Predicting PM$_{2.5}$ Concentrations Across USA Using Machine Learning

P. Preetham Vignesh[1], Jonathan H. Jiang[2], P. Kishore

[1.] University of California, Los Angeles, USA

[2.] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA.

[3.] Retired, University of California, Irvine, USA

Correspondence: Jonathan.H.Jiang@jpl.nasa.gov

Abstract:

Fine particulate matter with a size less than 2.5 µm (PM$_{2.5}$) is increasing due to economic growth, air pollution, and forest fires in some states in the United States. Although previous studies have attempted to retrieve the spatial and temporal behavior of PM$_{2.5}$ using aerosol remote sensing and geostatistical estimation methods the coarse resolution and accuracy limit these methods. In this paper the performance of machine learning models on predicting PM$_{2.5}$ is assessed with Linear Regression (LR), Decision Tree (DT), Gradient Boosting Regression (GBR), AdaBoost Regression (ABR), XG Boost (XGB), k-nearest neighbors (KNN), Long Short-Term Memory (LSTM), Random Forest (RF), and support vector machine (SVM) using PM$_{2.5}$ station data from 2017-2021. To compare the accuracy of all the nine machine learning models the coefficient of determination (R^2), root mean square error (RMSE), Nash-Sutcliffe efficiency (NSE), root mean square error ratio (RSR), and percent bias (PBIAS) were evaluated. Among all nine models the RF and SVM models were the best for predicting PM$_{2.5}$ concentrations. Comparison of the PM$_{2.5}$ performance metrics displayed that the models had better predictive behavior in the western United States than that in the eastern United States.

1. Introduction:

Air pollution has had negative effects on human health and has interfered with social functions; particles with diameters less than 2.5 $\boldsymbol{\mu}$ m (PM$_{2.5}$) have especially been the primary pollutants in many cities in the USA. Among air pollutants, PM$_{2.5}$ is among the most harmful and can easily cross the human defense barrier, enter the lungs, and cause human disease and even death because of its small particle size and potential for long-term exposure (Wu et al., 2018; Chen et al., 2019c; Wei et al., 2019). The PM$_{2.5}$ observations were from environmental monitoring stations, however, the quantity of available PM$_{2.5}$ data presented regional differences due to the uneven station distribution.

He et al. (2016) conducted research that indicates the $PM_{2.5}$ pollution index was positively correlated with the emergency admission rate of female acute myocardial infarction and with the increased incidence of diabetes and hypertension. According to the latest urban air quality database, 98% of low and middle income countries with more than 100,000 inhabitants do not meet the World Health Organization (WHO) air quality guidelines [2].

Several researchers have used satellite remote sensing data for spatial monitoring coverage in their studies to estimate $PM_{2.5}$ concentrations (Fang et al., 2016; Hu et al., 2017; Park et al., 2019). One way of using remote sensing satellites for estimating $PM_{2.5}$ levels is through the aerosol optical depth (AOD) parameter, which refers to the solar radiation attenuation due to the scattering and absorption characteristics of aerosols within the atmosphere (Hutschison et., 2005; Van Donkelaar et al., 2010; Soni et al., 2018). Wang and Christoper (2003) was the first estimated $PM_{2.5}$ using AOD measurements from Moderate Resolution Imaging Spectrometer (MODIS). Several researchers noted that satellite AOD as well as monitoring sources and transport of aerosols are key variables in estimating $PM_{2.5}$ and air quality (Gupta and Christopher, 2009). Most have used linear regression models to correlate AOD and $PM_{2.5}$ (Gupta and Christopher, 2009). Grahremanloo et al., 2021 examined seasonal behavior of $PM_{2.5}$ over Texas using the Random Forest model. Liu et al. (2005) studied $PM_{2.5}$ levels in three different areas such as urban, suburban, and county in the Eastern United States using multiple linear regression (MLR). They concluded that the model performance may decrease since the satellite images have a relatively coarse spatial resolution since each pixel represents a large area on the ground.

The design of a model for time series prediction focuses on the application of algorithms to predict future events based on past trends. The model captures the variables with certain assumptions and represents the existing dynamic relations, summarizing them to better understand the process that produced the past data to better predict the future. Most of the above studies have used linear and

59   non-linear regressions to correlate various parameters with $PM_{2.5}$ concentrations over a particular

60   region. In our study we focused on the entire United States and predicted $PM_{2.5}$ concentrations over

61   various regions using different machine learning models.

62   Recently, due to an increase in the application of machine learning models to various fields

63   in order to increase the accuracy of predictions, machine learning has also been used to predict particle

64   concentrations (Kuremoto et al., 2014; Ong et al., 2016; Gui et al., 2020). However, the data mining

65   does not only differ from one study to another but also in terms of classification algorithms and used

66   features. The regression, boosting models, and deep learning-based methods display remarkable

67   performance in time-series data processing to make predictions (Hochreiter and Schmidhuber, 1997).

68   The estimation using traditional statistical methods requires a large amount of historical data to

69   construct the relationship between explanatory variables and target variables (Breiman, 2001b). Since

70   machine learning is a very promising tool to forecast pollution, we proposed applying this approach

71   to predict $PM_{2.5}$ concentrations in the USA. The model predictions based on ML algorithms were

72   checked by cross-validation and evaluated using appropriate metrics such as root mean square

73   (RMSE) and mean absolute error (MAE).

74   Earlier studies used a limited number of statistical models, but in our study, we used nearly six

75   machine learning models to find the best accuracy of predictions. In addition to this, our research

76   paper took a novel approach in $PM_{2.5}$ concentration research by exploring concentrations over USA

77   as opposed to China where many existing $PM_{2.5}$studies have already been conducted. The purpose of

78   this paper is to present the predictions of $PM_{2.5}$ over different states over the USA. The data collection

79   and different machine learning techniques applied in the context of time series predictions are adopted

80   for the present study as described in Section 2. Results and discussion are given in Section 3 and

81   finally the overall conclusions are drawn from the present study presented in Section 4.

2. Datasets:

2.1 Ground PM$_{2.5}$ Measurements:

Daily PM$_{2.5}$ observational data was collected from January 2015 to December 2021 from the openaq air quality database (https://openaq.org/). These datasets are available from nearly 1081 stations around the USA. The PM$_{2.5}$ concentrations of ground sites were taken as the dependent variable of the model. In this paper, the daily PM$_{2.5}$ concentration data of 1081 ground monitoring stations were sorted in to monthly and seasonal data from January 2015 to December 2021, and the data integrity exceeded 97%. The datasets were calibrated and quality-controlled according to national standards. Figure 1 shows the ground-level monitoring site coverage over the United States; these sites collected 7 years of daily continuous observations. From this figure, we can see that PM$_{2.5}$ monitoring sites are greater in number in the eastern part than in the western part of USA. We observed small data gaps and therefore applied linear interpolation for filling the gaps of PM$_{2.5}$ datasets. However, stations are sparsely located, therefore ground level PM$_{2.5}$ monitoring sites face difficulties in meeting the data requirements (Lin et al., 2015). As expected, the PM$_{2.5}$ concentrations were much lower at remote sites compared to urban areas, mainly due to the absence of anthropogenic sources.

This study aims to achieve the best statistical comparison of nine machine learning models: Linear Regression, K-Nearest Neighbors Regressor, Logistic Regression, Gradient Boosting Regressor, Ada Boost Regressor, Decision Tree Regressor, XG Boost, Support Vector Regressor, Random Forest, Support Vector Machine, and LSTM for estimating the PM$_{2.5}$ concentrations over the specified period. The datasets are split into 80% and 20% as training and testing datasets, respectively. The training datasets are used to build the model, and the testing dataset is used to verify the model performance of the trained model.

2.2 K Nearest Neighbors (K-NN):

The K-NN model is one of the earliest ML models (reference). The K-NN model categorizes each unknown instance in the training set by choosing the majority class label among its k nearest neighbors. Its performance is also crucially dependent on the Euclidean distance metric used to define the most immediate neighbors. After determining the Euclidean distance between the data, the database samples are sorted in ascending order from the least distance (maximal similarity) to maximum distance (minimal similarity) [Wu et al. 2008]. The k nearest distances are looked at, and the highest occurring class label of these k nearest points to the instance is decided to be the class label of the previously unknown instance in the training set. Selecting an optimal value of k becomes challenging since too low of a value for k can result in overfitting while a larger value of k can cause the opposite to occur.

2.3 Random Forest (RF):

RF is a machine learning algorithm and was proposed by Breiman (2001); it integrates multiple trees through the idea of ensemble learning, utilizes classification and regression tree (CART) as learning algorithms of decision trees. The RF is a set of decision trees, where the structure of each one, and the space of the variables is divided into smaller subspaces so that the data in each region is as uniform as possible [Hastie et al., 2005 and Breiman, 2001]. It uses the bootstrap resampling technique to randomly extract k samples (with replacement) from the original training set to generate new training samples. RF uses multiple base classifiers to obtain higher accuracy classification results by voting or averaging. RF excels because of its ability to leverage several different independent decision trees in order to classify better, thereby reducing the error from using a single decision tree because oftentimes viewing classification in independent directions can lead to lower error than a single decision tree's direction.

128 **2.4 XGBoost**:

129   This is a highly efficient and optimized distributed gradient boosting algorithm. XGBoost

130 supports a range of different predictive modeling problems such as classification and regression. It is

131 trained by minimizing the loss of an objective function against a dataset, and the loss function is a

132 critical hyperparameter which is tied directly to the type of problem being solved. Regular gradient

133 boosting, stochastic gradient boosting, and regularized gradient boosting are the three main forms of

134 gradient boosting. For efficiency, the system features include parallelization, distributed computing,

135 out-of-core computing, cache optimization, and optimization of data structures to achieve the best

136 global minimum and run time.

137 **2.5 Long Short-Term Memory (LSTM)**:

138   LSTM is well suited for prediction based on time-series data, with better performance, to learn

139 long-term dependency, and it deals with exploding and vanishing gradient problems [Alahi et al.,

140 2016, Kong et al., 2017]. LSTM is superior to traditional ML methods in processing large input data

141 and is a type of Recurrent Neural Network (RNN) [Rumelhart et al., 1986], that has been proposed

142 to predict future outputs using past inputs. LSTM is great at processing time-series data because the

143 $PM_{2.5}$ concentrations are time-dependent, and it can better predict future air pollution concentrations

144 by learning features contained in past air pollution concentration time-series data.

145 **2.6 Decision Tree (DT)**:

146   Decision Trees are one of the most commonly used machine learning models in classification and

147 regression problems. To split a node into two or more sub-nodes DT uses mean squared error (MSE).

148 It is a tree structure with three types of nodes. The root node is the initial node, which may get split

149 into further nodes of the branched tree that finally leads to a terminal node (leaf node) that represents

150 the prediction or final outcome of the model. The interior nodes and branches represent features of

151 a data set and decision rules respectively. The final prediction is the average of the value of the

152 dependent variable in that particular leaf node.

2.7 Gradient Boosting Regression (GBR):

154   The type of boosting that combines simple models called weak learners into a single composite

155 model. Gradient boosting involves optimizing the loss function and a weak learner which makes

156 predictions. Generally, the gradient descent procedure is used to minimize a set of parameters, such

157 as coefficients in a regression equation or weights in a neural network. After estimating loss or error,

158 the weights are updated to minimize that error. Gradient Boosting algorithms minimize the bias error

159 of the model. The Gradient Boosting algorithm predicts the target variable using a regressor and Mean

160 Square Error (MSE) as the cost function (for regression problems) or predicts the target variable with

161 a classifier using a Log Loss cost function (for classification problems).

2.8 Support Vector Regression (SVR):

163   The SVR model is widely applied to time series prediction problems. It is a novel forecasting

164 approach, which is trained independently based on the same training data with different targets. The

165 SVR can be used with functions that are linear or non-linear (called kernel functions). The linear

166 function is used for the linear regression model and evaluates results with metrics such as Root Mean

167 Square Error (RMSE) and Mean Absolute Error (MAE) to estimate the performance of the model.

2.9 AdaBoost Regressor (ABR):

169   AdaBoost (Adaptive Boosting) is a popular technique, as it combines multiple weak classifiers to

170 build one strong classifier. The boosting approach is a class of ensembles of ML algorithms and is

171 described by Schapire (1990). Generally, the boosting approach requires a large amount of training

172 data which is not possible for many cases, and one way of mitigating this issue is by using AdaBoost

173 (Freund and Schapire, 1997). The main difference of AdaBoosting from most of the other boosting

174 approaches is in computing loss functions using relative error rather than absolute error. AdaBoost

7

175 regressor fits the data set and adjusts the weights according to the error rate of the current prediction,

176 and reduces the bias as well as the variance for supervised learning.

177 **2.10 Linear Regression**:

178   Linear Regression is a great statistical tool that achieves to model and predict variables by fitting

179 the predicted values to the observed values with a straight line or surface. This fitting process is

180 implemented by reducing the average perpendicular distance from the straight line/surface (which

181 are the predictions) to the observed values which oftentimes are scattered. The lower this

182 perpendicular distance, the better the line of best fit; based on this line of best fit's equation future

183 values can be predicted. In this case, the line of best fit's equation uses the $PM_{2.5}$ values as the

184 dependent and output variable whereas time is the independent variable.

185 **3.0 Results and Discussion:**

186   Before proceeding to apply machine learning models on the $PM_{2.5}$ data we will first discuss the

187 $PM_{2.5}$ concentrations monthly mean structures, a common method of data exploration to better

188 understand the data and potentially adjust hyperparameters of the models. Figure 2 shows the USA

189 monthly anomalies and quantiles for four years using daily $PM_{2.5}$ values. The monthly anomalies are

190 in percent form, so we subtracted 100 to set the average value to zero. In addition, we estimated the

191 anomaly to be positive or negative. Using anomalies we estimated the minimum, maximum values,

192 the 25%, 75% quantiles, and the interquartile ranges for each month of the entire time period, and the

193 resultant plot is shown in Figure 2. During 2018, in USA, the highest levels of $PM_{2.5}$ were observed

194 in the inland locations and they declined nearly 20% in the year 2019. In the inland areas, $PM_{2.5}$

195 concentrations are primarily influenced by the secondary particles' formation resulting from the

196 oxidation of gaseous precursors (NOx, SOx,and NH3) (South Coast Air Quality Management

197 District, 2017). $PM_{2.5}$ concentrations show a drastic change before and during pandemic years. Before

198   pandemic years the PM2.5 concentrations are higher in the spring and summer months especially

199   towards the end of summer (August) and early fall (September) during summer years.

200   The monthly $PM_{2.5}$ concentrations are greatest in 2018 when compared to other years. The

201   positive anomalies are observed on a higher frequency in August 2018 whereas negative anomalies

202   are observed more in September 2018. This indicates that before COVID-19 the $PM_{2.5}$ concentrations

203   were a little higher than in other years throughout the USA. $PM_{2.5}$ values were also higher in the

204   Eastern USA than in Western USA (Figure not shown). The decrease was moderate (in absolute and

205   relative terms) in urban areas and progressively became lower from the urban to the rural sites. From

206   our review of recent sources, primary traffic emissions are highest at traffic sites in absolute and

207   relative terms (Masiol et al., 2015; Khan et al., 2016, Pietrogrande et al., 2016). Before proceeding

208   with applying machine learning models to the data, a preliminary statistical analysis was performed

209   for each state's $PM_{2.5}$ values and all time series values were freed of trend and outliers. This was done

210   because otherwise the time-series data values would give rise to several issues during training like

211   overfitting or significantly decreasing the performance of the model. The seasonal and annual

212   variations were removed from all states' time series data points from the entire time period. This

213   ensured stationarity in the time series data, which is a preprocessing prerequisite before applying

214   different machine learning algorithms. This is because it is better to observe statistical properties of

215   a time series which do not change over time, since statistical properties would have to be averaged

216   for the entire time period, which is not as accurate.

217   **3.1 Evaluation Parameters:**

218   For model evaluation, the errors between the estimated and true values were evaluated using

219   several evaluation indices (Chadalawada & Babovic 2017; Shahid et al., 2018; Yi et al., 2019). The

220   statistical metrics selected for comparing the performance of the models and error-values between

221   computed and observed data are evaluated by Root Mean Square Error (RMSE): square root of the

222   mean squared differences between observed and predicted, and suggests the dispersion of the sample.

223   Smaller RMSE indicates better performance, and as performance decreases, the RMSE increases.

224   The coefficient of determination (R^2) indicates the collinearity (relationship) between the observed

225   and predicted data. The R^2 value ranges from 0 to 1 (Santhi et al., 2001 and Van Liew et al., 2003).

226   Mean absolute error (MAE): average of the absolute differences between the observed and predicted

227   values where a small value of MAE indicates better performance. Mean absolute percentage error

228   (MAPE): this index indicates the ratio between errors and observations, the lower the MAPE the

229   higher the accuracy (Chen et al., 2018). Root mean square error ratio (RSR): the ratio of the RMSE

230   to the standard deviation of measured data (Stajkowski et al., 2020). RSR is classified into four

231   intervals: very good ($0.0 \leq$ RSR ≤ 0.50), good ($0.50 <$ RSR ≤ 0.60), acceptable ($0.60 <$ RSR ≤ 0.70),

232   and unacceptable (RSR > 0.70), respectively (Khosravi et al., 2018). Nash-Sutcliffe efficiency (NSE):

233   is a normalized statistical metric to determine the relative magnitude of the residual variance relative

234   to the variance or noise (Nash and Sutcliffe 1970). NSE performance ratings are very good ($0.75 <$

235   NSE ≤ 1.0), good ($0.65 <$ NSE ≤ 0.75), satisfactory ($0.50 <$ NSE ≤ 0.65), and unsatisfactory (NSE $\leq$

236   0.50). Percent bias (PBIAS): it measures the average percent of the predicted value that is smaller or

237   larger than the observed value (Malik et al., 2018; Nury et al., 2017). The PBIAS is classified into

238   four ranges, very good (PBIAS $< \pm10$), good ($\pm10 \leq$ PBIAS $< \pm15$), satisfactory ($\pm15 \leq$ PBIAS $<$

239    $+25$), and unsatisfactory (PBIAS $\geq \pm25$).

240
$$MSE = \frac{\sum_{i=1}^{n}(x_{oi} - x_{pi})^2}{N}$$

241
$$MAE = \frac{1}{N}\sum_{i=1}^{n}|x_{oi} - x_{pi}|$$

242

243
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_{oi} - x_{pi})^2}{\sum_{i=1}^{n}(x_{oi} - x_{mean})^2}$$

244

245

246
$$RSR = \frac{RMSE}{STDEV_{obj}} = \frac{\sqrt{\sum_{i=1}^{n}(x_{oi}-x_{pi})^2}}{\sqrt{\sum_{i=1}^{n}(x_{oi}-x_{mean})^2}}$$

247
248

249
$$PBIAS = \left|\frac{\sum_{i=1}^{n}(x_{oi}-x_{pi})}{\sum_{i=1}^{n}x_{oi}}\right| * 100$$

250
251

252
$$NORM = \sqrt{\sum_{i=1}^{n}(x_{oi}-x_{pi})^2}$$

253
254

255
$$MAPE = \frac{\sum_{i=1}^{n}\frac{|x_{oi}-x_{pi}|}{x_{oi}}}{N} * 100\%$$

256

257
$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}(x_{oi}-x_{pi})^2}{\sum_{i=1}^{n}(x_{oi}-x_{mean})^2}\right]$$

258

259   where N refers to the number of data points, x_{oi}, x_{pi} are the observed and predicted daily PM$_{2.5}$

260   concentrations, respectively.

261   The nine machine learning models can describe daily variations of observed and estimated values

262   of PM$_{2.5}$ concentrations as shown in Figure 3 and Figure 4, in which the blue curve represents the

263   observed PM$_{2.5}$ concentrations, while the red curve represents the estimated PM$_{2.5}$ concentrations.

264   We generated time series plots for all states but we showed one state from the western side of the

265   USA: California (Figure 3) and another state from east USA: New York (Figure 4). All nine machine

266   learning models show that the seasonal variability of PM$_{2.5}$ concentration is lower in the spring and

267   summer and higher in autumn and winter, maybe due to atmospheric circulation of autumn and

268   winter. The PM$_{2.5}$ concentrations in the autumn and winter are less accurate because air pollution is

269   more severe than that in spring and summer. The SVM and RF models give better agreement with

270   observed PM$_{2.5}$ concentrations. However, the California PM$_{2.5}$ estimations are less accurate than

271   those of the New York because pollution is more severe due to forest fires in the summer. Sulfate

272   concentrations may reflect regional influences of PM$_{2.5}$; these concentrations decreased from east to

273   west but with higher amounts in California (Meng et al. 2018).

274   Figures 5 and 6 display California and New York's scatter plots of the observed vs estimated

275   daily PM$_{2.5}$ concentrations during the period of observations using different machine learning models

276   respectively. The scatter plot of the two variables suggests a positive linear relationship between

277   them. All points on the scatter plot lie on a straight line; this indicates the differences are zero and

278   suggest a strong correlation between the observed and estimated PM$_{2.5}$ concentrations. Tables 1 and

279   2 indicate the performance and statistical metrics as estimated for New York and California. The

280   metrics of all models in Table 1 are for New York: Random Forest with R^2 = 0.899, MAE = 2.122,

281   and RMSE = 3.121 has less error than the other models. The next model with the lowest error is

282   Support Vector Machine with R^2 = 0.857, MAE = 2.145, and RMSE = 3.125.

283   The performance of the models at different states are good at most sites, as 73% of them show an

284   R2 > 0.62 and 10% show an R2 less than 0.3. Moreover, an average RMSE less than 4.5 Mg/m3 in

285   70% of the states and more than 5 Mg/m3 in rest of the states demonstrates good performance. PM$_{2.5}$

286   estimations are lower and higher than observations with high and low PM$_{2.5}$ concentration scenarios

287   respectively, indicating that estimation accuracy will decline in extreme cases in both states. Zhan et

288   al. (2017) also found similar behavior using PM$_{2.5}$ concentration in some parts of China. This may be

289   due to the model's lack of performanced caused by a smaller amount of training data, especially

290   during extreme PM$_{2.5}$ concentrations. Ghahremanloo et al. 2021 observed PM$_{2.5}$ levels in Texas are

291   maximal in the summer and are attributed to higher temperatures and humidity that accelerate the

292   formation of nitrate and sulfate from NO2 and SO2 (Lin et al., 2019). Overall, the performance of RF

293   is reasonable, with California's R^2, RMSE, and MAE values of 0.77, 3.051 mg/m3, and 2.233 mg/m3,

294   respectively. New York's R^2, RMSE, and MAE values were 0.899, 3.121 mg/m3, and 2.12 mg/m3,

295   respectively. Comparing California's to New York's results, we observe that the California PM$_{2.5}$

296 concentration values and biases were slightly higher. Overall, the average error values are slightly

297 lower in the Eastern states than in the Western states. Each state's R2, RMSE, MAE, and bias values

298 are estimated for each model and we observed RF and SVM models produce better estimates than

299 the other models. On average, the R2 of the SVM model is 5% higher than that of the RF model. The

300 biases are 15% lower in the Eastern states than in the Western states of the USA. The high sulfate

301 concentrations around Los Angeles and Long Beach may be due to the ship emissions, since these

302 two areas combined have one-fourth of all container cargo traffic in the Unites States

303 (http://www.dot.ca.gov) (Vutukuru and Dabdub, 2008). However, the $PM_{2.5}$ estimations in the

304 autumn and winter are less accurate because air pollution is more severe than that present in the spring

305 and summer. Among the nine machine learning models, only the SVM and RF models give desirable

306 results in the mildest air pollution cases. The LSTM model performs the outperformed among all

307 models, which can neither reflect the variations of $PM_{2.5}$ concentrations significantly nor estimate the

308 $PM_{2.5}$ concentrations accurately.

309   A Taylor diagram can display multiple metrics in a single plot and can be used to summarize

310 the relative skill with several states' $PM_{2.5}$ model outputs. The Taylor diagram characterizes the

311 statistical relationship between two fields (Taylor, 2001). In this paper, observed is representing the

312 values based on observations, and predicted indicates that the values were simulated by a machine

313 learning model. Figures 7 and 8 illustrate the Random Forest and Support Vector Machine of standard

314 deviation and correlation of all states of USA. Metrics of RF and SVM were computed at each state,

315 and a number was assigned to each state considered. The position of each number appearing on the

316 plot quantifies how closely model $PM_{2.5}$ values matches with different states. Consider state 50, for

317 example and its correlation is about 0.78. The centered standard deviation difference between the

318 observed and predicted patterns is proportional to the point on the x-axis identified as observed. The

319 dotted line contours indicate the normalized standard deviation values, and it can be seen that in the

320 case of state 50 it is centered at about 1.65. Predicted patterns that agree well with observed test data

321 will lie nearest to the observed marked point. The state values lie near or on the observed dotted line,

322 and it indicates a small predicted pattern difference. Some of the state values are slightly further from

323 the observed value, it also shows that the predicted values are larger than the observed.

4. Conclusion:

325 In this paper, we present the prediction of $PM_{2.5}$ concentrations over USA using various machine

326 learning algorithms with the goal of improving our understanding of the differences among them.

327 Machine learning algorithms are new approaches for analyzing large datasets due to the

328 computational speed and easy implementation for massive data. In this paper we studied and

329 examined nine machine learning models (Linear Regression, Decision Tree, Gradient Boost, Ada

330 Boost, XG Boost, K-Nearest Neighbors, LSTM, Random Forest, and SVM) and their performance

331 in predicting $PM_{2.5}$ concentrations.

332 The obtained machine learning-based methods' accuracies vary in all of USA's states, but the

333 performance of RF (California: R^2=0.77, NSE = 0.817, PBIAS=7.022, and RSR=0.355; New York:

334 R^2=0.899, NSE=0.811, PBIAS=2.989, and RSR=0.331) and SVM (California: R^2=0.71, NSE=0.897,

335 PBIAS=7.027, and RSR=0.424; New York: R^2= 0.857, NSE=0.280, PBIAS=3.011, and RSR=0.338)

336 were better than the other examined methods. Moreover, it should be noted that the accuracy and

337 performance of these machine learning methods are not constant in different climates and regions.

338 Both RF and SVM models' R^2 scores were between 0.71 and 0.899, RMSE scores ranged between

339 3.05 to 3.714, NSE values ranged between 0.811 to 0.899, PBIAS ranged between 2.989-7.027, and

340 RSR scores ranged between 0.331-0.424 for California and New York states. These metrics revealed

341 high model reliability and performed well for both RF and SVM and larger datasets produced better

342 prediction results.

343   Our study can also contribute to limiting human health exposure risks and helping future

344   epidemiological studies of air pollution. With the improved computational efficiency, machine

345   learning models improved prediction performance and served as a better scientific tool for decision-

346   makers to make sound $PM_{2.5}$ control policies. Real-time measurements of the chemical composition

347   of $PM_{2.5}$ taken as regulatory air quality measurements are needed in the future.

348   Several parameters affect $PM_{2.5}$ concentrations; in the future, it is possible to improve the

349   performance of our machine learning models with GDP per capita, urbanization data, and other

350   atmospheric parameters which would be investigated for model development. In the United States

351   more extensive ground monitoring is needed, as the total number of stations is 1000, suggesting the

352   network of stations is too sparse for a large nation (See Figure 1). This becomes much more apparent

353   in some states as also displayed in Figure 1. However, understanding the spatial and temporal

354   distribution of each region over the United States is helpful, especially over rural areas. Considering

355   these areas, a larger amount of data for these locations and other ground-based locations would

356   enhance predicting $PM_{2.5}$ concentrations. Furthermore, the machine learning models can always be

357   updated to yield better results as new data becomes available, therefore, the expansion of sources of

358   data becomes even more important as models can be updated.

363   **Data availability:** All $PM_{2.5}$ data used for this study can be downloaded from the public website

364   https://openaq.org. For additional questions regarding the data sharing, please contact the

365   corresponding author at Jonathan.H.Jiang@jpl.nasa.gov.

References:

Alahi, A., K. Goel, V. Ramanathan, A. Robicquest, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces", in Proc. IEEE Conf. Comput.Vis. Pattern Recognit., Jun, 2016, pp.961-971.

Breiman, L. Random Forests, Mach. Learn. 2001, 45, 5-32. https://doi.org/1.0.1023/A: 1010933404324.

Breiman, L., 2001b, Statistical modeling: the two cultures. Stat. Sci., 16 (3), 199-215, https://doi.org/10.1214/ss/1009213726.

Chadalawada, J., and Babovic, V., 2017. Review and comparison of performance indices for automatic model induction. J. of Hydroinformatics, 21, 13-31, https://doi.org/10.2166/hydro.2017.078.

Chen, S., Li, D.C., Zhang, H.Y., Yu, D.K., Chen, R., Zhang, B., Tan, Y.F. et al., 2019c. The development of a cell-based model for the assessment of carcinogenic potential upon long-term PM2.5 exposure. Environ. Int. 131, https://doi.org/10.1016/j.envint.2019.104943.

Fang, X., Zou, B., Liu, X., Sternberg, T., Zhai, I., 2016. Satellite-based ground PM2.5 estimation using timely structure adaptive modeling. Rem. Sens. Environ, 186, 152-163.

Freund, Y., Schapire, R.E., 1997. A decision-theoritic generalization og on-line learning and an application to boosting, J. computer and Ssystem Sciences, 55 (1), 119-139.

Ghahremanloo, M., Lops, Y., Choi, Y., Mousavinezhad, S., 2021. Impact of the COVD-19 outbreak on air pollution levels in East Asia. Sci. Total Environ. 142226.

Gui, K., Che, H., Zeng, Z., Wang, Y., Zhai, S., Wang, Z., Luo, M., Zhang, L., Liao, T., Zhao, H., Li, L., Zheng, Y., Zhang, X., 2020. Construction of a virtual PM2.5 observation network in China based on high-density surface meteorological observations using the extreme gradient boosting model. Environ. Int. 141, 105801. https://doi.org/10.1016/j.envint.2020.105801.

390   Gupta, P., Christopher, S.A., 2009. Particulate matter air quality assessment using integrated

391   surface, satellite, and meteorological products: 2. A neural network approach. J. Geophys. Res.

392   Atmosphere 114 (D20).

393   Hastie, T,; Tibshirani, R.; Friedman, J,; Franklin, J. The elements of statistical learning: Data mining,

394   inference and prediction. Math. Intell. 2005, 27, 83-85.

395   He, X. N., Chen, P., Zhang, C., Chen, J.Y. Study on the correlation between PM2.5 and onset of acute

396   myocardial infarction among female patients. Child Care China 31, 22, 4626-4629, 2016.

397   Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., et al. Estimating PM2.5 concentrations in

398   the conterminous United States using the Random Forest approach. Environ, Sci., Technol. 2017,

399   51, 6936-6944. https://doi.org.10.1021/acs.est.7b01210.

400   Hochreiter, S., and Schmidhuber, J., 1997. Long short-term memory, Neural Comput., 9, 8, 1735-

401   1780.

402   Hutschison, K.D., Smith, S., Faruqui, S.J., 2005. Correlating MODIS aerosol optical thickness data

403   with ground-based PM2.5 observations across Texas for use in a real time air-quality prediction

404   system. Atmos. Environ. 39 (37), 7190-7203.

405   Lin, C., Li, Y., Yuan, Z., Lau, A.K.H., Li, C., Fung, J.C.H., 2015. Using satellite remote sensing

406   data to estimate the high-resolution distribution of ground-level PM2.5. Remote Sens. Environ.

407   156, 117-128. https://doi.org/10.1016/j.rse.2014.09.015.

408   Khan, M.B., Masiol, M., Forementon, G., Gilio, A.D., de Gennaaro, G., Agostinelli, C., and

409   Pavoni, B, 2016. Carboneous PM2.5 and secondary organic aerosol across the Veneto region (NE

410   Italy). Sci. Total Environ. 542, 172-181, doi:10.1016/j.scitotenv.2015.10.103.

411   Khosravi, K ; Mao, L; Kisi, O; Yaseen, Z. M; Shahid, S. Quantifying hourly suspended sediment load

412   using data mining models: case study of a glacierized Andean catchment in Chile. J. Hydrol.

413   2018, 567, 165-179.

Kong, W., Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang. "Short-term residential load forecasting based on LSTM recurrent neural network", IEEE Trans. Smart Grid, vol. 10, no.1, pp. 841-851, Jan. 2017.

Kuremoto, T., Kimura, S., Kobayashi, K., and Obayashi, M., 2014. Time series forecasting using a deep belief networkwith restricted Boltzmann machines, Neurocomputing, 137, 47-56.

Liu, B,; Philip, S. Y.; Top 10 algorithms in data mining. Knowl. Inf. Syst. 2008, 14, 1-37.

Liu, Y., Sarnat, J.A., Kilaru, V., Jacob, D.J., Koutrakis, P., 2005. Estimating ground-level PM2.5 in the eastern United States using satellite remote sensing, Environ. Sci. Technol., 39, 3269-3278.

Malik, A.; Kumar. A.; Kisi, O. Daily pan evaporation estimation using heuristic methods with gamma test. J. Irrig. Drain. Eng. 2018, 144, 4018023.

Masiol, M., Benetello, F., Harrisom, R.M., Fornenton, G., Gaspari, F.D., and Pavoni, B., 2015. Spatial, seasonal trends and trans-boundary transport of PM2.5 inorganic ions in the Veneto region (northeastern Italy), Atmos. Environ., 117, 19-31, doi:10.1016/j.atmosenv.2015.06.044.

Meng, X., Garay, M.J., Diner, D.J., Kalashnikova, O.V., Xu, J., Liu, Y., 2018. Estimating PM2.5 speciation concentrations using prototype 4.4 km resolution misr aerosol properties over Southern California, Atmos. Environ., 181, 70-81.

Nash J. E., Sutcliffe, J. V. River flow forecasting through conceptual models part I – A discussion of principles. J. Hydrol. 1970, 10, 282-290.

Nury, A.H.; Hasan, K.; Alam, M. J. Bin comparative study of wavelet-ARIMA and wavelet-ANN models for temperature time series data in northeastern Bangladesh. J. King. Saud. Univ. Sci. 2017, 29, 47-61.

Ong, B.T., Sugiura, K., and Zettsu, K., 2016. Dynamically pre-trained deep recurrent neural networks using environmental monitoring for predicting PM2.5, Neural Comput. Appl., 27, 6, 1553-1566.

437 Park, Y., Kwon, B., Heo, J., Hu, X., Liu, Y., Moon, T. 2019. Estimating PM2.5 concentration of the

438   conterminous Unites states via interpretable convolutional neural netwoks. Environ. Pollut.

439   113395.

440 Pietrogrande, M.C., Bacco, D., Ferrari, S., Ricciardelli, I., Scotto, F., Trentini, A., and Visentin, M.:

441   2016. Characteristics and major sources of carbonaceous aerosols in PM2.5 in Emilia Romagna

442   Region (Northern Italy) from four-year observations. Sci. Total Environ., 553, 172-183,

443   doi:10.1016/j.scitotenv.2016.02.074.

444 Santhi, C; Arnold, J. G.; Williams, J. R.; Dugas, W. A; Srinivasan, R.; and Hauck, L. M. Validation

445   of the swat model on a large river basin with point and non-point sources, JAWRA. J. Am. Water

446   Resour. Assoc., 2001, 37, 1169-1188.

447 Schapire, R. (1990). The strength of weak learnability. Machine Learning, 5 (2), 197-227.

448 Freund, Y., Schpire, R (1997). A decision-theoretic generalisation of on-line learning and an

449   application of boosting. J. Computer and System Sciences, 55 (1), 119-139.

450   Soni, M., Payra, S., Verma, S., 2018. Particulate matter estimation over a semi-arid region Jaipur,

451   India using satellite AOD and meteorological parameters. Atmospheric Pollution Research 9 (5),

452   949-958.

453 Stajkowski, S; Kumar, D; Samui, P; Bonakdari, H; and Gharabaghi, B, Genetic algorithm-optimized

454   sequential model for water temperature prediction, Sustainability, 12, 13, 5374, 2020.

455 Rumelhart, D. E., G. E. Hinton, and R. J. Williams, "learning representations by back-propagating

456   errors, "Nature, Vol. 323, no.6088, pp. 533-536, 1986.

457 Taylor, K.E., Summarizing multiple aspects of model performance in a single diagram, J. Geophys.

458   Res., 106, 7183-7192, 2001.

459   Van Donkelaar, A., Martin, R.V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., Villeneuve, P.J.,

460   2010. Global estimates of ambient fine particulate matter concentrations from satellite-based

461   optical depth: development and application. Environ. Health Perspect. 118 (6), 847-855.

462   Van Liew, M. W; Arnold, J. G.; Garbrecht, J. D. Hydrologic simulation on agricultural watersheds:

463   Choosing between two models. Trans. ASAE 2003, 56, 1539.

464   Vutukuru, S., Dabdub, D., 2008. Modeling the effects of ship emissions on coastal air quality: a case

465   study of Southern California. Atmos. Environ. 42, 3751-3764.

466   Wang, J., Christopher, S.A., 2003. Intercomparison between satellite-derived aerosol optical

467   thickness and PM2.5 mass: Implications for air quality studies. Geophys. Res. Lett. 30 (21).

468   Wei, J., Huang, W., Li,, Z., Xue, W., Peng, Y., Sun, L., Gribb, M., 2019. Estimating 1-km resolution

469   PM2.5 concentrations across China using space-time random forest approach. Rem. Sens.

470   Environ. 231, 111221.

471   World Health Organization, media centre (2016). Air pollution levels are rising in many of the

472   world's poorest cities: http://www.int/mediacentre/news/releases/2016/air-pollution-raising/.

473   Wu, X.; Kumar, V.; Quinlan, J. R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G. J.; Ng, A.;

474   Yi. L., Mengfan, T., Kun, Y., Yu, Z., Xiaolu, Z., Miao, Z., Yan, S., 2019. Research on PM2.5

475   estimation and prediction method and changing characteristics analysis under long temporal and

476   large spatial scale – a case study in China typical regions. Sci. Total Enviro. 696, 133983,

477   https://doi.org/10.1016/j/scitotenv.2019.133983.

478   Zhang, Y., Cao,F., 2015. Fine particle matter (PM2.5)in China at a city level. Sci. Rep., 5, 14884.

479

PM2.5 Locations
Locations: 1691
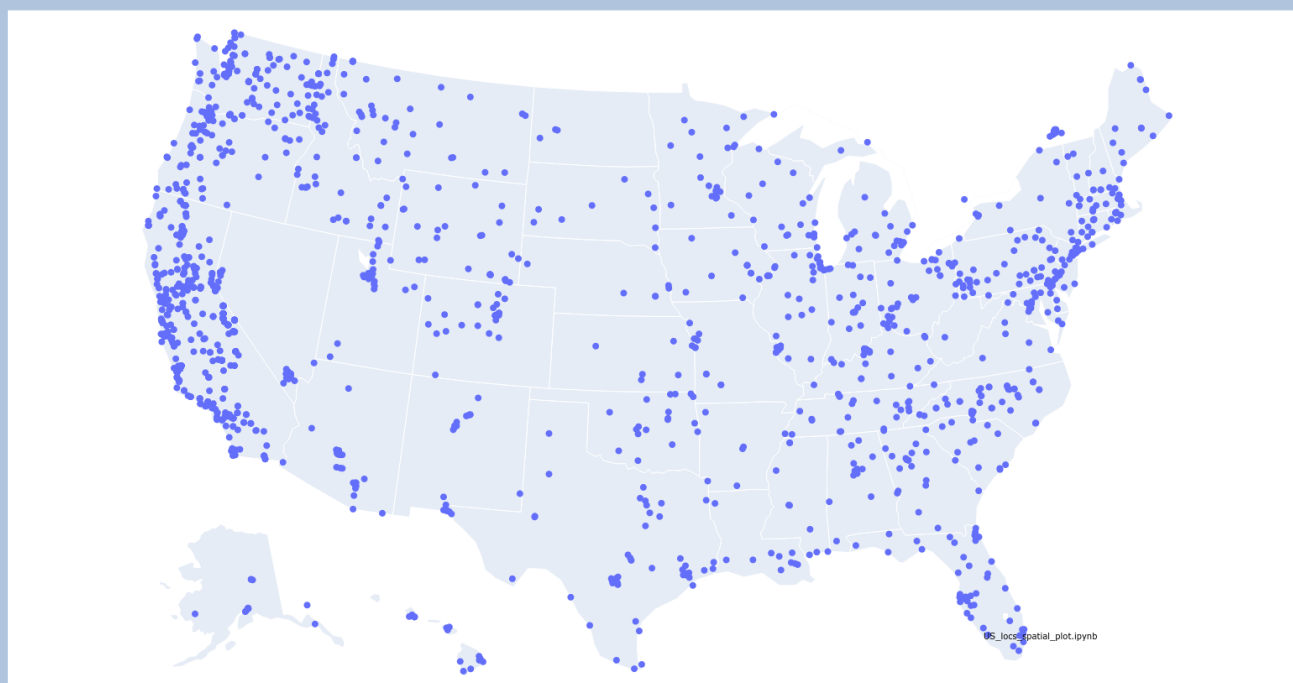
US_locs_spatial_plot.ipynb
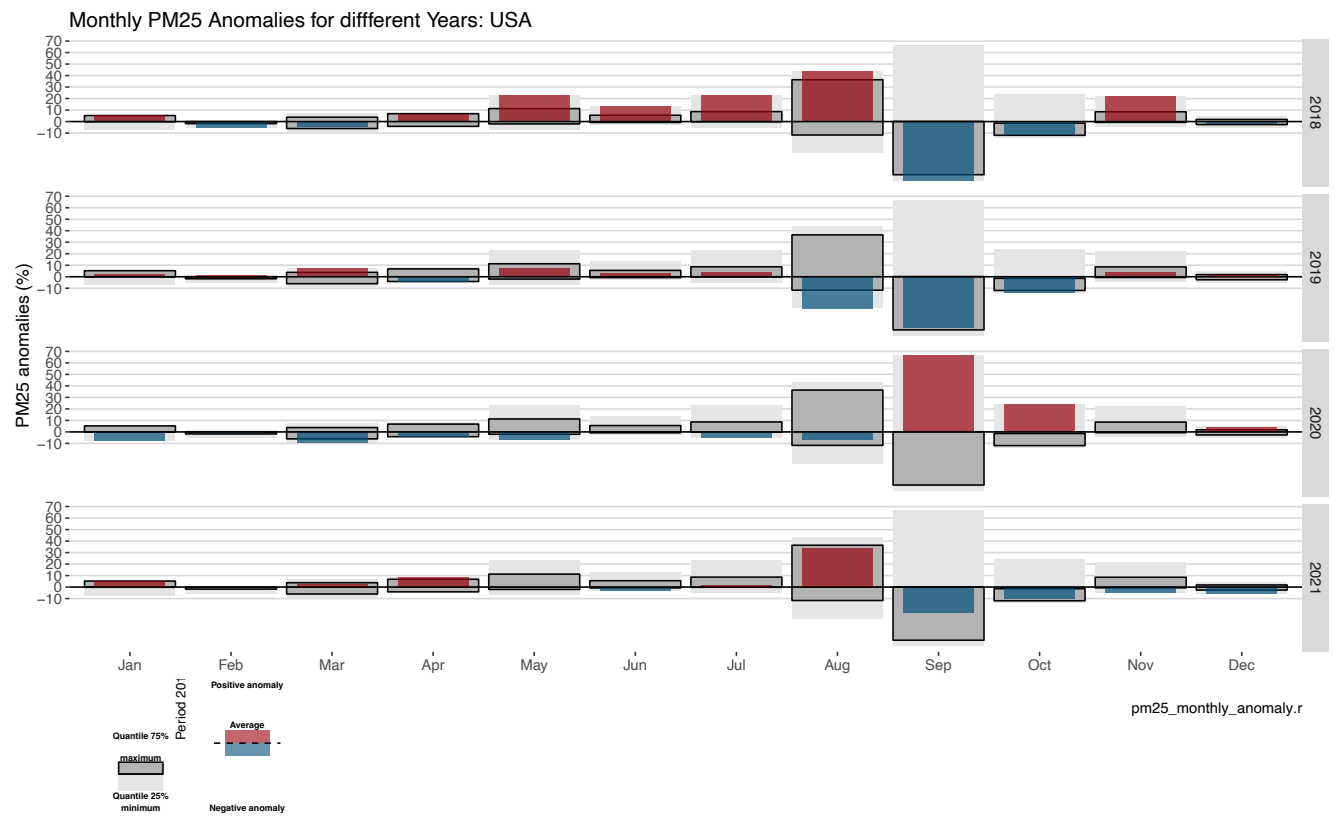
480

481 **Figure 1**. Locations of PM$_{2.5}$ monitoring sites over USA

Figure 2. Monthly anomalies and quantiles for the observed period (2018-2021) using daily PM$_{2.5}$ values over United States.
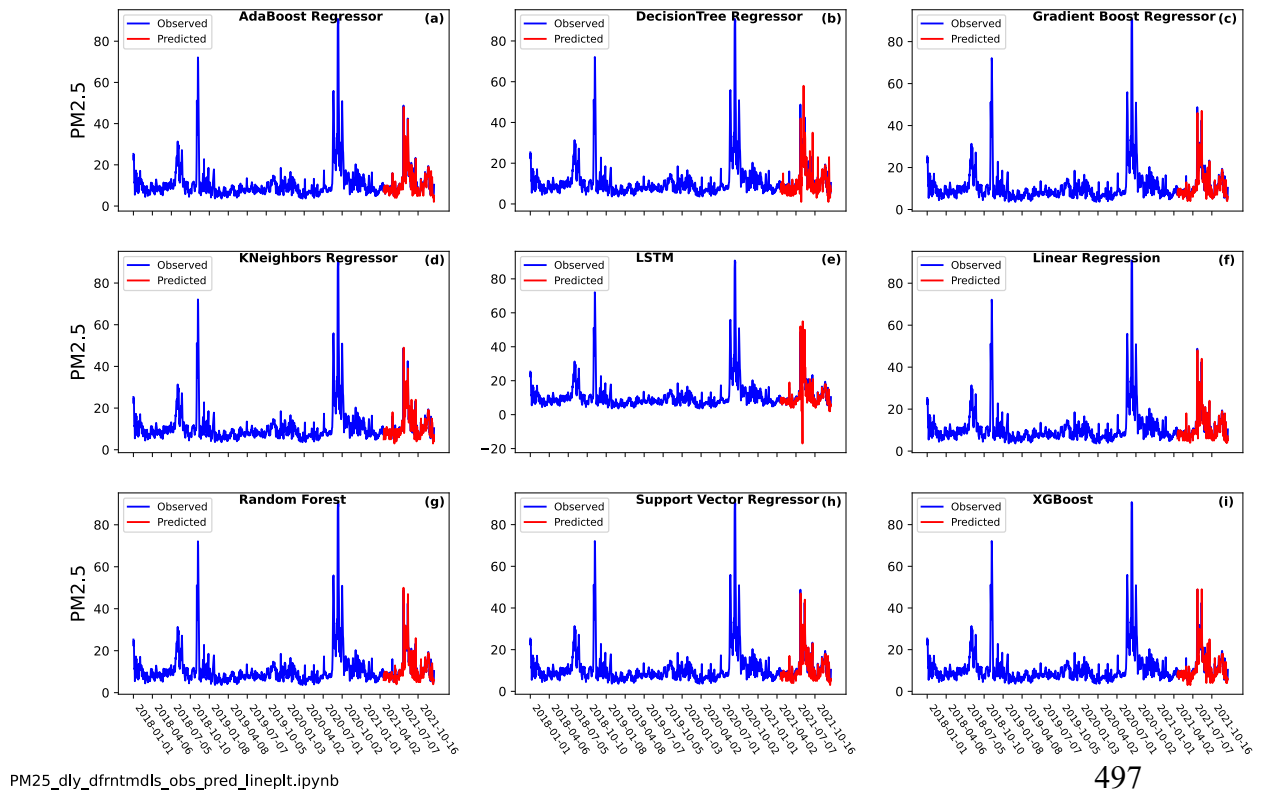
486

487

497

Figure 3. The comparison of the time series of estimated and observed PM$_{2.5}$ concentrations over California using different machine learning models: (a) AdaBoost regressor, (b) Decision Tree regression, (c) Gradient Boost regression, (d) K-neighbors regression (e) LSTM, (f) Linear regression, (g) Random Forest, (h) Support Vector regression, and (I) XGBoost.
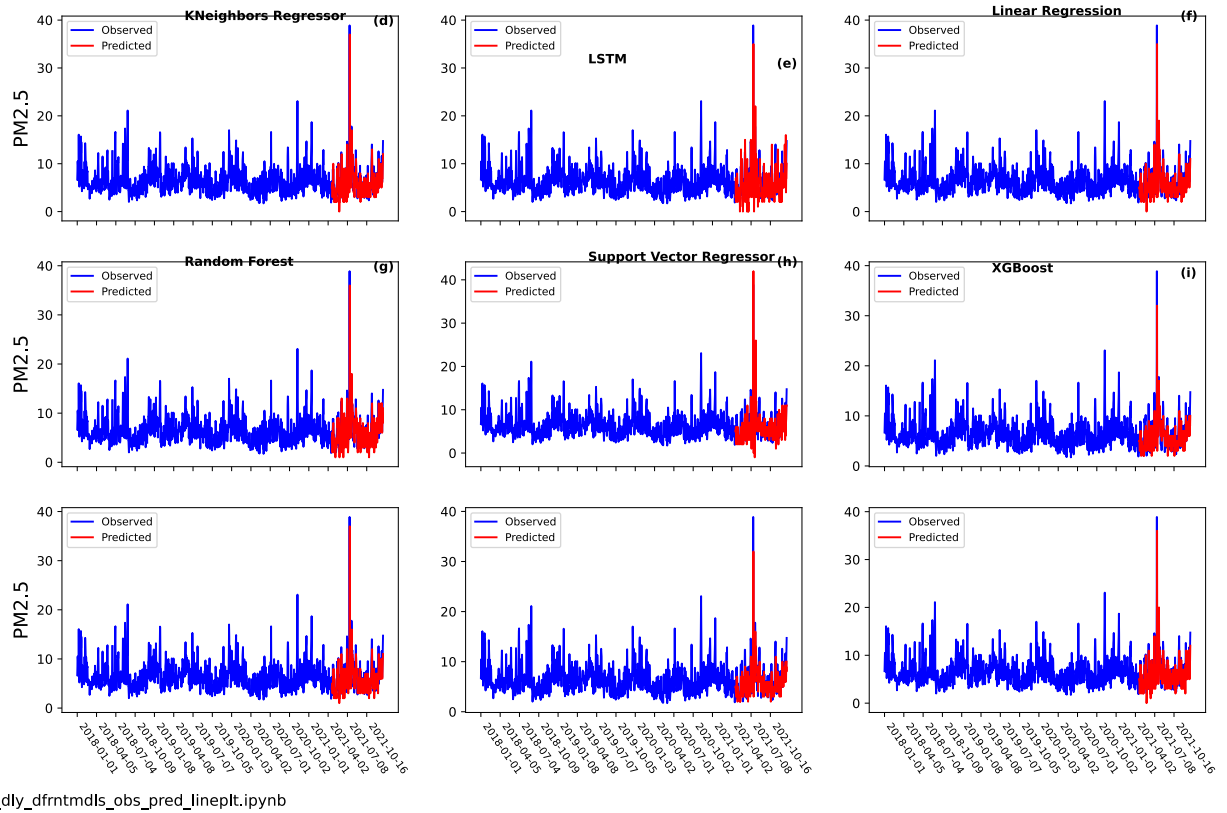
PM25_dly_dfrntmdls_obs_pred_lineplt.ipynb

Figure 4. The comparison of the time series of estimated and observed PM$_{2.5}$ concentrations over New York using different machine learning models: (a) AdaBoost regressor, (b) DecisionTree regression, (c) Gradient Boost regression, (d) Kneighbors regression (e) LSTM, (f) Linear regression, (g) Random Forest, (h) Support Vector regression, and (I) XGBoost.
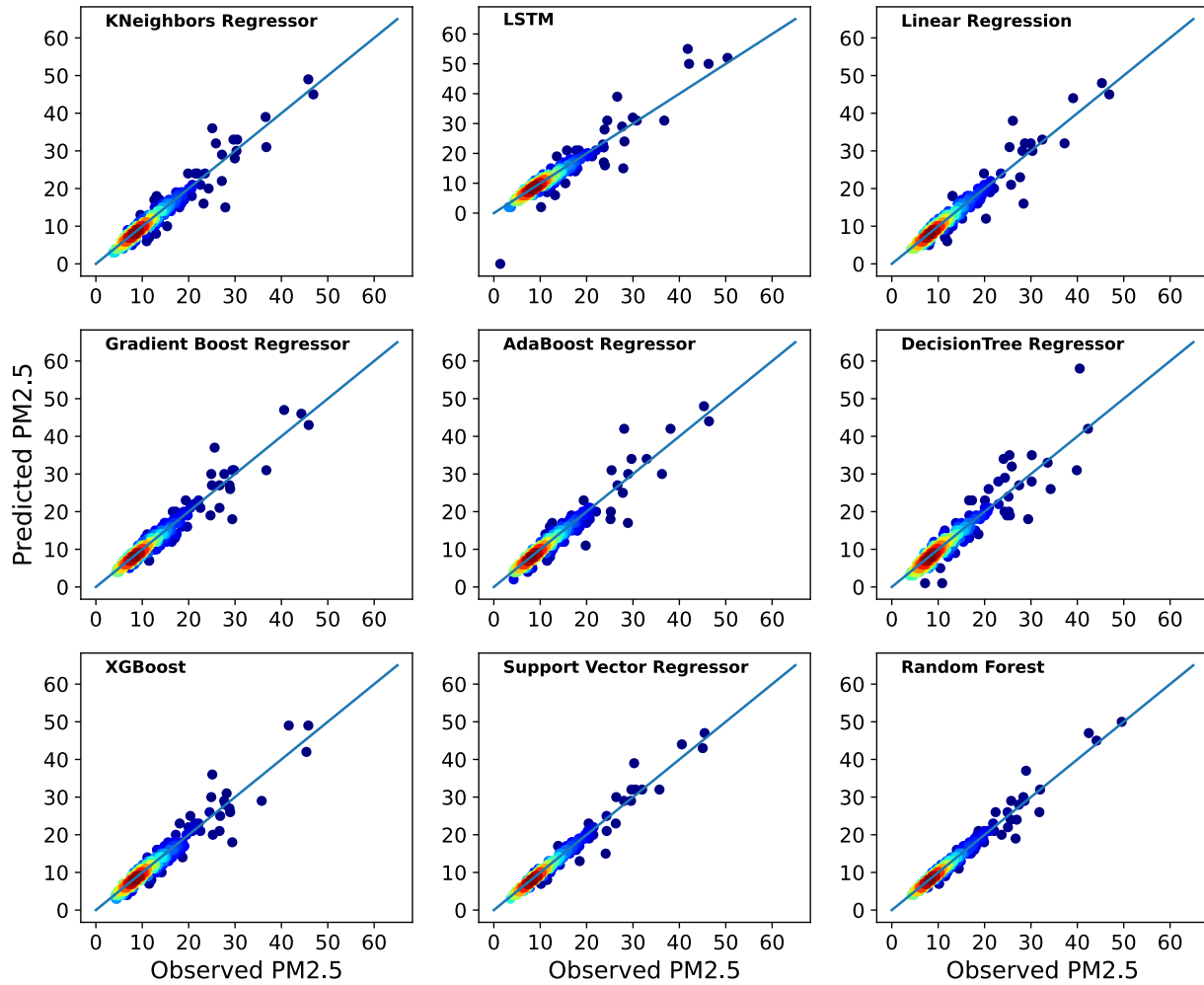
Figure 5. Scatter plots of observed and estimated daily PM$_{2.5}$ concentrations over California using different machine learning models: (a) AdaBoost regressor, (b)DecisionTree regression, (c) Gradient Boost regression, (d) Kneighbors regression (e) LSTM, (f) Linear regression, (g) Random Forest, (h) Support Vector regression, and (I) XGBoost.

Figure 6. Scatter plots of observed and estimated daily PM$_{2.5}$ concentrations over New York using different machine learning models: (a) AdaBoost regressor, (b)DecisionTree regression, (c) Gradient Boost regression, (d) Kneighbors regression (e) LSTM, (f) Linear regression, (g) Random Forest, (h) Support Vector regression, and (I) XGBoost.
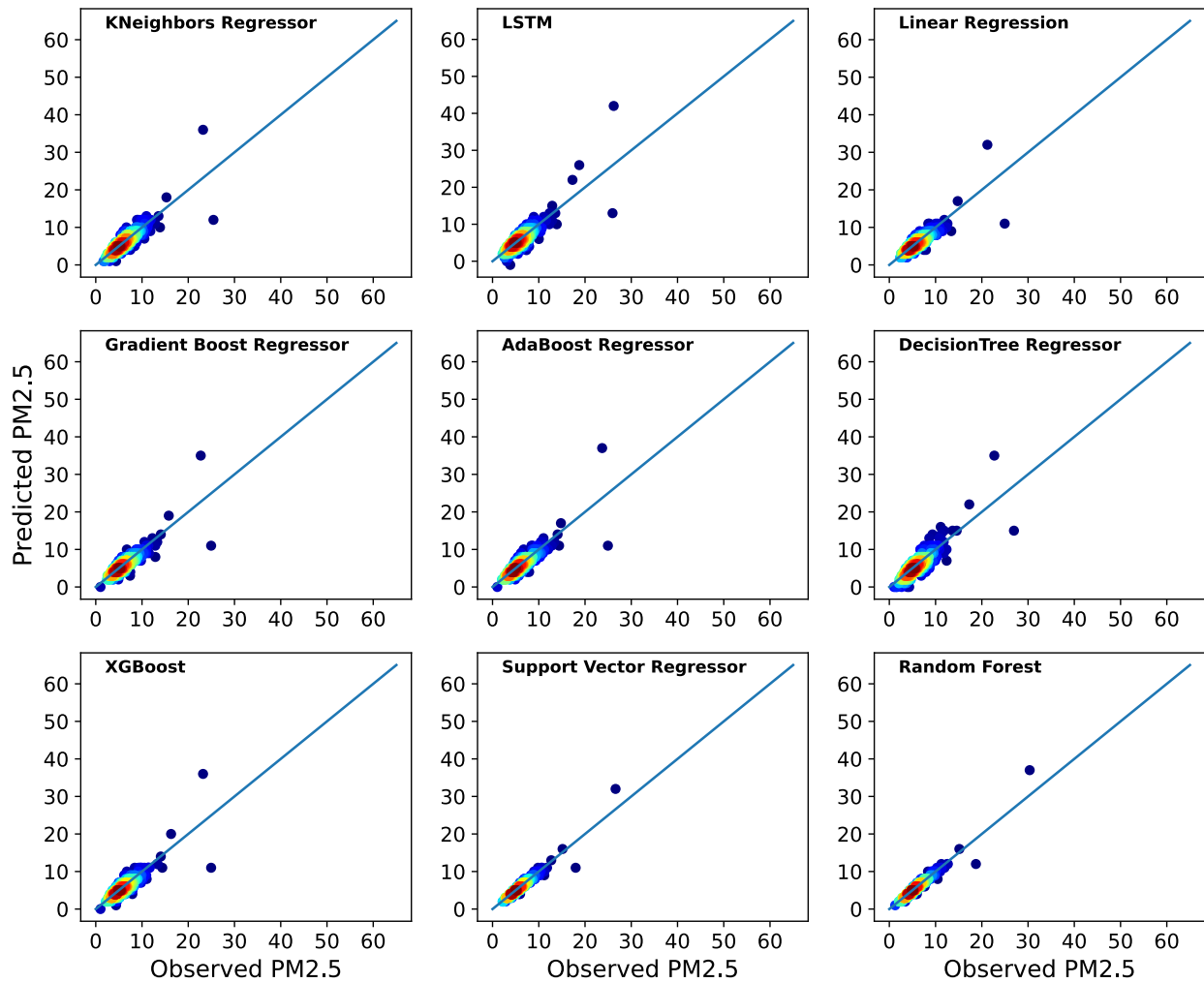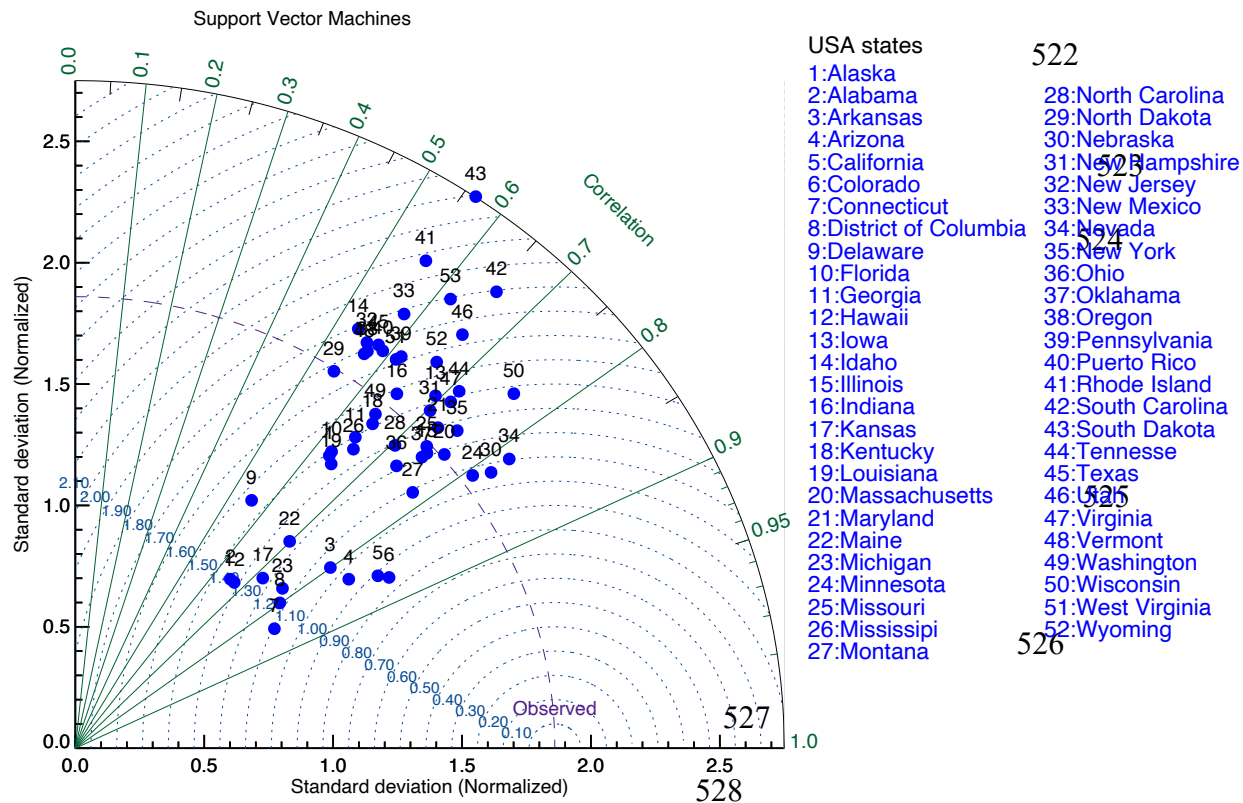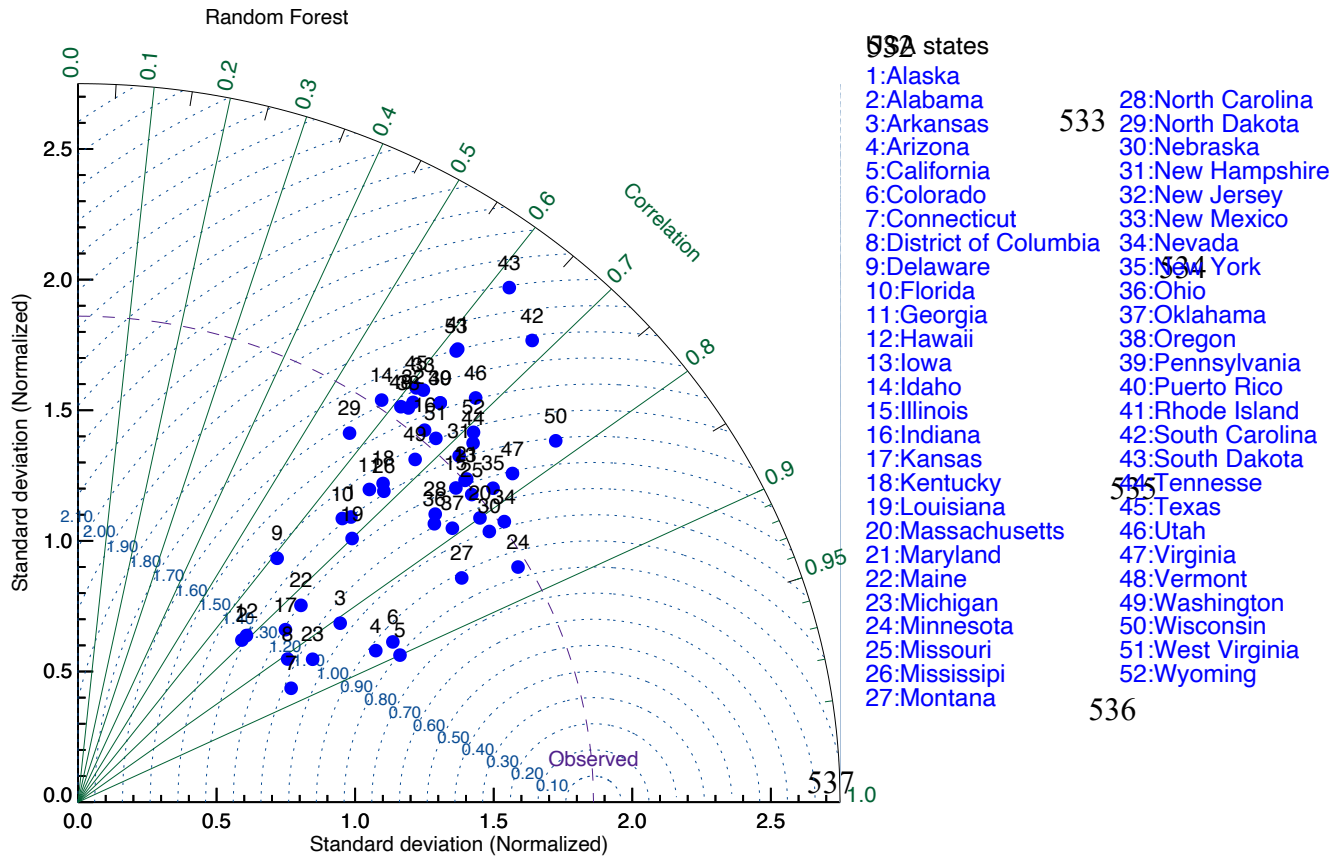
Figure 7. Taylor diagram of the Support Vector Machines (SVM) over each state of the United States.

Figure 8. Taylor diagram of the Random Forest (RF) over each state of the United States.

Table 1: Different Model Metrics for New York State

New York								
Model	**RMSE**	**MAE**	**MAPE**	**R2**	**NSE**	**NORM**	**PBIAS**	**RSR**
Linear Regression	3.883	2.309	0.285	0.688	0.613	60.156	11.24	0.561
Decision Tree	5.136	3.109	0.254	0.454	0.533	79.58	13.44	0.691
Gradient Boost Regressor	3.822	2.394	0.545	0.698	0.683	59.207	8.210	0.546
AdaBoost Regressor	3.961	2.316	0.188	0.676	0.683	61.369	9.653	0.576
XG Boost	3.898	2.501	0.202	0.686	0.681	60.393	8.342	0.559
KNeighbors Regressor	3.919	2.379	0.195	0.683	0.677	60.711	7.515	0.562
LSTM	7.487	3.359	0.218	0.158	0.455	115.991	6.020	0.812
Random Forest	3.121	2.122	0.182	0.899	0.811	38.671	2.989	0.331
SVM	3.125	2.145	0.183	0.857	0.820	39.161	3.011	0.338

RMSE = Root mean squared error

MAE = Mean absolute error

545   MAPE = Mean absolute percentage error
546   R^2 = The coefficient of determination
547   NSE = Nash-Sutcliffe efficiency
548   PBIAS = Percent Bias
549   RSR = root mean square error ratio
550
551

552   **Table 2:** Different Model Metrics for California State

California								
Model	**RMSE**	**MAE**	**MAPE**	**R^2**	**NSE**	**NORM**	**PBIAS**	**RSR**
Linear Regression	3.695	2.599	0.326	0.43	0.694	57.243	12.086	0.932
Decision Tree	5.481	3.743	0.467	0.23	0.576	84.917	19.901	0.732
Gradient Boost Regressor	4.051	2.736	0.340	0.28	0.461	62.758	16.891	1.017
AdaBoost Regressor	3.804	2.636	0.342	0.33	0.435	58.938	17.532	0.969
XG Boost	4.271	2.972	0.372	0.17	0.438	66.178	18.726	1.075
KNeighbors Regressor	4.394	3.062	0.392	0.22	0.286	68.071	17.076	1.106
LSTM	5.025	3.252	0.339	0.46	0.309	77.853	18.027	0.618
Random Forest	3.051	2.233	0.315	0.77	0.817	46.894	7.022	0.355
SVM	3.714	2.618	0.320	0.71	0.897	47.853	7.027	0.424

553
554   RMSE = Root mean squared error
555   MAE = Mean absolute error
556   MAPE = Mean absolute percentage error
557   R^2 = The coefficient of determination
558   NSE = Nash-Sutcliffe efficiency
559   PBIAS = Percent Bias
560   RSR = root mean square error ratio
561