**Machine Learning Estimation of Mixing Layer Height in the Chesapeake Bay**

Rahim Kamara1, Amanze Ejiogu2, Ruben Delgado2, Vanessa Caicedo2, and Belay Demoz2

1 - University of Maryland, College Park

2 - University of Maryland, Baltimore County

July 26, 2022

Abstract

The mixing layer height (MLH) is important to understand the transport and dispersion of pollutants emitted from different sources within the planetary boundary layer. Understanding its evolution is critical to air pollution, weather, and climate change. Under- and overestimation can lead to inaccurate forecasts. The MLH height can be determined by evaluating potential temperature (θ), relative humidity (RH), specific humidity (q) and atmospheric refractivity (N) by identifying turning points or gradients in a rawinsonde profile. However, considerable differences have been observed in these methods. To determine more efficiently the MLH from rawinsonde profiles, meteorological parameters gathered from rawinsonde data obtained from the Baltimore-Washington DC metropolitan area during June 29, 2021, to July 6, 2021 are used to make the best estimate for determining the MLH for the Chesapeake Bay area. Machine learning will also be used to draw inference from patterns in data to better estimate the MLH.

**1. Introduction**

Aerosols may be small and come in many forms such as fine particles, like smoke, ash, and pollution or coarse with particles larger than one micrometer, like dust and sea salt. These aerosols have the capability to affect large areas of the globe. Aerosols can both occur through natural processes such as ocean spray, volcanic activity, and anthropogenic activity, like biomass burning and pollution from fossil fuel combustion engines producing NO2 (Judd et al., 2020). Aerosols serve as cloud condensation nuclei where water droplets can form and precipitation to occur. Aerosols can both cool the surface by reflecting incoming solar radiation warm the atmosphere above due to their properties. The increase in aerosols have various radiative and cloud effects. As the aerosol concentration increases within the atmosphere, a warming or cooling effect can act to destabilize or stabilize the atmosphere respectively.

As the aerosol concentration increases within the atmosphere, the amount of solar radiation that reaches the surface may decrease and thus stabilize the atmosphere. This leads to both less evaporation of water vapor and less convective energy present near the surface. Resulting from this, warm clouds become suppressed which stabilizes the planetary boundary layer (PBL) (Li et al., 2017). The PBL is the layer of the atmosphere that has direct contact with and influence by the surface of the earth. Above the PBL is the free atmosphere (FA) as shown in Figure 1, which has little to no influence on Earth’s surface. The PBL is characterized by moisture, while the FA has a lack thereof. During the daytime, the location of atmosphere where the PBL height (PBLH) can be found is at or below the entrainment zone (EZ) where it meets the mixed layer (ML), and during the nighttime, the height at which the boundary layer is found is where the stable boundary layer (SBL) and residual layer (RL) meet. If wind shear is present due to the nighttime low level jet, the thickness of EZ and capping inversion (CI) can increase which complicates the PBLH estimate (Carroll et al., 2019). The mixing layer height (MLH) will be the focus of this study. The MLH describes the day-time convective boundary layer (CBL) which is affected by turbulent mixing.

Diagram, engineering drawing

Description automatically generated

**Figure 1**

In (a), the daytime layers include free atmosphere (FA), entrainment zone (EZ), capping inversion height (zi), Mixed layer (ML) which can be interchanged with Convective boundary layer (CBL), surface layer (SL). On the x-axis is temperature (T), potential temperature (θ), q is specific humidity (q), geostrophic winds (Vg). The top of the ML would be where the MLH would be found. In (b), the night layers include FA which is free atmosphere (Not influenced by the surface), capping inversion (CI), residual layer (RL) which contains dying or zero turbulence and the residual heat, moisture, and pollutants that were mixed there during the previous day, stable boundary layer (SBL) which forms at night near the ground in response to the cooling of the air. The top of the SBL would be where the MLH would be found. (Wallace and Hobbs, 2006).

There exist several ways to determine the PBLH for a given area and time of day. These instruments include the following: rawinsonde, which has small payload size and is launched by weather balloon to measure atmospheric parameters such as air pressure, height, temperature, dew point temperature, wind speed, and wind direction; ceilometer, which is a ground-based instrument used to measure both the height of a cloud’s base and top (Wiegner et al., 2014); and ground-based Light Detection and Ranging (lidar), which uses the concept of photons to measure the strength of scattering of particles in the atmosphere or the time delay between transmission and reception of the signal (Caicedo et al., 2020). There are advantages and disadvantages to these methods of retrieval. For example, rawinsondes are usually launched at 0 and 12 UTC which may not be ideal because depending on the location, the MLH may not have reached maximum height. Rawinsondes might be capturing inversions associated with nocturnal stable layers and/or residual layers. Ceilometers can measure throughout the entire day and can validate rawinsonde PBLH retrievals.

Section 2 will describe the data used to determine the MLH, section 3 will discuss the procedure for determining the MLH from vertical sounding profiles, section 4 will discuss the results of using rawinsonde profiles for MLH retrieval and uncertainties associated with it, section 5 will explain how machine learning is applied to attempt to make forecast on MLH, section 6 will explain the results of the machine learning tool, and section 7 gives conclusion to this study.

**2. Description of data**

This data was obtained from a previous project measuring mixing layer heights for Dulles, IAD during the period of June 29, 2021, to July 6, 202. During this period, the mixing layer was either found or could not be determined using the method in the journal article by Wang, X. Y., & Wang, K. C. (2014) excluding the 1-2-1 smoother. If this project were to be continued, the 1-2-1 smoother should be applied to the data. A total of 10 rawinsonde profiles during this period exhibited a clear indicator of mixing layer height. The dataset are columns are datehour (YearMonthDateUTCHour), pressure (hPa), height (meter), temperature (Celsius), dewpoint (Celsius), direction (degrees), speed (knot), u wind component (knot), v wind component (knot), Airport Station (number), latitude, longitude, elevation (feet), and pw. The units for the variable pw are not certain.

**3. Method to determine MLH from rawinsonde**

Turbulence, or unsteady movement of air, that exists due to the heating of the ground and convection creates an inversion between the upper FA and lower PBL in the troposphere. This occurrence during the daytime is commonly referred to as the CBL. The inversion between the FA and the PBL during the nighttime is due to radiative cooling at the ground which forms the stable boundary layer (SBL) (Liu & Liang, 2010).

There exist many methods for finding the MLH from sounding profile which typically involves observing gradients in potential temperature (θ), relative humidity (RH), or specific humidity (q) (Piironen and Eloranta, 1995; Hennemuth and Lammert, 2006; Liu and Liang; Caicedo et al., 2020). Furthermore, authors X.Y. Wang and K.C. Wang (2014) additionally observed gradients of vertical atmospheric reflectivity (n) profiles.

θ is the temperature a parcel of air would have if brought to 1000 hPa with no effect by heat. θ is useful to show atmospheric stability. The MLH is found at the transition from an unstable region below to a stable region above with an increase in height (Seidel et al., 2010). RH is the ratio of the actual mixing ratio of the air to the saturation mixing ratio occurring at the same temperature and pressure. The variable q is the ratio of the mass of water vapor to the total mass of the system. The PBL height is typically characterized by warm, moist air in the planetary boundary layer with cool, dry air in the free atmosphere. This change in moisture serves as an indicator to the height of the planetary boundary layer. The variable n is the speed in which light travels through a medium. Because light travels at different speeds through different mediums, such as this can serve as an indicator of the planetary boundary layer height.

The method of MLH retrieval from soundings by X.Y. Wang and K.C. Wang (2014) will be used in this study. The ten greatest gradients for θ and ten least gradients for RH, q, and n will be found at heights constrained to 3 km above ground level. Starting from the ten highest gradient values for θ and ten lowest values for RH, q, and n, we observed where at least three of those variables had the same height and were within a tolerable error of +/- 50 meters. That height is said to be the best estimate of the mixing layer height.

**4. Discussion of rawinsonde profiles for MLH retrieval and uncertainties**

The combination of the variables, θ, RH, q, and n used to determine the MLH. Using the method for finding MLH as described by K.C. Wang and X.Y. Wang (2014) was effective at the time of data capture. From the gradient profiles, it was common to observe RH, q, and n have ranks coinciding with each other. In cases where the MLH could not be determined, the height of the maximum gradient for θ was found at higher altitude than RH, q, and n. Figure 2 shows the success of MLH retrieved from the profiles of θ, RH, q, and n using the proposed method.

Chart, line chart

Description automatically generated(a)Chart, line chart

Description automatically generated(b)Chart, line chart

Description automatically generated(c)Chart, line chart

Description automatically generated(d)

**Figure 2**

Vertical profile θ, RH, q, and n with increasing height and MLH for all four variables. (a) - θ, (b) - RH, (c) - q, and (d) - n. This profile was retrieved from Sterling, Virginia, USA during July 6, 2021, at 0 UTC. The MLH during this time was 2793 meters.

Uncertainty in MLH retrieval can be caused by many factors. Rawinsondes can pass through clouds which causes an increase in moisture. Atmospheric soundings might be capturing inversions associated with nocturnal stable layers and/or residual layers which occur during dawn and dusk.

**5. Machine Learning Mixing Layer Height Forecast**

Machine learning can be used to draw inference from patterns in data, such as MLH, to make better estimates into the future. Various models have been used to make estimates of the MLH such as decision tree regression, random forest regression, least-squares regression, and gradient boosting regression (Rieutord et al., 2021; Krishnamurthy et al., 2021; de Arruda Moreira et al., 2021). The model best suited for making estimates into the future of PBLH is with a decision tree regressor since PBLH has a daily cycle. Decision tree is a model this is used to learn how to split the data into different branches for non-linear relationship, like temperature versus altitude. Random forest is a model that uses a bunch of decision trees to gather their outputs to determine the best solution. As with any machine learning tool, more data points and input variables will result in a more accurate tool that can make predictions about an outcome. This might include having input features like year, month, day, hour to be able to predict into the future, along with meteorological parameters of potential temperature, relative and specific humidity, refractivity, as well as the already calculated PBLH for any existing data. Steps for how to use machine learning on this data are numbered below.

1. Import Math, NumPy, Pandas, and Scikit Learn libraries into a code editor. Read in the data. All columns except the height column will be used for training. We then tell the script which features within the dataset are training and testing, otherwise known as splitting. We then standardize the data so that all the columns have a mean value of zero and a standard deviation of 1.
2. We then apply a fit and transform on the input samples and return a new array. Since the input variables have different units (scales), we transform them to decrease the difficulty of the problem being modeled and increase the performance during the learning. We then train the decision tree model using the scaled input variables (datehour, pressure, temperature, etc.) and the scaled output variable (altitude). We repeat this process for the random forest model.
3. We then want to calculate the mean squared error (mse), regression loss for the decision tree training model. This error tells us how close a set of points are above or below a regression line. The error values are then square to remove negative values. For the same regression line, more weight is given to error values of larger distances. When then take the average of these error values to get the mse. The lower the mse, the better the forecast. We can then calculate the root mean square error (rmse)
4. We then want to calculate the mean absolute error (mae) for the decision tree training model. Error values are all made positive through an absolute. These errors are then summed and averaged. We want to keep this value within a bracket for accurate forecasting. This is useful for comparing year by year results
5. We then repeat steps 2 and 3 for the random forest training model.
6. We then want to calculate the mse, mse, and rmse for the decision tree and random forest testing models.

**6. Results**

The decision tree training model all had values of 0 for the mse, mae, and rmse. The random forest training mse, mae, and rmse had values of 34440.64, 141.45, and 185.58 respectively. The decision tree testing mse, mae, and rmse had values of 1764770.5, 1020.5, and 1328.45. The random forest testing mse, mae, and rmse had values of 396043.7, 628.9, and 629.3.

The decision tree training data has near perfect prediction. The random forest training data had predictions that were not accurate. The random forest testing data is displaying even worse predictions than the training data.

**Training**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mean squared error** | **Mean absolute error** | **Root mean squared error** |
| **Decision tree** | 0.0 | 0.0 | 0.0 |
| **Random forest** | 34440.6 | 141.4 | 185.6 |

**Testing**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mean squared error** | **Mean absolute error** | **Root mean squared error** |
| **Decision tree** | 1764770.5 | 1020.5 | 1328.45 |
| **Random forest** | 396043.7 | 628.9 | 629.3 |

**Figure 3**

Table of the mean square error, mean absolute error, and root mean squared error of the decision tree and random forest models for the training and testing datasets.

**7. Conclusion**

Both the decision tree and random forest testing sets display large errors measured by the mean squared error, mean absolute error, and root mean square error. Something to consider if this project were to be continued is using this tool during different time regimes. For example, the mixing layer height is commonly found at lower altitudes during the night and higher altitudes during the day. We could instead separate the datasets into these 0 UTC and 12 UTC. This might improve the accuracy of the model.

Again, a 1-2-1 smoother should be applied to the data initially collected by rawinsonde. I attempted to create a tool that would predict the mixing layer height days in advance, but as this task has the word in it, it was currently too advanced for me to do. This project has further piqued my interest in computer sciences as I see that it can provide such great wealth in our understand of the natural world.

Computer code and comments explaining it for this entire project can be found on GitHub at https://github.com/RahimKamara/SeniorResearchProject

**References**

Caicedo, V., Delgado, R., Sakai, R., Knepp, T., Williams, D., Cavender, K., Lefer, B., &

Szykman, J. (2020). An automated common algorithm for planetary boundary layer retrievals using aerosol lidars in support of the U.S. EPA photochemical assessment monitoring stations program. Journal of Atmospheric and Oceanic Technology, 37(10), 1847–1864. <https://doi.org/10.1175/JTECH-D-20-0050.1>

Carroll, B. J., Demoz, B. B., & Delgado, R. (2019). An overview of low-level jet winds and

corresponding mixed layer depths during pecan. Journal of Geophysical Research: Atmospheres, 124(16), 9141–9160. <https://doi.org/10.1029/2019JD030658>

de Arruda Moreira, G., Sánchez-Hernández, G., Guerrero-Rascado, J. L., Cazorla, A., &

Alados-Arboledas, L. (2021). Estimating the urban atmospheric boundary layer height from remote sensing applying machine learning techniques. Atmospheric Research, 105962.

Hennemuth, B., & Lammert, A. (2006). Determination of the atmospheric boundary layer

height from rawinsonde and lidar backscatter. Boundary-Layer Meteorology, 120(1), 181–200.

Hicks, M., Sakai, R., &amp; Joseph, E. (2015). The evaluation of a new method to detect mixing

layer heights using lidar observations. Journal of Atmospheric and Oceanic Technology, 32(11), 2041–2051. <https://doi.org/10.1175/JTECH-D-14-00103.1>

Huff, A. K., Kondragunta, S., Zhang, H., Laszlo, I., Zhou, M., Caicedo, V., Delgado, R., &

Levy, R. (2021). Tracking smoke from a prescribed fire and its impacts on local air quality using temporally resolved GOES-16 abi aerosol optical depth (aod). Journal of Atmospheric and Oceanic Technology, 38(5), 963–976. <https://doi.org/10.1175/JTECH-D-20-0162.1>

Judd, L. M., Al-Saadi, J. A., Szykman, J. J., Valin, L. C., Janz, S. J., Kowalewski, M. G., Eskes,

H. J., Veefkind, J. P., Cede, A., Mueller, M., Gebetsberger, M., Swap, R., Pierce, R. B., Nowlan, C. R., Abad, G. G., Nehrir, A., & Williams, D. (2020). Evaluating Sentinal-5P TROPOMI tropospheric NO2 column densities with airborne and pandora spectrometers near New York City and Long Island sound. Atmospheric Measurement Techniques, 13(11), 6113–6140. <https://doi.org/10.5194/amt-13-6113-2020>

Krishnamurthy, R., Newsom, R. K., Berg, L. K., Xiao, H., Ma, P. L., & Turner, D. D. (2021). On

the estimation of boundary layer heights: a machine learning approach. Atmospheric Measurement Techniques, 14(6), 4403-4424.

Li, Z., Rosenfeld, D., & Fan, J. (2017). Oxford research encyclopedia of environmental

science. In Aerosols and their impact on radiation, clouds, precipitation, and severe

weather events. essay, Oxford University Press.

Liu, S., & Liang, X.-Z. (2010). Observed diurnal cycle climatology of planetary

boundary layer height. Journal of Climate, 23(21), 5790–5809.

Piironen, A. K., and Eloranta, E. W. (1995), Convective boundary layer mean depths and cloud

geometrical properties obtained from volume imaging lidar data, J. Geophys. Res., 100(D12), 25569– 25576, doi:10.1029/94JD02604.

Rieutord, T., Aubert, S., & Machado, T. (2021). Deriving boundary layer height from aerosol

lidar using machine learning: KABL and ADABL algorithms. Atmospheric Measurement Techniques, 14(6), 4335-4353.

Seidel, D. J., Ao, C. O., & Li, K. (2010). Estimating climatological planetary boundary layer

heights from rawinsonde observations: Comparison of methods and uncertainty analysis. Journal of Geophysical Research: Atmospheres, 115(D16).

Wallace, J. M., & Hobbs, P. V. (2006). Atmospheric science: an introductory survey (Vol. 92).

Elsevier.

Wang, X. Y., & Wang, K. C. (2014). Estimation of atmospheric mixing layer height from

rawinsonde data. Atmospheric Measurement Techniques Discussions, 7(2), 1247–1273. <https://doi.org/10.5194/amtd-7-1247-2014>

Wiegner, M., Madonna, F., Binietoglou, I., Forkel, R., Gasteiger, J., Geiß, A., Pappalardo, G.,

Schäfer K, & Thomas, W. (2014). What is the benefit of ceilometers for aerosol remote sensing? an answer from earlinet. Atmospheric Measurement Techniques Discussions, 7(3), 2491–2543. <https://doi.org/10.5194/amtd-7-2491-2014>

**Acknowledgements**

I would like to thank GitHub user srnghn and Jason Brownlee of MachineLearningMastery.com for their template on machine learning using decision tree and random forest models. Their comments on the script made it helpful in understanding a topic that is novel to me. I would like to thank Dr. Ruben Delgado for being an amazing research mentor and suggesting this topic. This was probably the most mentally challenging task that I have taken on so far, and I have come around to loving it! There has no growth without strain. I would like to thank Dr. Timothy Canty for being understanding of my situation and teaching his students very well! I would like to thank Dr. Alexandra Jones for taking a risk on me and sharing her abundance of resources, allowing me to be where I am at currently. I would like to thank Jeff Henrikson for teaching me how to code and the importance of patience. I would like to thank AOSC students for being an amazing support group.

I deeply thank NCAS-M for providing monthly workshops with professionals aligned with NOAA’s mission, funding for conferences, and education support. I would like to thank Dr. Xin-Zhong Liang for being my NCAS-M mentor and teaching the importance of having curiosity. I wish I would have shed the fear of being surrounded by established scientists so that I could ask more questions.

I would like to thank my family for raising me and teaching me well. Finally, I would like to thank Kaya for inspiring me to take a chance and risk with life to see tomorrow.