Atmospheric Research 279 (2022) 106401

Contents lists available at ScienceDirect

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmosres

Estimation of the atmospheric boundary layer height by means of machine learning techniques using ground-level meteorological data

Francisco Molero *, Rubén Barragán, Begoña Artíñano

Department of Environment, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avda Complutense, 40, Madrid 28040, Spain

ARTICLE INFO

Keywords:
Atmospheric boundary layer
Machine learning
Ceilometer
MLH
ABLH

ABSTRACT

Aerosols affect air quality, weather and climate through many mechanisms and are dangerous to human health. They are mostly concentrated within the atmospheric boundary layer (ABL) its height is affected by the radiation emitted by the surface, causing turbulence and evolving along the day, influencing the vertical mixing of the air pollutants generated near the surface and therefore, their ground-level concentration from local sources. Lidars have demonstrated their capabilities to study the aerosol vertical distribution and their spatio-temporal evolution can provide very complete information on the ABL dynamics. In this work, machine learning techniques are employed to predict the ABL height. The meteorological variables measured at ground-level are used as features of the algorithm and the ABL height estimated by the STRATfinder algorithm using ceilometer profiles, a small lidar instrument with enhanced characteristics for unassisted continuous operation, are considered the truth in the supervised regression algorithm. The machine learning models allow considering combination of features in the regression algorithm and also allow characterizing the importance of each of the predictors to determine the final result. This property is used to study different boundary layer regimes. The ABL is difficult to study in certain parts of the day due to transitions between atmospheric regimes. In order to improve the performance of the model, each day was divided in four parts (nighttime, morning, daytime and evening). The Madrid ceilometer profile database has been studied for the year 2020, splitting the training datasets for the machine learning algorithm into season and part of the day, and the importance of predictors analyzed. Major influence of temperature and relative humidity is found in most of the situations, but also wind velocity in certain circumstances and pressure. The influence of radiation is small, contrary to expected. The main advantage of the proposed method is that MLHs and ABLHs can be retrieved directly from widely available ground-level meteorological data. Future work will focus on more relevant predictors, as latent heat or turbulence.

1. Introduction

The atmospheric boundary layer (ABL) regulates the exchange of energy and moisture between the surface and the atmosphere, playing a critical role in air quality forecasts (Monks et al., 2009) and greenhouse gas concentration budgets (Gerbig et al., 2008). It is defined as the layer located at the lowermost region of the troposphere that is directly influenced by the Earth's surface and responds to surface forcing over a short period of time (Stull, 1988). It is mainly characterized by turbulent processes and presents a daily evolution cycle depending on the solar radiation (Mahrt, 1999). The cycle, in clear-sky situations, starts with the increase of ground surface temperature after sunrise, which intensifies the convection, producing ascension of warm air masses and downward displacement of colder air masses, which creates a growing mixing layer (ML), named after the vertical mixing process generated by the ascending air parcels (White et al., 2009). During the early evening transition period, the gradual reduction of incoming solar irradiance causes a reduction of the convective processes and a weakening of the turbulence, producing a transition of the ML into two layers, a stable stratified boundary layer called the nocturnal boundary layer (NBL) close to the surface and a residual layer (RL) which is a remnant of the daytime ML and is just above the NBL. The next day starts with an early morning transition period from sunrise until the time when the NBL is eroded and a new ML begins to grow rapidly. The underlying surface plays also a crucial role for the ABL development, influenced by the surface albedo because different surfaces respond differently to the solar heating (Sailor, 1995). Another aspect that plays a role in shaping the boundary layer height is the orography, with different behavior between

* Corresponding author.
E-mail address: f.Molero@ciemat.es (F. Molero).

https://doi.org/10.1016/j.atmosres.2022.106401
Received 25 April 2022; Received in revised form 17 August 2022; Accepted 17 August 2022
Available online 19 August 2022

mountainous terrains with respect to urban or rural flat environment (Trentmann et al., 2009). The pollutants emitted by ground-level local sources into the ML are dispersed by turbulence, both horizontal and vertically, until they are completely mixed providing enough time is given and no significant sinks are present (Seibert et al., 2000). A detailed understanding of ABL processes would improve forecasting of pollution dispersion and cloud dynamics in the context of future climate scenarios, as it determines the available volume that the anthropogenic pollutants emitted at surface can occupy, affecting their concentration and consequently the air quality (Geiß et al., 2017). A strong aerosol gradient normally occurs at the ABL top at daytime due to turbulent vertical mixing process, the primary process by which aerosol particles are transported vertically in the atmosphere (Pal et al., 2010), and it can be used as proxy of the ABLH estimation. During other periods, the ABLH is more difficult to determine from the aerosol gradient, as the strongest aerosol gradient sometimes corresponds to the top of RL and others to growing layers within, yielding inaccuracy results.

Nowadays, the most common dataset used for ABLH determination are radiosoundings, the unique officially accepted measurements at global scale, but they are usually taken only twice per day (00:00 and 12:00 UTC) mainly at airports in western countries in the Northern Hemisphere, in the frame of the World Meteorological Organization (WMO) radiosounding global network (Durre and Yin, 2008). The infrequent observation and sparse spatial coverage compromise their representativeness for urban and regional scales. Better temporal coverage in order to capture the ABL diurnal cycle is needed to improve mesoscale analyses that are used to drive short-term model predictions of aerosol dispersion and reanalysis (Haeffelin et al., 2012). Different remote sensing techniques have been introduced into ABLH studies over the past several decades, improving the spatial and temporal capabilities thanks to their continuous operation and network developments (Baars et al., 2008). Among them, aerosol lidar has developed as a powerful remote sensing instrument to retrieve the ABLH through detecting aerosol vertical profiles (Pal et al., 2013). As mentioned above, aerosols can be used to trace the ABLH from the vertically resolved profiles. The aerosol lidar-derived ABLH is actually the height of aerosol layers and it can disagree with the radiosounding-derived ones due to the inconsistency between the thermal profiles and the aerosol profile, especially during morning or evening transitions (Emeis and Schäfer, 2006). The spatial coverage of lidar instruments has improved recently with the organization of networks, such as EARLINET (Pappalardo et al., 2014) and *E*-PROFILE (Illingworth et al., 2015; E-profile, 2021). This last network employs ceilometers, a small lidar with enhanced characteristics for unassisted continuous operation. The recently available high temporal and spatial density of these observations has driven a development of the ABLH retrieval algorithms for these instruments. The determination of the atmospheric layers using the aerosol profiles as proxy is based in two assumptions: firstly, the aerosols emitted at surface level are well-mixed within the ABL, with a cleaner upper free atmosphere, producing a strong negative gradient clearly observable in the backscatter profiles (Flamant et al., 2001); secondly, the interface presents considerable fluctuations in aerosol concentration, due to constant interexchange of airmasses, some clear from the free troposphere, moving downward, some polluted with aerosols, moving upward. This fact increases the variance of the backscatter profile at that height (Menut et al., 1999). The first assumption is applied in classical methodologies, such as gradient method (Flamant et al., 1997) and wavelet covariance transform (Brooks, 2003). The second is applied in the standard deviation analysis (Hooper and Eloranta (1986)). Further developments have applied both assumptions in combination (Emeis et al., 2008).

Considering the four different dynamic regimes mentioned above, it can be established that during the daytime period, with the mixed layer well developed, the ABLH is equal to the MLH because the distribution of aerosol concentration is dominated by turbulent mixing. On the other hand, when the surface heating is weaker than that of the previous day, the MLH may not reach the height attained by the previous day's ABL, showing some stratification. It can be determined by applying both assumptions, as the variance is particularly strong at the top of the mixed layer during daytime when air from the free troposphere is being mixed into the lowest atmospheric layer. But for the nighttime period, the height determined by aerosol lidar is either the nocturnal layer developed near the surface when it cools down after sunset or the top of a residual layer from the previous mixed layer left behind when the turbulence is “switched off”. The transition periods are more difficult to study. For instance, during the morning growth, variance can be particularly strong at the top of the ML because air from the residual layer may be mixed, but the gradients are weaker than when the air masses come from the free troposphere.

One recent advance in the field is the use of edge detection method (Poltera et al., 2017) based on temporal gradients in the attenuated backscatter signal. It started with the structure of the atmosphere (STRAT) method (Morille et al., 2007), that was extended in two dimensions (Temporal and vertical) by the STRAT-2D (Haeffelin et al., 2012) in order to guarantee temporal consistency of the resulted ABLH. A recent development, called pathfinder (de Bruine et al., 2017), applies graph theory to track the diurnal evolution. The combination of these methodologies has produced a reliable method, called STRATfinder, which applies a backward propagating layer, from the end of the day, and decides the type of layer from the forward and backward determinations by minimizing a cost function. This layer attribution is the most uncertain step (Haeffelin et al., 2012) and it can be assisted by commonly available surface measurements of radiation and temperature. More details are provided in (Kotthaus et al., 2020). Recent developments follow the line of two dimensional analysis, such as morphological image processing techniques (Vivone et al., 2021).

The MLH and ABLH provided by the STRATfinder algorithm have been employed as true values in machine learning algorithms in order to predict them from ground-based meteorological data. The main advantage of machine learning algorithms is the contribution from multiple features simultaneously (McGovern et al., 2017). Traditional fitting algorithms (for instance, linear regression) cannot consider the influence of multiple meteorological variables at the same time. This constraint is solved using machine learning approach, which has the potential for a fast, robust, accurate, and automated ABLH estimation. Machine learning models are becoming increasingly important due to the availability of large datasets, difficult to analyze with traditional methods. For instance, these models have proved faster and more reliable in automated inspection, defect detection, autonomous cars and predictive maintenance (Hastie et al., 2001). Beyond these most commonly recognized ones, they are starting to be used in other fields, such as atmospheric remote sensing, thanks to their generalization and fast training speed (Wei et al., 2019; de Moreira et al., 2022). Developing a predictive model follows a determined workflow, from data preparation, including cleansing, selection of the best algorithm for the problem, splitting the database into a dataset for training of the model and another for testing, and some feature discrimination and performance analysis in an iterative scheme. Identifying the right algorithm is often a process of trial and error because every problem requires a properly tuned machine learning algorithm. The adequate selection of the features is critical in the performance of the model, as models with larger number of features require more computational resources during the training stage, and using too many features leads to overfitting. Removing features without useful information or redundant optimizes performance of a simpler model less likely to overfit and with reduced computational cost. Feature selection is a process of selection the most relevant features for the specific problem that will capture the essential patterns in the data.

In this paper, we present an estimation of the MLH and ABLH from ground-based measurements using machine learning methods, validated with ceilometers profiles and STRATfinder estimations. The capability of machine learning models to consider multiple features to establish correlation between variables, and the use of widely available ground-level meteorological data are the main advantages of the proposed method.

Tha manuscript is organized as follows: Section 2 is a description of the location as well as a summary of all the instruments, datasets and the algorithm used in this work. Section 3 describes the statistical indicators (RMSE, R^2, MAE and predictors importance) for the different datasets, when machine learning is applied to the ABLH or MLH estimations. Finally, the results and discussion along with the main conclusions are presented in section 4.

2. Material and methods

2.1. Experimental site: Madrid

The experimental site belonging to the Department of Environment of the CIEMAT is located in the center of the Iberian Peninsula (40.45 N; 3.72 W; 669 m. a.s.l.) in the northwest part of Madrid city. The Madrid air basin is bordered to the north-northwest by a high mountain chain, Sierra de Guadarrama (max. Altitude 2420 m. a.g.l.), 40 km from the metropolitan area; to the south by another mountain system, Montes de Toledo, and finally to the northeast and east by lower mountainous terrain. The population of this metropolitan area including the Madrid city and surrounding towns is nearly 6 million inhabitants, one of the most populated regions in Spain. The Madrid air pollution plume is considered as typically urban, fed by traffic emissions and residential heating, given that industrial activity is comprised of light factories and it does not represent an important atmospheric pollutant source (Artíñano et al., 2003). The Madrid climate is continental Mediterranean, with hot dry summers and cold winters, and most days present clear-sky conditions (López et al., 2019). The Azores high-pressure system governs the atmospheric situation in these latitudes during a great part of the year and in winter high pressure systems over the Madrid area produces periods of stagnation with high stability, poor ventilation and increases in air pollution.

2.2. Ceilometer profiles and ABLH estimation algorithm (STRATfinder)

The site has in operation a Lufft CHM15k-Nimbus ceilometer since December 2019. This instrument employs the infrared light at 1064 nm from a pulsed Nd: YAG laser, emitting 60 mW per pulse of output power at a repetition frequency ranging between 5 and 7 kHz. Vertical profiles are obtained with a temporal resolution of 15 s and a vertical resolution of 15 m, reaching a maximum height of 15 km. Regarding the minimum height, the system is biaxial, so the laser beam enters into the telescope's field of view gradually, obtaining a complete overlap at about 1.5 km. Taking into account the laser beam divergence (0.3 mrad) and the telescope field of view (0.45 mrad), a correction function is applied to reduce the incomplete overlap, producing a useful signal down to 240 m (Molero and Jaque, 1999). The backscattered signal is detected with an avalanche photodiode in photon-counting mode, allowing the study of aerosols, which produce a return signal weaker than clouds.

The daily files are processed by means of the STRATfinder algorithm, available under the GNU General Public License v3.0. The temporal resolution is reduced to 10 min, in order to match the meteorological data resolution before analyzing both dataset using the tree regression algorithm. The STRATfinder algorithm provides estimations of the MLH and ABLH. It also estimates an auxiliary layer height that is tracked backwards in time from midnight to noon to assist ABLH detection during the evening decay of the mixed layer. The Dijkstra algorithm (Dijkstra, 1959) is applied to track MLH, ABLH and the auxiliary layer and individual paths are connected to determine MLH and ABLH for the whole 24 h period, merging the auxiliary layer and the preliminary ABLH to provide the final ABLH estimate. Cases of rain, snow and low clouds aren't included in the dataset, as they are filtered by the STRATfinder algorithm. The algorithm employs a fast Fourier transform function to avoid high signals associated with rain, snow and low clouds by setting a threshold and exclude periods when MLH cannot reliably be determined from attenuated backscatter profile observations from the analysis. Each day is divided into four parts, namely Nighttime (NT: Sunset +2 h until sunrise), Morning (MO: Sunrise until Sunrise +4 h), Daytime (DT: Sunrise +4 h until sunset – 2 h) and Evening (EV: Sunset – 2 h until Sunset +2 h) in order to simplify the model predictions taking into account the distinct characteristics of the ABL and ML development during the diurnal cycle.

Fig. 1 shows an example of the STRATfinder estimations for one day, 16 July 2020 in this case. The MLH (black crosses) is close to the minimum (232 m agl) during night and part of the morning, while the ABLH (red circles) follows the residual layer from the previous day, with some difficult decisions between 3:00 and 9:00 am due to layer appearance and disappearance. The most pronounced layer edge usually occurs between the ABL and the free troposphere above, but sometimes aerosol-rich layers appear in the ceilometer field of view, as it happens this day between 3:00 and 10:00, with a layer located between 500 m and 1 km. The convection starts at sunrise (4:59 am), observable in the figure as vertical cyan lines between 5:00 and 10:00, detecting the strong upward movement of airmasses. When the convective growing layer extends over the whole ABL, both estimations are equal, as it happens at 10:00, although later on, the algorithm estimates lower MLH from 11:00 to 15:00, probably due to some inner structure in the aerosol layer. Both estimations agree between 15:00 and 19:48, sunset time, when the MLH estimation drops to the first detected layer close to the ground, while the ABLH tracks the residual layer that slowly transitions into the next day. The parts of the day are labelled in white, showing the more difficult task of identifying the layers in the morning (MO) and evening (EV) parts, where the largest transitions occur.

2.3. Ground-based meteorological data

Meteorological information in Madrid was obtained at CIEMAT (Molero et al., 2014). Surface meteorological parameters, such as temperature, relative humidity, wind speed and wind direction were obtained from an automatic meteorological station (U3-NRC, Onset HOBO, USA) and recorded every 10 min, obtaining a total of 48,839 sets of 10 min averaged data. The temperature is measured at two heights (4 and 50 m) in order to characterize the inversion. Fig. 2 shows the temporal evolution of the estimated MLH (black crosses) and the ABLH (red circles) in the top panel (Fig. 2-A) for eight days of July 2020. As it can be seen, the boundary layer during those days of summer usually reaches 2 km in height, with a daily evolution starting at 7:00 UTC (9:00 local time, day saving hour applied) and growing along the morning. A sharp decrease is observed after sunset. Regarding the ABLH (red circles), it loosely follows the same daily dynamics, but remains at heights between 1 and 2 km overnight, as a reminder of the previous day mixing layer. Several prediction variables are plotted in Fig. 2.B (4 m temperature), 2.C (Relative humidity) and 2.D (Solar radiation), in order to compare the daily evolution of those variables with the atmospheric layers. As it can be observed, the temperature (Fig. 2.B) and solar radiation (Fig. 2.D) visually correlate with the MLH directly, although a delay in the growth of the ML is observed. The delay of the ABL growth each morning respect to the solar radiation and rise of temperature is a relevant feature, with this last one better correlated with the ABL peak but wider curve, while the solar radiation shows similar width curves but displaced respect to the ABL. Regarding the water vapor relative humidity, the correlation is inverse, but no delay is observed, with the peak of the ABL nearly coincident with the RH minimum. As it will be shown later, this has a distinct effect on the results, as the machine learning algorithm will take advantage of all these correlations to weight the importance of each predictor.

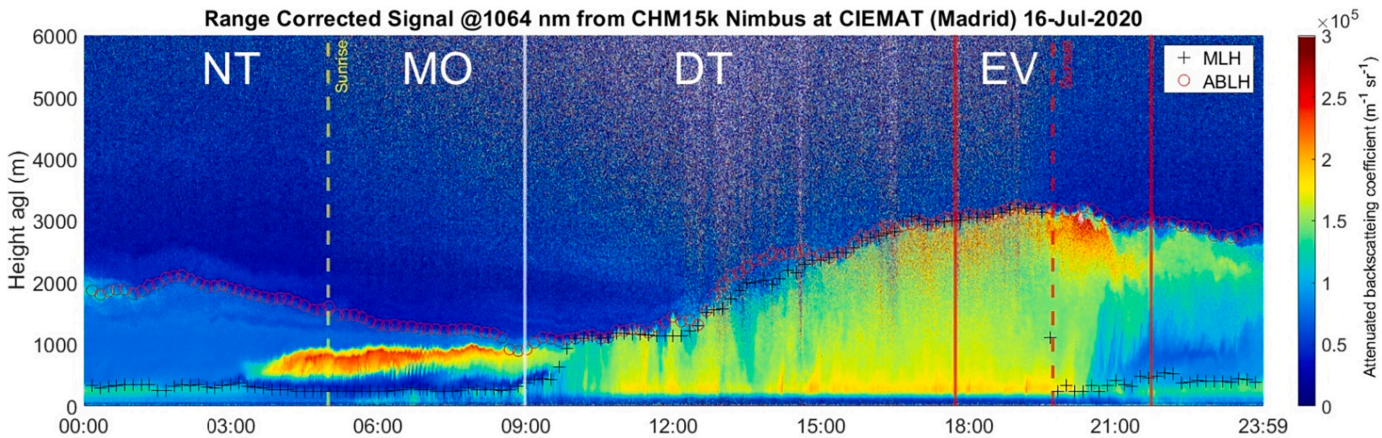

Fig. 1. Temporal evolution of the ABL the day 16 July 2020, obtained by plotting the attenuated backscatter in color scale, against UTC time (x-axis) and height (y-axis). Black crosses and red circles are the MLH and ABLH, respectively, estimated by the STRATfinder algorithm. Parts of the day are labelled and separated by solid lines, except sunrise, highlighted as a dash white line, and sunset as dash red line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

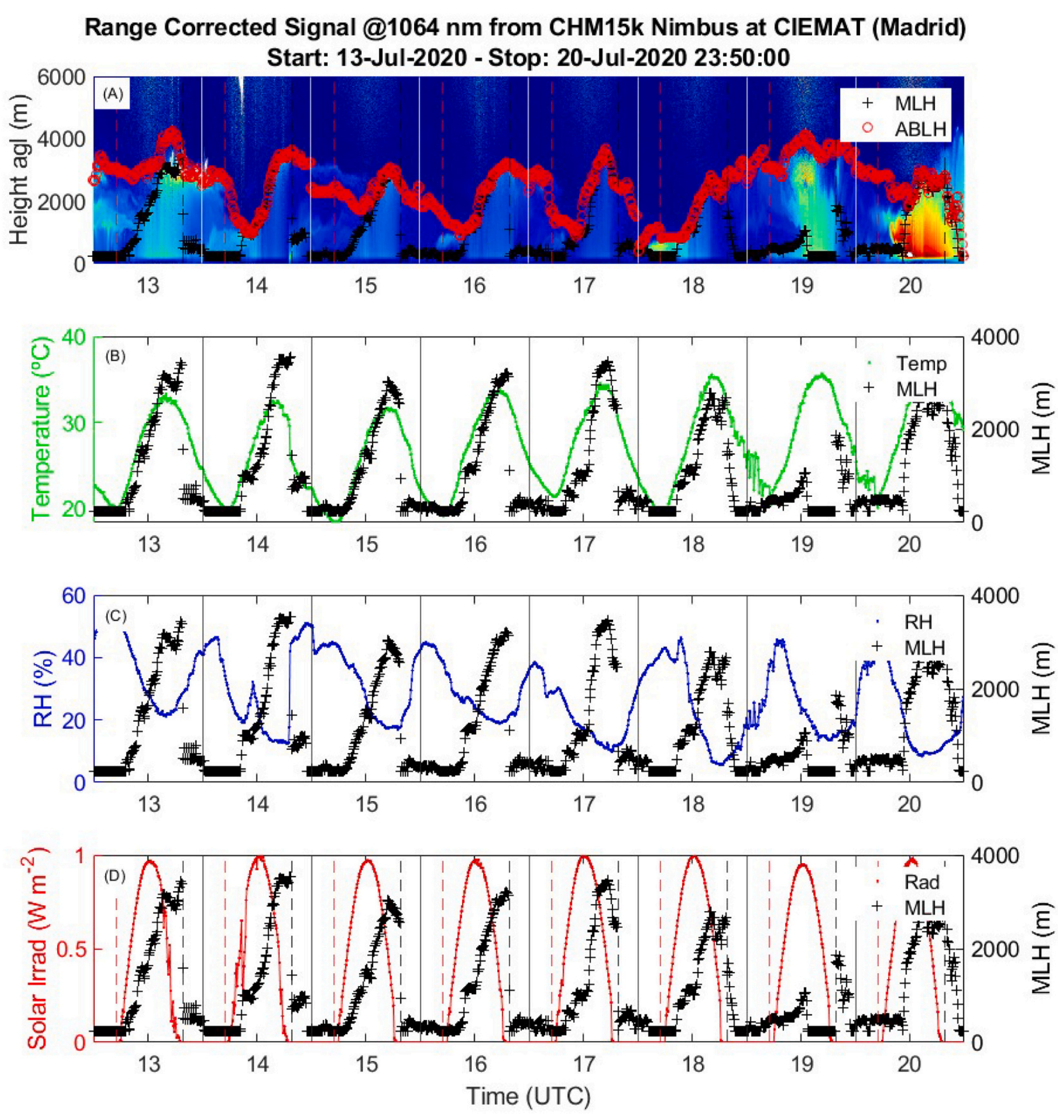

Fig. 2. Temporal evolution of atmospheric layers (A) and predictors: Temperature (B), RH (C) and solar irradiance (D) from the 13 to the 20 July 2020.

2.4. Machine learning

Since the algorithm will predict results within a continuous output, the MLH and ABLH, with a "true" response determined by the STRATfinder algorithm, it is a supervised regression problem. The one-year dataset from January 2020 to December 2020 is break down into three datasets, with percentages: training dataset: 60%, cross validation dataset: 20% and test dataset: 20% selected randomly, using cross-validation by the standard K-fold error estimation method. The testing dataset is only used to evaluate model performance, without being involved in the training of the model. The input variables provided by the meteorological station, are selected as features and normalized to avoid scale problems, using the mean and standard deviation of the training dataset. Several algorithms were tested (linear regression, regression tree, support vector machine and Gaussian process regression), and for each model, the performance was analyzed by inspecting the RMSE (Validation) score and also several diagnostic scores such as model accuracy, and plots, such as a response plot or residuals plot, obtaining the best results for regression tree. The function fitrtree within the Matlab environment is employed to obtain a regression tree based on the ground-level meteorological variables (also referred as predictors or features) and the MLH and ABLH values provided by the STRATfinder algorithm as true response. The returned tree is a binary tree where each branching node is split based on the values of a column of features. The trained model yields the importance of each predictor in a tree by summing changes in the node risk due to splits on every predictor, and then dividing the sum by the total number of branch nodes. These predictors importance will be used to determine the boundary layer dynamics. The accuracy of the model is usually measured by the mean absolute error (MAE) or the root mean square error (RMSE), accounting for the deviation between the retrieved results and the true values, and the squared correlation coefficient (R^2), characterizing the goodness of the fitting.

2.5. Dataset distribution

Statistics are split by part of the day using the above-mentioned division (NT, MO, DT & EV), and they are also split by season, with winter corresponding to December, January and February months, and labelled as DJF, and the other season accordingly as spring: MAM, summer: JJA and autumn: SON. Taking into account that data is averaged each 10 min. Table 1 summarizes the measurements employed in the study.

Data was obtained each 10 min and 48,439 datasets were recorded during 2020. It corresponds to 92.2% of the total possible data (52560). Some instrument failures occurred on 28–29 January, 4–14 April, 21. 26 May, 7–18 September and 29–31 October, explaining the slight differences in the amount of data for the four seasons, with the largest amount of data recorded in summer (JJA) and the smallest in spring (MAM), but the distribution is fairly constant along the year. Regarding the amount of data recorded during the different parts of the day, the morning (MO) and Evening (EV) periods show a constant percentage of the recorded data, around 16%, as it will be expected for those constant four hours periods. The nighttime (NT) and daytime (DT) periods show variation with the season, with the largest percentage of data recorded for the NT period during the winter (50.6%) and autumn (45.7%) seasons, and the opposite occurs for the DT period, with largest percentages in spring (30.1%) and summer (35.4%), due to the length of day changes along the year. Data availability by season and part of the day are sufficient for adequate model training and statistical analysis.

Table 1
Number of measurements, and percentage of total in brackets, by season and part of day. Percent values for the total measurements by season (All) refers to the total number of measurements, while those for each part of day refers to the measurements in the corresponding season.

	All	NT	MO	DT	EV
Winter (DJF)	12,871	6510	2136	2067	2158
	(26.6)	(50.6)	(16.6)	(16.0)	(16.8)
Spring	11,040	4051	1824	3317	1848
(MAM)	(22.8)	(36.7)	(16.5)	(30.1)	(16.7)
Summer	13,152	4115	2198	4655	2184
(JJA)	(27.2)	(31.3)	(16.7)	(35.4)	(16.6)
Autumn	11,376	5203	1896	2381	1896
(SON)	(23.5)	(45.7)	(16.7)	(20.9)	(16.7)
TOTAL	48,439 out of 52,560				
	(92.2)				

Firstly, the dataset is analyzed splitting by the part of the day (Table 2), taking into account the whole year. As indicators of the fitting quality, the root mean square error (RMSE), mean absolute error (MAE) and the coefficient of determination (R^2) are applied to quantify the agreement between STRATfinder and machine learning estimation, in order to assess the performance of each model. RMSE represents the difference between predicted and actual values, and the coefficient of determination quantifies the amount of variance explained by the model. After training the model, these scores are provided for the train dataset, shown at the top of Table 2, but these values are usually artificially good due to overfitting of the data. Also, the error of the dataset with which the model is trained will be lower than the error on any other dataset. In order to better assess the quality of the predictions, a second dataset, named test dataset, is used and the results are shown at the bottom of the table. Both set of scores allow the estimation of the effect of the overfitting and the amount corrected by the test dataset. The importance of the predictors, expressed as percentage, is also shown.

3. Results

3.1. MLH estimations

Fig. 3 shows the results for the training (3.A) and testing datasets (3.B) for the case of daytime and MLH. As it can be observed, the linear fit of the predicted data, respect to the true response, or STRATfinder values, shows a reasonable good agreement, with a R^2 of 0.96 for the training dataset and 0.66 for the testing dataset. The effect of the overfitting caused by the model is clearly observable in these two figures, with the lower correlation coefficient of the test dataset. The main discrepancies are due to wrong predictions of the lowest height (232.15 m, limited by the partial overlap of the laser and telescope fields of view in the ceilometer). The importance of the different predictors is shown in the bottom left panel (Fig. 3.C), with a strong importance of the relative humidity, due to the correspondence of the MLH highest value with the RH minimum, as it was mentioned before, and contributions from pressure, temperature and radiation. The residuals, not shown, produce

Table 2
Model quality indicators and predictors importance for MLH estimations split by parts of the day.

	NT	MO	DT	EV
Train rmse	108.81	87.82	168.30	193.17
Train R^2	0.90	0.86	0.96	0.95
Train MAE	47.19	38.11	90.83	94.66
Predictors (%)				
Temp 50 m	14.42	21.23	9.93	9.81
Temp 4 m	12.78	10.20	3.85	14.84
RH	20.49	20.82	45.51	26.52
Wind dir	17.95	7.91	5.50	11.88
Wind speed	7.07	7.21	5.31	6.38
Precipitation	0.12	0.39	0.15	0.01
Pressure	27.17	24.74	20.73	21.65
Radiation	0.00	7.49	9.01	8.91
Test rmse	267.69	174.99	454.15	474.24
Test R^2	0.36	0.28	0.66	0.69
Test MAE	128.57	95.16	261.56	244.32

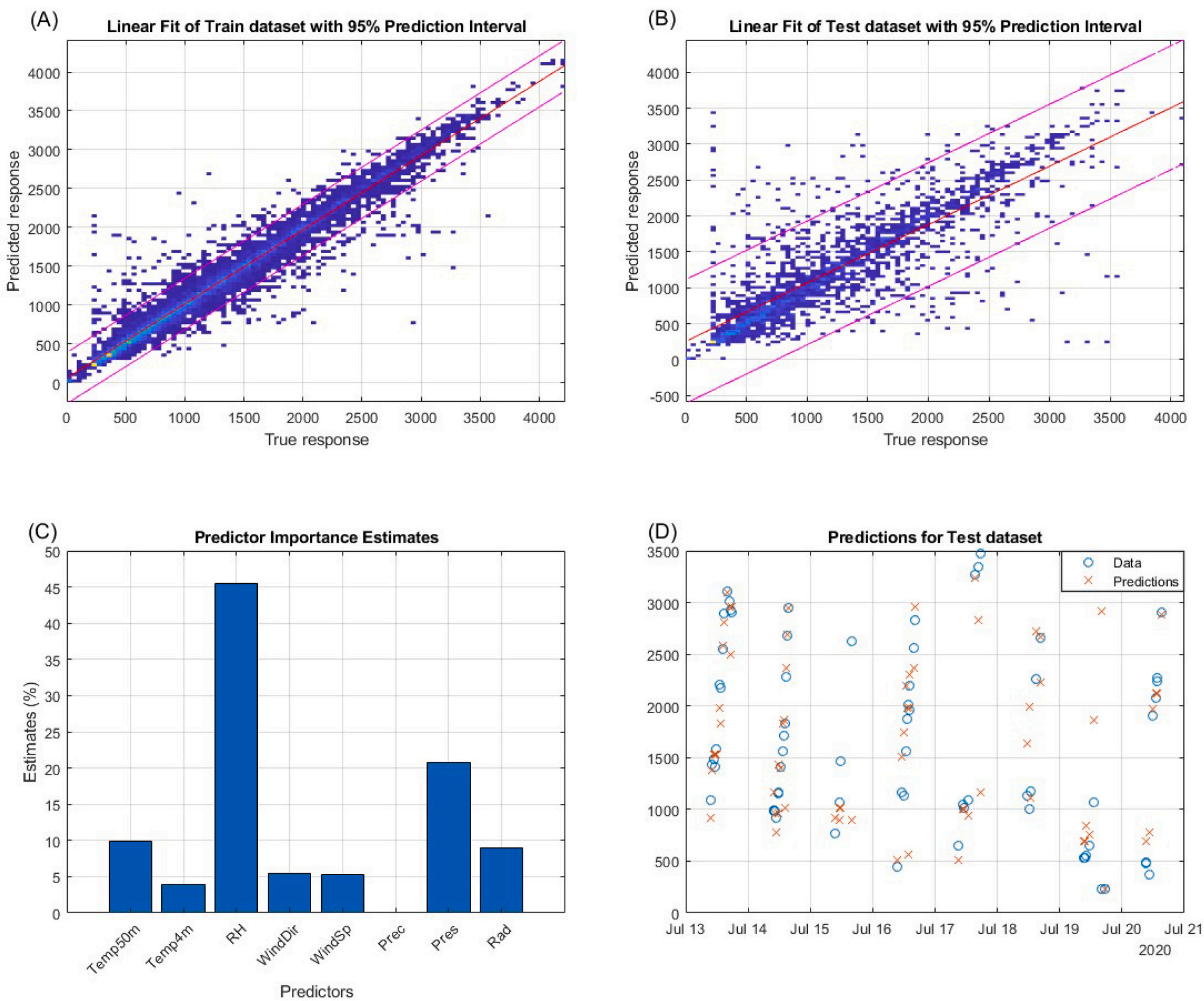

Fig. 3. Results from fitting the MLH for daytime cases for the whole year. (A) Predicted response, provided by the tree regression algorithm, vs true response, provided by the STRATfinder code, for the training dataset. (B) Same as A but for the testing dataset. (C) Predictors importance. (D) Comparison of MLH predictions vs true response for the time period 13–20 July, same as in Fig. 2.

reasonable Gaussian fit between 0.1 and 0.9 probability, and the tails of the distribution of the residuals diverging from the random error due to the effect of the minimum estimation value of the algorithm and the partial overlap of the ceilometer instrument at low heights. Fig. 3 D compares the predictions obtained from the model for the testing dataset between the 13 and 20 July, same period selected in Fig. 2. Reasonable agreement is obtained for the growing MLH from 1 to 3 km along the day, with occasional mistakes as in day 15, with predictions close to 1 km, while the ceilometer profiles reached 2.5 km, and opposite, on day 19, when the ceilometer data produces a strange low MLH, between 200 and 700 m probably due to algorithm error, but the model predicts a similar growth as the other days, reaching 3 km at 15:00 local time. The same analysis has been applied to each of the cases and the results summarized in the following tables (Tables 2, 3 and 4).

Table 2 shows the machine learning results for the MLH estimations for the different parts of the day, taking into account all the year. It can be observed that the performance of the model for NT and MO is rather poor, with R^2 equal 0.36 for NT and 0.277 for MO, when the test dataset is predicted. It improves for the other two parts of the day, with R^2 equals 0.66 for DT and 0.69 for EV. This indicates that the model

Table 3
Model quality indicators and predictors importance for MLH estimations split by season.

	DJF	MAM	JJA	SON
Train rmse	77.35	128.66	209.67	97.72
Train R^2	0.95	0.96	0.96	0.96
Train MAE	39.73	64.92	96.31	49.18
Predictors (%)				
Temp 50 m	17.06	14.14	5.48	8.06
Temp 4 m	7.03	7.25	34.38	18.33
RH	27.41	27.66	16.20	19.39
Wind dir	8.00	10.38	8.99	12.44
Wind speed	8.43	6.48	6.63	11.42
Precipitation	0.04	0.62	0.00	0.00
Pressure	22.65	17.93	9.60	21.22
Radiation	9.38	15.54	18.72	9.13
Test rmse	182.03	330.29	525.91	240.96
Test R^2	0.73	0.70	0.73	0.79
Test MAE	100.19	175.67	262.11	127.80

Table 4
Model quality indicators and predictor's importance for MLH estimations split by both the part of the day and season. Only Test dataset quality indicators are shown.

	Temp 50 m	RH	Wind sp	Pressure	Radiation	Rmse	R^2
DJF – NT	14.09	24.91	7.38	30.43	0.00	180.53	0.63
DJF – MO	14.93	32.79	6.06	29.65	6.14	111.42	0.54
DJF – DTs	8.22	26.90	39.42	10.24	6.87	212.53	0.67
DJF – EV	12.77	22.10	21.86	22.40	3.37	215.59	0.74
MAM – NT	14.87	26.94	9.93	18.82	0.00	284.11	0.42
MAM – MO	17.33	10.11	8.37	25.68	9.55	127.93	0.65
MAM – DT	23.26	29.53	8.72	20.37	8.55	323.26	0.69
MAM – EV	22.41	22.80	5.27	18.72	9.23	447.80	0.57
JJA – NT	14.23	23.80	12.71	13.22	0.00	287.18	0.03
JJA – MO	44.06	14.38	7.48	17.05	6.14	194.44	0.34
JJA – DT	13.99	27.51	9.54	14.66	20.55	567.31	0.54
JJA – EV	10.54	17.59	5.63	16.08	31.05	796.47	0.51
SON – NT	6.55	11.49	8.96	33.81	0.00	235.07	0.48
SON – MO	22.46	21.63	6.02	17.57	12.14	117.15	0.57
SON – DTs	7.69	38.10	17.07	23.03	5.40	257.46	0.75
SON – EV	11.32	24.41	17.03	25.49	6.95	292.87	0.77

performs better when the mixing layer is fully developed, justifying the discrimination of the different parts of the day. The predictor's importance indicates that at NT and MO, no main predictor influences the model, with similar importance for temperature, RH and pressure. On the other hand, when DT and EV are trained, the main predictor is RH, with less influence of pressure. As a curious note, it can be observed that radiation increases its importance as day progresses, with zero influence at NT, obviously, but also small influence at MO, due to the above-mentioned shift between the radiation intensity and the mixing layer growth.

Table 3 shows similar results than Table 2, but separating the results by season. In this case, the model performs reasonably well in all cases, with R^2 values always above 0.7. The predictors' importance is still shared between the temperature, RH and pressure, with more influence of temperature and radiation during spring and summer, and during autumn and winter, of relative humidity and pressure.

Finally, Table 4 summarizes the results when the analysis is discriminated by both the part of the day and the season. Due to space reasons, the table is reorganized in transpose mode, with the predictors' importance as columns and only the performance indicators for the test dataset are shown. The separation in both criteria improves the performance of the model for NT and MO, with R^2 values rising above 0.4 except for summer, with very low value at NT (0.03). This dataset was further analyzed in order to understand the difficulty for the model and it is related with the minimum MLH assigned by the STRATfinder algorithm, due to the partial overlap at low heights of the ceilometer instrument. The model overweights these values, yielding a high prediction of low MLH when the test dataset is employed. Further development, related with the handling of the low MLH, very frequent at nighttime, is required to improve this prediction.

As a relevant feature, the most important predictor for DT in winter is the wind velocity, with 39.42% of estimate of importance. This is the only case in which it happens, indicating that turbulence driven by wind is more important than that driven by heat. For the other seasons, the expected temperature contribution occurs in spring and summer, and a balanced combination of temperature, RH and pressure for autumn. The radiation is never the most important predictor, probably due to the mismatch in the duration mentioned above. Its importance grows along the day for summer, from NT, when it is zero as there is no radiation in this part of the day, to EV, when it reaches its largest value. For the other seasons, the progression is less clear, with largest value in the morning for autumn and spring and in the winter case, the DT value is larger than the evening one.

3.2. ABLH estimations

The STRATfinder algorithm also provides estimations for the ABLH. In this case, it follows better the aerosol layer detected by the ceilometer, although the layer identification using the back-propagating layer is also applied. The same machine learning analysis and study of the importance of the predictors can be done for this estimation.

Table 5 shows the performance of the model when the dataset is discriminated by the parts of the day. In this case, the four parts of the day attain reasonable performance, with R^2 values close to or larger than 0.7. The most important predictors for all the cases are temperature, followed by RH. It is noteworthy to observe that the radiation does not follow the pattern of increasing as day progresses, with largest value for DT and decreasing from there. This is caused by the different behavior of the ABL, in respect to the ML, as it can be seen in Fig. 1. The ML is estimated as the lowest gradient observed in the aerosol profile by the STRATfinder algorithm, after sunset, but the ABL remains estimated as the strongest gradient, normally signaling the separation of the aerosol layer and the cleaner free troposphere. The radiation cannot contribute to this last estimation, producing the difference respect to the MLH results.

When the dataset is separated by season, shown in Table 6, the model performs reasonably well, with R^2 values above 0.7, except for the case of summer, with $R^2 = 0.57$. The predictor's importance is still shared between the temperature, RH and pressure, with more influence of the temperature during summer, and pressure during autumn and winter. This feature can be a lousy proxy of the synoptic situation, but a better predictor should be found to establish such relationship.

Additional insight can be obtained by discriminating according to both the part of the day and the season, shown in Table 7. The

Table 5
Model quality indicators and predictors importance for ABLH estimations split by parts of the day.

	NT	MO	DT	EV
Train rmse	191.56	166.18	196.16	172.64
Train R^2	0.96	0.97	0.96	0.97
Train MAE	98.55	87.16	105.99	91.39
Predictors (%)				
Temp 4 m	7.88	48.79	44.88	51.33
Temp 50 m	41.13	4.89	4.54	3.34
RH	14.22	9.90	13.83	9.16
Wind dir	8.62	6.95	7.15	7.64
Wind speed	7.06	5.91	4.53	4.42
Precipitation	0.02	0.13	0.04	0.07
Pressure	21.07	18.90	19.74	22.53
Radiation	0.00	4.53	5.29	1.50
Test rmse	514.44	537.36	493.75	489.16
Test R^2	0.73	0.69	0.74	0.78
Test MAE	277.91	277.07	289.05	246.01

Table 6
Model quality indicators and predictors importance for ABLH estimations split by season.

	DJF	MAM	JJA	SON
Train rmse	166.52	170.35	199.13	132.61
Train R^2	0.95	0.96	0.94	0.97
Train MAE	77.64	92.28	111.59	70.34
Predictors				
Temp 50 m	10.17	22.38	34.19	23.18
Temp 4 m	8.83	6.06	4.94	7.20
RH	19.69	23.68	15.16	19.11
Wind dir	10.74	13.09	12.58	8.30
Wind speed	18.27	6.36	8.53	12.42
Precipitation	0.10	0.10	0.00	0.00
Pressure	28.09	23.04	16.86	24.26
Radiation	4.12	5.30	7.75	5.52
Test rmse	376.27	463.49	546.36	358.61
Test R^2	0.73	0.72	0.57	0.78
Test MAE	196.74	259.97	315.10	186.32

Table 7
Model quality indicators and predictor's importance for ABLH estimations split by both the part of the day and season. Only Test dataset quality indicators are shown.

	Temp 4 m	RH	Wind sp	Pressure	Radiation	Rmse	R^2
DJF – NT	11.18	27.17	18.73	19.71	0.00	397.93	0.69
DJF – MO	9.64	22.08	7.78	38.63	3.19	260.08	0.87
DJF - DTs	14.87	10.99	26.17	18.79	12.28	358.63	0.75
DJF - EV	10.74	26.08	15.34	26.29	1.15	326.07	0.77
MAM – NT	18.65	21.03	4.68	29.16	0.00	392.47	0.81
MAM - MO	27.41	16.50	4.90	30.19	8.28	408.66	0.75
MAM - DT	32.49	13.57	8.39	24.08	6.44	424.26	0.71
MAM - EV	25.00	15.27	7.52	34.22	0.91	475.51	0.71
JJA – NT	9.07	18.57	6.35	21.95	0.00	547.36	0.52
JJA – MO	13.43	14.32	10.02	18.09	7.08	539.91	0.65
JJA – DT	34.01	16.18	4.89	18.21	11.84	560.77	0.52
JJA – EV	27.26	20.07	8.45	24.60	1.22	483.98	0.52
SON – NT	11.20	17.47	5.11	29.46	0.00	399.67	0.72
SON – MO	31.49	12.79	8.43	21.61	5.03	303.86	0.83
SON - DTs	25.97	22.91	9.01	21.04	9.32	407.19	0.69
SON - EV	15.09	11.72	16.17	24.98	0.91	379.73	0.81

performance of the model is again poor for summer, with R^2 values below 0.6 except for the MO case. Studying the predictor's importance, it can be seen that in this case, the temperature is the most important predictor, while in the other cases, Pressure, RH and temperature attain similar values. Further study of this behavior is required to improve the prediction for summer. Similarly, to the MLH, the wind speed predictor is important at daytime in winter, with 26.17% of estimate of importance, indicating that turbulence driven by wind is more important than that driven by heat for both the MLH and ABLH.

4. Discussion and conclusions

This study proposes a machine learning approach for atmospheric boundary layer heights estimation using meteorological data. ABLH estimations provided by the STRATfinder algorithm using ceilometer vertical profiles are considered the true values for the supervised regression model. Ground-level meteorological variables (temperature, pressure, relative humidity, solar radiation, wind direction and speed and precipitation) are employed as predictors in the machine learning model. The complete year 2020 is analyzed, with data averaged 10 min, obtaining a total number of points of 48,480 (92.2% of total possible data). In order to study the different atmospheric regimes, diurnal and seasonal variations have been investigated. For the MLH, the performance of the model was better for DT and EV than for NT and MO, indicating that the model performs better when the mixing layer is fully developed. This is supported by the results for the ABLH, where the four parts of the day attain reasonable performance. When the results are separated by season, the model performs reasonably well in all cases for the MLH, but all except summer for the ABLH. This may be related to synoptic situation, but a better predictor should be found to establish such relationship. The importance of predictors, an estimation of the error provided by changes in the order, was used to analyze the results. Several relevant aspects were obtained; an important contribution to the regression was obtained for relative humidity in most of the cases. A potential explanation can be the better correlation between these data and the MLH and ABLH estimations than for the other variables, especially radiation, although it is negatively correlated in this case. The influence of temperature is smaller than expected, although major contribution was obtained in daytime. Only a different dynamic was obtained in winter at daytime, with the largest importance of the wind speed, indicating a mechanically driven mixing layer. Finally, an unexpectedly limited influence of radiation was found, probably caused by mismatch in periodic duration, which affects the training of the model. The major advantage of the proposed method is that MLHs and ABLHs can be retrieved directly from widely available ground-level meteorological data. Future work includes the study of other features more relevant to the ABL dynamics, such as latent heat, turbulence or meteorological situations. One challenge is the input of that information into the machine learning algorithm in adequate form. Also, no time correlation was considered among the measurements, as the data partition was done randomly. It can be considered in future studies with a better designed partition. Additionally, further efforts should focus on improving predictions in the transition times, when growing or decaying turbulent mixing occur. This approach may help the remote sensing techniques in the challenging task of layer attribution.

CRediT authorship contribution statement

Francisco Molero: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. **Rubén Barragán:** Conceptualization, Formal analysis, Investigation, Methodology, Writing – review & editing. **Begoña Artíñano:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by H2020 programme from the European Union (grant 654109, ACTRIS-2 project), the Spanish Ministry of Economy and Competitiveness (CRISOL, CGL2017-85344-R, AtPollen-Fluo, PID2020-117873RB-I00) and Madrid Regional Government (TIGAS-CM, Y2018/EMT-5177).

References

Artíñano, B., Salvador, P., Alonso, D.G., Querol, X., Alastuey, A., 2003. Anthropogenic and natural influence on the PM10 and PM2.5 aerosol in Madrid (Spain). Analysis of high concentration episodes. Environ. Pollut. 125 (3), 453–465. https://doi.org/10.1016/S0269-7491(03)00078-2.

Baars, H., Ansmann, A., Engelmann, R., Althausen, D., 2008. Continuous monitoring of the boundary-layer top with lidar. Atmos. Chem. Phys. 8 (23), 7281–7296. https://doi.org/10.5194/acp-8-7281-2008.

Brooks, I.M., 2003. Finding Boundary Layer top: Application of a Wavelet Covariance Transform to Lidar Backscatter Profiles. J. Atmos. Ocean. Technol. 20 (8), 1092–1105. https://doi.org/10.1175/1520-0426(2003)020<1092:FBLTAO>2.0.CO;2.

de Bruine, M., Apituley, A., Donovan, D.P., Klein Baltink, H., de Haij, M.J., 2017. Pathfinder: applying graph theory to consistent tracking of daytime mixed layer height with backscatter lidar. Atmos. Measure. Tech. 10 (5), 1893–1909. https://doi.org/10.5194/amt-10-1893-2017.

de Moreira, G.A., Sánchez-Hernández, G., Guerrero-Rascado, J.L., Cazorla, A., Alados-Arboledas, L., 2022. Estimating the urban atmospheric boundary layer height from remote sensing applying machine learning techniques. Atmos. Res. 266, 105962. ISSN 0169-8095. https://doi.org/10.1016/j.atmosres.2021.105962.

Dijkstra, E.W., 1959. A note on two problems in connexion with graphs. Numer. Math. 1 (1), 269–271. https://doi.org/10.1007/BF01386390.

Durre, I., Yin, X., 2008. Enhanced radiosonde data for studies of vertical structure. Bull. Am. Meteorol. Soc. 89 (9), 1257–1262. http://www.jstor.org/stable/26220887.

Emeis, S., Schäfer, K., 2006. Remote sensing methods to investigate boundary-layer structures relevant to air pollution in cities. Bound.-Layer Meteorol. 121 (2) https://doi.org/10.1007/s10546-006-9068-2.

Emeis, S., Schäfer, K., Münkel, C., 2008. Surface-based remote sensing of the mixing-layer height a review. Meteorol. Z. 17 (5), 621–630. https://doi.org/10.1127/0941-2948/2008/0312.

E-profile, 2021. available at. https://www.eumetnet.eu/activities/observations-programme/current-activities/e-profile/. Accessed: 28 January 2022.

Flamant, C., Georgelin, M., Menut, L., Pelon, J., Bougeault, P., 2001. The atmospheric boundary-layer structure within a cold air outbreak: comparison of in situ, lidar and satellite measurements with three-dimensional Simulations. Bound.-Layer Meteorol. 99 (1), 85–103. https://doi.org/10.1023/A:1018722923349.

Flamant, C., Pelon, J., Flamant, P.H., Durand, P., 1997. Lidar determination of the entrainment zone thickness at the top of the unstable marine atmospheric boundary layer. Bound.-Layer Meteorol. 83 (2), 247–284.

Geiß, A., Wiegner, M., Bonn, B., Schäfer, K., Forkel, R., von Schneidemesser, E., Münkel, C., Chan, K.L., Nothard, R., 2017. Mixing layer height as an indicator for urban air quality? Atmos. Measure. Tech. 10 (8), 2969–2988. https://doi.org/10.5194/amt-10-2969-2017.

Gerbig, C., Körner, S., Lin, J.C., 2008. Vertical mixing in atmospheric tracer transport models: error characterization and propagation. Atmos. Chem. Phys. 8 (3), 591–602. https://doi.org/10.5194/acp-8-591-2008.

Haeffelin, M., Angelini, F., Morille, Y., Martucci, G., Frey, S., Gobbi, G.P., Lolli, S., O'Dowd, C.D., Sauvage, L., Xueref-Rémy, I., Wastine, B., Feist, D.G., 2012. Evaluation of mixing-height retrievals from automatic profiling lidars and ceilometers in view of future integrated networks in Europe. Bound.-Layer Meteorol. 143 (1), 49–75. https://doi.org/10.1007/s10546-011-9643-z.

Hastie, T.J., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media, Stanford, California, USA.

Hooper, W.P., Eloranta, E.W., 1986. Lidar measurements of wind in the planetary boundary layer: the method, accuracy and results from joint measurements with Radiosonde and Kytoon. J. Appl. Meteorol. Climatol. 25 (7), 990–1001. https://doi.org/10.1175/1520-0450(1986)025<0990:LMOWIT>2.0.CO;2.

Illingworth, A.J., Cimini, D., Gaffard, C., Haeffelin, M., Lehmann, V., Löhnert, U., O'Connor, E.J., Ruffieux, D., 2015. Exploiting existing ground-based remote sensing networks to improve high-resolution weather forecasts. Bull. Am. Meteorol. Soc. 96 (12), 2107–2125. https://doi.org/10.1175/BAMS-D-13-00283.1.

Kotthaus, S., Haeffelin, M., Drouin, M.-A., Dupont, J.-C., Grimmond, S., Haefele, A., Hervo, M., Poltera, Y., Wiegner, M., 2020. Tailored algorithms for the detection of the atmospheric boundary layer height from common automatic lidars and ceilometers (ALC). Remote Sens. 12, 3259.

López, V., Salvador, P., Artíñano, B., Gómez-Moreno, F.J., Fernández, J., Molero, F., 2019. Influence of the origin of the air mass on the background levels of atmospheric particulate matter and secondary inorganic compounds in the Madrid air basin. Environ. Sci. Pollut. Res. 26, 30426–30443.

Mahrt, L., 1999. Stratified atmospheric boundary layers. Bound.-Layer Meteorol. 90 (3), 375–396. https://doi.org/10.1023/A:1001765727956.

McGovern, A., Elmore, K.L., Gagne, D.J., Haupt, S.E., Karstens, C.D., Lagerquist, R., Smith, T., Williams, J.K., 2017. Using artificial intelligence to improve real-time decision-making for high-impact weather. Bull. Am. Meteorol. Soc. 98 (10), 2073–2090. https://doi.org/10.1175/BAMS-D-16-0123.1.

Menut, L., Flamant, C., Pelon, J., Flamant, P.H., 1999. Urban boundary-layer height determination from lidar measurements over the Paris area. Appl. Opt. 38 (6), 945–954. https://doi.org/10.1364/AO.38.000945.

Molero, F., Jaque, F., 1999. The laser as a tool in environmental problems. Opt. Mater. 13 (1), 167–173. https://doi.org/10.1016/S0925-3467(99)00026-9.

Molero, F., Andrey, F.J., Fernandez, A.J., Parrondo, M.C., Pujadas, M., Córdoba-Jabonero, C., Revuelta, M.A., Gomez-Moreno, F.J., 2014. Study of vertically resolved aerosol properties over an urban background site in Madrid (Spain). Int. J. Remote Sens. 35 (6), 2311–2326. https://doi.org/10.1080/01431161.2014.894664.

Monks, P.S., Granier, C., Fuzzi, S., Stohl, A., Williams, M.L., Akimoto, H., Amann, M., Baklanov, A., Baltensperger, U., Bey, I., Blake, N., Blake, R.S., Carslaw, K., Cooper, O.R., Dentener, F., Fowler, D., Fragkou, E., Frost, G.J., Generoso, S., von Glasow, R., 2009. Atmospheric composition change – global and regional air quality. Atmos. Environ. 43 (33), 5268–5350. https://doi.org/10.1016/j.atmosenv.2009.08.021.

Morille, Y., Haeffelin, M., Drobinski, P., Pelon, J., 2007. STRAT: an Automated Algorithm to Retrieve the Vertical Structure of the Atmosphere from Single-Channel Lidar Data. J. Atmos. Ocean. Technol. 24 (5), 761–775. https://doi.org/10.1175/JTECH2008.1.

Pal, S., Behrendt, A., Wulfmeyer, V., 2010. Elastic-backscatter-lidar-based characterization of the convective boundary layer and investigation of related statistics. Ann. Geophys. 28 (3), 825–847. https://doi.org/10.5194/angeo-28-825-2010.

Pal, S., Haeffelin, M., Batchvarova, E., 2013. Exploring a geophysical process-based attribution technique for the determination of the atmospheric boundary layer depth using aerosol lidar and near-surface meteorological measurements. J. Geophys. Res. Atmos. 118, 9277–9295.

Pappalardo, G., Amodeo, A., Apituley, A., Comeron, A., Freudenthaler, V., Linné, H., Ansmann, A., Bösenberg, J., D'Amico, G., Mattis, I., Mona, L., Wandinger, U., Amiridis, V., Alados-Arboledas, L., Nicolae, D., Wiegner, M., 2014. EARLINET: towards an advanced sustainable European aerosol lidar network. Atmos. Measure. Tech. 7 (8), 2389–2409. https://doi.org/10.5194/amt-7-2389-2014.

Poltera, Y., Martucci, G., Collaud Coen, M., Hervo, M., Emmenegger, L., Henne, S., Brunner, D., Haefele, A., 2017. PathfinderTURB: an automatic boundary layer algorithm. Development, validation and application to study the impact on in situ measurements at the Jungfraujoch. Atmos. Chem. Phys. 17 (16), 10051–10070. https://doi.org/10.5194/acp-17-10051-2017.

Sailor, D.J., 1995. Simulated urban climate response to modifications in surface albedo and vegetative cover. J. Appl. Meteorol. (1988–2005) 34 (7), 1694–1704. http://www.jstor.org/stable/26187281.

Seibert, P., Beyrich, F., Gryning, S.-E., Joffre, S., Rasmussen, A., Tercier, P., 2000. Review and intercomparison of operational methods for the determination of the mixing height. Atmos. Environ. 34 (7), 1001–1027. https://doi.org/10.1016/S1352-2310(99)00349-0.

Stull, R.B., 1988. An Introduction to Boundary Layer Meteorology. Xiii + 666 Pp, 126. Kluwer Academic Publishers, Dordrecht, Boston, London, pp. 308–309. https://doi.org/10.1017/S0016756800022433 (3).

Trentmann, J., Keil, C., Salzmann, M., Barthlott, C., Bauer, H.-S., Schwitalla, T., Lawrence, M.G., Leuenberger, D., Wulfmeyer, V., Corsmeier, U., Kottmeier, C., Wernli, H., 2009. Multi-model simulations of a convective situation in low-mountain terrain in Central Europe. Meteorog. Atmos. Phys. 103 (1), 95–103. https://doi.org/10.1007/s00703-008-0323-6.

Vivone, G., D'Amico, G., Summa, D., Lolli, S., Amodeo, A., Bortoli, D., Pappalardo, G., 2021. Atmospheric boundary layer height estimation from aerosol lidar: a new approach based on morphological image processing techniques Atmos. Chem. Phys. 21, 4249–4265. https://doi.org/10.5194/acp-21-4249-2021.

Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., Cribb, M., 2019. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. Remote Sens. Environ. 231, 111221 https://doi.org/10.1016/j.rse.2019.111221.

White, J.M., Bowers, J.F., Hanna, S.R., Lundquist, J.K., 2009. Importance of using Observations of Mixing Depths in order to Avoid large Prediction Errors by a Transport and Dispersion Model. J. Atmos. Ocean. Technol. 26 (1), 22–32. https://doi.org/10.1175/2008JTECHA1134.1.