Low-Cost Sensor Performance Intercomparison, Correction Factor Development, and 2+ Years of Ambient $PM_{2.5}$ monitoring in Accra, Ghana

SUPPORTING INFORMATION FOR PUBLICATION

Authors: *Garima Raheja* * [1,2], *James Nimo*[3,4], *Emmanuel K.-E. Appoh*[5] ,*Benjamin Essien*[5], *Maxwell Sunu*[5], *John Nyante*[5], *Mawuli Amegah*[5], *Reginald Quansah*[6], *Raphael E Arku*[7], *Stefani L. Penn*[8], *Michael R. Giordano*[9], *Zhonghua Zheng*[10], *Darby Jack*[11], *Steven Chillrud*[11], *Kofi Amegah*[12], *R Subramanian*[9,13,14], *Robert Pinder*[15], *Ebenezer Appah-Sampong*[5], *Esi Nerquaye Tetteh*[5], *Mathias A. Borketey*[5], *Allison Felix Hughes*[3], *Daniel M. Westervelt* * [2,16]

Affiliations:

1. Department of Earth and Environmental Sciences, Columbia University, New York, NY
2. Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY
3. Department of Physics, University of Ghana, Legon , Ghana
4. African Institute of Mathematical Sciences, Kigali, Rwanda
5. Ghana Environmental Protection Agency, Accra, Ghana
6. School of Public Health, University of Ghana, Accra, Ghana
7. Department of Environmental Health Sciences, School of Public Health and Health Sciences, University of Massachusetts, Amherst, MA
8. Industrial Economics, Inc, Cambridge, MA
9. Univ Paris Est Creteil, CNRS UMS 3563, Ecole Nationale des Ponts et Chaussés, Université de Paris, OSU-EFLUVE – Observatoire Sciences de L'Univers-Envelopes Fluides de La Ville à L'Exobiologie, F-94010 Créteil, France
10. Department of Earth and Environmental Sciences, The University of Manchester, Manchester, UK
11. Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY
12. University of Cape Coast, Cape Coast, Ghana
13. Present Address: Qatar Environment and Energy Research Institute, Hamad Bin Khalifa University, Doha, Qatar
14. Kigali Collaborative Research Centre, Kigali, Rwanda
15. Environmental Protection Agency, Raleigh, North Carolina
16. NASA Goddard Institute for Space Science, New York, NY

SUPPORTING INFORMATION SUMMARY

Number of pages: 11
Figures: S1 - S20
Tables: S1

Figure S1. Timeseries of raw and GMR-corrected $PM_{2.5}$ from 17 Clarity nodes in Accra, Ghana, using the 4 month UG colocation as model training data.

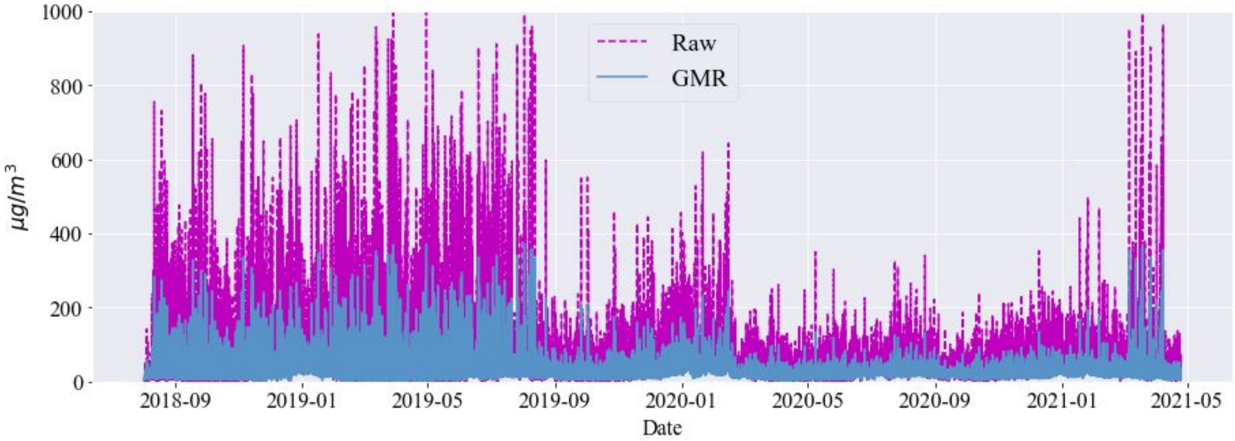


Figure S2. Timeline of of each Clarity node deployed across Accra, with the number of days of valid data indicated next to the node name. Some monitors malfunctioned, interrupting or terminating data collection. The Jamestown monitors were deployed on May 1, 2019.

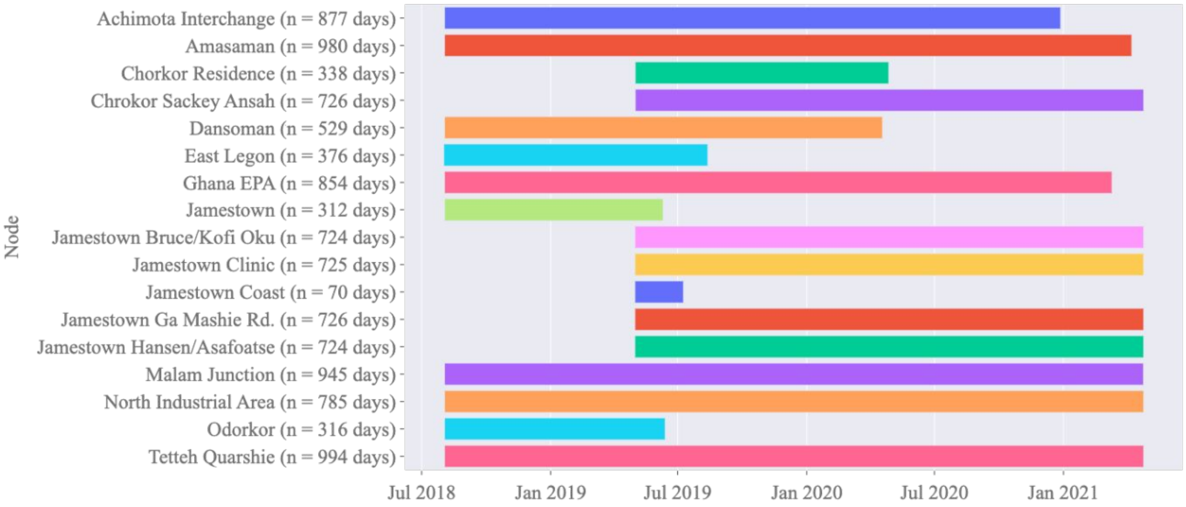


*Figure S3. S*catter plots of hourly LCS $PM_{2.5}$ measurements compared to hourly reference monitor measurements, shaded by temperature

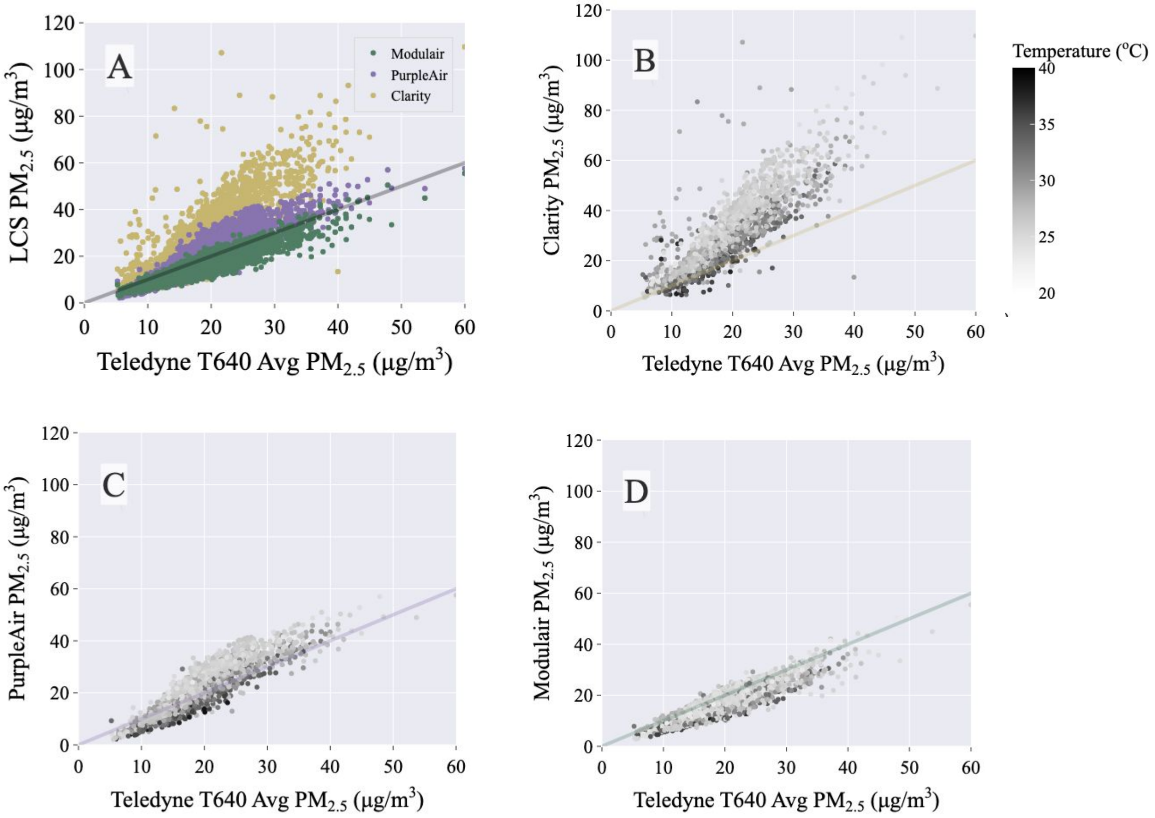
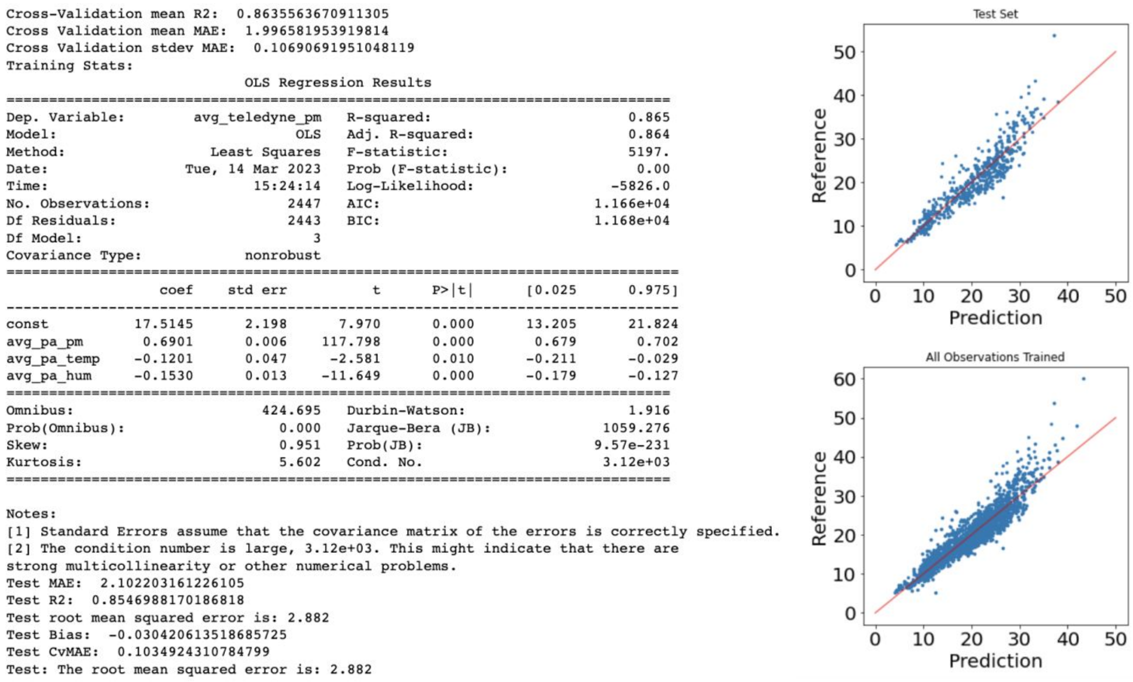


Table S1. Coefficients and Standard Errors for Multiple Linear Regression correction of colocated LCS with reference monitor. To apply MLR to observations, use Equation 1 with respective coefficient values.

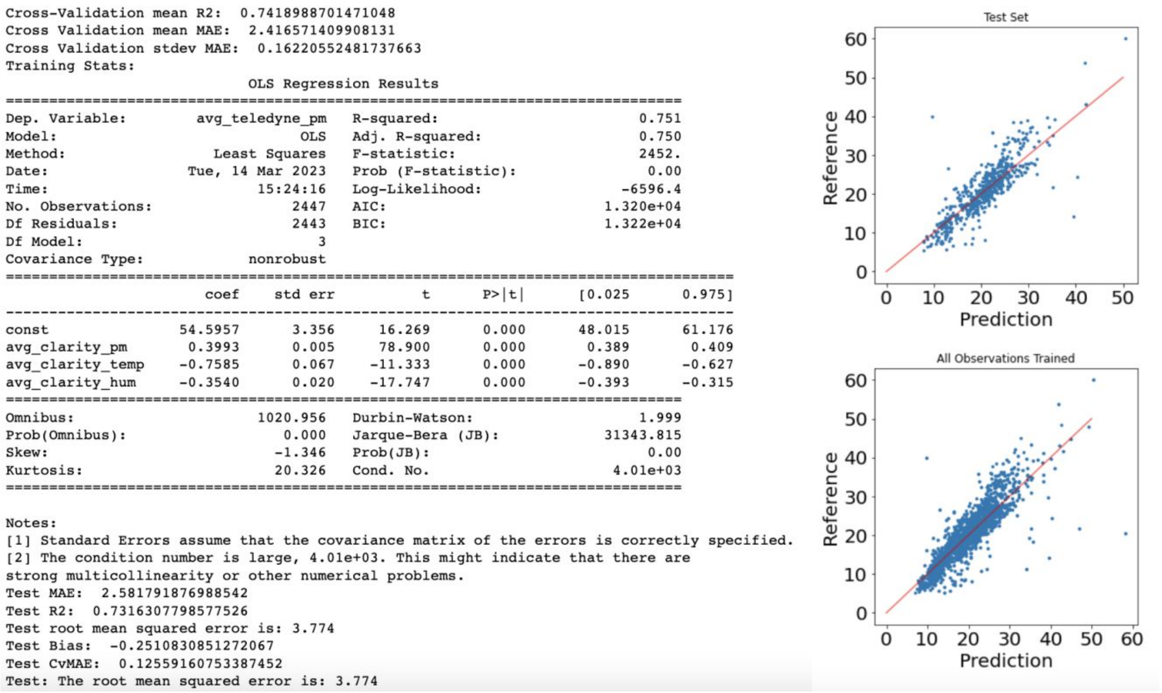
	Purple Air		**Clarity**		**Modulair-PM**	
	Coefficient	*Std Error*	*Coefficient*	*Std Error*	*Coefficient*	*Std Error*
Constant	17.51	2.20	54.60	3.36	19.82	1.92
$PM_{2.5}$	0.69	0.01	0.40	0.01	0.94	0.01
Temperature	-0.12	0.05	-0.76	0.07	-0.34	0.05
Relative Humidity	-0.15	0.01	-0.35	0.02	-0.08	0.01

Figure S4. Multiple Linear Regression statistics for PurpleAir, with $PM_{2.5}$ (Channel A/B average), temperature, and relative humidity as measured by the PurpleAir monitors. Red line is 1:1 line.



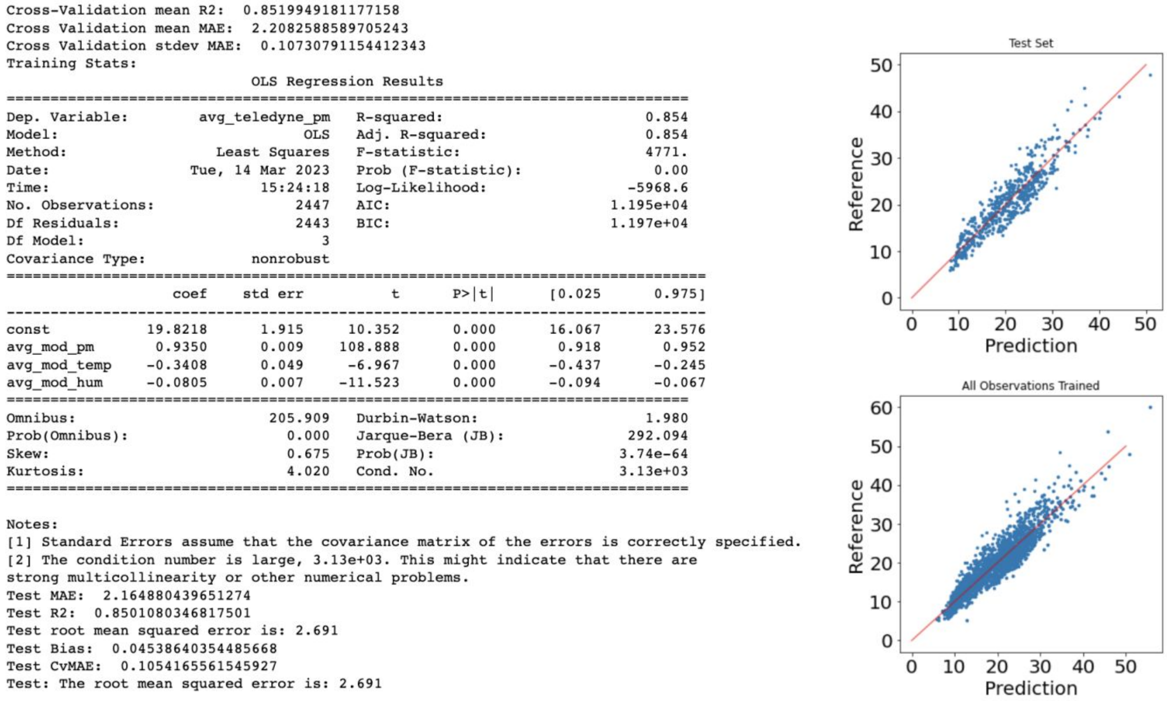
```
Cross-Validation mean R2:  0.8635563670911305
Cross Validation mean MAE:  1.996581953919814
Cross Validation stdev MAE:  0.10690691951048119
Training Stats:
                            OLS Regression Results
==============================================================================
Dep. Variable:        avg_teledyne_pm   R-squared:                       0.865
Model:                            OLS   Adj. R-squared:                  0.864
Method:                 Least Squares   F-statistic:                     5197.
Date:                Tue, 14 Mar 2023   Prob (F-statistic):               0.00
Time:                        15:24:14   Log-Likelihood:                -5826.0
No. Observations:                2447   AIC:                         1.166e+04
Df Residuals:                    2443   BIC:                         1.168e+04
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          17.5145      2.198      7.970      0.000      13.205      21.824
avg_pa_pm       0.6901      0.006    117.798      0.000       0.679       0.702
avg_pa_temp    -0.1201      0.047     -2.581      0.010      -0.211      -0.029
avg_pa_hum     -0.1530      0.013    -11.649      0.000      -0.179      -0.127
==============================================================================
Omnibus:                      424.695   Durbin-Watson:                   1.916
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1059.276
Skew:                           0.951   Prob(JB):                    9.57e-231
Kurtosis:                       5.602   Cond. No.                     3.12e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.12e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
Test MAE:  2.102203161226105
Test R2:  0.8546988170186818
Test root mean squared error is: 2.882
Test Bias:  -0.030420613518685725
Test CvMAE:  0.1034924310784799
Test: The root mean squared error is: 2.882
```

Figure S5. Multiple Linear Regression statistics for Clarity, with $PM_{2.5}$, temperature and relative humidity as measured by the monitors. Red line is 1:1 line.



```
Cross-Validation mean R2:  0.7418988701471048
Cross Validation mean MAE:  2.416571409908131
Cross Validation stdev MAE:  0.16220552481737663
Training Stats:
                            OLS Regression Results
==============================================================================
Dep. Variable:        avg_teledyne_pm   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.750
Method:                 Least Squares   F-statistic:                     2452.
Date:                Tue, 14 Mar 2023   Prob (F-statistic):               0.00
Time:                        15:24:16   Log-Likelihood:                -6596.4
No. Observations:                2447   AIC:                         1.320e+04
Df Residuals:                    2443   BIC:                         1.322e+04
Df Model:                           3
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const               54.5957      3.356     16.269      0.000      48.015      61.176
avg_clarity_pm       0.3993      0.005     78.900      0.000       0.389       0.409
avg_clarity_temp    -0.7585      0.067    -11.333      0.000      -0.890      -0.627
avg_clarity_hum     -0.3540      0.020    -17.747      0.000      -0.393      -0.315
==============================================================================
Omnibus:                     1020.956   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            31343.815
Skew:                          -1.346   Prob(JB):                         0.00
Kurtosis:                      20.326   Cond. No.                     4.01e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.01e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
Test MAE:  2.581791876988542
Test R2:  0.7316307798577526
Test root mean squared error is: 3.774
Test Bias:  -0.2510830851272067
Test CvMAE:  0.12559160753387452
Test: The root mean squared error is: 3.774
```

Figure S6. Multiple Linear Regression statistics for Modulair-PM, with PM$_{2.5}$, x2 temperature and relative humidity as measured by the Modulair-PM monitors. Red line is 1:1 line.



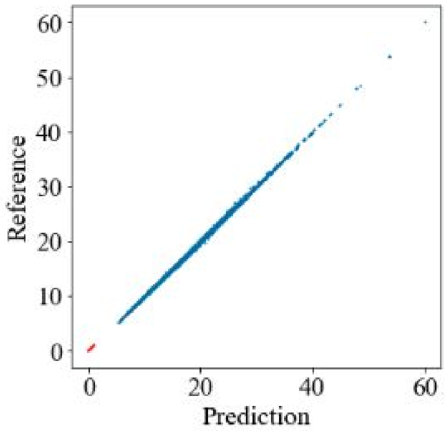
```
Cross-Validation mean R2:  0.8519949181177158
Cross Validation mean MAE:  2.2082588589705243
Cross Validation stdev MAE:  0.10730791154412343
Training Stats:
                            OLS Regression Results
==============================================================================
Dep. Variable:        avg_teledyne_pm   R-squared:                       0.854
Model:                            OLS   Adj. R-squared:                  0.854
Method:                 Least Squares   F-statistic:                     4771.
Date:                Tue, 14 Mar 2023   Prob (F-statistic):               0.00
Time:                        15:24:18   Log-Likelihood:                -5968.6
No. Observations:                2447   AIC:                         1.195e+04
Df Residuals:                    2443   BIC:                         1.197e+04
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           19.8218      1.915     10.352      0.000      16.067      23.576
avg_mod_pm       0.9350      0.009    108.888      0.000       0.918       0.952
avg_mod_temp    -0.3408      0.049     -6.967      0.000      -0.437      -0.245
avg_mod_hum     -0.0805      0.007    -11.523      0.000      -0.094      -0.067
==============================================================================
Omnibus:                      205.909   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              292.094
Skew:                           0.675   Prob(JB):                     3.74e-64
Kurtosis:                       4.020   Cond. No.                     3.13e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.13e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
Test MAE:  2.164880439651274
Test R2:  0.8501080346817501
Test root mean squared error is: 2.691
Test Bias:  0.04538640354485668
Test CvMAE:  0.1054165561545927
Test: The root mean squared error is: 2.691
```

Figure S7. XGBoost training for PurpleAir. Red line is 1:1 line.

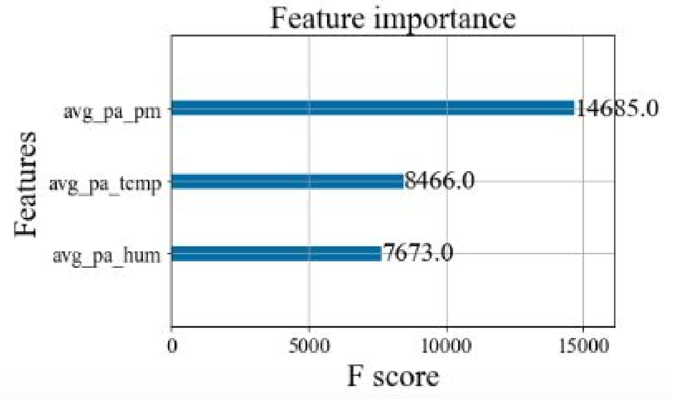
Grid search space:

'learning_rate': [0.01,0.05,0.1], 'max_depth': [5,6], 'n_estimators': [100,1000]

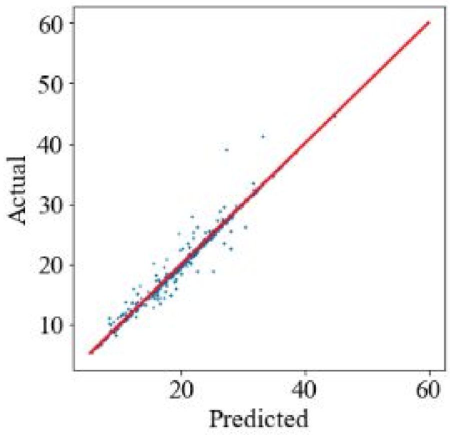
```
Start grid search
Finish grid search, it takes 171.42178916931152
learning_rate 0.1
max_depth 6
n_estimators 1000
Evaluate the models for training
```



```
Training set: The coefficient of determination is: 1.000
Training set: The index of agreement is: 0.380
Training set: The root mean squared error is: 0.146
Training set: The mean absolute error is: 0.095
```



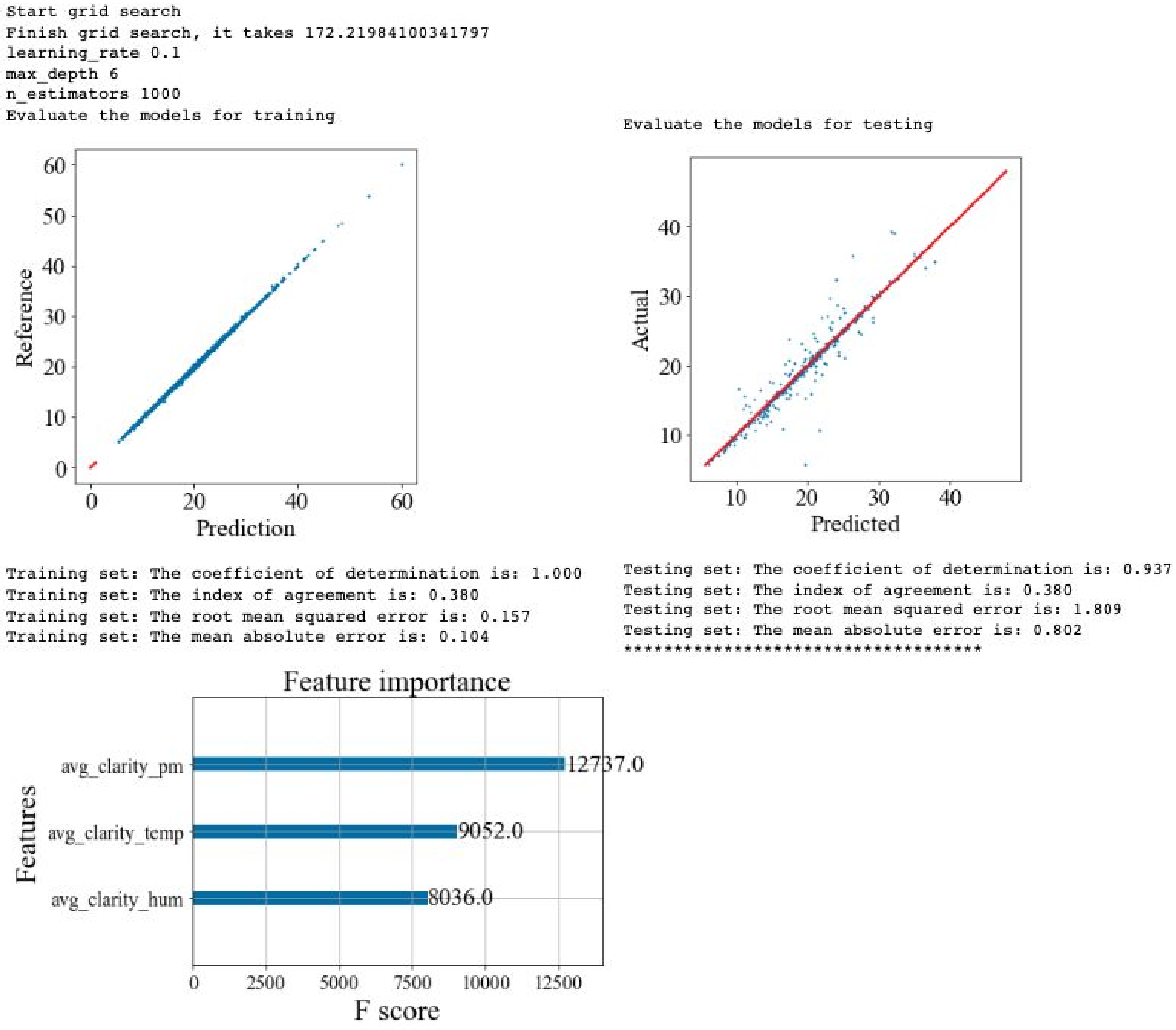
```
Evaluate the models for testing
```



```
Testing set: The coefficient of determination is: 0.970
Testing set: The index of agreement is: 0.377
Testing set: The root mean squared error is: 1.235
Testing set: The mean absolute error is: 0.557
*************************************
```

Figure S8. XGBoost training for Clarity. Red line is 1:1 line.

Grid search space:
'learning_rate': [0.01,0.05,0.1], 'max_depth': [5,6], 'n_estimators': [100,1000]



```
Start grid search
Finish grid search, it takes 172.21984100341797
learning_rate 0.1
max_depth 6
n_estimators 1000
Evaluate the models for training
```

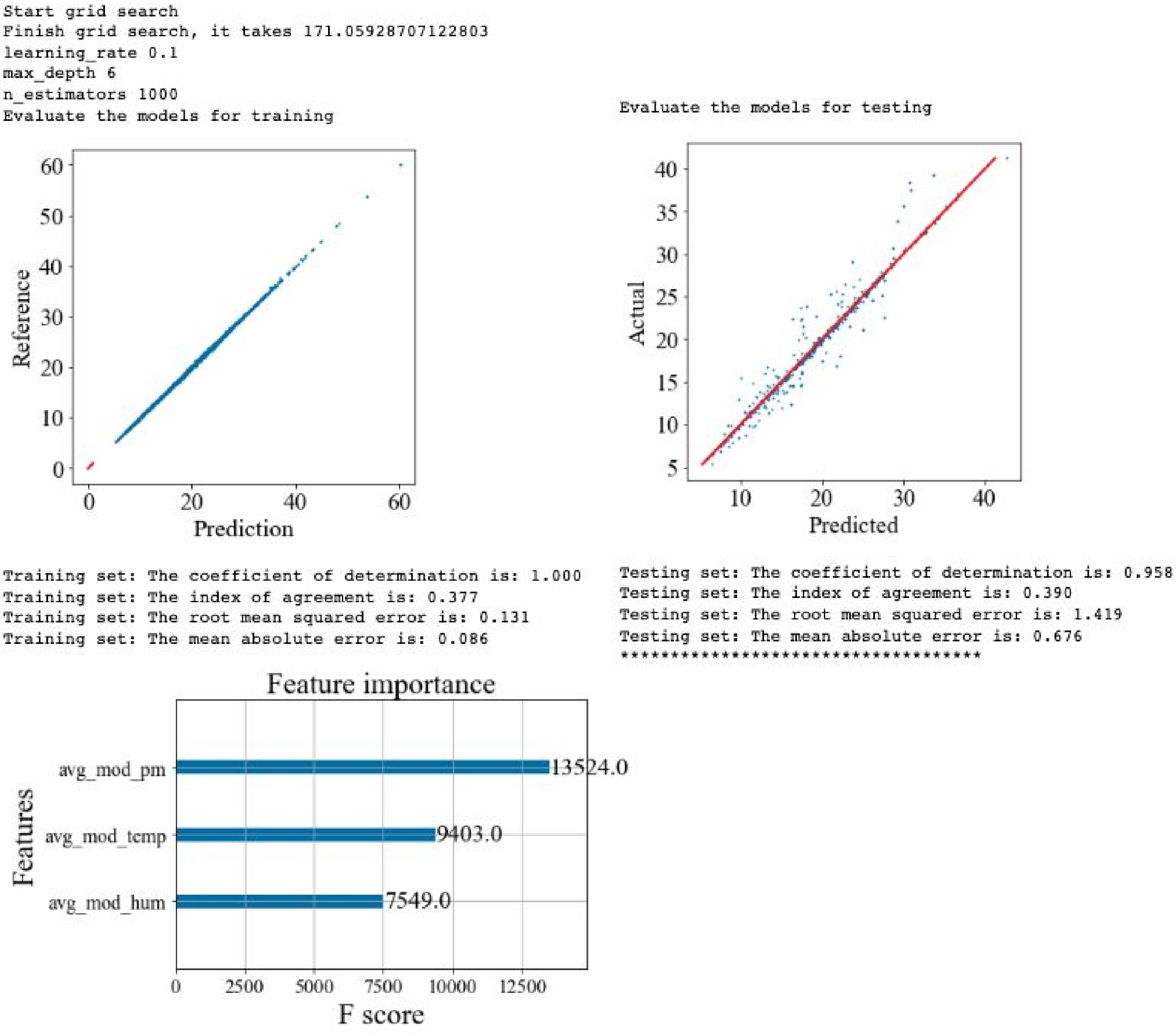
```
Evaluate the models for testing
```

```
Training set: The coefficient of determination is: 1.000
Training set: The index of agreement is: 0.380
Training set: The root mean squared error is: 0.157
Training set: The mean absolute error is: 0.104
```

```
Testing set: The coefficient of determination is: 0.937
Testing set: The index of agreement is: 0.380
Testing set: The root mean squared error is: 1.809
Testing set: The mean absolute error is: 0.802
************************************
```

Figure S9. XGBoost training for Modulair-PM. Red line is 1:1 line.

Grid search space:
'learning_rate': [0.01,0.05,0.1], 'max_depth': [5,6], 'n_estimators': [100,1000]



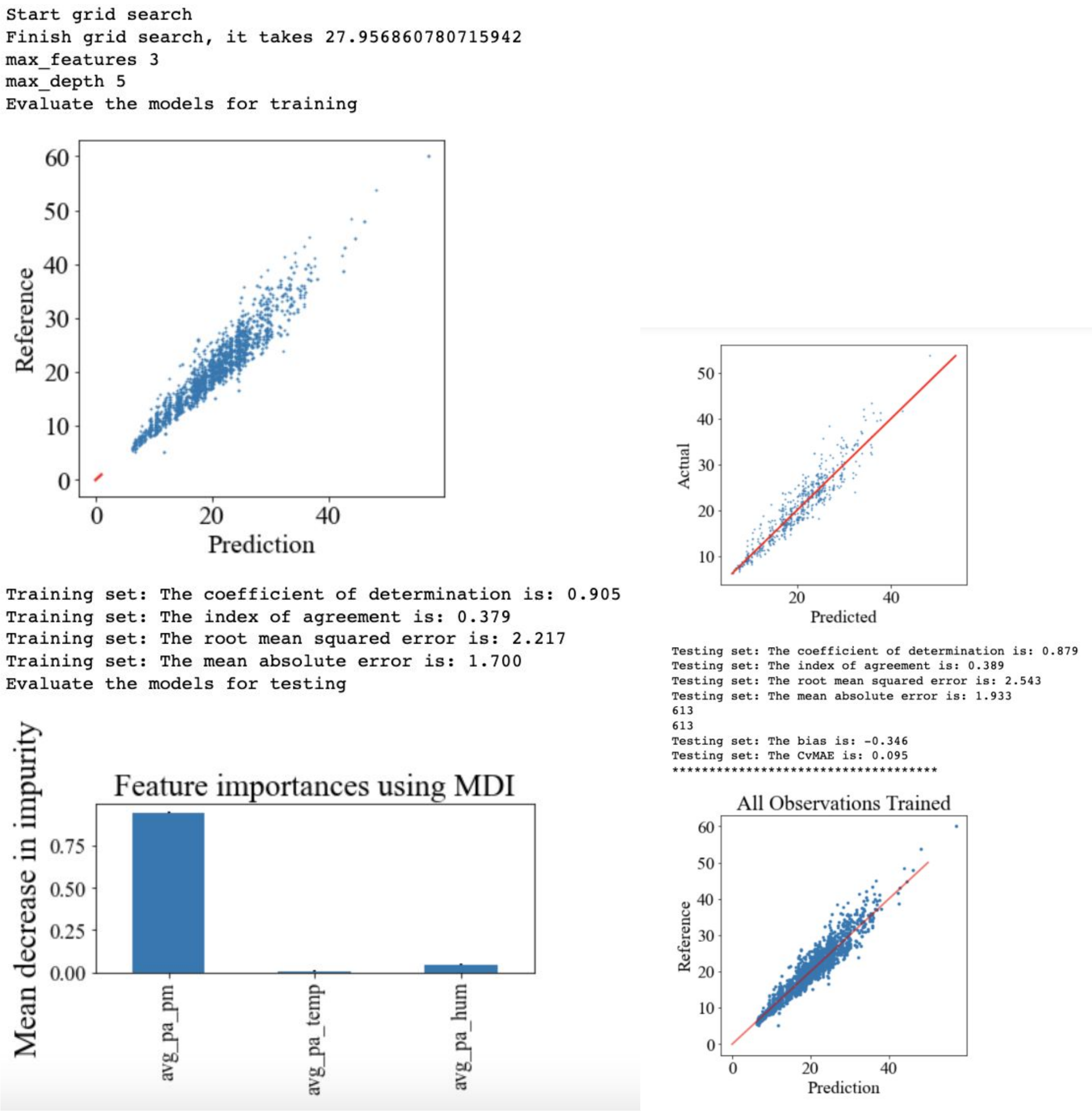
```
Start grid search
Finish grid search, it takes 171.05928707122803
learning_rate 0.1
max_depth 6
n_estimators 1000
Evaluate the models for training
```

```
Training set: The coefficient of determination is: 1.000
Training set: The index of agreement is: 0.377
Training set: The root mean squared error is: 0.131
Training set: The mean absolute error is: 0.086
```

```
Evaluate the models for testing
```

```
Testing set: The coefficient of determination is: 0.958
Testing set: The index of agreement is: 0.390
Testing set: The root mean squared error is: 1.419
Testing set: The mean absolute error is: 0.676
************************************
```

Figure S10. Random Forest training for PurpleAir. Red line is 1:1 line.

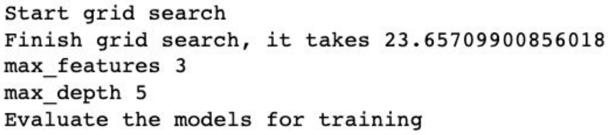


```
Start grid search
Finish grid search, it takes 27.956860780715942
max_features 3
max_depth 5
Evaluate the models for training
```

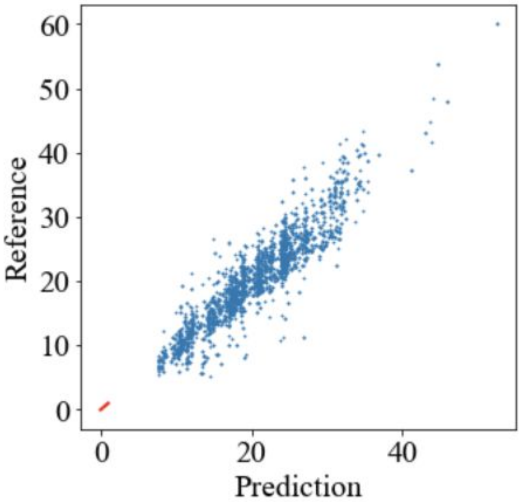
```
Training set: The coefficient of determination is: 0.905
Training set: The index of agreement is: 0.379
Training set: The root mean squared error is: 2.217
Training set: The mean absolute error is: 1.700
Evaluate the models for testing
```

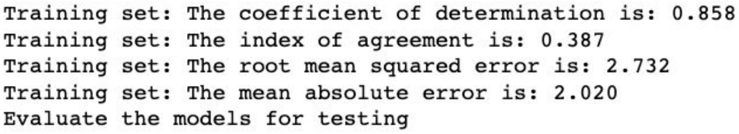
```
Testing set: The coefficient of determination is: 0.879
Testing set: The index of agreement is: 0.389
Testing set: The root mean squared error is: 2.543
Testing set: The mean absolute error is: 1.933
613
613
Testing set: The bias is: -0.346
Testing set: The CvMAE is: 0.095
************************************
```

Figure S11. Random Forest Training for Clarity. Red is 1:1 line.

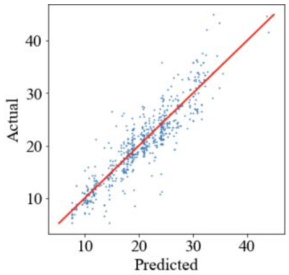


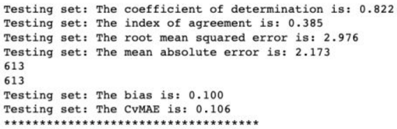
```
Start grid search
Finish grid search, it takes 23.65709900856018
max_features 3
max_depth 5
Evaluate the models for training
```



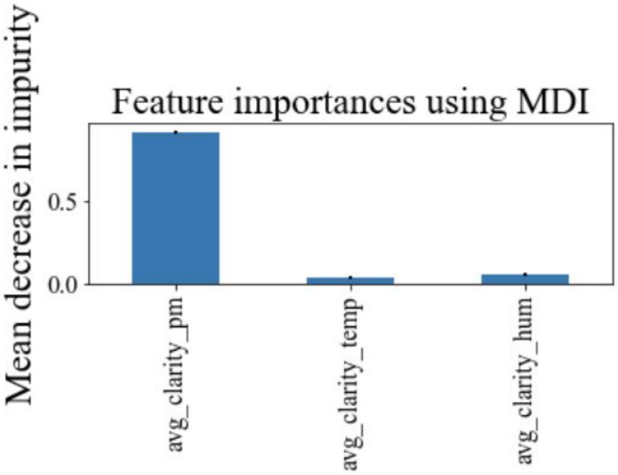


```
Training set: The coefficient of determination is: 0.858
Training set: The index of agreement is: 0.387
Training set: The root mean squared error is: 2.732
Training set: The mean absolute error is: 2.020
Evaluate the models for testing
```





```
Testing set: The coefficient of determination is: 0.822
Testing set: The index of agreement is: 0.385
Testing set: The root mean squared error is: 2.976
Testing set: The mean absolute error is: 2.173
613
613
Testing set: The bias is: 0.100
Testing set: The CvMAE is: 0.106
************************************
```



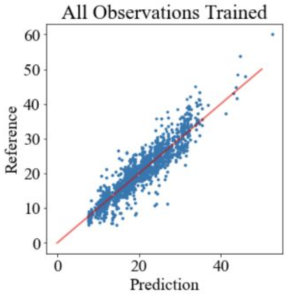
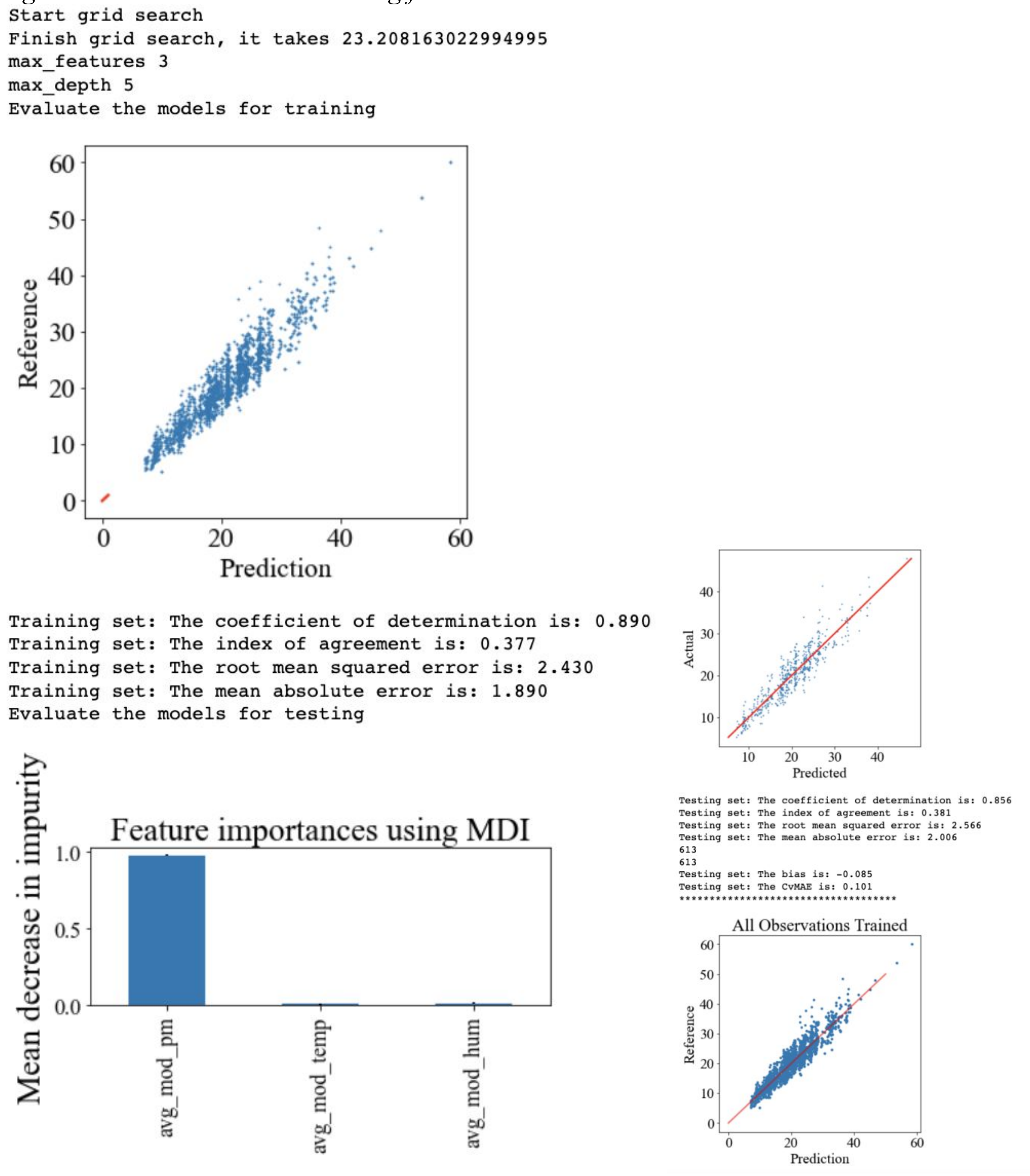


Figure S12. Random Forest training for Modulair.-PM Red is 1:1 line.

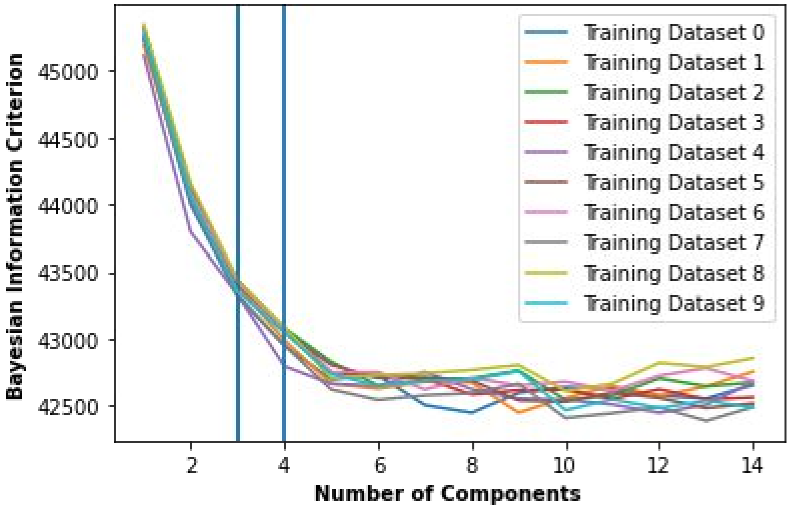


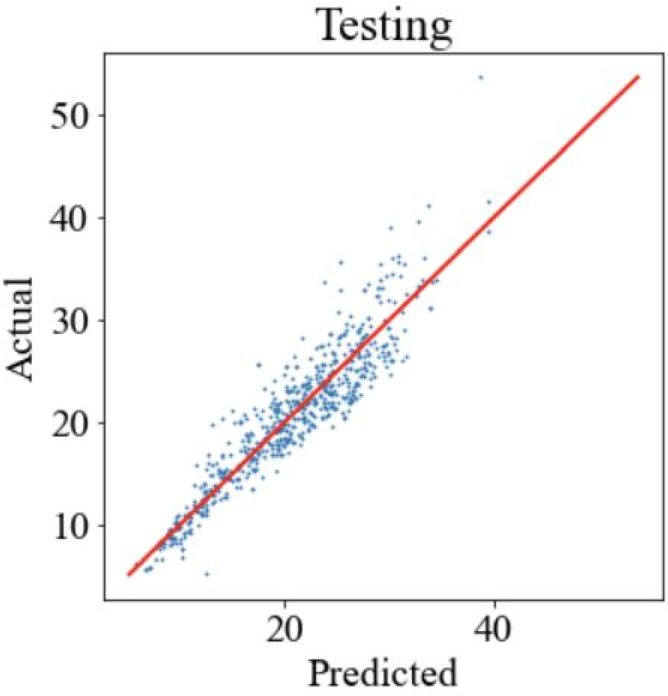
```
Start grid search
Finish grid search, it takes 23.208163022994995
max_features 3
max_depth 5
Evaluate the models for training
```

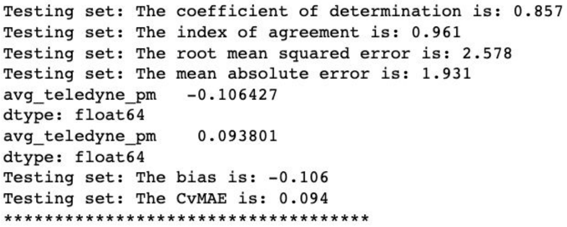
```
Training set: The coefficient of determination is: 0.890
Training set: The index of agreement is: 0.377
Training set: The root mean squared error is: 2.430
Training set: The mean absolute error is: 1.890
Evaluate the models for testing
```

```
Testing set: The coefficient of determination is: 0.856
Testing set: The index of agreement is: 0.381
Testing set: The root mean squared error is: 2.566
Testing set: The mean absolute error is: 2.006
613
613
Testing set: The bias is: -0.085
Testing set: The CvMAE is: 0.101
***********************************
```

Figure S13. GMR Training for PurpleAir. Red diagonal line is 1:1 line.







```
Testing set: The coefficient of determination is: 0.857
Testing set: The index of agreement is: 0.961
Testing set: The root mean squared error is: 2.578
Testing set: The mean absolute error is: 1.931
avg_teledyne_pm   -0.106427
dtype: float64
avg_teledyne_pm    0.093801
dtype: float64
Testing set: The bias is: -0.106
Testing set: The CvMAE is: 0.094
************************************
```

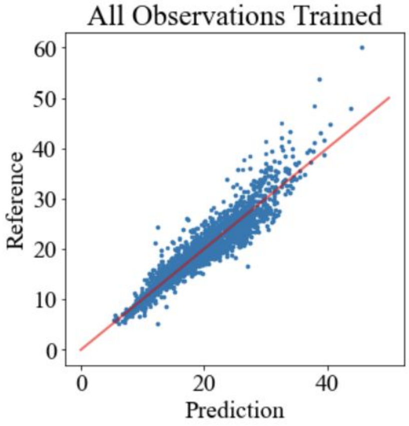
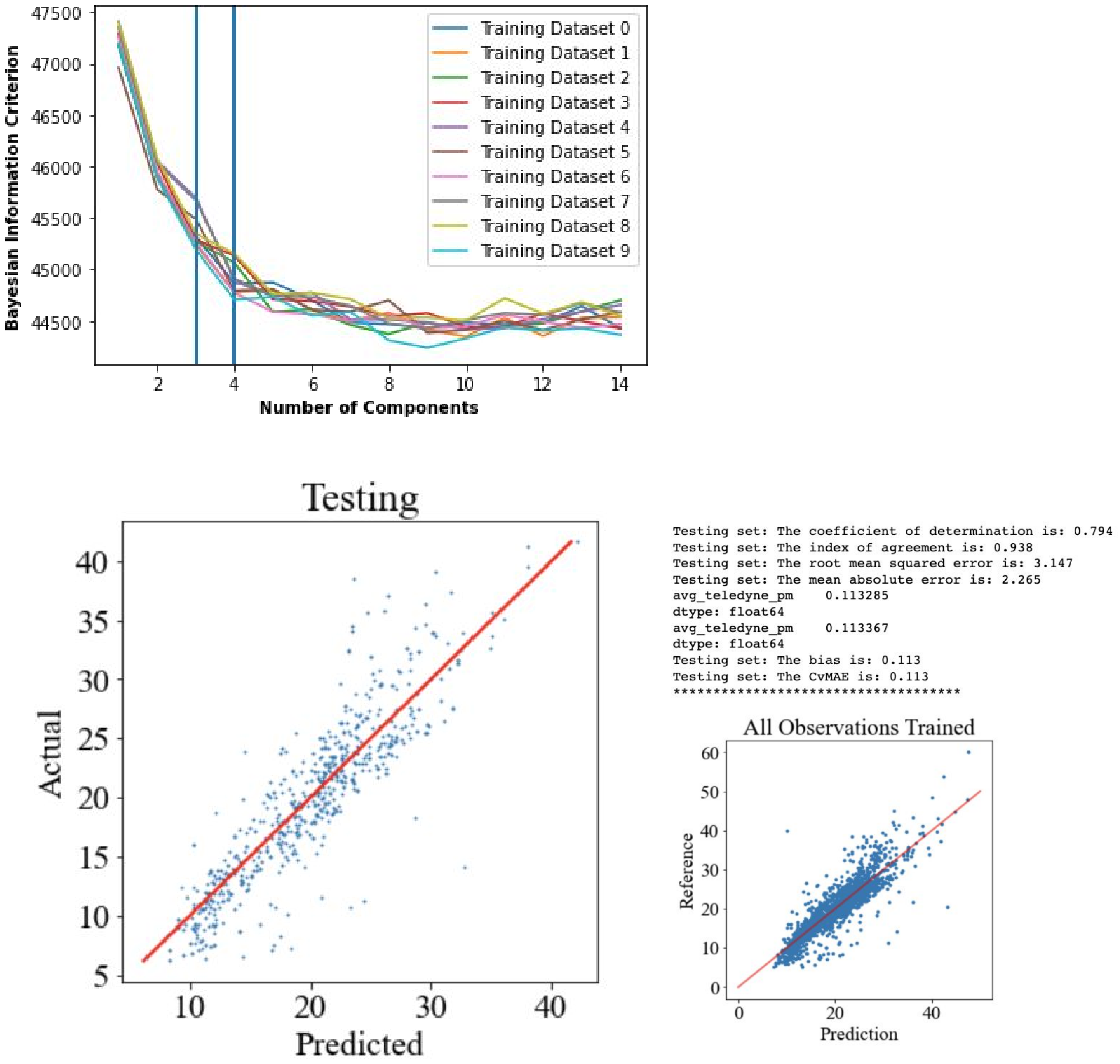
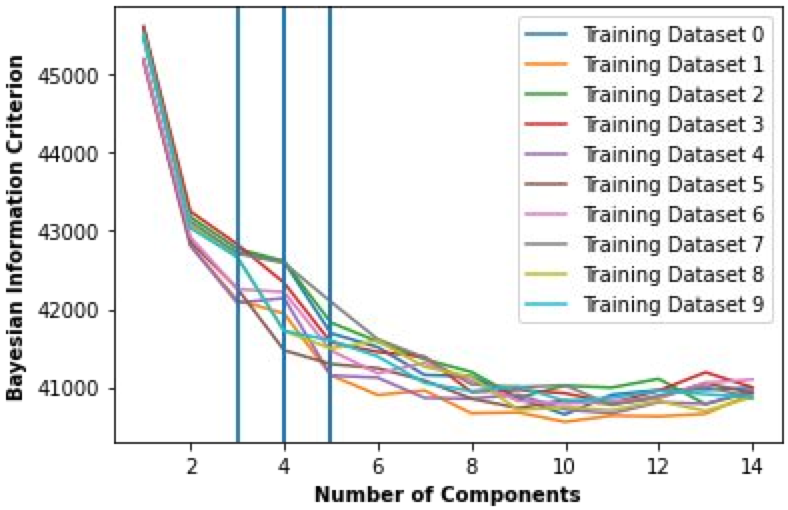


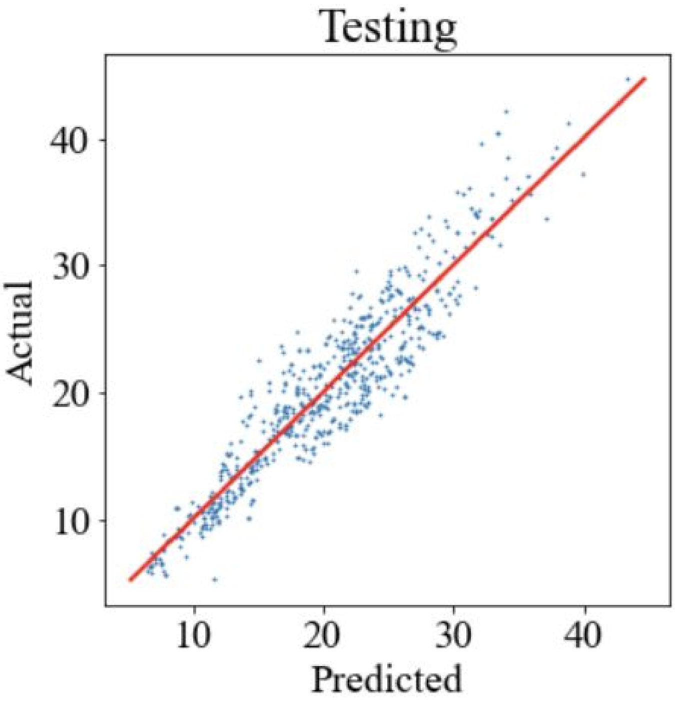
Figure S14. GMR Training for Clarity. Red diagonal line is 1:1 line.

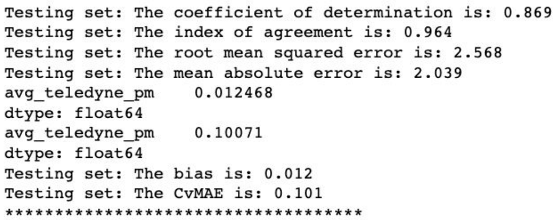


```
Testing set: The coefficient of determination is: 0.794
Testing set: The index of agreement is: 0.938
Testing set: The root mean squared error is: 3.147
Testing set: The mean absolute error is: 2.265
avg_teledyne_pm    0.113285
dtype: float64
avg_teledyne_pm    0.113367
dtype: float64
Testing set: The bias is: 0.113
Testing set: The CvMAE is: 0.113
************************************
```

Figure S15. GMR Training for Modulair-PM. Red diagonal line is 1:1 line.







```
Testing set: The coefficient of determination is: 0.869
Testing set: The index of agreement is: 0.964
Testing set: The root mean squared error is: 2.568
Testing set: The mean absolute error is: 2.039
avg_teledyne_pm    0.012468
dtype: float64
avg_teledyne_pm    0.10071
dtype: float64
Testing set: The bias is: 0.012
Testing set: The CvMAE is: 0.101
************************************
```

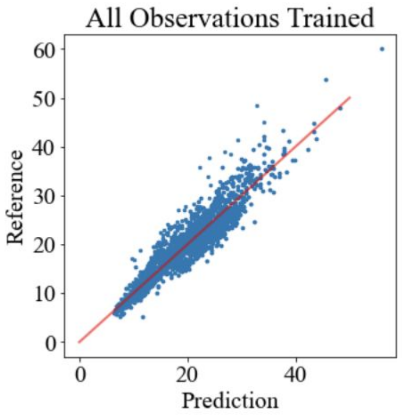


Figure S16. Distributions of raw $PM_{2.5}$ measured during the Accra deployment.

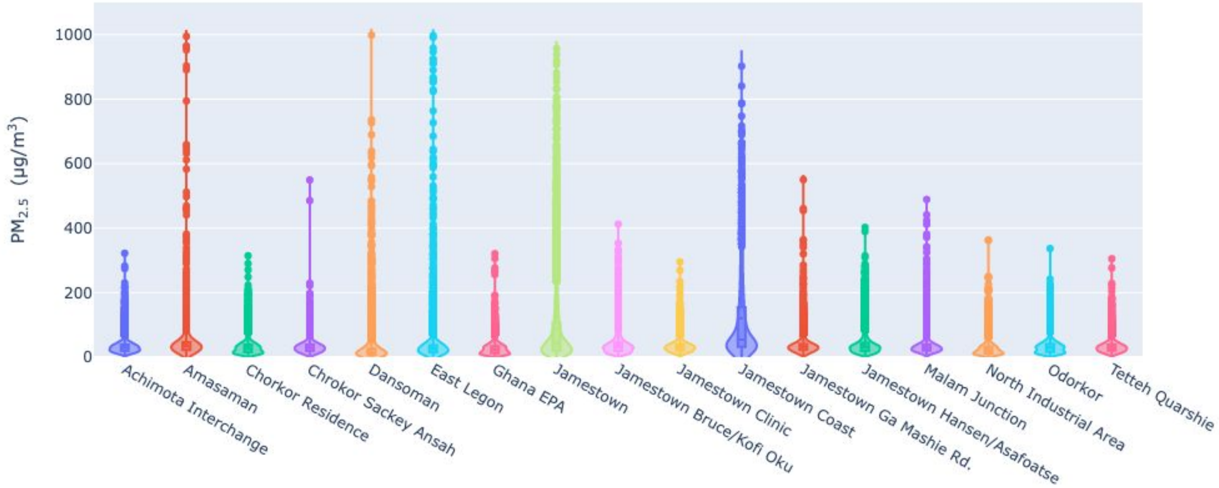


Figure S17. Comparison of $PM_{2.5}$ observations measured by Clarity monitors deployed around Accra, Ghana, corrected using four machine learning models, using the 4 month UG colocation as model training data.

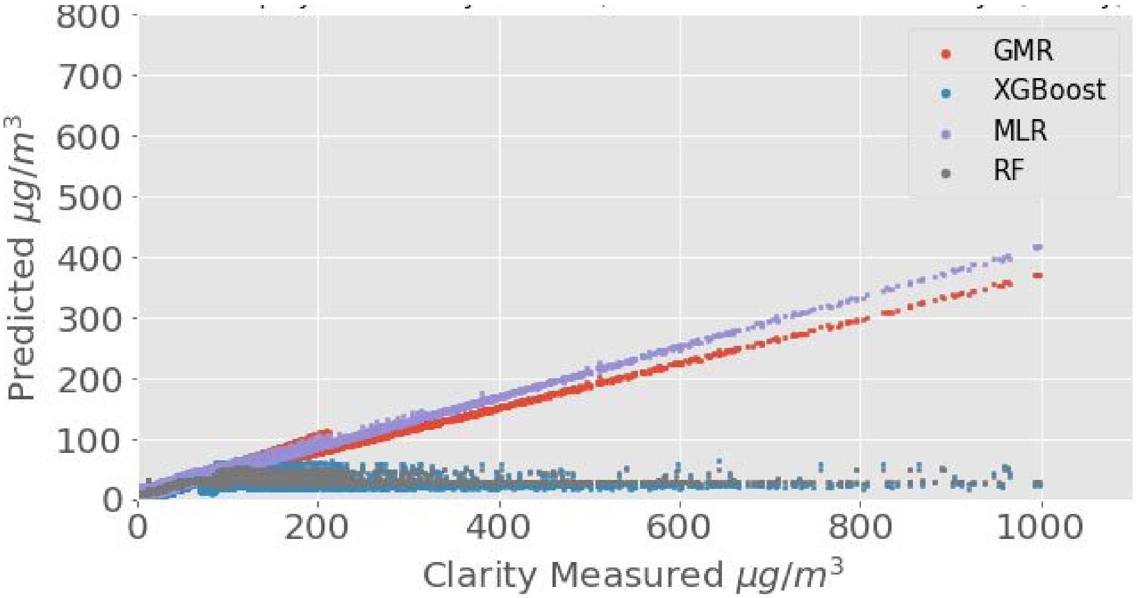


Figure S18. Distributions of raw $PM_{2.5}$ measured during the University of Ghana colocation.

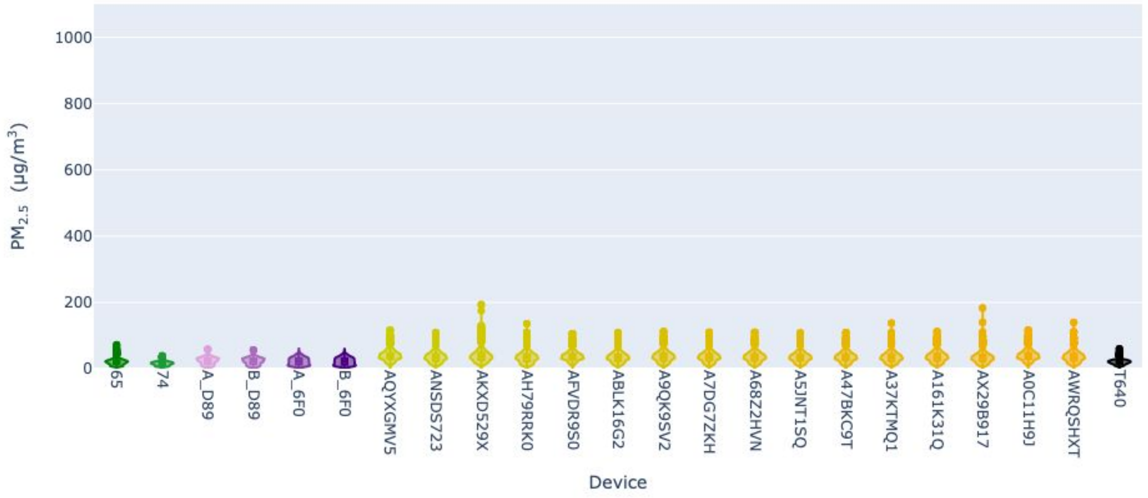


Figure S19. GMR-corrected annual $PM_{2.5}$ averages across the Accra Clarity network. Note that 2018 and 2021 are incomplete years (see Figure S10 for measurement date ranges).

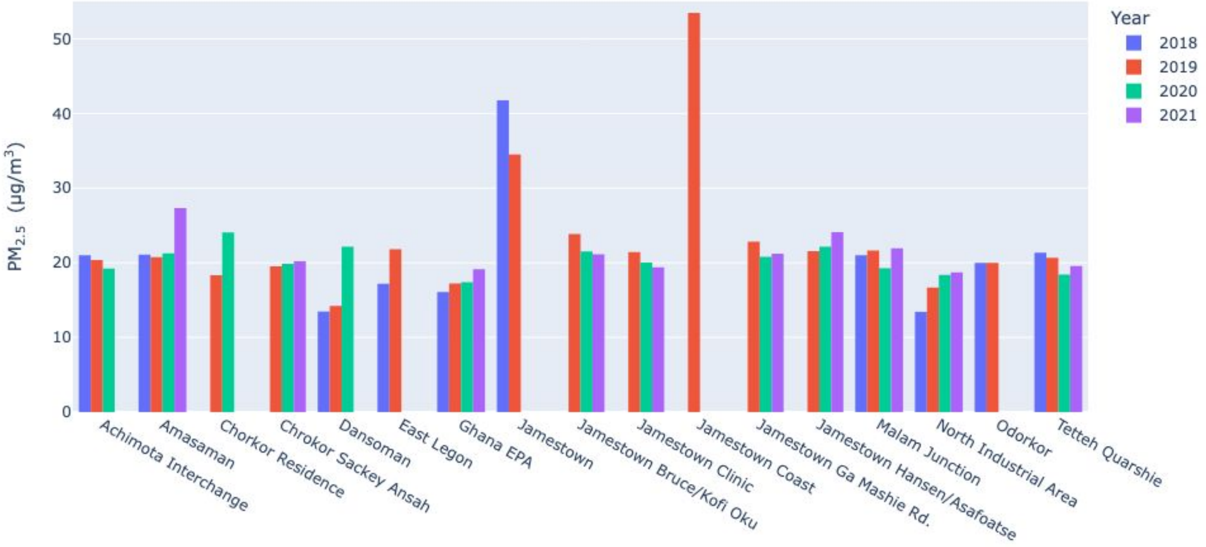
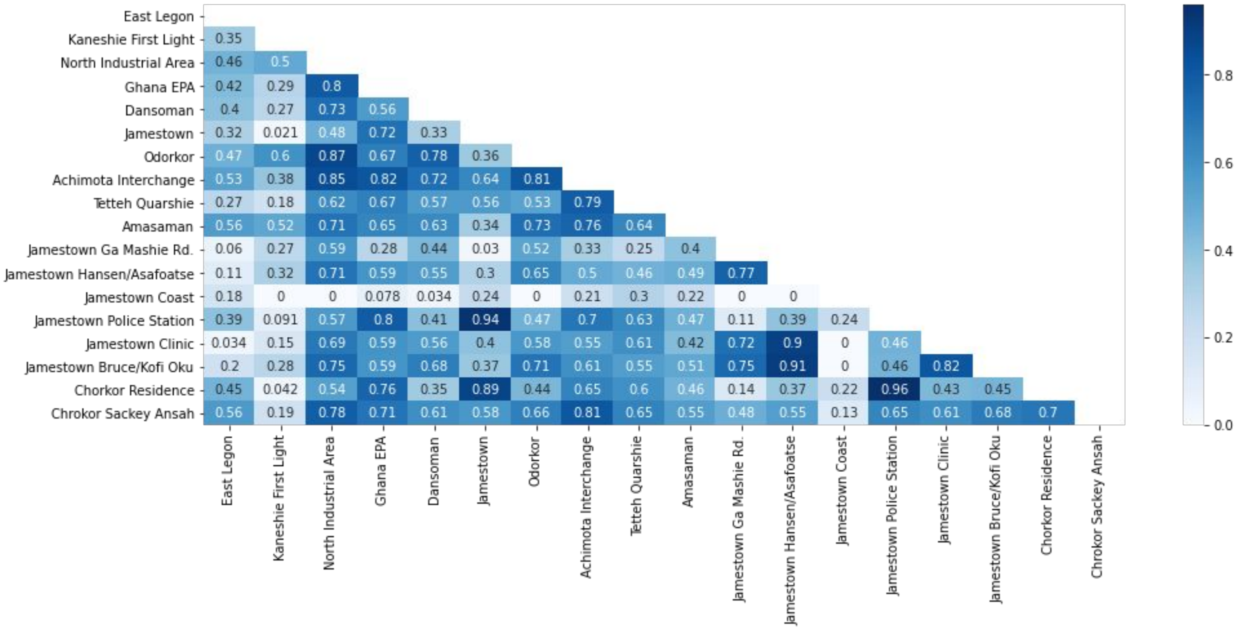


Figure S20. Pearson correlation coefficients of comparisons of hourly data to assess inter-site variability across the Accra Clarity network, where 0 represents no correlation and 1 represents perfect correlation.



	East Legon	Kaneshie First Light	North Industrial Area	Ghana EPA	Dansoman	Jamestown	Odorkor	Achimota Interchange	Tetteh Quarshie	Amasaman	Jamestown Ga Mashie Rd.	Jamestown Hansen/Asafoatse	Jamestown Coast	Jamestown Police Station	Jamestown Clinic	Jamestown Bruce/Kofi Oku	Chorkor Residence	Chrokor Sackey Ansah
East Legon																		
Kaneshie First Light	0.35																	
North Industrial Area	0.46	0.5																
Ghana EPA	0.42	0.29	0.8															
Dansoman	0.4	0.27	0.73	0.56														
Jamestown	0.32	0.021	0.48	0.72	0.33													
Odorkor	0.47	0.6	0.87	0.67	0.78	0.36												
Achimota Interchange	0.53	0.38	0.85	0.82	0.72	0.64	0.81											
Tetteh Quarshie	0.27	0.18	0.62	0.67	0.57	0.56	0.53	0.79										
Amasaman	0.56	0.52	0.71	0.65	0.63	0.34	0.73	0.76	0.64									
Jamestown Ga Mashie Rd.	0.06	0.27	0.59	0.28	0.44	0.03	0.52	0.33	0.25	0.4								
Jamestown Hansen/Asafoatse	0.11	0.32	0.71	0.59	0.55	0.3	0.65	0.5	0.46	0.49	0.77							
Jamestown Coast	0.18	0	0	0.078	0.034	0.24	0	0.21	0.3	0.22	0	0						
Jamestown Police Station	0.39	0.091	0.57	0.8	0.41	0.94	0.47	0.7	0.63	0.47	0.11	0.39	0.24					
Jamestown Clinic	0.034	0.15	0.69	0.59	0.56	0.4	0.58	0.55	0.61	0.42	0.72	0.9	0	0.46				
Jamestown Bruce/Kofi Oku	0.2	0.28	0.75	0.59	0.68	0.37	0.71	0.61	0.55	0.51	0.75	0.91	0	0.46	0.82			
Chorkor Residence	0.45	0.042	0.54	0.76	0.35	0.89	0.44	0.65	0.6	0.46	0.14	0.37	0.22	0.96	0.43	0.45		
Chrokor Sackey Ansah	0.56	0.19	0.78	0.71	0.61	0.58	0.66	0.81	0.65	0.55	0.48	0.55	0.13	0.65	0.61	0.68	0.7	