to rely more on the measurements. Further, the sensitivity analysis presented here for KABL needs to be performed for ADABL.

5.3 Real time ~~estimation~~estimations

~~Although it has not been~~ Even though it was not necessary for this study, all ~~algorithms~~ of the algorithms studied here can be used in real time~~: as soon as the~~ . As soon as a backscatter profile is available, BLH estimations can be performed instantaneously[4]. However, ~~it has been shown that~~ KABL suffers from undesirable oscillations from one profile to the next. A method to filter these oscillations ~~will be needed , but it can also divest~~ is needed but would disable the "real-time" ~~property.~~ feature of

[4]Both KABL and ADABL need less than 1 ~~second~~ s to run ~~on~~ a single profile.

the algorithm. In addition, the hour of the day needs to be explicitly passed in a periodic function. This has not been done here because we worked only on 24-hour time periods.

5.4 Quality of the evaluation

~~Although we had~~ Even though we made an effort to sort ~~meteorological conditions with~~ the meteorological conditions using ancillary data, the ~~comparison over 2 years~~ two-year comparison still mixes heterogeneous ~~situations. More~~ conditions. In addition, the results are clearly different at the sites studied here, emphasizing the importance of local conditions. A more precise casting of ~~meteorological situation (~~the meteorological conditions with atmospheric stability indices or ~~large scale insight, for example)~~ large-scale insights would lead to a better understanding of ~~algorithm~~ the strengths and weaknesses ~~.~~ of the algorithms. The importance of the site needs to be investigated by extending the study to a larger number of sites with different environments. A more careful examination of cloudy days also needs to be performed. Cases where the cloud bases are below to 3 km were filtered from our study. However, cases where clouds reside inside the inversion should be detected by KABL as an extra cluster; further studies are required to confirm this behavior. In addition, ADABL was not specifically trained to deal with cloudy situations. Further studies to determine how ADABL behaves without training and how it could be appropriately trained would be interesting.

5.5 Quality of the reference

~~Radiosoundings are unquestionably the~~ Radiosondes unquestionably provide the proper reference for altitude measurements. However, the derivation of BLH from ~~the measurements is more questionable, as several methods exists and (strongly)~~ such measurements is contentious because several methods exist and some strongly disagree. Moreover, ~~they~~ RS measurements cannot be used to assess the full diurnal ~~cycle of the BLH . It~~ BLH cycle. This is a clear limitation of this study ~~: it cannot tell~~ because the RS measurements cannot determine if the difference between ~~ADABL's diurnal cycle and the others is an improvementof not. Therefore~~the diurnal cycle of ADABL and those of the other methods represents an improvement. Therefore, a very interesting ~~prospects~~ project would be to use a dedicated field experiment with ~~high frequency RS~~ high-frequency radiosonde or other continuously running instruments as a reference. For example, microwave radiometers are good candidates ~~, as they bring an information~~ because they provide information that is not based on aerosols and BLH ~~derivation is usual~~ derivations from these instruments are routine (Cimini et al., 2013).

5.6 ~~Training of~~ ADABL: Training

ADABL ~~shows already good performances while it has been trained on two days only. It has been shown that most~~ already shows good performance when trained on only two days. Most of its bad ~~estimations come~~ estimates result from the short length of its training period. ~~Hence, a short term prospect~~ Therefore, a short-term project would be to label more days ~~,~~ with various meteorological conditions. However, ~~its dependence~~ the dependence of ADABL on training makes it ~~more~~ sensitive to instrumentation settings and ~~calibration. Although~~ calibrations. Even though the effect of a calibration or ~~an evolution of the~~

the evolution of an instrumental device has not been studied, it is likely that ~~the training must~~ training needs to be repeated after each calibration or change in the instrumental device. Therefore, two strategies are possible for ~~the training of~~ training ADABL: remove the influence of calibration ~~before training (it would require to know~~ prior to training (this would require knowing the instrumental constants for all of the devices) or train it to deal with differences (~~it would require to include~~ this would require including as many different devices as possible in the training set, which would then become very large). In any case, the main limitation will be the need to label ~~all the~~ the entire dataset (a priori by human ~~expert~~experts).

5.7 KABL~~is "trainingless"~~: Training-less

KABL appeared to ~~be the less well performing algorithm~~ perform the least well in this study~~, but it has very interesting ways of improvement. It~~ ; however, there are interesting prospects to improve its performance. KABL does not require any training~~, therefore~~; therefore, it is less dependent on instrumentation settings and ~~calibration. As it is poorly dependent on instrumental device, one could use it~~ calibrations. Because it is not strongly dependent on the instrumental devices, it can be used on backscatter profiles made by other instruments (e.g., ceilometers). Moreover, ~~one can imagine to add other profiles than backscatter intensity : after normalisation, it is just an additional predictor~~ other profiles besides the backscatter intensity can be added as additional predictors for unsupervised learning ~~. Thus the idea~~ after normalization. Therefore, the concept of KABL can be ~~pushed further to make instrument synergy between~~ advanced further to create synergy between multiple remote sensing instruments. Microwave radiometers are ~~(again) good candidate,~~ good candidates because they have ~~a~~ comparable time resolution ~~and they bring an independent information on~~ to lidar and provide independent information concerning the thermal stratification of the boundary layer. Cloud ~~radar have also~~ radars also have comparable time resolution ~~and they bring another~~ to lidar and provide additional independent information.

5.8 Quality flags

Currently, no quality flags for the estimation are provided. One approach would be to use the internal scores (i.e., silhouette, Davies–Bouldin, and Calinski–Harabasz defined in Sect. 3.4) as quality flags; however, further study is required to determine whether these metrics can serve as reliable quality flags.

6 Conclusions

This paper ~~has~~ described two algorithms based on machine learning to estimate the mixing layer height from ~~aerosols lidars measurements. One of them~~ aerosol lidar measurements. The first, KABL is based on the K-means algorithm~~: it is named KABL (K-means for Atmospheric Boundary Layer). The other~~ , the second, ADABL, is based on ~~AdaBoost algorithm: it is named ADABL (AdaBoost for Atmospheric Boundary Layer). Both~~ the AdaBoost algorithm. Both algorithms take the same input file: one day of data generated by the *raw2l1* routine ~~; and give the same output: the time series of BLH for this~~ and produce a similar output, a BLH time series for the input day. KABL is a non-supervised algorithm ~~. It will look~~ that looks for a natural separation in the ~~basekscatter~~ backscatter signals between the boundary layer and the free atmosphere. ADABL

is a supervised algorithm ~~. It~~ that fits a large number of ~~decisions trees on~~ decision trees in a labelled dataset and aggregates them ~~smartly to give~~ in an intelligent manner to provide a good prediction. KABL, ADABL and the lidar manufacturer's algorithm ~~(unknown) have been~~ were tested on a ~~2-years~~ two-year dataset taken from ~~Meteo-France operational network of lidars. The sites of~~ Météo-France operational lidar network. The Trappes and Brest ~~have been chosen for their different climate~~ sites were chosen because of their different climates and the availability of regular ~~radiosoundings~~RS measurements, which were used as ~~the reference. KABL has the largest error for both sites, the lowest correlation for Brest and the second lowest for Trappes. ADABL has the best performance (highest correlation, lowest error) for Trappes and medium performances for Brest. Manufacturer~~a reference.

A large discrepancy in the results was observed between the two sites. At the Trappes site, KABL and ADABL outperformed the manufacturer's algorithm ~~has the best performance for Brest , medium error and lowest correlation for Trappes. Overall, ADABL has better performances than KABL , and so has the~~ while the opposite occured at the Brest site. At both sites, ADABL performed better than KABL (higher correlation and lower error) and manufacturer's ~~algotithm. However, it is not clear which one between the manufacturer and ADABL is better. By analysing~~ algorithm (using s wavelet covariance transform) performed well. By analyzing the seasonal and diurnal cycles, we ~~can see that~~ determined that the KABL and manufacturer's ~~estimations have similar behaviour, but KABL is always higher of about 200m. ADABL has the most marked~~ estimates have similar behavior; however, the KABL estimates are always higher by approximately 200 m. ADABL generates the most pronounced diurnal cycle, with a ~~look~~ pattern that is very similar to the expected diurnal cycle~~, but it depends too much~~ ; however, its results depend greatly on the days it has been trained on. ~~Especially~~In particular, the sunset and sunrise ~~time~~ times of these days ~~are over-influencing the estimations. On~~ over-influenced the ADABL estimation. In the case study, we ~~can see~~ saw that both algorithms perform ~~globally well , but we have also illustrated some algorithmic limitations: KABL has an unfortunate trend~~ well overall; however, we identified several algorithmic limitations, e.g., KABL tended to oscillate between several candidates for ~~boundary layer top (surface layer~~ the top of the boundary layer (surface layers or clouds) ~~, ADABL is too much constained~~ and ADABL was overly constrained by the days it ~~has been~~ was trained on (~~night estimation~~ e.g., the night estimate and morning transition).

~~This experiment show that, despite few training and no maturity on the application of boosting on this problem, ADABL is competing with~~ In summary, ADABL is promising but has training issues that need to be resolved, KABL has a lower performance but is much more versatile, and the manufacturer's algorithm ~~. However, it is dependent on a trustworthy training set to get improved. Although KABLestimations do not always match with RS ones, it has the valuable advantage to not being dependent on a training set . Therefore it might easily be extended on backscatter profiles from other instruments (like ceilometers)~~using a wavelet covariance transform performs well with little tuning but is not open source. A wide range of future developments is available for ADABL and KABL, the most immediate being that the training set of ADABL can be enhanced, time and altitude continuity can be enforced in the KABL estimation, and both can be compared to high temporal resolution RS measurements.

Code availability. The KABL source code is available to and usable by all users, including commercial users. The code is freely available under an open-source license at the following link: https://github.com/ThomasRieutord/kabl. It is made in Python 3.7 with regular statistics and machine learning packages, namely, Scikit-learn 0.20 (Pedregosa et al., 2011) and SALib 1.3.7 (Herman and Usher, 2017), which are open source and available under free licenses. The repository contains all the necessary features to run the code on *raw2l1* outputs. Several days of data are also provided as examples.

Author contributions. Tiago Machado implemented an initial version of the KABL code and performed the first comparisons to the RS data. Sylvain Aubert extracted and processed all the data (lidar, radiosonde, and ancillary), made one of the hand-labelled BLHs, and participated actively in the writing of the manuscript. Thomas Rieutord implemented the current versions of KABL and ADABL, made one of the hand-labelled BLHs, produced the figures, and actively participated in the writing of the manuscript.

Competing interests. The authors declare that they have no conflicts of interest.

Acknowledgements. The authors would like to thank Alexandre Paci, Alain Dabas, and Olivier Traullé for their helpful reading and comments. ~~Thanks to~~ We would like to thank Marc-Antoine Drouin, from ~~SIRTA for giving us~~ the Site Instrumental de Recherche par Télédétection Atmosphérique, for providing us with the link to *raw2l1* ~~. Thanks to all Meteo-France~~ and all the Météo-France personnel agents who install and ~~sustain~~ maintain the lidar network. We thank Martha Evonuk from Evonuk Scientific Editing (http://evonukscientificediting.com) for editing a draft of this manuscript.

References

Arciszewska, C. and McClatchey, J.: The importance of meteorological data for modelling air pollution using ADMS-Urban, Meteorological Applications, 8, 345–350, 2001.

Arthur, D. and Vassilvitskii, S.: k-means++: The advantages of careful seeding, in: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

Besse, P., Guillouet, B., and Laurent, B.: Wikistat 2.0: Educational Resources for Artificial Intelligence, arXiv preprint arXiv:1810.02688, 2018.

Breiman, L., Friedman, J., Olshen, R., and Stone, C.: Classification and Regression Trees, Wadsworth, 1984.

Brooks, I. M.: Finding boundary layer top: Application of a wavelet covariance transform to lidar backscatter profiles., Journal of Atmospheric & Oceanic Technology, 20, 2003.

Caicedo, V., Rappenglück, B., Lefer, B., Morris, G., Toledo, D., and Delgado, R.: Comparison of aerosol lidar retrieval methods for boundary layer height detection using ceilometer aerosol backscatter data, Atmospheric Measurement Techniques, 10, 2017.

Caliński, T. and Harabasz, J.: A dendrite method for cluster analysis, Communications in Statistics-theory and Methods, 3, 1–27, 1974.

Campbell, J. R., Hlavka, D. L., Welton, E. J., Flynn, C. J., Turner, D. D., Spinhirne, J. D., Scott, III, V. S., and Hwang, I. H.: Full-Time, Eye-Safe Cloud and Aerosol Lidar Observation at Atmospheric Radiation Measurement Program Sites: Instruments and Data Processing, Journal of Atmospheric and Oceanic Technology, 19, 2002.

Cimini, D., De Angelis, F., Dupont, J.-C., Pal, S., and Haeffelin, M.: Mixing layer height retrievals by multichannel microwave radiometer observations, 2013.

Cohn, S. A. and Angevine, W. M.: Boundary layer height and entrainment zone thickness measured by lidars and wind-profiling radars, Journal of Applied Meteorology, 39, 1233–1247, 2000.

Collaud Coen, M., Praz, C., Haefele, A., Ruffieux, D., Kaufmann, P., and Calpini, B.: Determination and climatology of the planetary boundary layer height above the Swiss plateau by in situ and remote sensing measurements as well as by the COSMO-2 model, Atmospheric Chemistry and Physics, 14, 13 205–13 221, 2014.

Davies, D. L. and Bouldin, D. W.: A cluster separation measure, IEEE transactions on pattern analysis and machine intelligence, pp. 224–227, 1979.

Davison, A. C. and Hinkley, D. V.: Bootstrap methods and their application, 1, Cambridge university press, 1997.

De Bruine, M., Apituley, A., Donovan, D., Klein Baltink, H., and de Haij, M.: Pathfinder: Applying graph theory for consistent tracking of daytime mixed layer height with backscatter lidar, Atmospheric Measurement Techniques Discussions, 10, 1893–1909, 2017.

Desgraupes, B.: Clustering indices, University of Paris Ouest-Lab Modal'X, 1, 34, 2013.

Dupont, J.-C., Haeffelin, M., Badosa, J., Elias, T., Favez, O., Petit, J., Meleux, F., Sciare, J., Crenn, V., and Bonne, J.: Role of the boundary layer dynamics effects on an extreme air pollution event in Paris, Atmospheric environment, 141, 571–579, 2016.

Flynn, C. J., Mendoza, A., Zheng, Y., and Mathur, S.: Novel polarization-sensitive micropulse lidar measurement technique, Opt. Express, 15, 2785–2790, 2007.

Freund, Y. and Schapire, R. E.: A decision-theoretic generalization of on-line learning and an application to boosting, Journal of computer and system sciences, 55, 119–139, 1997.

Gamage, N. and Hagelberg, C.: Detection and analysis of microfronts and associated coherent events using localized transforms, Journal of the atmospheric sciences, 50, 750–756, 1993.

Guo, J., Miao, Y., Zhang, Y., Liu, H., Li, Z., Zhang, W., He, J., Lou, M., Yan, Y., Bian, L., et al.: The climatology of planetary boundary layer height in China derived from radiosonde and reanalysis data, Atmospheric Chemistry and Physics, 16, 13 309, 2016.

Haefele, A., Hervo, M., Turp, M., Lampin J-L, Haeffelin, M., Lehmann, V., et al.: The E-PROFILE network for the operational measurement of wind and aerosol profiles over Europe, Proceedings of WMO Technical Conference on Meteorological and Environmental Instruments and Methods of Observation (CIMO TECO 2016, Madrid), 2016.

Haeffelin, M., Angelini, F., Morille, Y., Martucci, G., Frey, S., Gobbi, G., Lolli, S., O'dowd, C., Sauvage, L., Xueref-Rémy, I., et al.: Evaluation of mixing-height retrievals from automatic profiling lidars and ceilometers in view of future integrated networks in Europe, Boundary-Layer Meteorology, 143, 49–75, 2012.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Science & Business Media, 2009.

Hayden, K., Anlauf, K., Hoff, R., Strapp, J., Bottenheim, J., Wiebe, H., Froude, F., Martin, J., Steyn, D., and McKendry, I.: The vertical chemical and meteorological structure of the boundary layer in the Lower Fraser Valley during Pacific'93, Atmospheric Environment, 31, 2089–2105, 1997.

Hennemuth, B. and Lammert, A.: Determination of the atmospheric boundary layer height from radiosonde and lidar backscatter, Boundary-Layer Meteorology, 120, 181–200, 2006.

Herman, J. and Usher, W.: SALib: An open-source Python library for Sensitivity Analysis, The Journal of Open Source Software, 2, https://doi.org/10.21105/joss.00097, https://doi.org/10.21105/joss.00097, 2017.

Hintze, J. L. and Nelson, R. D.: Violin Plots: A Box Plot-Density Trace Synergism, The American Statistician, 52, 181–184, 1998.

Iooss, B. and Lemaître, P.: A review on global sensitivity analysis methods, in: Uncertainty management in simulation-optimization of complex systems, pp. 101–122, Springer, 2015.

Jain, A. K., Murty, M. N., and Flynn, P. J.: Data clustering: a review, ACM computing surveys (CSUR), 31, 264–323, 1999.

Kotthaus, S. and Grimmond, C. S. B.: Atmospheric boundary-layer characteristics from ceilometer measurements. Part 1: A new method to track mixed layer height and classify clouds, Quarterly Journal of the Royal Meteorological Society, 144, 1525–1538, 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105, 2012.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, nature, 521, 436–444, 2015.

Melfi, S., Spinhirne, J., Chou, S., and Palm, S.: Lidar observations of vertically organized convection in the planetary boundary layer over the ocean, Journal of climate and applied meteorology, 24, 806–821, 1985.

Menut, L., Flamant, C., Pelon, J., and Flamant, P. H.: Urban boundary-layer height determination from lidar measurements over the Paris area, Applied Optics, 38, 945–954, 1999.

Mohan, M., Bhati, S., Sreenivas, A., and Marrapu, P.: Performance evaluation of AERMOD and ADMS-urban for total suspended particulate matter concentrations in megacity Delhi, Aerosol and Air Quality Resarch, 11, 883–894, 2011.

Morille, Y., Haeffelin, M., Drobinski, P., and Pelon, J.: STRAT: An automated algorithm to retrieve the vertical structure of the atmosphere from single-channel lidar data, Journal of Atmospheric and Oceanic Technology, 24, 761–775, 2007.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Pollard, D. et al.: Strong consistency of k-means clustering, The Annals of Statistics, 9, 135–140, 1981.

Poltera, Y., Martucci, G., Collaud Coen, M., Hervo, M., Emmenegger, L., Henne, S., Brunner, D., and Haefele, A.: PathfinderTURB: an automatic boundary layer algorithm. Development, validation and application to study the impact on in-situ measurements at the Jungfraujoch, Atmospheric Chemistry and Physics Discussions, 2017.

Rieutord, T.: Sensitivity analysis of a filtering algorithm for wind lidar measurements, Ph.D. thesis, 2017.

Rieutord, T., Brewer, W. A., and Hardesty, R. M.: Automatic detection of boundary layer height using Doppler lidar measurements, 2014.

Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics, 20, 53–65, 1987.

Schapire, R. E.: Explaining adaboost, in: Empirical inference, pp. 37–52, Springer, 2013.

Seibert, P., Beyrich, F., Gryning, S.-E., Joffre, S., Rasmussen, A., and Tercier, P.: Review and intercomparison of operational methods for the determination of the mixing height, Atmospheric environment, 34, 1001–1027, 2000.

Seidel, D. J., Ao, C. O., and Li, K.: Estimating climatological planetary boundary layer heights from radiosonde observations: Comparison of methods and uncertainty analysis, Journal of Geophysical Research: Atmospheres, 115, 2010.

Seidel, D. J., Zhang, Y., Beljaars, A., Golaz, J.-C., Jacobson, A. R., and Medeiros, B.: Climatology of the planetary boundary layer over the continental United States and Europe, Journal of Geophysical Research: Atmospheres, 117, 2012.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V.: The AROME-France convective-scale operational model, Monthly Weather Review, 139, 976–991, 2011.

Selim, S. Z. and Ismail, M. A.: K-means-type algorithms: a generalized convergence theorem and characterization of local optimality, Pattern Analysis and Machine Intelligence, IEEE Transactions on, pp. 81–87, 1984.

Senff, C., Bösenberg, J., Peters, G., and Schaberl, T.: Remote sensing of turbulent ozone fluxes and the ozone budget in the convective boundary layer with DIAL and Radar-RASS: A case study, Contributions to atmospheric physics, 69, 161–176, 1996.

Sobol, I. M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Mathematics and computers in simulation, 55, 271–280, 2001.

Stull, R. B.: An introduction to boundary layer meteorology, vol. 13, Springer, 1988.

Toledo, D., Córdoba-Jabonero, C., and Gil-Ojeda, M.: Cluster analysis: A new approach applied to lidar measurements for atmospheric boundary layer height estimation, Journal of Atmospheric and Oceanic Technology, 31, 422–436, 2014.

Toledo, D., Córdoba-Jabonero, C., Adame, J. A., De La Morena, B., and Gil-Ojeda, M.: Estimation of the atmospheric boundary layer height during different atmospheric conditions: a comparison on reliability of several methods applied to lidar measurements, International journal of remote sensing, 38, 3203–3218, 2017.

Ware, J., Kort, E. A., DeCola, P., and Duren, R.: Aerosol lidar observations of atmospheric mixing in Los Angeles: Climatology and implications for greenhouse gas observations, Journal of Geophysical Research: Atmospheres, 121, 9862–9878, 2016.