Low-Cost Air Quality Sensor Evaluation and Calibration in Contrasting Aerosol Environments

Pawan Gupta[1,2], Prakash Doraiswamy[3], Jashwanth Reddy[4,1], Palak Balyan[5,6], Sagnik Dey[5], Ryan Chartier[3], Adeel Khan[7], Karmann Riter[3], Brandon Feenstra[8], Robert C. Levy[9], Nhu Nguyen Minh Tran[4], Olga Pikelnaya[8], Kurinji Selvaraj[8], Tanushree Ganguly[7], Karthik Ganesan[7]

[1]STI-Universities Space Research Associations, Huntsville, AL, USA.
[2] NASA Marshall Space Flight Center, Huntsville, AL, USA
[3]RTI International, Research Triangle Park, NC, USA.
[4]The University of Alabama in Huntsville, AL, USA.
[5]Indian Institute of Technology, New Delhi, India
[6]Health Effects Institute (HEI), Boston, U.S.A.
[7]Council on Energy, Environment, and Water (CEEW), New Delhi, India
[8]South Coast Air Quality Management District, CA, USA.
[9]NASA Goddard Space Flight Center, Huntsville, AL, USA.

Correspondence to: Pawan Gupta (pawan.gupta@nasa.gov)

Supplementary Material

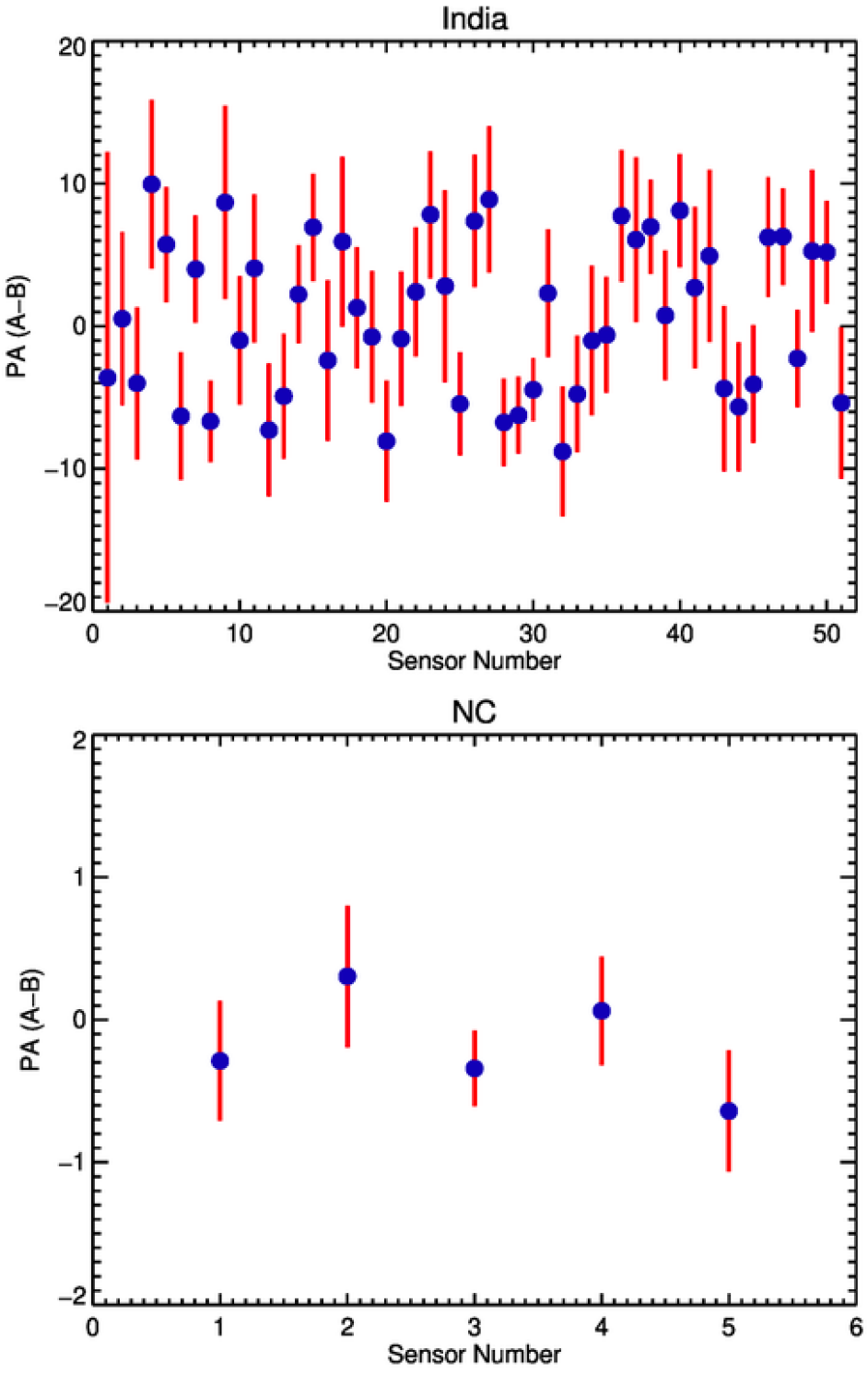

Figure S1. The plots showing mean difference in A & B measurements from each sensor for the two regions.

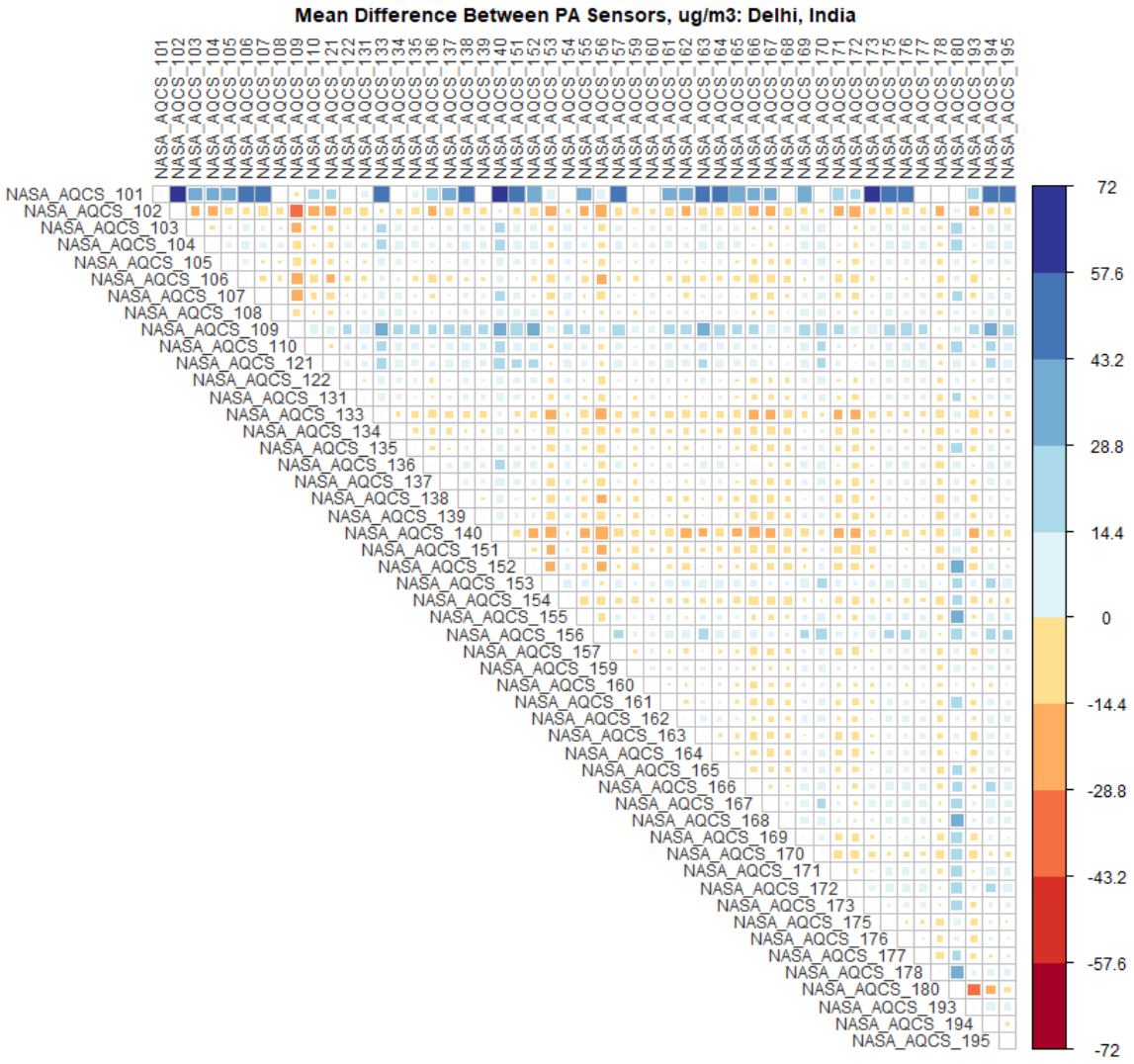

Figure S2.1. Mean Absolute Difference (µg/m^3) Between Any Two Sensors in Delhi

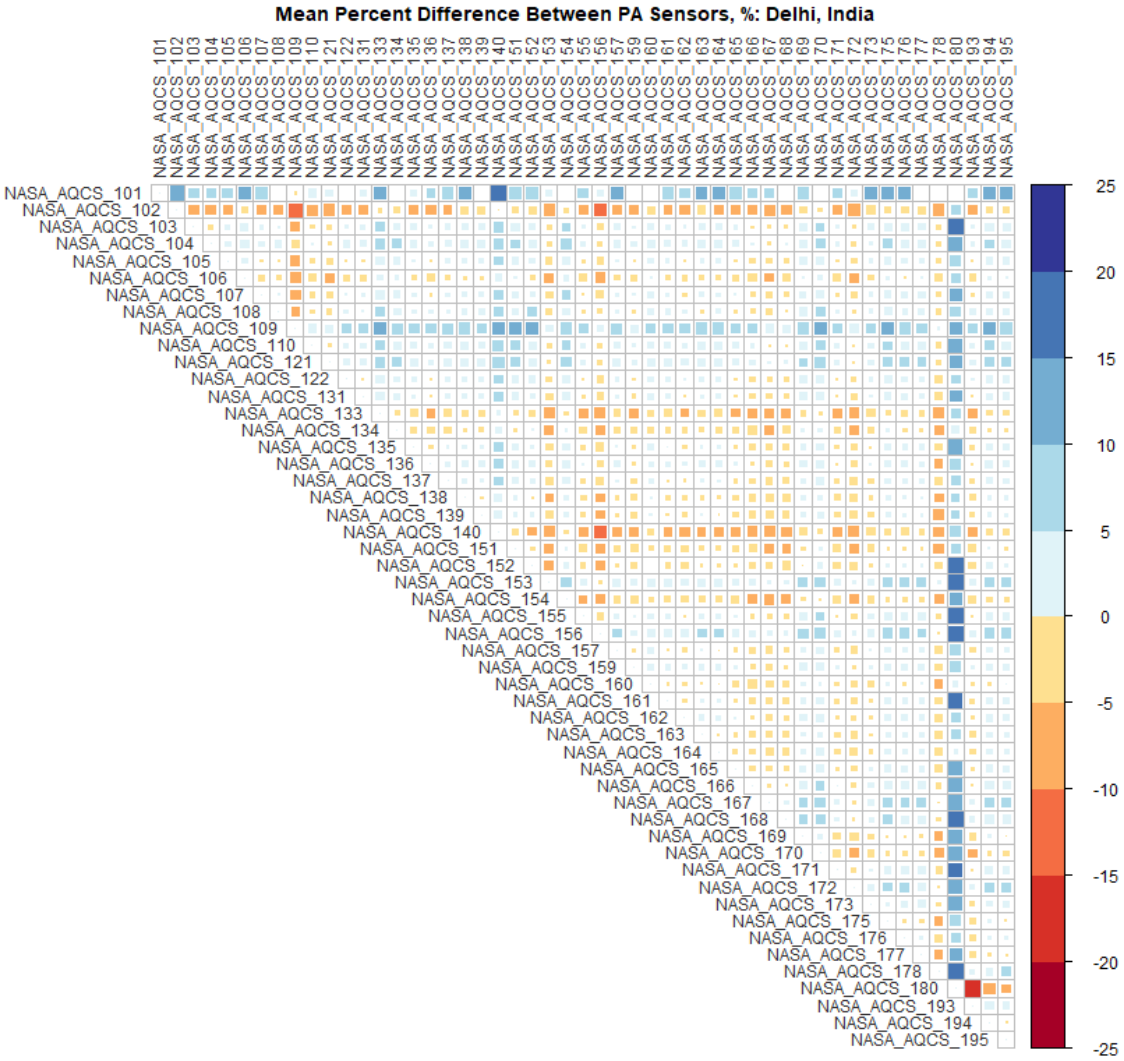

Figure S2.2. Mean Percent (%) Difference Between Any Two Sensors in Delhi

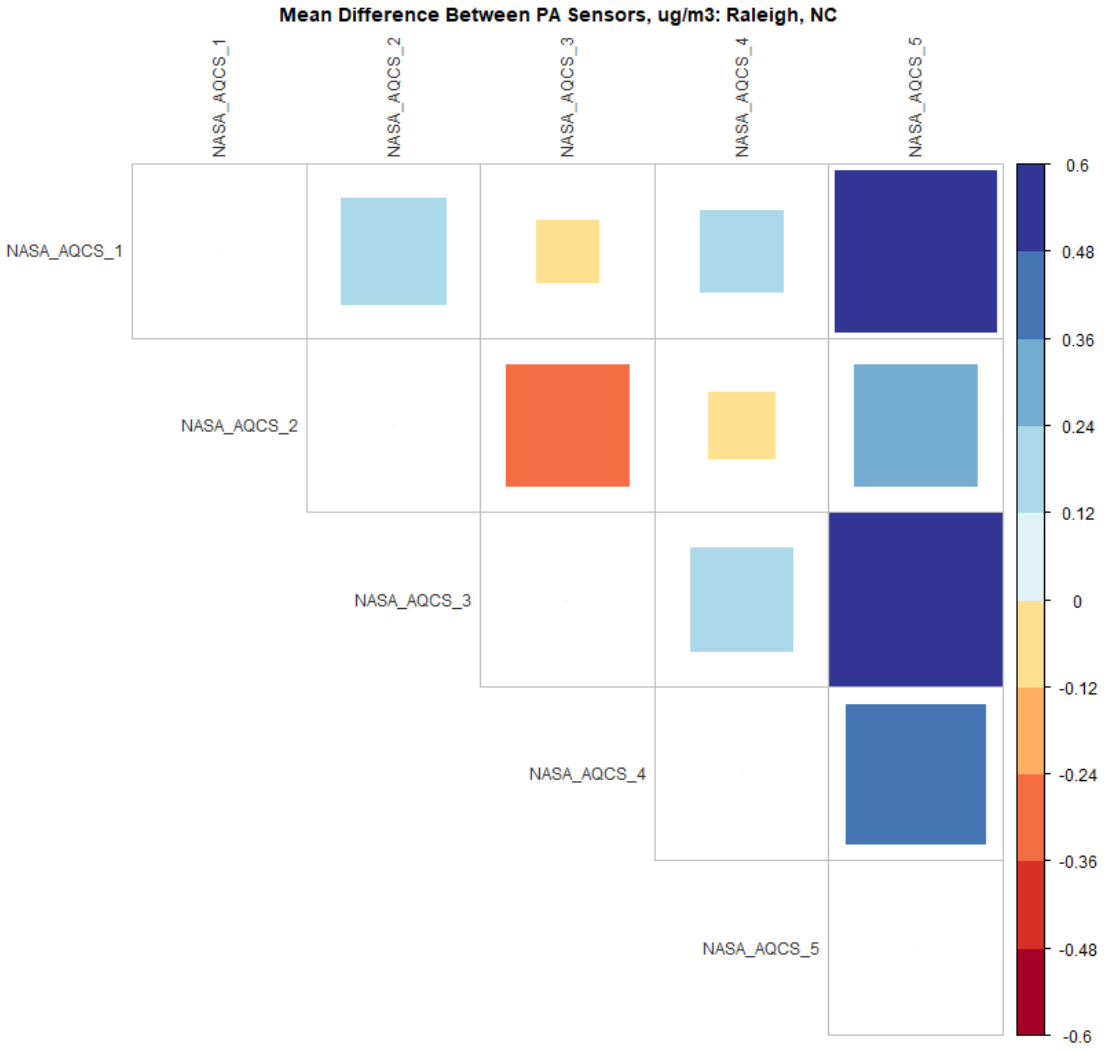

Figure S2.3. Mean Absolute Difference (µg/m^3) Between Any Two Sensors in Raleigh

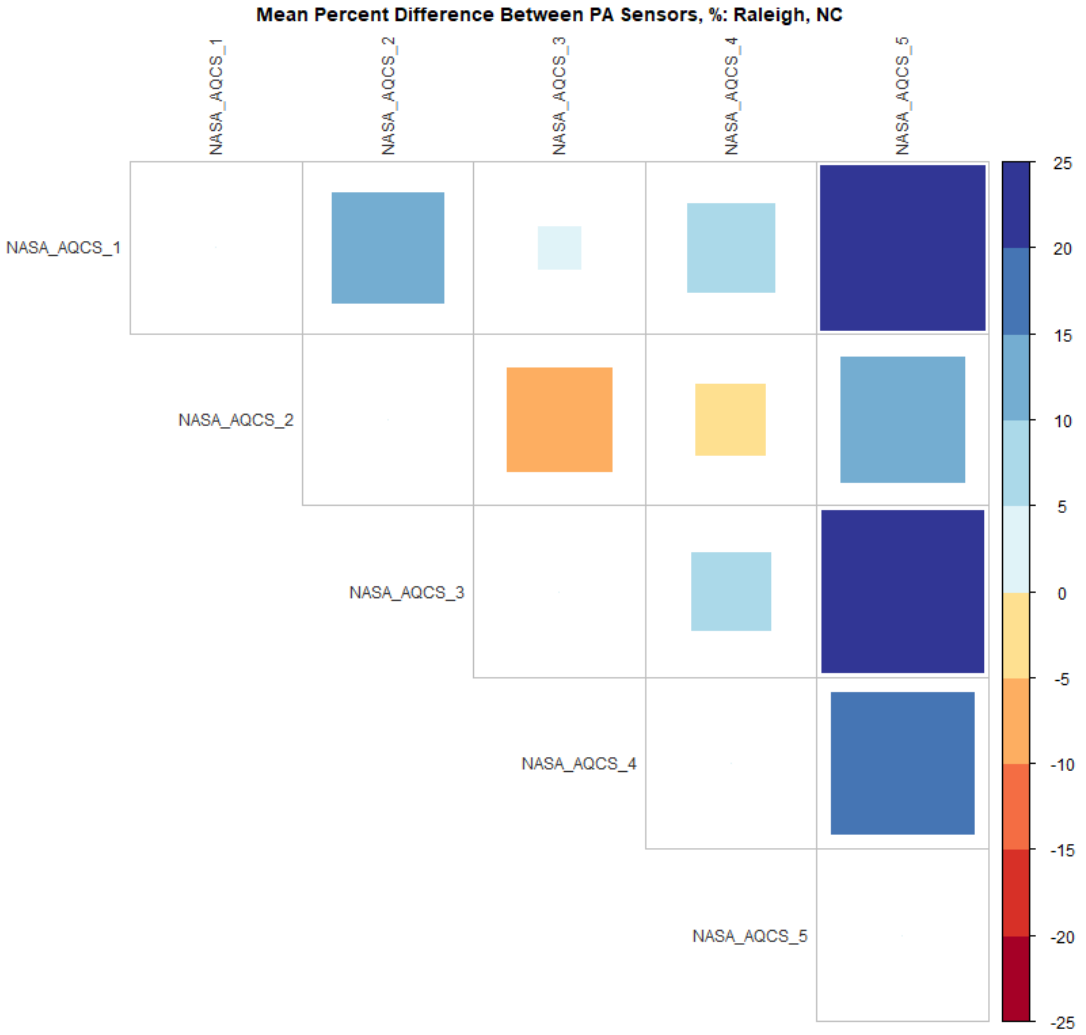

Figure S2.4. Mean Percent (%) Difference Between Any Two Sensors in Raleigh

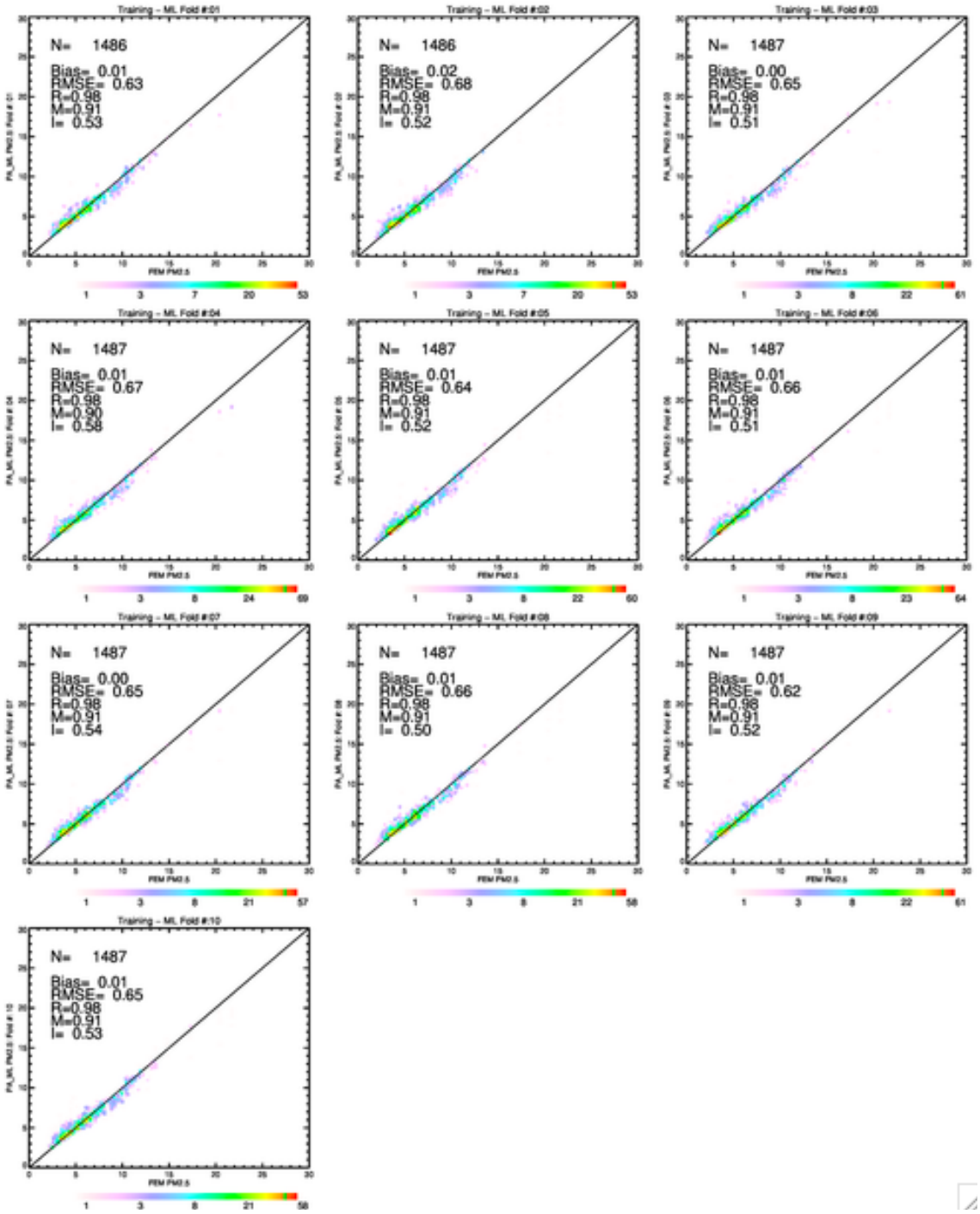

Figure S3.1 – Inter-comparison between FEM and ML output for Raleigh during 10-fold training of MLA. Each density scatter plot represents 1-fold.

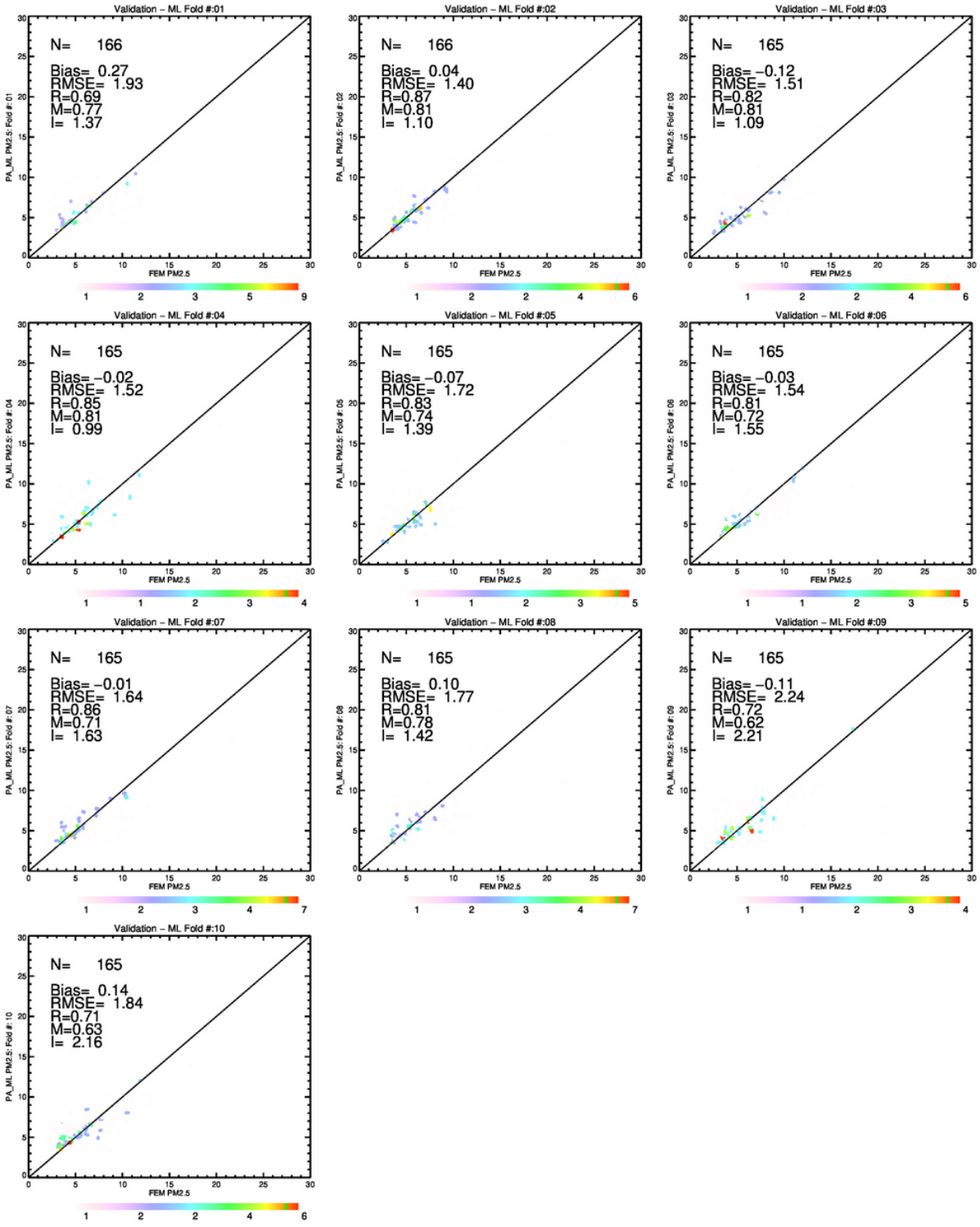

Figure S3.2 – same as Figure S3.1 but for 10-fold validation in Raleigh.

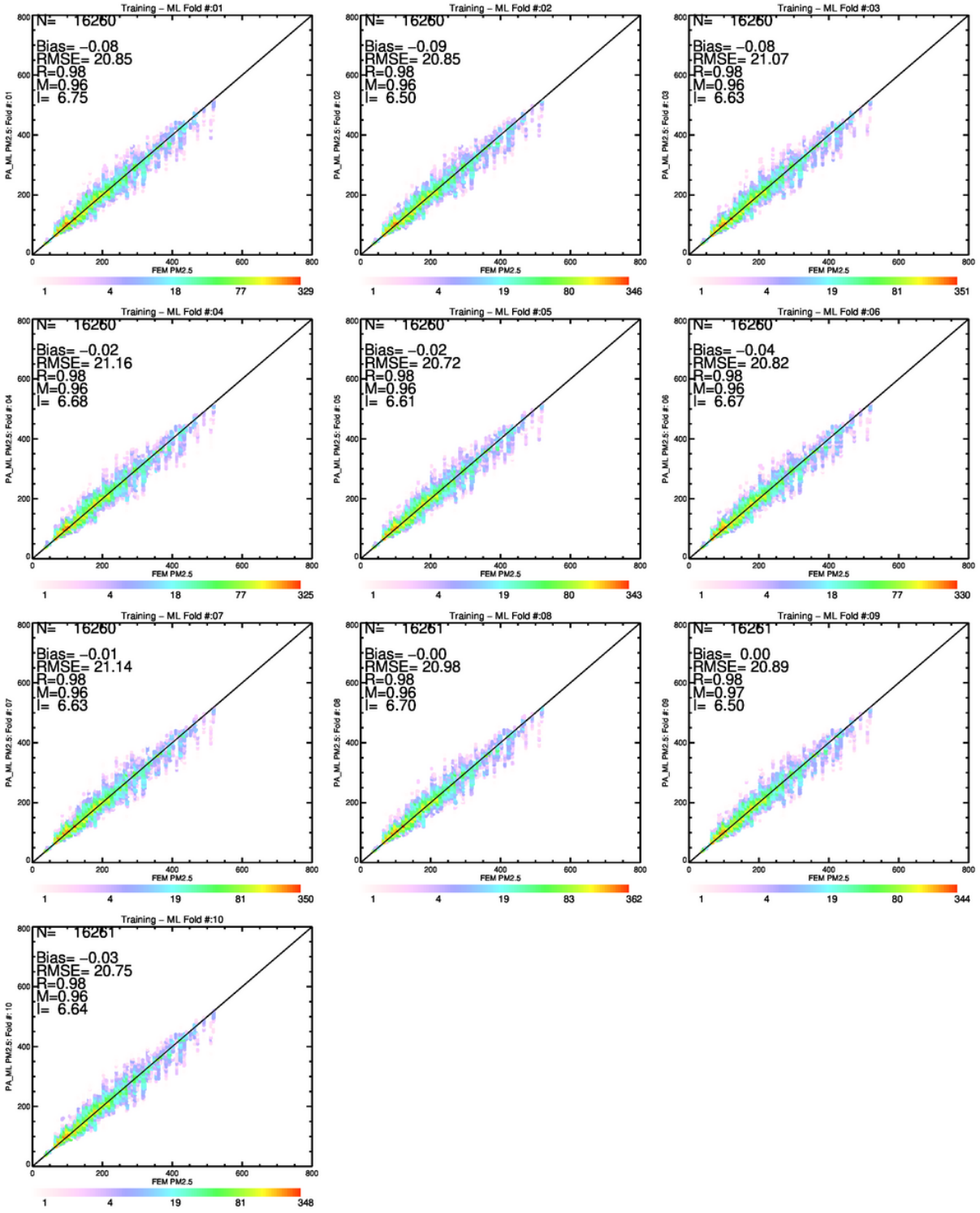

Figure S3.3 – same as Figure S3.1 but for 10-fold training for Delhi.

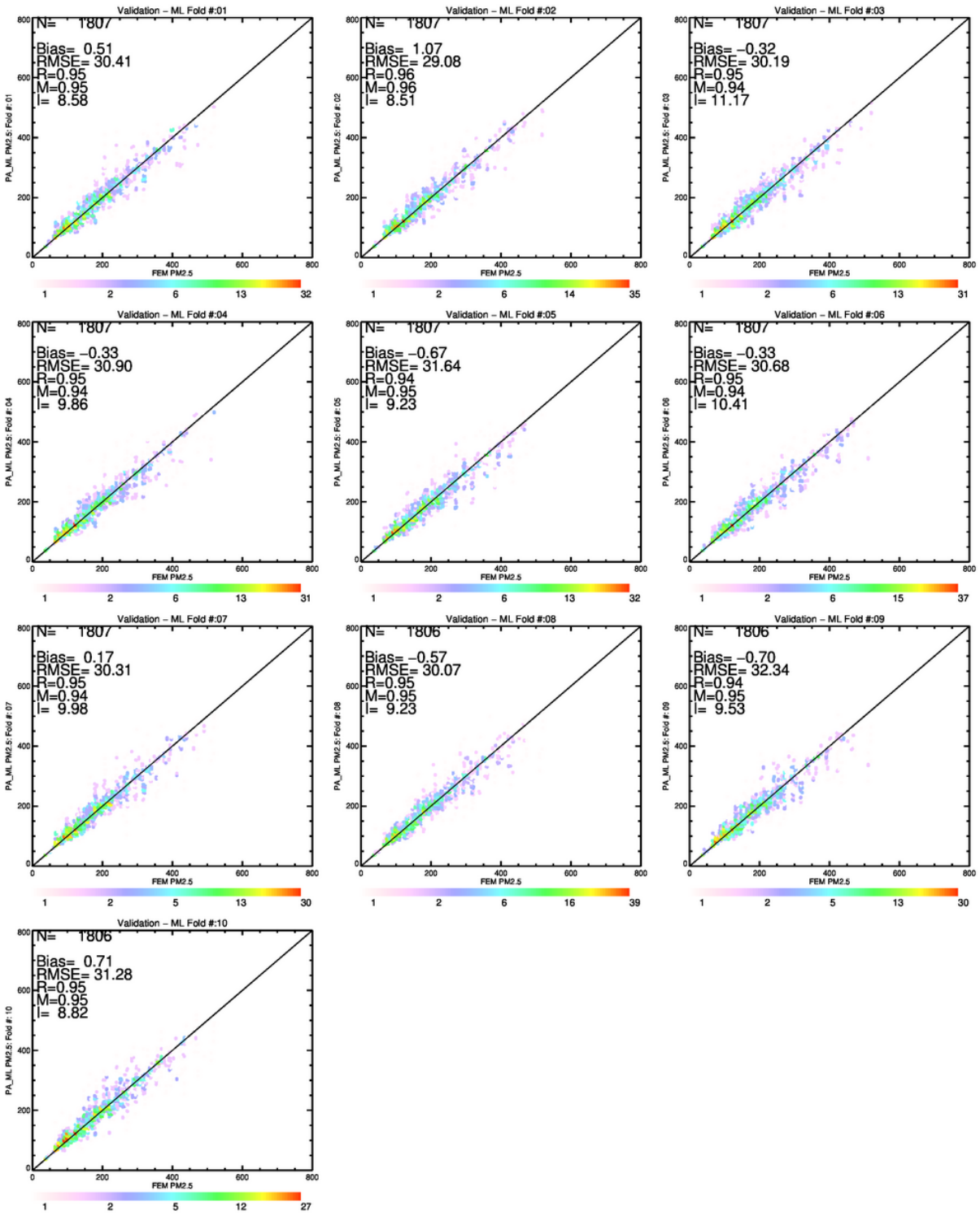

Figure S3.4 – same as Figure S3.1 but for 10-fold validation for Delhi.

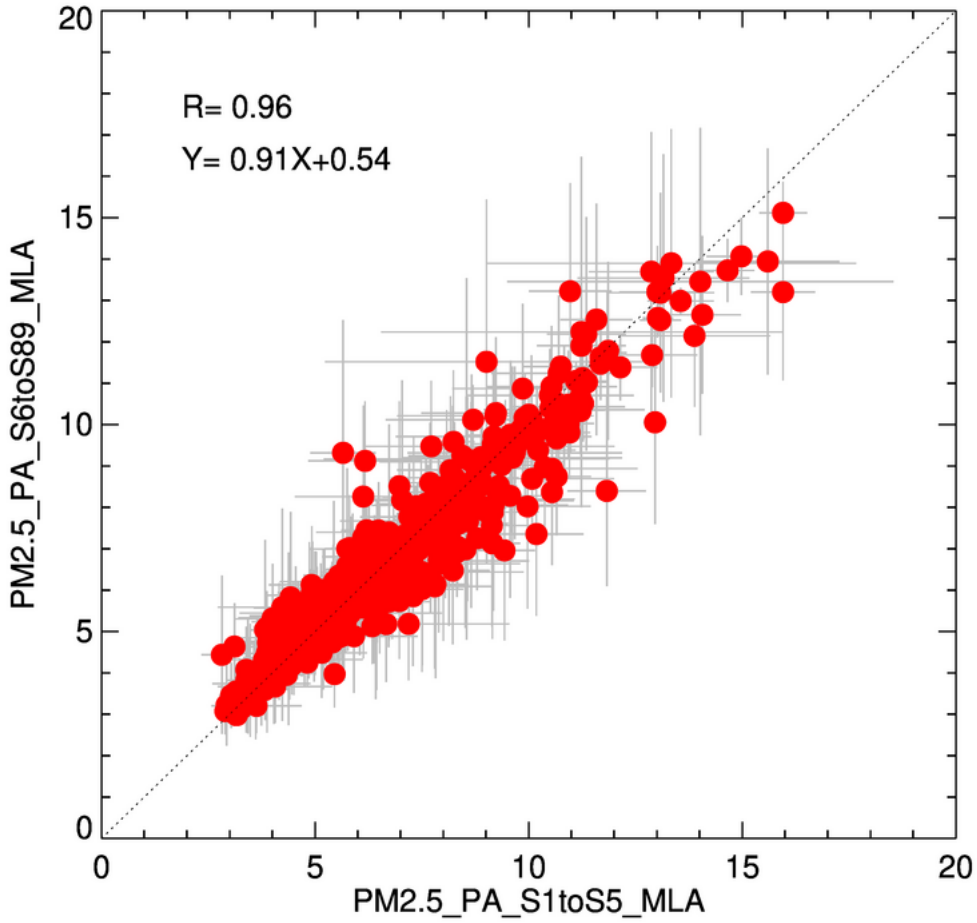

Figure S4. ADD THIS FIGURE on Raleigh ML comparison for 84 vs. 5 sensors.

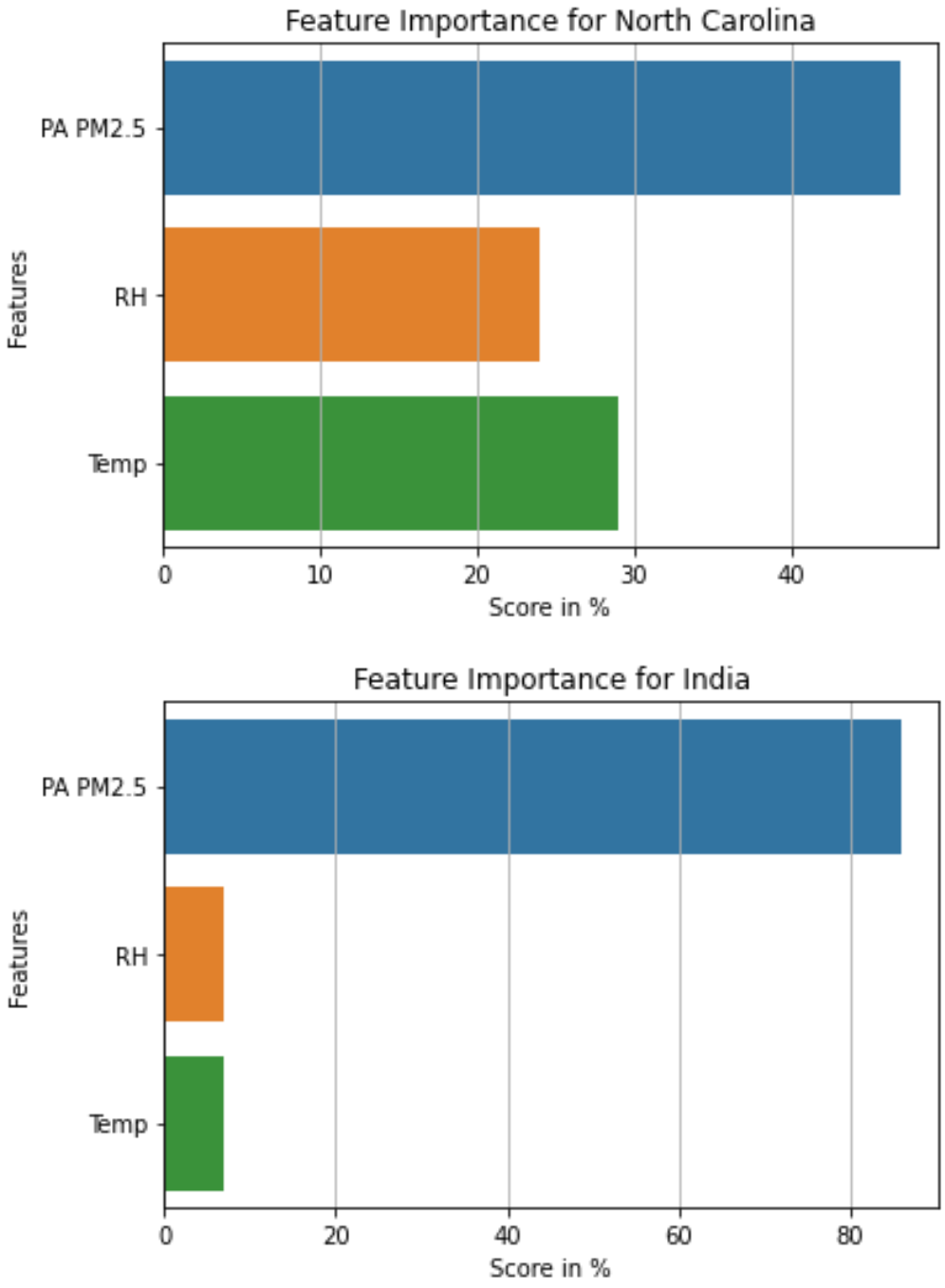

Figure S5. The input feature importance for RF model for Raleigh (top) and Delhi (bottom)

```python
# sample code used to train RF model

# Import packages
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor

# read already cleaned and quality controllled Input parameters
df = pd.read_csv('Extracted_Data_latest.csv')
X = df[['PAB_Pm25','Humidity','Temperature']]
Y = df['BAM_Pm25']

# call the function to randomly select training & testing data
# splits 75% data to train and 25% to test, test_size = 0.25
X_train, X_test, y_train, y_test = train_test_split(X, Y, random_state=0)

# call the function to train the model
regr = RandomForestRegressor(max_depth=20, random_state=2,n_estimators=50,bootstrap=True)
regr.fit(X_train, y_train)

# run trained model on training and testing data
pred_trn = regr.predict(X_train)
pred_tst = regr.predict(X_test)

# save the trained model in a file
filename = 'model_for_PAB.sav'
pickle.dump(regr, open(filename, 'wb'))
```

Figure S6. The sample python code to train and test a random forest model.

Table S1. The mean bias (MB), mean percentage bias (MB%), and mean percentage absolute bias (|MB%| for hourly and daily averages. The biases before and after corrections are provided.

Delhi	**Hourly**		**MB**	**MB%**	**|MB%|**
		Raw Data	**35.1**	**23.8**	**28.9**
		Corrected Data	**0.22**	**2.0**	**9.1**
	Daily	**Raw Data**	**37.3**	**23.7**	**25.3**
		Corrected Data	**1.3**	**1.8**	**5.4**
Raleigh	**Hourly**	**Raw Data**	**-0.8**	**-4.9**	**604**
		Corrected Data	**0.02**	**4.1**	**10.9**
	Daily	**Raw Data**	**-0.8**	**-11.2**	**27.7**
		Corrected Data	**0.03**	**2.2**	**5.0**