

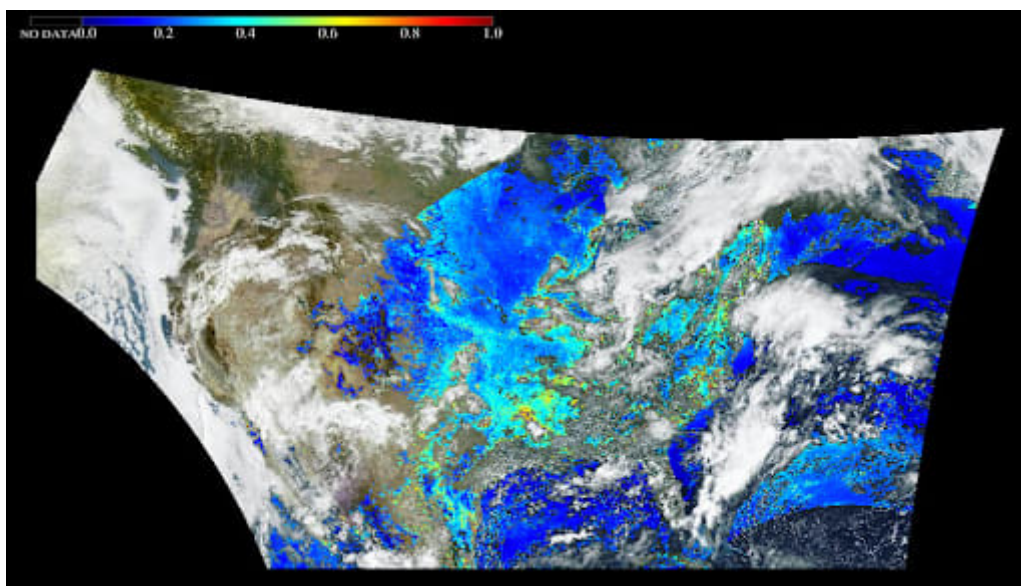
## 1. INTRODUCTION

Tropospheric aerosols impact weather and climate through scattering and absorption of radiation (Ramaswamy et al., 2019), and degrade air quality (Hoff & Christopher, 2009), thereby adversely impacting human health (Cohen et al., 2017). Fine particles, or aerosols with aerodynamic diameter less than  $2.5\ \mu\text{m}$  ( $\text{PM}_{2.5}$ ), pose the greatest risk to health as they are easily inhalable and impact respiratory and cardiovascular functions (Shaddick et al., 2020). Direct measurements of surface  $\text{PM}_{2.5}$  concentrations are made by ground-based stations, representing air quality at a specific geographic location, which may not provide a holistic representation of the air quality over a given area. The limitation in spatial coverage of ground-based stations can be mitigated through the use of satellite remote sensing retrievals, which typically estimate the aerosol loading in an atmospheric column, known as aerosol optical depth (AOD). While the AOD- $\text{PM}_{2.5}$  relation facilitates contiguous  $\text{PM}_{2.5}$  estimation from satellites, it is not robust in case of aerosols residing aloft. To mitigate this issue meteorological and surface processes are included in the development a more complex relation (Liu et al., 2009). The availability of a large number of data in the last decade has enable the use of machine learning (ML) methods in order to obtain a data-driven relation between  $\text{PM}_{2.5}$ , AOD and aforementioned variables. Prior studies focused on estimating  $\text{PM}_{2.5}$  from AOD have used these variables as input to different machine learning (ML) models to infer  $\text{PM}_{2.5}$  concentration on polluted regions, however, few such studies have been published concentrating on the Contiguous United States (CONUS) (Di et al. 2016; Hu et al. 2017). The AOD- $\text{PM}_{2.5}$  relation can be highly variable in the space domain, therefore estimating  $\text{PM}_{2.5}$  concentration on a continental scale may necessitate a large number of ML models (Chudnovsky et al. 2012).

The aim of this initiative is twofold. Firstly, to implement and compare different ML methods to estimate surface  $\text{PM}_{2.5}$  over the CONUS, leveraging data from geostationary satellites, atmospheric models and surface measurements. Secondly, to introduce a novel technique to decrease the large number of regional-based models utilizing a statistical measure.

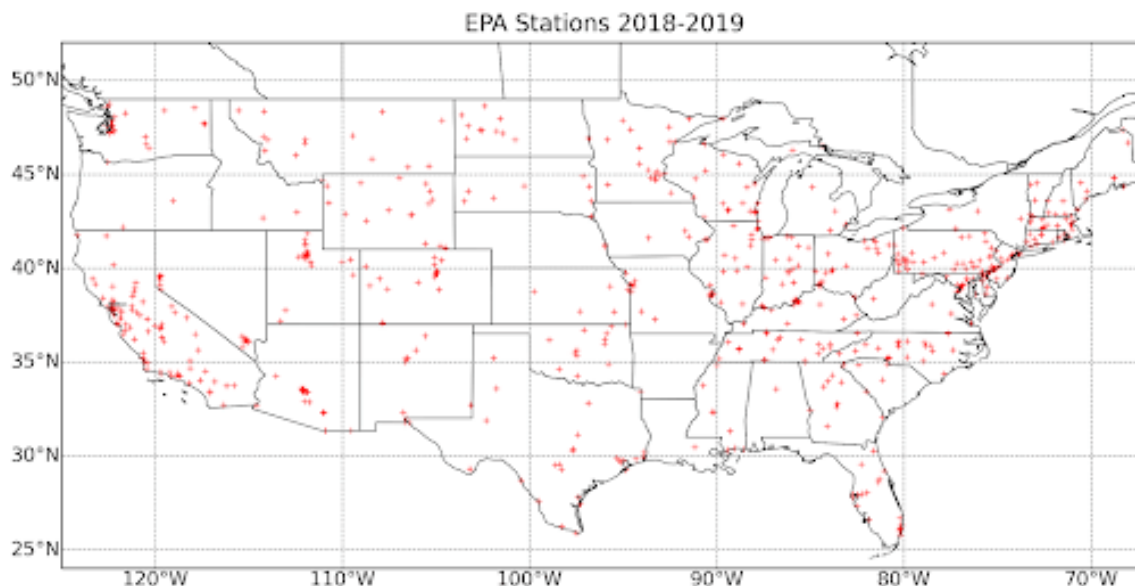
## 2. DATA AND PREPROCESSING

Aerosol Optical Depth GOES-16 2019/08/06 15:41 UTC



**Figure 1.** Aerosol optical depth (AOD) over the CONUS for August 6th, 2019.

The National Oceanic and Atmospheric Administration's (NOAA's) Geostationary Operational Environmental Satellite-16 (GOES-16) has been providing valuable information for modeling and monitoring atmospheric pollution. Level-2+ AOD over CONUS (Kondragunta et al., 2020; GOES-R AOD ATBD, 2018) at 550 nm product is available in 5 minutes interval time and 2-km spatial resolution, which also includes quality flags to isolate high and medium retrievals. AOD is not retrieved for pixels containing clouds, and for retrievals over snow/ice, bright land surface, or areas of sun glint (Figure 1). PM<sub>2.5</sub> concentration is obtained by the Environmental Protection Agency (EPA) air quality system (AQS) for 317 monitoring stations across the CONUS, which provide observations at hourly intervals (Figure 2). Meteorological data is collected by NOAA's High Resolution Rapid Refresh (HRRR; Benjamin et al. 2016) model at 3-km spatial resolution.



**Figure 2.** EPA stations over the CONUS.

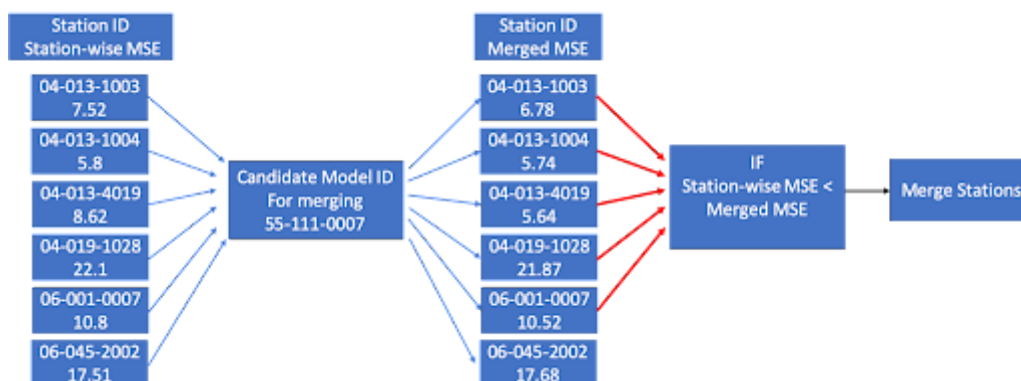
The National Oceanic and Atmospheric Administration's (NOAA's) Geostationary Operational Environmental Satellite-16 (GOES-16) has been providing valuable information for modeling and monitoring atmospheric pollution. Level-2+ AOD over CONUS (Kondragunta et al., 2020; GOES-R AOD ATBD, 2018) at 550 nm product is available in 5 minutes interval time and 2-km spatial resolution, which also includes quality flags to isolate high and medium retrievals. AOD is not retrieved for pixels containing clouds, and for retrievals over snow/ice, bright land surface, or areas of sun glint (Figure 1). Importantly, AOD is not retrieved over thick smoke (Satellite and ground based assessment of the impact of Canadian fires on U.S. Air Quality (<http://agu.confex.com/agu/fm19/meetingapp.cgi/Paper/490194>)). PM<sub>2.5</sub> concentration is obtained by the Environmental Protection Agency (EPA) air quality system (AQS) on an hourly basis and for 317 monitoring stations across CONUS (Figure 2). Meteorological data is extracted from the NOAA's High Resolution Rapid Refresh (HRRR; Benjamin et al. 2016) model at 3-km spatial resolution. More specifically, we use surface temperature (T), relative humidity (RH), u/v-wind components and planetary boundary layer height (PBLH) to account for different atmospheric conditions at the lowest portion of the troposphere. The aforementioned data are collected over the time period 2018-2019.

Pre-processing of data includes collocation of all the dataset over the EPA stations at the closest time. Further subsetting in the time domain between 14-23 UTC is performed to mitigate sun angle effects. Data for which the AOD is negative or the quality flag is larger than 2 (low quality) are removed. In the space domain, AOD retrievals might not always be available at the closest point from an EPA-station location, therefore AOD pixels are extracted within a circle of radius 10 km (euclidean distance) and their average is calculated.

The final dataset (59055), which encompasses all the aforementioned parameters that are collocated with the ground-based (EPA) stations, is then split in training (80%) and test (20%) size.

## 4. APPLICATIONS

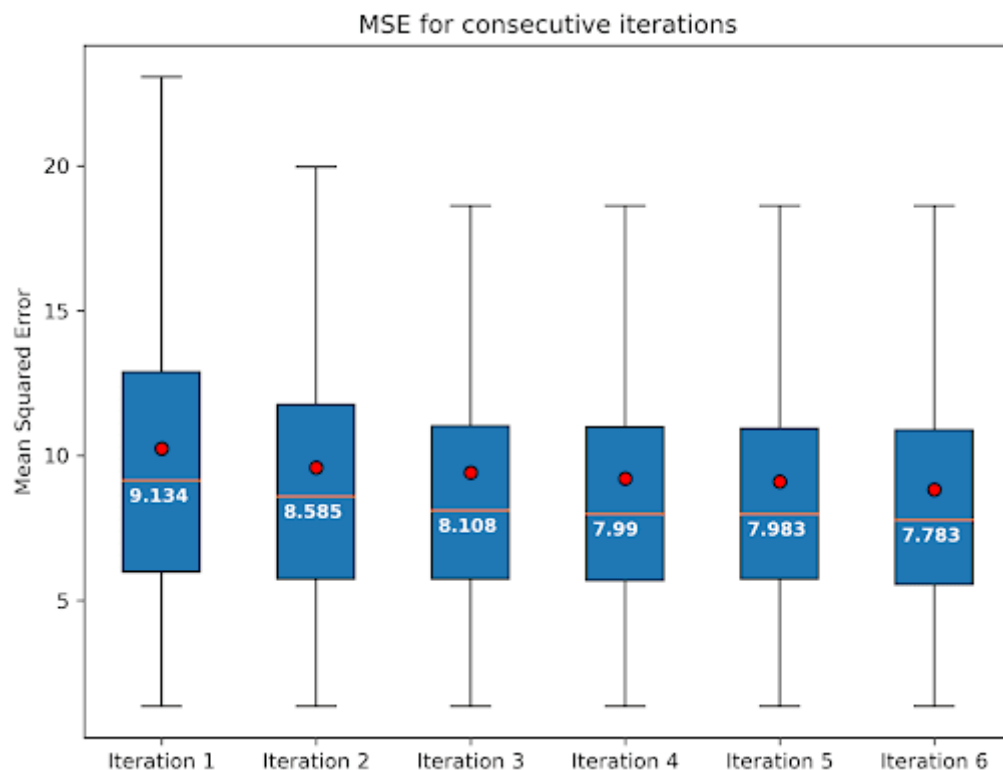
### 4.1 Fleet Model



**Figure 5.** Model architecture for the fleet model.

Feature selection using ensemble models has been utilized in the ML community and has shown to help in identification of best initialized stochastic models. To investigate the spatial patterns of PM<sub>2.5</sub> concentration against the generalized model (XGBoost), the LSTM architecture was chosen as a base model to train station-wise models. LSTM models are therefore developed for each of the EPA stations and trained on accumulated dataset from the individual station throughout the period 2018-2019.

Early studies recommended that the effects of extrinsic factors that influence the AOD-PM<sub>2.5</sub> relationship should be examine separately for each region (Hoff & Christopher, 2009; Chudnovsky et al. 2012). Therefore, a station-wise model's approach might indicate that there are extrinsic characteristics that might not be captured by generalized models. In a given area, the height of the boundary layer and relative humidity maybe less variant, thus leading to a more robust AOD-PM<sub>2.5</sub> relation than other areas. A new technique, termed fleet training (Figure 5), is proposed to account for these common, unknown characteristics as well as iteratively improve the model. To identify common characterisitcs between different stations the following technique is developed: Each of the models trained on data from an individual station is subsequently trained on the rest of the stations iteratively, by merging the corresponding data. If the model trained on the merged data performed better (decrease in the MSE) than the model trained on the individual station, then merging of the two datasets occurs. This process is performed for multiple iterations, until the average MSE is not improved any further. A total of 6 iterations are performed in this case leading to a decrease in the average MSE from 9.134 to nearly 7.783 and from 317 to 129 models (Figure 6).

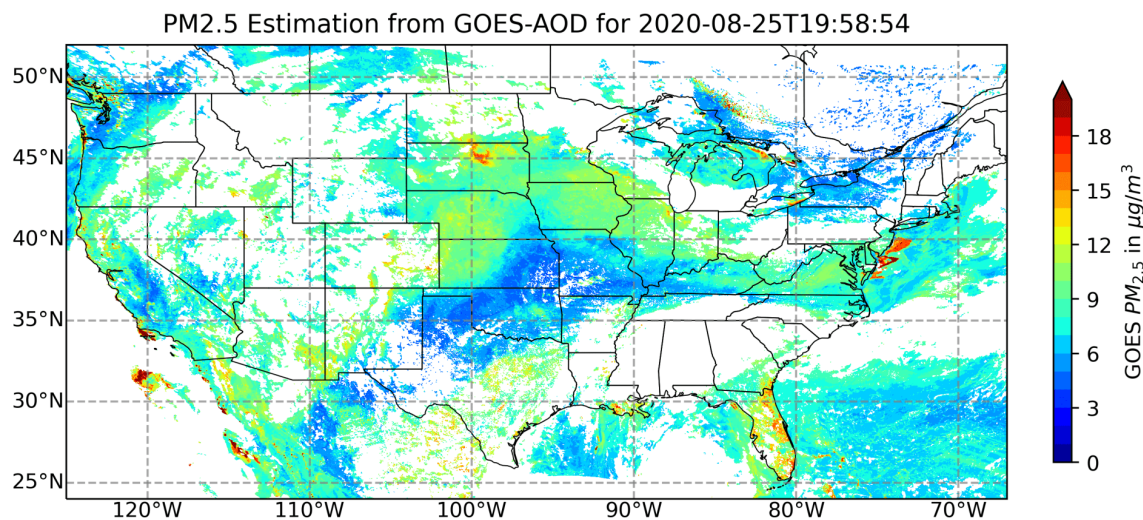


**Figure 6.** PM<sub>2.5</sub> concentration measurements from EPA

ground-based stations over CONUS for 2020-08-25 at

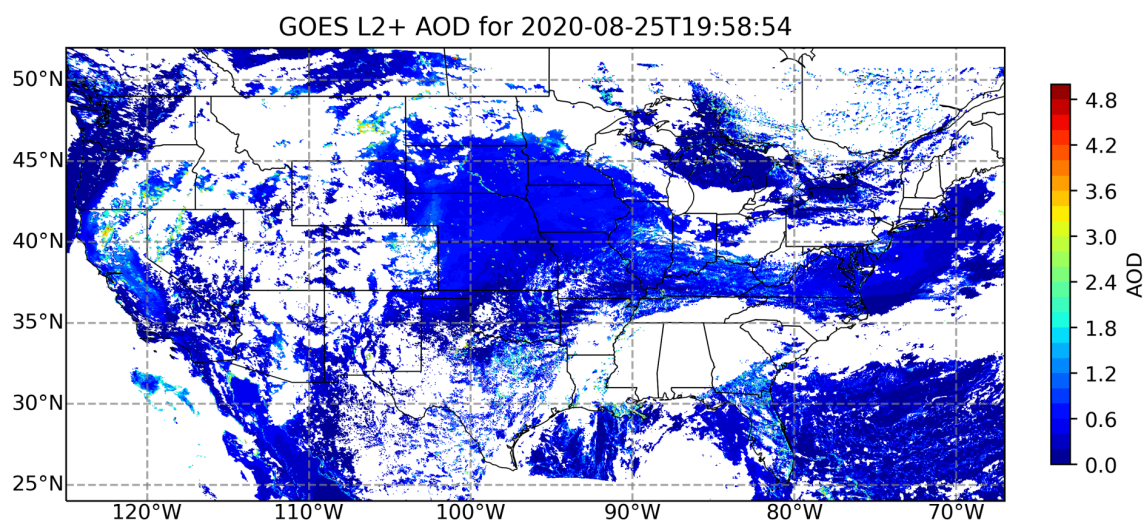
20:00 UTC.

#### 4.2 PM<sub>2.5</sub> Estimates

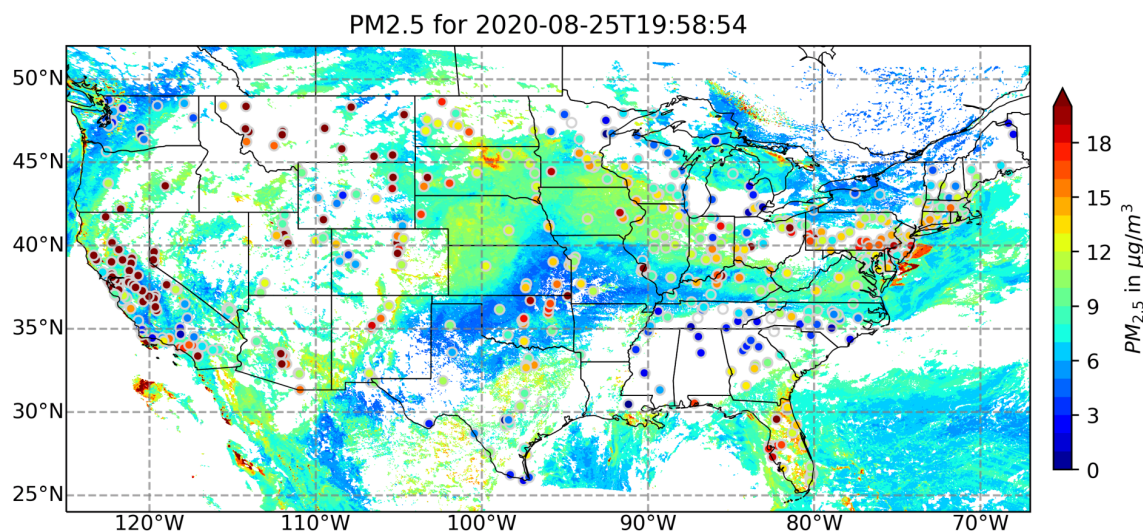


**Figure 7.** PM<sub>2.5</sub> estimation from the generalized XGBOOST model over CONUS for 2020-08-25 at 19:58:54 UTC.

The 2020 California wildfire season set a new record for the number of wildfires in the state (wiki ([https://en.wikipedia.org/wiki/2020\\_California\\_wildfires](https://en.wikipedia.org/wiki/2020_California_wildfires))). The generalized model was used to calculate average  $PM_{2.5}$  concentrations for each grid cell on August 25, 2020 at 19:58UTC. For this time stamp, Figure 7 shows a map of the spatial distribution of  $PM_{2.5}$  estimates. The maximum magnitude of  $PM_{2.5}$  estimates is  $44 \mu m m^{-3}$  and the 85th percentile has a value of  $9.06 \mu m m^{-3}$ . From a qualitative perspective, several hotspots are prevalent over the CONUS in this map, especially over Florida, South Dakota and Northern California. AOD retrievals from GOES-16 (Figure 8) are also elevated over California, Texas and Northern California. AOD trends also agree with  $PM_{2.5}$  estimates over New Orleans. However, low AOD is observed over South Dakota, where  $PM_{2.5}$  estimate is high. To obtain a better understanding of the trends and pattern of  $PM_{2.5}$ , Figure 9 is created, which presents  $PM_{2.5}$  estimates from the model and  $PM_{2.5}$  measurements from the EPA ground-based stations over plotted on the same map.  $PM_{2.5}$  measurements from the EPA ground-based stations have a maximum of  $112 \mu m m^{-3}$  and 85th percentile of  $20 \mu m m^{-3}$  (Figure 10). Elevated concentration of  $PM_{2.5}$  over the California fires is not well captured by the model estimates which might be attributed to low number of samples of extreme  $PM_{2.5}$  in comparison to the total number of training samples or to erroneous AOD retrievals over very thick smoke.



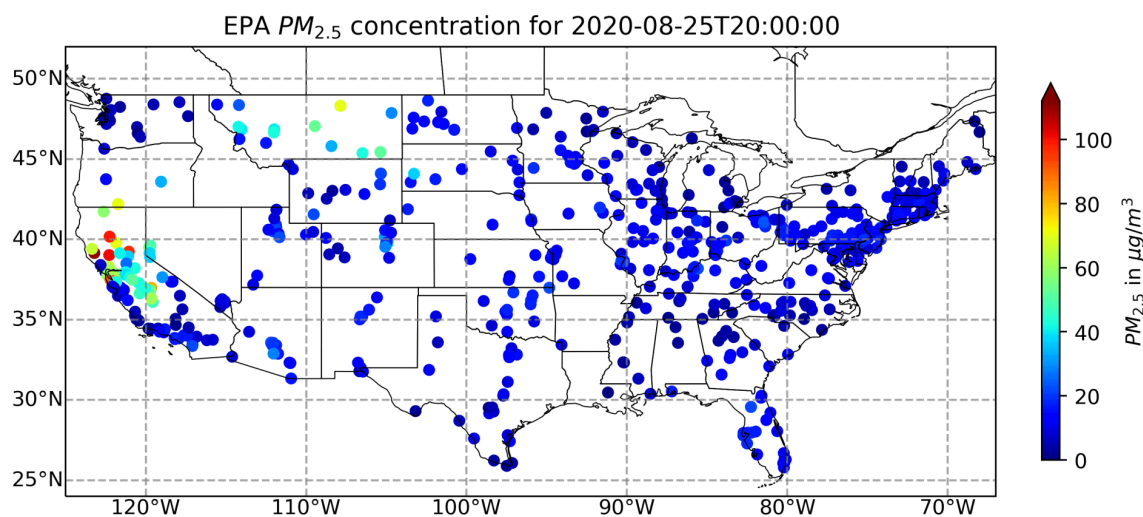
**Figure 8.** Aerosol optical depth retrievals from GOES-16 for 2020-08-25 at 19:58:54 UTC.



**Figure 9.** PM<sub>2.5</sub> estimation from the generalized

XGBOOST model over CONUS for 2020-08-25 at 19:58:54 UTC. Scattered points correspond to PM<sub>2.5</sub> concentration obtained from the EPA ground-based stations at 20:00 UTC. PM<sub>2.5</sub> values are capped at 20

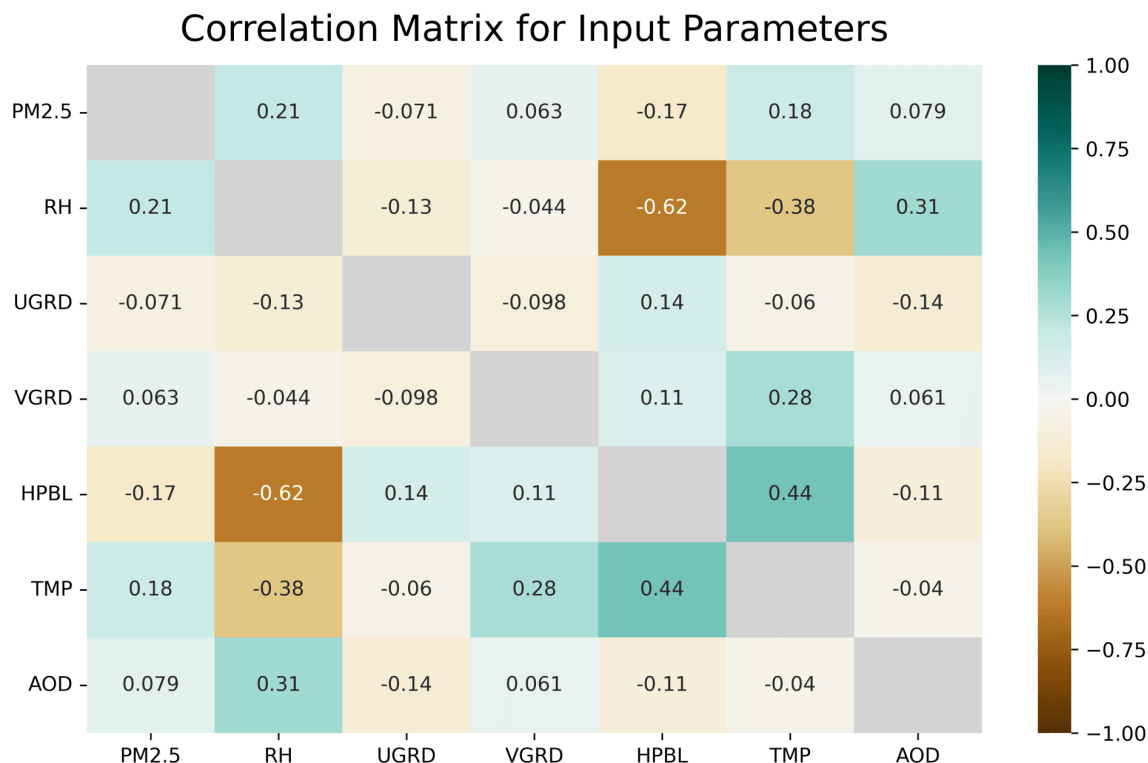
µm m<sup>-3</sup>.



**Figure 10.** PM<sub>2.5</sub> concentration measurements from EPA ground-based stations over CONUS for 2020-08-25 at 20:00 UTC.

### 3. MODEL'S INPUT AND ARCHITECTURE

#### 3.1 Model's Inputs

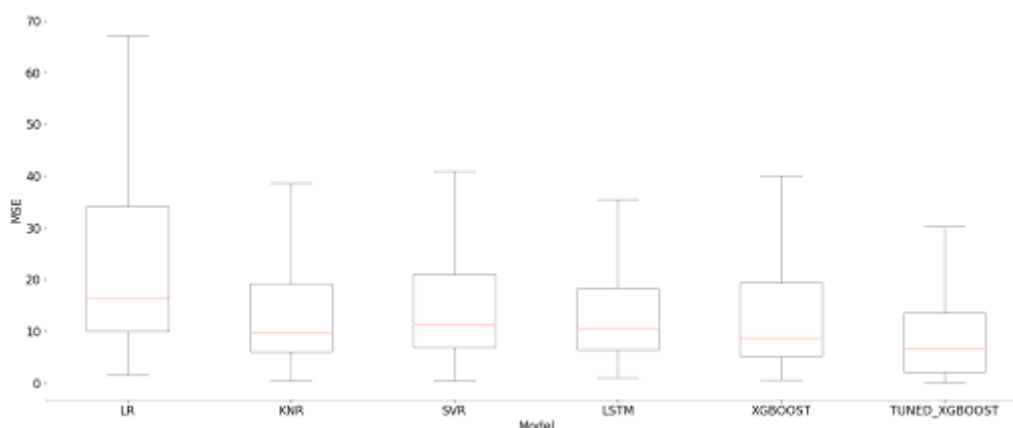


**Figure 3.** Correlation matrix for the variables used in this study. EPA PM<sub>2.5</sub> concentration (PM<sub>2.5</sub>); Relative humidity (RH); U-component of wind vector (UGRD); V-component of wind vector (VGRD); Height of planetary boundary layer (HPBL); Temperature (TMP); Aerosol optical depth retrievals (AOD).

The estimation of PM<sub>2.5</sub> from satellites is based on the AOD, which is a columnar estimate of the depth of aerosols in the atmosphere over a grid box. However, under certain atmospheric conditions this estimate might deviate from the "truth" measurement obtained by ground-based stations. To account for such cases, meteorological parameters from atmospheric models (i.e. HRRR) have been included. To visualize the relations among the different variables used as input to the ML model a correlation matrix is created (Figure 3). Temperature (TMP) can affect the formation of (PM<sub>2.5</sub>) particles by enhancing the photochemical reactions between precursors (Gupta and Christopher 2009). Temperature inversion can also reduce the vertical mixing and therefore increase chemical concentration of precursors. A correlation of 0.18 among the two parameters is obtained throughout the time range of this study and for all station data. Relative humidity (RH) affects the partition of semi-volatile species, with increased RH facilitating the partition of semi-volatile species into the aerosol phase, thus resulting in an increase in PM concentration (Hu et al., 2008). A correlation of 0.21 indicates a relatively weak dependence of PM<sub>2.5</sub> on the RH. Wind speed affects the horizontal dispersion (small- to large-scale transport) of pollutants, thus modulating their concentration. The correlation matrix indicates a very weak relation of the U- (-0.071) and V-wind (0.063) components with the PM<sub>2.5</sub> concentration. The planetary boundary layer height (HPBL) has a direct effect on the concentration of aerosols in lower part of the atmosphere. As the HPBL grows, pollution levels will be diluted within the boundary

layer, resulting in lower  $PM_{2.5}$  concentration measured by surface instruments (Gupta & Christopher 2009). Once again the correlation matrix indicates a slightly negative correlation coefficient (-0.17) which capture this inverse relation. The small positive correlation is found between the AOD and  $PM_{2.5}$  (0.079), which might be attributed to differences in the magnitude of correlation between the western (low) and eastern (high) CONUS (Paciorek et al. 2008), as well as the amount of available measurements (number of stations). Nevertheless, a correlation coefficient value of near zero does not mean that there is no correlation, as there could be a nonlinear relation. Additionally, the aggregated data over all stations may be indicative of the difficulty to account for the regional differences of the atmosphere on a continental scale.

### 3.2 Generalized model



**Figure 4.** Mean Squared Error (MSE) for different machine learning models. Linear regression (LR); k-nearest neighbour (KNR); Support vector machine (SVR); Long-short term memory (LSTM); Extreme-gradient boosting (XGBOOST).

To further explore the relation of  $PM_{2.5}$  concentration with AOD retrievals and the aforementioned atmospheric variables, different machine learning methods are used, which are trained on the aggregated datasets for the development of a generalized model. Those methods range from simple linear regression model to more complex tree-based and deep learning methods (Figure 4). The Linear Regression (LR) model attains a high median MSE (18) and a large range (0-68), thus failing to capture the complex relation between the input variables and the target ( $PM_{2.5}$ ). The rest of the machine learning models, that is k-Nearest Neighbour (kNR), Support vector machine (SVR), Long-Short Term Memory (LSTM), eXtreme-Gradient Boosting (XGBOOST) and tuned XGBOOST outperformed the LR with small MSE (~10) and range less than 40. kNR and XGBOOST have similar performance with a median MSE of approximately 10, however the latter has a smaller MSE range (0-35). SVR and XGBOOST attain the same range of MSE (0-40), however the XGBOOST has a smaller median (8) from all the aforementioned models. Although LSTM is adequate to handle sequence dependence prediction problems, tree-based models such as the XGBOOST can perform better (smaller MSE) with less amount of available data. This is in accordance with Xu et al. (2018) who also found that the XGBoost was among the three best performing models against a total of eight for the estimation of monthly average ground-level  $PM_{2.5}$  concentrations estimation over British Columbia. In order to develop a robust model, the XGBOOST is tuned (e.g.

number of iterations) using the learning curves as a diagnostic tool, providing the final model (TUNED\_XGBOOST). This model is characterized by a median MSE of 5 and a range 0-32.

## 5.CONCLUSIONS AND ACKNOWLEDGEMENTS

Tropospheric aerosols can pose a significant threat for human health among other effects. This study focuses on the estimation of surface PM<sub>2.5</sub> concentration combining remote sensing, in-situ and atmospheric model datasets. To obtain a robust relation between satellite retrieved AOD and PM<sub>2.5</sub> concentration, a multi-parametric model needs to be defined. A comparison of different ML models concludes that the XGBOOST outperformed the rest models. A generalized XGBOOST model is trained by accumulating data from multiple ground-based stations, satellite AOD retrievals and atmospheric parameters, the latter obtained from atmospheric models. PM<sub>2.5</sub> estimates were obtained on August 25, 2020. The evaluation of model output with AOD and PM<sub>2.5</sub> measurements indicates that the generalized model was able to capture the overall patterns, with the exception of California fires, where the estimates were low. To obtain a better representation of PM<sub>2.5</sub> spatial characteristic station-wise LSTM models are created and trained on each individual station. The large number of these models is decreased from 317 to 129 by introducing the fleet training method, which in turn leads to a decrease in the average MSE over all remaining models. This method is based on merging individual station data into a station-wise model iteratively if the MSE after merging is decreased in comparison with the station-wise model.

Future work will focus on evaluating the fleet training model on specific events, as well as on a 7-day average map in order to obtain continuous estimates over the CONUS.

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

# DESIGN YOUR IPOSTER

Background

insert image

remove image

Title Text Font

Default

Title Text Color

Textbox Title Color

Side Textbox Title Background Color

Fill Color

Subtitle Text Font

Middle Textbox Title Background Color

Gradient Color

Default

Subtitle Text Color

Textbox Text Font

OPEN Arrow Color

Default

Side Textbox Text Color

Light Gray

Textbox Title Font

Side Textbox Background Color

Default

Middle Textbox Text Color

Middle Textbox Background Color

reset all styles to default

Notice!

Your iPoster has now been unpublished and will not be displayed on the iPoster Gallery.

You need to publish it again if you want to be displayed.

**You must be registered to publish your poster.**

Please visit [service@agu.org](mailto:service@agu.org) (<https://www.agu.org/Fall-Meeting/Pages/Register-Housing/Registration-Rates>) to get registered </p> <p>If you have recently registered, you may have to wait up to 10 minutes to be able to publish. </p> <p>If you are still unable to publish, please contact us at: <a href=>

Because of maintenance we have just saved your content and will within a few minutes logout all users and restart our server. We will be back in a moment.

Sorry for the inconvenience!

Because of maintenance we will within a few minutes restart our server. We will be back in a moment.

Sorry for the inconvenience!

LINK:

# AUTHOR INFORMATION

Georgios Priftis email: [gp0014@uah.edu](mailto:gp0014@uah.edu)

## ABSTRACT

Particulate matter (PM) with a diameter of less or equal to  $2.5\text{ }\mu\text{m}$ , known as  $\text{PM}_{2.5}$ , affects human health as it penetrates the respiratory system. The Environmental Protection Agency (EPA) measures the atmospheric concentration of  $\text{PM}_{2.5}$  using air quality monitors stationed throughout the Continental United States (CONUS). Such measurements are points on a spatial domain and therefore, might not be representative of the air quality at nearby areas considering that the composition of the atmosphere is highly variable from place to place. Satellite based AOD permits a spatially uniform means of estimating  $\text{PM}_{2.5}$  and new geostationary satellites provide high temporal and spatial resolution estimation of AOD. However, the concentration of  $\text{PM}_{2.5}$  is non-linearly dependent on other atmospheric parameters that include relative humidity, temperature, and height of the planetary boundary layer. This information may be estimated at similar spatial and temporal resolutions as AOD 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 estimation of  $\text{PM}_{2.5}$  concentration is a multi-parametric problem that considers the effect of temporal dependencies among the different parameters. Deep learning approaches are appropriate for such complex estimation problems as they intrinsically capture relations among multiple non-linear parameters. This study compares deep-learning methods to traditional regression analysis to demonstrate the capabilities of these methods in predicting  $\text{PM}_{2.5}$  concentrations. Additionally, a novel ensemble learning approach is employed to identify scientific processes that could further improve the estimation of  $\text{PM}_{2.5}$  concentration. Utilizing Long Short-Term Memory (LSTM) neural networks, which are suitable for multivariate time series estimation problems as they are capable of learning long-term dependencies, individual models are created for each EPA station and trained on the aforementioned dataset collocated over each station. Individual station models are merged if the model's performance is improved by reducing the root mean squared error (RMSE) metric. This ensemble training method ultimately reduces the RMSE value. Evaluation of these results provide insights into physical processes and related observable parameters that may contribute to  $\text{PM}_{2.5}$  concentrations. Identified parameters evaluated to be statistically different between the merged and unmerged models are expected to improve overall performance. These new parameters are then utilized for reevaluation of the deep learning methods with an extreme gradient boosting model with an RMSE of 5.5 providing the best results.

## REFERENCES

- Benjamin, S. G., Weygandt, S. S., Brown, J. M., Hu, M., Alexander, C. R., Smirnova, T. G., ... & Manikin, G. S. (2016). A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Monthly Weather Review*, 144(4), 1669-1694.
- Chudnovsky, A. A., Lee, H. J., Kostinski, A., Kotlov, T., & Koutrakis, P. (2012). Prediction of daily fine particulate matter concentrations using aerosol optical depth retrievals from the Geostationary Operational Environmental Satellite (GOES). *Journal of the Air & Waste Management Association*, 62(9), 1022-1031.
- Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., ... & Forouzanfar, M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, 389(10082), 1907-1918.
- Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., & Schwartz, J. (2016). Assessing PM<sub>2.5</sub> exposures with high spatiotemporal resolution across the continental United States. *Environmental science & technology*, 50(9), 4712-4721.
- Hoff, R. M., & Christopher, S. A. (2009). Remote sensing of particulate pollution from space: have we reached the promised land?. *Journal of the Air & Waste Management Association*, 59(6), 645-675.
- Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., & Liu, Y. (2017). Estimating PM<sub>2.5</sub> concentrations in the conterminous United States using the random forest approach. *Environmental science & technology*, 51(12), 6936-6944.
- Xu, Y., Ho, H. C., Wong, M. S., Deng, C., Shi, Y., Chan, T.-C., & Knudby, A. (2018). Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level pm<sub>2.5</sub>. *Environmental Pollution*, 242, 1417-1426.
- Gupta, P., & Christopher, S. A. (2009). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. *Journal of Geophysical Research: Atmospheres*, 114(D14).
- Kondragunta, S., Laszlo, I., Zhang, H., Ciren, P., and Huff, A.: Air Quality Applications of ABI Aerosol Products from the GOES-R Series, in: The GOES-R Series: A New Generation of Geostationary Environmental Satellites, Elsevier, Amsterdam, the Netherlands, Oxford, UK, Cambridge MA, USA, 203-217, 2020.
- Liu, Y., Paciorek, C. J., & Koutrakis, P. (2009). Estimating regional spatial and temporal variability of PM<sub>2.5</sub> concentrations using satellite data, meteorology, and land use information. *Environmental health perspectives*, 117(6), 886-892.
- Paciorek, C. J., Liu, Y., Moreno-Macias, H., & Kondragunta, S. (2008). Spatiotemporal associations between GOES aerosol optical depth retrievals and ground-level PM<sub>2.5</sub>. *Environmental science & Technology*, 42(15), 5800-5806.
- Ramaswamy, V., Collins, W., Haywood, J., Lean, J., Mahowald, N., Myhre, G., ... & Storelvmo, T. (2019). Radiative forcing of climate: the historical evolution of the radiative forcing concept, the forcing agents and their quantification, and applications. *Meteorological Monographs*, 59, 14-1.
- Seinfeld, J. H., & Pandis, S. N. (2006). Atmospheric chemistry and physics from air pollution to climate change.

Shaddick, G., Thomas, M. L., Mudu, P., Ruggeri, G., & Gumy, S. (2020). Half the world's population is exposed to increasing air pollution. *NPJ Climate and Atmospheric Science*, 3(1), 1-5.

Xu, Y., Ho, H. C., Wong, M. S., Deng, C., Shi, Y., Chan, T.-C., & Knudby, A. (2018). Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level pm2.5. *Environmental Pollution*, 242, 1417–1426.