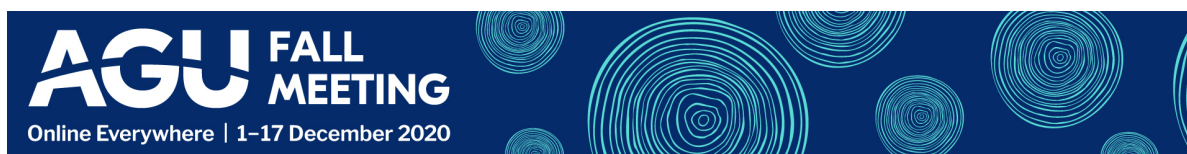


A Deep Learning Approach for Surface PM_{2.5} Estimations from Geostationary Satellite and Numerical Model Data

Manisha Khatri, Muthukumaran Ramasubramanian, Iksha Gurung, Aaron S Kaulfus, George Priftis, Peiyang Cheng, Pawan Gupta, Manil Maskey, Rahul Ramachandran, Sundar Christopher, Haeyong Chung

University of Alabama in Huntsville, Huntsville, AL; NASA Marshall Space Flight Center, Huntsville AL

PRESENTED AT:



INTRODUCTION

Fine particles released from activities such as vehicle exhausts, burning of fuels, forest fires are known to cause severe impacts on public health. Particulate matter (PM) with a diameter less or equal to $2.5\text{ }\mu\text{m}$, known as $\text{PM}_{2.5}$, is known to cause or exacerbate cardiovascular and respiratory illnesses. In the United States, the Environmental Protection Agency (EPA) monitors the levels of particulate matter and regulates surface level concentrations of $\text{PM}_{2.5}$ to protect human health and property. However, monitoring is sparse and observations are representative only of the surrounding area. Furthermore, regulations dictate that at many locations observations must only be collected every third day.

The spatial and temporal limitations in surface monitoring can be overcome by using Aerosol Optical Depth (AOD), a representation of the amount of aerosols in the atmosphere, products from satellite remote sensing. Geostationary satellites in particular have the ability to provide AOD observations multiple times throughout the day and at high spatial resolutions. While AOD may be related to the concentrations of aerosols at the surface, it is not a direct measurement. In addition, meteorological conditions including relative humidity, temperature, the height of the planetary boundary layer nonlinearly affect surface $\text{PM}_{2.5}$ concentrations. While readily observed, advancements in high resolution modeling provide accurate analysis of meteorological conditions where observations are absent. Traditional approaches to estimating surface $\text{PM}_{2.5}$ relies on linear regression of these meteorological factors and AOD. The volume of high resolution datasets and complex relationships between factors make deep learning approaches appropriate for such complex estimation problems. We demonstrate a deep learning neural network that utilizes the AOD retrievals from geostationary satellites and high resolution meteorological analysis fields to estimate surface $\text{PM}_{2.5}$.

MODEL FORMULATION

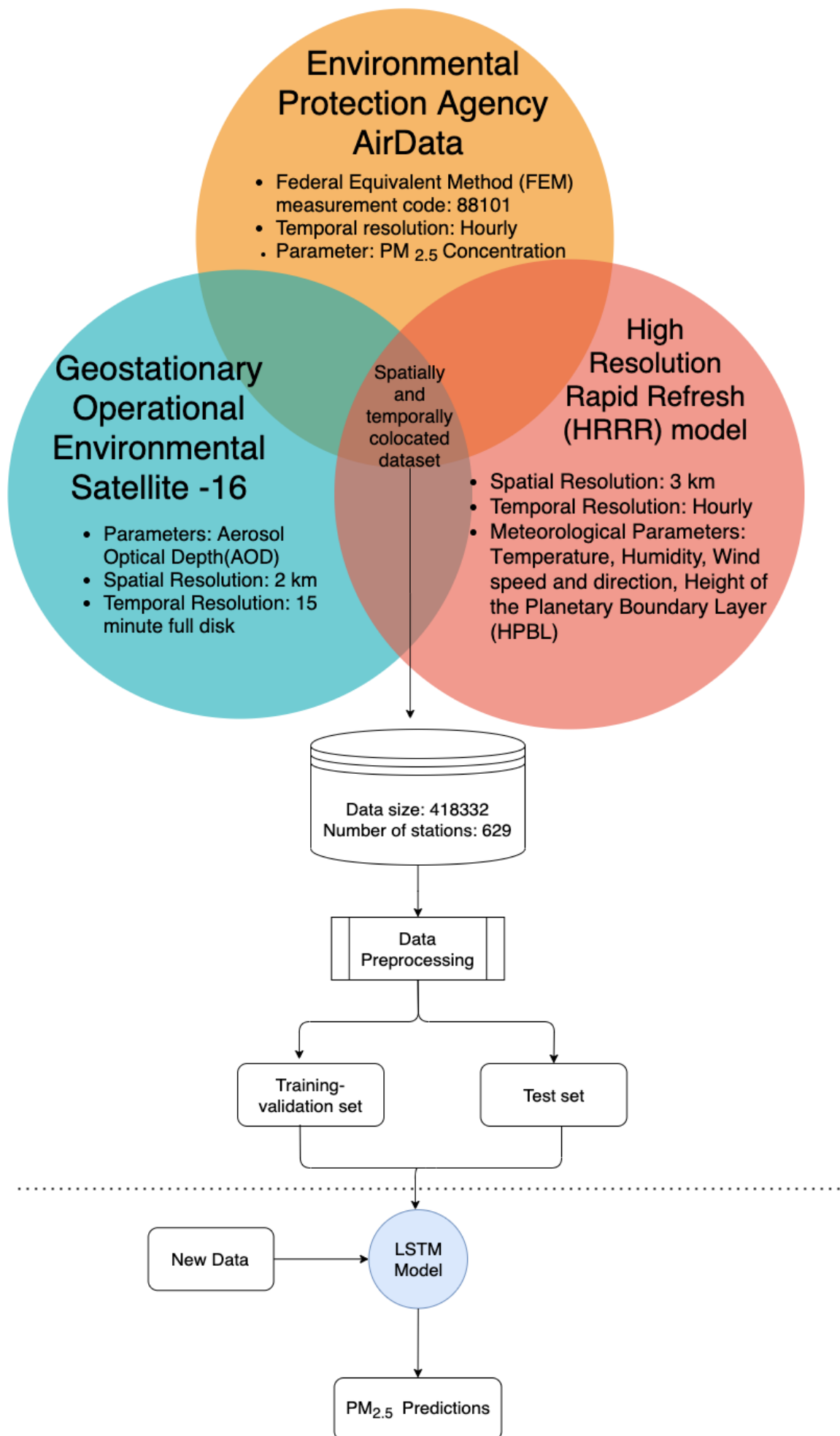
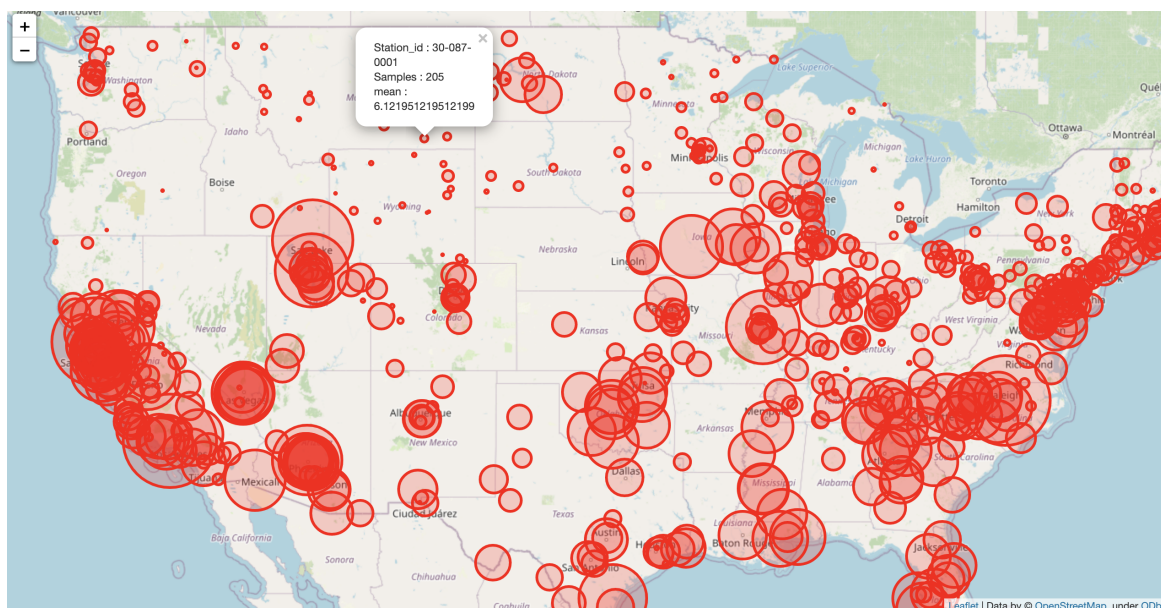


Figure 1. Data sources and processing pipeline.

AOD from the Geostationary Environmental Operational Satellite -16, Relative Humidity (RH), Temperature (TMP), Height of Planetary Boundary Layer (HPBL), Wind speed vectors (UGRD, VGRD) from the High Resolution Rapid Refresh (HRRR) atmospheric model, and surface PM_{2.5} concentrations from EPA monitors collected during 2018 and 2019 are utilized in this effort (Figure 1). Data are split into a training, validation, and testing dataset 60-20-10 for developing and testing the Long short-term memory (LSTM) model, an artificial recurrent neural network architecture that can maintain information in memory for long periods of time

The results from the LSTM are compared with Linear Regression (LR), Support Vector Regression (SVR), and K Nearest Neighbors - Regression (KNN). In contrast to LR, SVR and KNN are non-parametric models and can model the nonlinearity in the data and therefore have been selected as the baseline models for the experiment. The LSTM model has memory cells that can process long sequences of data and model long term information.

Figure 2. PM_{2.5} monitoring locations scaled by the number of observations in our sample dataset.

The majority of observations identified for this study are found to be in urban areas (Figure 2). Also, the density of samples is particularly high around California and the east coast.

DATA VISUALIZATIONS

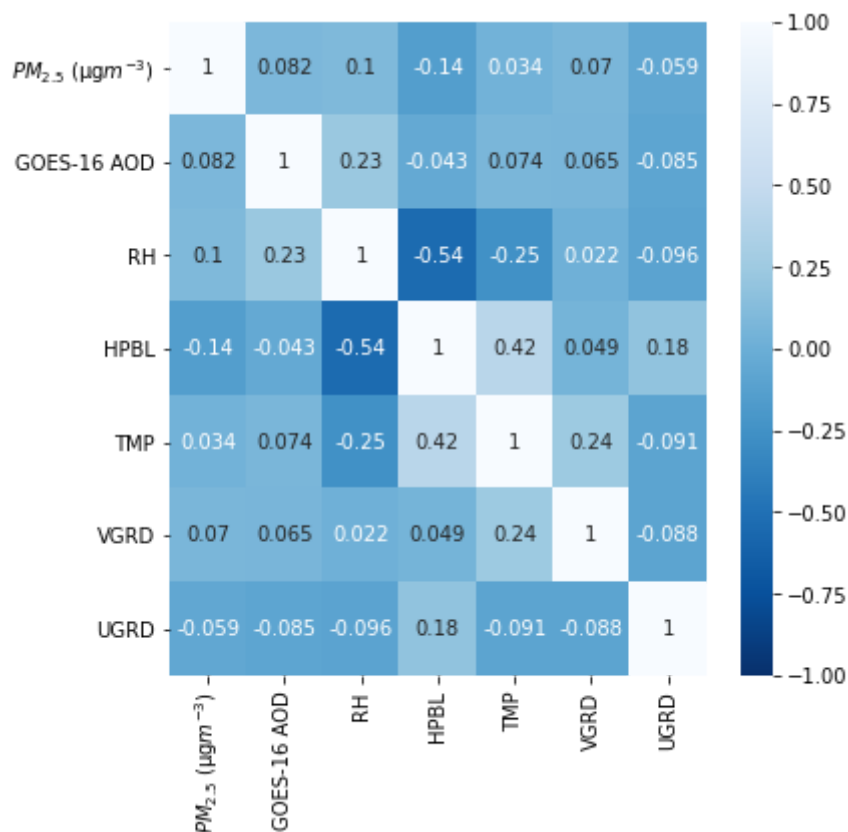
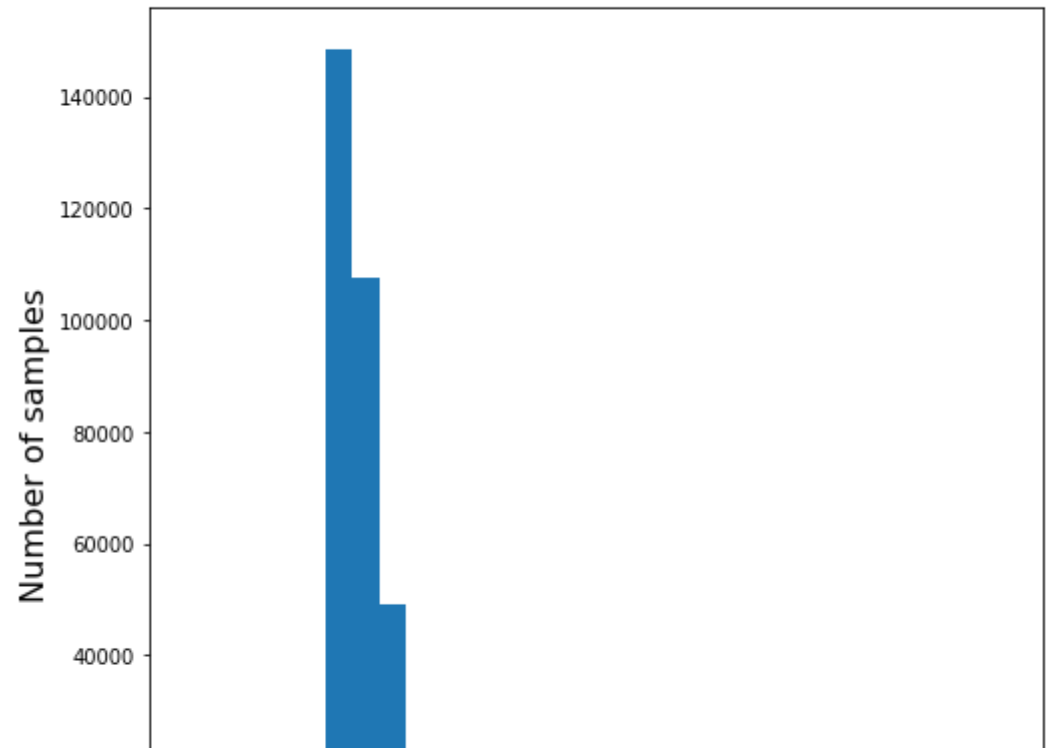
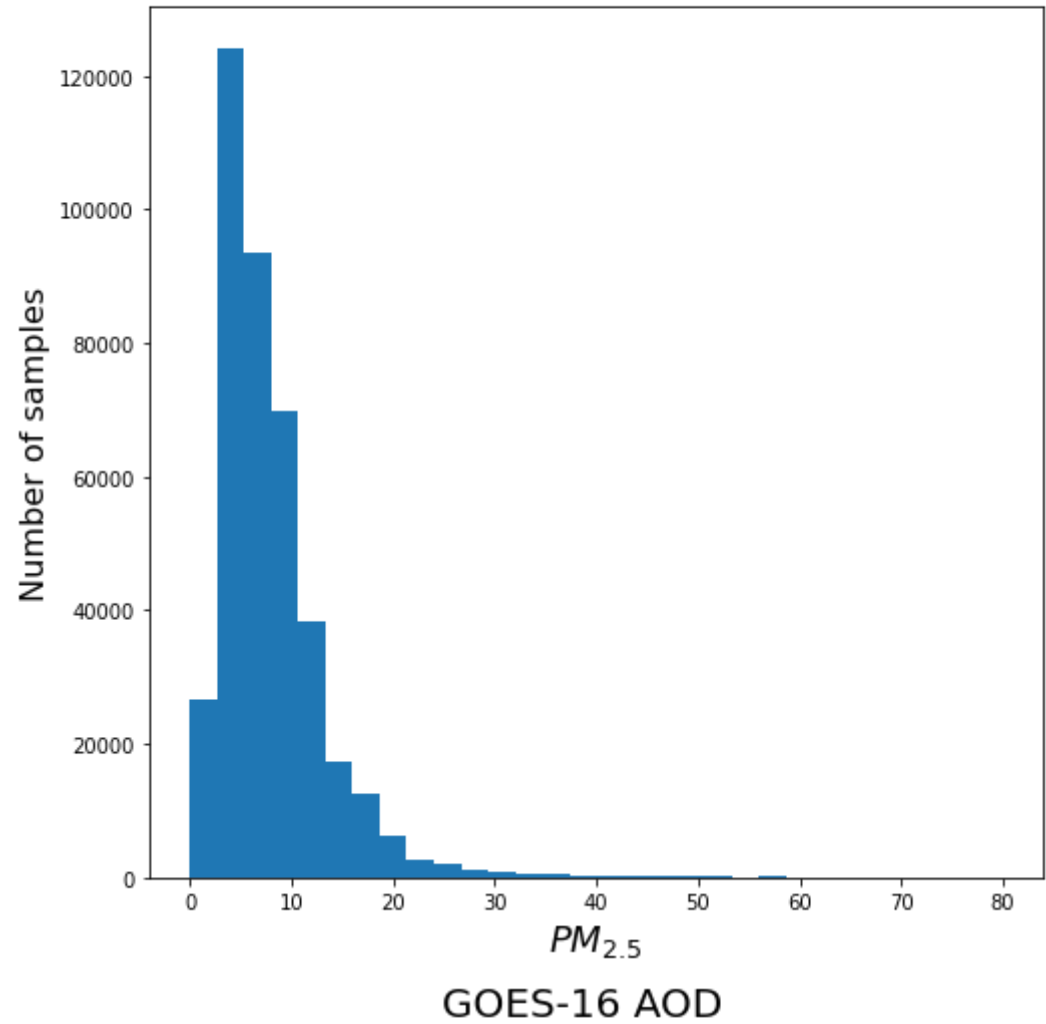
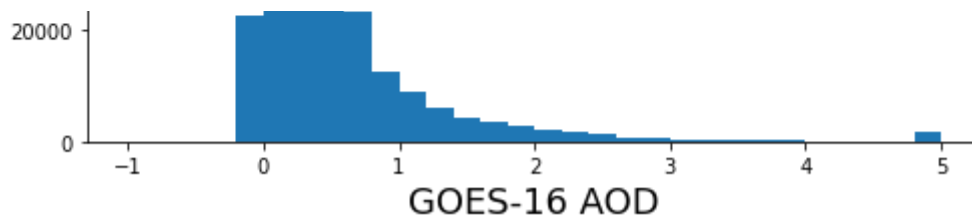


Figure 3. Correlation matrix representing the relations between GOES R AOD and HRRR variables and $PM_{2.5}$ ($\mu g m^{-3}$).

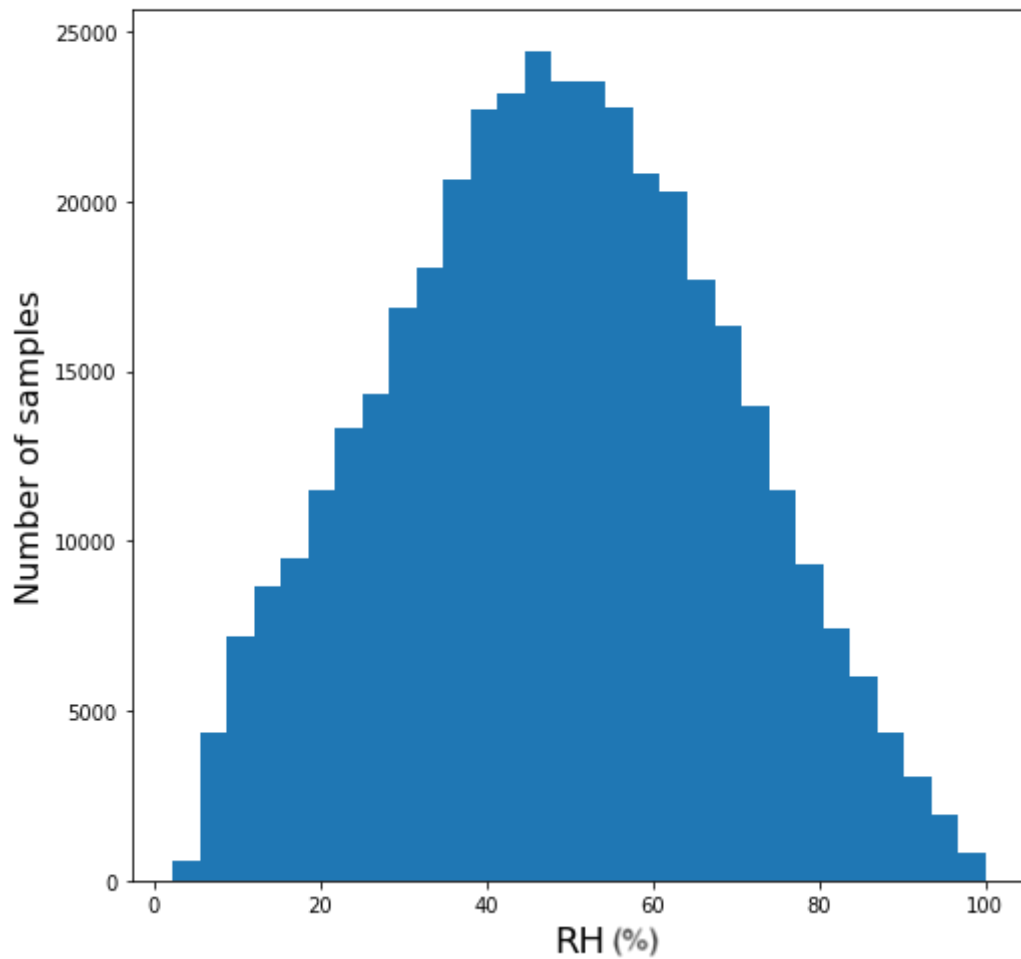
The matrix represents the correlation coefficients between the different parameters considered for the experiment (Figure 3). It was observed that the AOD and ground $PM_{2.5}$ measurements were weakly correlated, with a coefficient of 0.082. There is a negative correlation between HPLB height and $PM_{2.5}$ concentration indicating an overall decrease in surface concentrations for deep, well-mixed boundary layers. Also, wind vectors (UGRD and VGRD) were weakly correlated possibly due to local and regional emissions which are station-dependent.

EPA Air data $PM_{2.5}$ frequency

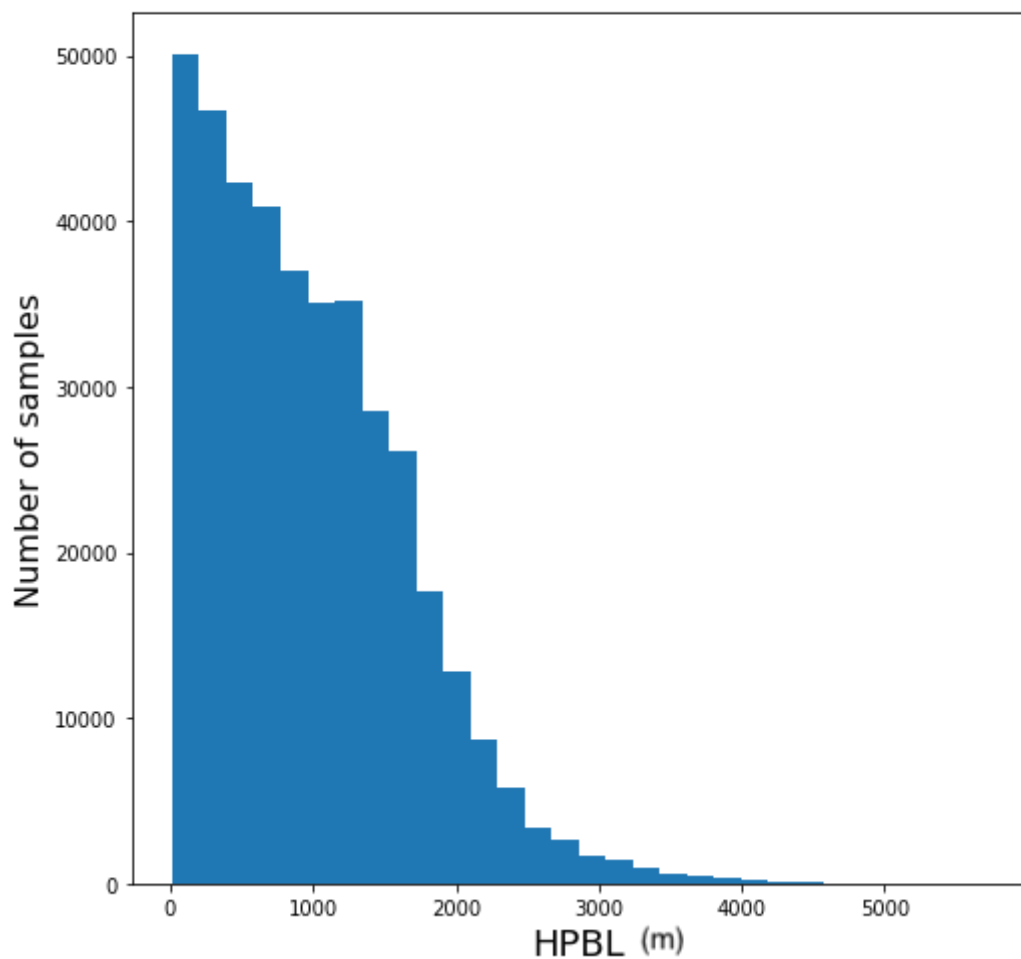




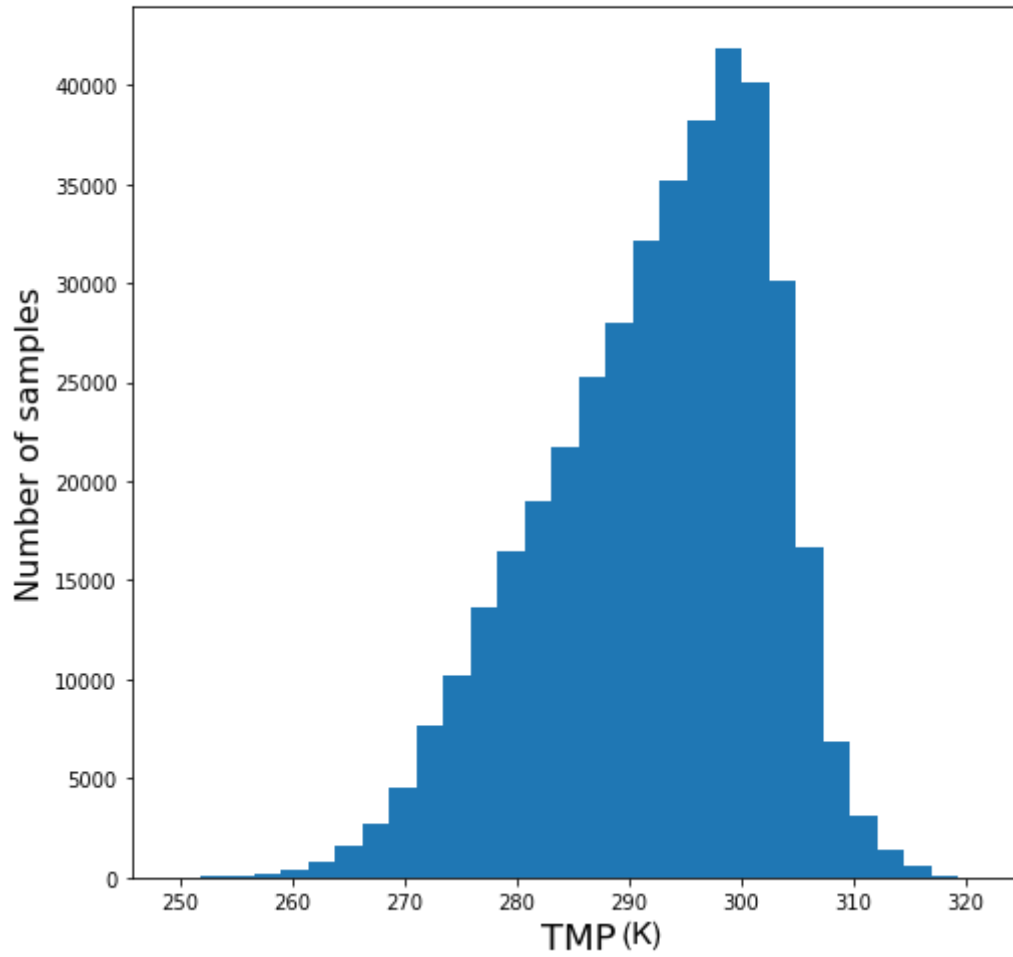
Relative Humidity



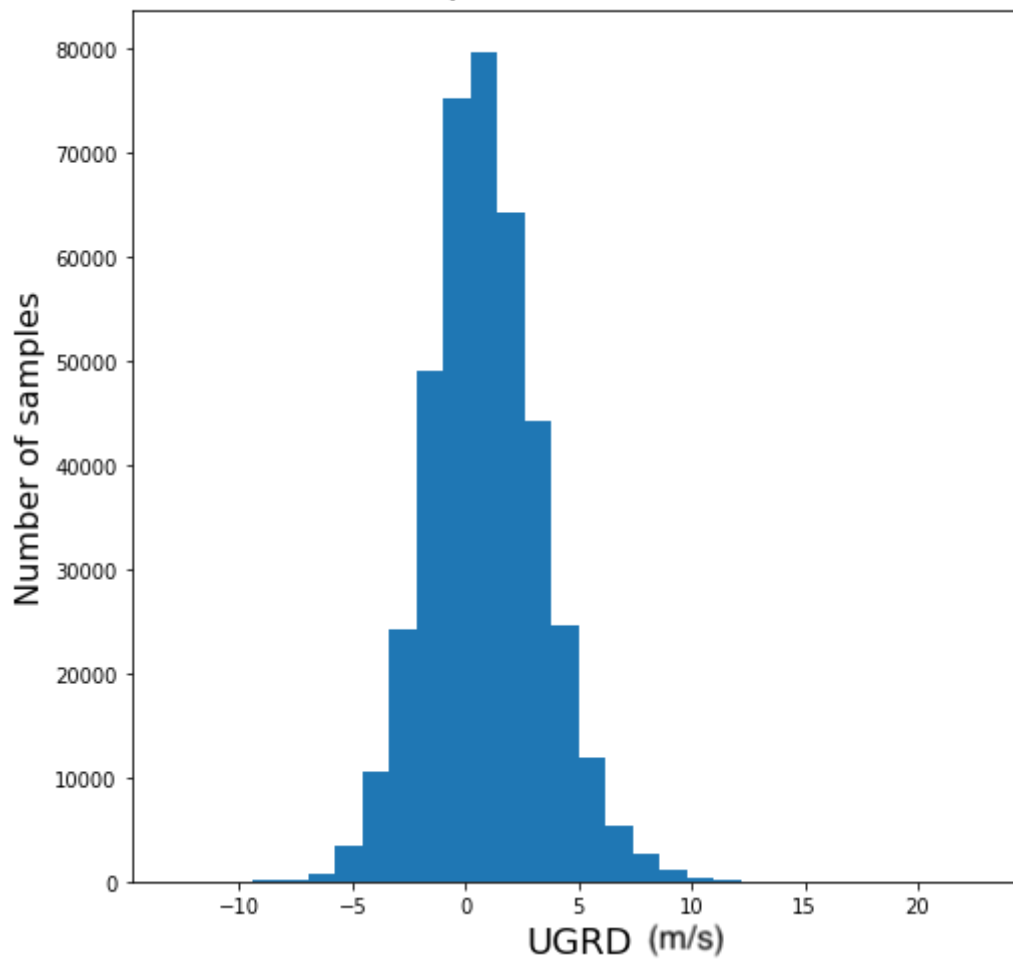
Height of Planetary Boundary layer



Temperature



U Component of Wind



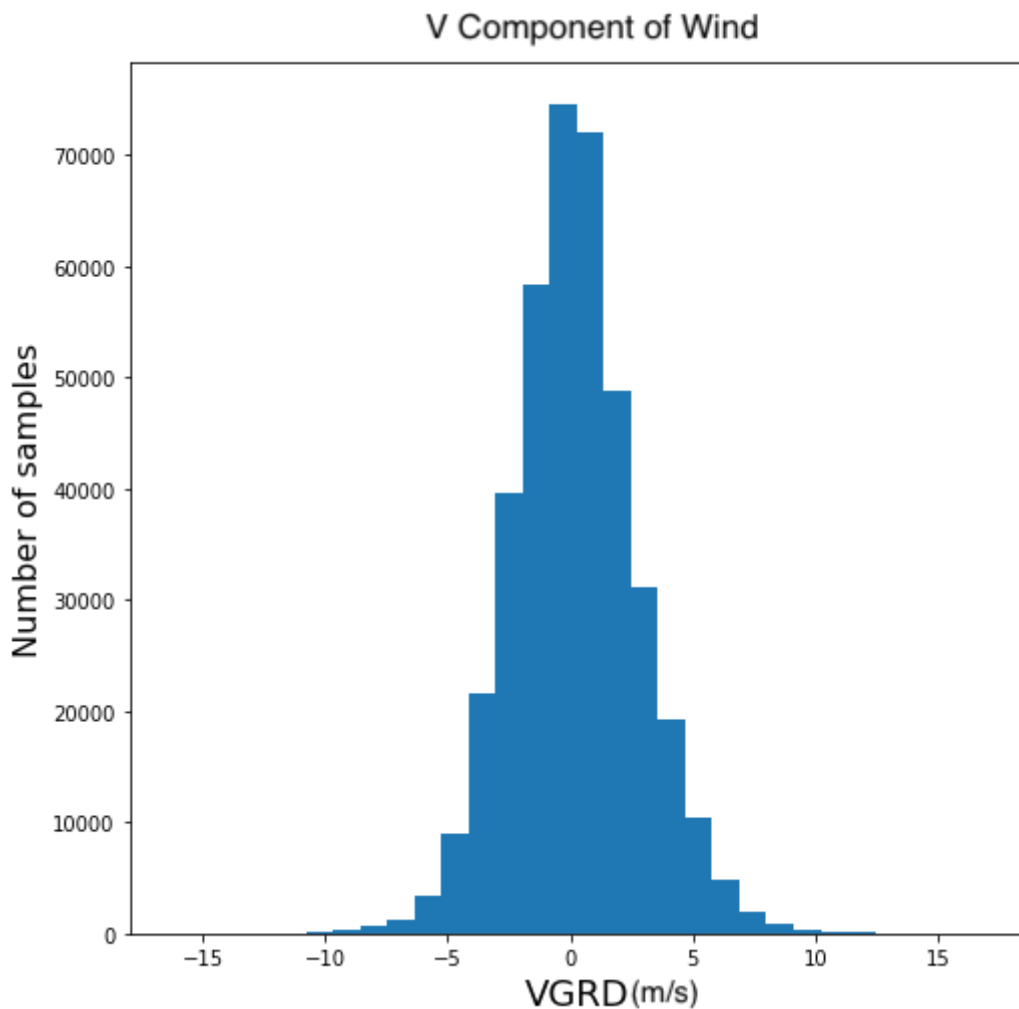
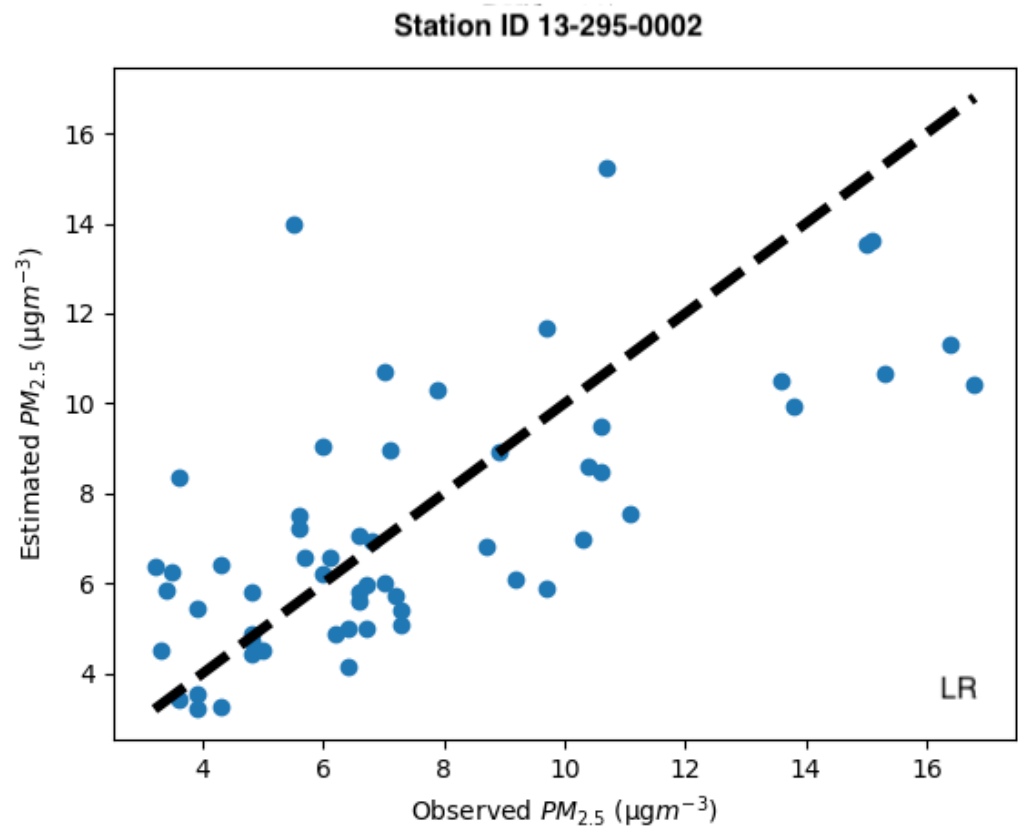


Figure 4. Distribution of dataset parameters EPA PM_{2.5}, GOES AOD, Relative Humidity, Height of Planetary Boundary Layer, Temperature, U and V components of Wind vector.

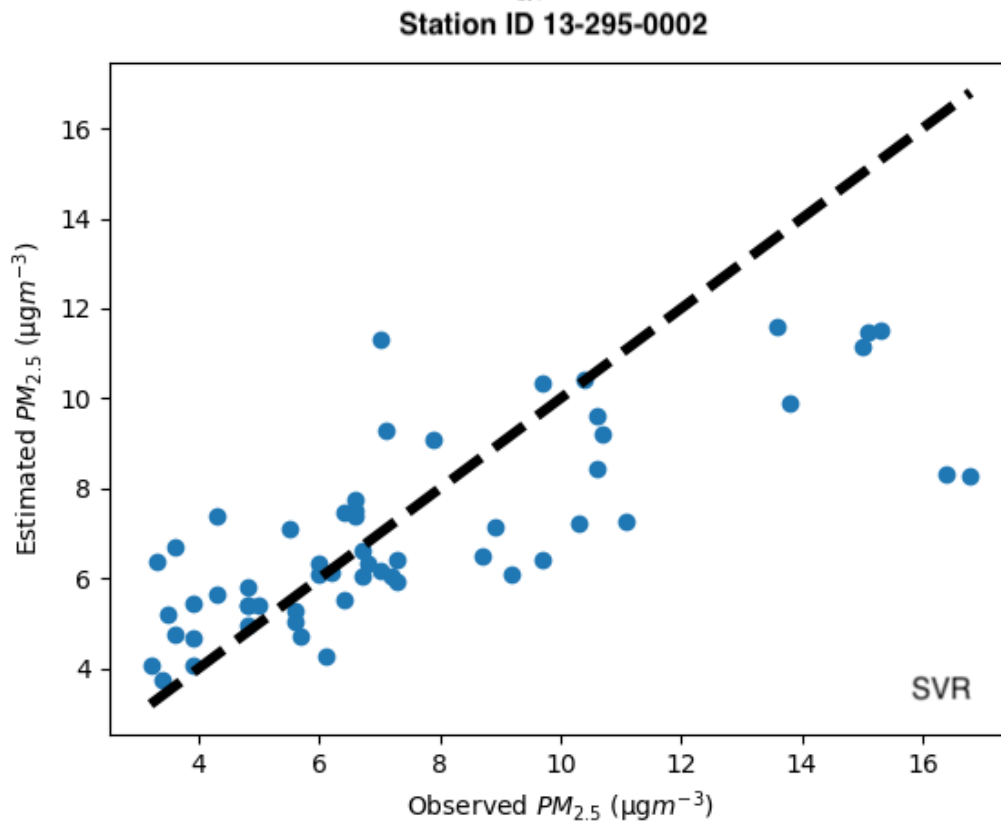
AOD and surface PM_{2.5} have similar shaped distributions including a pronounced positive skew (Figure 4). Relative humidity has a near-normal distribution with a mean at approximately 50%. Similarly, U and V components of wind are normally distribution about 0 m/s. Both temperature and boundary layer height have few observations at the high end of the distribution.

RESULTS

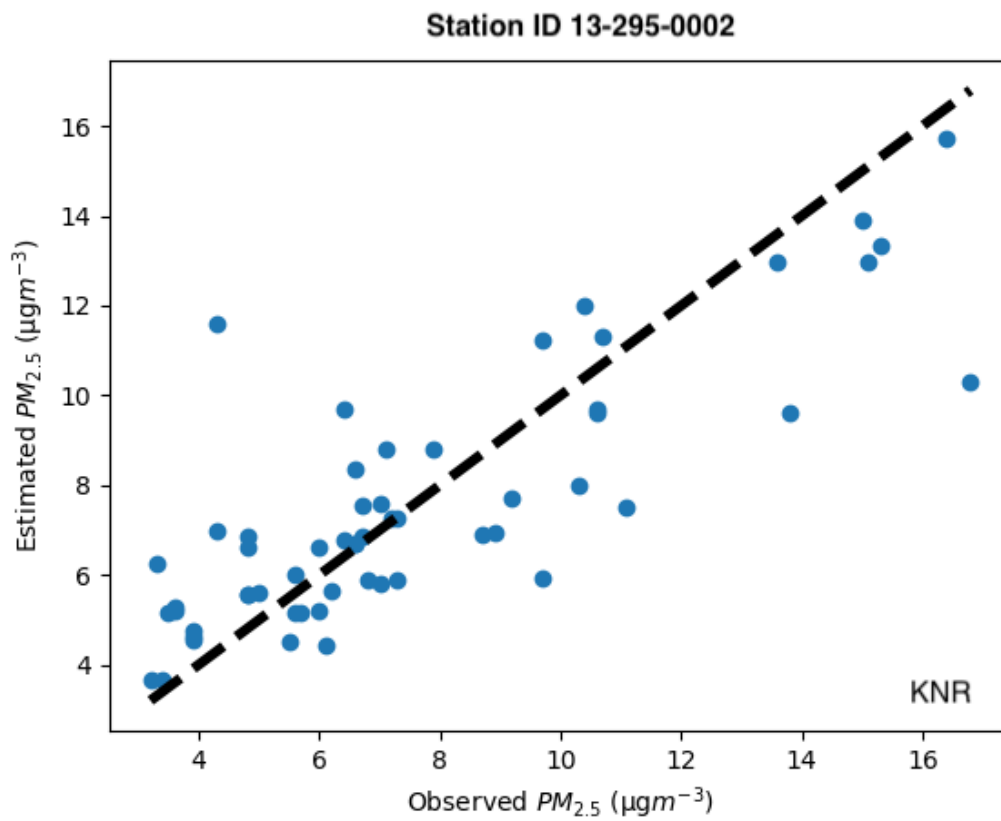
Linear Regression



Support Vector Regression



K Nearest Neighbors - Regression



Long short-term memory

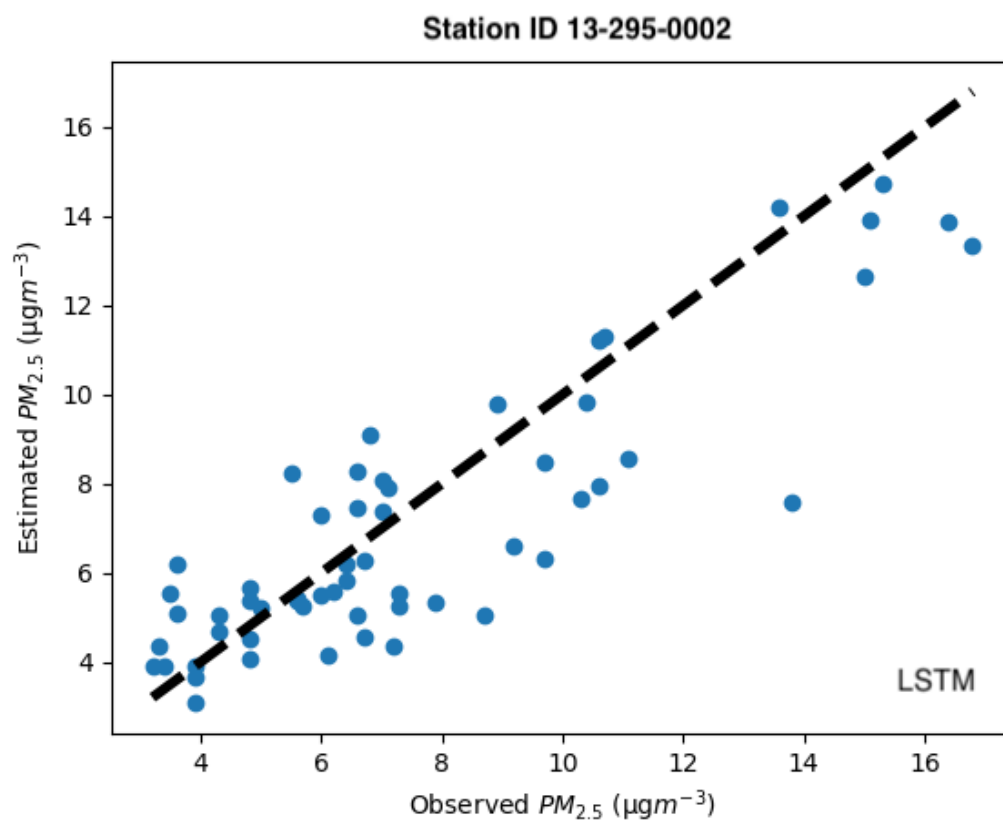


Figure 5. Example of regression for Linear regression (LR), Support Vector Regression (SVR), and K Nearest Neighbors - Regression (KNN), Long short-term memory (LSTM) models.

	Mean Squared Error (MSE)	Correlation Coefficient (R^2)
Long Short-Term Memory (LSTM)	3.3716	0.7316
K-nearest Neighbor Regression (KNR)	4.1465	0.6699
Support Vector Regression (SVR)	6.0235	0.5205
Linear Regression (LR)	6.9596	0.4460

Table 1. Mean Squared Error and Correlation Coefficient for Linear regression (LR), Support Vector Regression (SVR), and K Nearest Neighbors - Regression (KNR), Long short-term memory (LSTM) model.

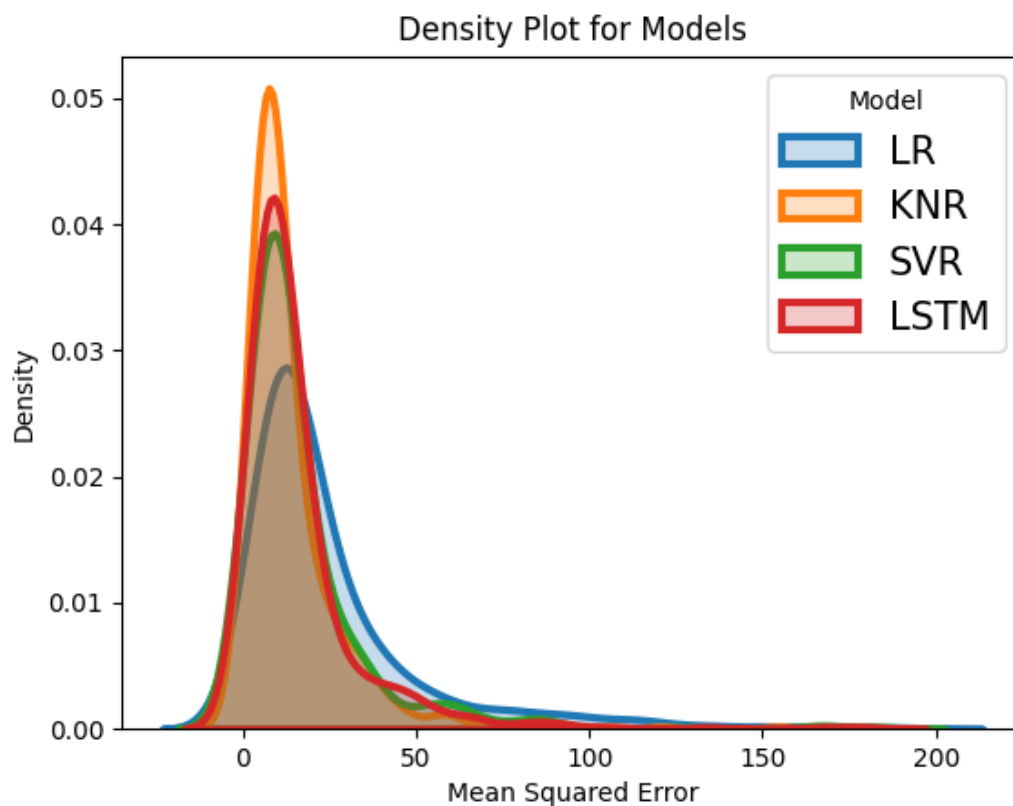


Figure 6. Density plot for Linear regression (LR), Support Vector Regression (SVR), and K Nearest Neighbors - Regression (KNR), Long short-term memory (LSTM) models.

LSTM model seemed to be more adaptive to the patterns in data, followed by KNR (Figure 5). The mean squared error decreased as the complexity of the model increased (Table1). Nonparametric models performed better as attributed by denser lower MSE (Figure 6).

DISCUSSION

Real-time estimation and analysis of $PM_{2.5}$ levels can be critical for various environments and health agencies. A large volume of data is required to produce accurate estimations by factoring in the dependent parameters. A data processing pipeline based on LSTM to consume the AOD measurements from GOES-16 can potentially provide higher frequency estimations of surface $PM_{2.5}$ levels and surpass the spatial and temporal limitations. Here we estimate surface $PM_{2.5}$ concentrations for stations across the United States. From the results, the following can be inferred:

- Extremely high $PM_{2.5}$ concentrations and a low number of samples can result in poor performance for all the models
- Local and regional effects are implicitly included in the model
- LSTM requires 4 consecutive hours of observations which:
 - further limits the training and testing data available
 - limits the earliest possible prediction time during the day

Generalised LSTM model results

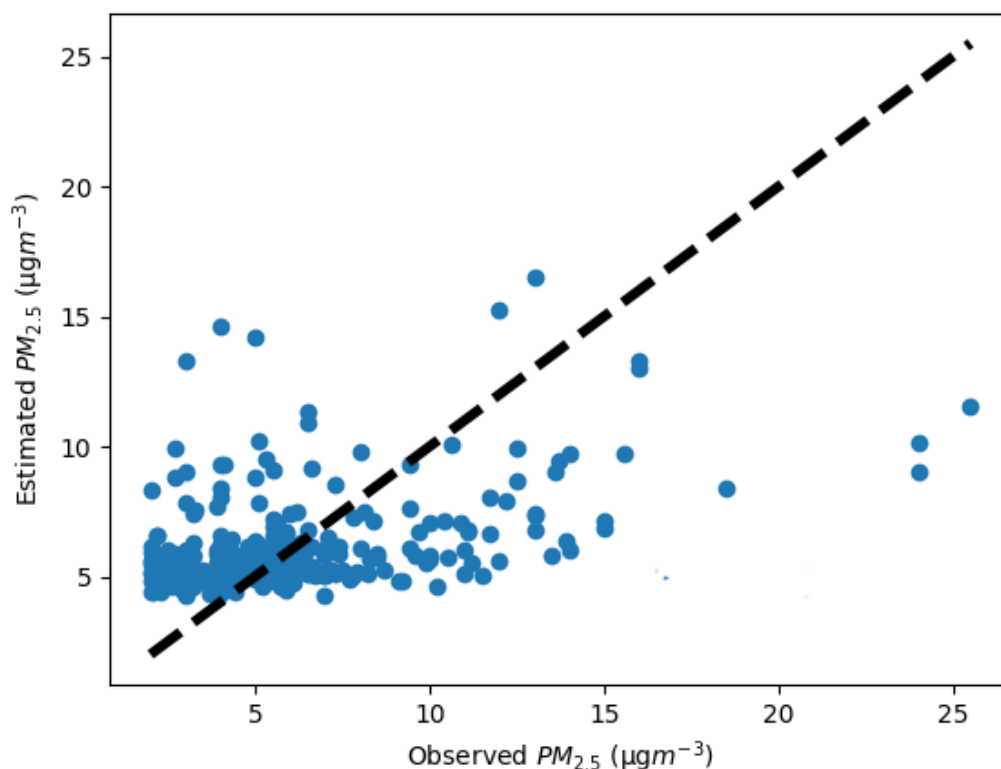


Figure 7. Scatter plot for Long short-term memory (LSTM) model for CONUS region.

Generalised KNR model results

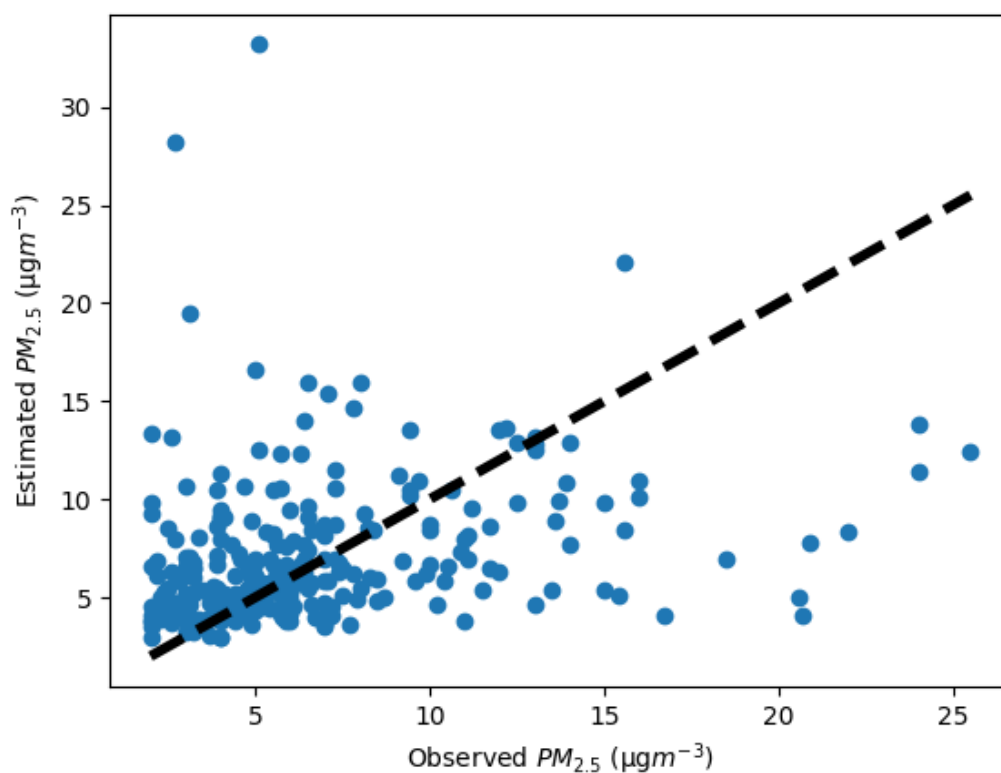


Figure 8. Scatter plot for K Nearest Neighbors - Regression (KNR) model for CONUS region.

The MSE for LSTM and KNR were 15.79 and 22.52 respectively, which is notably higher than the stationwise models. This indicates that additional information, such as indicators of local and regional emissions, may need to be included in the models. Although KNR performed better at station-level (Figure 6), LSTM can potentially provide a more generic model.

FUTURE WORK

- Refine the current model configuration by including additional parameters that introduce site-specific features such as local anthropogenic emissions and further refine the model configurations for better estimations
- Provided the utility of using advanced methods for $PM_{2.5}$ surface concentration information has been demonstrated, our goal is to have a single LSTM model applicable to the entire US.
- Quantify the impact that high spatial and temporal resolution satellite derived AOD has on estimating surface air quality across the United States

This research is supported by NASA Grant NNM11AA01A as part of the Interagency Implementation and Advanced Concepts Team (IMPACT) program

ABSTRACT

Fine particles released from activities such as vehicle exhausts, burning of fuels, forest fires are known to cause severe impacts on public health. Particulate matter (PM) with diameter less or equal to $2.5\text{ }\mu\text{m}$, known as PM_{2.5}, is known to cause or exacerbate cardiovascular and respiratory illnesses. The Environmental Protection Agency (EPA) monitors the levels of particulate matter using air quality monitors stationed throughout the Continental United States (CONUS). The measurements indicate concentrations from point locations, which may not always be available and may not represent the concentrations of surrounding areas. To overcome these limitations in air quality monitoring, Aerosol Optical Depth (AOD) products at high spatiotemporal resolution from geostationary satellites may be used to estimate the levels of surface PM_{2.5}. However, additional meteorological factors that nonlinearly affect surface PM_{2.5} concentrations such as relative humidity, temperature, height of the planetary boundary layer must be considered. This information is accessible from numerical modeling such as from the National Oceanic and Atmospheric Administration's (NOAA) High Resolution Rapid Refresh (HRRR) model which resolves near real-time atmospheric conditions over the CONUS.

The analysis of the data from these different sources and derivation of the relationship among them to estimate surface PM_{2.5} can be a tedious task and may require significant computation and processing. Deep learning approaches are appropriate for such complex estimation problems as it defines relations among multiple non-linear parameters. We propose a deep learning neural network model that would utilize the AOD retrievals from Geostationary Operational Environmental Satellite (GOES) 16 and factor in the data from the other aforementioned sources to estimate the PM_{2.5} levels. The presentation will illustrate the development of the neural network model and its ability to provide estimations of surface PM_{2.5} over the CONUS based on the high spatial temporal AOD retrievals from GOES 16.