Aerosol-cloud interactions: Predicting Cloud Condensation Nuclei from AOS Measurements

Project: total_smokeshow

ARM Team Mentor
Maria Zawadowicz

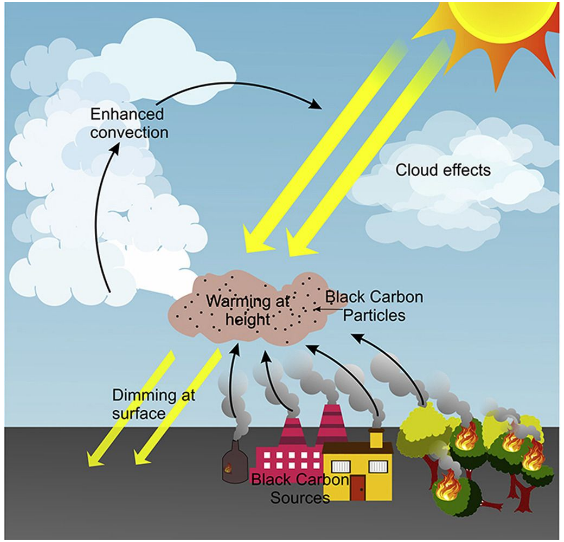
Student Team
- Abhigyan Chakraborty
- Ajinkya Desai
- Benjamin Marosites
- William Zhai

Science Questions

- Can we predict CCN from other measurable aerosol quantities using a machine learning model?
  - Accuracy
  - Efficiency (parsimonious use of observations)

- Does the model align with aerosol-cloud interaction physics?

Arif et al., 2018



Project Goal and Data

Develop and train a ML model to obtain 10-m estimates of CCN from aerosol properties

Use explainable AI to test model understanding of physics

Location: ARM Observatory, Southern Great Plains (SGP)

Time Range: Nov 1 - Dec 31, 2019
(Winter: Inorganic compounds more hygroscopic)

Geophysical Variables

Variable	Physical Explanation
Aerosol Size Distribution	
Particle Number Size Distribution [PNSD] (**merged_dN_dlogDp**)	SMPS and APS are merged for a wider size range, Strong indicator of CCN activation: larger particles, with high curvature, more hygroscopic
Diameter bounds for each bin	Track bin-to-count association
Cloud Condensation Nuclei (CCN)	
Mean number concentration of activated condensation nuclei (**N_CCN**)	Pass air through column with supersaturated water vapor (0.4%) Condensed particles counted…
Droplet count by bin size	(...and sized) Dimensionality reduction

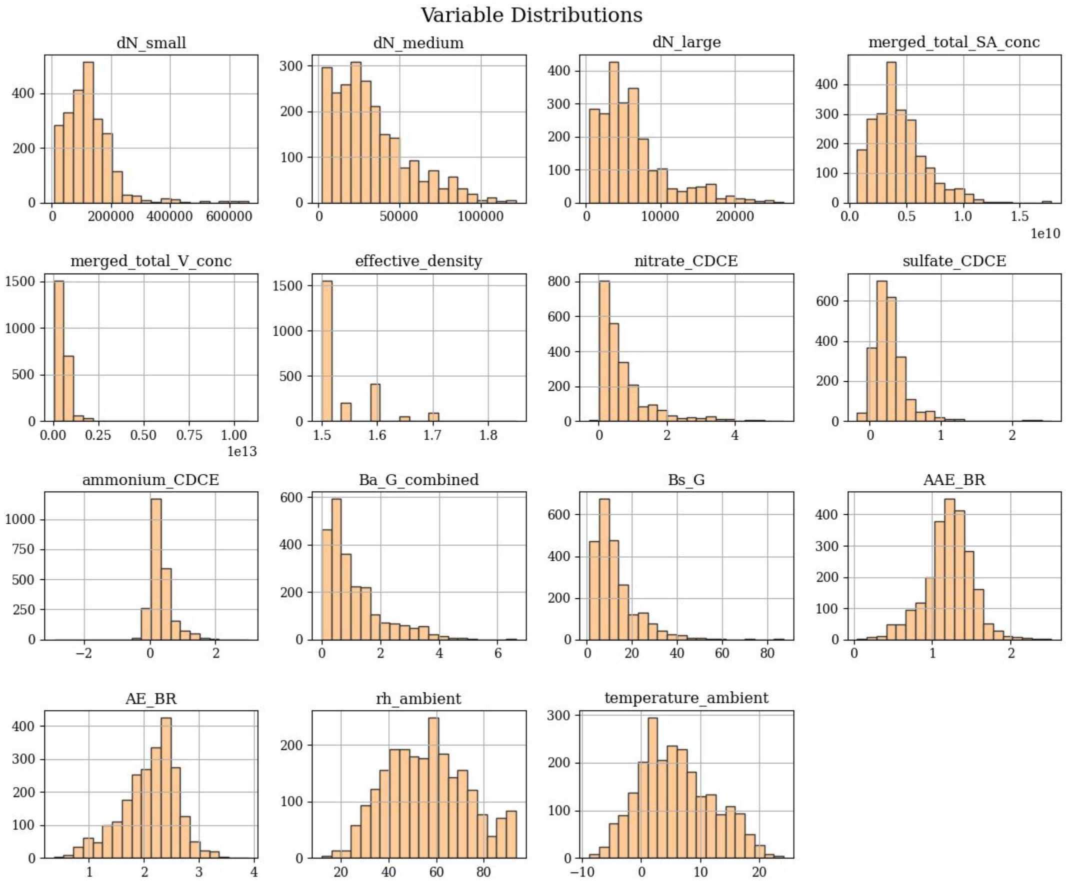
Geophysical Variables

Variable	Physical Explanation
Thermodynamic Variables	
Relative Humidity	Governs aerosol hygroscopicity and water uptake
Temperature	
Aerosol Chemical Composition	
Nitrates (**nitrate_CDCE**)	Inorganic compounds that are most hygroscopic
Sulfates (**sulfate_CDCE**)	
Ammonium (**ammonium_CDCE**)	
Chloride	Hygroscopic but negligible at SGP (Zabala et al. 2026)
Organic Compounds	Less likely to be hygroscopic

Geophysical Variables

Variable	Physical Explanation
Aerosol Optical Properties	
Absorption coefficient for green wavelength (**Ba_G_combined**)	Br, Black Carbon concentration information (composition without hygroscopic species)
Scattering coefficient for green wavelength (**Bs_G**)	Coarse vs. fine size information
Absorption Angstrom Exponent for blue/red wavelength (**AAE_BR**)	Proxy for aerosol composition
Scattering Angstrom exponent blue/red wavelengths (**AE_BR**)	Proxy for particle size
Aerosol Optical Depth	Variable repetition

Exploratory Data Analysis: Variables Distributions



Gauge the distribution of variables

Note gaps in the range of values

Other EDA Tools:

- ARM Data Discovery
- ARM Data Advisor

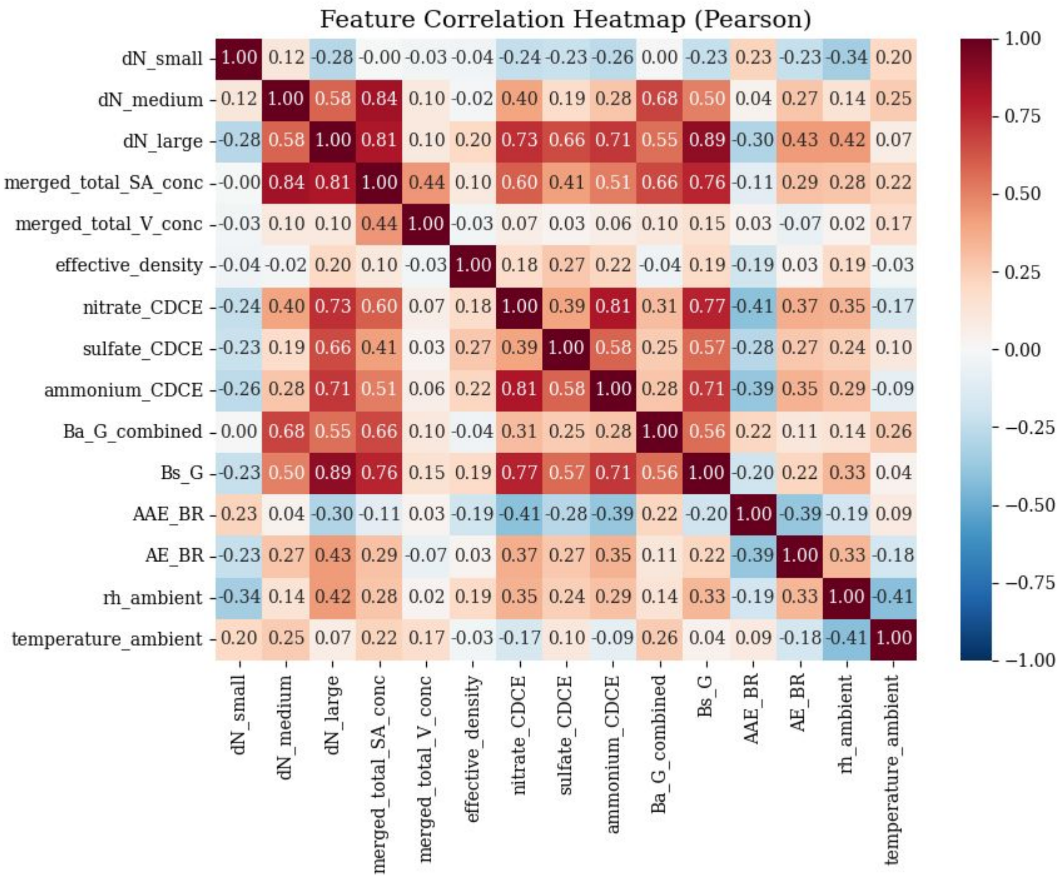
Exploratory Data Analysis: Pearson Correlation Heatmap

Testing for multicollinearity

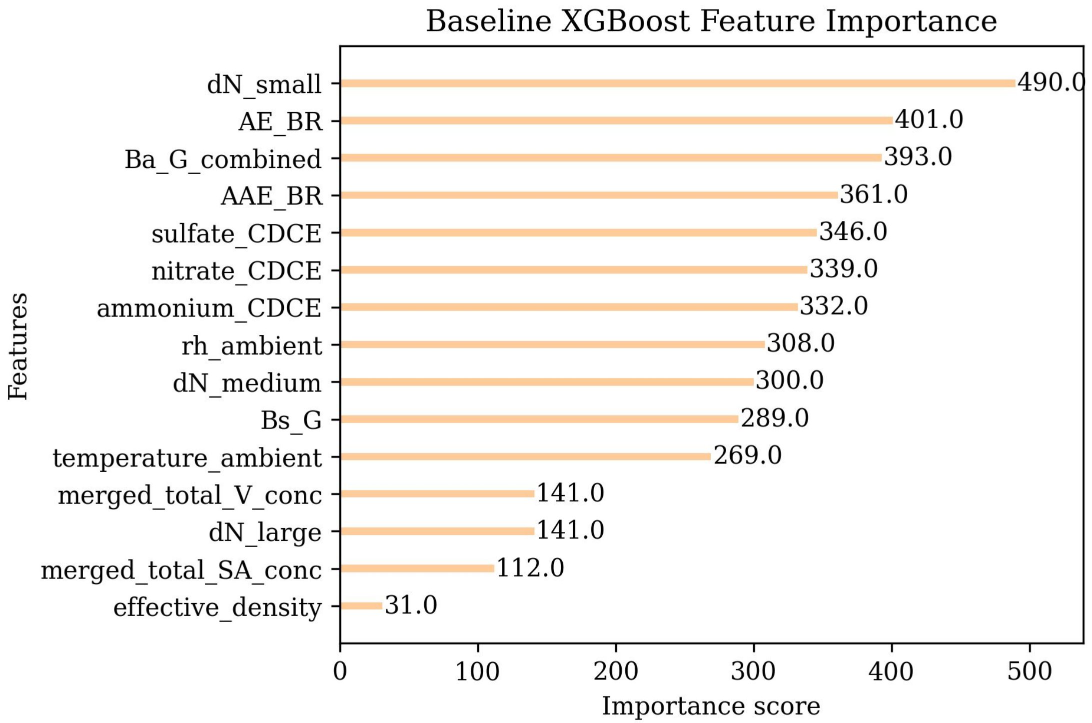
User defined threshold of 0.7-0.85

We could also test with other methods

- Scatter plots
- Variance Inflation Factor (VIF)



Feature Engineering



Iterative process:

Keep variables with high feature importance,

From Correlation heatmap identify features with high correlation with each other (>0.7). Drop the feature with lower Feature importance

The importance score indicates how frequently XGBoost selected a feature to make decisions while constructing the boosted trees.

Tuned XGBoost

```
param_grid =
{
    'n_estimators': [100, 200],
    'learning_rate': [0.03, 0.05, 0.1],
    'max_depth': [3, 4, 5],
    'subsample': [0.7, 0.8, 1.0],
    'colsample_bytree': [0.7, 0.8, 1.0]
}
```

Hyperparameter tuning using GridsearchCV With R^2 score

Best Parameters:

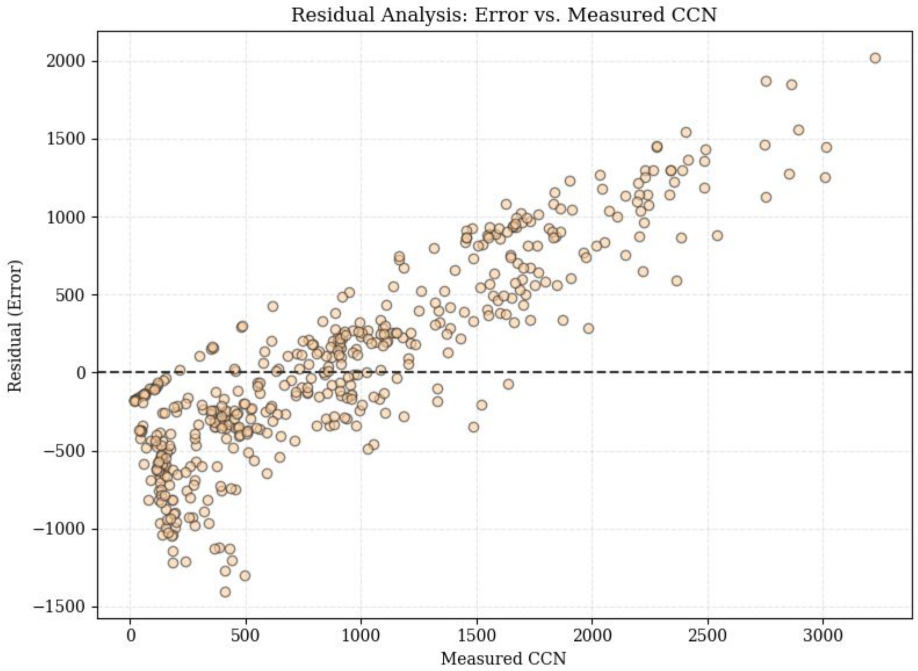
- 'colsample_bytree': 1.0,
- 'learning_rate': 0.03,
- 'max_depth': 3,
- 'n_estimators': 100,
- 'subsample': 0.7

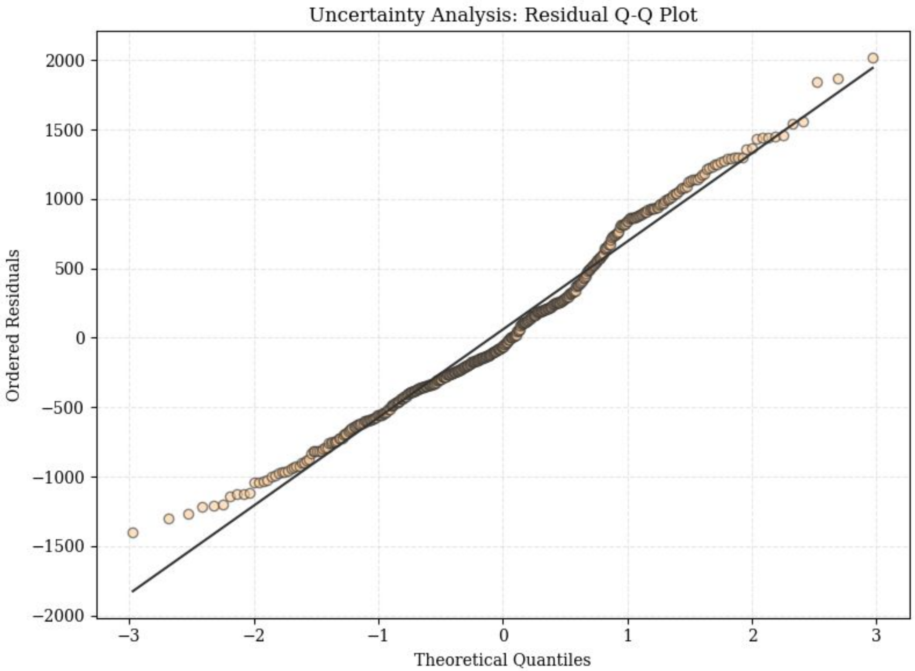
Model Performance:

R2 Score: 0.210

RMSE: 639.965

Tuned XGBoost: Model Uncertainty





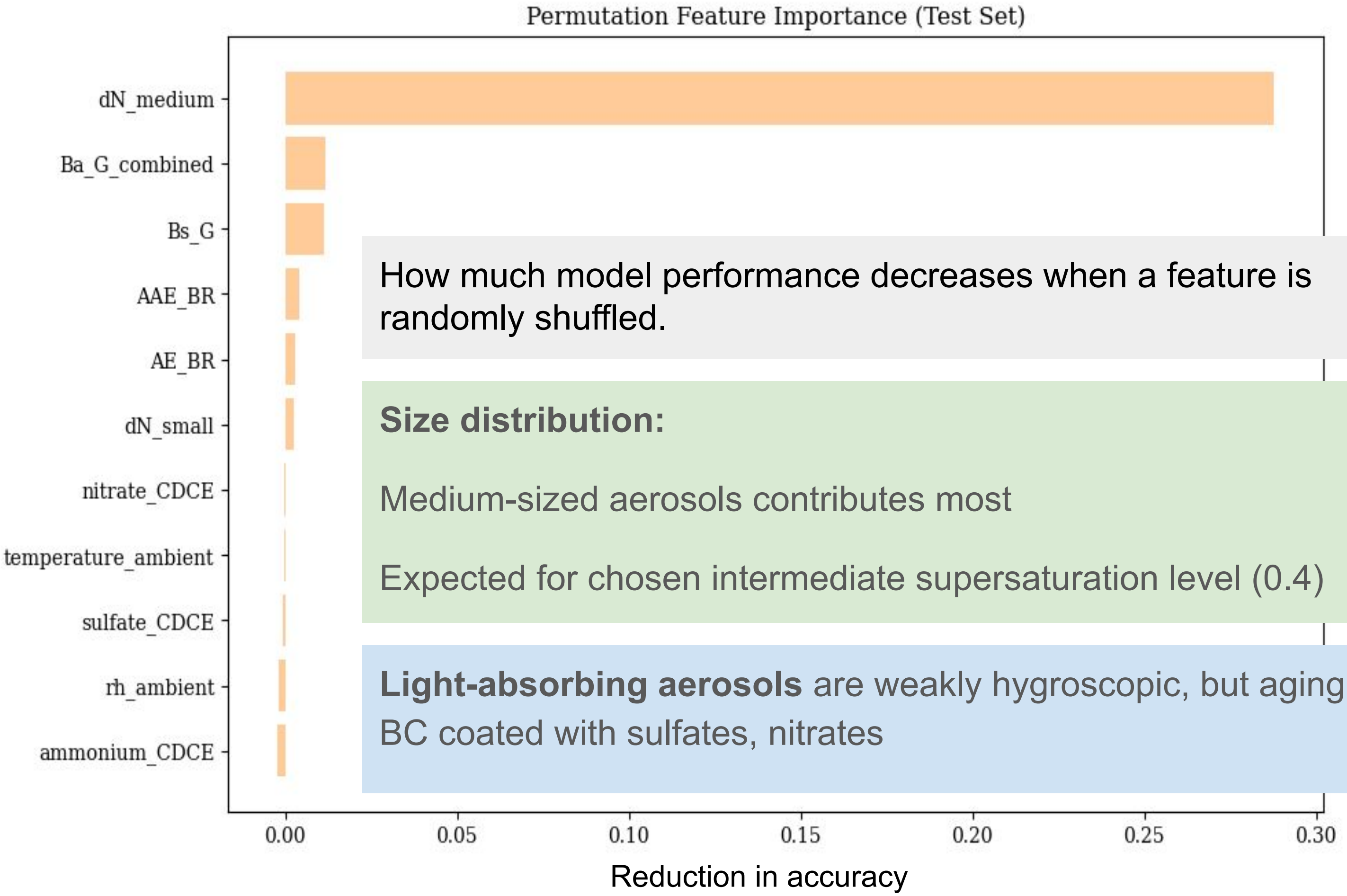
Ideally should be be scattered around the zero line

Over estimating at lower quantiles

XAI: Explainable AI helps us understand **why** a machine learning model makes predictions

Instead of treating the model as a “black box,” XAI techniques reveal:

- which features matter most
- how much they influence predictions
- whether they increase or decrease outcomes.



How much model performance decreases when a feature is randomly shuffled.

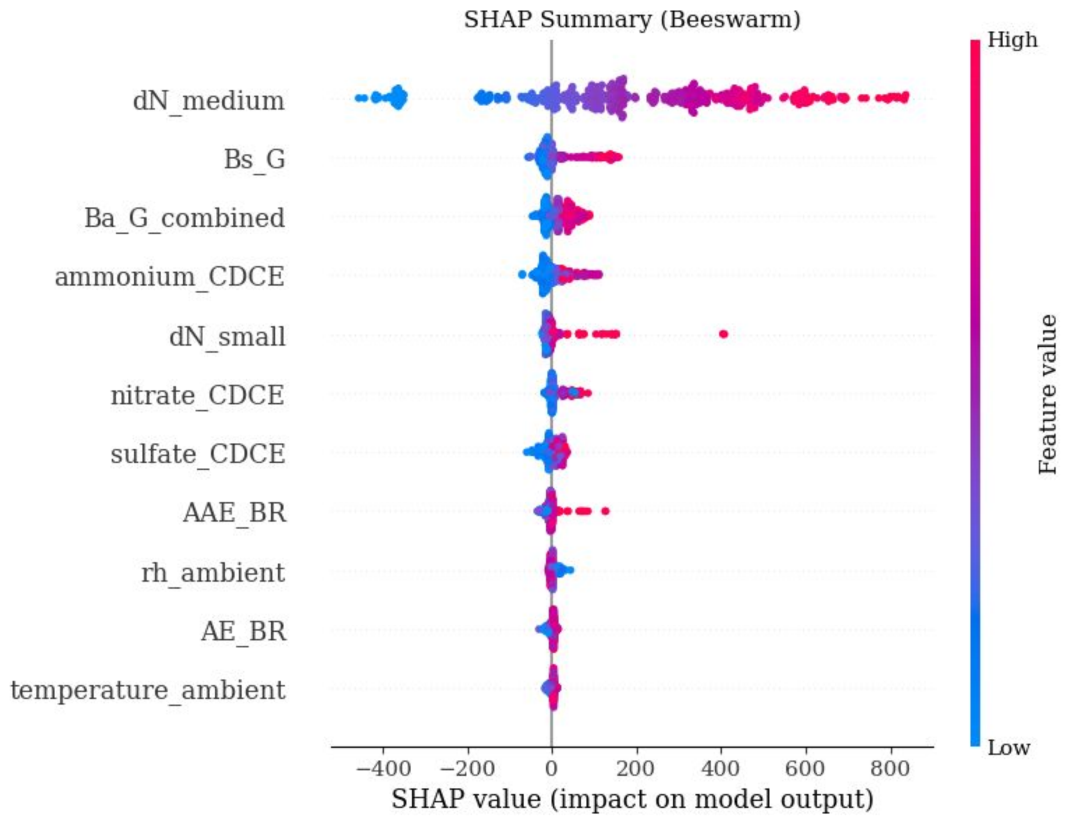
Size distribution:

Medium-sized aerosols contributes most

Expected for chosen intermediate supersaturation level (0.4)

Light-absorbing aerosols are weakly hygroscopic, but aging BC coated with sulfates, nitrates

XAI: SHAP Beeswarm



The SHAP beeswarm plot shows how features influence model predictions using SHAP values.

A SHAP value represents a feature's contribution to increasing or decreasing the prediction.

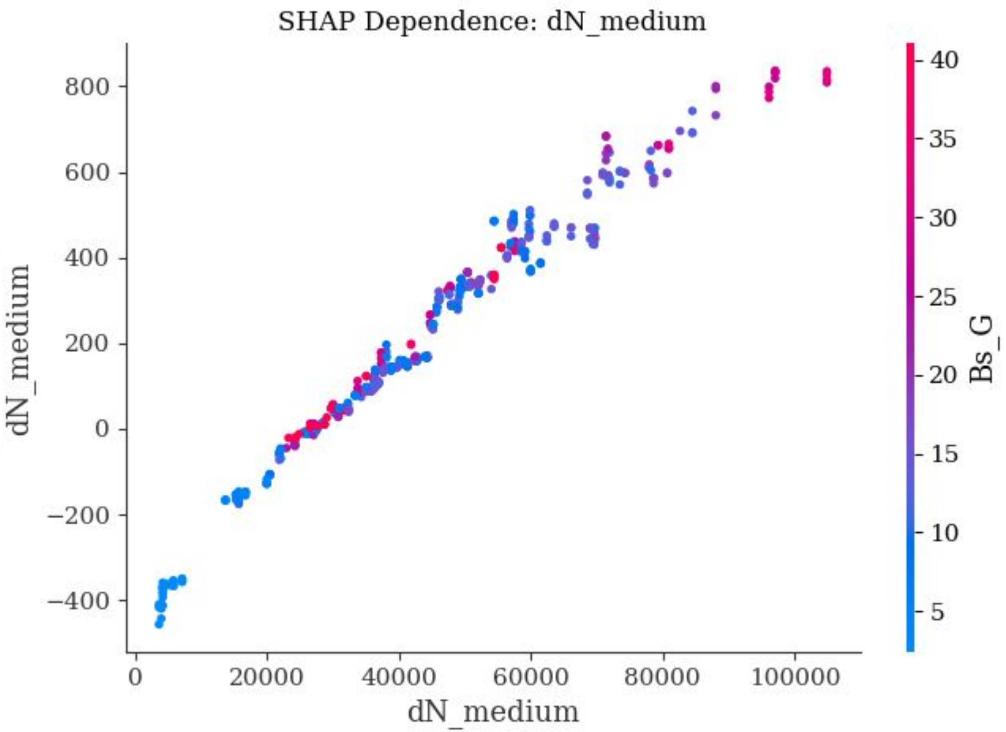
Features farther from zero have stronger effects, while color represents low-to-high feature values.

`dN_medium` has the strongest positive impact, showing high sensitivity to medium-sized aerosol concentration.

`Bs_G` and `Ba_G_combined` indicate additional influence from aerosol scattering and light-absorbing particles.

XAI: SHAP Dependence

SHAP value for



Combined effect of dN_medium and Bs_g on CCN prediction

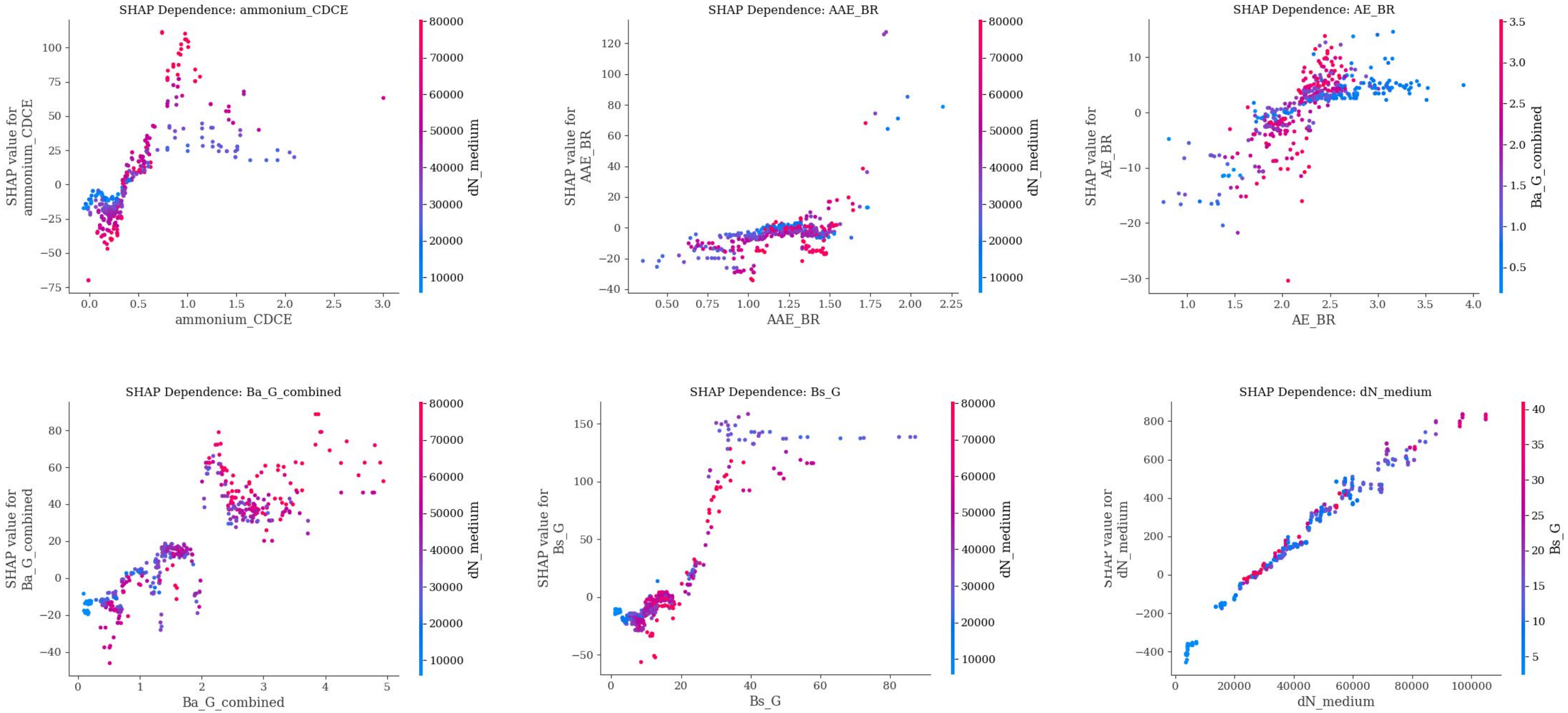
Combination has effect on predicted CCN:

Both low brings down predicted CCN

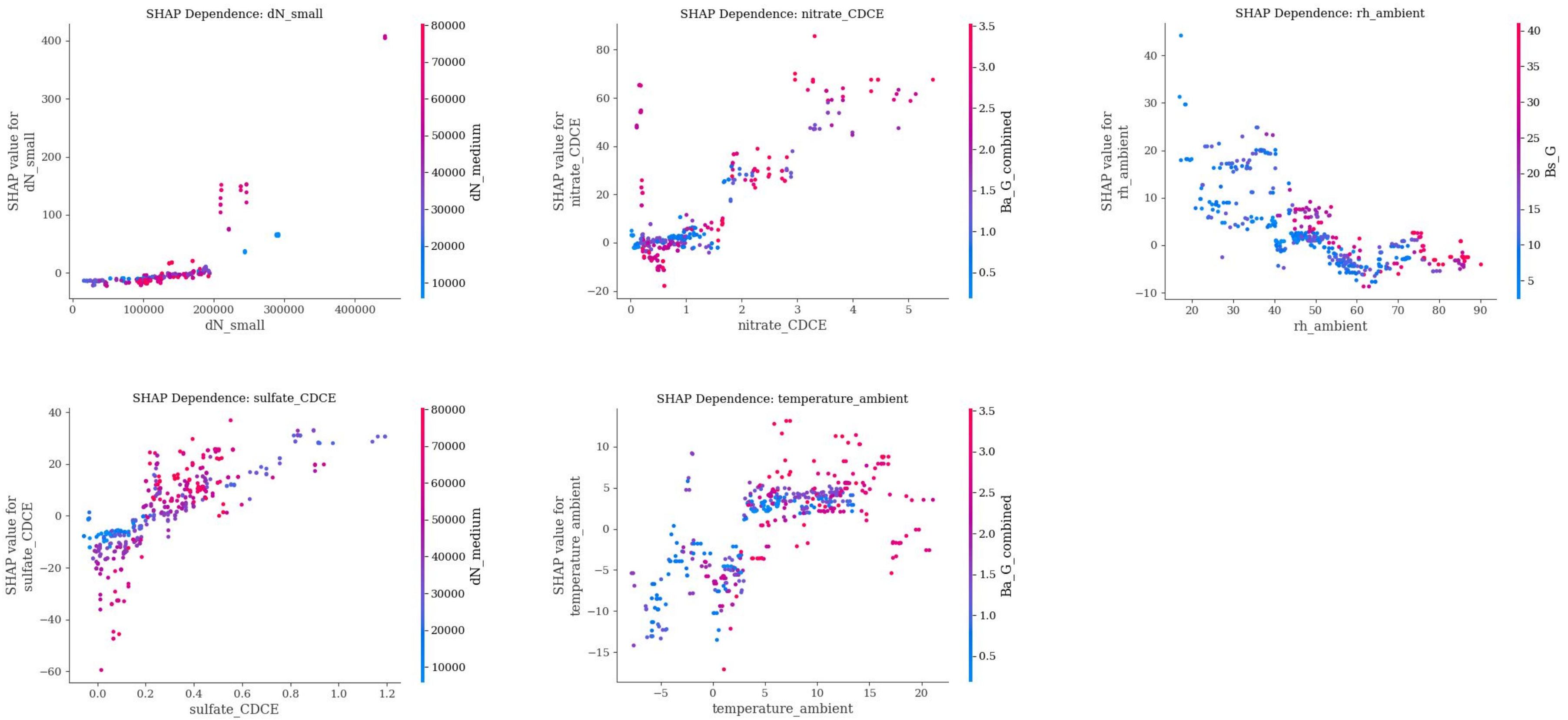
Both very high raises predicted CCN

It shows **how changes in dN_medium affect the model prediction** and whether its effect interacts with another feature (Bs_G).

XAI: SHAP Dependence



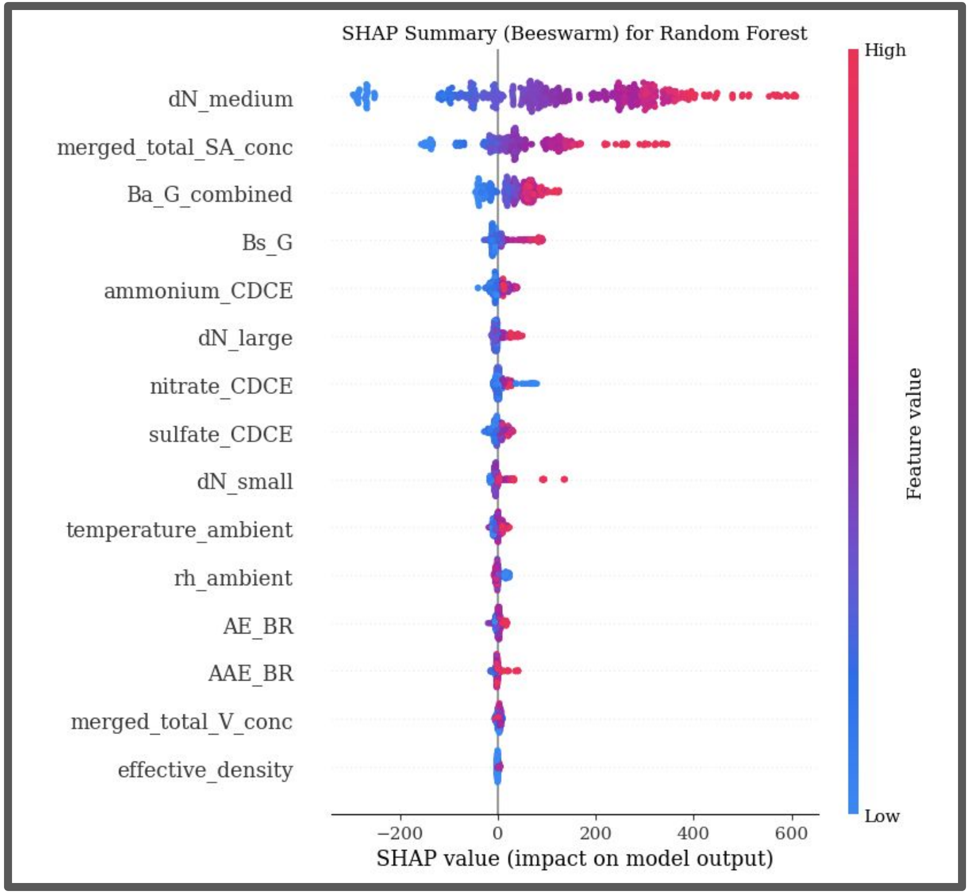
XAI: SHAP Dependence

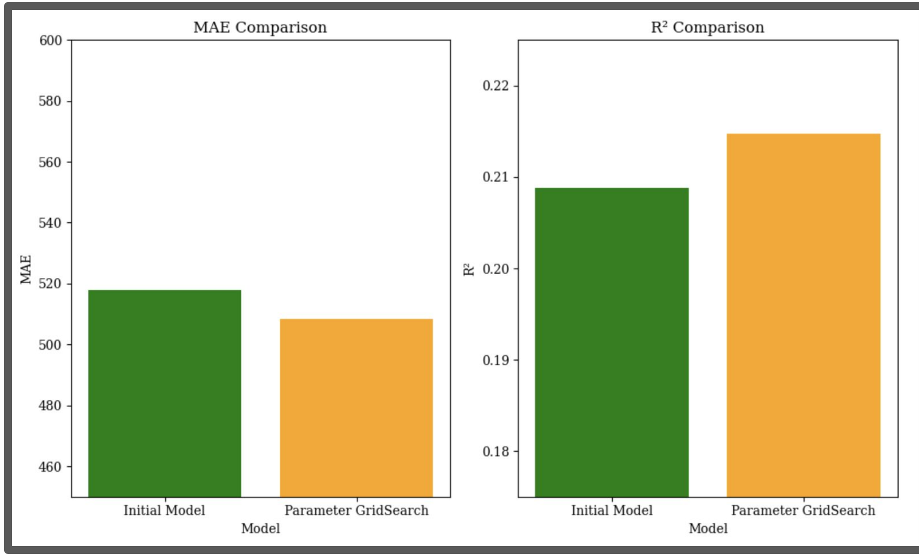


Next Steps….

- More complex ML models
  - Neural networks
  - Probabilistic and/or Distributional
- Selecting a longer time range and exploring seasonality
- Robust selection of the train-test split
- Including temporal memory: Lagged relation

Next Steps: Other ML Methods (e.g. Random Forest)





References

Zabala, I., Casquero-Vera, J. A., Andrews, E., Casans, A., Carrillo-Cardenas, G., Gannet Hallar, A., & Titos, G. (2026). Cloud condensation nuclei phenomenology: predictions based on aerosol chemical and optical properties. Atmospheric Chemistry and Physics, 26(5), 3697-3722. https://doi.org/10.5194/acp-26-3697-2026

Arif, Mohammad, Rajesh Kumar, Ramesh Kumar, Eric Zusman, Ramesh P. Singh, and Akhilesh Gupta. "Assessment of indoor & outdoor black carbon emissions in rural areas of Indo-Gangetic Plain: Seasonal characteristics, source apportionment and radiative forcing." Atmospheric Environment 191 (2018): 227-240. https://doi.org/10.1016/j.atmosenv.2018.07.057

Koontz, A., Uin, J., Andrews, E., Enekwizu, O., Hayes, C., & Salwen, C. Cloud Condensation Nuclei Particle Counter (AOSCCN2COLB), 2019-11-01 to 2019-12-31, Southern Great Plains (SGP), Lamont, OK (Extended and Co-located with C1) (E13). Atmospheric Radiation Measurement (ARM) User Facility. https://doi.org/10.5439/1323893

Koontz, A., Uin, J., Andrews, E., Enekwizu, O., Hayes, C., & Salwen, C. Cloud Condensation Nuclei Particle Counter (AOSCCN2COLAAVG), 2019-11-01 to 2019-12-31, Southern Great Plains (SGP), Lamont, OK (Extended and Co-located with C1) (E13). Atmospheric Radiation Measurement (ARM) User Facility. https://doi.org/10.5439/1323894

Koontz, A., Flynn, C., Shilling, J., & Flynn, C. Aerosol Optical Properties (AOPPSAP1FLYNN1M), 2019-11-01 to 2019-12-31, Southern Great Plains (SGP), Lamont, OK (Extended and Co-located with C1) (E13). Atmospheric Radiation Measurement (ARM) User Facility. https://doi.org/10.5439/1369240

Koontz, A., Flynn, C., & Shilling, J. Aerosol Optical Depth (AOD) derived from MFRSR measurements (MFRSRAOD1MICH), 2019-11-01 to 2019-12-31, Southern Great Plains (SGP), Lamont, OK (Extended and Co-located with C1) (E13). Atmospheric Radiation Measurement (ARM) User Facility. https://doi.org/10.5439/1356805

Kyrouac, J., Springston, S., & Tuftedal, M. Meteorological Measurements associated with the Aerosol Observing System (AOSMET), 2019-11-01 to 2019-12-31, Southern Great Plains (SGP), Lamont, OK (Extended and Co-located with C1) (E13). Atmospheric Radiation Measurement (ARM) User Facility. https://doi.org/10.5439/1984920

Shilling, J., & Levin, M. merged size distribution from SMPS and APS (MERGEDSMPSAPS), 2019-11-01 to 2019-12-31, Southern Great Plains (SGP), Lamont, OK (Extended and Co-located with C1) (E13). Atmospheric Radiation Measurement (ARM) User Facility. https://doi.org/10.5439/1871375

Zawadowicz, M., Howie, J., Shilling, J., & Levin, M. ACSM, corrected for composition-dependent collection efficiency (ACSMCDCE), 2019-11-01 to 2019-12-31, Southern Great Plains (SGP), Lamont, OK (Extended and Co-located with C1) (E13). Atmospheric Radiation Measurement (ARM) User Facility. https://doi.org/10.5439/1763029

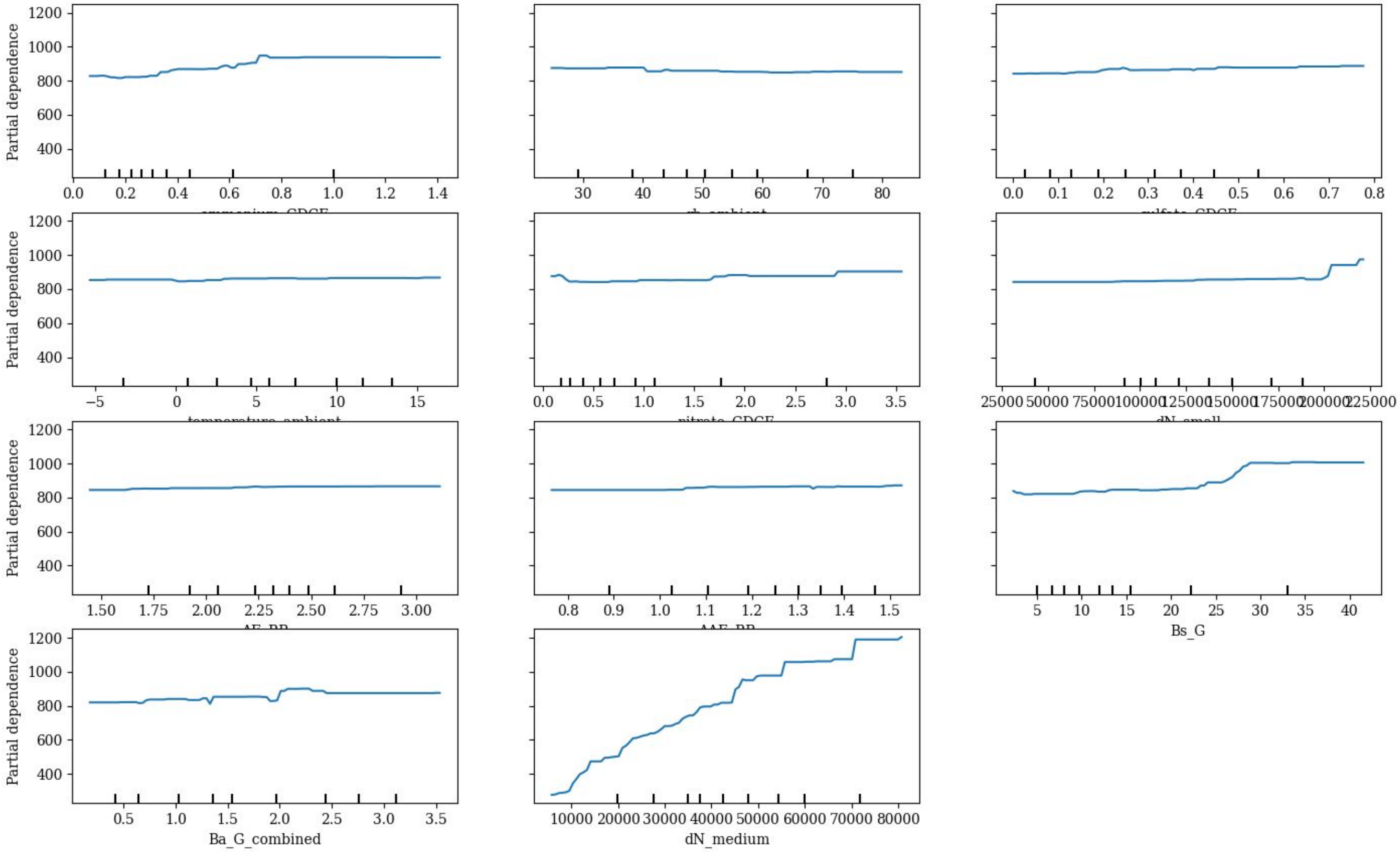
Zhang, D., Ermold, B., & Morris, V. Ceilometer (CEIL), 2019-11-01 to 2019-12-31, Southern Great Plains (SGP), Central Facility, Lamont, OK (C1). Atmospheric Radiation Measurement (ARM) User Facility. https://doi.org/10.5439/1181954

THANK YOU!

Credit: Scott Collis

Supplementary Slides

Partial Dependence Plots



	Variable_1	Variable_2	Correlation	Abs_Correlation
39	dN_large	Bs_G	0.889043	0.889043
17	dN_medium	merged_total_SA_conc	0.836435	0.836435
32	dN_large	merged_total_SA_conc	0.811940	0.811940
97	nitrate_CDCE	ammonium_CDCE	0.810564	0.810564
99	nitrate_CDCE	Bs_G	0.768200	0.768200
54	merged_total_SA_conc	Bs_G	0.760737	0.760737
35	dN_large	nitrate_CDCE	0.732934	0.732934
129	ammonium_CDCE	Bs_G	0.707849	0.707849
37	dN_large	ammonium_CDCE	0.707728	0.707728