STi | Sonoma Technology

Use of Air Quality Sensor Data in Data Fusion Applications for Current and Forecast Air Quality Mapping

Nathan R. Pavlovic, Daniel H. King, Anondo D. Mukherjee,
Anthony M. Cavallaro, Fred W. Lurmann, and Jennifer L. DeWinter

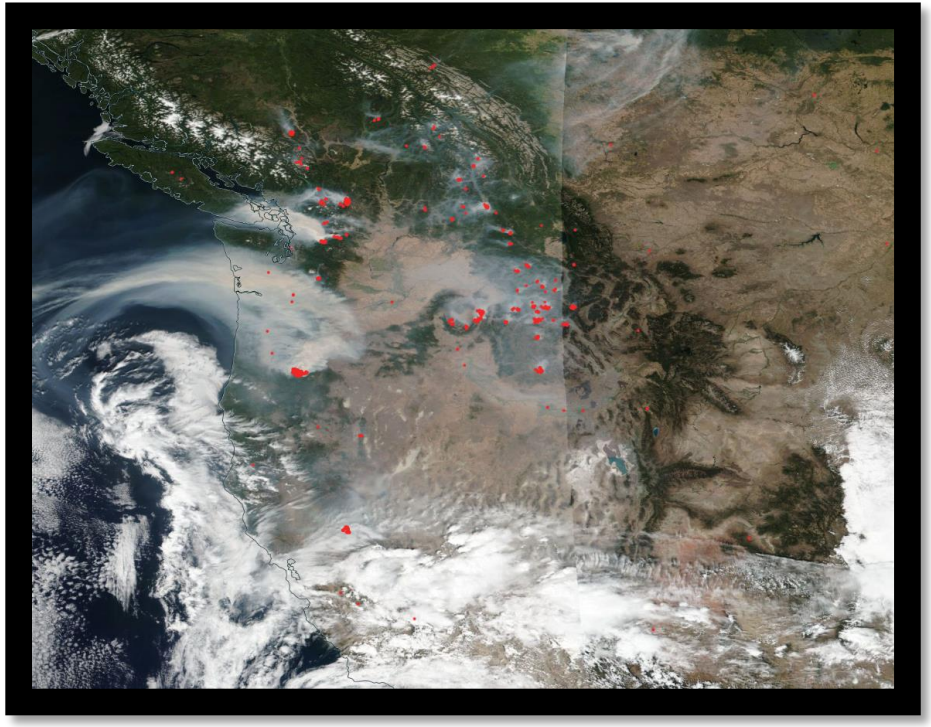
EPA Air Sensor QA Workshop

July 2023

Need for High-Resolution Air Quality Information

Real-time community-scale air quality information is critical for decision making during smoke and other events to reduce exposure and limit impacts to daily life and the economy.

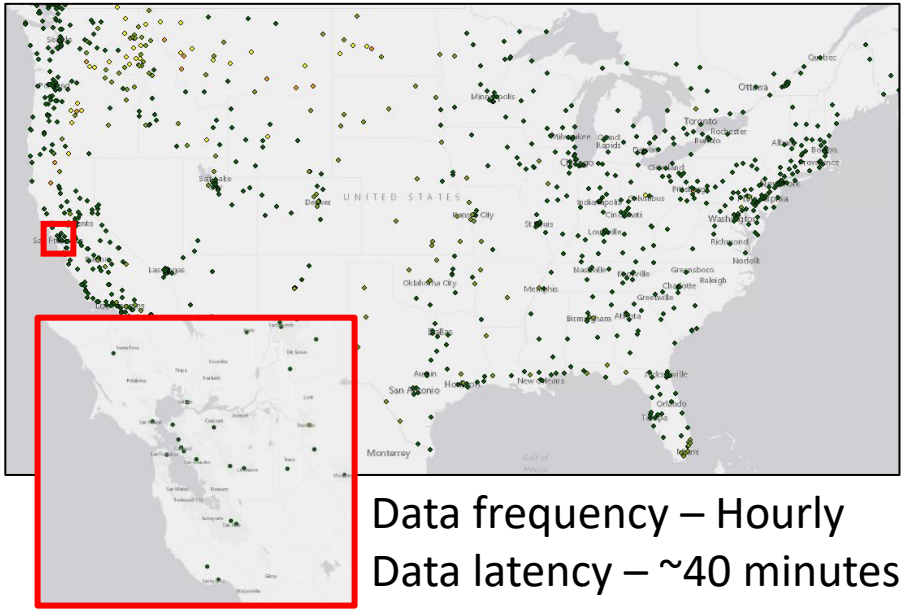
The accuracy and resolution of real-time and forecast air quality information during events can be improved.



Sensor Data Provide Improved Spatial Coverage and Timeliness of Air Quality Information

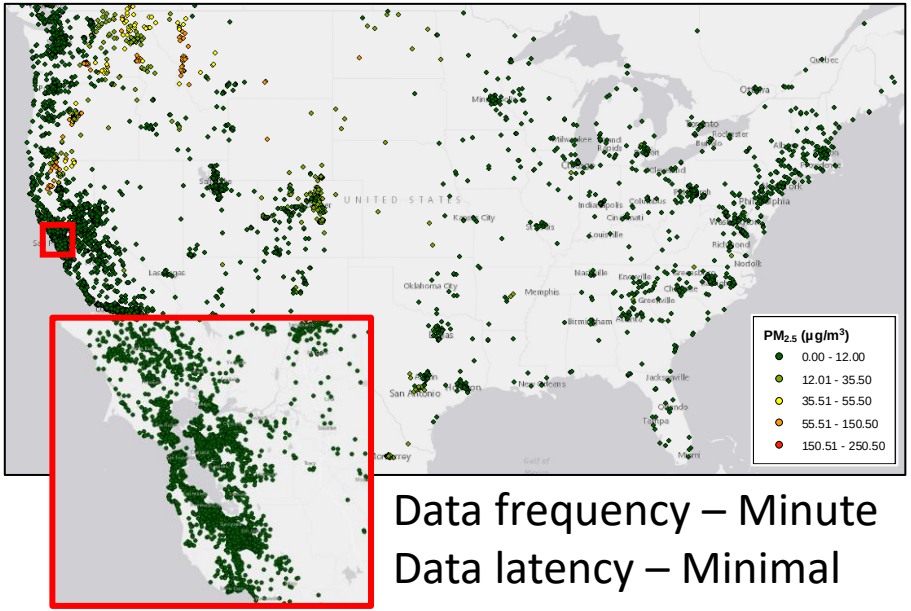
Maps show data for 2021-09-10 17:00 UTC

AirNow Network



Data frequency – Hourly
Data latency – ~40 minutes

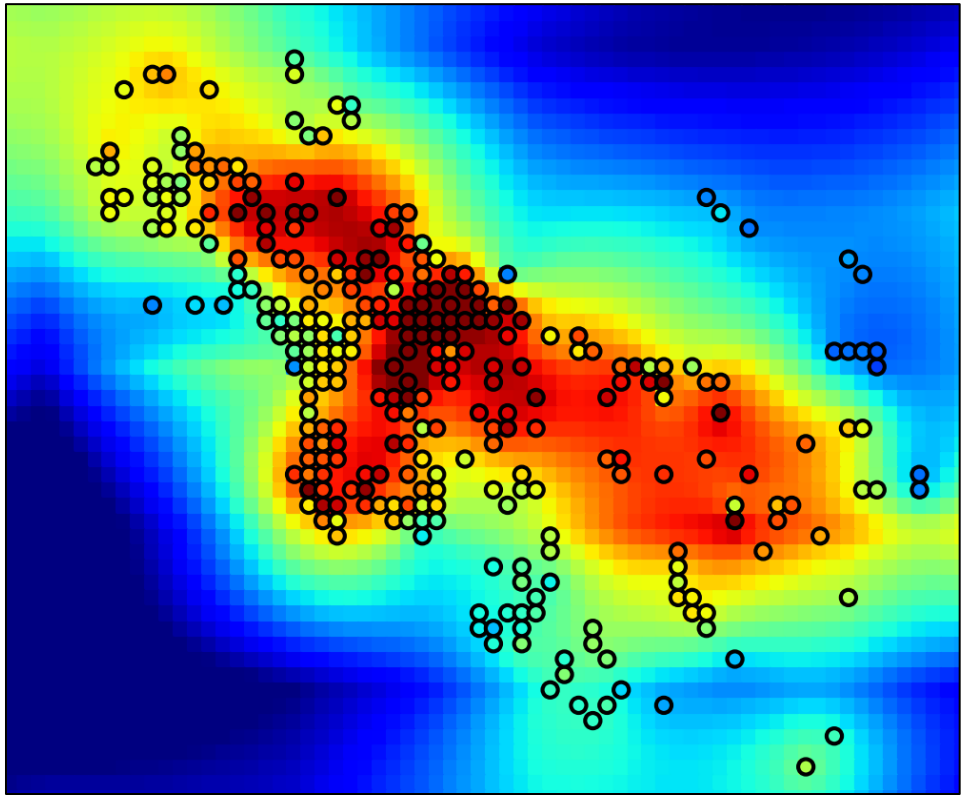
PurpleAir (PA) Network



Data frequency – Minute
Data latency – Minimal

Presentation Overview

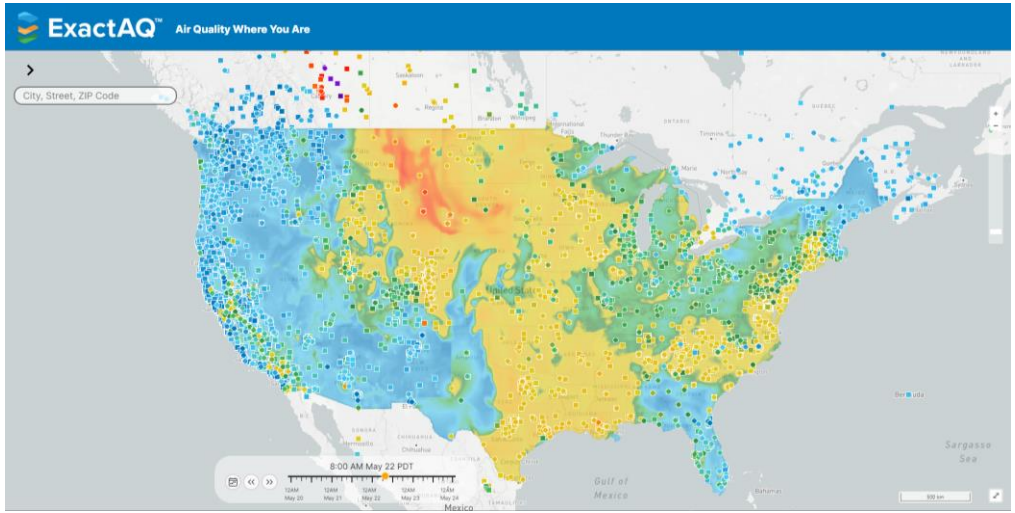
- Summary of input data sets with focus on sensor QA
- Data fusion
  - Current conditions
  - Forecast
- Conclusions and future directions



Introduction to Exact AQ

- Sonoma Technology's ExactAQ is an operational modeling system that provides high-resolution information about current and forecast air quality during pollution events, including wildfire smoke

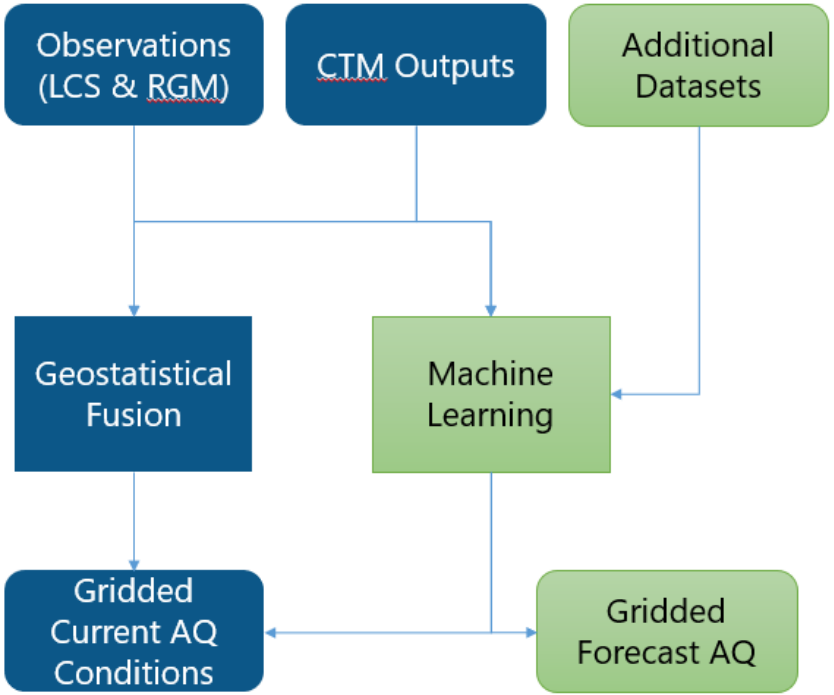
- Operational and runs in real-time
- 1-km gridded maps
- Latest information and 48-hour forecasts
- Combines data from reference-grade monitors, sensors, fire/smoke models, and other data sources



https://www.exactaq.com

ExactAQ's Fusion Processes

- Fusion combines air quality observations from low-cost sensors (LCS) and reference grade monitors (RGM) with chemical transport model (CTM) outputs
- Two fusion approaches
  - Current air quality (AQ) conditions
    - Geostatistical fusion of observations with model data
  - Forecast AQ
    - Machine learning fusion of observations, model data, and additional data sources



LCS Data QA Approach Summary: Purple Air

1. **Screen suspect monitors**
2. QA measurements
3. Apply Correction

<u>Screen out sensors that may be unrepresentative or suspect</u>

- Exclude sensors that are flagged as indoor/inside in their name or metadata
- Exclude sensor where previously corrected hourly $PM_{2.5}$ values have a median value >400 µg/m^3 using at least a month of recent data

LCS Data QA Approach Summary: Purple Air (2)

1. Screen suspect monitors
2. **QA measurements**
3. Apply Correction

<u>Applying quality control steps on minute data</u> – using "cf_1" conversion factor

- Exclude minute data if relative humidity is greater than 95%, or if RH data is missing
- Exclude minute data if channel A/B difference is more than 5 µg/m^3 and greater than 70% of A/B channel average value
- Exclude minute data if channel A or B value is greater than 3,000 µg/m^3
- Require 27 records of minute data when we take hourly average, i.e., 90% completeness assuming 2-minute frequency

LCS Data QA Approach Summary: Purple Air (3)

1. Screen suspect monitors
2. QA measurements
3. **Apply Correction**

Apply U.S. EPA CF_1 correction method to hourly $PM_{2.5}$ data

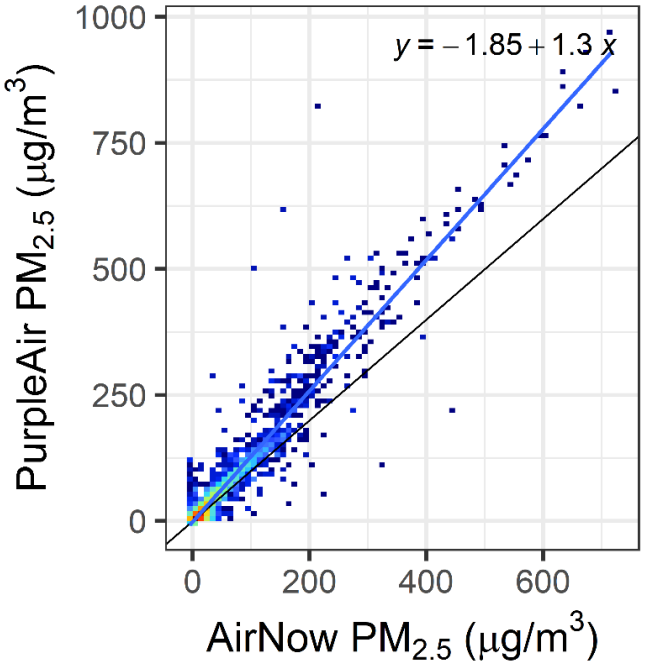
Linear correction applied at low range, quadratic at high range

Cutoffs and correction developed by U.S. EPA (Barkjohn et al., 2021)

Range Applied	Correction
$PA_{cf\_1} \leq 343$ µg/m³ ~176-185 µg/m³ after correction	$PM_{2.5} = 0.52 \times PA_{cf\_1} - 0.086 \times RH + 5.75$
$PA_{cf\_1} > 343$ µg/m³ ~207 µg/m³ after correction	$PM_{2.5} = 0.46 \times PA_{cf\_1} + 3.93 \times 10^{-4} \times (PA_{cf\_1})^2 + 2.97$

Agreement Between LCS and RGM Measurements

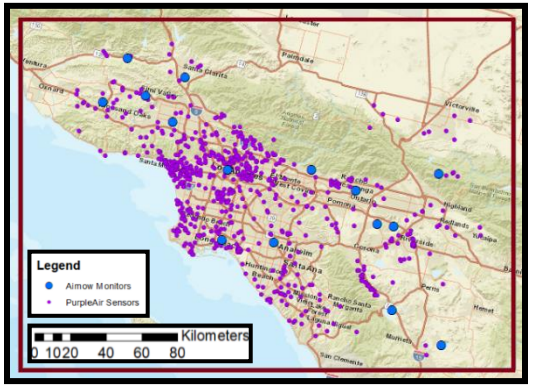
- Select large wildland fire smoke event in Western U.S.
- Bin monitor locations on 1-km grid
- Compare median corrected LCS measurement to median RGM observation
- Close agreement with some bias



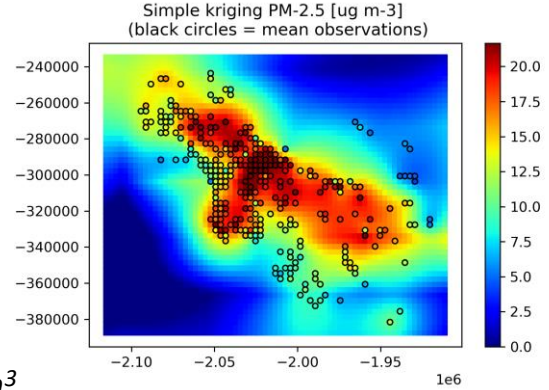
Washington, Oregon, California, Nevada, Idaho, Utah, Arizona, Montana, and Wyoming

Data is from western states from September 1-17, 2022, strongly impacted by wildfire smoke

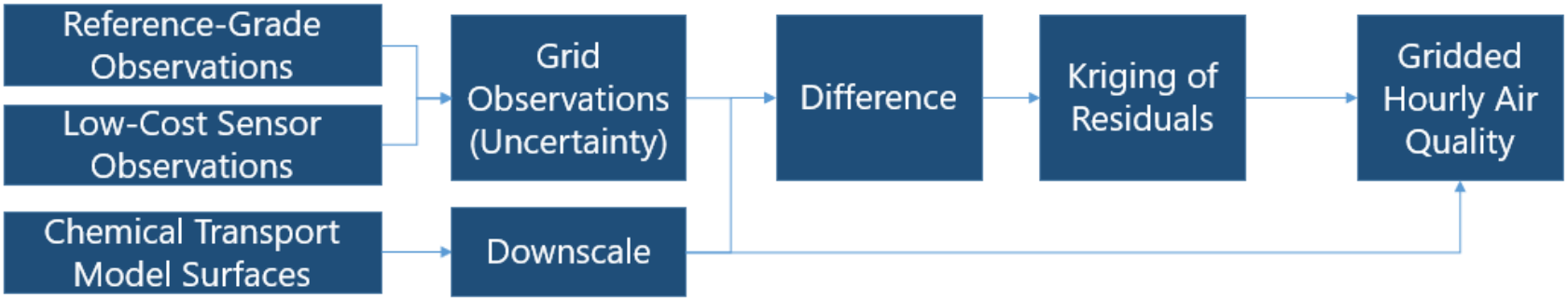
ExactAQ's Geostatistical Data Fusion – Current Conditions



Kriging of Residuals



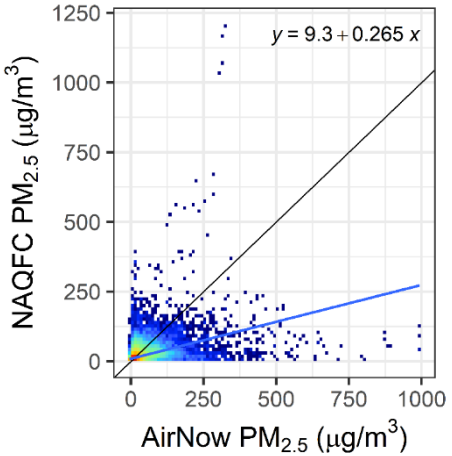
LCS error assumed to be ±4.07 µg/m³ (Schulte et al., 2020) or ±22% (Evans et al., 2021)

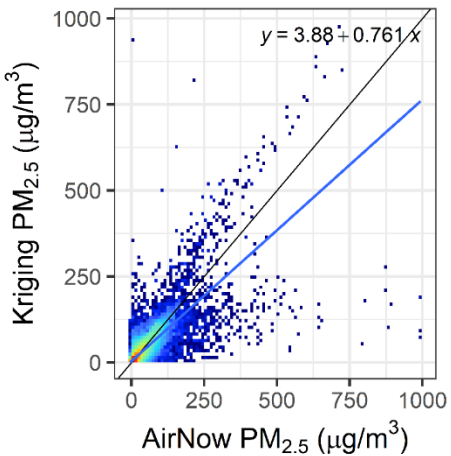


Geostatistical Fusion: 10-Fold Cross Validation

Cross validation results for Bias-corrected NOAA NAQFC (left) and ExactAQ's kriging (right).

The same data from the 10-fold cross validation for Western States is shown on both plots.





Color scale indicates the log-density of plotted points from blue (less dense) to red (more dense).

	NAQFC	ExactAQ Kriging	NAQFC	ExactAQ Kriging
All Data	0.204	0.644	26.4	17.6
No Smoke	0.167	0.427	8.6	6.4
Smoke	0.188	0.637	30.2	20.1

Mean R^2 values from 10 folds of AirNow vs. model/ExactAq for smoke and no-smoke conditions from HMS smoke polygons. Results shown for Western States for Sept. 1-17, 2022.

Mean RMSE values (µg/m^3) from 10 folds for the same data.

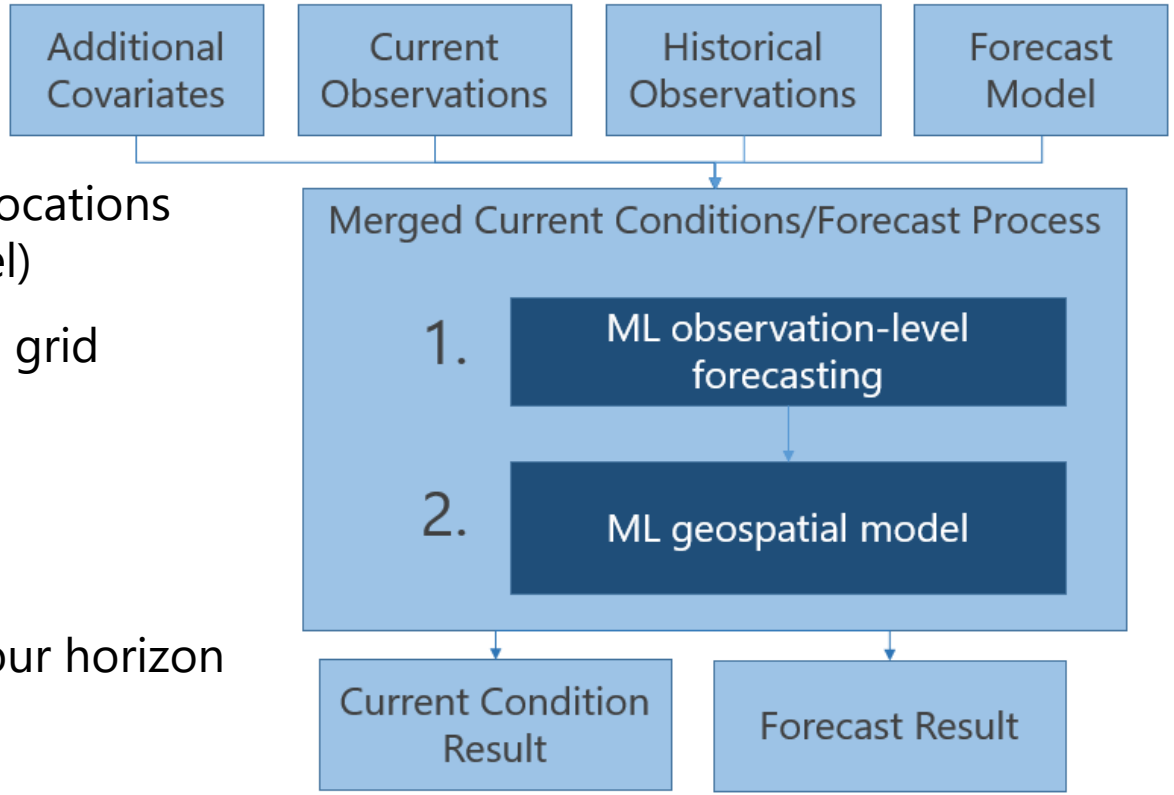
ExactAQ's Machine Learning (ML) Data Fusion – Gridded Forecasting

Modeling Steps

1. Forecasting at monitor locations (observation-level model)
2. Forecasting across 1-km grid (geospatial model)

Implementation Details

- Full CONUS Domain
- Hourly forecast with 48-hour horizon
- Initialized hourly

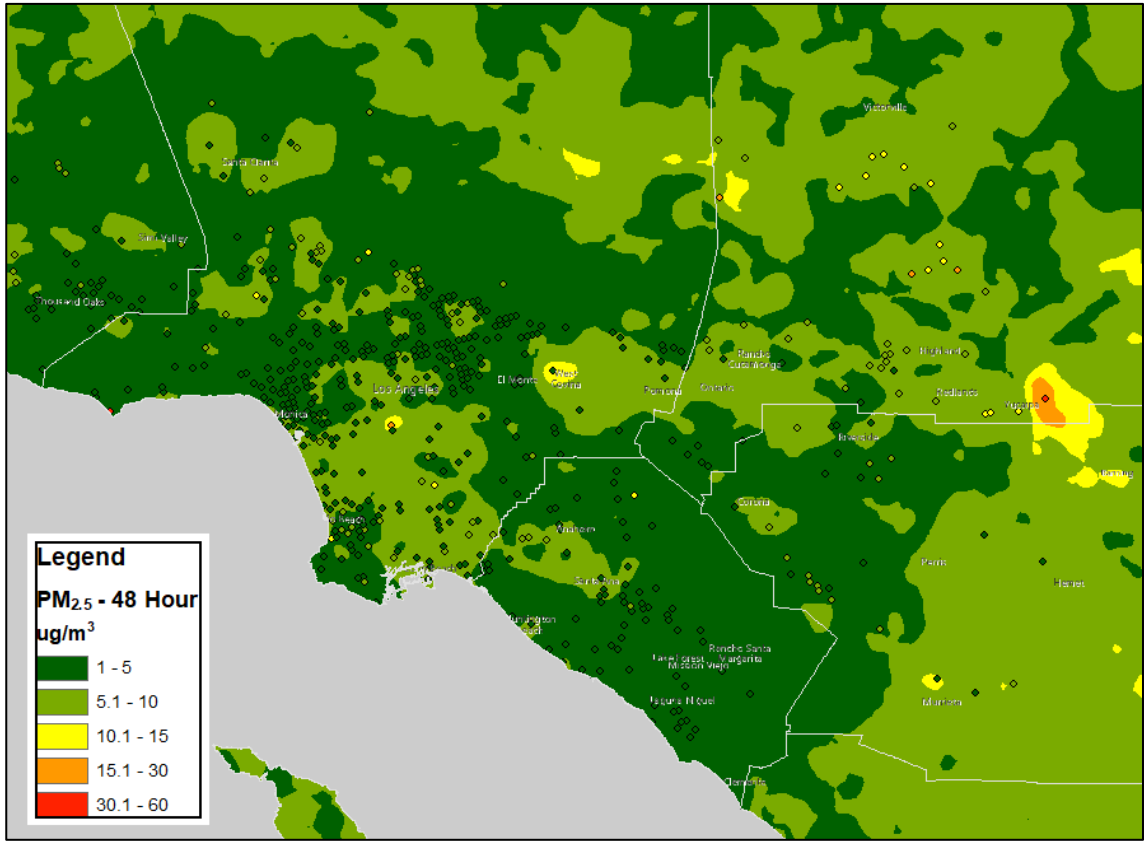


Gridded Forecast Results

48-Hour-Ahead Hourly Forecast for:

December 1, 2022
02:00 UTC

Southern California

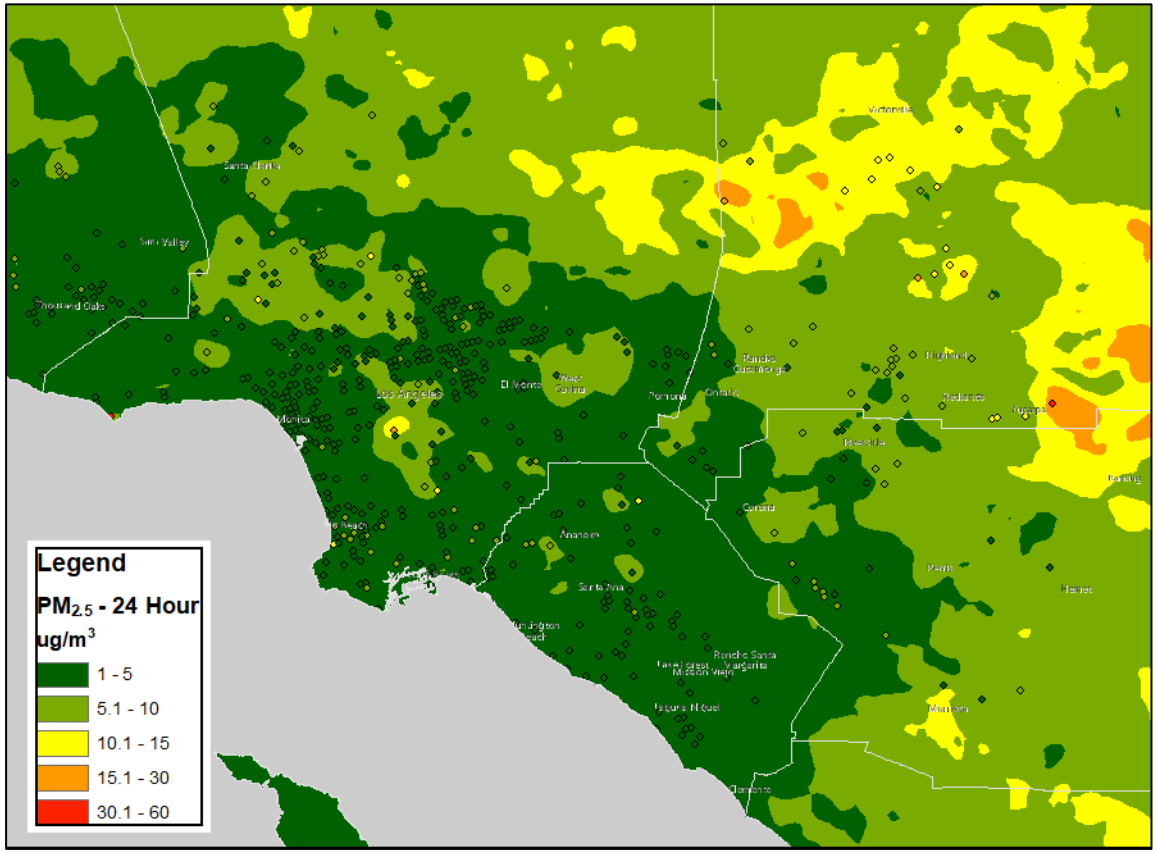


Gridded Forecast Results (2)

24-Hour-Ahead Hourly Forecast for:

December 1, 2022
02:00 UTC

Southern California

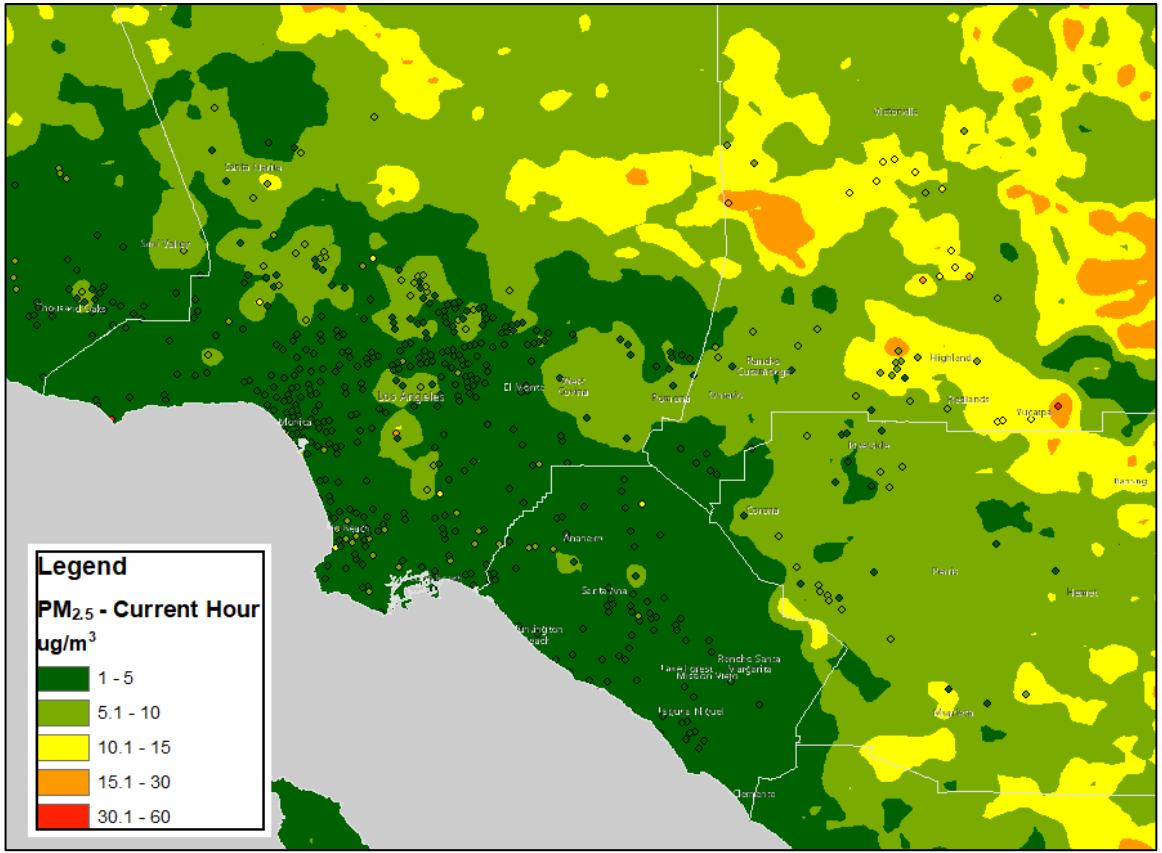


Gridded Forecast Results (3)

Current Hour Forecast for:

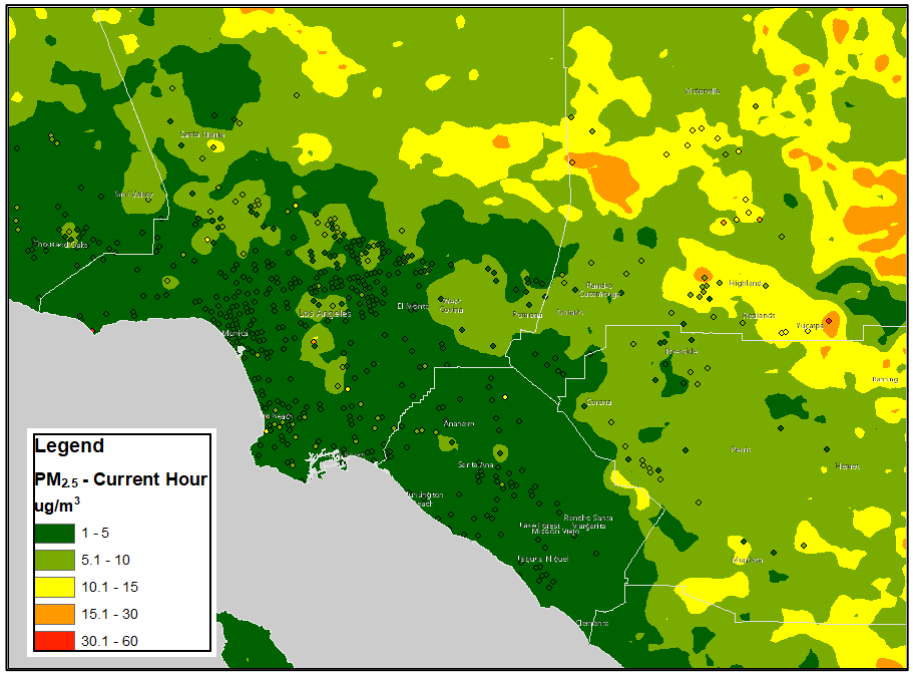
December 1, 2022
02:00 UTC

Southern California

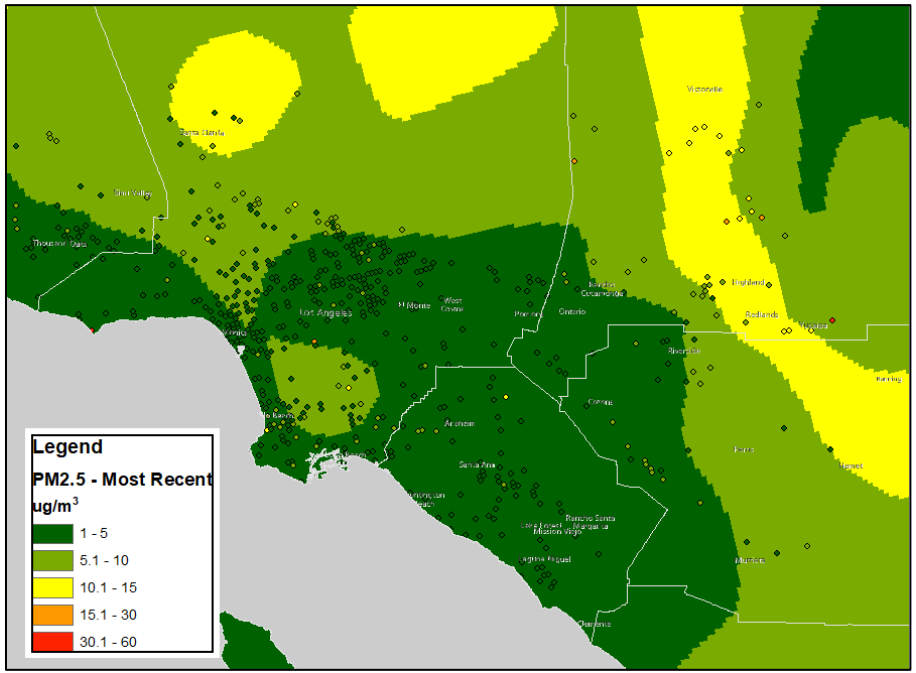


Gridded Forecast Results (Current hour)

Machine Learning Forecast
(1 km Resolution)

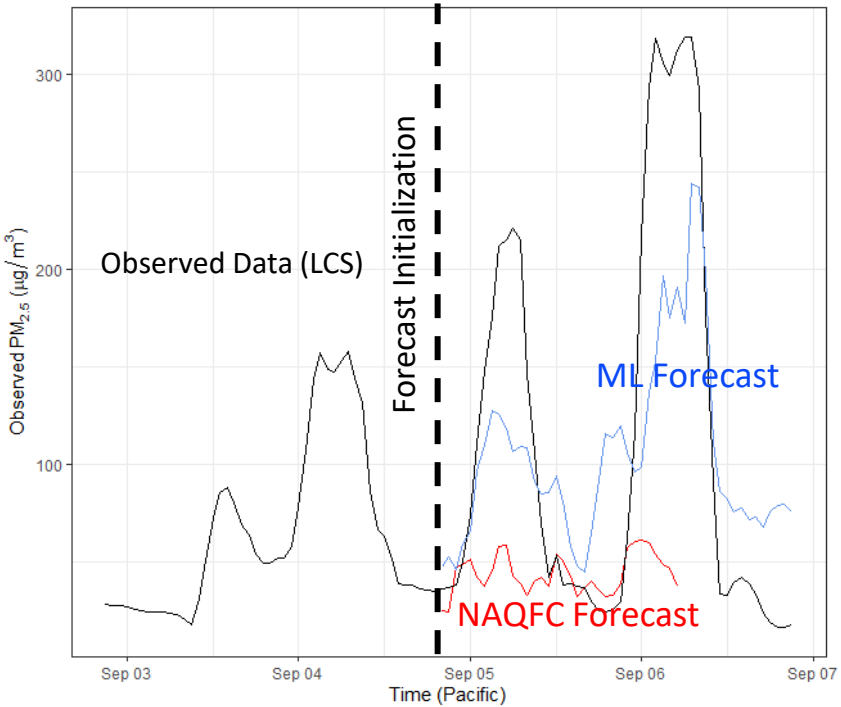


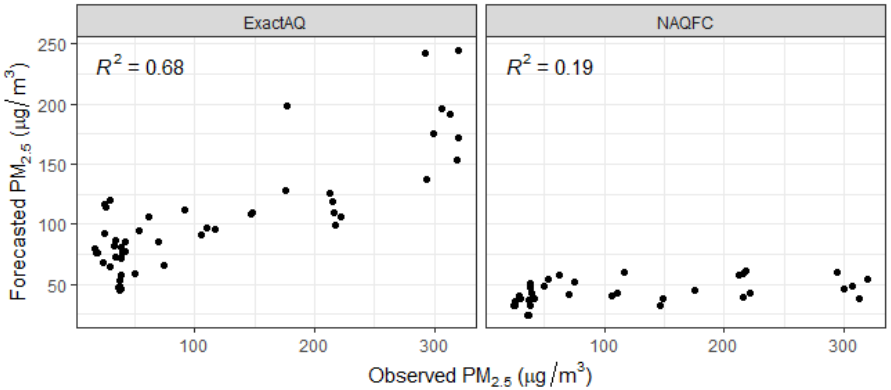
Bias-Corrected NAQFC
(Bilinearly Interpolated to 1 km Resolution)



Southern California, December 1, 2022 02:00 UTC

Machine Learning (ML) Forecast: Cross Validation Example





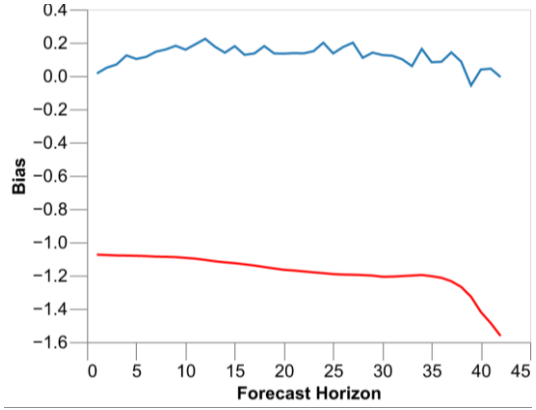
In this example, for hours with both ML and NAQFC forecasts, ML shows:

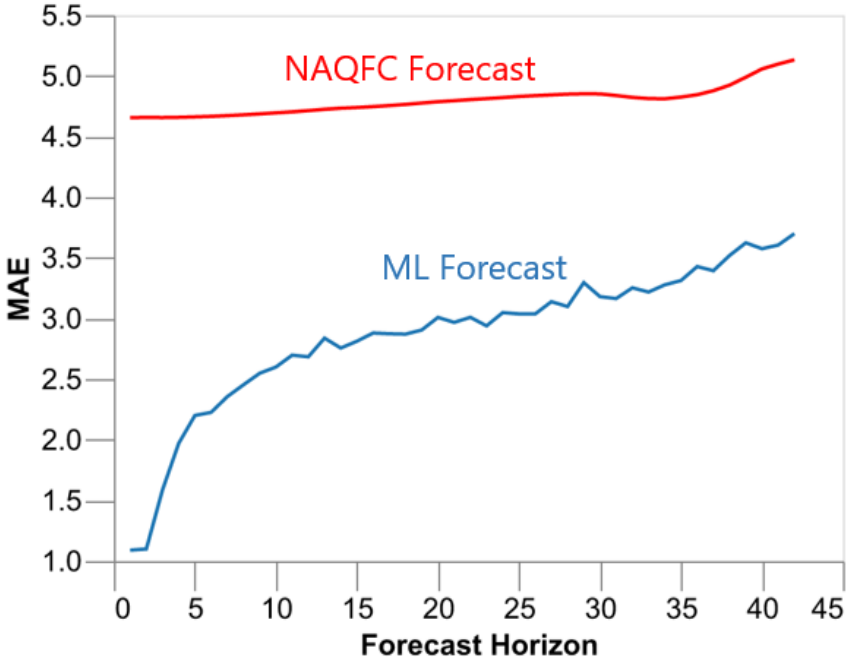
- higher R^2 (0.65 vs. 0.19)
- lower RMSE (75.7 vs 125.1)
- more frequent predictions of actual unhealth conditions (100% vs. 32%)

ML forecast initialized on September 4, 2021, at 8 p.m. Results shown for PurpleAir monitor located in Central California.

Machine Learning (ML) Forecast: Complete Cross Validation

- Across CONUS, ML forecasts are within 3 ug/m^3 on average over forecast horizons up to 24 hours (right)
- ML forecasts have minimal bias across forecast horizons (below)





Cross-validation results based on independent test data from all CONUS RGM and LCS data for April and September 2021, and February, June, and December 2022.

Conclusions and Next Steps

- Sensor data provide critical spatial and temporal information for fusion applications
- QA approaches for use of sensors in fusion are needed improve outcomes
- Sensor timeliness provides key advantage over what is possible with reference-grade monitors
- Machine learning forecasts incorporating low-cost sensor data can provide high-quality air quality conditions maps
- Next steps:
  - Investigate methods that may further reduce bias and error when event type is known or unknown

Sonoma Technology

Nathan R. Pavlovic

Lead Geospatial Data Scientist

npavlovic@sonomatech.com

Jennifer DeWinter

Vice President/ExactAQ Product Manager

jdewinter@sonomatech.com

Acknowledgements and References

Data Sources:

Low-cost sensor data obtained from PurpleAir (https://www2.purpleair.com/).

Reference-grade monitor data courtesy of U.S. Environmental Protection Agency (EPA) AirNow.

Bias-Corrected National Air Quality Forecasting Capability (NAQFC) data courtesy of U.S. National Oceanic and Atmospheric Administration (NOAA).

References:

Barkjohn, K., Holder, A., Clements, A., Fredrick, S., Evans, R, 2021. Sensor Data Cleaning and Correction: Application on the AirNow Fire and Smoke Map, U.S. EPA ORD.

Evans, R., Larkin, S., Barkjohn, K., Clements, A., Holder, A., 2021, AirNow Fire and Smoke Map: Extension of the US-Wide Correction for Purple PM2.5 Sensors, U.S. EPA ORD.

Schulte, N., Li, X., Ghosh, J.K., Fine, P.M., Epstein, S.A., 2020. Responsive high-resolution air quality index mapping using model, regulatory monitor, and sensor data in real-time. Environ. Res. Lett. 15, 1040a7. https://doi.org/10.1088/1748-9326/abb62b.