Multi-fidelity Assessment of Atmospheric Boundary Layer Observations Through Clustering Machine Learning

Eric Enright,* Austin Aguilar, † and Daniel Foti ‡
Department of Mechanical Engineering, University of Memphis, Memphis, TN, 38152, USA

Atmospheric conditions rapidly change due to many factors. Long-time observations from the US Department of Energy Atmospheric Radiation Measurement and Climate (ARM) Southern Great Plains (SGP) user facility site provide various experimental tools that measure conditions at various space and time scales. Vast quantities of data are continuously accumulated and are readily available for investigation, but locating particular conditions in the data is nontrivial due to the data size. Data assimilation, observational analysis, and computational model validation are used in a wide range of tasks such as weather prediction, climate models, and renewable energy utilization models but require an understanding of atmospheric conditions and their statistics. Due to the disparate space and time scales in observations, a multi-fidelity clustering approach is employed to identify and organize atmospheric conditions including repetitive events and outliers. A k-means clustering method is used to classify sparse, 30-minute, time-averaged observational data. High-resolution data is then assessed within a cluster. Observations at SGP over several months are selected and clustered based on eddy correlation flux measurement system variables of sensible heat flux, momentum heat flux, latent heat flux, and wind speed. Clusters lead to broad classes of atmospheric conditions and stability. Once clustered, high-resolution data from observations of sonic anemometers on meteorological towers and radar are assessed within the classification. This enables a view of data over narrow time ranges mitigating costs associated with large data set volumes. The developed methods allow for the selection of time ranges for an event and ease prediction and energy utilization modeling tasks.

I. Nomenclature

$\bar{u'}$	=	time-averaged velocity component in the u direction
$\bar{v'}$	=	time-averaged velocity component in the v direction
$\bar{w'}$	=	time-averaged velocity component in the w direction
K	=	Turbulent Kinetic Energy
x_i	=	Data point in a given dataset being analyzed by the K-means method.
z_{ik}	=	Binary variable ($zik \in \{0, 1\}$) denoting if a point is in the kth cluster.
a_k	=	Cluster center for the kth cluster.

II. Introduction

The atmospheric boundary layer (ABL) is the lowest part of the atmosphere located on the Earth's surface, and its height ranges on the order of one kilometer. Its height and turbulence intensity are due to various factors that cause variable mixing and turbulence within the ABL. Understanding these conditions is beneficial for applications such as weather prediction, developing climate models, and sustainable energy production. Atmospheric dynamics often play a critical role in the sustainability and reliability of diverse forms of energy production. This is especially true for the growing number of renewable energy deployments that harness aspects of the environment for power production. One of the major challenges for understanding and developing energy systems and management platforms is accurate modeling/forecasting of atmospheric conditions across disparate spatial and temporal scales. These conditions are often

*Undergraduate Student, Department of Mechanical Engineering, Senior
†Undergraduate Student, Department of Mechanical Engineering, Senior
‡Assistant Professor, Department of Mechanical Engineering

required to understand the lowest levels of the ABL. Long-term atmospheric observations and measurements at the U.S. Department of Energy Atmospheric Radiation Measurement (ARM) sites are compiled.

Machine learning is a powerful tool for classification and data analysis and has been applied in many studies [1][2][3]. We will employ clustering as our primary tool to create categories. The k-means algorithm is an unsupervised method of machine learning and is one of the most well-known and regularly utilized clustering algorithms. It is robust and can be easily applied to many domains [4]. The application of k-means on atmospheric data is relatively new; Zhenxing and Chang et. al. and Toledo et. al. [5] [6] have recently conducted studies using k-means to predict ABL height. These studies attest to the reliability of clustering methods. When compared to other numerical methods, Toledo found that the classic k-means clustering method yielded the same results as six other methods that are traditionally used to determine ABL height. Through a multi-fidelity assessment of data collected from other experimental tools, k means clustering will be used to correlate outlier behaviors of the data to create a predictive model of ABL conditions.

The eddy correlation station at SGP measures gas exchanges between the atmosphere and the Earth's surface and various additional factors. Being located at the lowest part of the troposphere, the ABL is highly susceptible to interactions with many factors, which can greatly affect ABL characteristics such as height and stability. The compiled flux data from the eddy correlation station helps determine gas and thermal exchanges with the atmosphere at a local level. This is very important for agricultural and sustainability reasons. A clustering approach used on this data can create a correlation between findings found using light detection and ranging (LiDar) and surface meteorology systems (MET) data. Met towers record wind characteristics at a certain site. Lidars determine distances to aerosol particles by emitting light waves. The lidar at ARM determines vertical velocity and cloud statistics. There is a variety of LiDar tools available at the SGP ARM site. Krishnamurthy et. al. [7] utilized a variety of lidars, including the doppler lidar, to train a machine-learning model to predict ABL height. The Doppler lidar makes a clear determination of ABL height by measuring turbulence directly at high temporal resolutions. The availability of Lidar data in a variety of fidelities allows making connections with the results from the data obtained by the tools available at the SGP data site can lead to a clear categorization of these variables.

This research employs data collected at the Southern Great Plains (SGP) site of the ARM facilities for further study through k-means clusters. The study focuses on variables of fluxes, friction velocity, Monin Obukhov length, and Turbulence Kinetic Energy (TKE) to create classifications of atmospheric stability and turbulence levels. By categorizing the key variables being assessed by the broad range of experimental tools available at SGP, a more concise categorizing of the data can be done. The developed methods will expedite the process of scientific exploration of the data and create a tool for ABL researchers to easily navigate the surplus of data ARM has readily available.

III. Methods

A. Data Acquisition

The data of interest is obtained from the Eddy Correlation station equipped with a Smart Flux processor(ESCORSF14) [8]. This tool has acquired a broad range of data. In this paper, flux variables of Latent Heat Flux, Momentum Flux, and Sensible Heat Flux are examined. Latent Heat Flux represents the heat exchange between the surface and the atmosphere due to water vapor phase change [9]. Sensible heat flux is the representation of the energy exchange due to turbulence of heat between the surface and the ABL and represents the surface-atmosphere temperature difference. Momentum flux is the measure of wall-normal shear stress due to turbulence. These parameters are directly related to thermal and mechanical turbulence and are vital in determining the stability of the ABL [10]. Aside from the fluxes, this dataset contains friction velocity, Monin-Obukhov length (Mo length), and turbulence kinetic energy. Turbulent kinetic energy is the representation of the intensity using time-averaged velocity components and can be equated as follows (1)[11]. Friction velocity is the scaling variable of the ABL parameter that characterizes turbulence strength and laminar sub-layer thickness. Mo length examines behaviors of buoyancy and turbulence at the lower portions of the ABL.

$$K = \frac{1}{2}(\bar{u'} + \bar{v'} + \bar{w'})^2 \tag{1}$$

Met tower data from the SGP site of Lamont, Oklahoma was obtained for the month of July 2020 [12]. The measurements from this instrument provide high-fidelity data averaged over one-minute increments. Findings from these measurements are referenced back to the findings from the eddy correlation station data which is averaged over thirty-minute periods. This categorizing of data between different spatiotemporal scales provides different perspectives on varying atmospheric conditions.

The analyzed LiDar data was obtained for the month of July 2020 at Lamont, Oklahoma[13]. This data was mainly used to determine the conditions during the time period of the data being analyzed from the Met tower and Eddy Correlation station. As mentioned by Zhenxing and Chang et. al [5], the data observed from k-means can be skewed by the presence of a residual layer. The data from the lidar is simply cross-referenced with the clusters from the Met Tower and the Eddy Correlation station to determine the cause of any outlier data or anomalies in the clusters.

B. Clustering Algorithm

The data clusters evaluated in this paper are based on the characteristics of friction velocity, Monin-Obukhov length, turbulent kinetic energy, latent heat flux, momentum flux, and sensible heat flux to classify atmospheric stability and turbulence levels. In the higher fidelity, compare observation from the met tower and radar to assess the classification of methods.

The clustering algorithm that is being employed on the data is k-means. The k-means algorithm is one of the most popular unsupervised machine learning algorithms. Being unsupervised, it will make inferences based strictly on the inputs provided and no other adjustment or supervision is required. The algorithm employed in this study makes use of an agglomerative approach by clustering data iteratively based on their similarity [14]. The number k refers to the number of centroids that will be in the dataset, and the centroid in this case is the center of the cluster. The evaluation of the algorithm by Sinaga et. al [4] describes the process of k means with the following equations. $J(z, A)$ (2) is the objective function of the algorithm taking in the binary argument of z and cluster center a.

$$J(z, A) = \sum_{i=1}^{n} \sum_{i=1}^{c} z_{ik} \|x_i - a_k\|^2 \tag{2}$$

The equation for a_k (3) is a minimizing characteristic of the objective function, used to calculate the cluster center a for the kth point in the iteration of the algorithm.

$$a_k = \frac{\sum_{i=1}^{n} z_{ik} x_{ij}}{\sum_{i=1}^{n} z_{ik}} \tag{3}$$

Similarly, z_{ik}(4) is a value used to minimize the objective function. When iterating through a data set it changes its binary value based on the Euclidean distance between the kth point and a cluster center.

$$z_{ik} \begin{cases} 1, & \text{if } \|x_i - a_k\|^2 = \min\limits_{1 \le k \le c} \|x_i - a_k\|^2 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

To solidify the validity of the number of clusters, different methods will be additionally employed. For this paper, these methods are known as internal clustering validity indices[15]. Specifically focusing on two of these called silhouette width, and elbow method. The silhouette width utilizes the partition from the k-means and a Euclidean distance, $\|x_i - a_k\|$, to determine the accuracy of the cluster [16]. The elbow works differently by first calculating the sum of squares of the cluster, then plotting against the number of clusters [17]. The product of this plot is a line gradually decreasing until reaching a pivot, or elbow, then flattening out with respect to an increasing number of clusters. This elbow is the key to determining the theoretical amount of clusters suggested.

IV. Results

We start by focusing on 30-minute averaged flux data from the eddy correlation station over the month of June 2020. The averaging of the data provides a low-cost approach that employs relatively small quantities of data that need to be assessed. The changes in fluxes in this data are important factors for determining specific boundary layer conditions and provide an assessment of the overall instantaneous events that are occurring in the ABL. Figures 1, 2, and 3 show the variables of latent heat flux, momentum flux, and sensible heat flux, respectively, changes over the course of a day. The changes in the fluxes fluctuate over each day over the month. The diurnal cycle is clearly present in sensible and latent heat fluxes, where large increases are present during afternoon and into the night when the fluxes start to decrease. The sensible heat flux indicates that a convective boundary layer is likely in the afternoon and a stable boundary layer is likely at night. The changes in the momentum flux are much less evident than the energy-based fluxes. While there may be large fluxes at a given time, the diurnal cycle is not shown clearly. This is evidence that the average wind speed

does not change drastically based on the time of day or the stability of the ABL. With these plots, consistencies were discovered that enabled a reference for plotting specific months and weeks using k-means clustering.

The time frame places it during the summer months specifically towards the hottest parts of the year. The beginning of July presents low sensible heat flux but as the month passes on the sensible heat flux increases drastically.

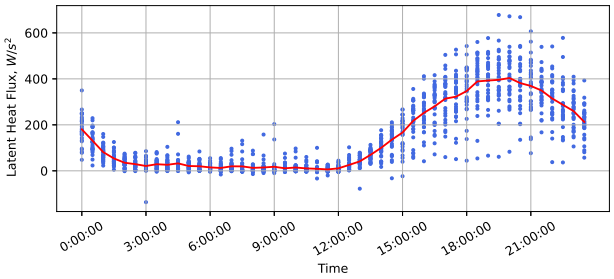

Fig. 1 Thirty-minute averaged latent heat flux based on time of day for the 31 days of July 2020 obtained from sgpecorrsfE31 [8]. The markers represent the flux at a specific time of day. The red line is the average for each time.

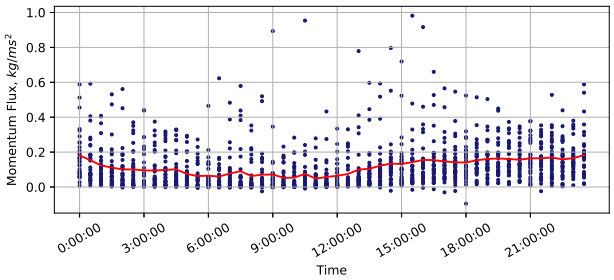

Fig. 2 Thirty-minute averaged momentum flux based on time of day for the 31 days of July 2020 obtained from sgpecorrsfE31 [8]. The markers represent the flux at a specific time of day. The red line is the average for each time.

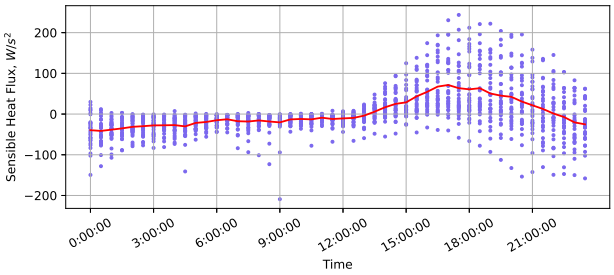

Fig. 3 Thirty-minute averaged sensible heat flux based on time of day for the 31 days of July 2020 obtained from sgpecorrsfE31 [8]. The markers represent the flux at a specific time of day. The red line is the average for each time.

Before clustering is employed, the elbow method is used to analytically assess the potential number of clusters. The elbow method is a heuristic based on the number of clusters. While subjective in the sense that the number of clusters chosen is based on the curve or "elbow" of the curve, it provides a tool to determine the number of clusters. First, clustering will be based on using combinations of two of the three fluxes. Figure 4 shows the variations or distortion

with the increased number of clusters. Figure 4(a) shows the elbow method based on the momentum flux and the sensible heat flux, Fig 4(b) shows the variations for the sensible heat flux and the latent heat flux, and Fig. 4(c) shows the variations for the latent heat flux and the momentum flux. Overall, the elbow method suggests that the potential number of clusters in the data is between 3 and 4 clusters regardless of the two fluxes used. The silhouette score method is shown in Fig. 5 where the silhouette score is evaluated over 2-5 clusters. The silhouette score results for the flux values extend the findings of the elbow method, showing the data has the highest silhouette score between 2-3 clusters. In what follows, we will investigate the clustering based on 3 clusters.

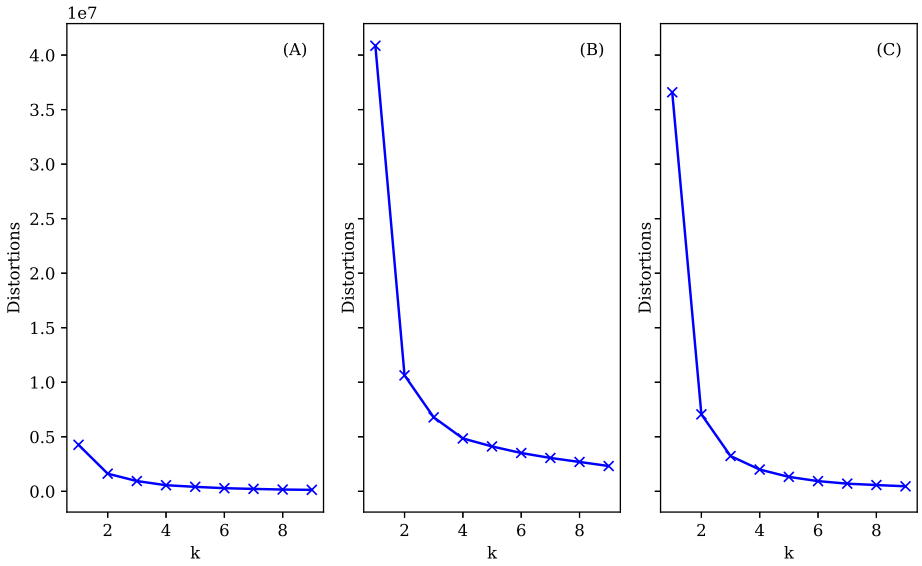

Fig. 4 Elbow method on (a) momentum flux and sensible heat flux, (b) sensible heat flux and latent heat flux, and (c)latent heat flux and momentum flux.

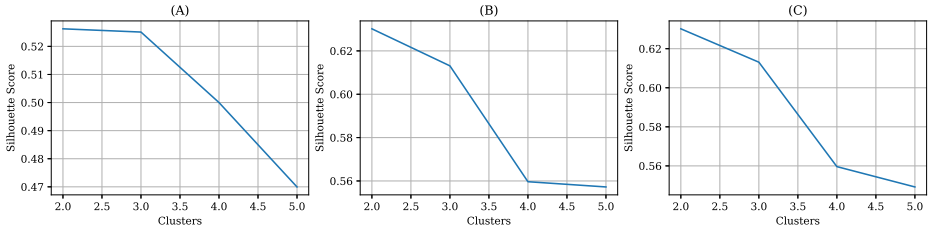

Fig. 5 Silhouette score method on determining optimal clusters for (a) momentum flux and sensible heat flux, (b) sensible heat flux and latent heat flux, and (c) latent heat flux and momentum flux.

Clusters and potential categories are revealed for each combination of two fluxes in Fig. 6. Each observation is normalized by the maximum and minimum values such that the normalized flux is between 0 and 1. Normalization is important for the observations because the k-mean algorithm is based on a distance metric. Due to normalization, the true value of the sensible heat flux and where it changes from convective to stable is at one-half. The normalized sensible heat flux is convective between 0.5 and 1 and stable between 0 and 0.5. In Fig. 6(a) the clustering reveals that two of the three clusters are in the stable regime, while one is mainly in the convective regime. There are two regimes of clustering that is in a stable ABL, those conditions with low shear stress and friction velocity and those with high shear stress and friction velocity. On the other hand, the convective regime has a low to medium shear stress. While conditions are not entirely independent of one another, we can see that categories of ABL conditions can be organized. Similarly, the clustering of the other two combinations in Fig. 6(B) and Fig. 6(C) show similar behavior. High latent heat flux is many associated with low shear stress, while the categories in the sensible and latent heat flux show less conclusive results. In all cases, more clusters could play a role in categorizing more sub-regimes in the ABL conditions

but will be left to future work.

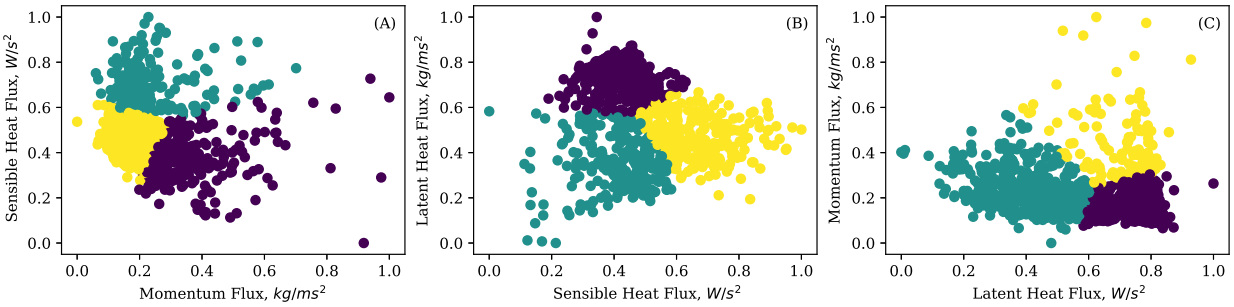

Fig. 6 Clusters of the 30-minute averaged data of the parameters of (a) momentum flux and sensible heat flux, (b) sensible heat flux and latent heat flux, and (c) latent heat flux and momentum flux.

Figure 7 shows elbow analysis for the clustering of all three fluxes together. The change in the distortion with increased clusters occurs between 3-5 clusters. Additionally, a silhouette analysis was employed in the compilation of fluxes shown in figure 8 where the silhouette score reveals the ideal number of clusters for the algorithm is between 2 and 3. As shown in Fig. 9, k-means is undertaken on the three fluxes using 3 - 5 clusters. The largest factor in clustering is the latent heat flux which at high values contains two different clusters, while low latent heat flux is clustered into one regime that encompasses all of the ABL stability conditions. To decide on the desired condition the method would be employed finally by accessing the data presented by the color of the data. The exploitation of this feature of the scikit-learn python module proves beneficial in achieving proper data for higher fidelity data.

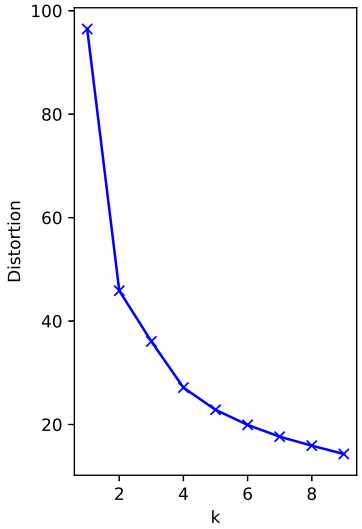

Fig. 7 Elbow analysis of sensible heat flux, latent heat flux, and momentum flux parameters obtained from sgpecorrsfE31 [8] data set for the month of July 2020.

The main goal of clustering the 30-minute average is to employ low temporal resolution data. The final part of the method involves taking the reviewed k-means clustering data and applying it to a higher temporal fidelity data set. In this paper, measurements from the met tower were utilized. Specifically, the variables that were reviewed are temperature and wind speed, from which the sensible and momentum flux data are derived. However, due to measurements at much higher frequencies, there is substantially more data per day than the eddy correlation station measurements. Plotting these two observations against the time of day displays ABL conditions essential for further analysis. The temperature in the month of July 2020, proves many similarities to sensible heat flux during the same times as shown in see Fig. 10. Temperatures start high during this month, decrease, and then begin to increase more rapidly as the

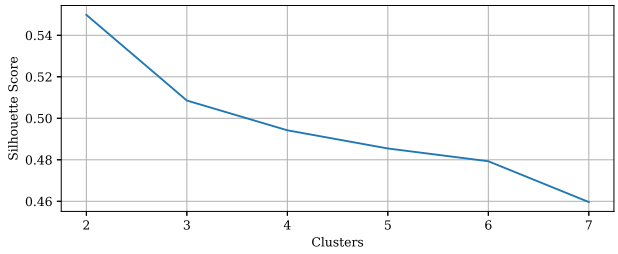

Fig. 8 Silhouette analysis of sensible heat flux, latent heat flux, and momentum flux parameters obtained from sgpecorrsfE31 [8] data set for the month of July 2020.

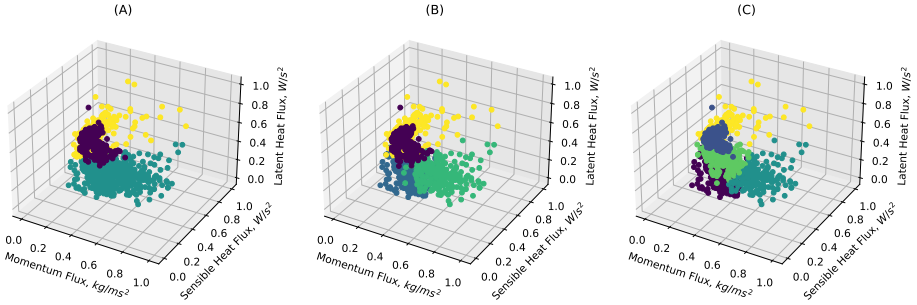

Fig. 9 K-means clustering of momentum flux, sensible heat flux, and latent heat flux data for the month of July 2020 utilizing (a) 3 clusters, (b) 4 clusters, (c) and 5 clusters.

month progresses. Wind speed, shown in Fig. 11, fluctuates rapidly throughout the day and month. Similarities can be analyzed in momentum flux during similar times. The importance behind these two figures is the result of clustering low-fidelity data. Although Fig. 10 and Fig. 11 are in the same month a more efficient step would be to utilize the previously mentioned categories. Obtaining higher fidelity data with the utilization of categories will grant access to more preferred data than measurements directly from the met tower.

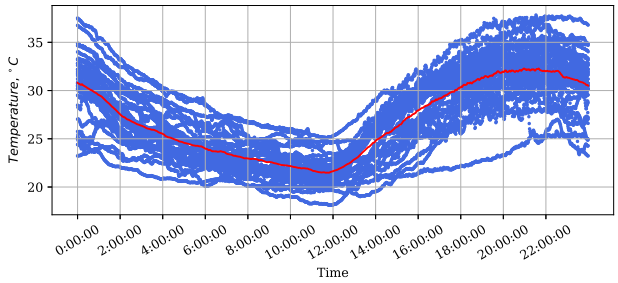

Fig. 10 July 2020 temperature data averaged over 1 minute obtained from sgpmetE31 [12]. The blue markers represent the temperatures at a specific time of day, and the red line is the average temperature over all 31 days.

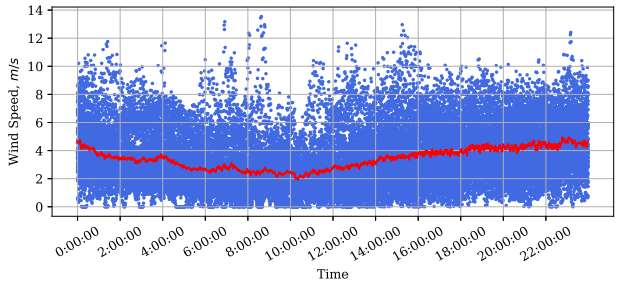

Fig. 11 July 2020 wind speed data averaged over 1 minute obtained from sgpmetE31 [12]. The blue markers represent the wind speed at a specific time of day, and the red line is the average temperature over all 31 days.

V. Conclusion

Conditions in the ABL are challenging to analyze from long-term observations due to the sheer size of the data collected, especially when instantaneous data or outlier events are desired. Low-fidelity 30-minute average samples are employed to asses and cluster conditions to find sub-categories of the data. A data set containing one month of experimental data was selected from the eddy correlation station at the SGP site in Lamont, Oklahoma. The key fluxes, momentum, sensible heat, and latent heat, are used to categorize the conditions. Using a k-means algorithm, the fluxes are clustered. Elbow analysis reveals that there are only a few clusters in the data. Despite the 30-minute averaged data, the data is very dense, possibly due to the ergodicity of turbulence. While elbow analysis shows few key clusters, the k-means clusters select physically relevant categories such as stable ABL with low friction velocity, convective ABL with low friction velocity, and convective ABL with high friction velocity. These categories provide a first attempt to sort the data and locate times of interest in higher fidelity, high temporal resolution data. We identify the met tower data as particularly interesting as instantaneous wind speed and temperatures can provide information about the conditions at specific times. This data can also be used to assess turbulent scales and energy. In the future, we will assess additional clustering methods such as DB-SCAN or non-parametric Dirichlet processes. These alternatives may identify different and/or more clusters of physical interest.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under Award Number DE-SC0023267.

References

[1] MacQueen, J., "Classification and analysis of multivariate observations," *5th Berkeley Symp. Math. Statist. Probability*, University of California Los Angeles LA USA, 1967, pp. 281–297.

[2] Steinhaus, H., "Sur la division des corps matériels en parties: Bulletin de l'Académie polonaise des sciences," 1957.

[3] Lloyd, S., "Least squares quantization in PCM," *IEEE transactions on information theory*, Vol. 28, No. 2, 1982, pp. 129–137.

[4] Sinaga, K. P., and Yang, M.-S., "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, Vol. 8, 2020, pp. 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796.

[5] Liu, Z., Chang, J., Li, H., Chen, S., and Dai, T., "Estimating Boundary Layer Height from LiDAR Data under Complex Atmospheric Conditions Using Machine Learning," *Remote Sensing*, Vol. 14, No. 2, 2022. https://doi.org/10.3390/rs14020418, URL https://www.mdpi.com/2072-4292/14/2/418.

[6] 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*, Vol. 31, No. 2, 2014, pp. 422–436.

[7] Krishnamurthy, R., Newsom, R. K., Berg, L. K., Xiao, H., Ma, P.-L., and Turner, D. D., "On the estimation of boundary layer heights: a machine learning approach," *Atmospheric Measurement Techniques*, Vol. 14, No. 6, 2021, pp. 4403–4424.

[8] Shi, Y., "Eddy Correlation Flux Measurement System (ECORSF)," , ???? https://doi.org/10.5439/1494128.

[9] Ledley, T., "ENERGY BALANCE MODEL, SURFACE," *Encyclopedia of Atmospheric Sciences*, edited by J. R. Holton, Academic Press, Oxford, 2003, pp. 747–754. https://doi.org/https://doi.org/10.1016/B0-12-227090-8/00150-0, URL https://www.sciencedirect.com/science/article/pii/B0122270908001500.

[10] Daoo, V., Panchal, N., Sunny, F., and Venkat Raj, V., "Scintillometric measurements of daytime atmospheric turbulent heat and momentum fluxes and their application to atmospheric stability evaluation," *Experimental Thermal and Fluid Science*, Vol. 28, No. 4, 2004, pp. 337–345. https://doi.org/https://doi.org/10.1016/j.expthermflusci.2003.06.006, URL https://www.sciencedirect.com/science/article/pii/S0894177703001146.

[11] Gorlé, C., van Beeck, J., Rambaud, P., and Van Tendeloo, G., "CFD modelling of small particle dispersion: The influence of the turbulence kinetic energy in the atmospheric boundary layer," *Atmospheric Environment*, Vol. 43, No. 3, 2009, pp. 673–681. https://doi.org/https://doi.org/10.1016/j.atmosenv.2008.09.060, URL https://www.sciencedirect.com/science/article/pii/S1352231008009084.

[12] Kyrouac, J., and Shi, Y., "Surface Meteorological Instrumentation (MET)," , ???? https://doi.org/10.5439/1786358.

[13] Shippert, T., Newsom, R., and Riihimaki, L., "Doppler Lidar Wind Statistics Profiles (DLPROFWSTATS4NEWS)," , ???? https://doi.org/10.5439/1178583.

[14] Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., and Heming, J., "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, Vol. 622, 2023, pp. 178–210. https://doi.org/https://doi.org/10.1016/j.ins.2022.11.139, URL https://www.sciencedirect.com/science/article/pii/S0020025522014633.

[15] Rendón, E., Abundez, I., Arizmendi, A., and Quiroz, E. M., "Internal versus external cluster validation indexes," *International Journal of computers and communications*, Vol. 5, No. 1, 2011, pp. 27–34.

[16] Rousseeuw, P. J., "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, Vol. 20, 1987, pp. 53–65.

[17] Bholowalia, P., and Kumar, A., "EBK-means: A clustering technique based on elbow method and k-means in WSN," *International Journal of Computer Applications*, Vol. 105, No. 9, 2014.