PurpleAir $PM_{2.5}$ performance across the U.S.

Karoline (Johnson) Barkjohn[1,2], Brett Gantt[3], Ian VonWald[1,2], Andrea L. Clements[2]

[1]ORISE fellow hosted by EPA Office of Research and Development

[2]EPA Office of Research and Development, Center for Environmental Measurement & Modeling, Air Methods & Characterization Division, Source & Fine Scale Branch, Research Triangle Park, NC

[3]EPA Office of Air Quality Planning and Standards, Air Quality Assessment Division, Air Quality Analysis Group, Research Triangle Park, NC

Webinar with project partners

December 9th, 2019

Introductions

Karoline (Johnson) Barkjohn, Ph.D.

ORISE Post-doc

johnson.karoline@epa.gov

Brett Gantt, Ph.D.

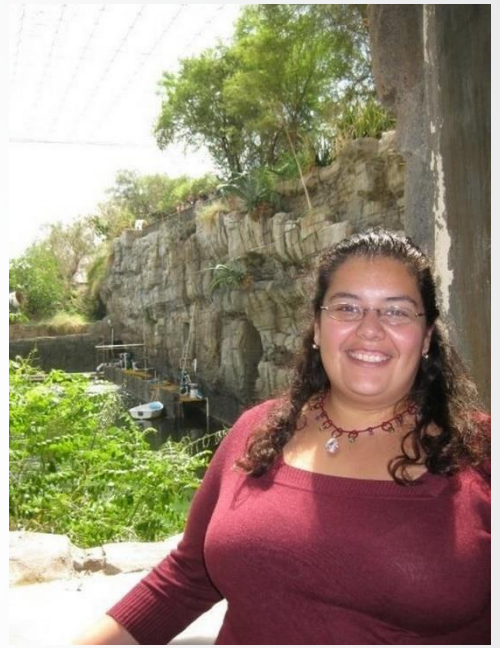
EPA OAQPS

Gantt.Brett@epa.gov

Ian VonWald, Ph.D.

ORISE Post-doc

vonwald.ian@epa.gov



Andrea Clements, Ph.D.

EPA ORD

clements.andrea@epa.gov

Overview

- Background
- Goals
- Methods
  - Dataset
  - Quality assurance methods
  - Correction methods
- Results
  - Error
  - Reported Air Quality Index (AQI) values
  - Results by state
- Conclusions
- Next Steps

PurpleAir Map

- Cost: ~$250
- 1,000's deployed worldwide

purpleair.com/map

Goal

To develop a correction that can improve PurpleAir $PM_{2.5}$ measurements across the U.S.

- To be used by PurpleAir and/or various users (local agencies, community groups, exposure scientists)

Balance:

- Broad applicability
- Simplicity of use
- Best correction method

Dataset

All sensors collocated by air monitoring agencies (Thank you!)

- Communicating with all partner agencies to ensure collocation and to glean additional details on sensor siting
- **Note:** This removes some of the uncertainty associated with using PurpleAir data as usually you cannot confirm they are correctly located on the map

24-hour averaged Federal Equivalent Method (FEM) and Federal Reference Method (FRM) data downloaded from Air Quality System (AQS)

Previous work:

- typically small number of sensors (~<10) in a single region
- some studies not collocated

Collocation sites

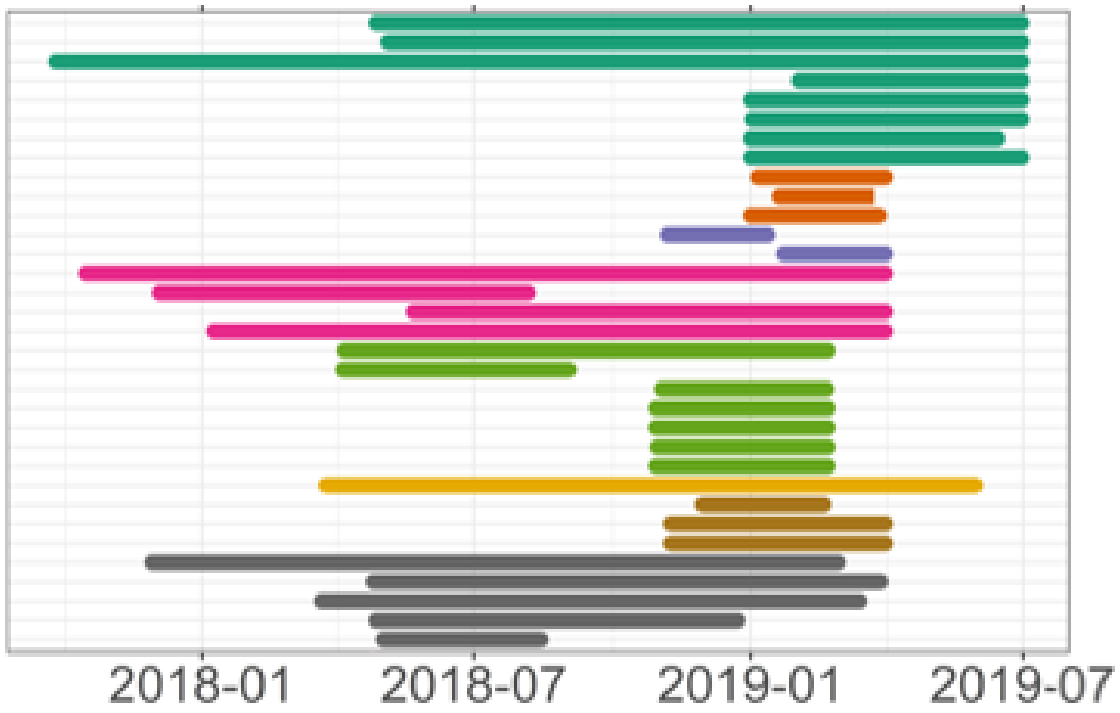
Phase 1: Sites included in this presentation
Phase 2: Sites added soon

Batchgeo.com

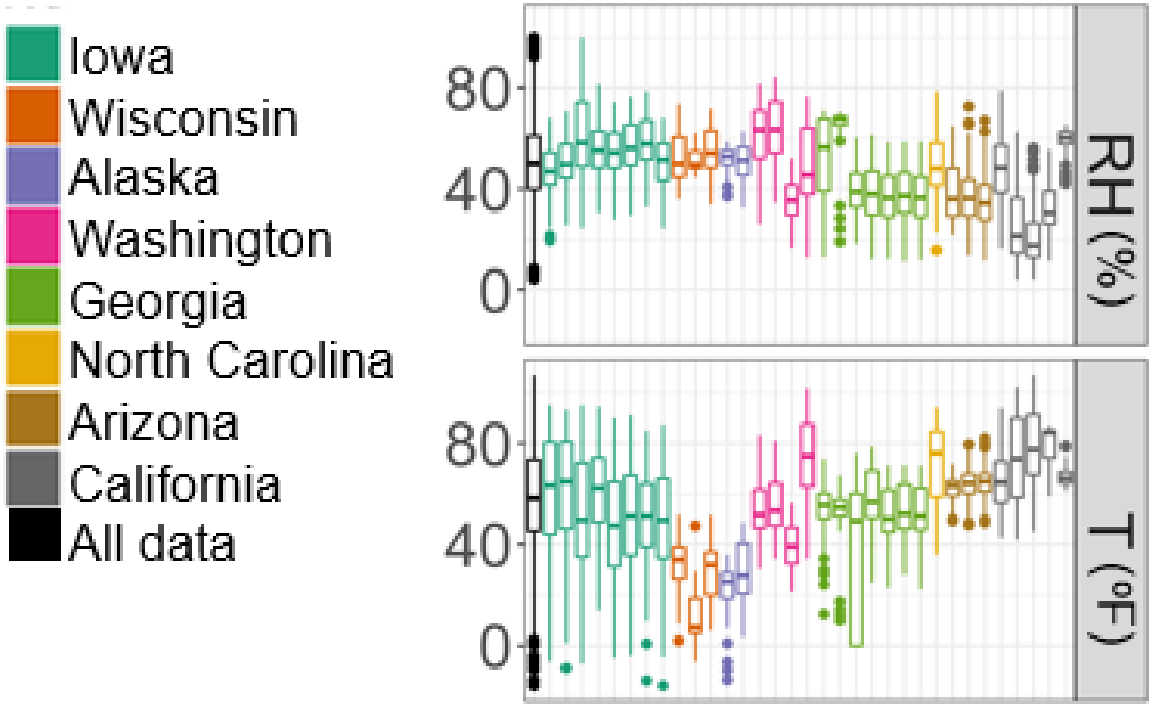
Notes about the dataset

- Primarily online but a few offline
- Public and private sensors
- Operating for different time periods and lengths of time

Dataset time periods



Sensors operated over different time periods from a few months to almost 2 years



24-hr temperature (T) and relative humidity (RH) across sites as measured by the PurpleAir sensors. Some sites ran for a single season resulting in T & RH not representative of the typical annual climate

Cf=1 and Cf=atm corrections

PurpleAir provides PM data with two corrections

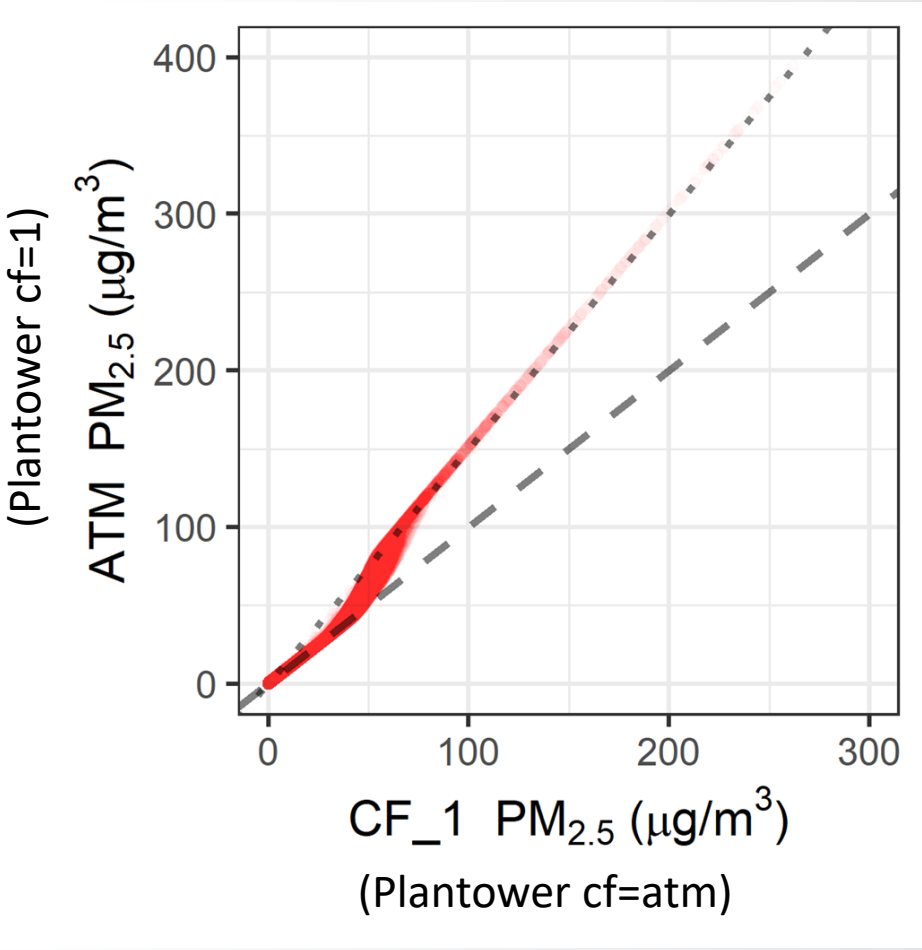
- Displayed on the map: cf=1
  - Plantower cf=atm output
- Also: cf=atm
  - Plantower cf=1 output

Note: since flipped labels were identified, the column names have been flipped in the sensor list output

- A fix may be coming to the other datasets soon

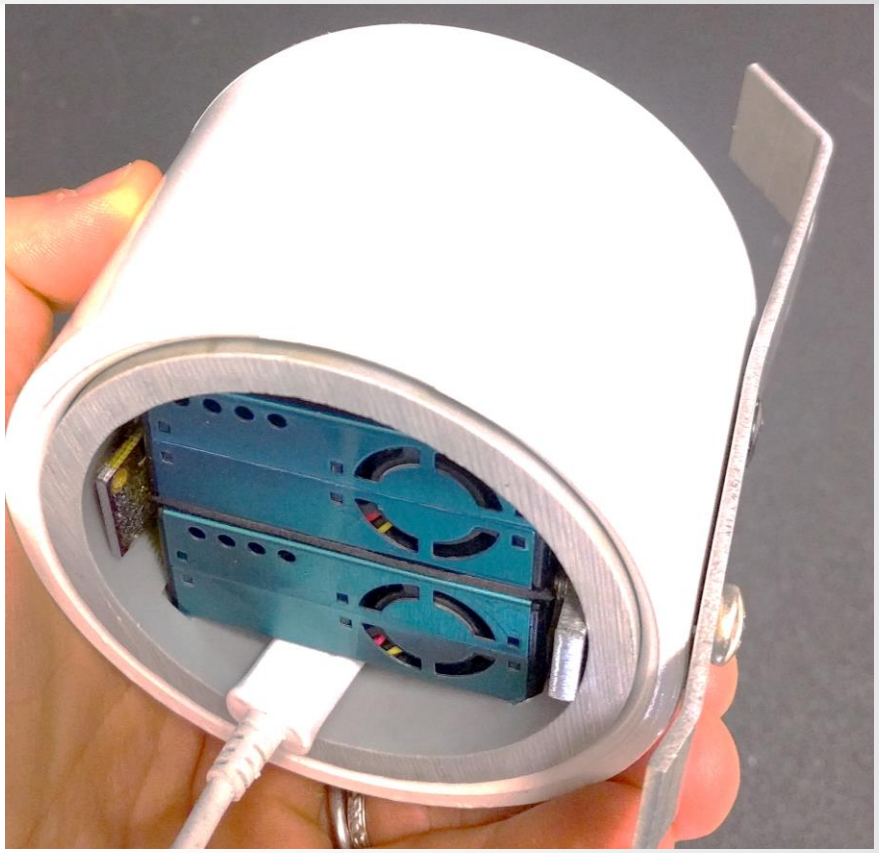
PurpleAir cf=atm selected for this work

- cf=atm R^2=0.65 > cf=1 R^2=0.64



QA: A & B channels

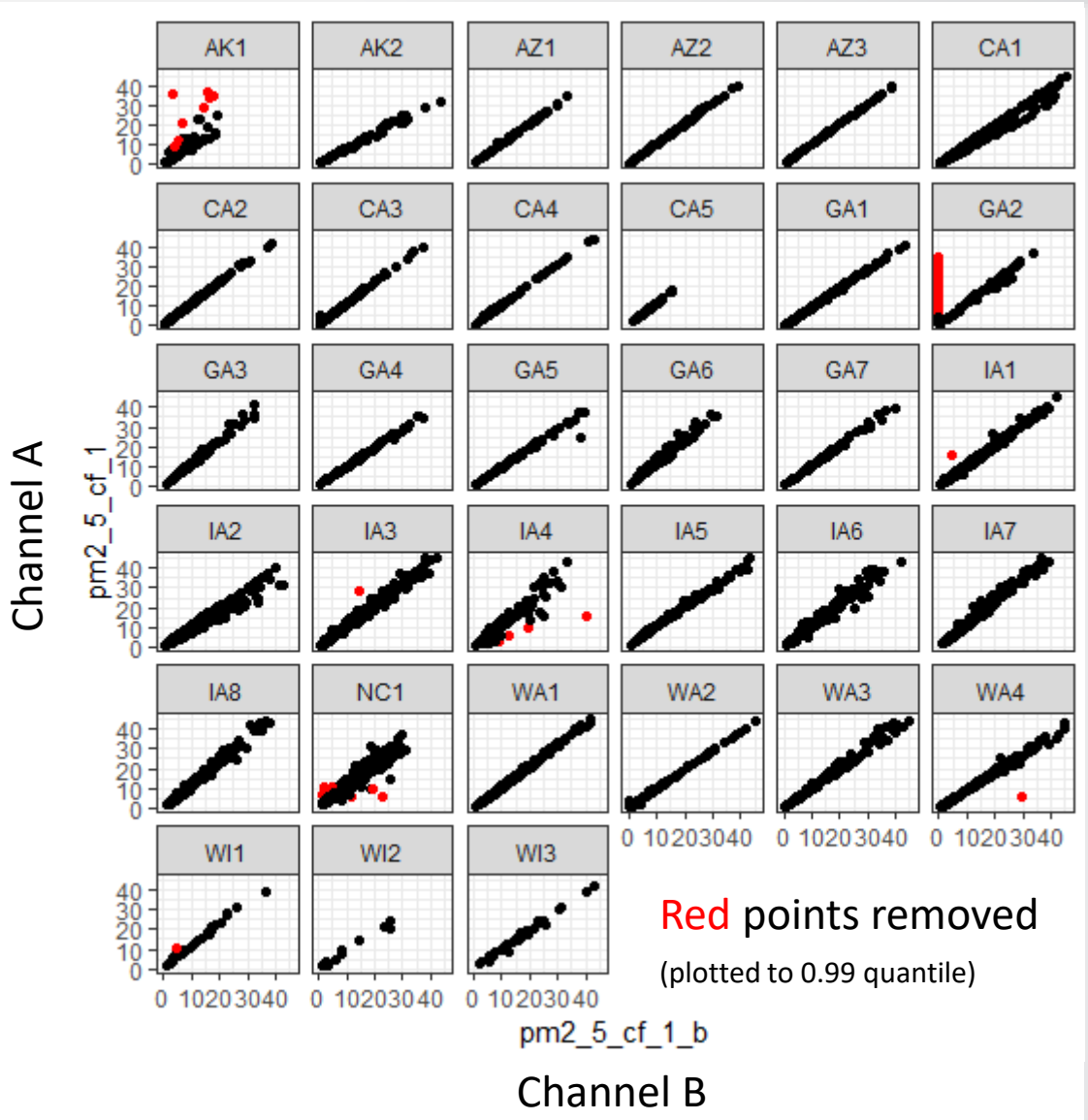
- 2 Plantower PMS5003 (channels A & B)
- Sample for alternating 10-second intervals
- Generate 2-minute averages
  - previously 80-second



PurpleAir underside view

QA steps

- Only 24-hr averages with at least 90% of the data present were used
- Points removed if 24-hr averaged A & B $PM_{2.5}$ differ by:
  - $\geq \pm 5$ µg m^{-3} AND
  - $\geq \pm$ 16 %
    - 2*sd(% error)
  - 2% removed
- A & B channels averaged
  - Slightly improves overall R^2 and confidence on linear regression
  - More important for individual sensors



Equations considered

1. **Raw:** $PM_{2.5}$ = PA (raw PurpleAir $PM_{2.5}$ cf=atm)
2. **Linear:** $PM_{2.5}$ = a*PA + b
3. **T & RH:** $PM_{2.5}$ = a*PA + b*T + c*RH + d

Units:
$PM_{2.5}$= µg m^{-3}
T=ºF
RH=%

Other parameters considered:

- Pressure (not available for all sensors)
- Other T & RH functions
  - Selected equation is more broadly applicable than T*RH*PM
  - Strong nonlinear RH influence not seen in most states
- Size bin data: minimal improvements
- Duration deployed: More data needed from sensors operating ≥ 1 year
- State: Seasonal influences may dominate since <1 year of data in many states

Equations considered

1. **Raw:** $PM_{2.5}$ = PA (raw PurpleAir $PM_{2.5}$ cf=atm)

2. **Linear:** $PM_{2.5} = 0.38*PA + 2.94$, $R^2=0.69$

3. **T & RH:** $PM_{2.5} = 0.39*PA +0.0024*T -0.050*RH + 5.19$, $R^2=0.72$

Units:

$PM_{2.5}$= µg m^{-3}

T=ºF

RH=%

Statistics used for evaluation

Mean Bias Error (MBE): Accuracy

- The average difference between the PurpleAir and FEM or FRM measurements

Root Mean Square Error (RMSE): Precision

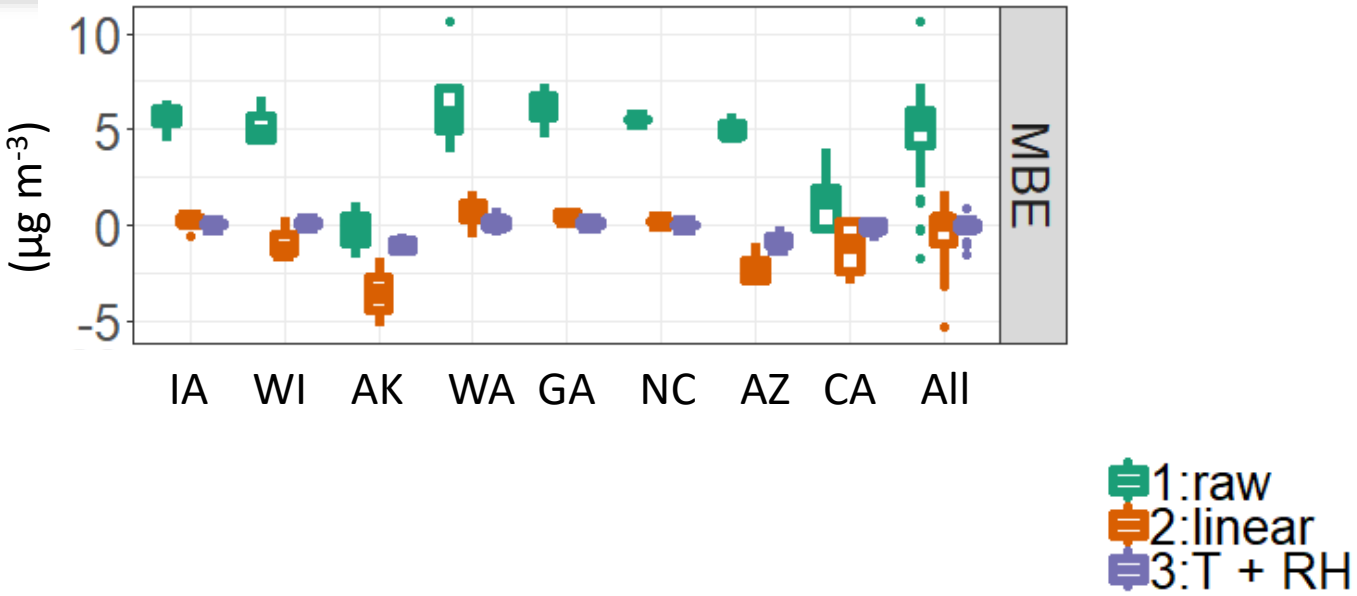
- A measure of the spread between the 24-hr PurpleAir and FEM or FRM measurements
- Since it is squared it penalizes outliers

Spearman correlation: Measure of association

- Similar to pearson correlation or R^2 but for non normally distributed datasets

Results

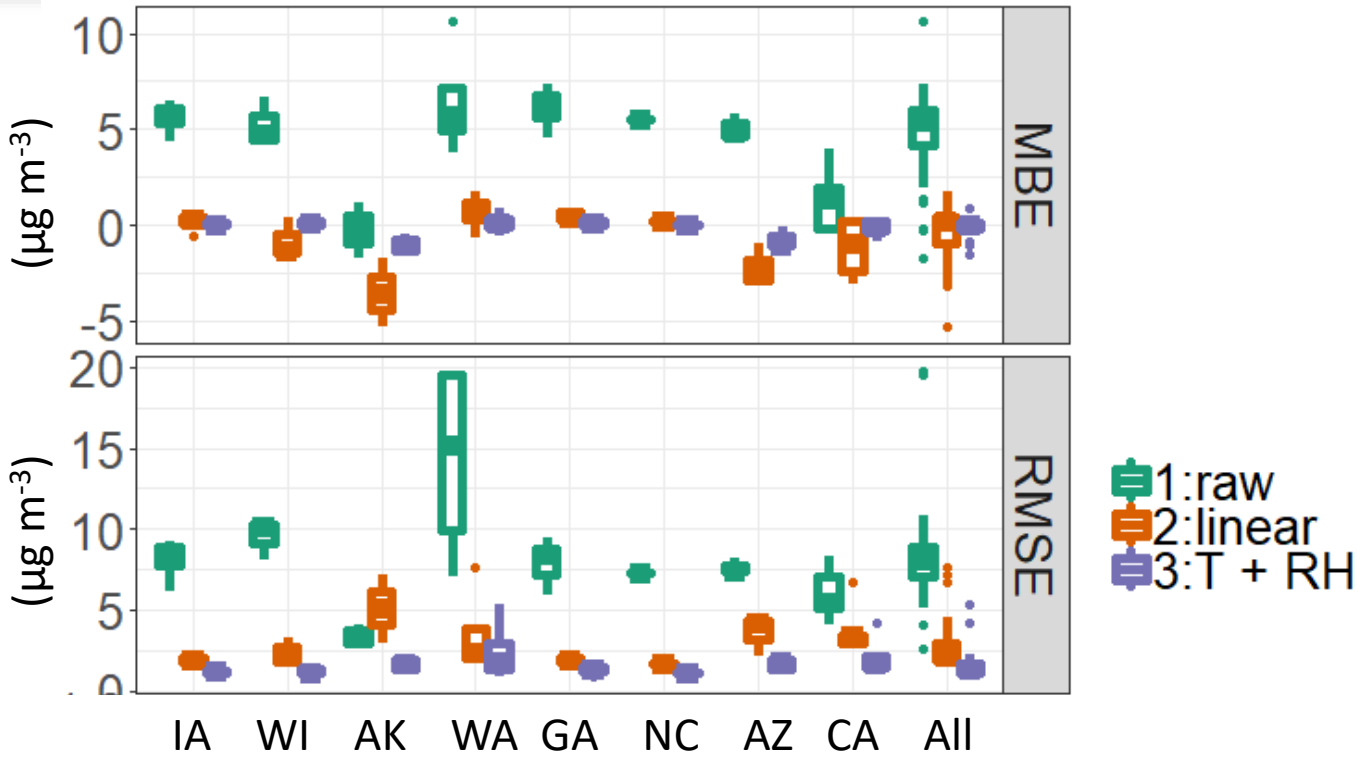
- Linear correction reduces bias (MBE) **except**:
  - AK
- T + RH correction reduces state bias in all states **especially**:
  - AK, CA, AZ



Each point on the boxplot is a single PurpleAir sensor

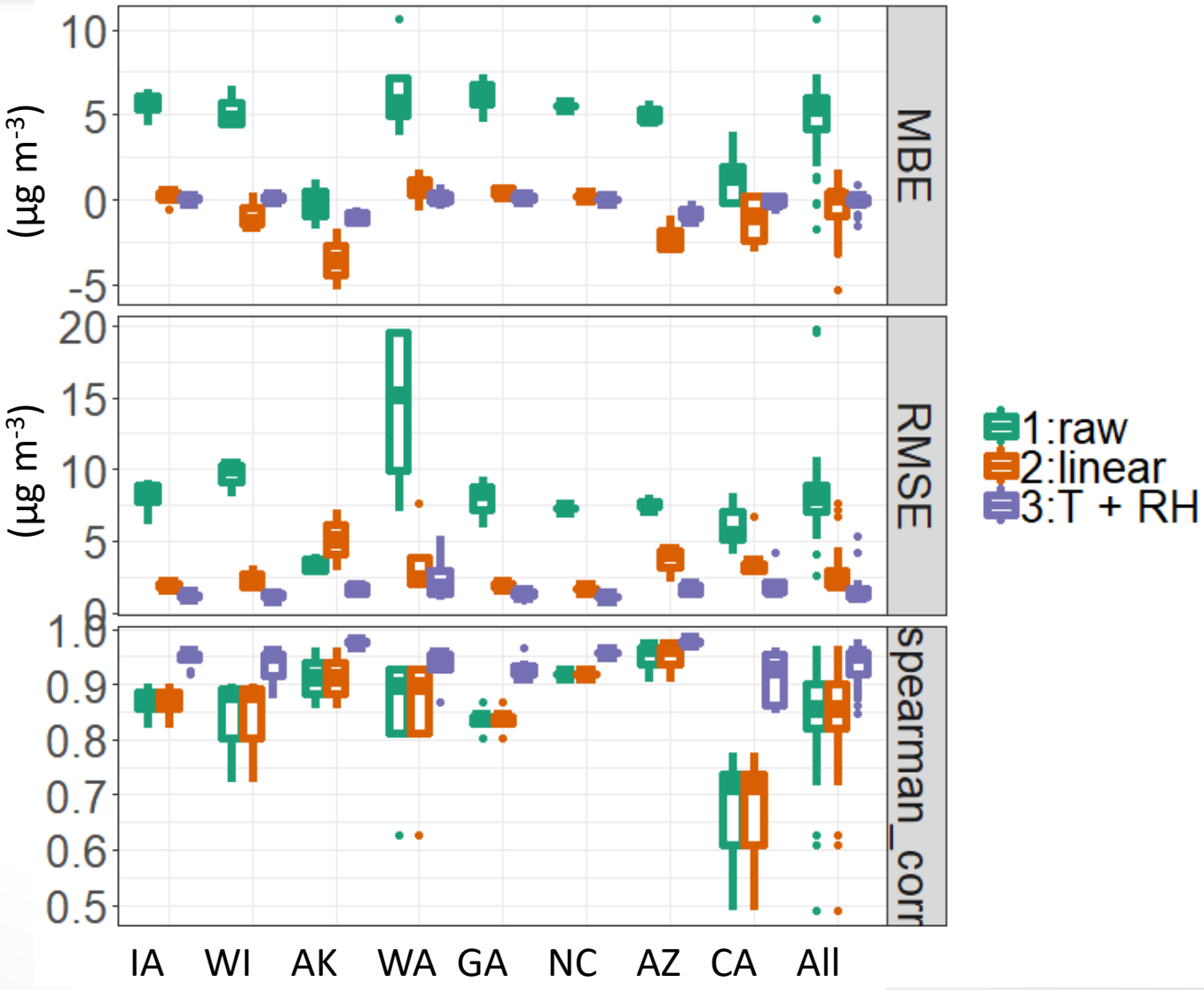
Results

- Linear correction reduces RMSE except:
  - AK
- T + RH correction reduces RMSE in all states

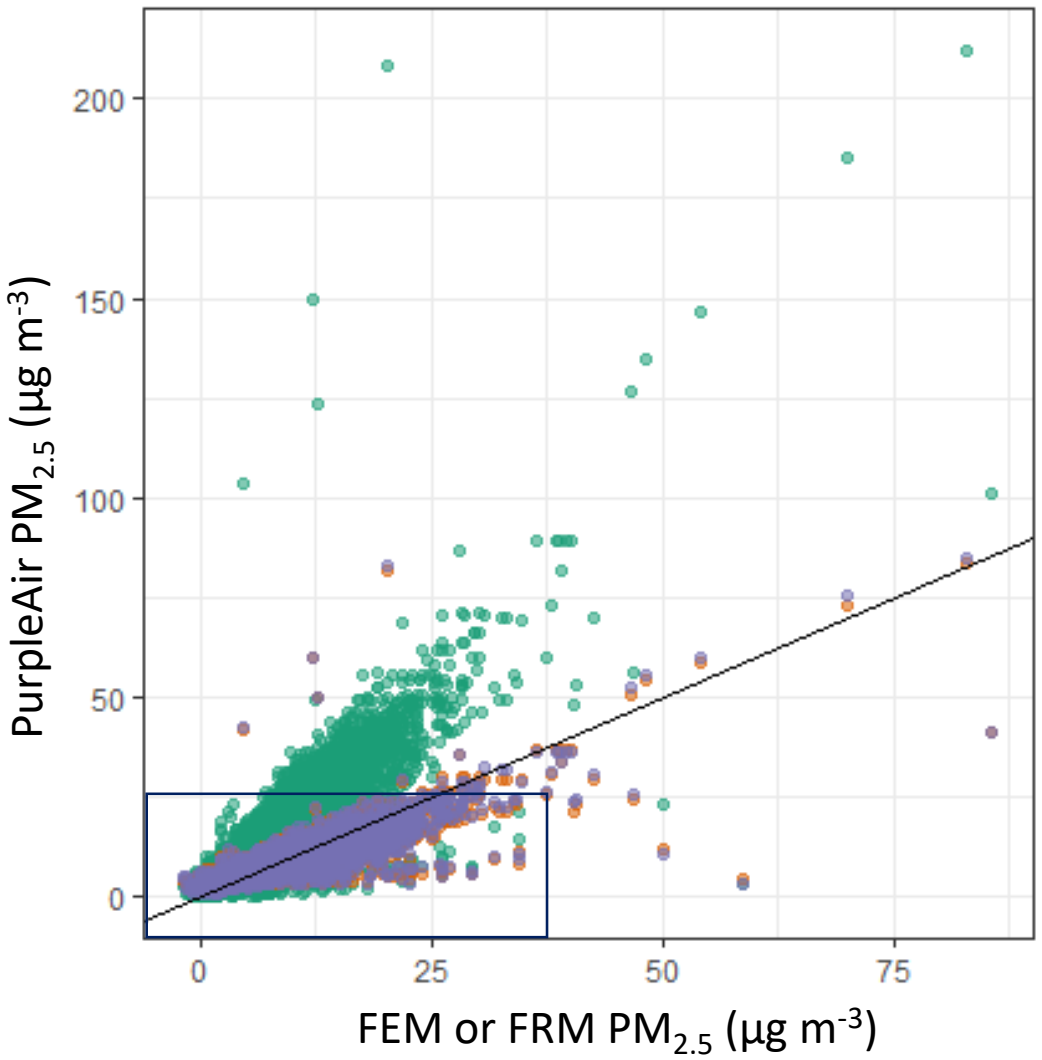


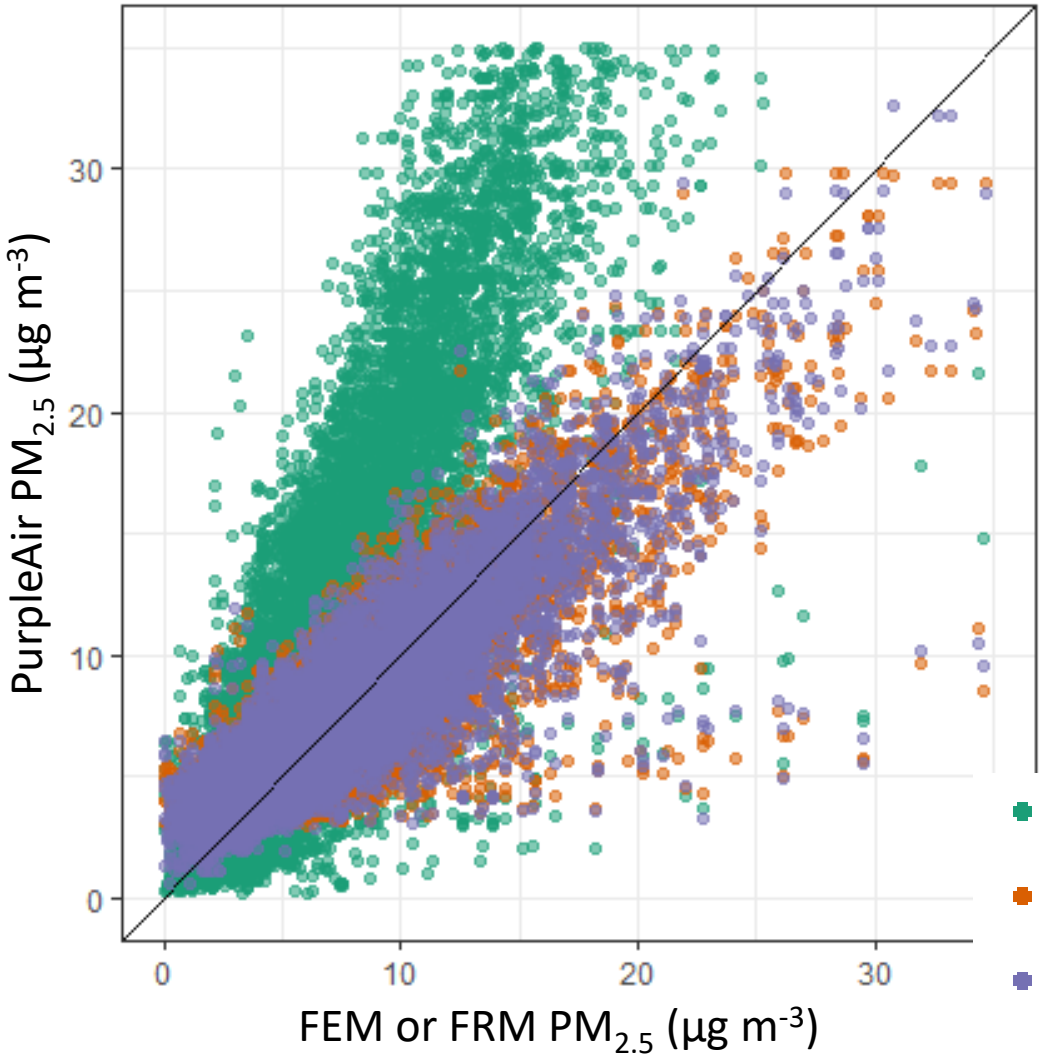
Results

- Spearman correlation improved using T + RH correction in all states



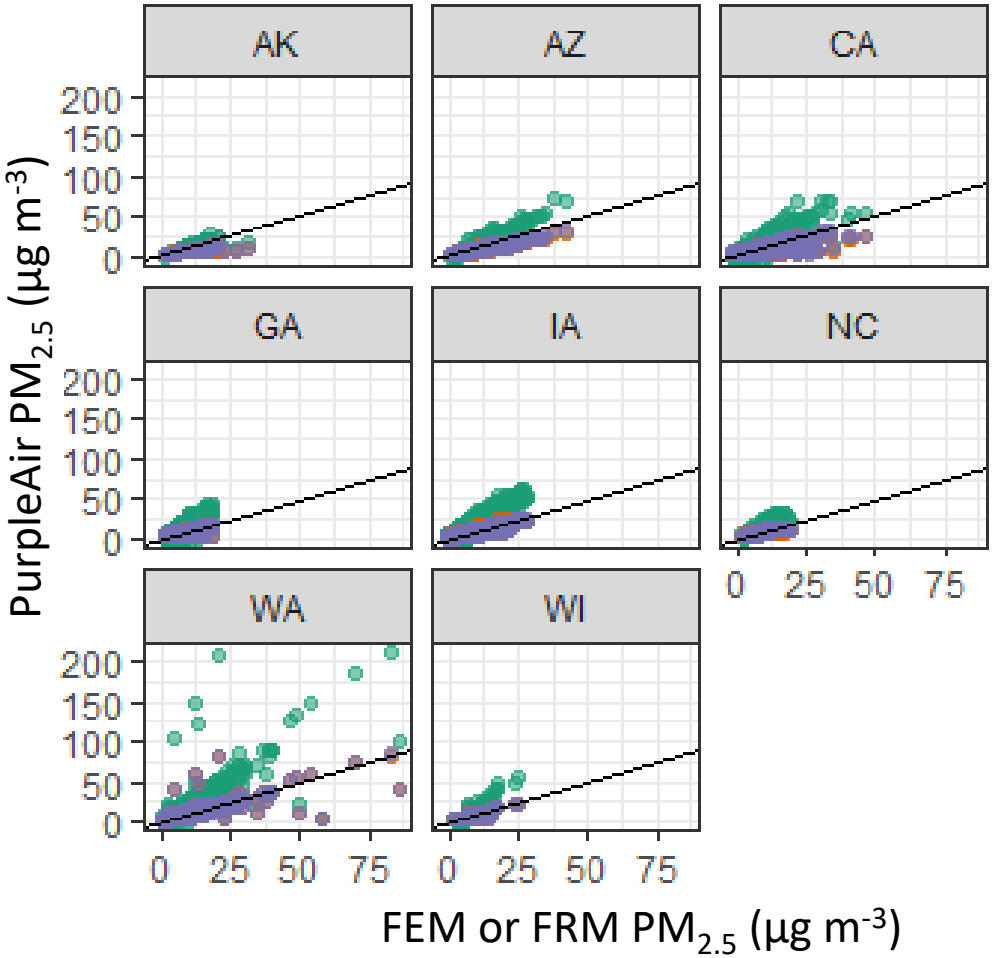
Results

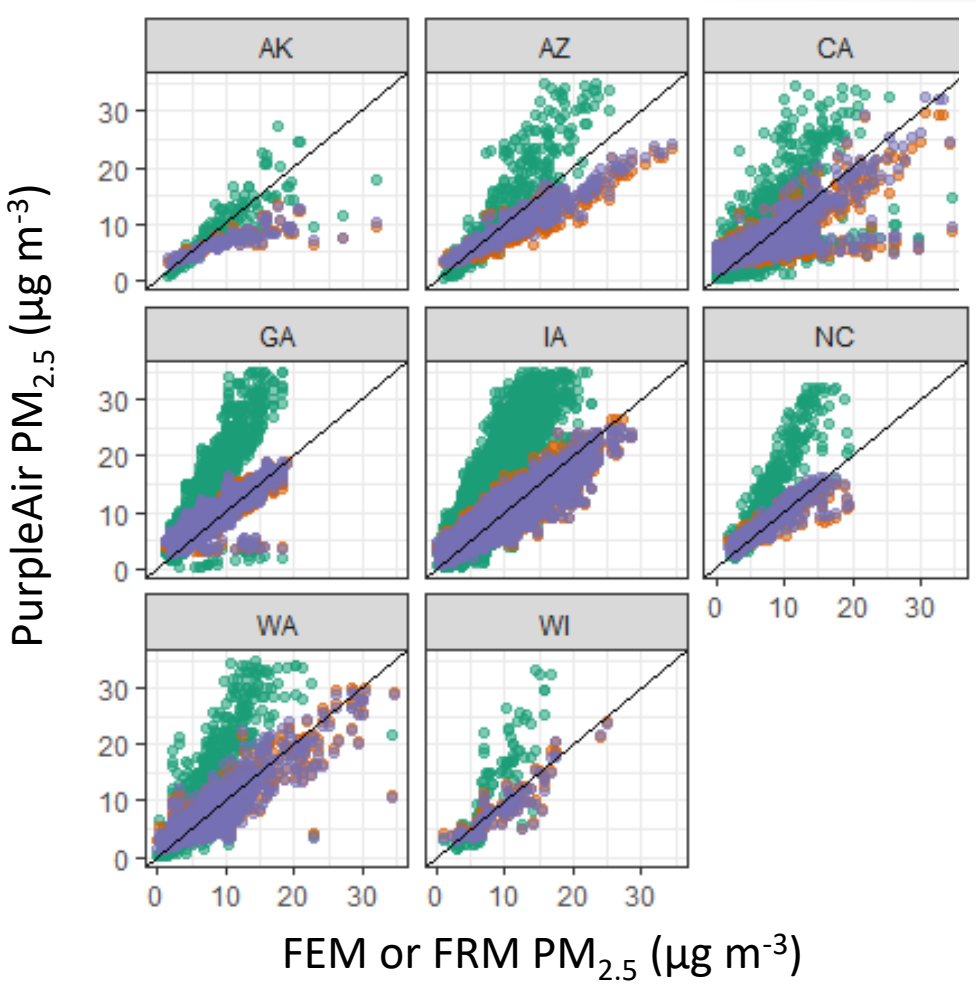




- Raw (cf=atm)
- Linear
- T+RH

Results by state



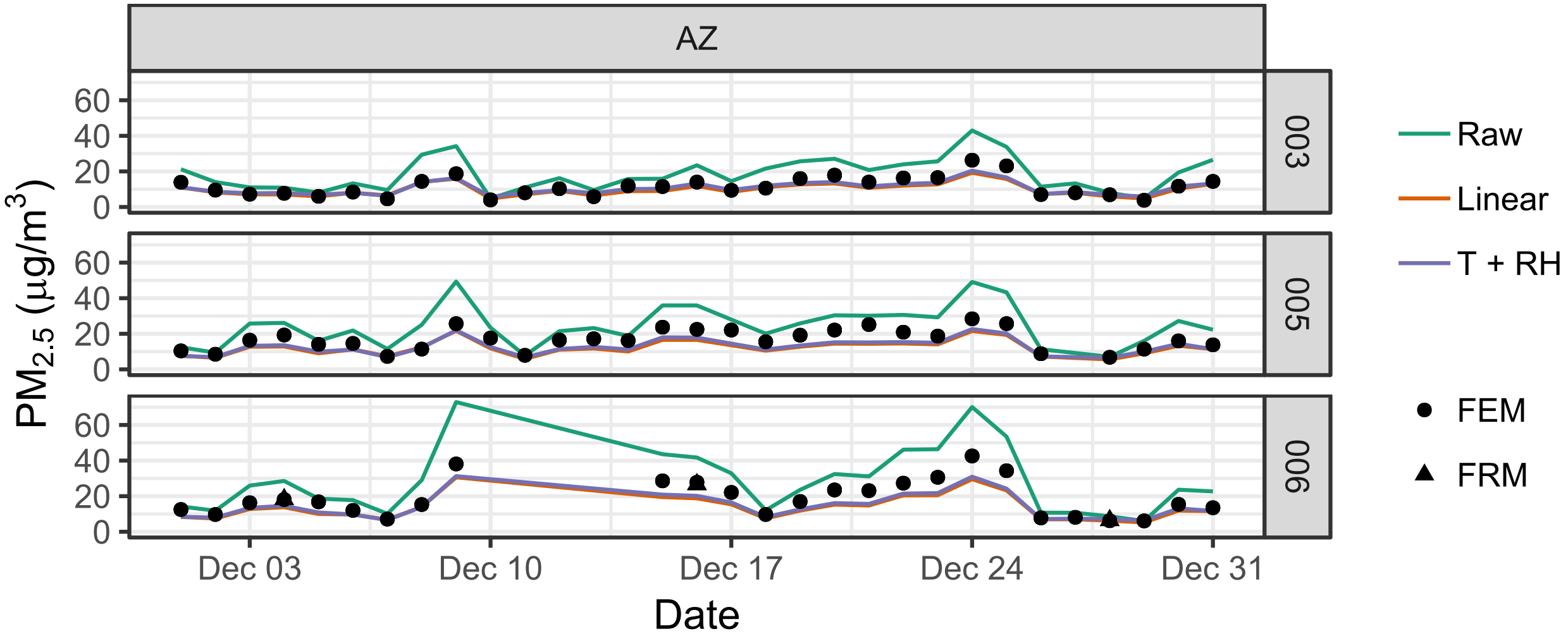


- Raw (cf=atm)
- Linear
- T+RH

- Linear and T+RH very similar
- Does not capture higher concentrations in AK and AZ
- Captures higher concentrations in WA
- 2 prong distribution in CA and GA

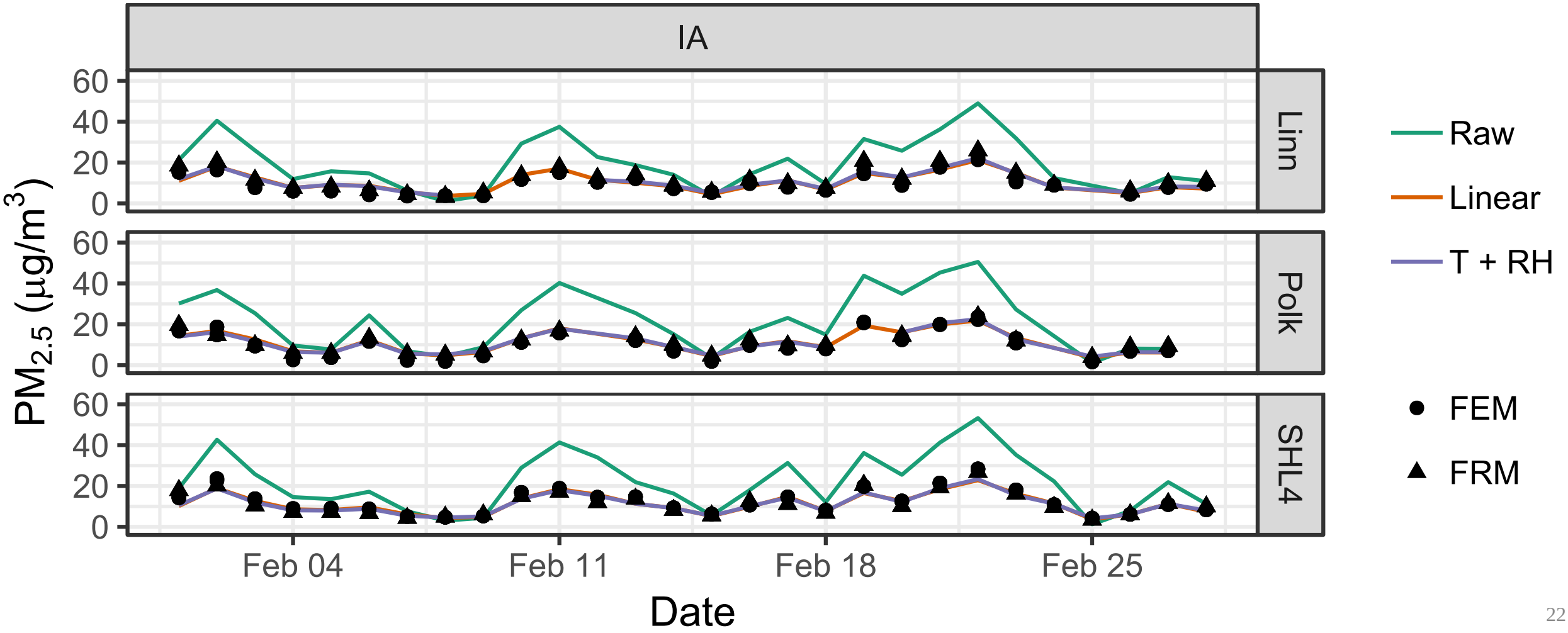
Arizona

3 example sites from Phoenix Arizona



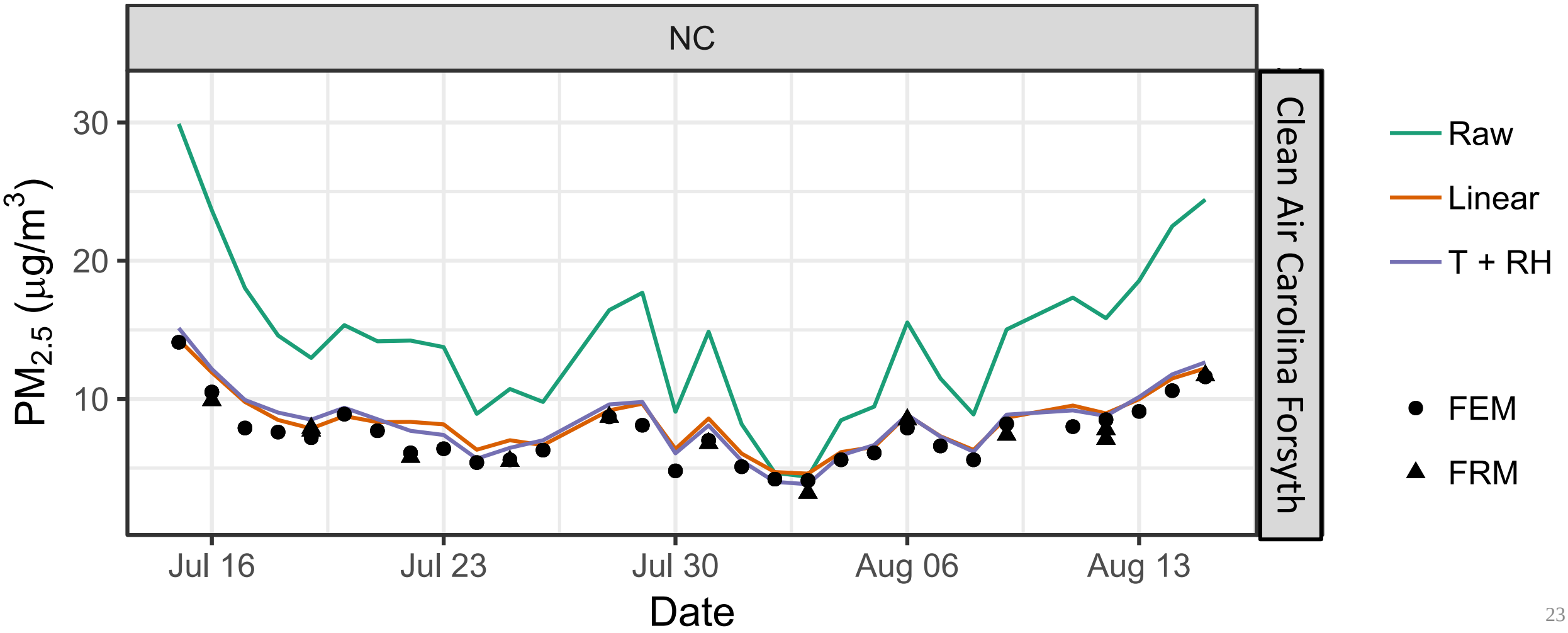
Iowa

3 example sites from Iowa

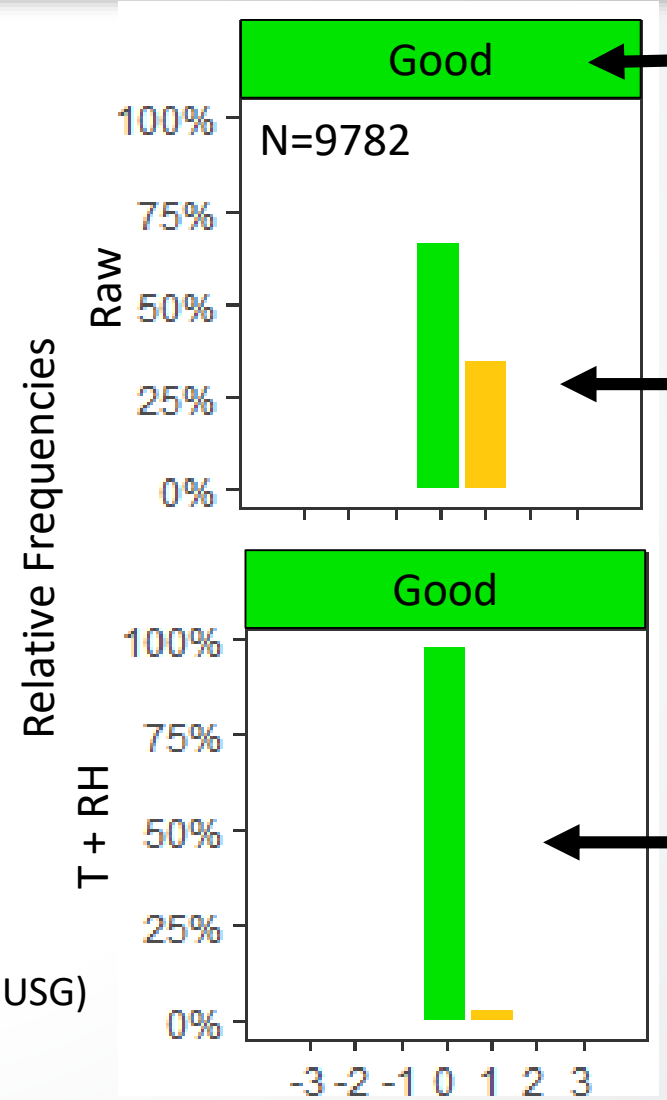


North Carolina

example site from Forsyth County, North Carolina



AQI reporting: Good-Moderate



While the FRM/FEM reports

The PurpleAir reports

After correction the PurpleAir reports

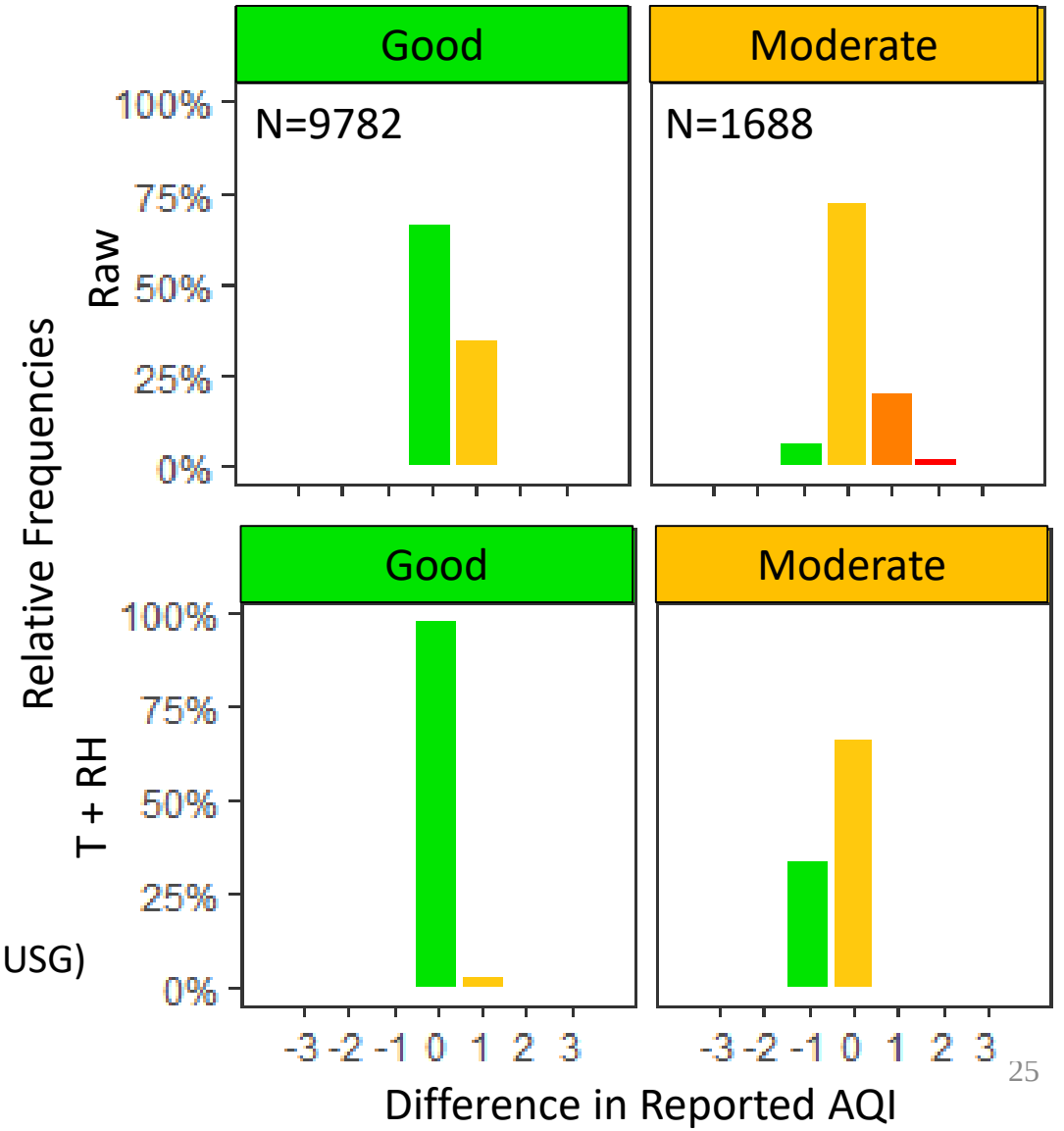
- Good
- Moderate
- Unhealthy for sensitive groups (USG)
- Unhealthy
- Very unhealthy

AQI reporting: Good-Moderate

Raw PurpleAir often over reports AQI

- This is important to note as this is what is being displayed and viewed by many of the PurpleAir map

Correcting causes under reporting at moderate AQI

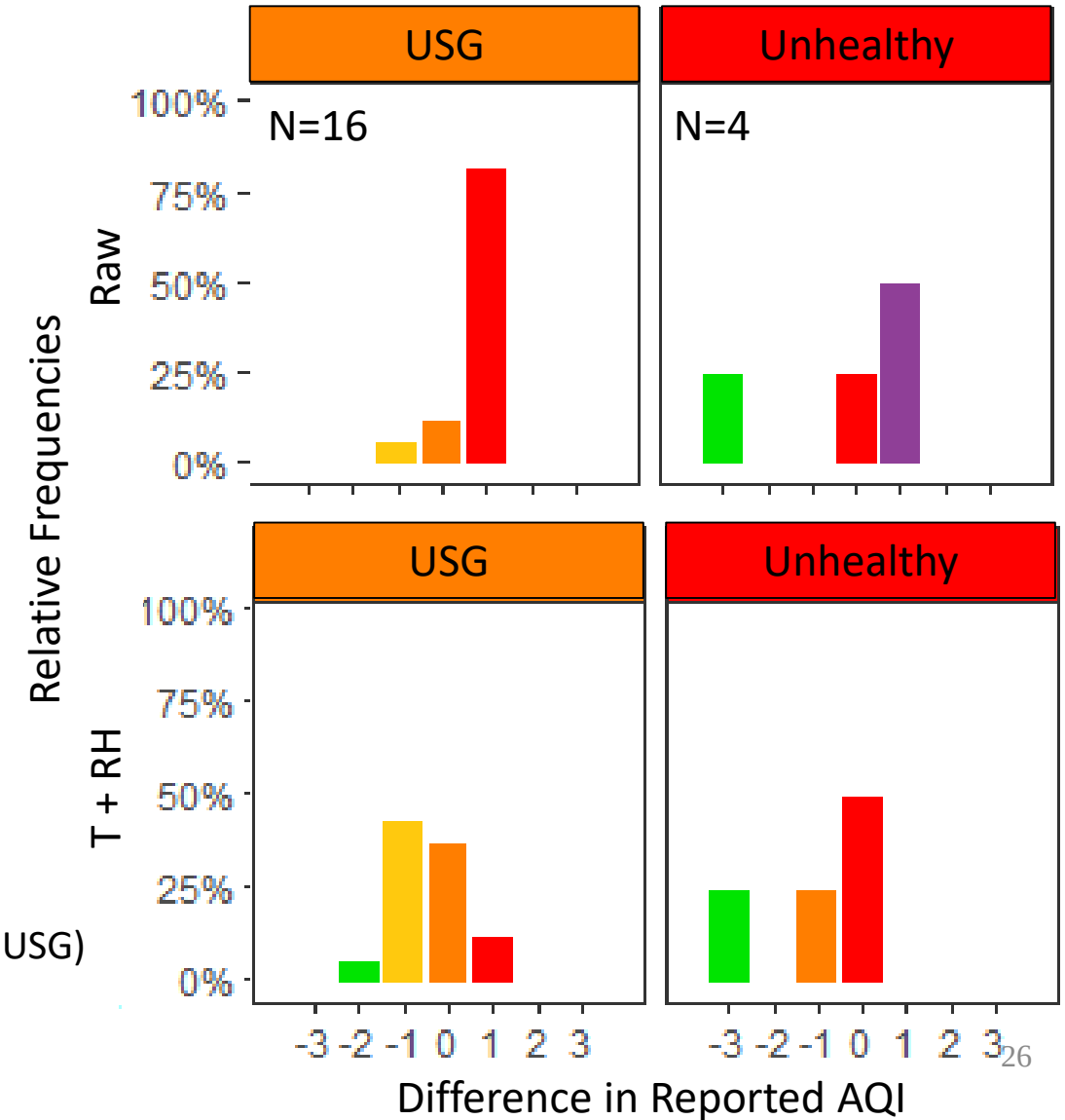


- Good
- Moderate
- Unhealthy for sensitive groups (USG)
- Unhealthy
- Very unhealthy

AQI reporting: USG-Unhealthy

Limited data at higher AQI

Correction also may contribute to under reporting at higher AQI



Conclusions

- A U.S. wide correction including T & RH could improve PurpleAir performance in terms of MBE, RMSE, and correlation across the U.S.
- Even a simple linear correction would significantly improve performance in most parts of the U.S.
- Currently PurpleAir often over reports AQI category
- Proposed corrections may not improve higher category AQI reporting
  - Amara Holder (EPA) is specifically exploring performance under wildfire conditions
  - Working to add more high concentration data to improve model performance at high concentration

Limitations:

- Limited high concentration data
- Unknown applicability at shorter time intervals
- Real-world PurpleAir data has additional uncertainty with reported location

Next steps

- Explore additional collocated data provided by partners
- Explore higher time resolutions data (1-hr)
  - Are agencies interested in 2-minute data?
- Further explore performance over time as more sensors operate for 1 year
- Explore data from wildfires
- Explore the long term performance of additional sensor types across the U.S.

Acknowledgements

AK: State of Alaska, Citizens for Clean Air

AZ: Maricopa County Air Quality Department

CA: San Luis Obispo County Air Pollution Control District, Mojave Desert Air Quality Management District, Antelope Valley Air Quality Managment District, California Air Resources Board, Santa Barbara County Air Pollution Control District, Air Quality Sensor Performance Evaluation Center, Ventura County Air Pollution Control District

CO: Colorado Department of Public Health and Environment

DE: Delaware Division of Air Quality

FL: Sarasota County Government

GA: Region 4, Georgia Environmental Protection Division

IA: Iowa Air Quality Bureau

MT: Missoula County, Montana Department of Environmental Quality

NC: Forsyth County Office of Environmental Assistance & Protection, Clean Air Carolina, UNC Charlotte, North Carolina Department of Environmental Quality

OH: Akron Regional Air Quality Management District

OK: Quapaw Nation, Oklahoma Department of Environmental Quality

UT: University of Utah

VA: Virginia Department of Environmental Quality

VT: State of Vermont

WA: Washington Department of Ecology, Puget Sound Clean Air Agency

WI: Wisconsin Department of Natural Resources

Questions?

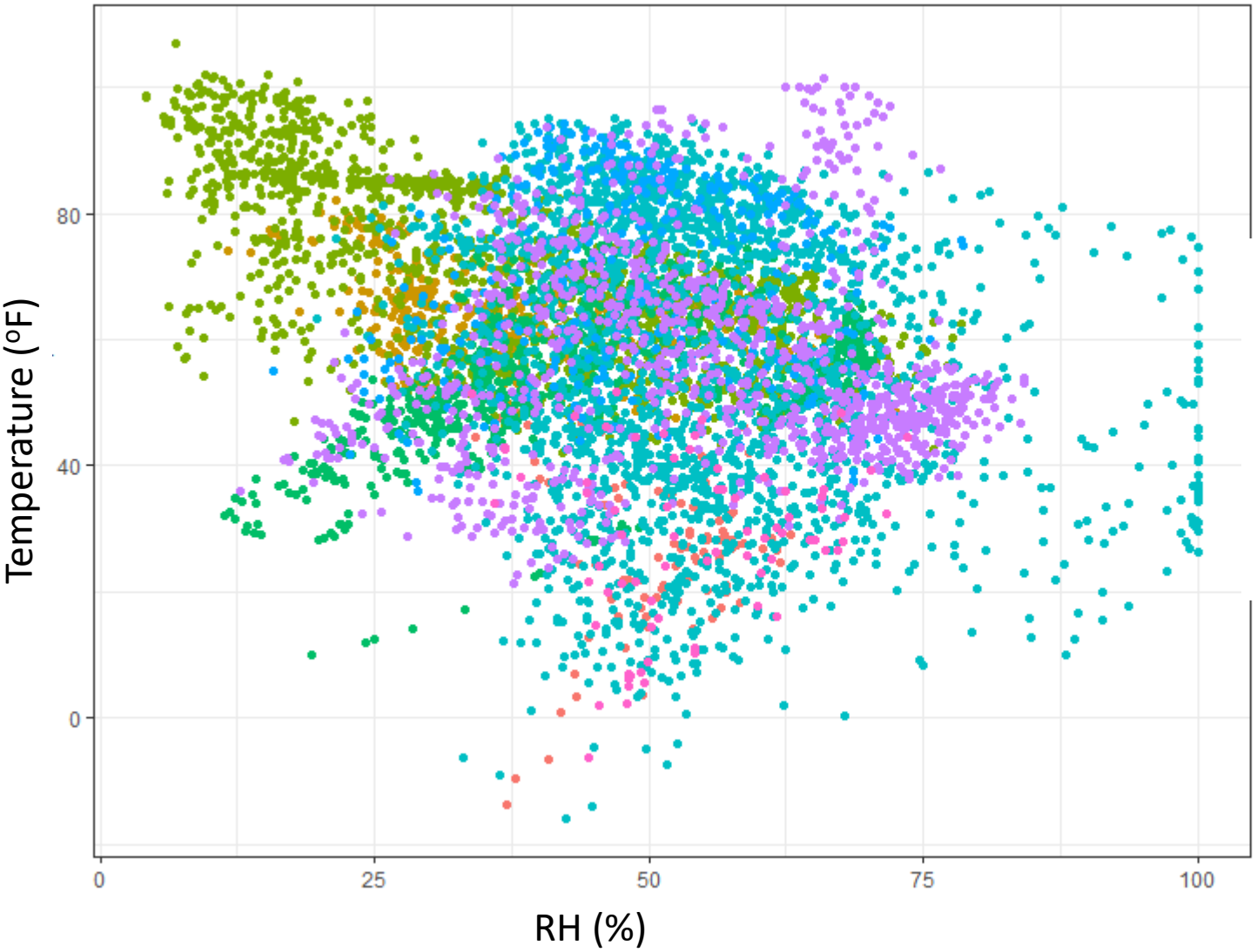
- Contact: Johnson.karoline@epa.gov

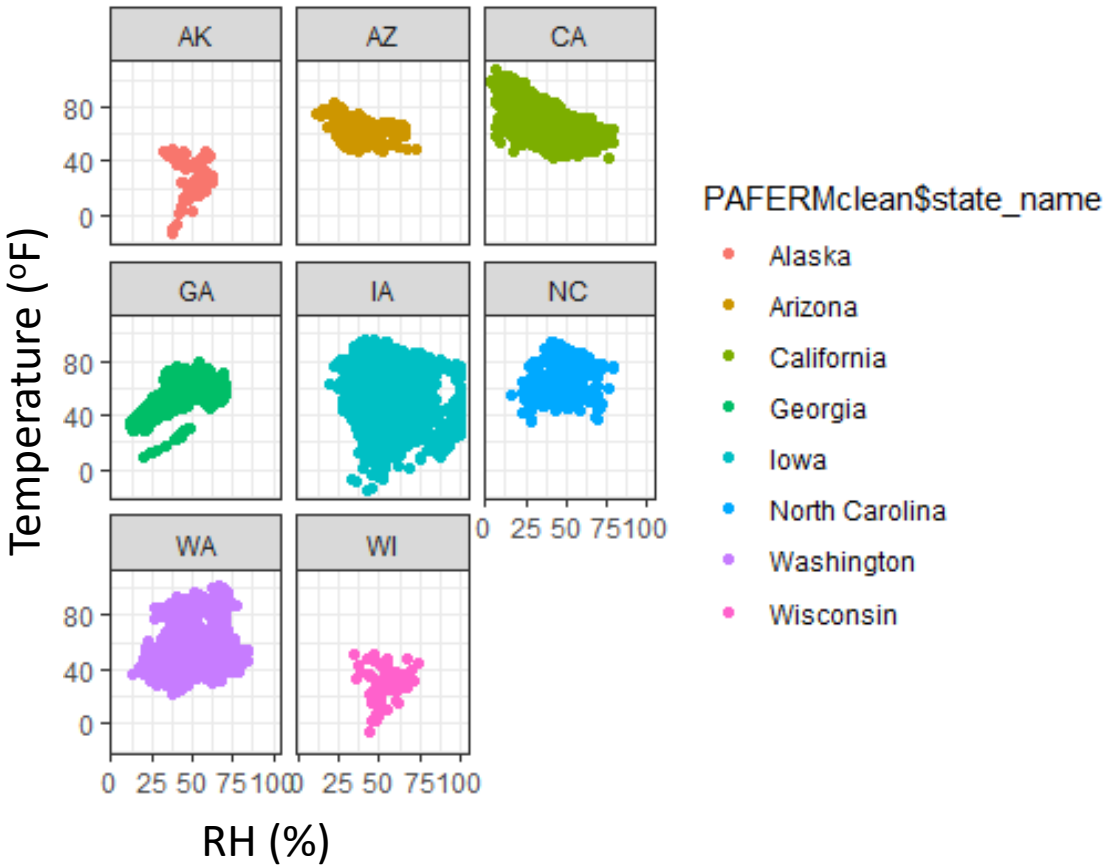
Discussion

Partner agencies: Any initiative or project you would like to discuss?

Extra slides

Dataset: temperature and RH



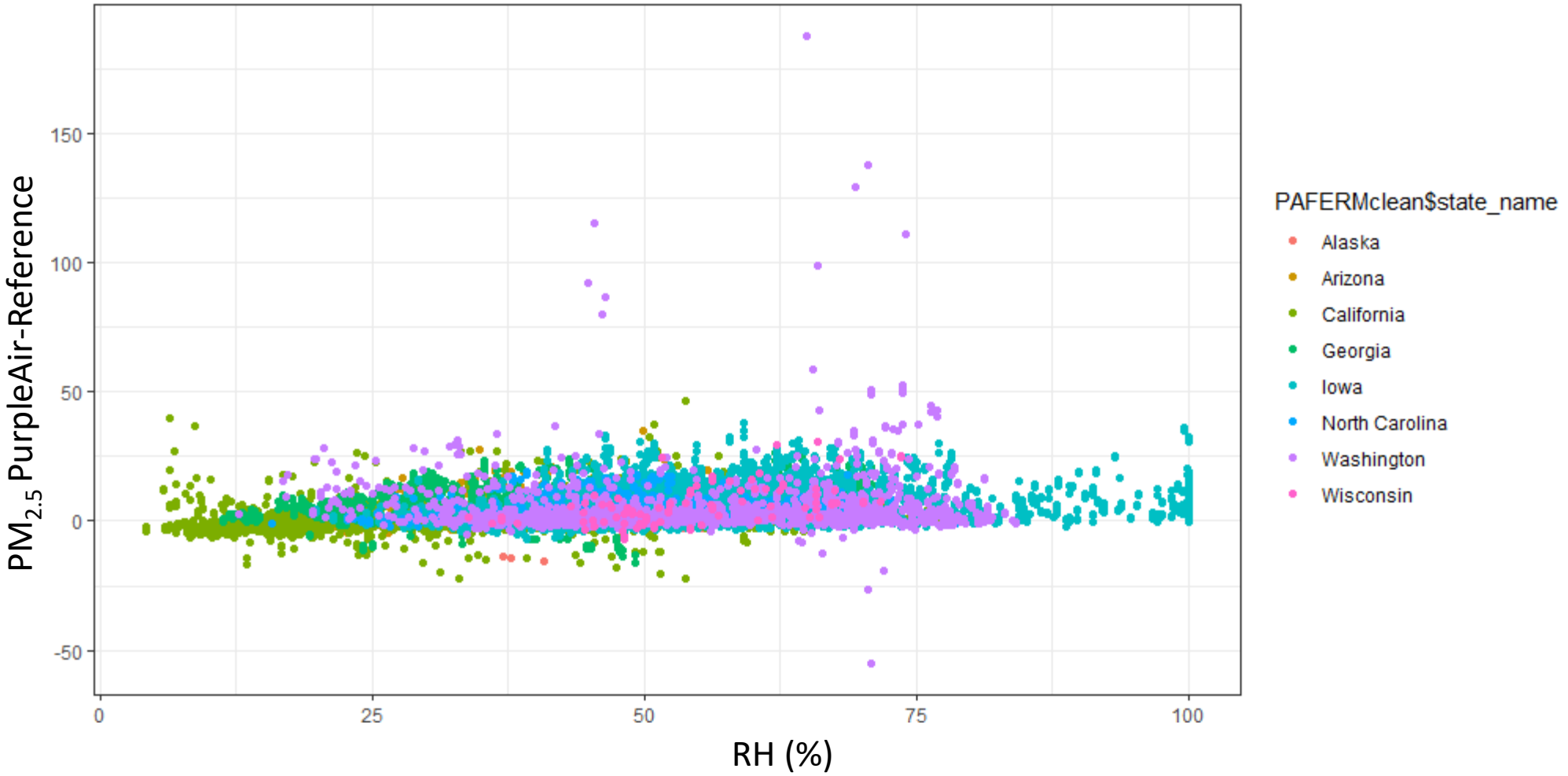


Parameters considered for multilinear regression

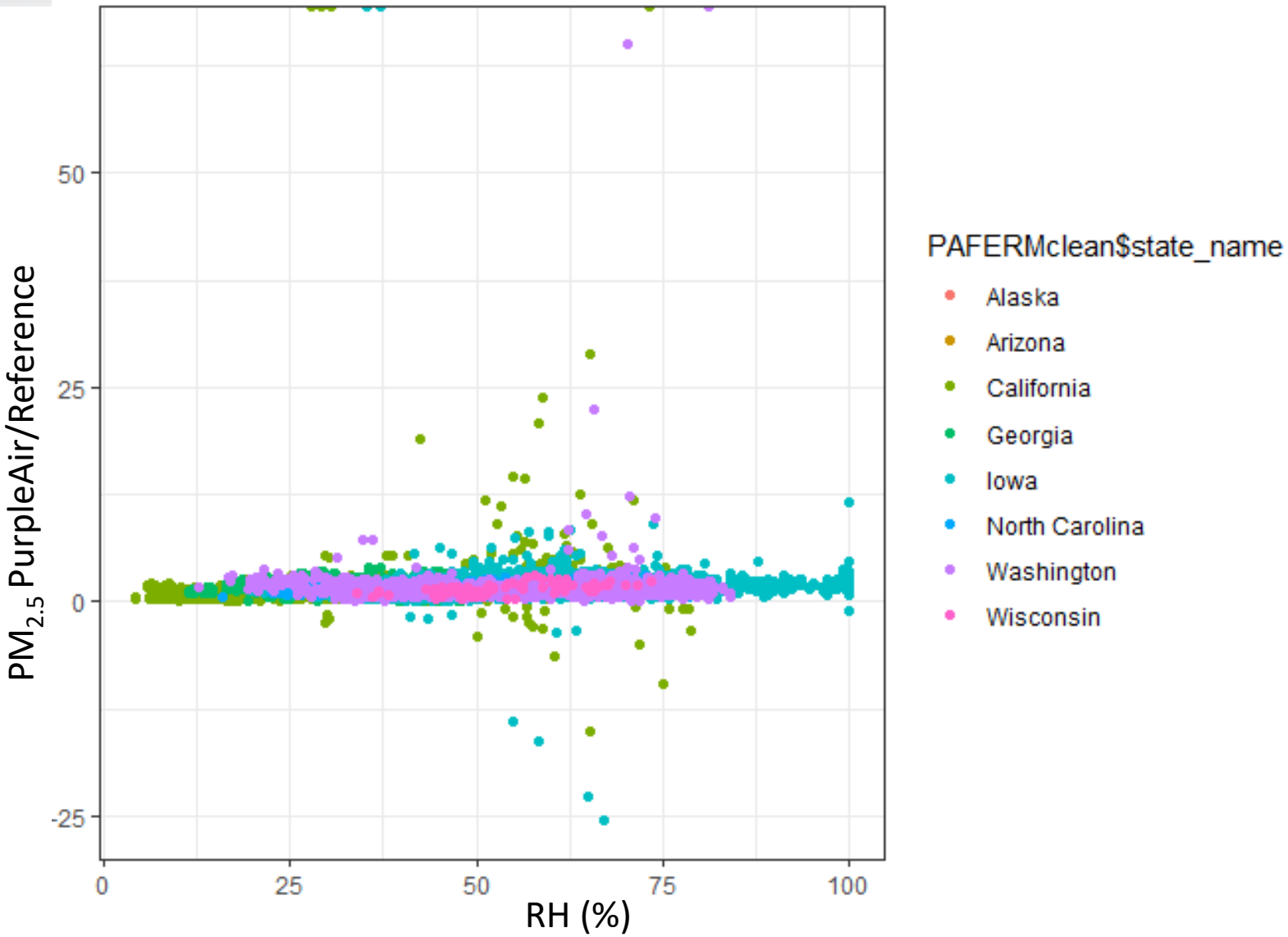
- $PM_{2.5}$ cf=atm and cf=1 channels
  - Cf=atm slightly stronger correlation (cf=1 from Plantower)
- Relative humidity
- Dewpoint
  - RH explained more error
- Temperature
  - T in both ^{o}F and K considered but both provide the same results just different coefficients
  - Marginally significant may remove based on results of full dataset (p=0.05)
- Pressure
  - Missing from 22% of the dataset
- Size bin information (reported to reduce detection limit)
  - reduces error <1%
- Duration deployed
  - full dataset suggests intercept decreases by 0.77 µg m^{-3} after a year of use
- Location (state, region, etc.)
  - Correction less broadly applicable especially since many state datasets do not have all seasons

Only considering parameters available from PurpleAir

RH: sensor-reference



RH: sensor/reference



No distinct pattern between A& B channels over time

- R^2=0 (p=0.02)
- Also no significant relationship time deployed vs. ref-PA difference