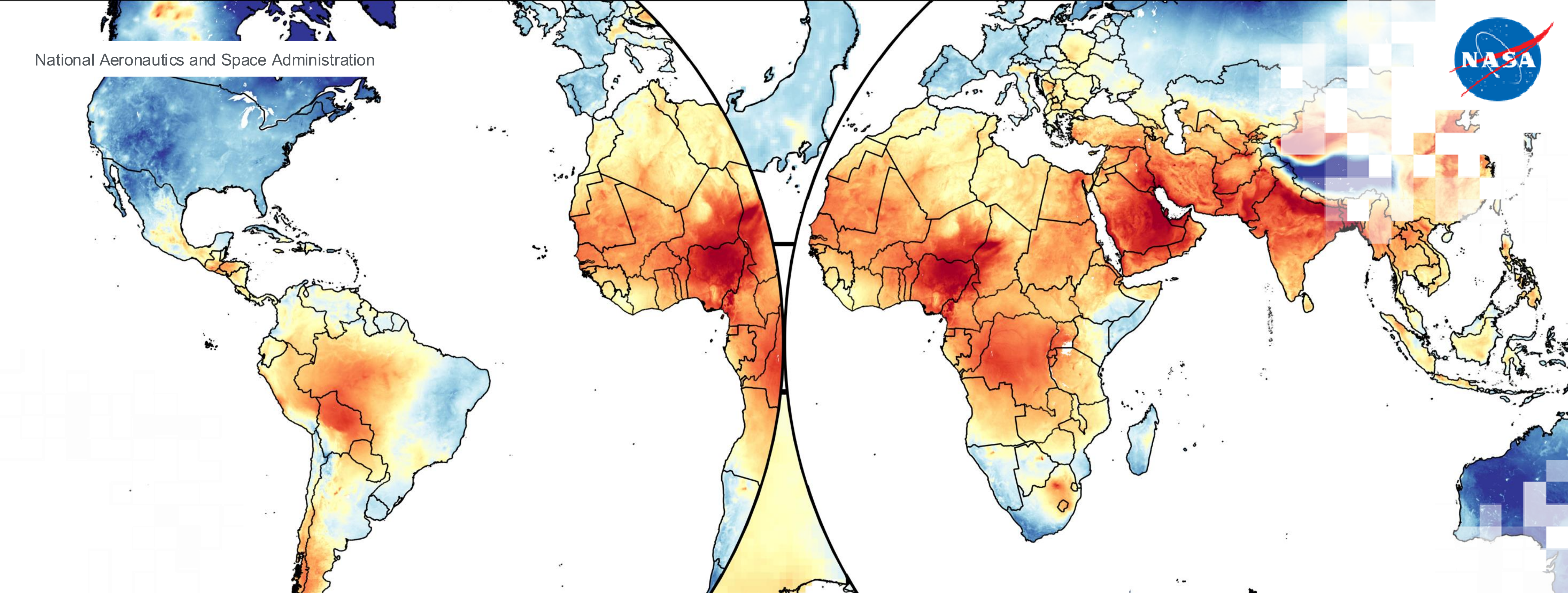
National Aeronautics and Space Administration



Estimating Surface $PM_{2.5}$ using Satellite Data and Other Information Sources

Part 2: Estimation of $PM_{2.5}$ from AOD - methodologies and available datasets

Carl Malings (Morgan State University, MSU), Pawan Gupta (NASA GSFC), Aaron van Donkelaar (Washington University in St. Louis), Junhyeon Seo (MSU), Sebastián Diez (Universidad del Desarrollo) & Ana Prados (University of Maryland Baltimore County)

July 15, 2026

Part 2 – Trainers

Aaron van Donkelaar
Research Associate
Washington University in St. Louis

Pawan Gupta
Research Scientist
Goddard Space Flight Center

Junhyeon Seo
Assistant Research Scientist
Morgan State University

Part 2 Objectives

By the end of Part 2, participants will be able to:

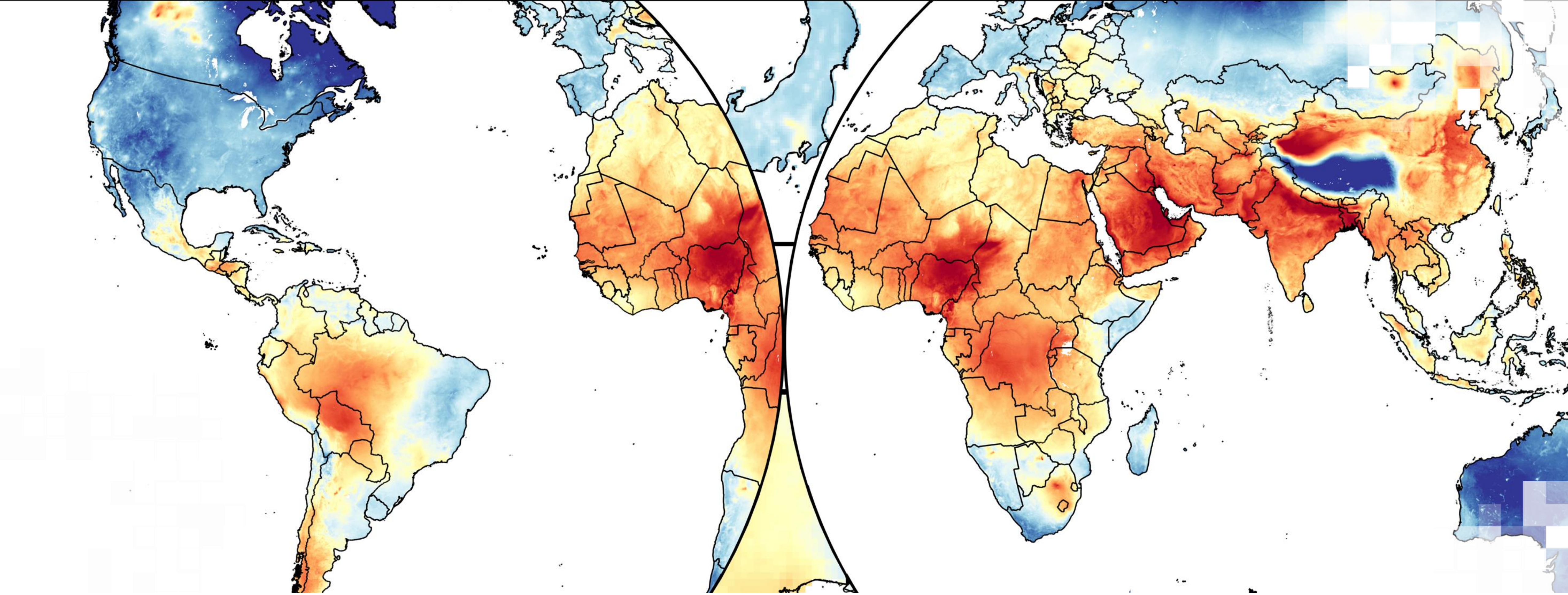
- Explain geophysical, hybrid, and machine learning methods of inferring surface $PM_{2.5}$ from satellite AOD.
- Differentiate between available data products for surface $PM_{2.5}$ from NASA and Washington University in St. Louis based on their methodologies, strengths, and weaknesses.
- Use NASA tools and the SatPM website to access these surface $PM_{2.5}$ data products.

Review of Prior Knowledge

- $PM_{2.5}$ is a measure of the in-situ mass concentration of particles smaller than 2.5 microns.
- Aerosol Optical Depth (AOD) is an optical measure of the atmospheric column aerosol loading.
- AOD can be well related to $PM_{2.5}$ under certain conditions, but the relationship breaks down if:
  - Aerosols are prevalent well above ground level
  - Coarse aerosol particles are dominant
  - Aerosols absorb water in high-humidity conditions
  - Atmospheric or surface conditions prevent reliable AOD retrievals
  - The spatial resolution of AOD data do not capture local $PM_{2.5}$ variability
  - The timing of $PM_{2.5}$ measurements do not coincide with the satellite overpasses

How to Ask Questions

- To ensure we see your question, please locate the three dots ... to the bottom right of your training window.
- You will see a Slido/Q&A option. You may also see a standalone Slido tab or app.
- Please place all questions there and we will answer them during the Q&A session.
- We will try to get to all of the questions during the Q&A session after the webinar. The remainder of the questions will be answered in the Q&A document, which will be posted to the training website about a week after the training.

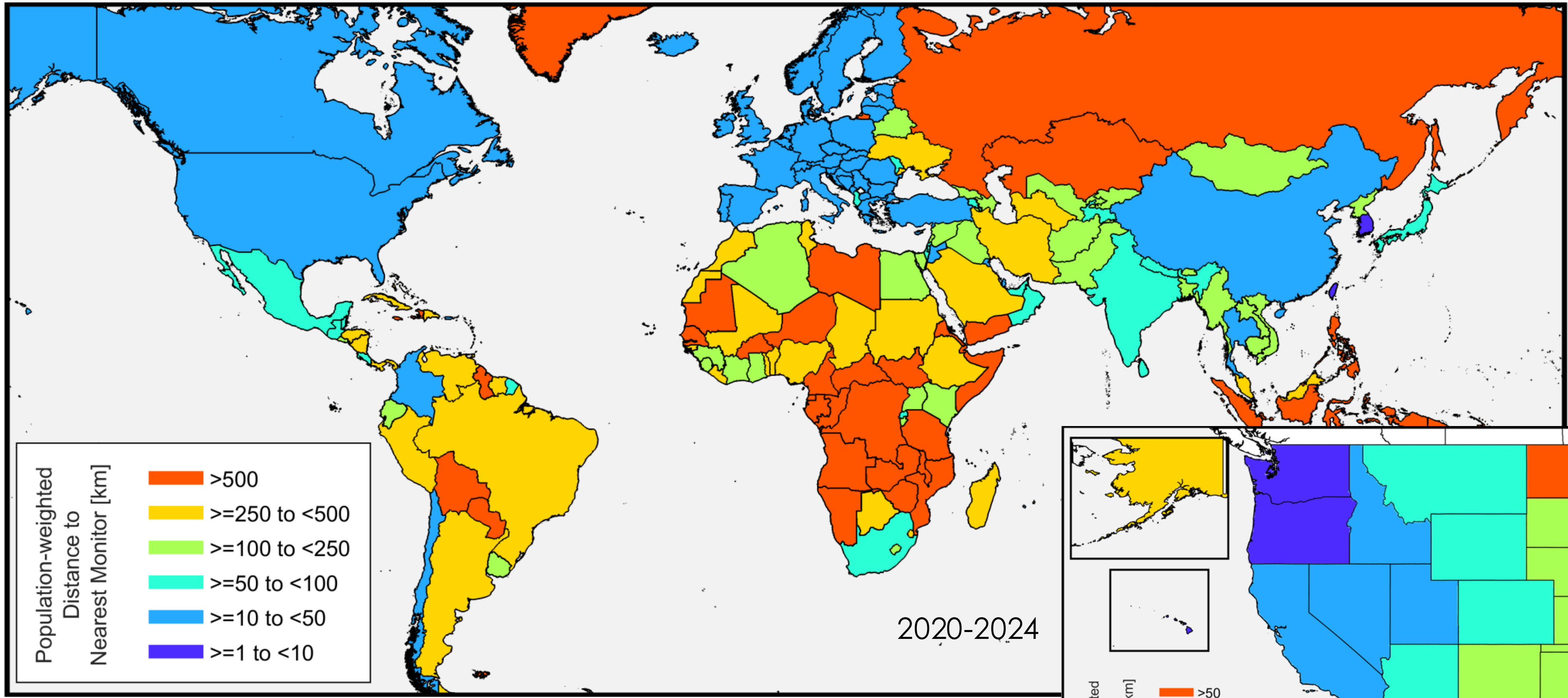


Geophysical and Hybrid $PM_{2.5}$ Estimation

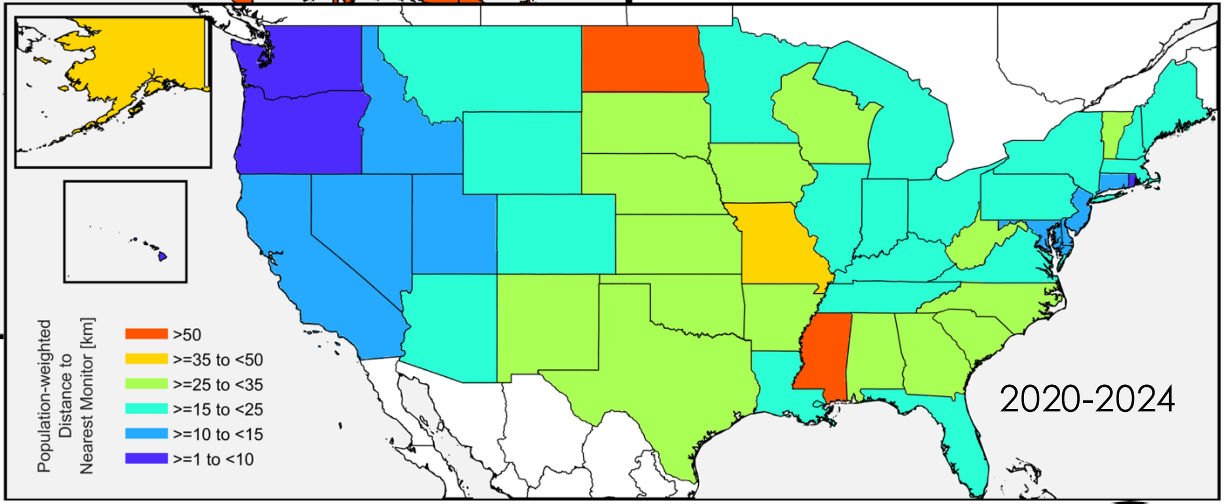
Ground-based monitors can be far from population.

Motivates the use of AOD to represent $PM_{2.5}$, but challenges methods that rely heavily on monitored locations

Population-Weighted Distance to Nearest Monitor	
Global	200 km
North America	20 km
North Africa/Middle East	260 km
South Asia	70 km
Central Latin America	130 km



Updated from: Martin et al., 2019



How to Relate AOD to $PM_{2.5}$ without Local Monitors

From Last Week:

Mixing (boundary layer) height

Extinction coefficient for dry aerosols

$$\boxed{\mathrm{AOD}_{550\,nm} = \mathrm{PM}_{2.5}\ \mathrm{H}\ \ \mathrm{f}(RH)\,\frac{3\,Q_{ext,dry}}{4\,\rho\,r_{eff}}}$$

Function of aerosol scattering change with relative humidity

Aerosol dry mass density

Particle effective radius

Assumes:

- Cloud-free skies
- Well mixed boundary layer
- No aerosols above the boundary layer
- Aerosols are optically similar throughout column

Chrisopher and Hoff, 2009
van Donkelaar et al., 2006

What if we relaxed our assumptions?

If we remove the assumption that there are no aerosols above the boundary layer and only consider the AOD between the surface and a near-surface height of Δz:

Between surface and Δz and associated with particles < 2.5 μm

$$\boxed{\mathrm{PM}_{2.5} = \mathrm{g}(RH)\left[\frac{4\,\rho_{2.5,\Delta z}\, r_{eff,2.5,\Delta z}}{3Q_{ext,dry,2.5,\Delta z}\Delta z}\right]\mathcal{F}_{2.5,\Delta z}\mathrm{AOD}_{550\;nm}}$$

Function of aerosol scattering change with relative humidity

Fraction of AOD near the surface

Assumes:
- Cloud-free skies
- ~~Well mixed boundary layer~~
- ~~No aerosols above the boundary layer~~
- ~~Aerosols are optically similar throughout column~~

Chrisopher and Hoff, 2009
van Donkelaar et al., 2006

Geophysical $PM_{2.5}$ relies on the theory-based, physically-driven relationship between total column AOD and $PM_{2.5}$.

- Requires local assumptions/representation of parameters driving the relationship
- Independent of local $PM_{2.5}$ measurements

Growth Factor, f(Type, RH)

Density, f(Type)

Effective Radius, f(Type)

Near-Surface Fraction, f(Vertical Profile)

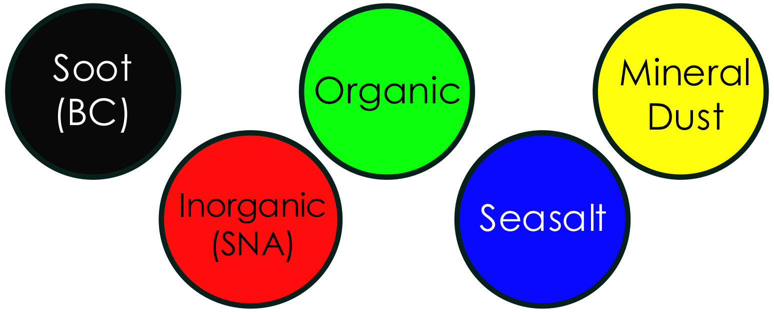
$$\mathrm{PM}_{2.5} = \mathrm{g}(RH)\left[\frac{4\,\rho_{2.5,\Delta z}\,r_{eff,2.5,\Delta z}}{3Q_{ext,dry,2.5,\Delta z}\Delta z}\right]\mathcal{F}_{2.5,\Delta z}\mathrm{AOD}_{550\,nm}$$

Extinction Coefficient, f(Type)

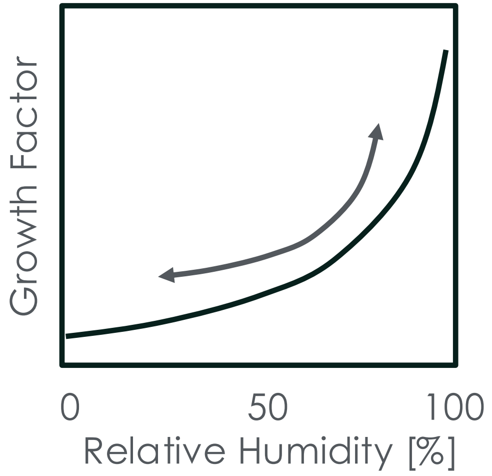
$$\boxed{PM_{2.5} = \eta \times AOD}$$

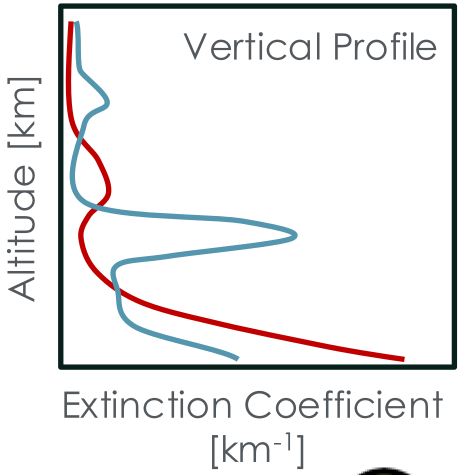
For simplicity, the geophysical relationship is usually represented to a single scalar term, η.

Aerosol Type:



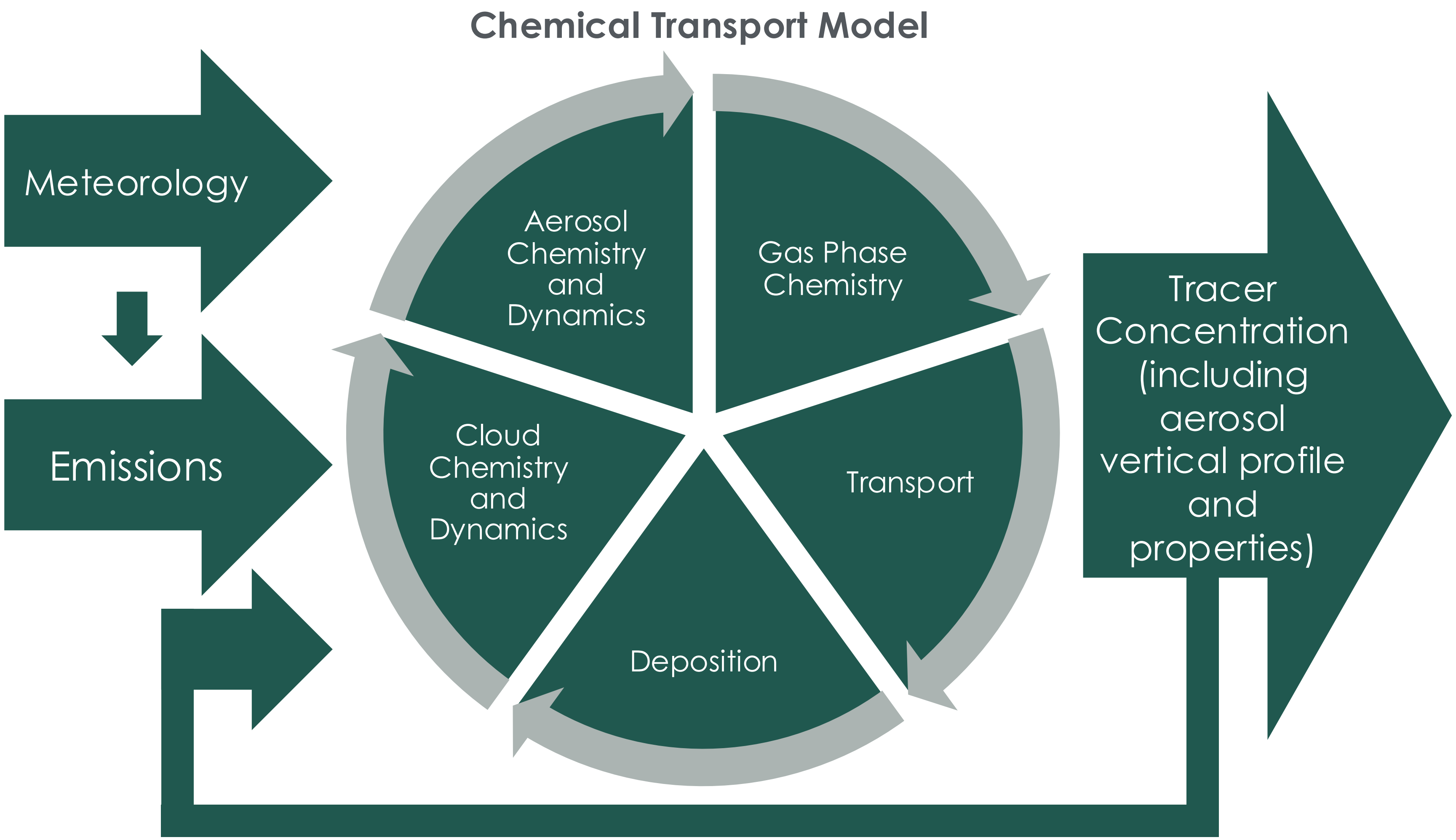
Smaller Particles → Larger Particles



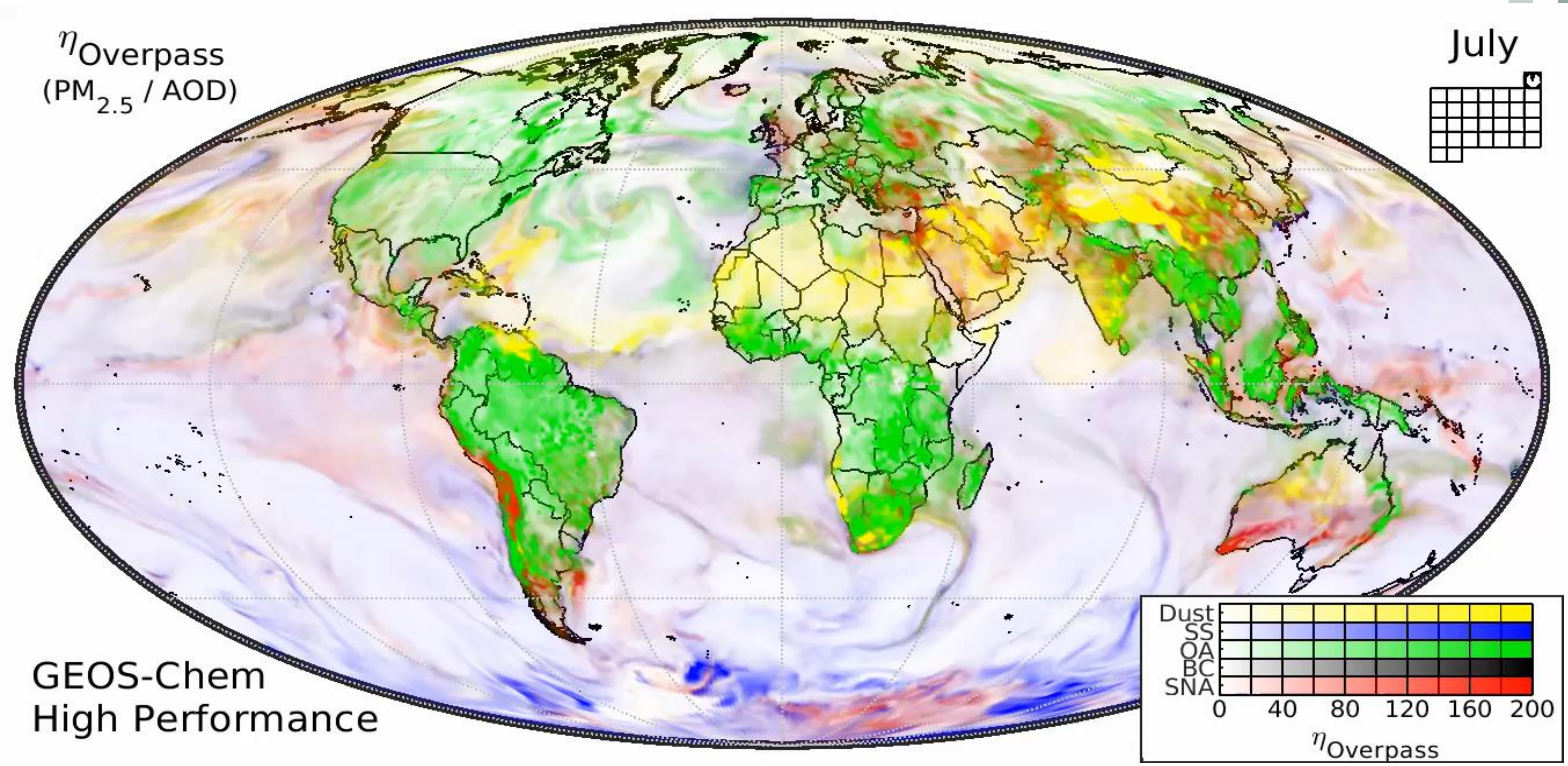


Chemical Transport Models (CTMs) can provide a spatiotemporally complete atmospheric representation of AOD to $PM_{2.5}$ relationship (η).

- A CTM is a computer model used to simulate atmospheric composition.
- CTMs combine the equations that govern atmospheric physics and chemistry with global meteorology and emissions.
- CTMs provide a complete representation of the atmosphere.
  - ➢ **Gives all the information needed to represent η**



CTMs capture the many factors that impact *η*.



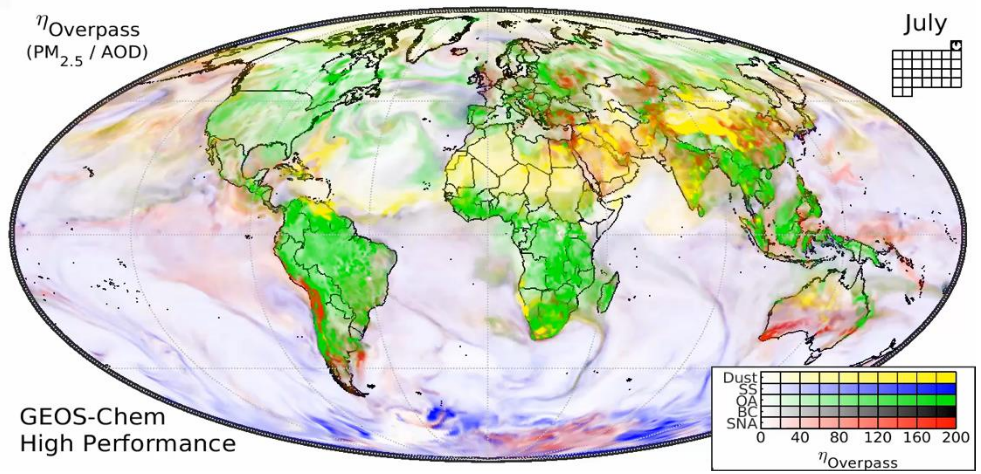
Simulations represent the spatiotemporal variation of η.

Strengths:

- Spatial/Vertical variability driven by local emissions and meteorology
- Temporal variability; compensates for satellite overpass and sampling
- Globally applicable without additional ground-based constraints

Limitations:

- Coarse resolution (50-100 km)
- Only as good is its inputs (e.g., emissions)
- Challenging to represent individual events
- Intensive to run/dependent on available model results



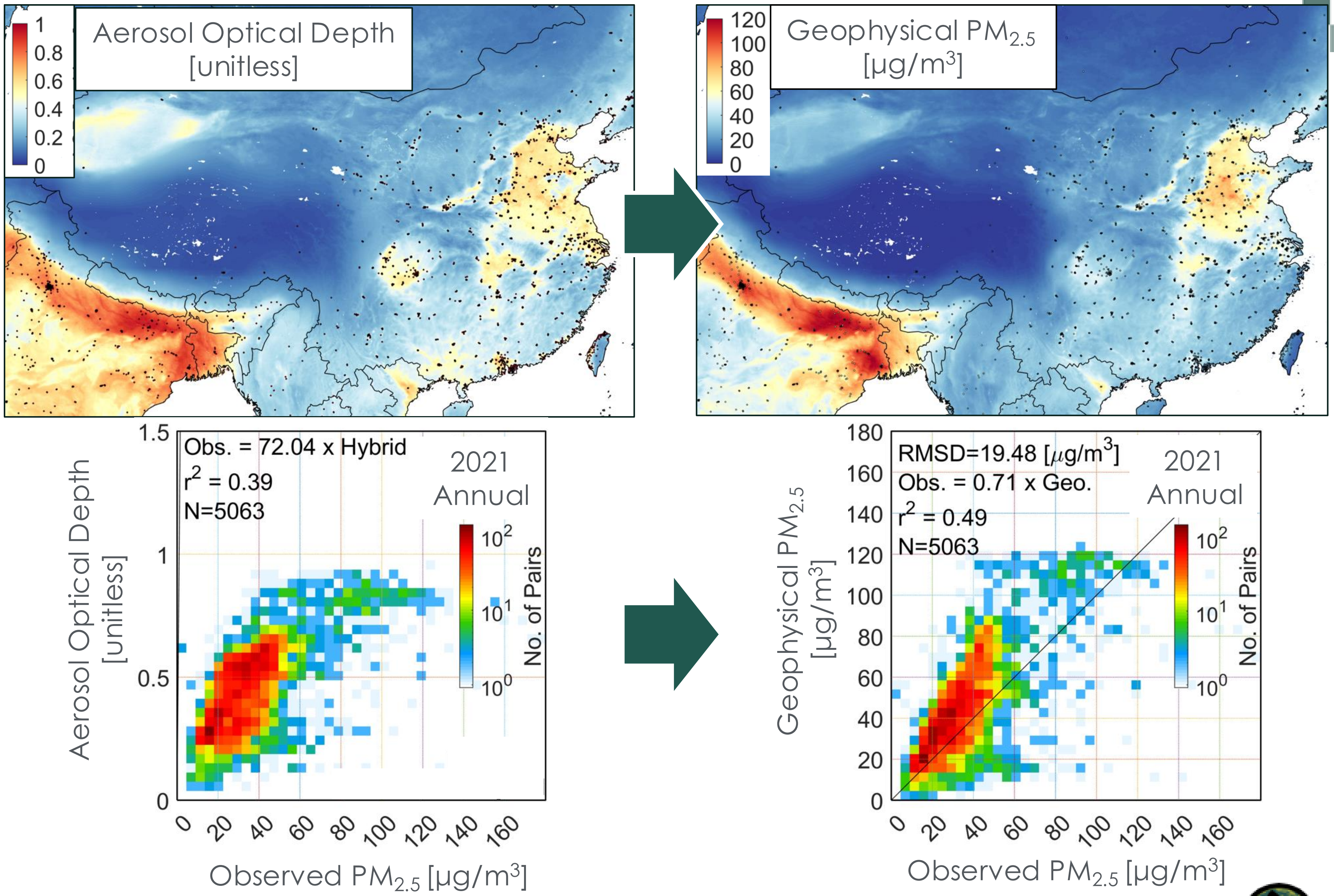
Geophysical $PM_{2.5}$ can use a CTM-based η to relate AOD to $PM_{2.5}$.

Strengths:

- Successfully captures major spatial features
- Allows AOD to be directly converted into $PM_{2.5}$ (multiply by η)
- Often improves correlation with measured $PM_{2.5}$ compared to AOD alone

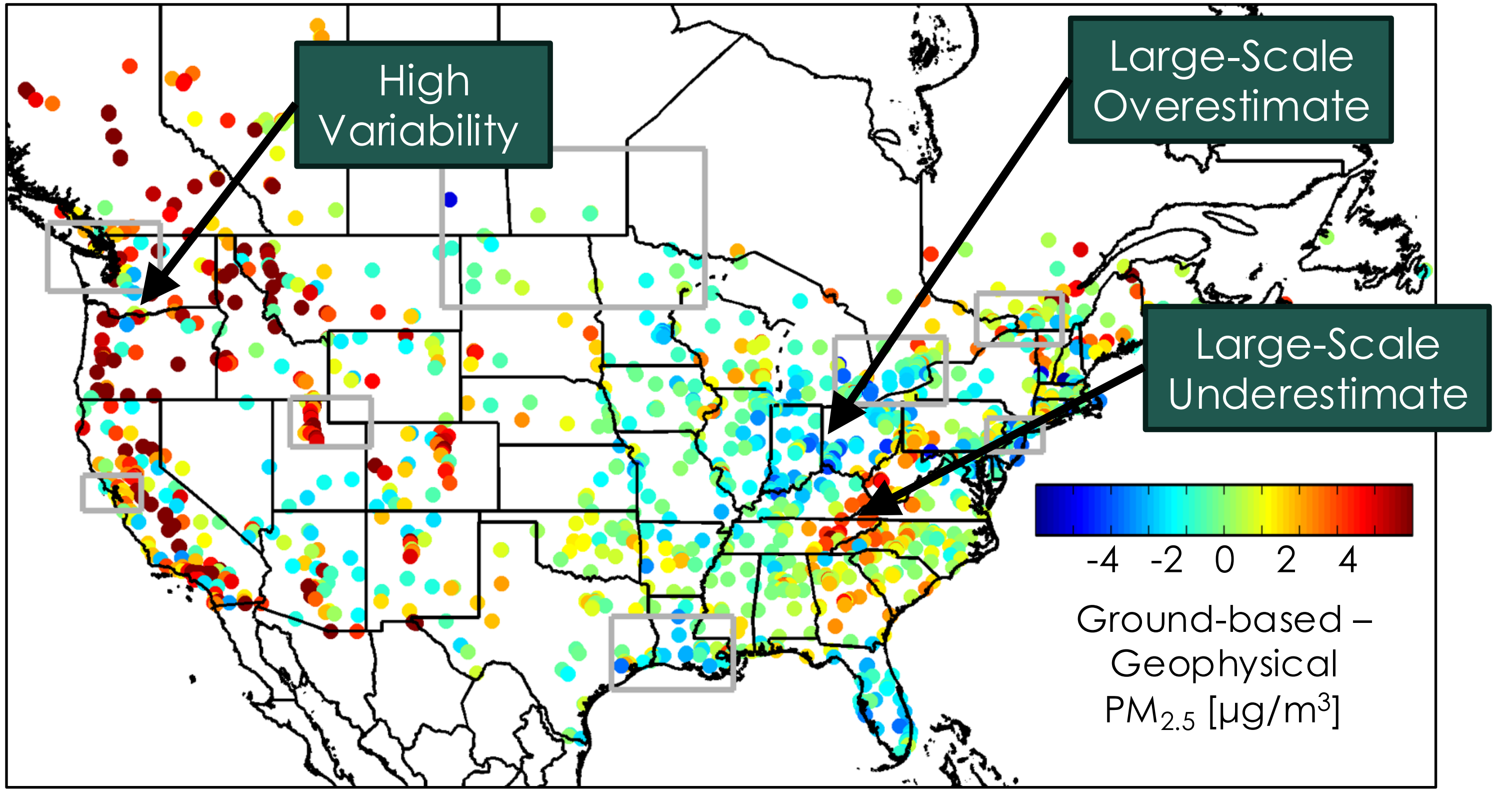
Challenges:

- Potential for bias
- Large amount of scatter with observations
- Most effective for larger spatiotemporal scales
- Difficult to identify accuracy without validation



Hybrid $PM_{2.5}$ builds on Geophysical $PM_{2.5}$ using ground-based observations.

- Hybrid methods use statistical training to predict **the bias** between geophysical $PM_{2.5}$ and ground-based observations.
- Uses statistical model to predict limitations associated with AOD retrievals or η simulation.



van Donkelaar et al., ES&T, 2015

Multiple Linear Regression (MLR) can predict geophysical bias in $PM_{2.5}$.

- Multiple Linear Regression (MLR) provides an intuitive approach to predicting the geophysical bias with independent variables

$$\underbrace{Y}_{\text{Dependent Variable (outcome)}} = \underbrace{\beta_0}_{\text{Intercept}} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \underbrace{\varepsilon}_{\text{Error Term}}$$

Dependent Variable (outcome)

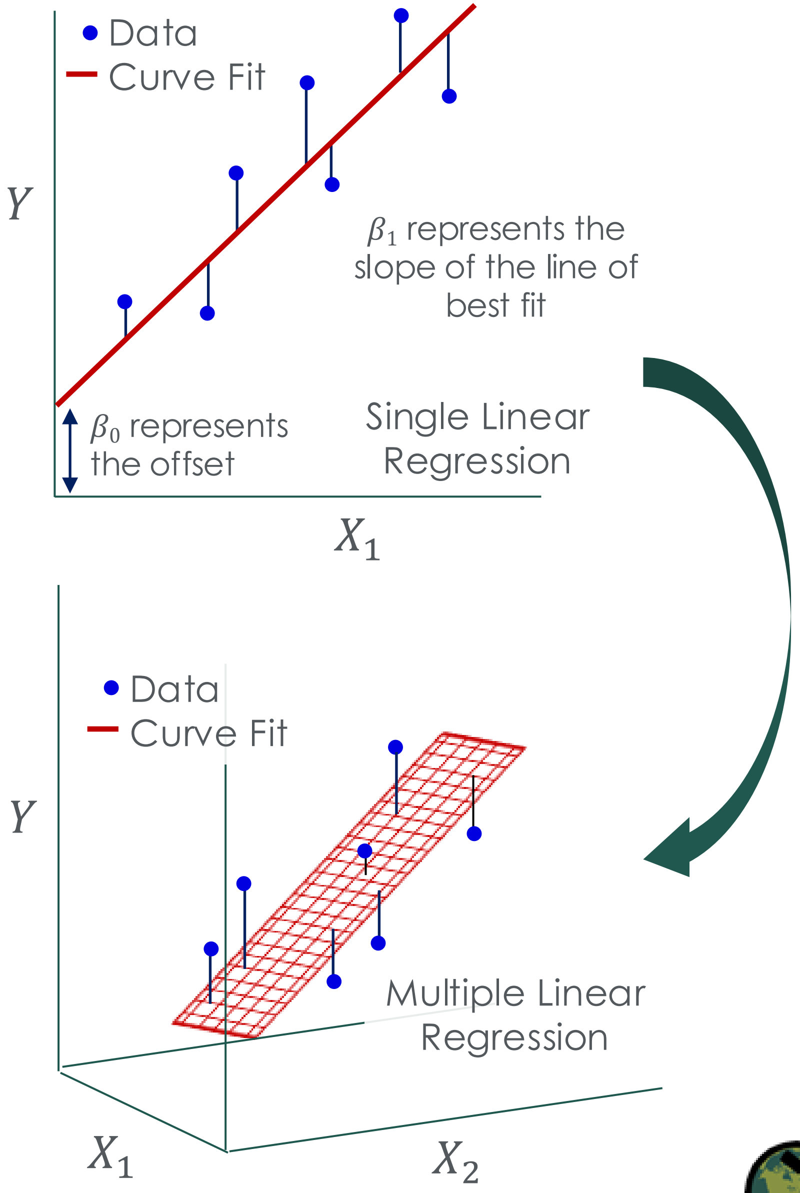
Intercept

Predictor Coefficients

Independent Variables (Predictors)

Error Term

- MLR extends single linear regression (i.e., one predictor)
  - Slopes of each independent variable, X_n, is represented by β_n.
  - β_n can be thought of as the linear response in Y to changes in predictor X_n.



Geographically Weighted Regression (GWR) allows spatially varying predictor coefficients.

Dependent Variable$_{i,j}$ (outcome)

Intercept$_{i,j}$

Value at Location *i,j*

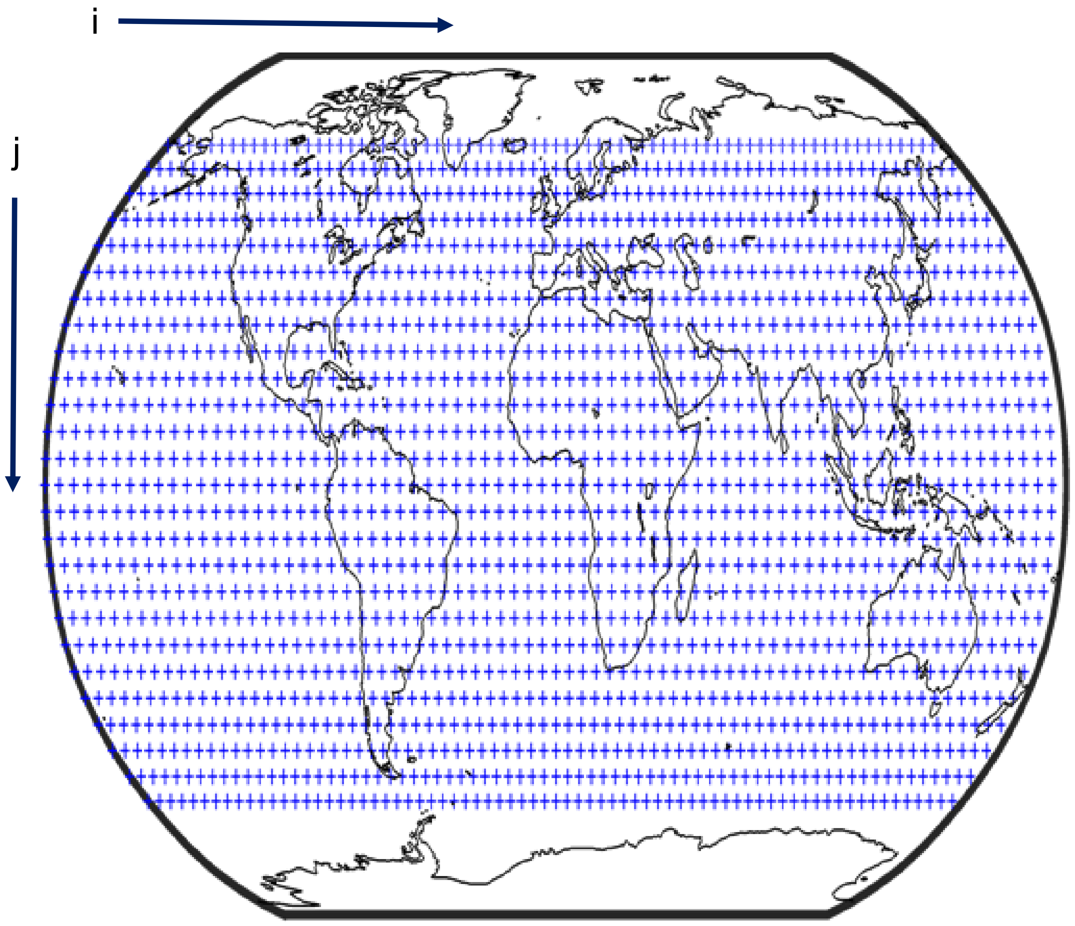
Predictor Coefficients$_{i,j}$

$$Y_{i,j} = \beta_{0,i,j} + \beta_{1,i,j}X_{1,i,j} + \beta_{2,i,j}X_{2,i,j} + \cdots + \beta_{n,i,j}X_{n,i,j} + \varepsilon_{i,j}$$

Independent Variables$_{i,j}$ (Predictors)

Error Term$_{i,j}$

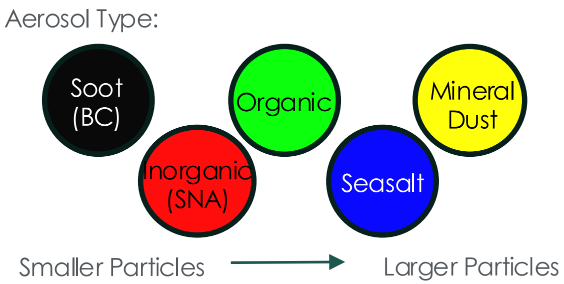
- Individual predictor variables not spatially constant
- Map of predictor coefficients are applied to map of predictor variables
- MLR calibrated at discrete locations on a global grid
  - Sites weighted by distances from each monitor
  - Sites beyond a distance threshold are excluded
  - Grid can be constant or vary with training site density



Each + represents a discrete locations considered for GWR.

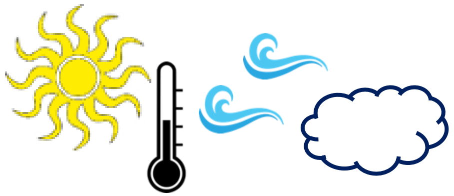
Effective predictors in GWR relate to uncertainties in AOD and η.

Common predictors related to AOD or *η* uncertainty include:



Composition

Representative of different emissions sources, vertical profiles, growth factors



Meteorology

Potential impacts on vertical profile, aerosol formation, emissions

Land type

Potential impacts on AOD retrieval, emissions

Elevation/Elevation changes

Potential impacts on vertical profile, high-resolution (subgrid CTM) features

Predictor insights combine for total geophysical bias prediction.

$$(GM\ PM_{2.5} - SAT\ PM_{2.5})_{i,j} = \beta_{0,i,j} + \Sigma\beta_{SPECi,j,k}SPEC_{i,j,k} + \Sigma\beta_{METi,j,k}\,MET_{i,j,k} + \beta_{ELEV,i,j}ELEV_{i,j} + \beta_{LTi,j,k}LT_{i,j,k}$$

Monitor-Geophysical $PM_{2.5}$ Difference

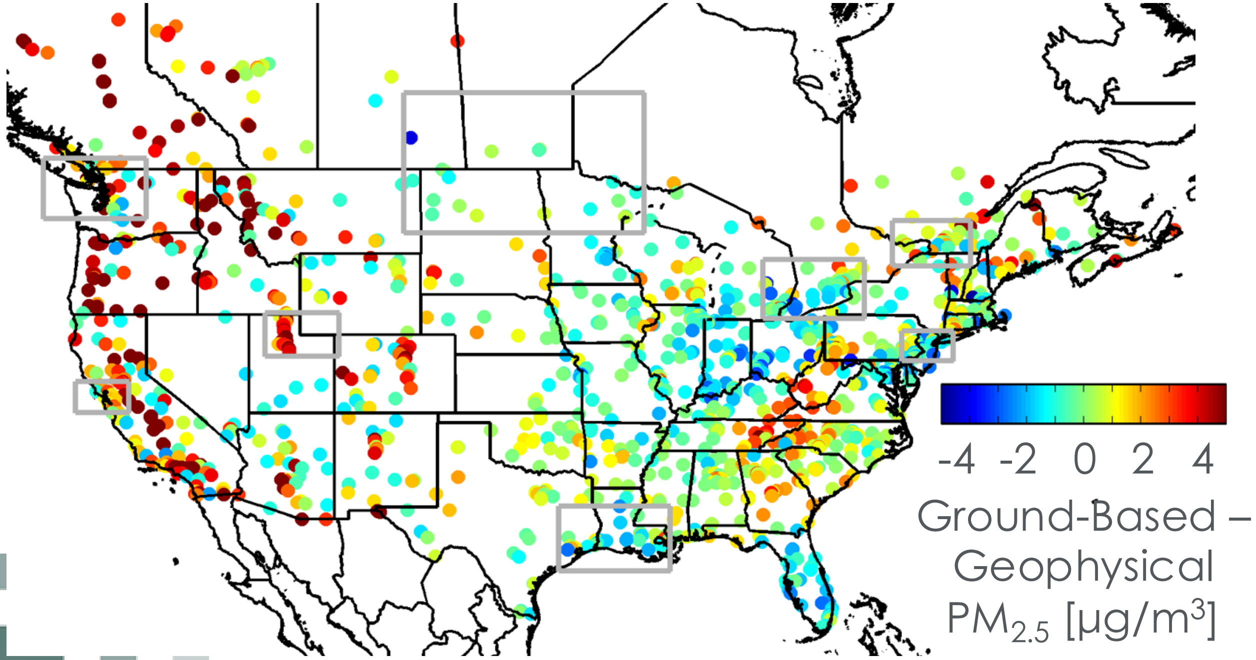
Simulated Composition

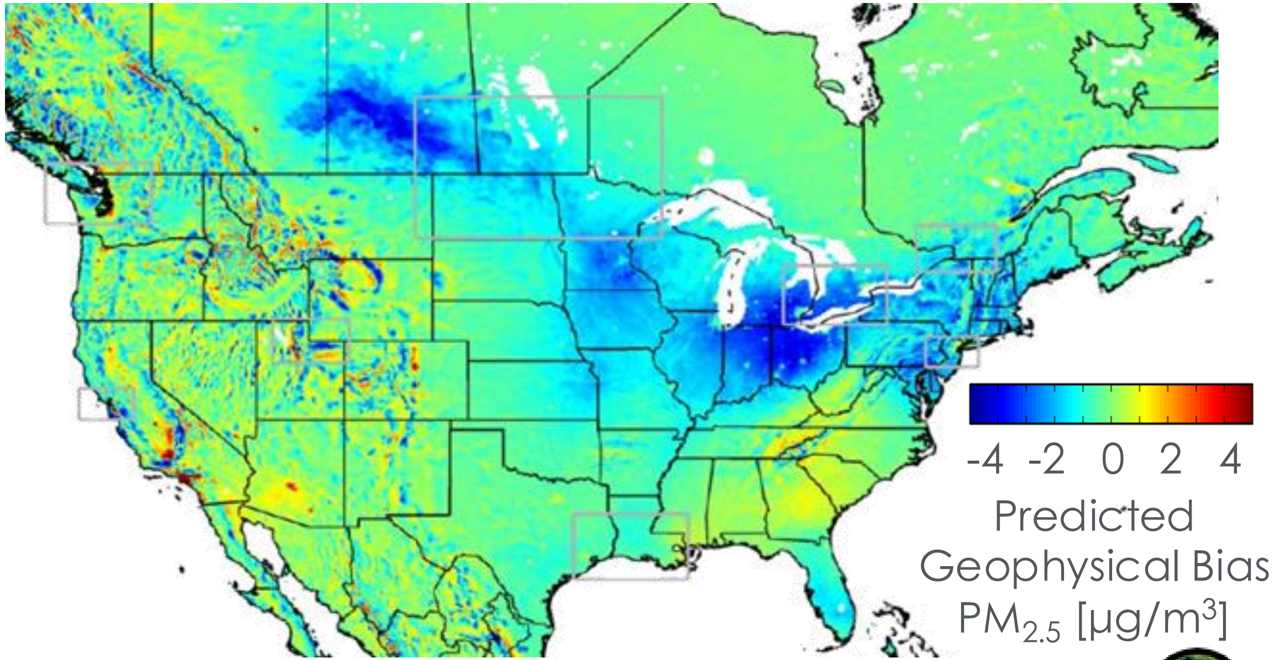
Meteorology Parameters

Elevation Parameters

Land Type Parameters

- Captures observed bias features
- Predicts features in unobserved locations





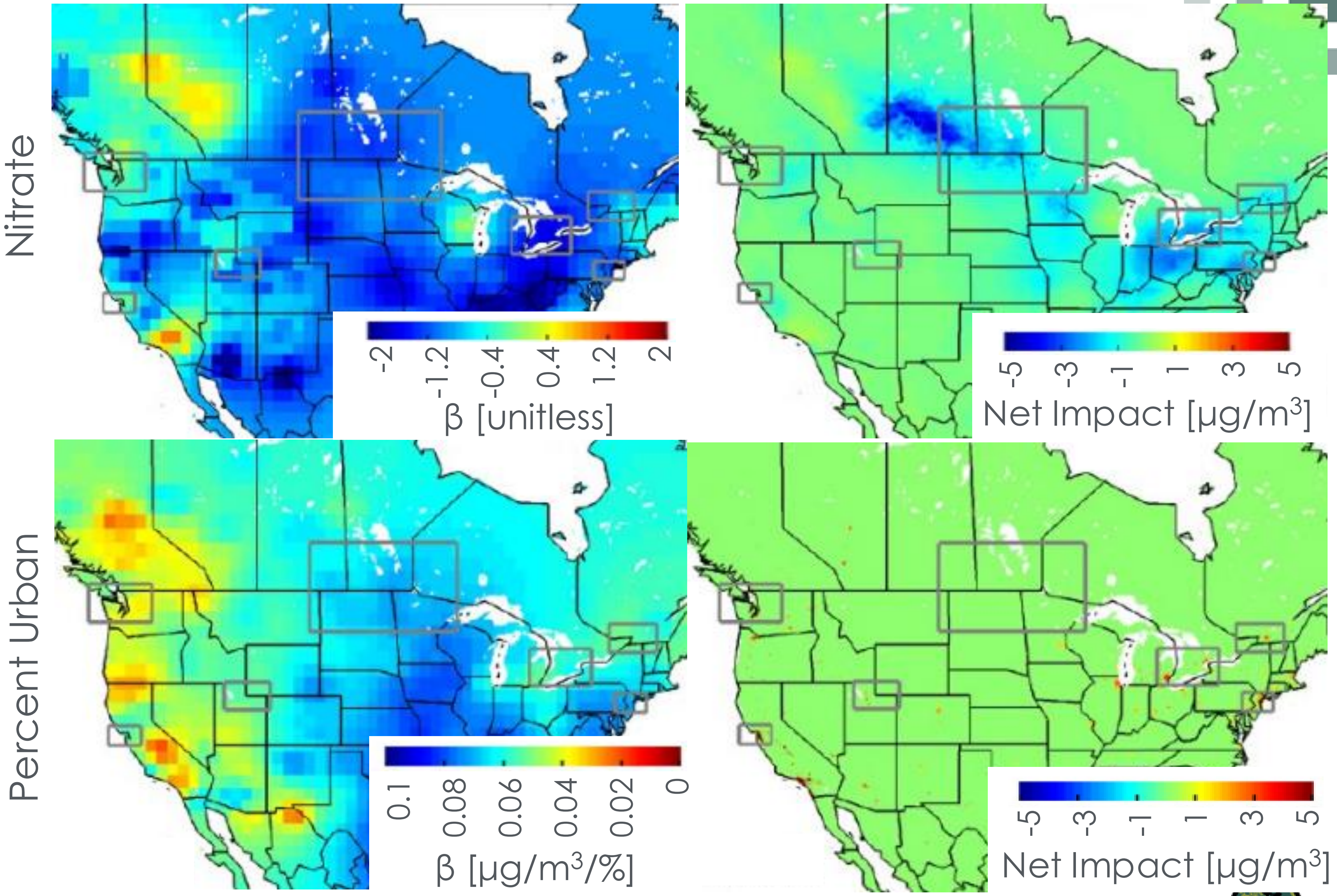
Individual predictors offer unique insights and directions.

Nitrate:

- Fairly constant β over impactful regions
- Explains regional geophysical overestimate
- Chemical/Emission Effect?

Percent Urban:

- Clear East-West gradient in β
- Impacts fine-scale geophysical bias
- η Resolution/AOD retrieval effect?



van Donkelaar et al., ES&T, 2015

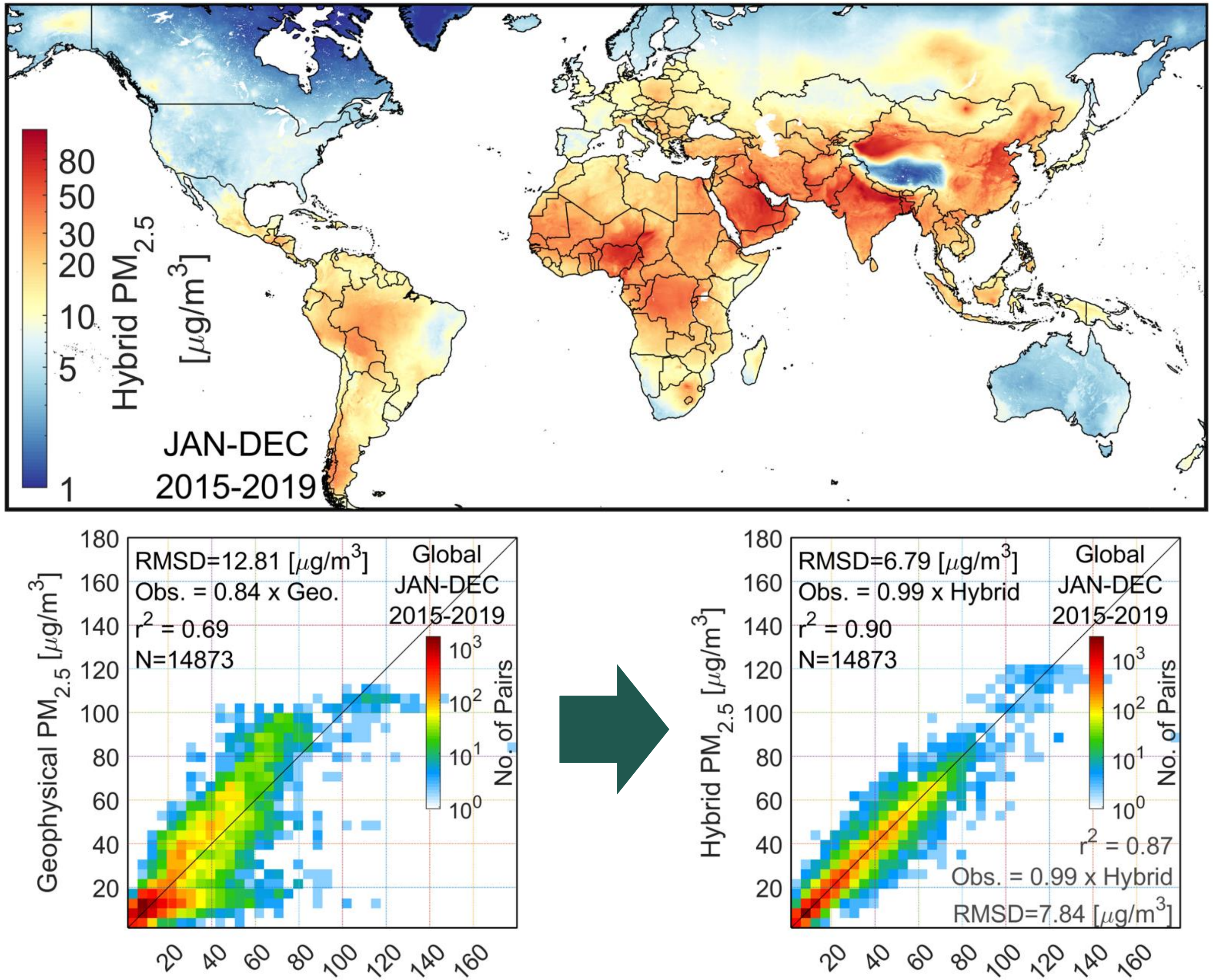
Hybrid $PM_{2.5}$ globally refines purely Geophysical $PM_{2.5}$.

Strengths:

- Maintains global applicability of geophysical methods
- Includes on ground-based constraints
- Improves geophysical $PM_{2.5}$

Limitations:

- Requires ground-based observations
- Effective model setup is complicated
  - Model framework
  - Predictor variable selection
- Models need to be globally applicable
  - Avoid overfitting
  - Appropriate **cross validation** is key!

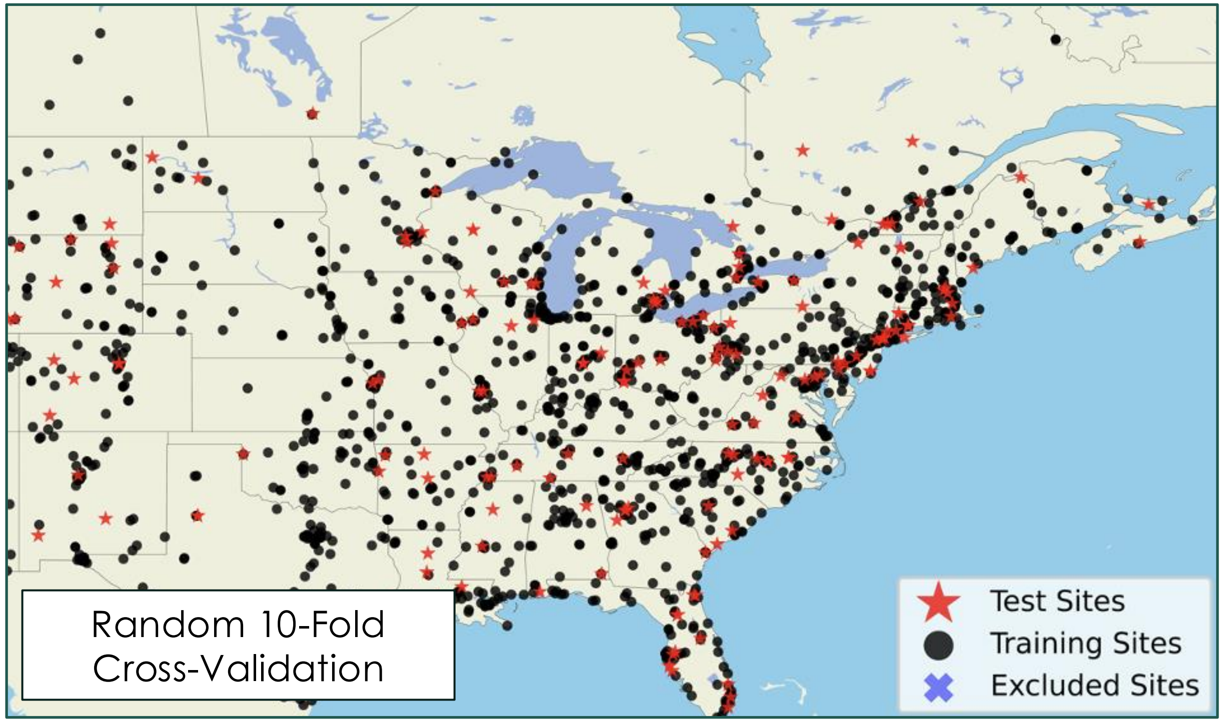


Cross Validation (CV) allows for independent evaluation of uncertainty.

Method selection must balance complexity, accuracy, and representation, based on observation site properties – must maintain independence of CV dataset.

Random k-Fold Cross-Validation:

- Train separate models k times (e.g. k=10), randomly withholding data sites from training for subsequent validation
- Simple implementation
- Clustered or related data sites will cause an over-representation of accuracy
  - Collocated sites aren't independent!



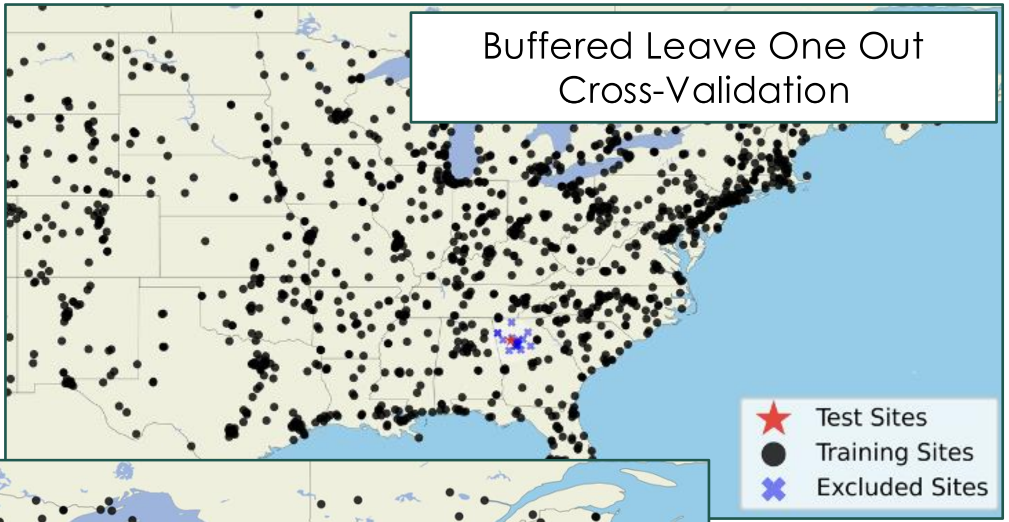
Shen et al., ES&T, 2026

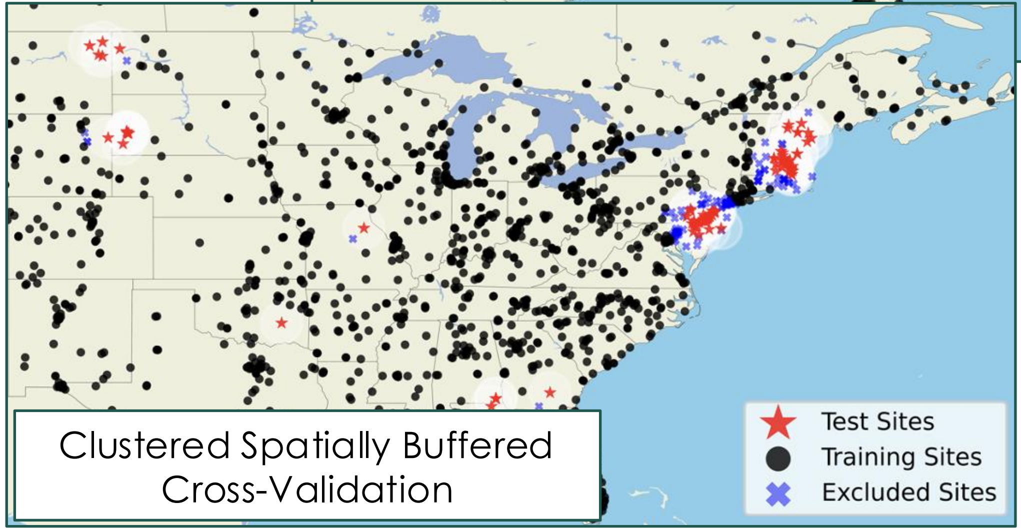
Spatially buffered CV reduces the impact of spatial auto-correlation.

Method selection must balance complexity, accuracy, and representation, based on observation site properties – must maintain independence of CV dataset.

Spatially Buffered Cross-Validation:

- Adds a spatial buffer to increase independence of CV sites
- Train separate models, withholding data from training for subsequent validation, plus exclude additional monitors within spatial buffer from training
- Multiple Implementations:
  - Buffered Leave One Out (BLOO)
    - Withholds one site out at a time
    - Effective, but computationally expensive
  - Buffered Leave Cluster Out (BLeCO)/Buffered Leave Isolated Sites and Cluster Out (BLISCO)
    - Withholds groups of sites plus buffer
    - Enhanced computational efficiency





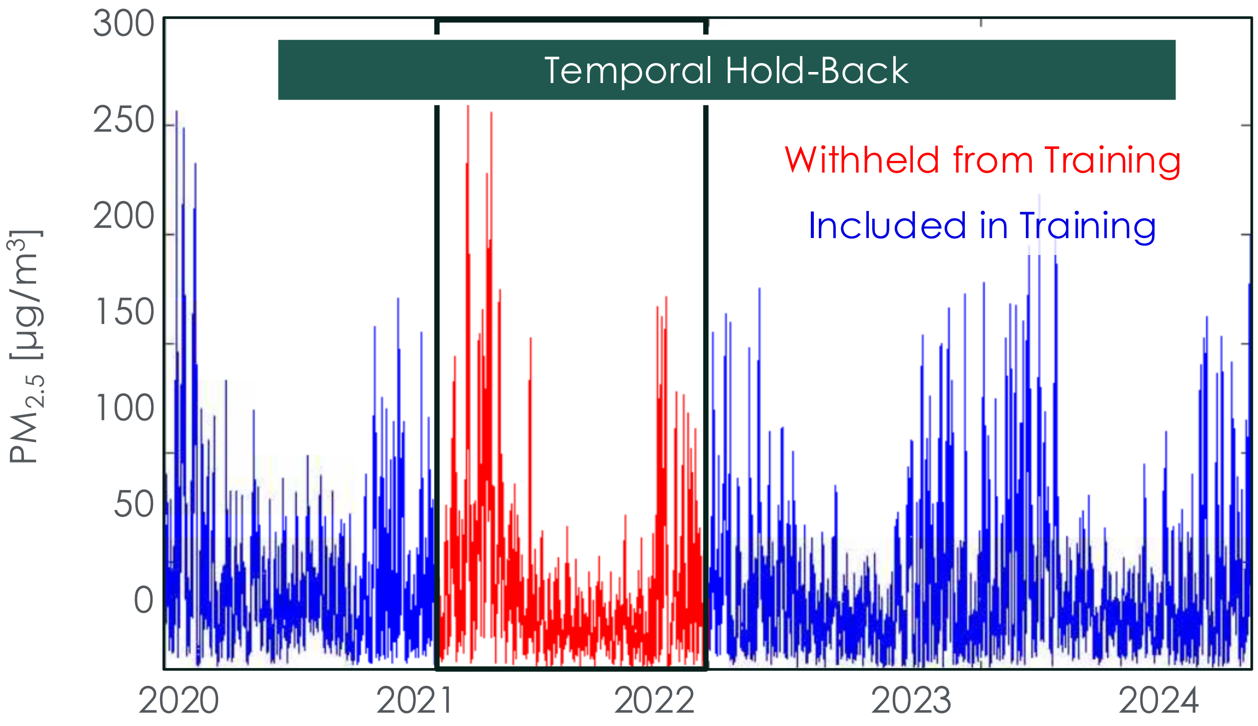
Shen et al., ES&T, 2026

Temporal hold-back CV reduces the impact of temporal auto-correlation.

Method selection must balance complexity, accuracy, and representation, based on observation site properties – must maintain independence of CV dataset

Temporal Hold-Back

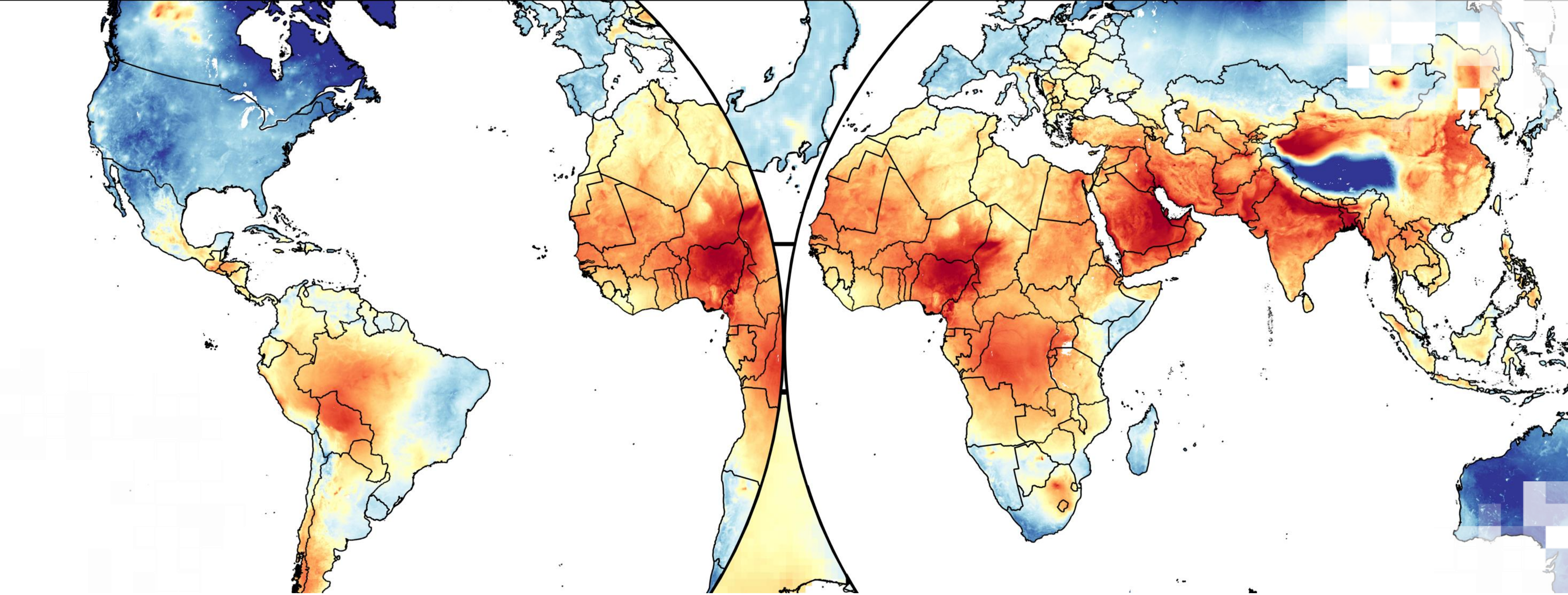
- Specific to cases where models are applied across time
- Informative for temporal uncertainties (i.e., time series)
- Withhold all observations taken during a specified period for subsequent validation
- Can be combined with spatially buffered analysis



Shen et al., ES&T, 2026

Summary – Geophysical and Hybrid $PM_{2.5}$ Estimation

- Geophysical methods to estimate $PM_{2.5}$ from satellite retrievals of AOD avoid the use of ground-based monitors by using the theory-based, physically-driven relationship between total column AOD and $PM_{2.5}$.
- Chemical Transport Models (CTMs) can be used to globally provide the parameters needed to enable geophysical methods.
- Hybrid methods build on geophysical methods by using ground-based $PM_{2.5}$ observations to understand, predict, and correct bias in geophysical values using a statistical framework.
- Statistical frameworks require effective cross validation for accurate representation of uncertainty and therefore must ensure strong independence of validation points.

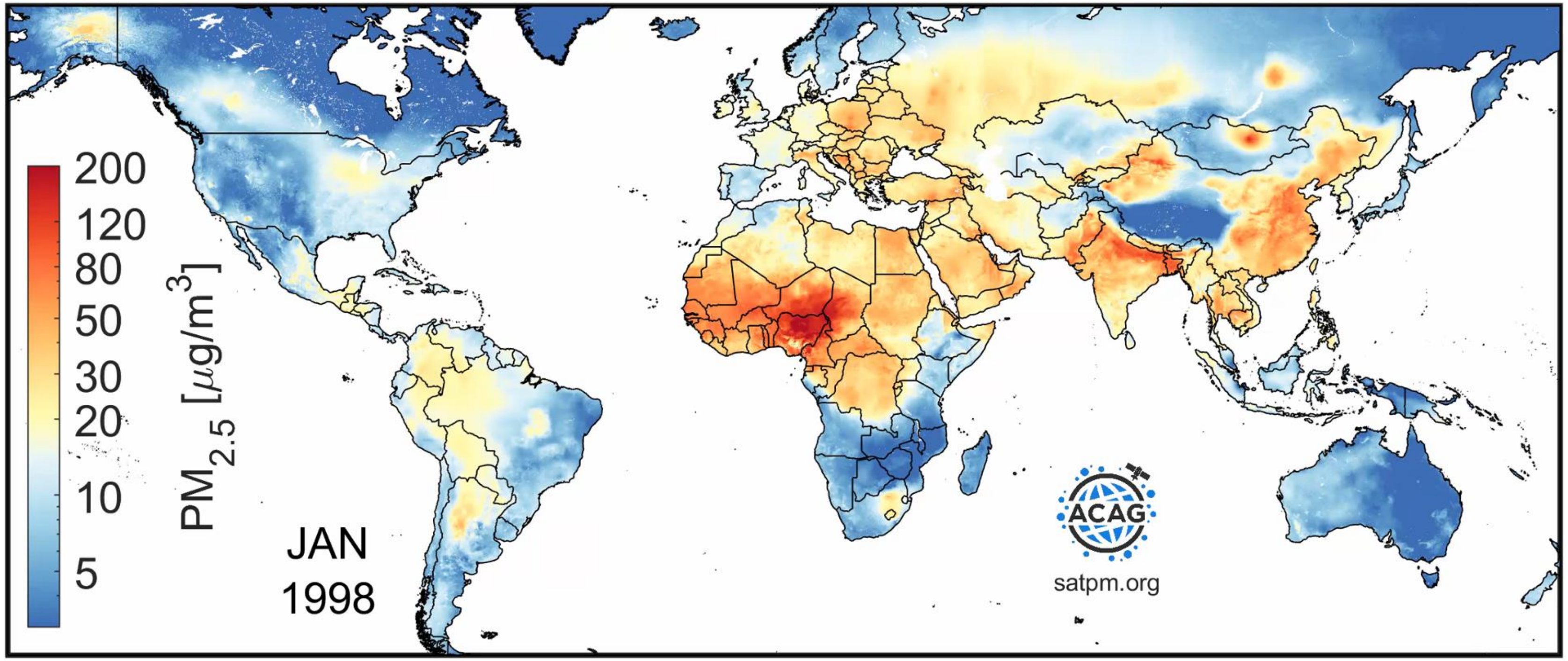


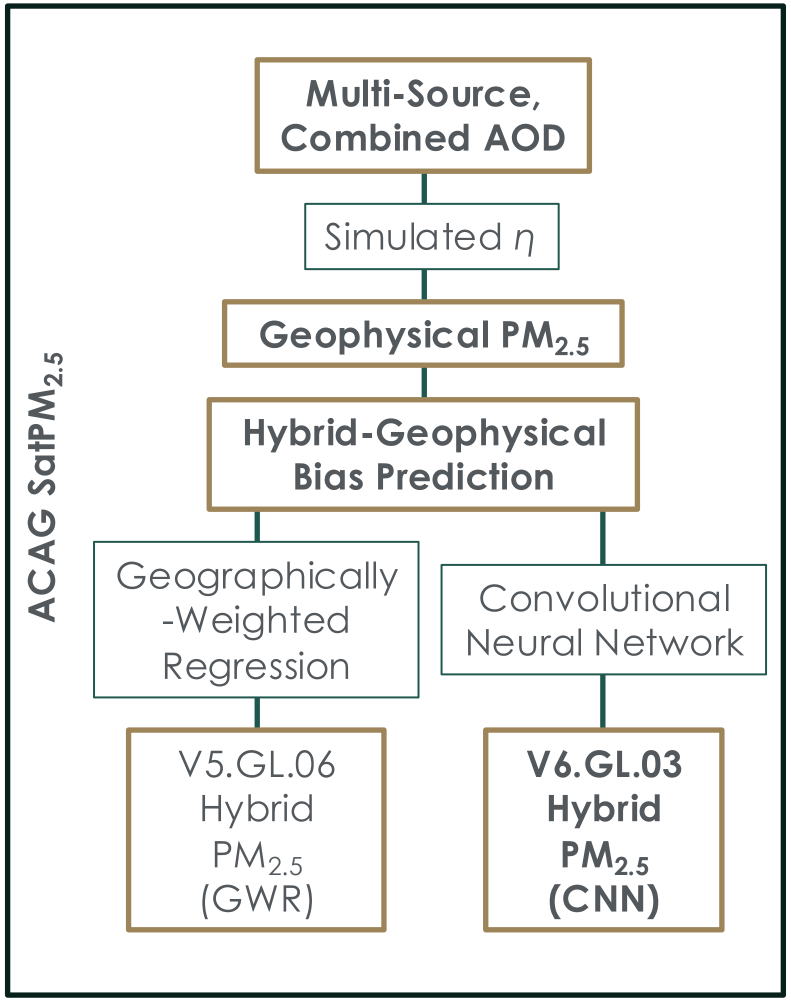
SatPM$_{2.5}$ @ ACAG Washington University in St. Louis

Atmospheric Composition Analysis Group

WashU

ACAG SatPM$_{2.5}$ provides public access to a global PM$_{2.5}$ dataset.



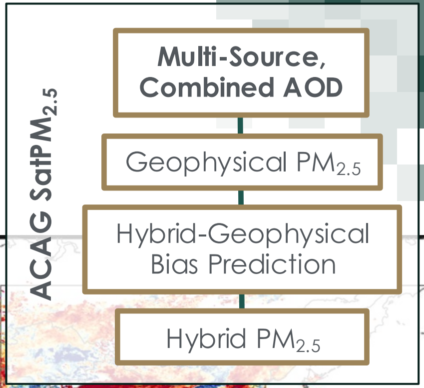


SatPM$_{2.5}$ provides a consistent, global, long-term, PM$_{2.5}$ dataset that is:

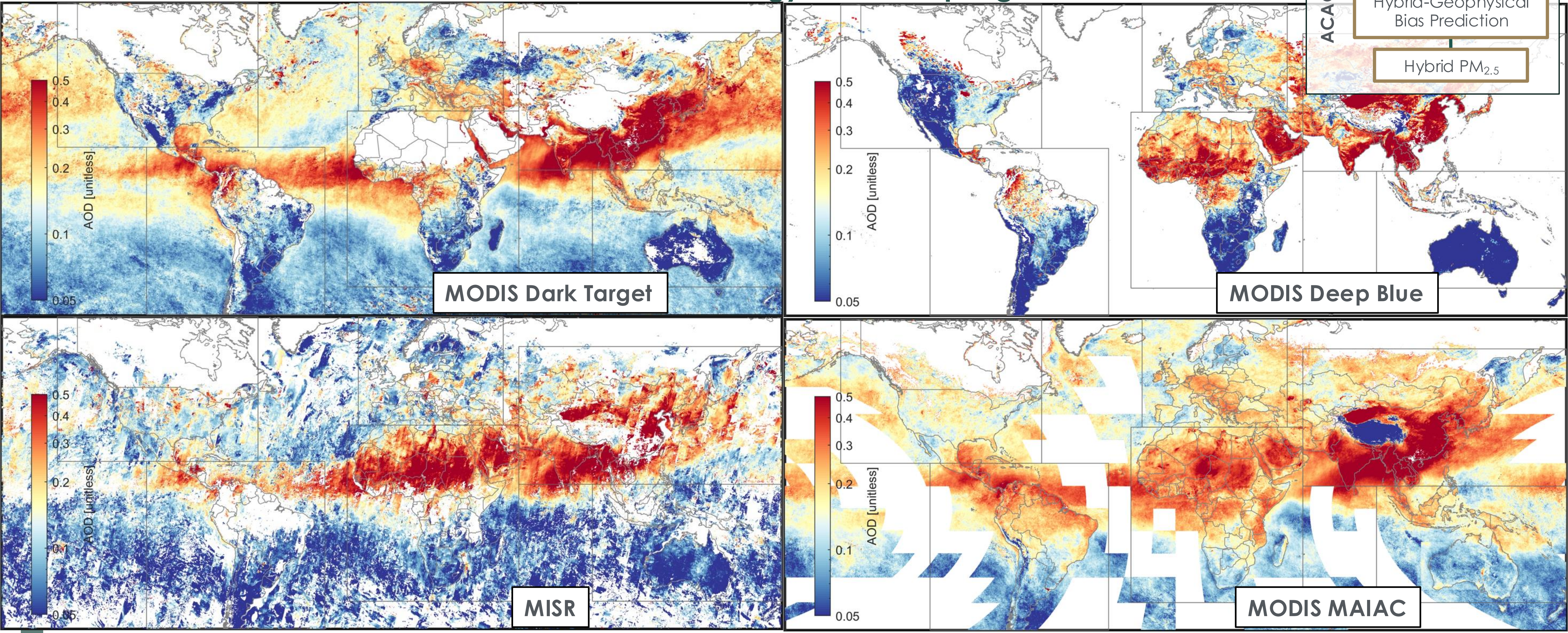
- Observationally constrained using satellite AOD and ground-based monitors
- Based on a hybrid-geophysical framework
- Monthly, 1998-2024 (updated annually)
- High-resolution (0.01°×0.01° grid)
- Publicly available from: http://satpm.org

Atmospheric Composition Analysis Group

Different AOD retrievals are designed for different conditions.

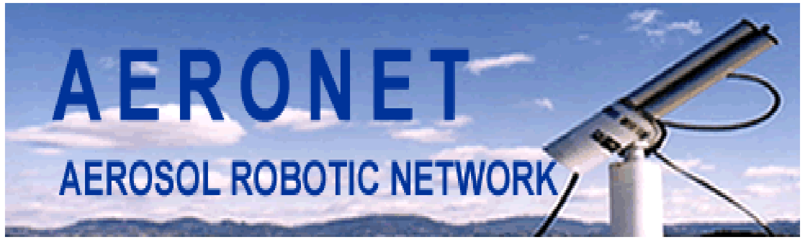


Differences result from instrumentation, methodology, and sampling.

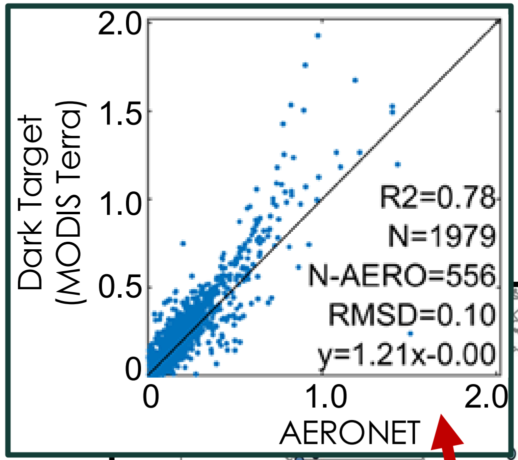


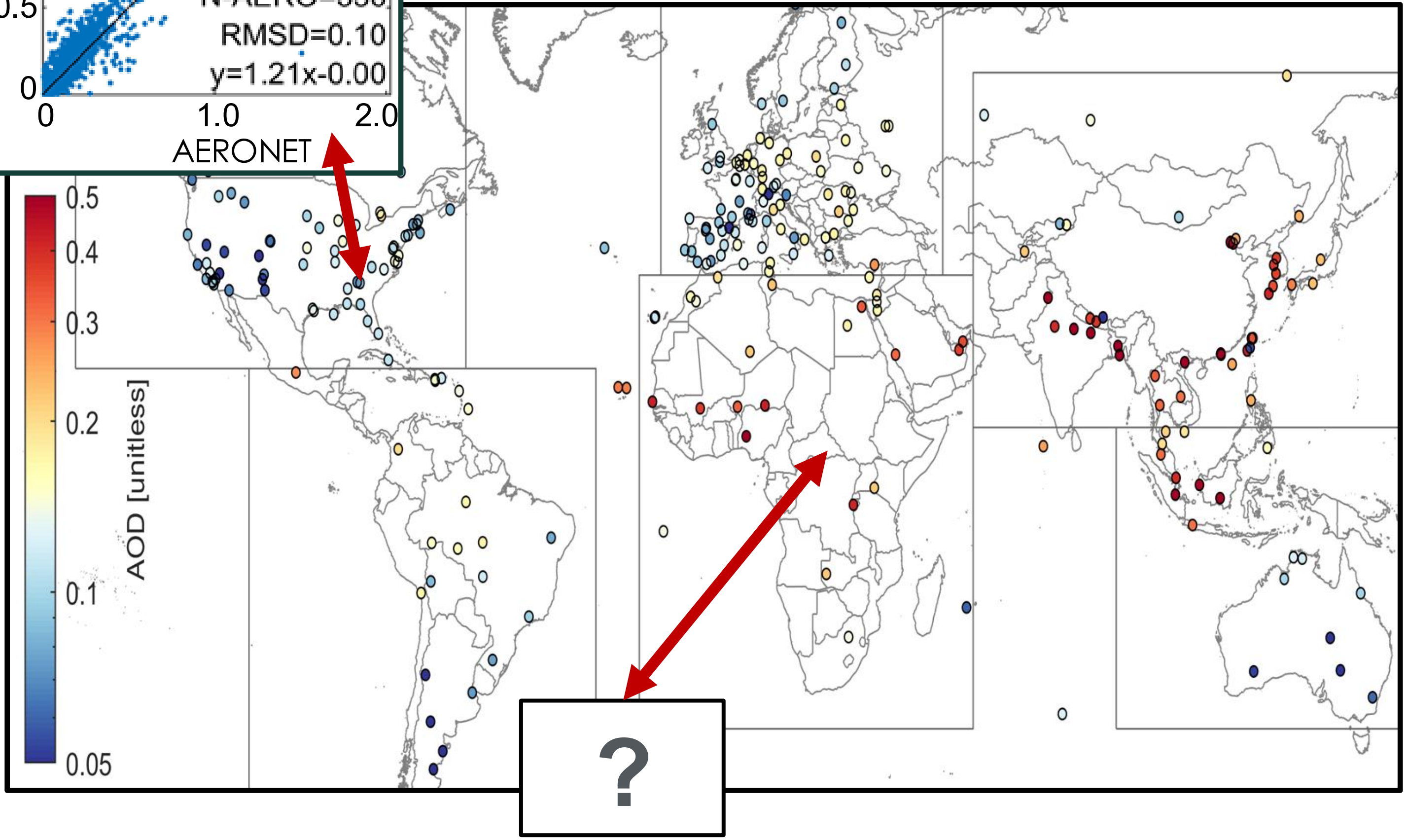
Mean Terra AOD for April 2014

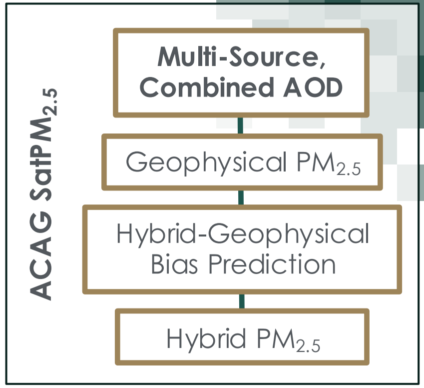
AERONET allows consistent evaluation of AOD uncertainty.



AERONET AOD:
- Global
- Accurate
- Long-term
- allows direct evaluation of satellite retrievals

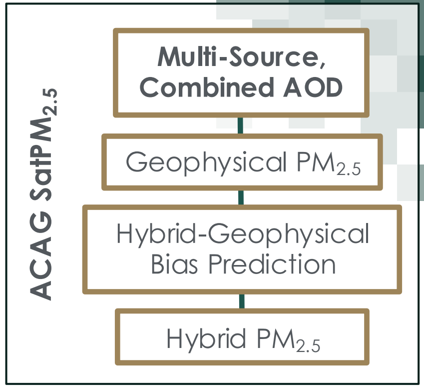






➢ **How to globally extend location-specific comparisons outside network locations/times?**

Local uncertainty can be inferred using locations with similar conditions.

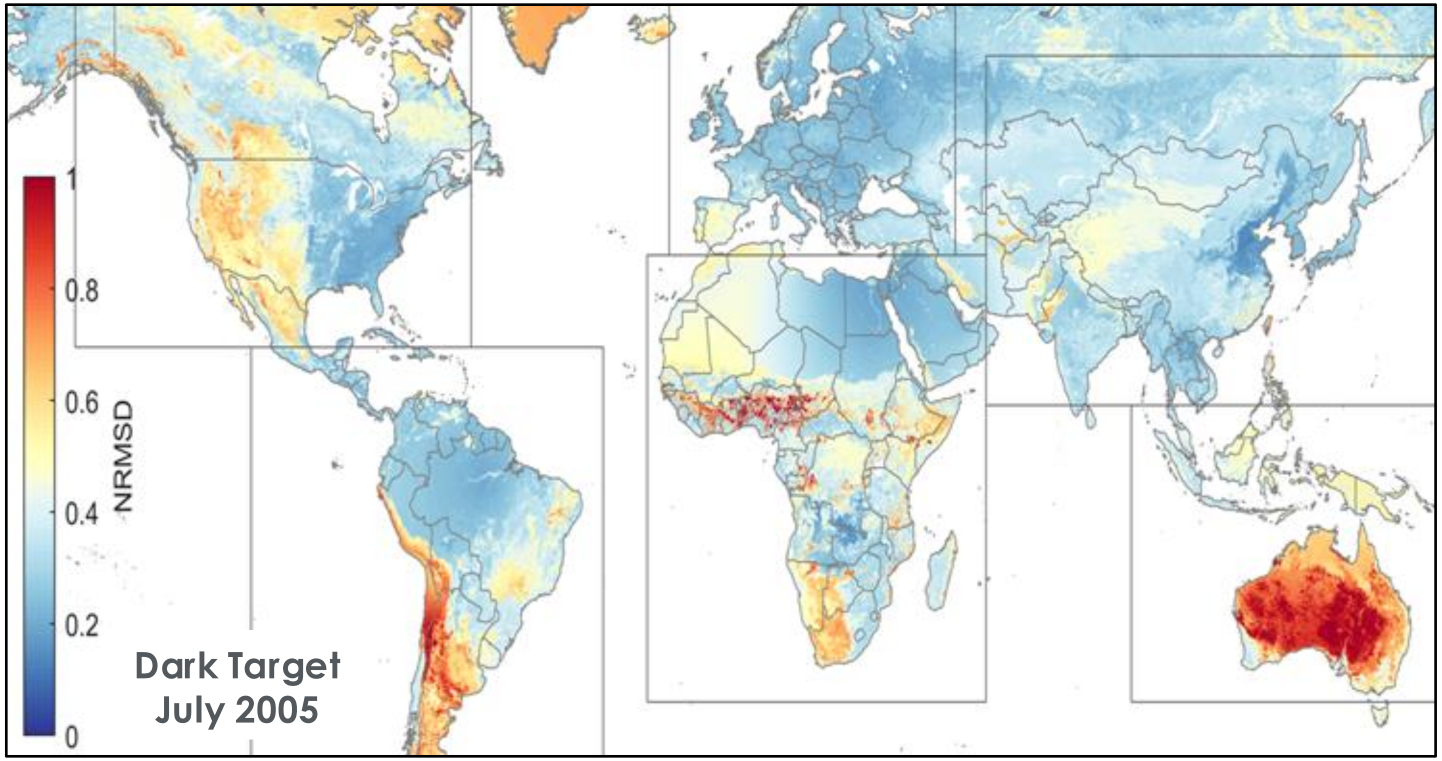


Perform AOD comparisons using all AERONET sites based on:

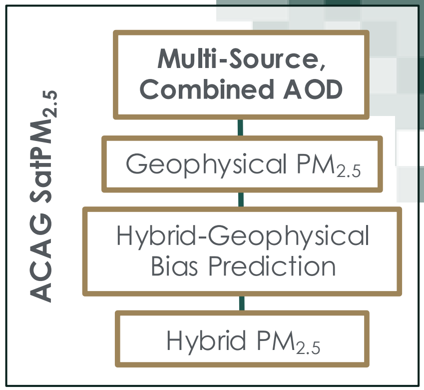
- Retrieval
- Month
- Land Type
- Normalized Difference Vegetation Index
- Proximity to Location of Interest

➢ **Recombine based on local conditions**

Continuous, Consistent, Global Error Definition



Global Evaluation: Use Each AOD Dataset at its Best

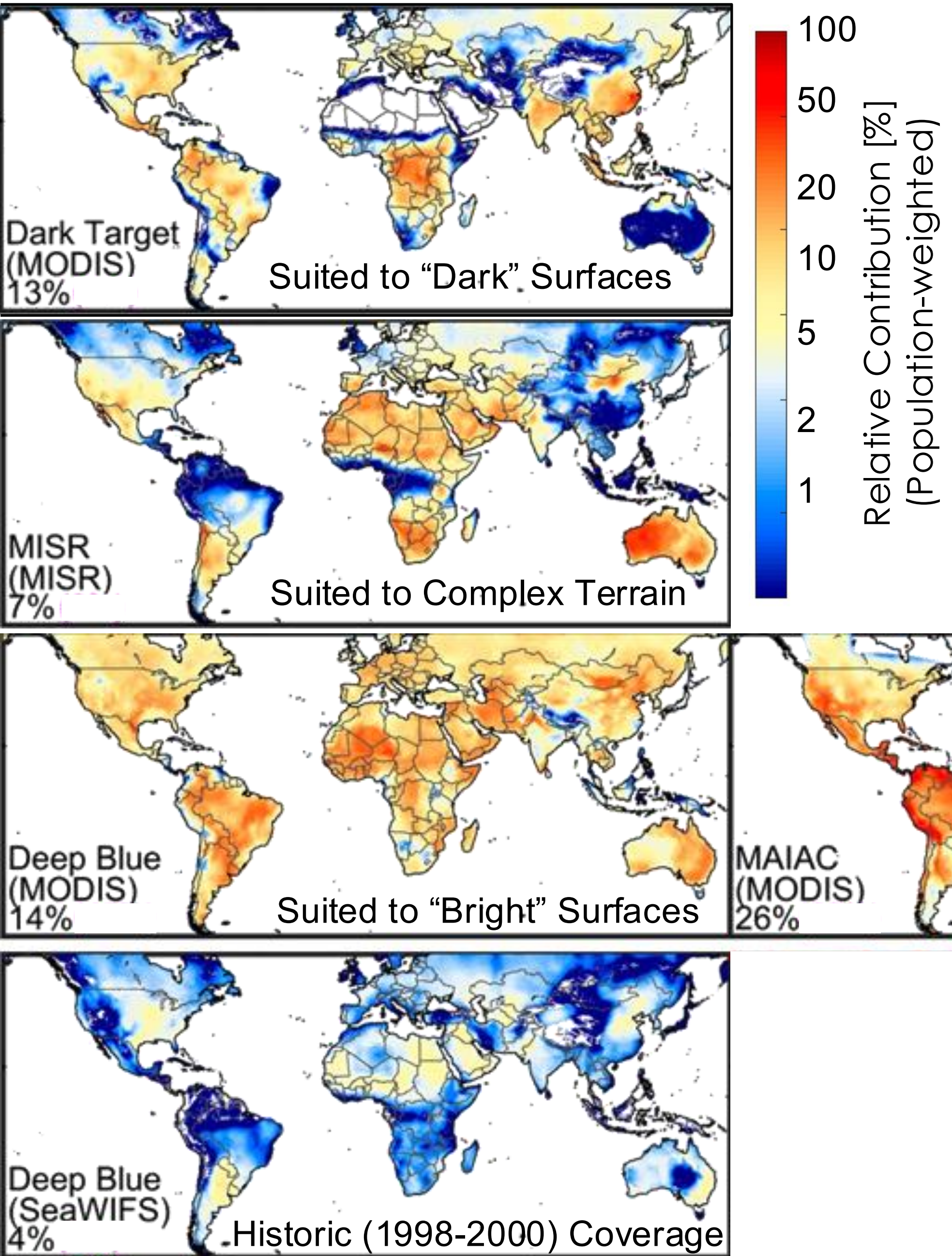


Different AOD sources combined using a weighted average based on local uncertainties:

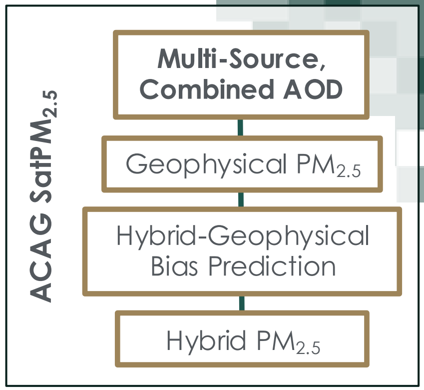
- NRMSD
- Bias (i.e., slope)
- Data Density (number of observations)

$$AOD = \frac{\sum_{n=1}^{N} \frac{1}{NRMSD_n} \times \left[\frac{\Delta AOD_{adj,n}}{AOD_n}\right]^{-1} \times N_{obs,n}^2 \times AOD_n}{\sum_{n=1}^{N} \frac{1}{NRMSD_n} \times \left[\frac{\Delta AOD_{adj,n}}{AOD_n}\right]^{-1} \times N_{obs,n}^2}$$

Global Evaluation: Use Each AOD Dataset at its Best



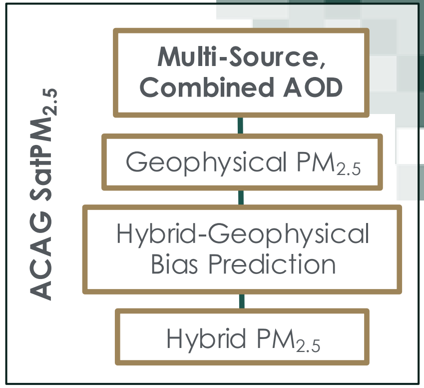
The resulting weights for each dataset are intuitive.

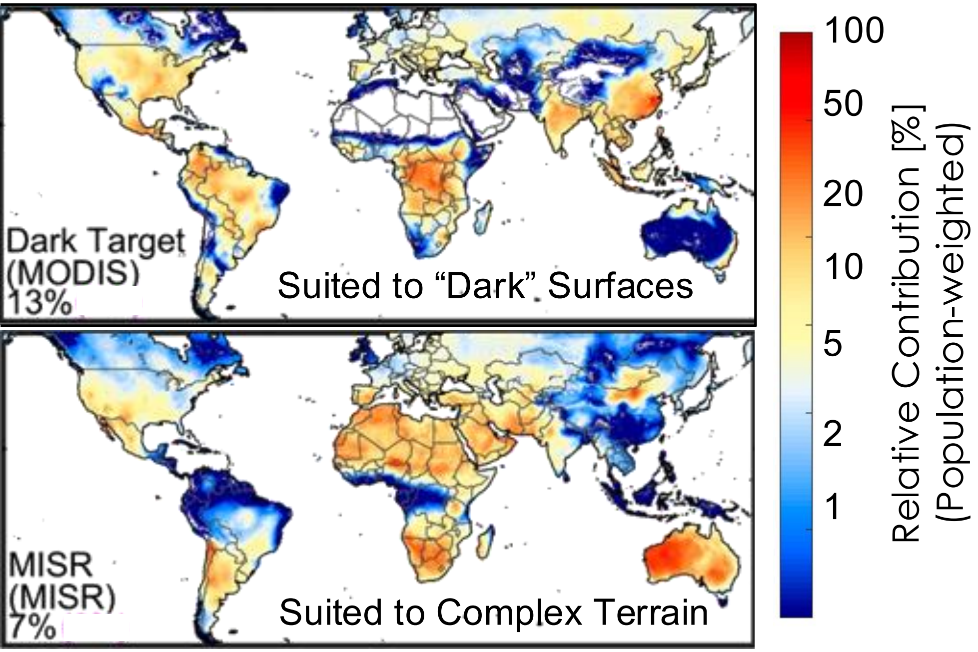


- **Combine all AOD sources weighted by $NRMSD^{-1}$, bias correction^{-1}, data density:**

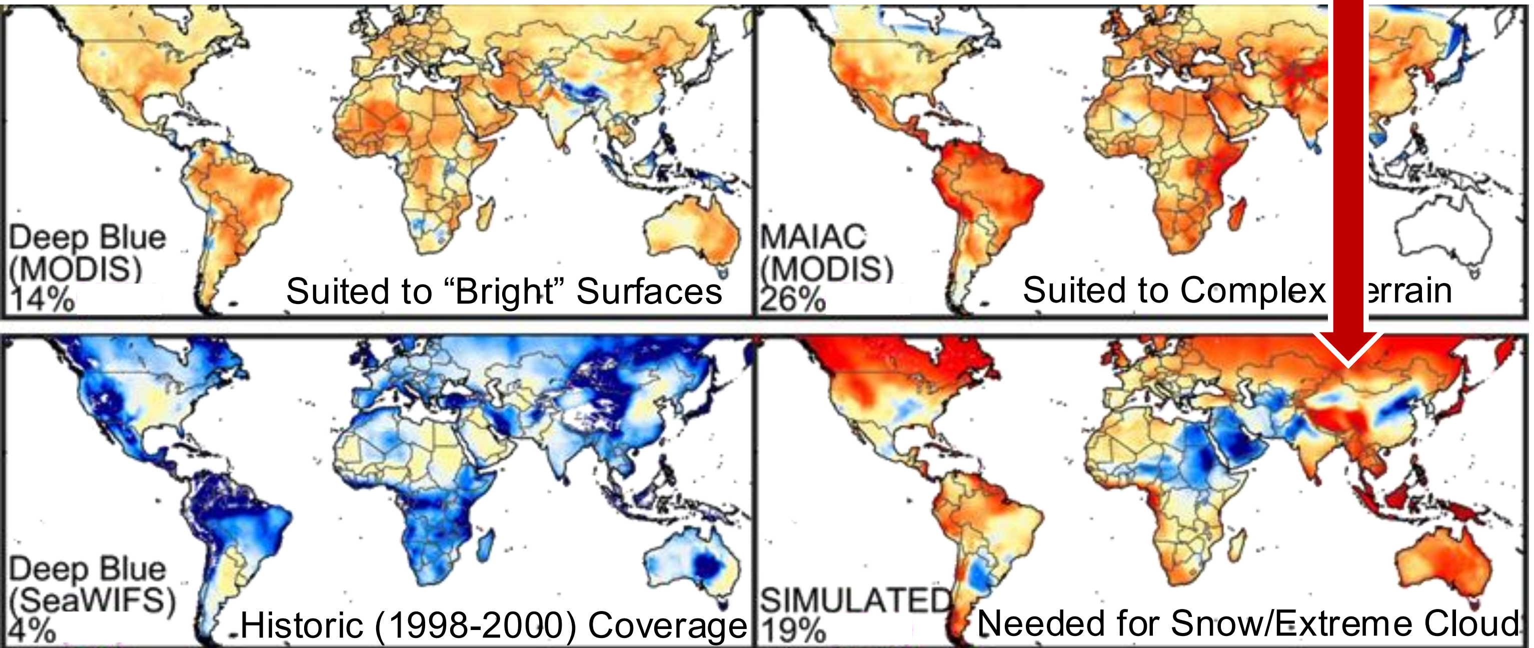
$$AOD = \frac{\sum_{n=1}^{N} \frac{1}{NRMSD_n} \times \left[\frac{\Delta AOD_{adj,n}}{AOD_n}\right]^{-1} \times N_{obs,n}^2 \times AOD_n}{\sum_{n=1}^{N} \frac{1}{NRMSD_n} \times \left[\frac{\Delta AOD_{adj,n}}{AOD_n}\right]^{-1} \times N_{obs,n}^2}$$

Global Evaluation: Use Each AOD Dataset at its Best





Simulated AOD can help over unretrieved conditions.



➢ **Combine all AOD sources weighted by $NRMSD^{-1}$, bias correction^{-1}, data density:**

$$AOD = \frac{\sum_{n=1}^{N} \frac{1}{NRMSD_n} \times \left[\frac{\Delta AOD_{adj,n}}{AOD_n}\right]^{-1} \times N_{obs,n}^2 \times AOD_n}{\sum_{n=1}^{N} \frac{1}{NRMSD_n} \times \left[\frac{\Delta AOD_{adj,n}}{AOD_n}\right]^{-1} \times N_{obs,n}^2}$$

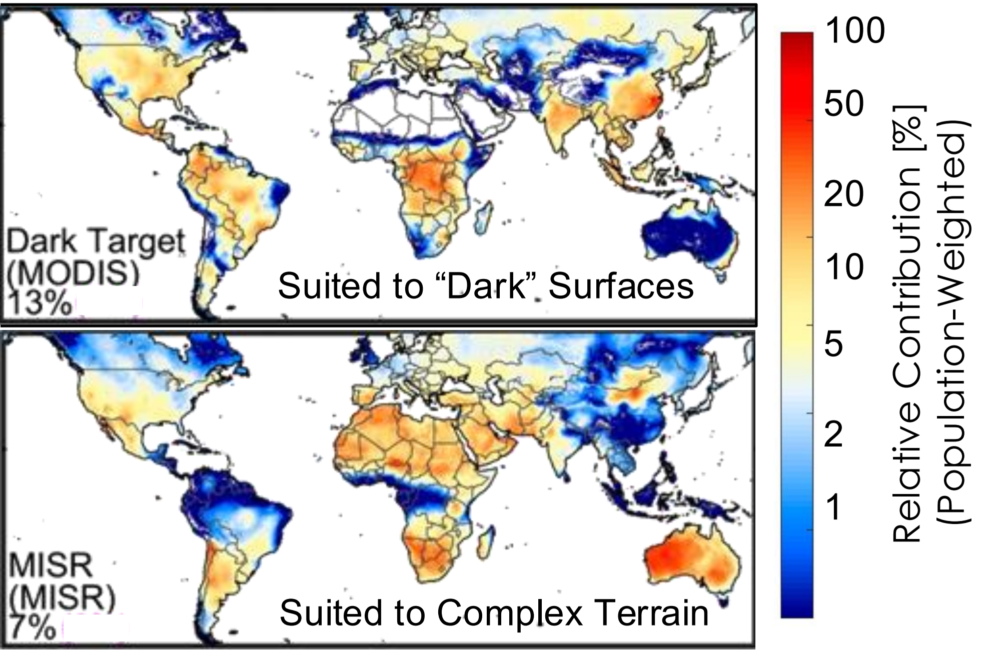
Sim. Error →

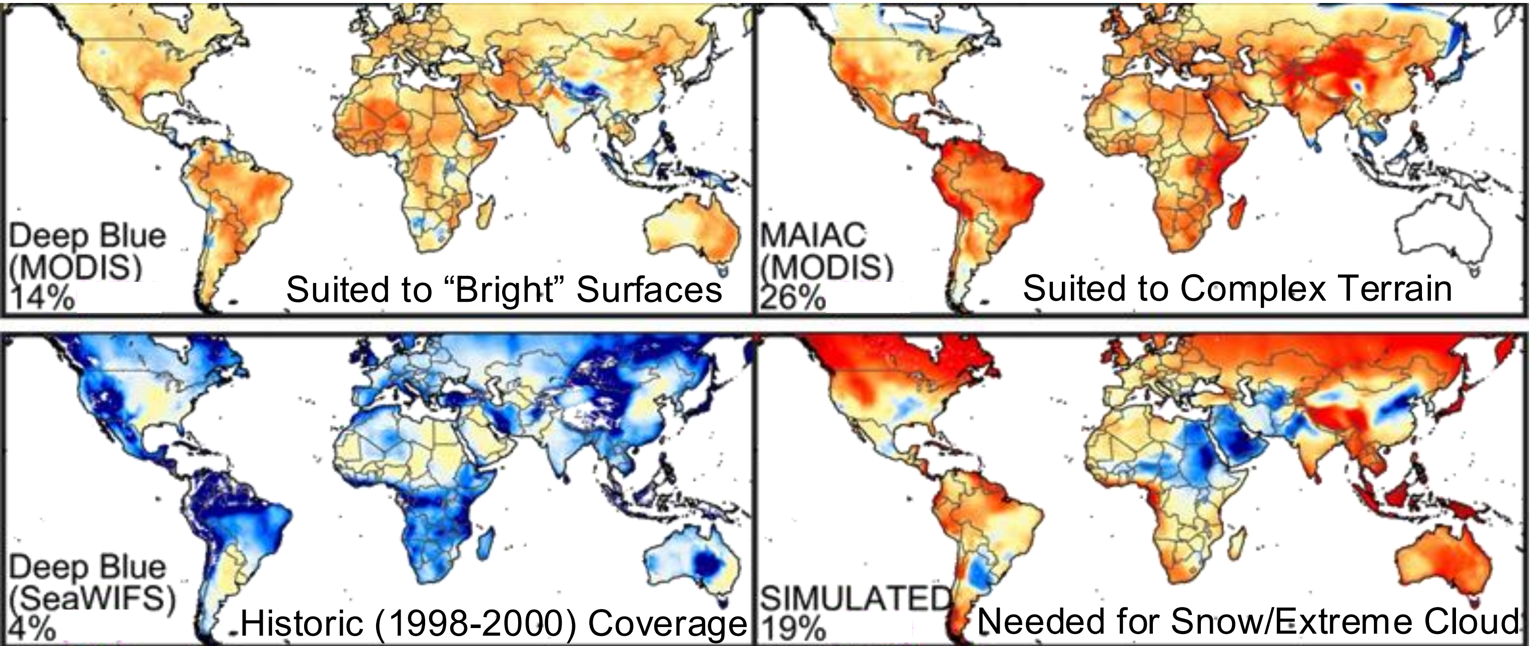
- Emissions
- Chemistry
- Optics
- Meteorology

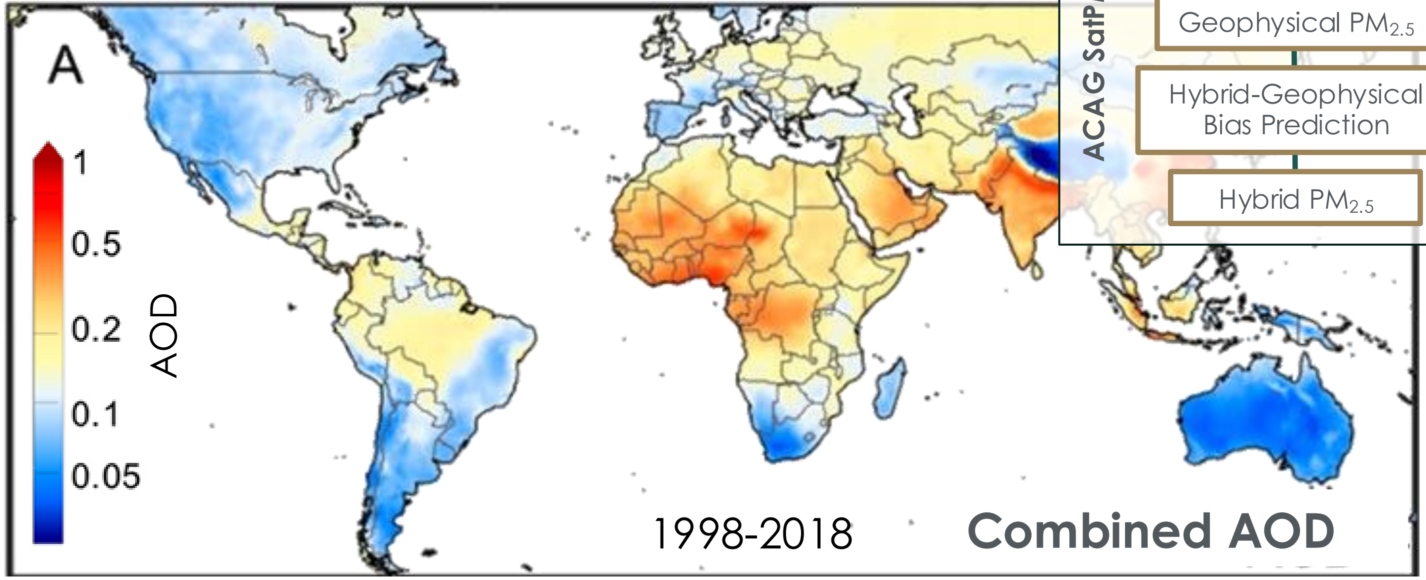
Weight AERONET → **comparison by**

- Species
- Proximity
- Elevation
- Similarity

Global Evaluation: Use Each AOD Dataset at its Best







Hammer et al., ES&T, 2020

➢ **Combine all AOD sources weighted by $NRMSD^{-1}$, bias correction^{-1}, data density:**

$$AOD = \frac{\sum_{n=1}^{N} \frac{1}{NRMSD_n} \times \left[\frac{\Delta AOD_{adj,n}}{AOD_n}\right]^{-1} \times N_{obs,n}^2 \times AOD_n}{\sum_{n=1}^{N} \frac{1}{NRMSD_n} \times \left[\frac{\Delta AOD_{adj,n}}{AOD_n}\right]^{-1} \times N_{obs,n}^2}$$

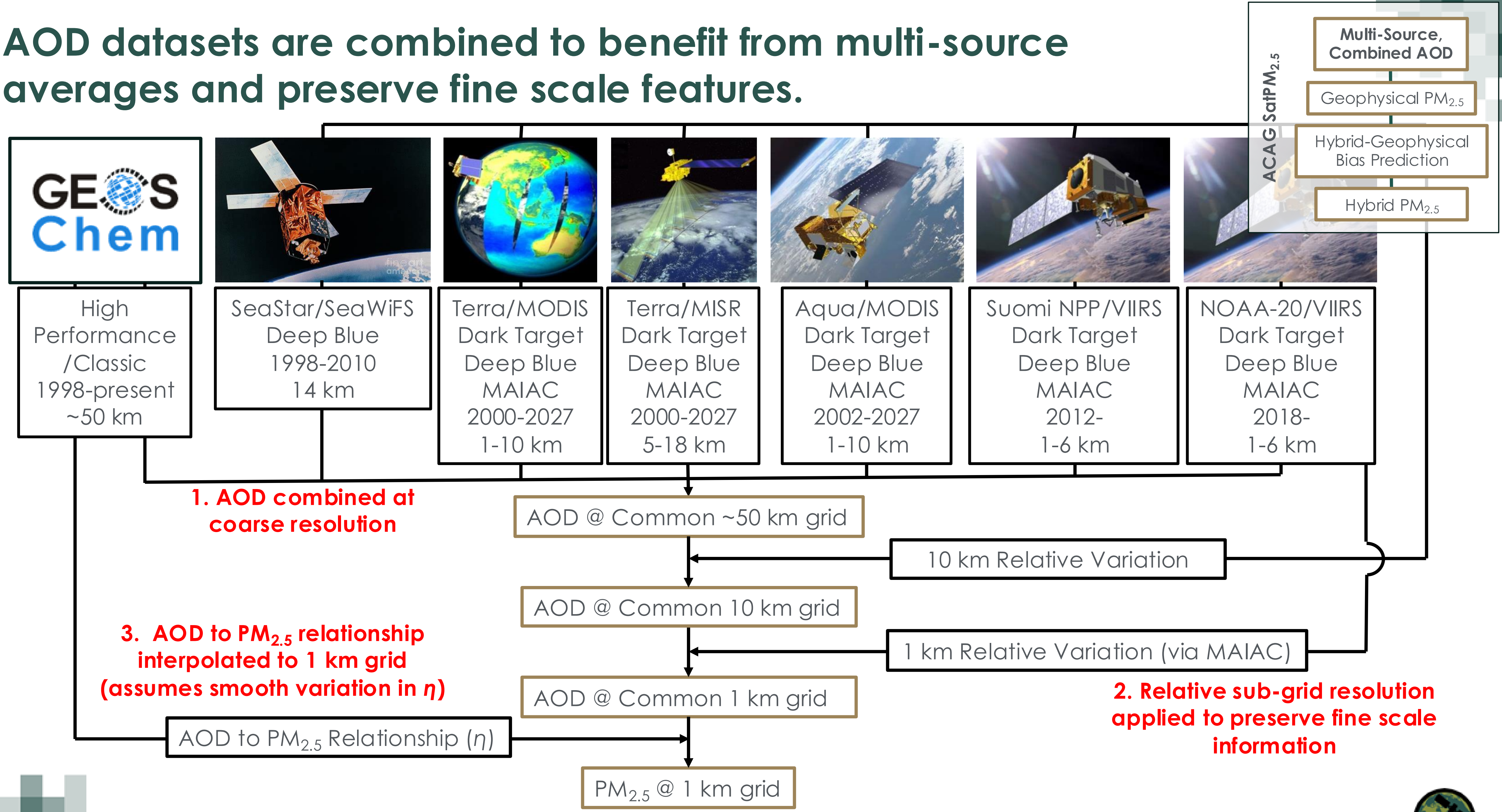
Sim. Error

- Emissions
- Chemistry
- Optics
- Meteorology

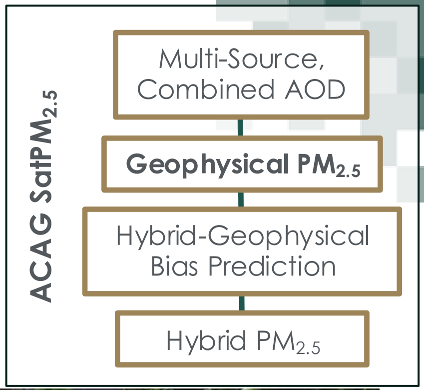
Weight AERONET comparison by

- Species
- Proximity
- Elevation Similarity

AOD datasets are combined to benefit from multi-source averages and preserve fine scale features.

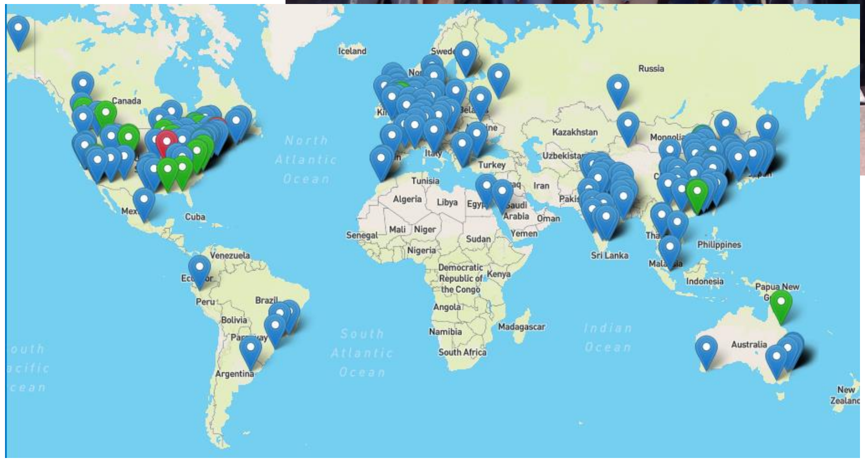


ACAG at WashU is a proud part of the GEOS-Chem community.



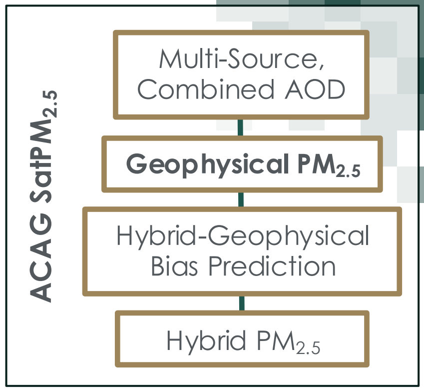
GEOS-Chem Community Mission: To advance understanding of human and natural influences on the environment through a comprehensive, state-of-the-science, readily-accessible global model of atmospheric composition.

GEOS Chem



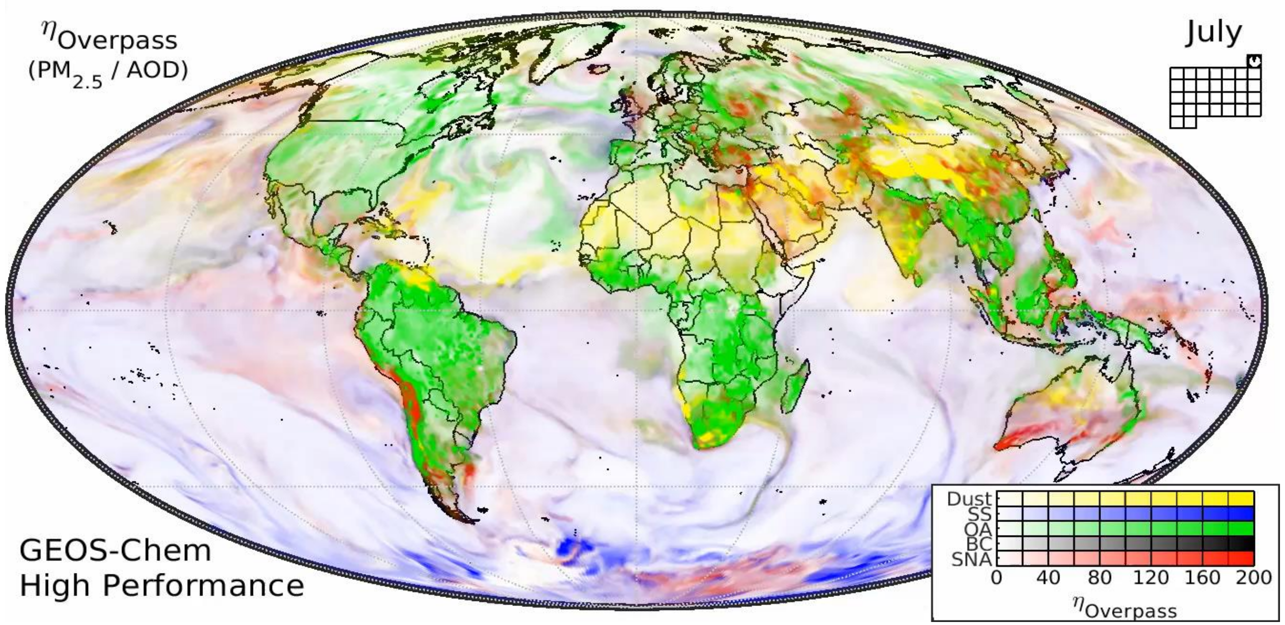
The 11th International GEOS-Chem meeting at WashU in June 2024 had 278 registrants from 96 institutions and 22 countries.

Geophysical Relationship (η) is Produced using GEOS-Chem CTM

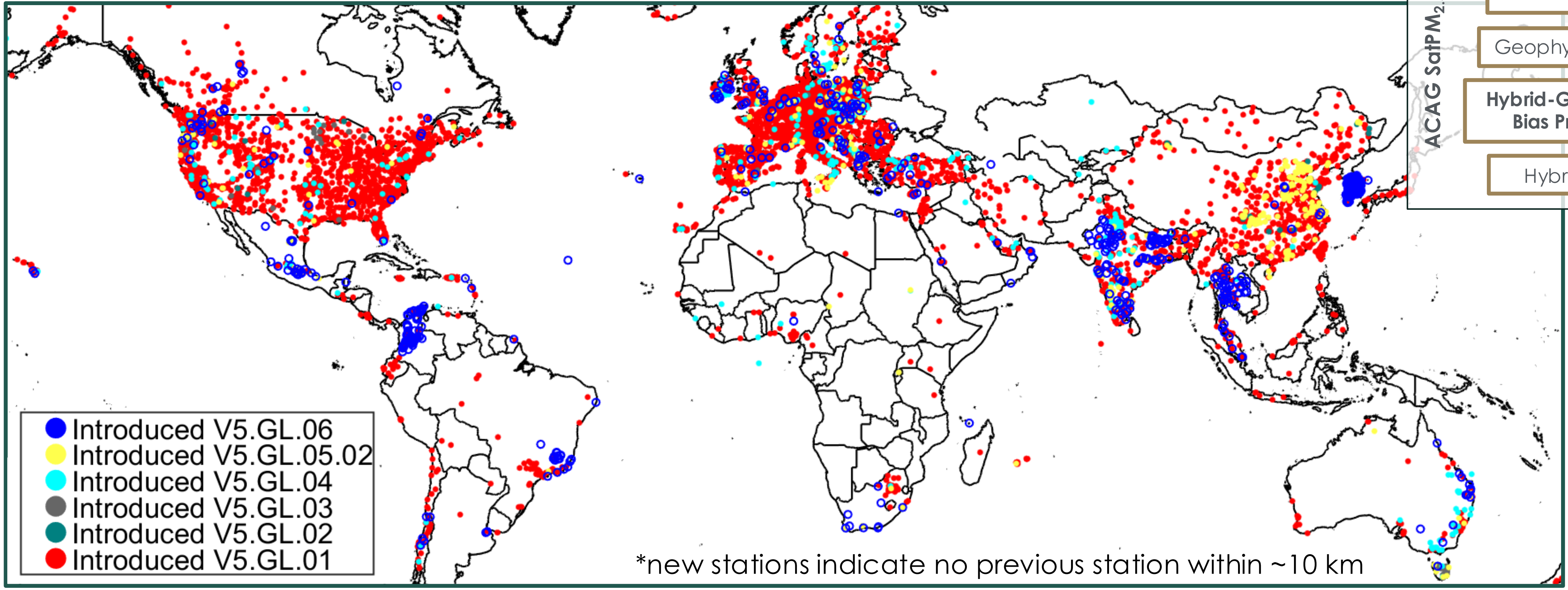


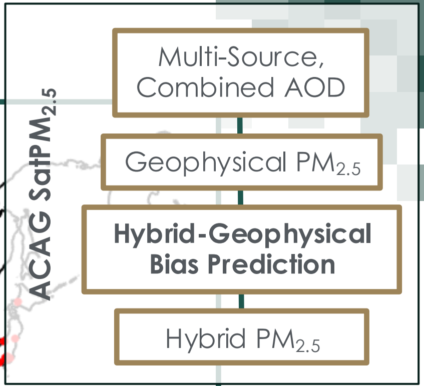
Relevant Features of GEOS-Chem CTM to $SatPM_{2.5}$:

- Open-source, community-driven
- Detailed aerosol-oxidant model
  - Sulfate-Nitrate-Ammonium System
  - Carbonaceous (primary and secondary)
  - Mineral Dust (natural and anthropogenic)
  - Seasalt
- Emissions Include:
  - Meteorologically-driven dust aerosol, lightning NOx, biogenic VOCs, soil NOx, and sea salt aerosol
  - High-resolution anthropogenic emissions via regional and the CEDS v2 inventories
  - Open Fires – Temporally varying GFED4.1s, GFAS, etc.
- Assimilated meteorology (MERRA2, GEOS-IT, etc.)
- Typical resolutions of ~50×50 km^2



Global Hybrid $PM_{2.5}$ requires global ground-based observations.



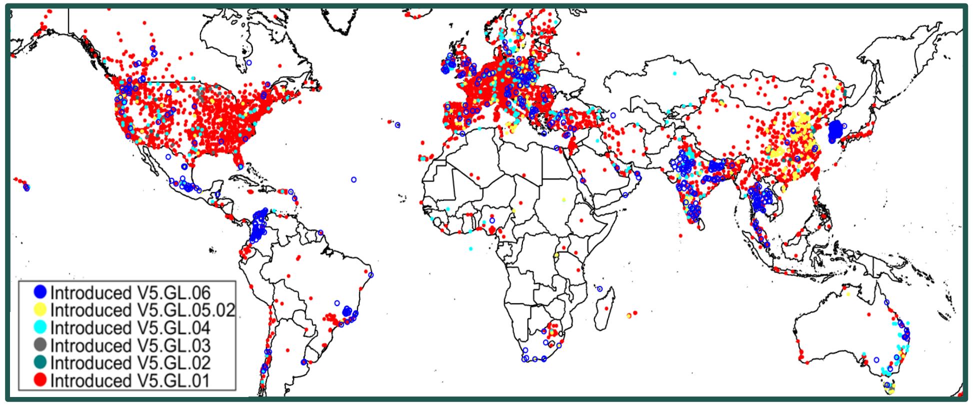


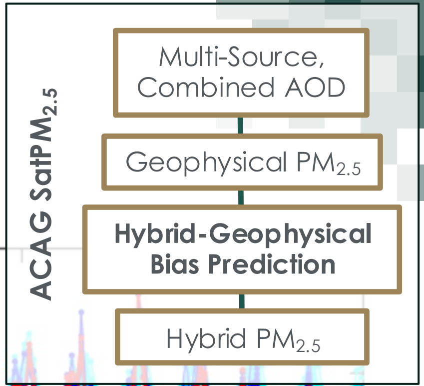
We currently include ground-based $PM_{2.5}$ from approximately 15,000 sites, using observations collected from:

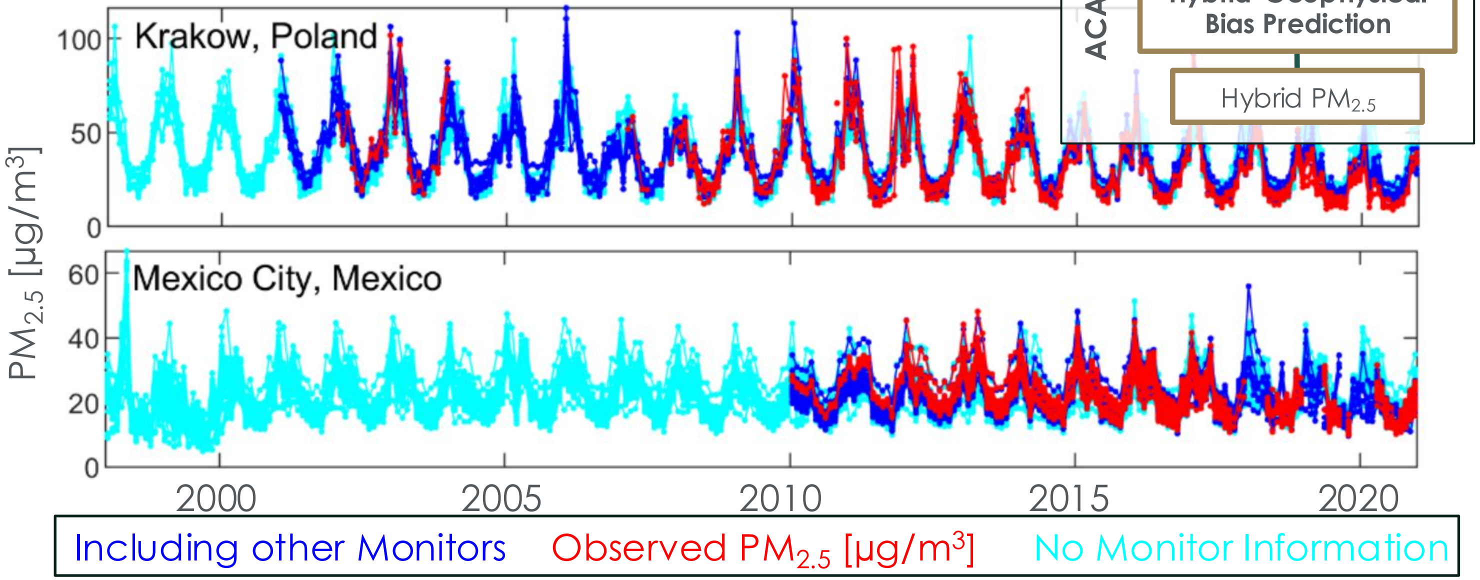
- USA (EPA/AQS)
- Canada (NAPS)
- Europe (Airbase)
- Australia (NT,Q,NSW)
- Taiwan (Taiwanese EPA/OpenAQ)
- Global (GBD/WHO, OpenAQ)
- China (Chinese EPA)
- India (CPCB)
- South Korea (MoE)
- Oman (EA)
- Hashemite Kingdom of Jordan (MoE)
- Argentina (ACUMAR)
- Brazil (IEMA)
- Chile (SINCA)
- Colombia (SISAIRE)
- Ecuador (REMMAQ)
- Uruguay (Montevideo)

Ground-based observations are not temporally complete.

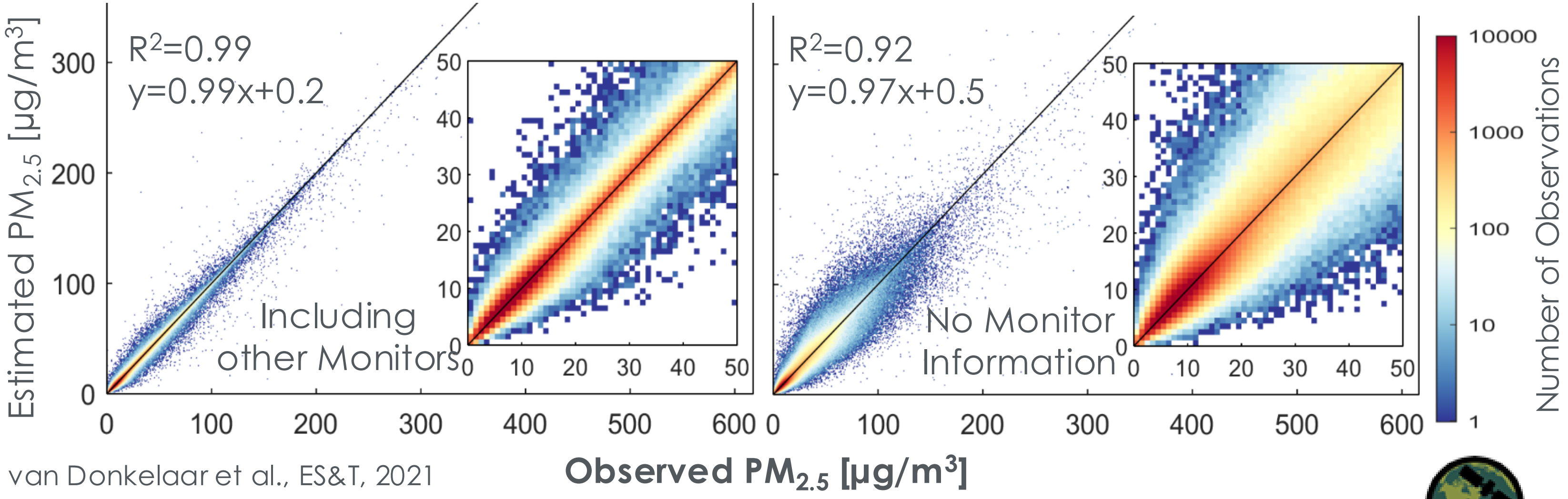
Need to represent missing time periods for consistent time series.







- Infer ground-based observations using:
  - Ground-based monitors
  - Simulations
  - Geophysical PM$_{2.5}$.
- Combine inferred values using observed uncertainty
- Temporally complete ground-based time-series limits impact of coverage changes on SatPM$_{2.5}$ time-series



van Donkelaar et al., ES&T, 2021

ACAG Hybrid $PM_{2.5}$ uses a Convolutional Neural Network (CNN).

- A CNN is a state-of-the-science machine learning algorithm.
  - Specially designed to identify and learn from images.
  - Highly effective for spatial tasks, such as relationship of geophysical $PM_{2.5}$ bias with surrounding predictors.
- Trained separately each month
  - Capture temporal variability in relationships between predictors and bias
- Geographically Weighted Regression (GWR) also available, as version V5.

CNN Predictor Variables

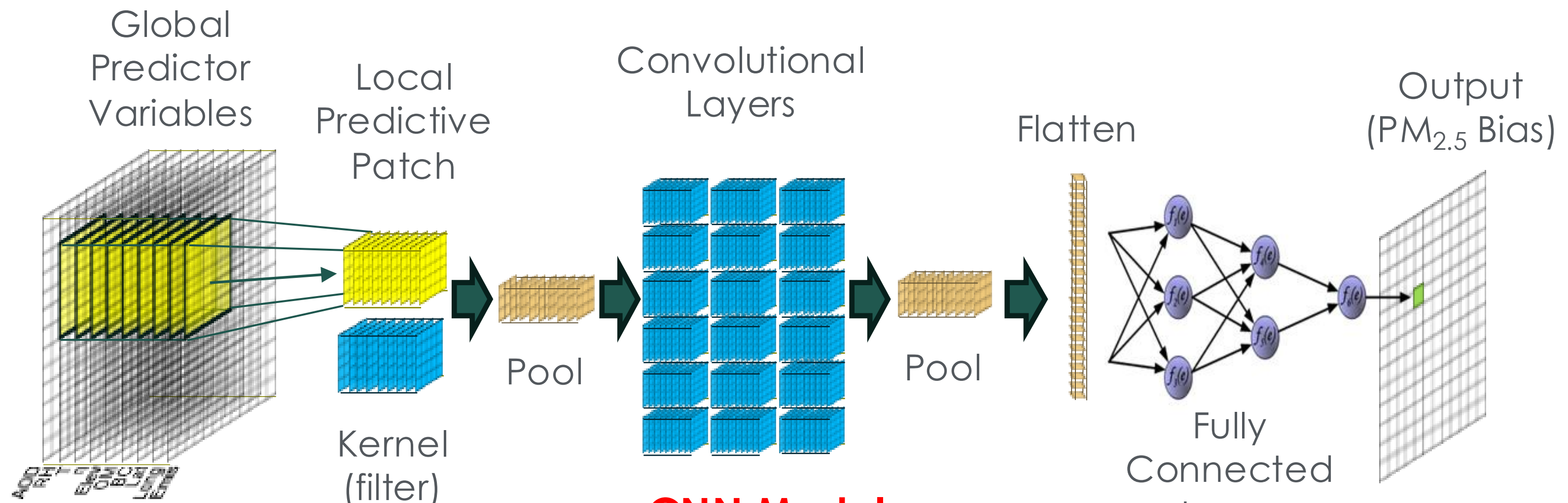
- **Satellite Products**
  - AOD
  - Geophysical $PM_{2.5}$
- **CTM output**
  - $PM_{2.5}$
  - components
  - η
  - η uncertainty
- **Emissions**
  - Anthropogenic
    - Organic Carbon
    - Black Carbon
  - Dust
  - Sea Salt
- **Meteorology**
  - PBLH
  - Relative Humidity
  - Temperature (2m)
  - V wind speed (10m)
  - U wind speed (10m)
- **Site Locations**
- **Geology**
  - Land type
  - Elevation

ACAG SatPM$_{2.5}$

Multi-Source, Combined AOD
→ Geophysical $PM_{2.5}$
→ **Hybrid-Geophysical Bias Prediction**
→ Hybrid $PM_{2.5}$

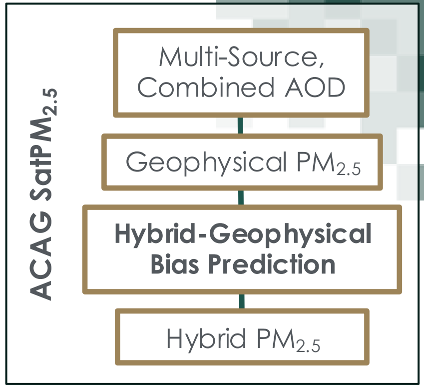
Hybrid-Geophysical Bias Prediction

Geographically-Weighted Regression	**Convolutional Neural Network**
V5.GL.06 Hybrid $PM_{2.5}$ (GWR)	**V6.GL.03 Hybrid $PM_{2.5}$ (CNN)**



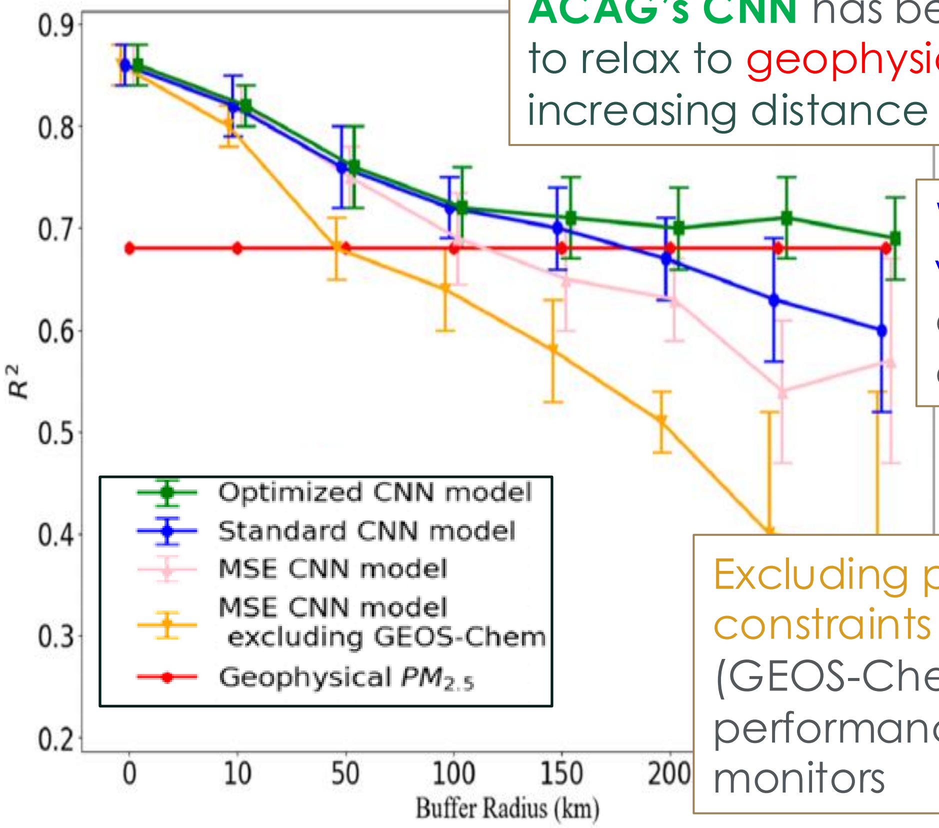
CNN Model

Geophysical and process-based predictors drive agreement away from ground-based constraints.



Demonstrated by both spatially-buffered cross validation and SHapley Additive exPlanations (SHAP) Analysis

Processed-based predictors are those that include the atmospheric response to physical and chemical drivers, such as simulated component concentrations and η.

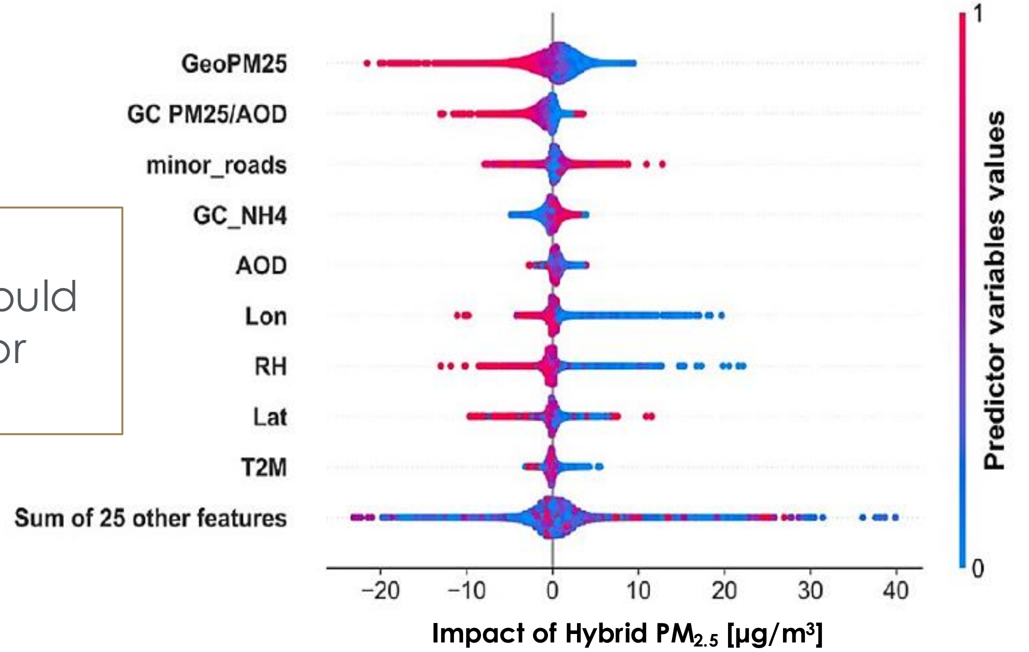


ACAG's CNN has been optimized to relax to geophysical PM$_{2.5}$ with increasing distance from monitor

Without geophysical weighting, quality would degrade with monitor distance

Excluding process-based constraints from a CTM (GEOS-Chem) reduces performance away from monitors

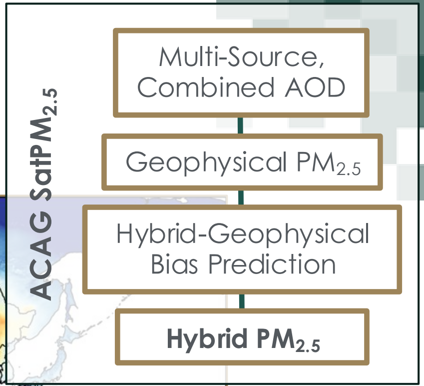
SHAP Analysis: Variables are ordered by importance



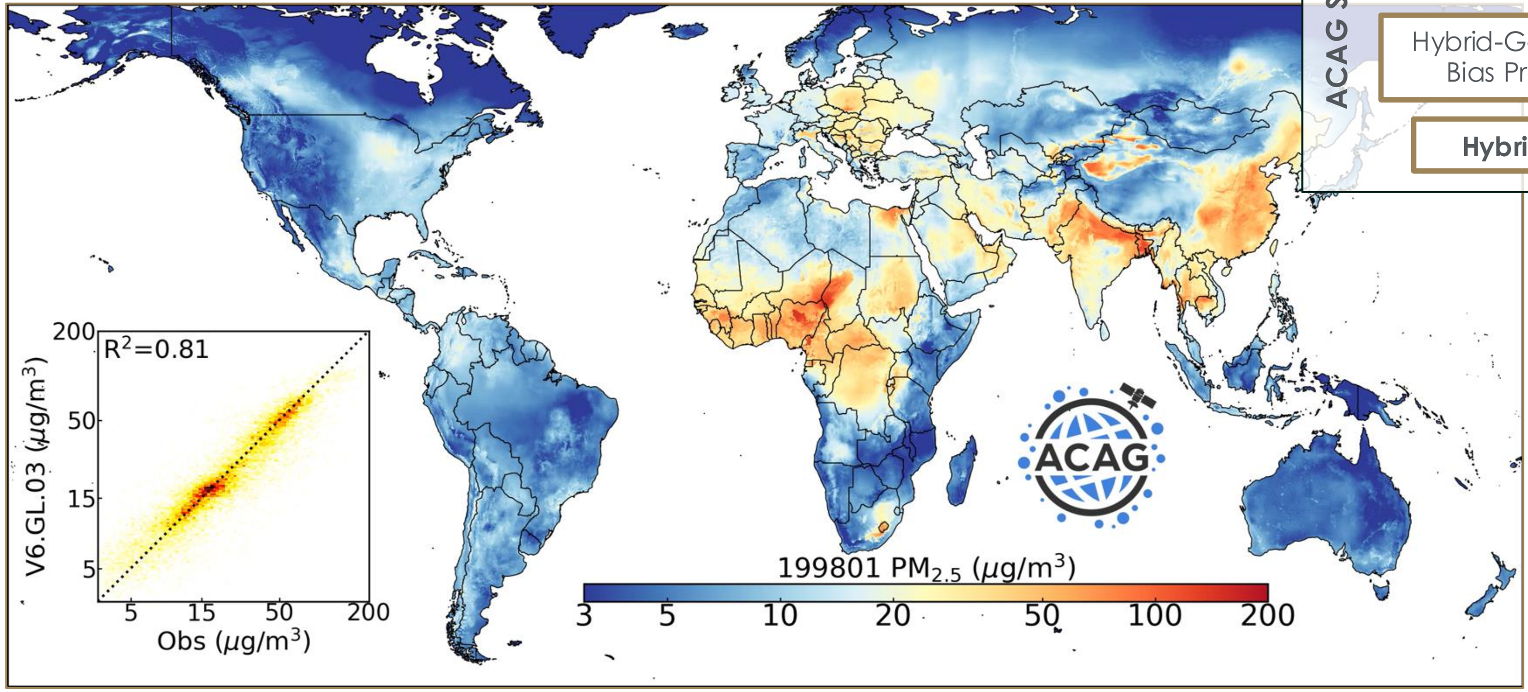
- Geo indicates geophysical a priori constrained by GEOS-Chem output
- GC indicates GEOS-Chem

Shen et al., ES&T, 2024

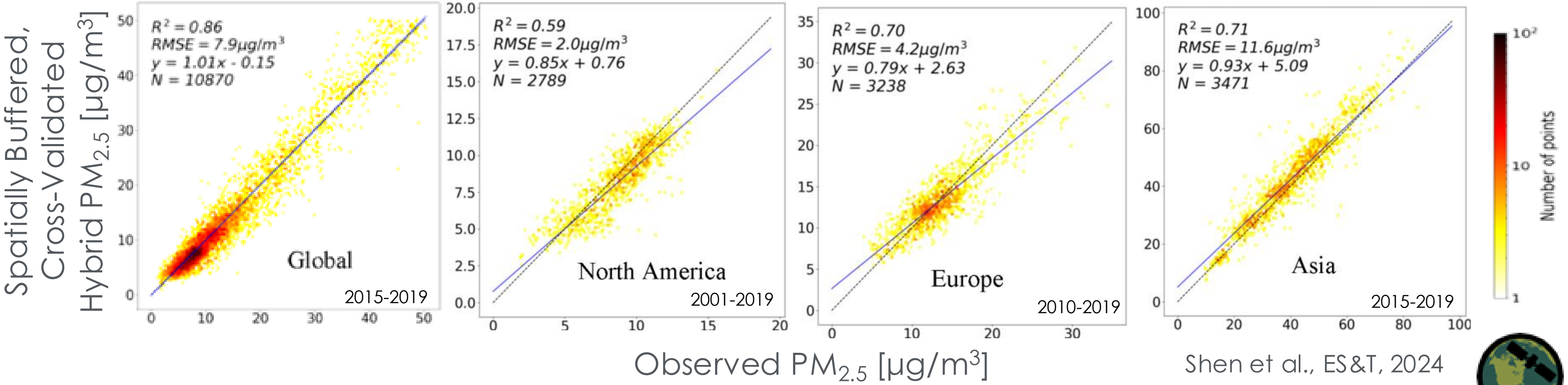
SatPM$_{2.5}$ provides high quality, global PM$_{2.5}$ representation.



- SatPM$_{2.5}$ V6.GL.03 refines the geophysical PM$_{2.5}$ surface using ACAG's optimized CNN model.
- Evaluated using the highest cross-validation standards – temporally and spatially buffered
- SatPM$_{2.5}$ are particularly well-suited for applications that require:
  - High spatial resolution
  - Consistent, long-term temporal coverage
  - Consistency away from ground-based constraints



Atmospheric Composition Analysis Group



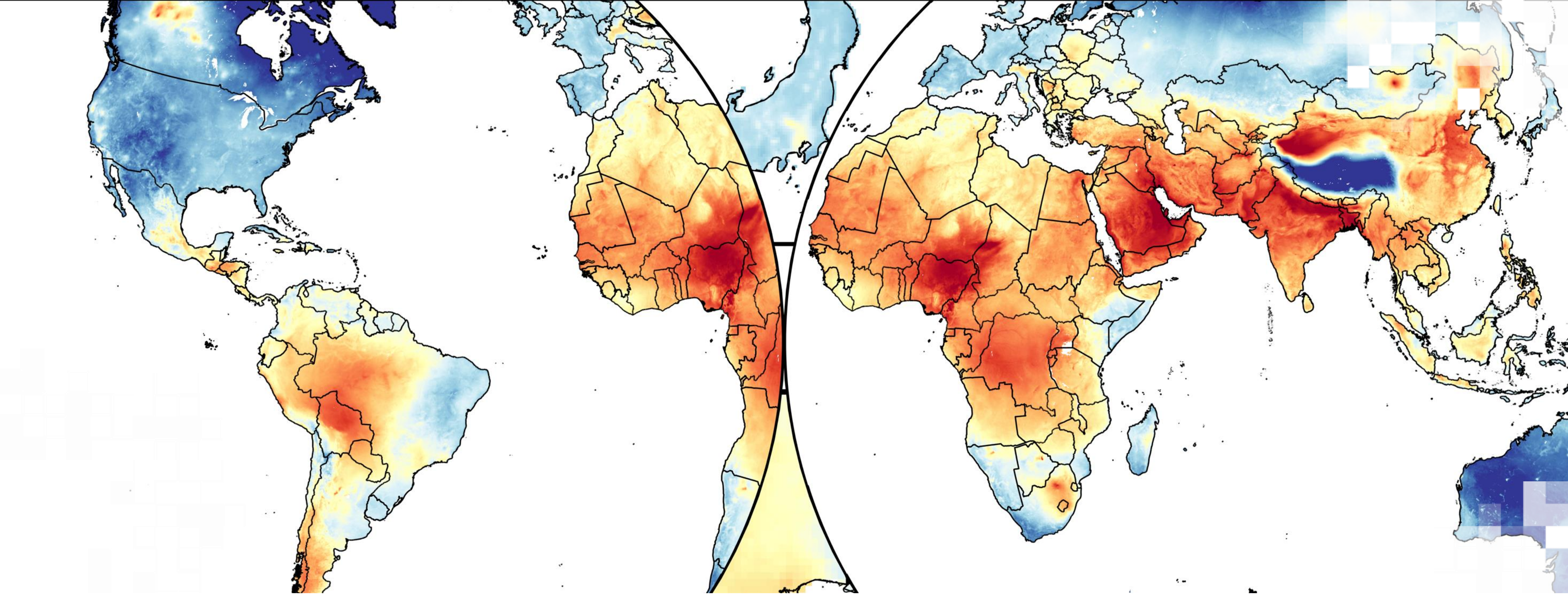
Shen et al., ES&T, 2024

Summary – SatPM$_{2.5}$ @ ACAG Washington University in St. Louis

In this section, we've focused on the ACAG SatPM$_{2.5}$ dataset.

This dataset:

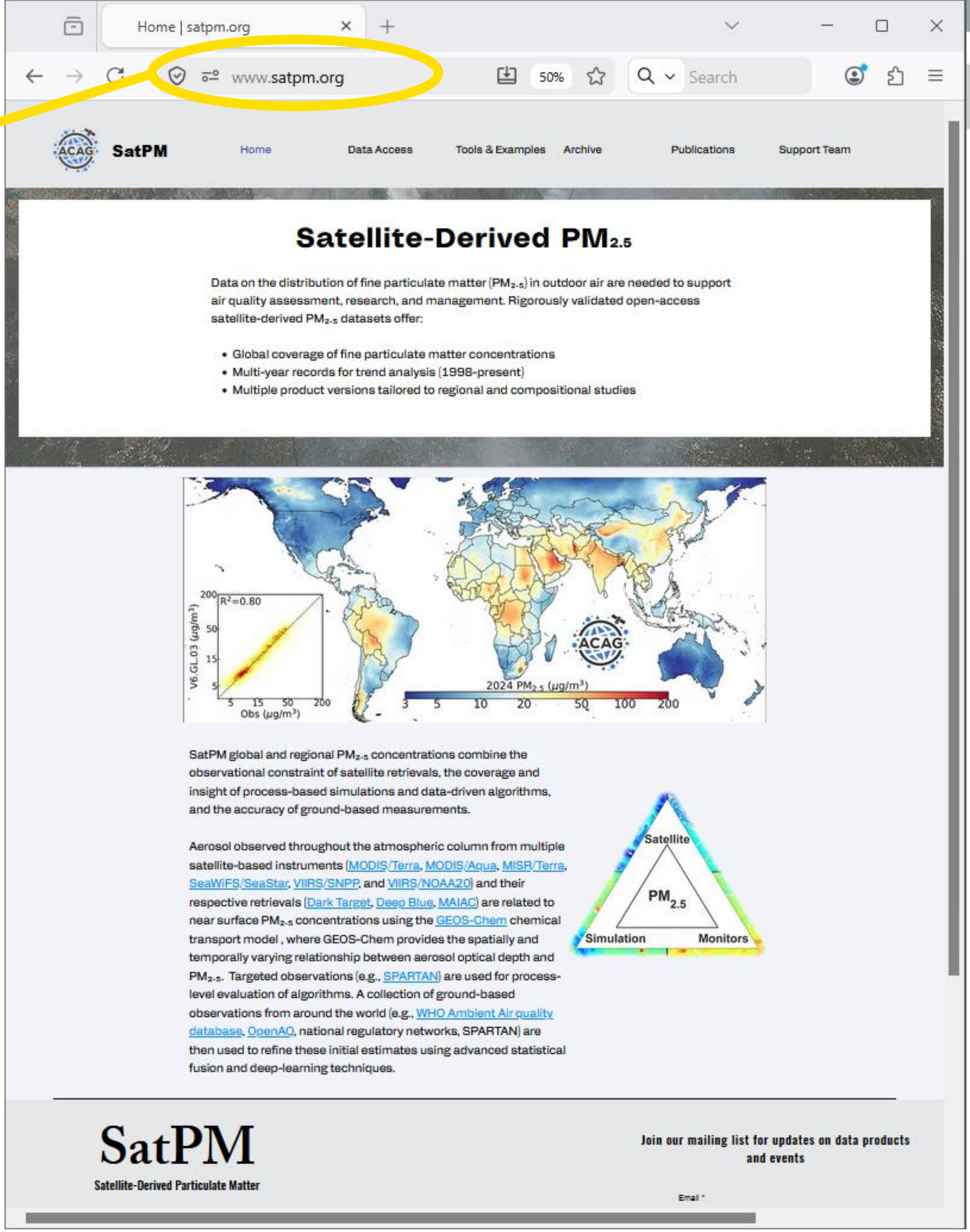
- Is based on the Geophysical-Hybrid method.
- Combines multiple AOD products, using AERONET to identify which retrievals are most effective under specific conditions, locations, and times.
- Incorporates and actively maintains a database of ~15,000 ground-based monitors.
  - Accounts for variations in temporal coverage of these ground-based observations.
- Uses a Convolutional Neural Network (CNN) to understand, predict, and correct geophysical bias. Process-based constraints essential to maintain quality away from monitors.
- Focuses on long-term consistency across the entire time series and away from ground-based constraints.
- Is freely and publicly available.



Accessing SatPM$_{2.5}$ Data Products – Website Demo

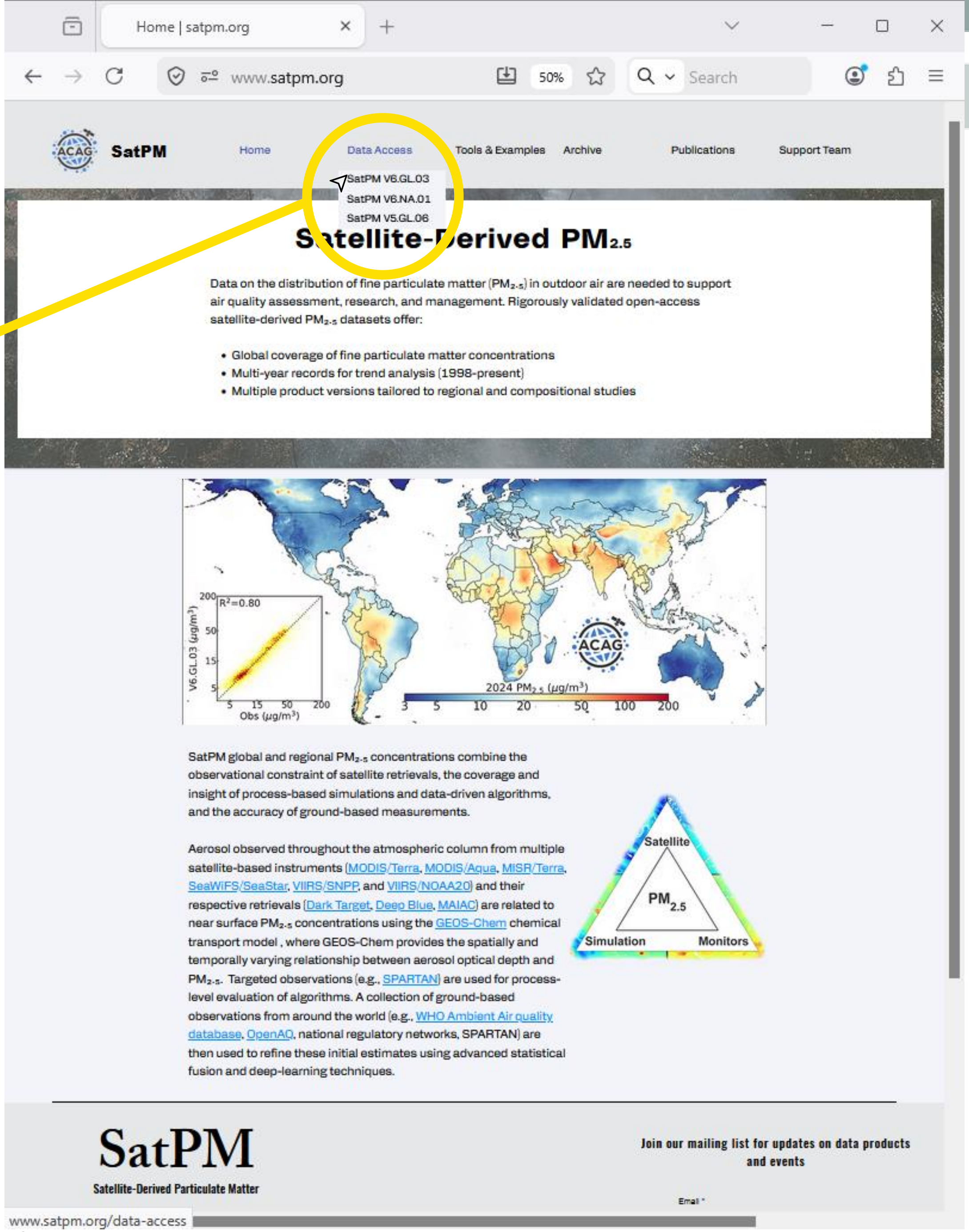
http://satpm.org

Dedicated Website (www.satpm.org) Provides Access to Latest SatPM$_{2.5}$ Versions



Dedicated Website (www.satpm.org) Provides Access to Latest SatPM$_{2.5}$ Versions

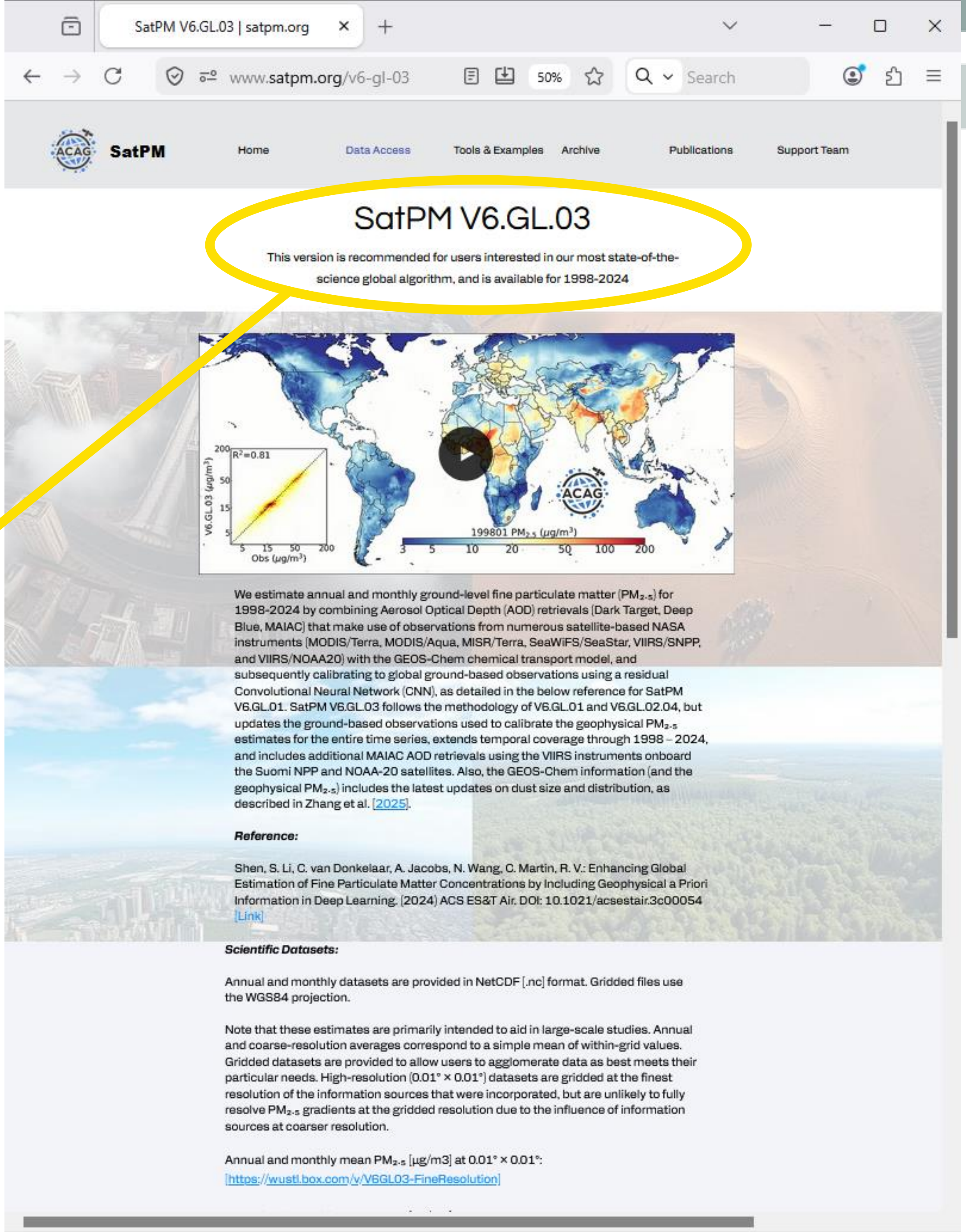
Data Access Tab Links to Latest SatPM$_{2.5}$ Datasets



Dedicated Website (www.satpm.org) Provides Access to Latest SatPM$_{2.5}$ Versions

Data Access Tab Links to Latest SatPM$_{2.5}$ Datasets

Individual Pages Provide Details of Each Version

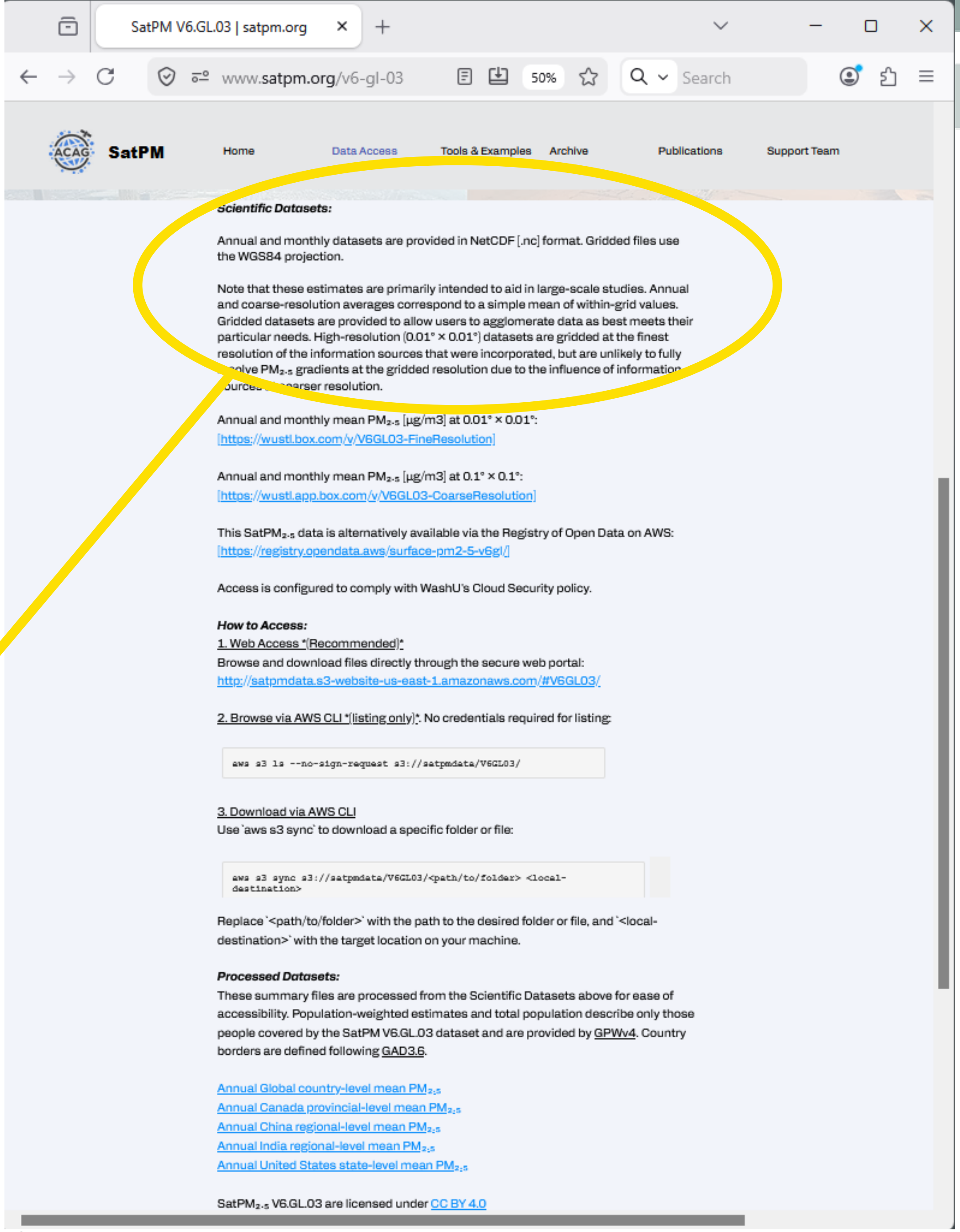


Dedicated Website (www.satpm.org) Provides Access to Latest SatPM$_{2.5}$ Versions

Data Access Tab Links to Latest SatPM$_{2.5}$ Datasets

Individual Pages Provide Details of Each Version

Includes Dataset/Access Instructions


SatPM V6.GL.03 | satpm.org
www.satpm.org/v6-gl-03
50%
Search
SatPM
Home
Data Access
Tools & Examples
Archive
Publications
Support Team

Scientific Datasets:

Annual and monthly datasets are provided in NetCDF [.nc] format. Gridded files use the WGS84 projection.

Note that these estimates are primarily intended to aid in large-scale studies. Annual and coarse-resolution averages correspond to a simple mean of within-grid values. Gridded datasets are provided to allow users to agglomerate data as best meets their particular needs. High-resolution (0.01° × 0.01°) datasets are gridded at the finest resolution of the information sources that were incorporated, but are unlikely to fully resolve PM$_{2.5}$ gradients at the gridded resolution due to the influence of information sources at coarser resolution.

Annual and monthly mean PM$_{2.5}$ [μg/m3] at 0.01° × 0.01°:
[https://wustl.box.com/v/V6GL03-FineResolution]

Annual and monthly mean PM$_{2.5}$ [μg/m3] at 0.1° × 0.1°:
[https://wustl.app.box.com/v/V6GL03-CoarseResolution]

This SatPM$_{2.5}$ data is alternatively available via the Registry of Open Data on AWS:
[https://registry.opendata.aws/surface-pm2-5-v6gl/]

Access is configured to comply with WashU's Cloud Security policy.

How to Access:
1. Web Access *(Recommended)*
Browse and download files directly through the secure web portal:
http://satpmdata.s3-website-us-east-1.amazonaws.com/#V6GL03/

2. Browse via AWS CLI *(listing only)*: No credentials required for listing:

```
aws s3 ls --no-sign-request s3://satpmdata/V6GL03/
```

3. Download via AWS CLI
Use `aws s3 sync` to download a specific folder or file:

```
aws s3 sync s3://satpmdata/V6GL03/<path/to/folder> <local-destination>
```

Replace `<path/to/folder>` with the path to the desired folder or file, and `<local-destination>` with the target location on your machine.

Processed Datasets:
These summary files are processed from the Scientific Datasets above for ease of accessibility. Population-weighted estimates and total population describe only those people covered by the SatPM V6.GL.03 dataset and are provided by GPWv4. Country borders are defined following GADM3.6.

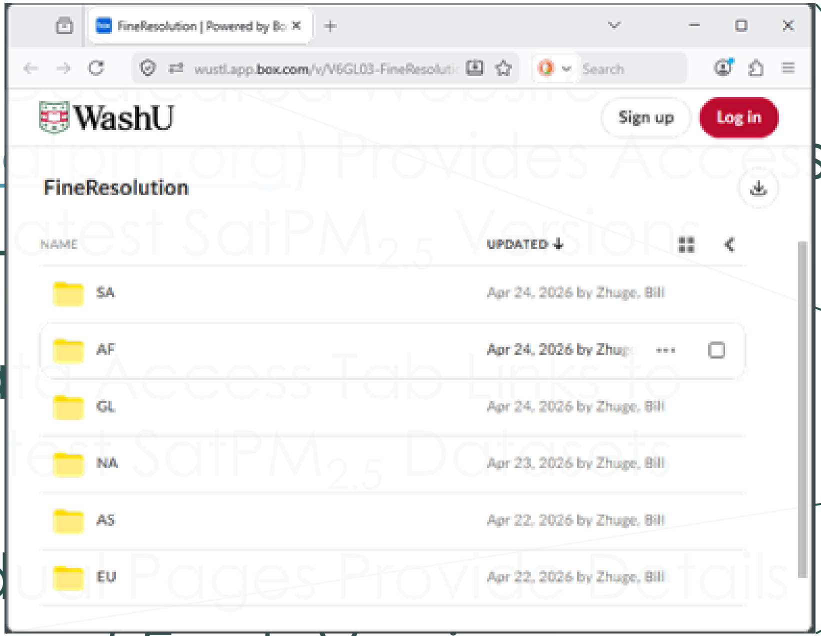
Annual Global country-level mean PM$_{2.5}$
Annual Canada provincial-level mean PM$_{2.5}$
Annual China regional-level mean PM$_{2.5}$
Annual India regional-level mean PM$_{2.5}$
Annual United States state-level mean PM$_{2.5}$

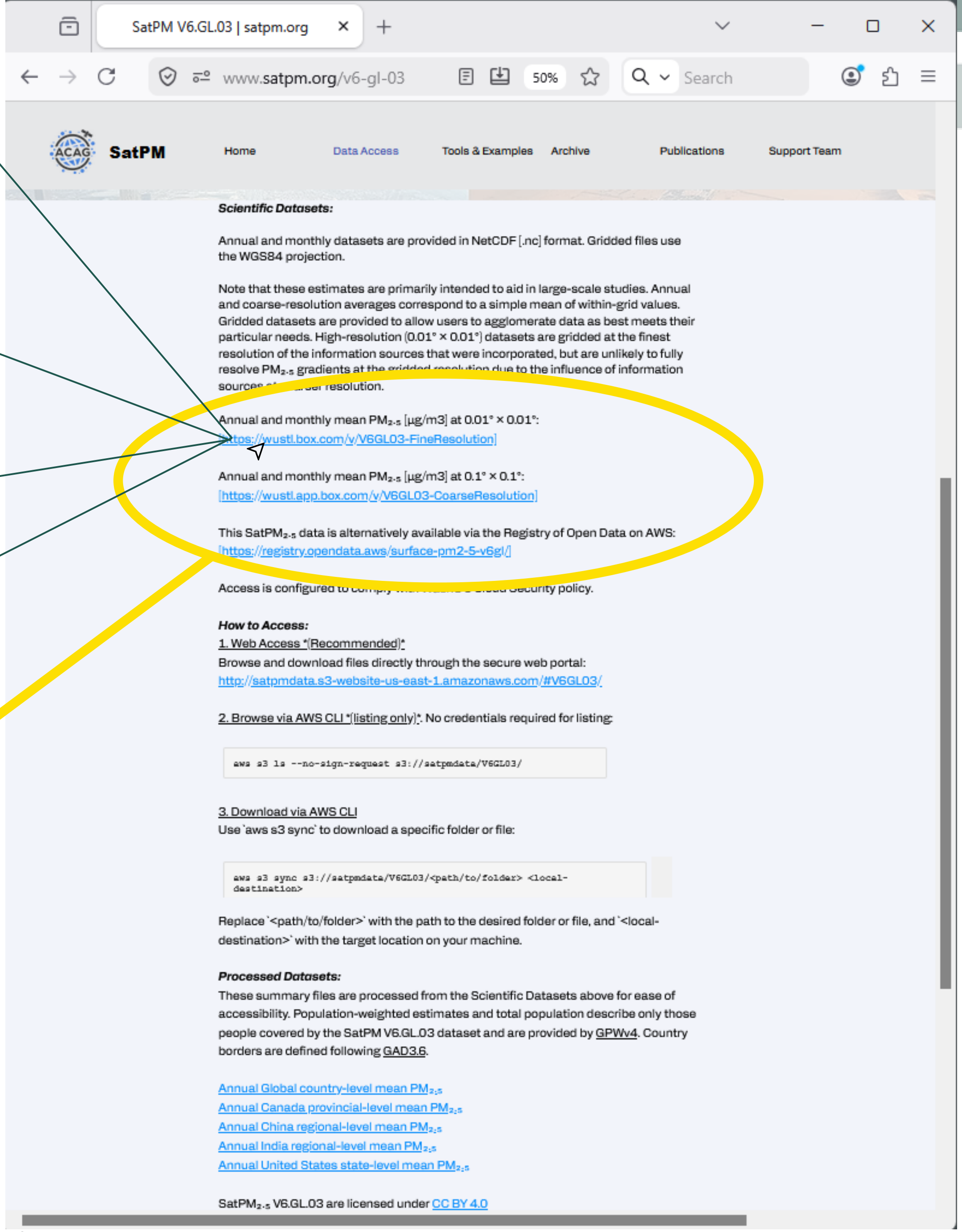
SatPM$_{2.5}$ V6.GL.03 are licensed under CC BY 4.0

(www.satpm.org) Provides Access to Latest SatPM$_{2.5}$ Versions

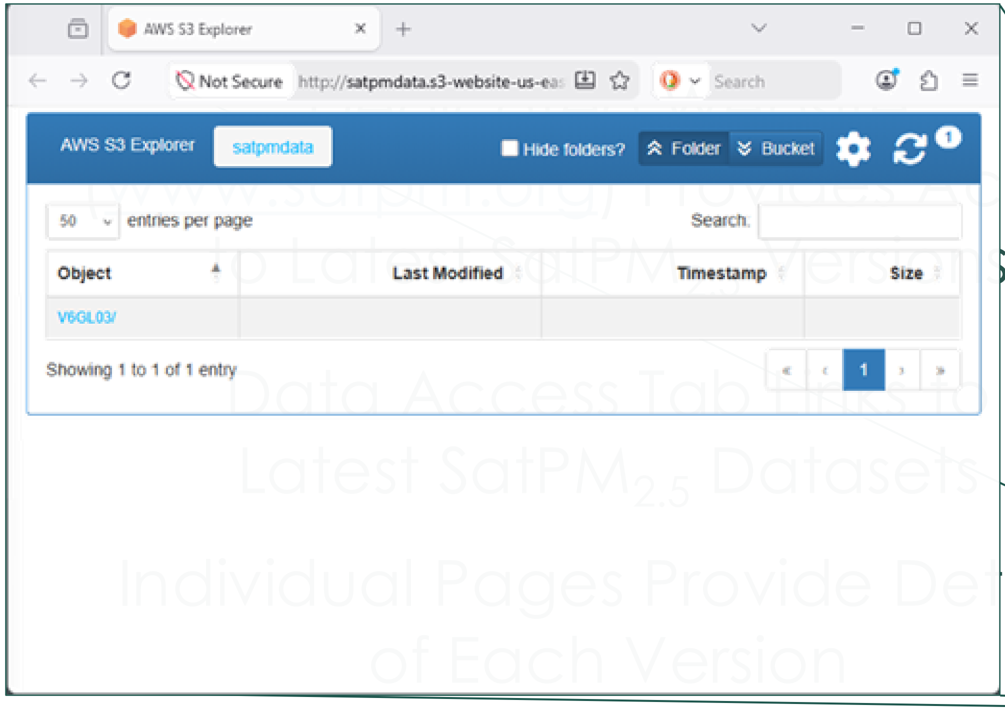
Data Access Tab Links to Latest SatPM$_{2.5}$ Datasets

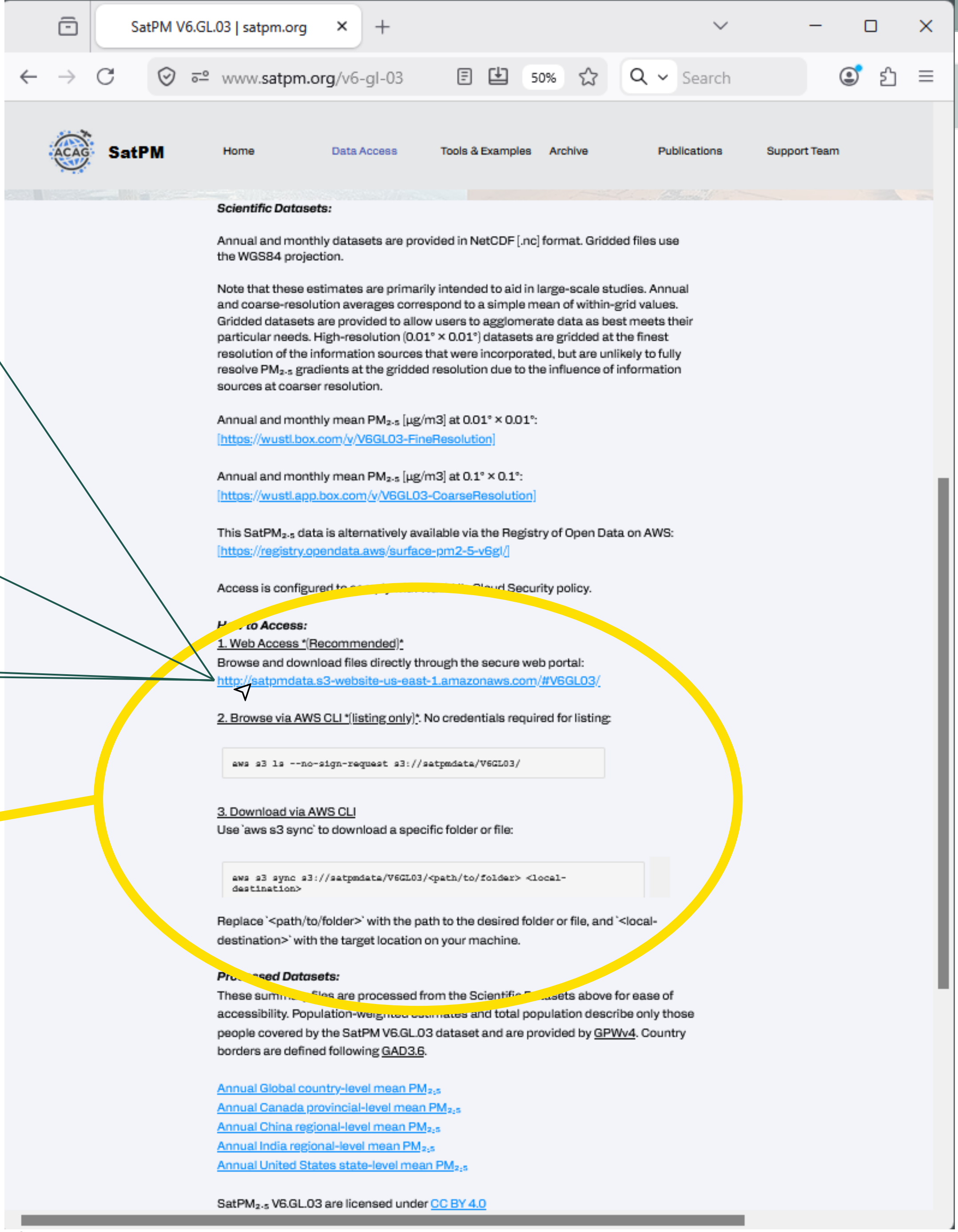
Individual Pages Provide Details of Each Version





Includes Dataset/Access Instructions **via Box Folders**





Includes Dataset/Access Instructions via Box Folders

and via AWS Registry of Open Data

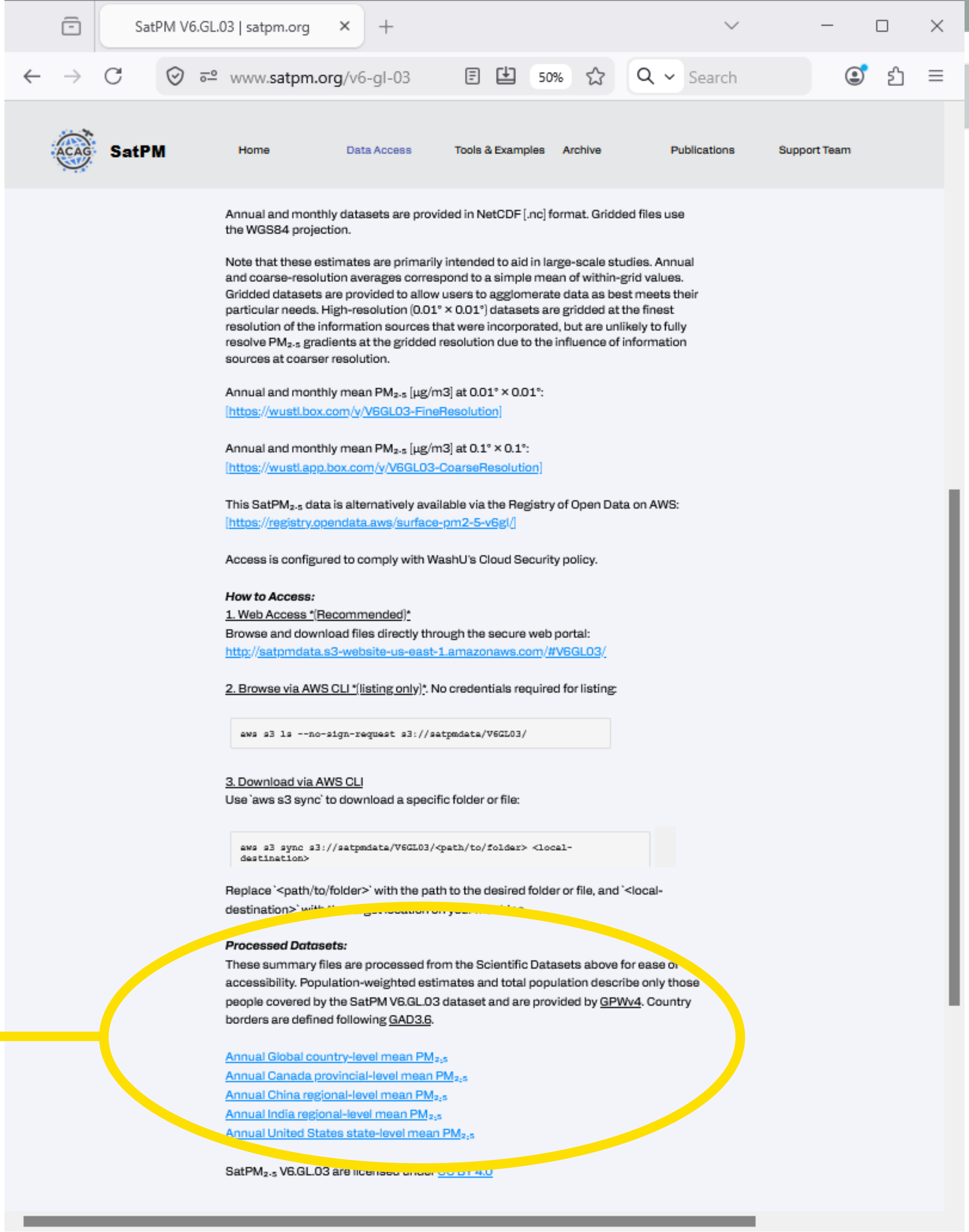
Dedicated Website (www.satpm.org) Provides Access to Latest SatPM$_{2.5}$ Versions

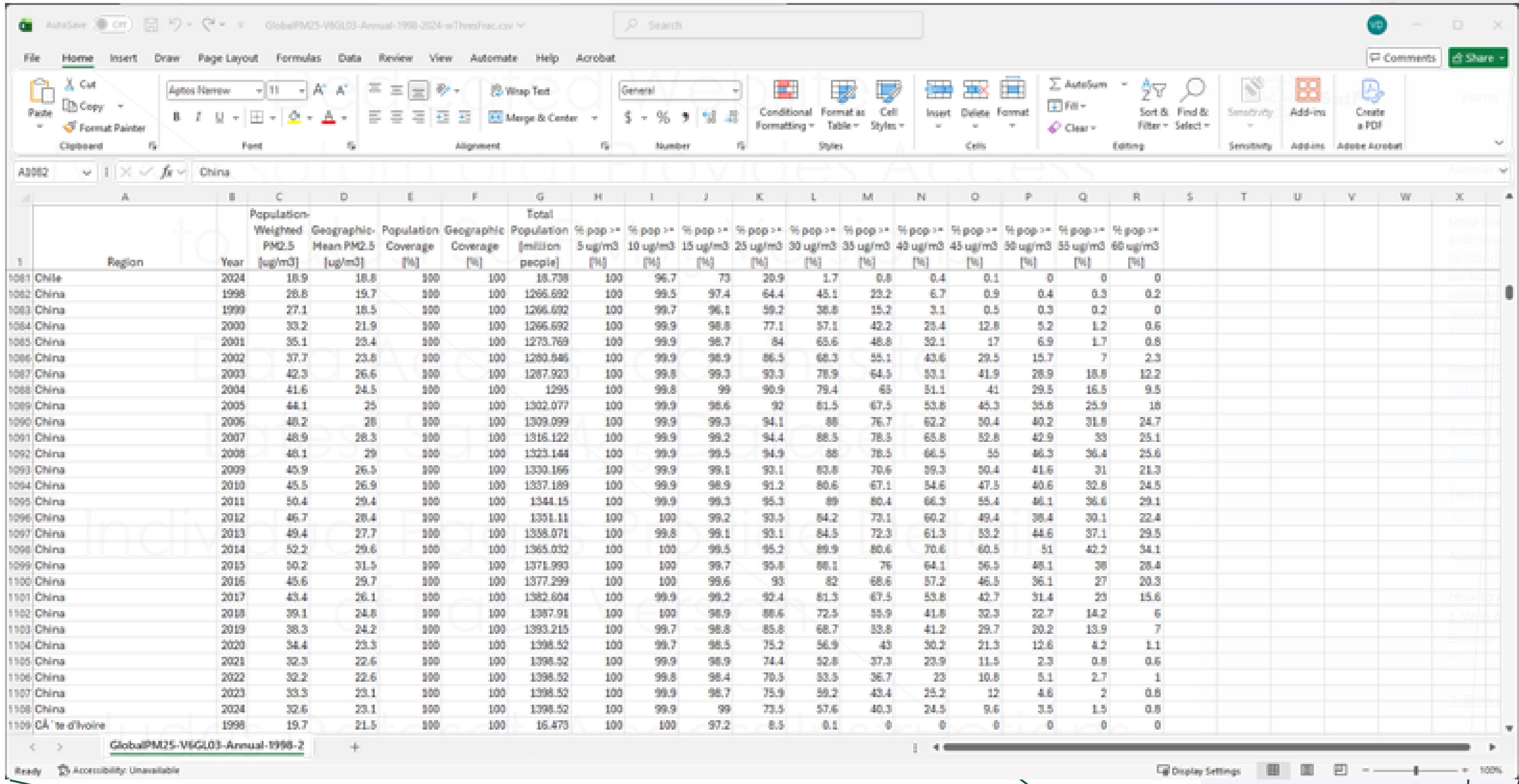
Data Access Tab Links to Latest SatPM$_{2.5}$ Datasets

Individual Pages Provide Details of Each Version

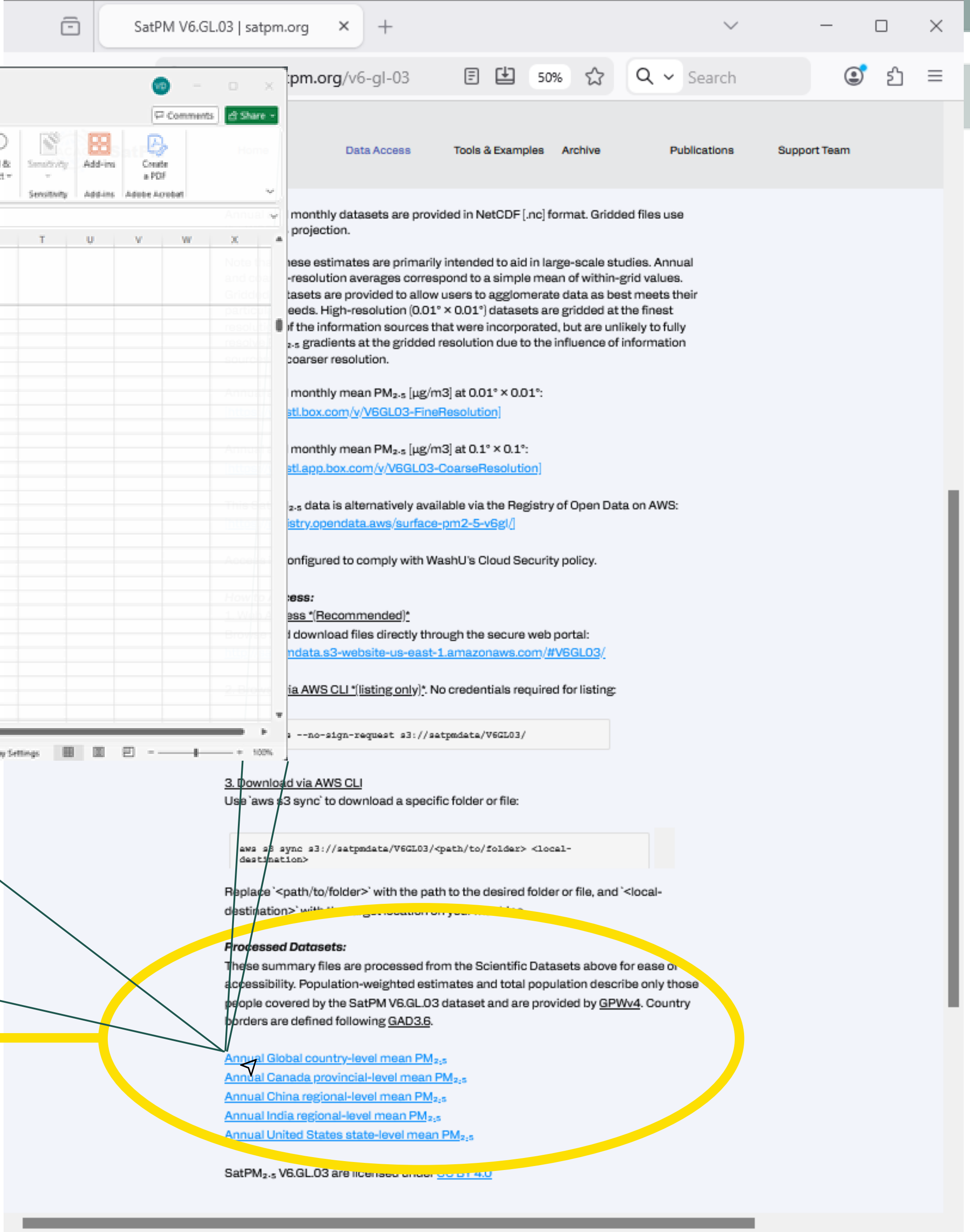
Includes Dataset/Access Instructions via Box Folders and via AWS Registry of Open Data

Processed Datasets Provide Simplified Summaries



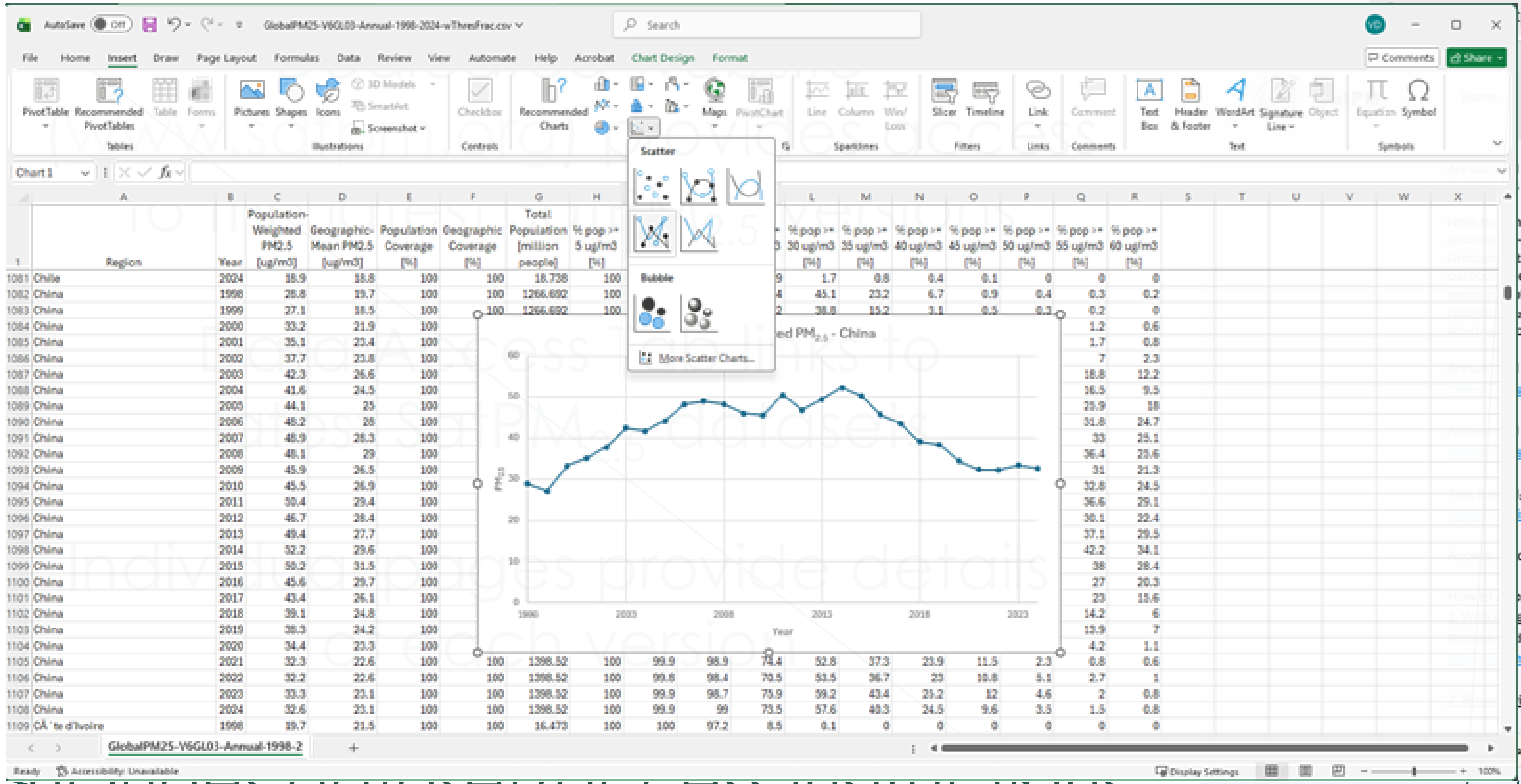


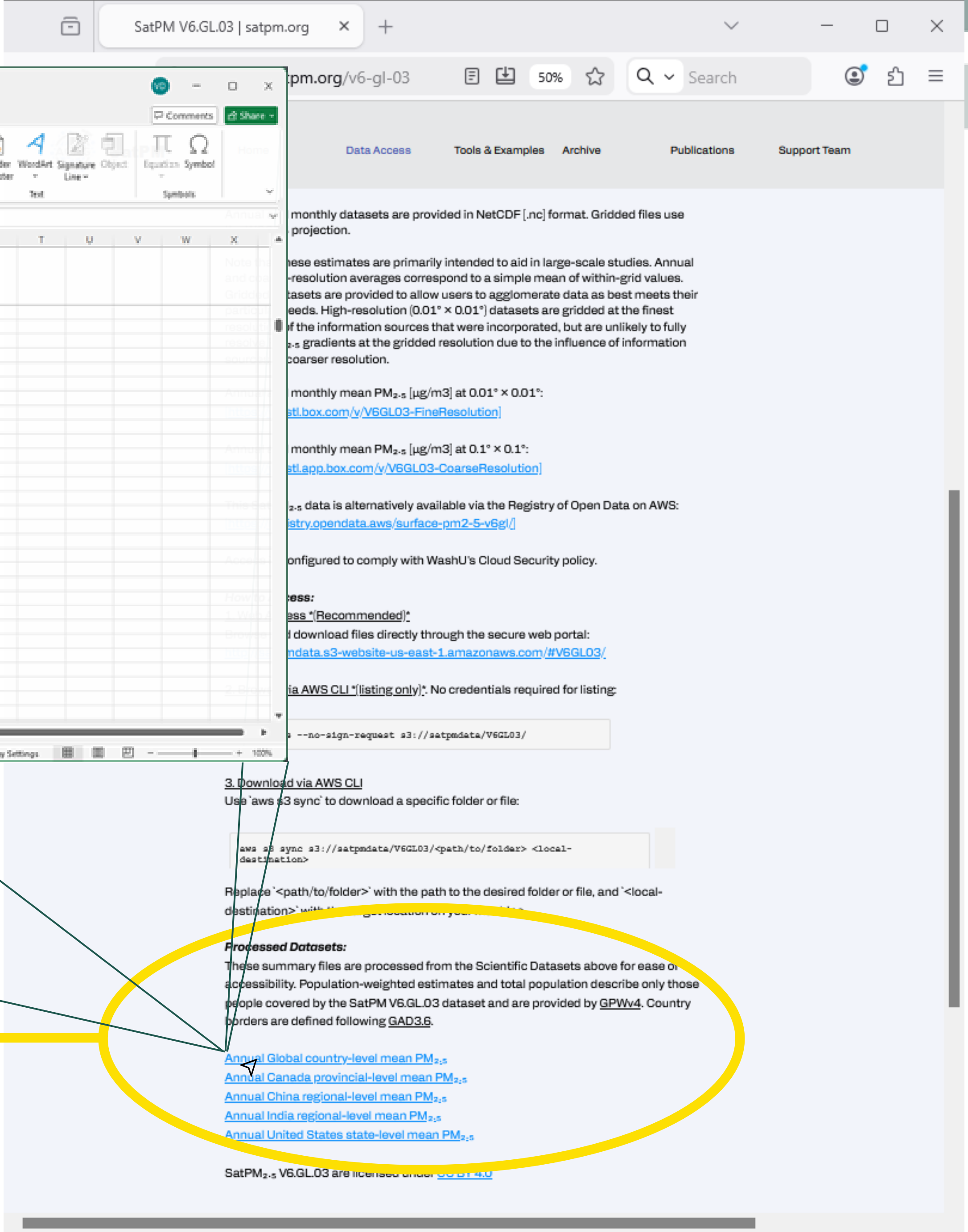
1	Region	Year	Population-Weighted PM2.5 [ug/m3]	Geographic-Mean PM2.5 [ug/m3]	Population Coverage [%]	Geographic Coverage [%]	Total Population [million people]	% pop >= 5 ug/m3 [%]	% pop >= 10 ug/m3 [%]	% pop >= 15 ug/m3 [%]	% pop >= 25 ug/m3 [%]	% pop >= 30 ug/m3 [%]	% pop >= 35 ug/m3 [%]	% pop >= 40 ug/m3 [%]	% pop >= 45 ug/m3 [%]	% pop >= 50 ug/m3 [%]	% pop >= 55 ug/m3 [%]	% pop >= 60 ug/m3 [%]
1081	Chile	2024	18.9	18.8	100	100	18.738	100	96.7	73	20.9	1.7	0.8	0.4	0.1	0	0	0
1082	China	1998	28.8	19.7	100	100	1266.692	100	99.5	97.4	64.4	45.1	23.2	6.7	0.9	0.4	0.3	0.2
1083	China	1999	27.1	18.5	100	100	1266.692	100	99.7	96.1	59.2	38.8	15.2	3.1	0.5	0.3	0.2	0
1084	China	2000	33.2	21.9	100	100	1266.692	100	99.9	98.8	77.1	57.1	42.2	23.4	12.8	5.2	1.2	0.6
1085	China	2001	35.1	23.4	100	100	1273.769	100	99.9	98.7	84	65.6	48.8	32.1	17	6.9	1.7	0.8
1086	China	2002	37.7	23.8	100	100	1280.846	100	99.9	98.9	86.5	68.3	55.1	43.6	29.5	15.7	7	2.3
1087	China	2003	42.3	26.6	100	100	1287.923	100	99.8	99.3	93.3	78.9	64.5	53.1	41.9	28.9	18.8	12.2
1088	China	2004	41.6	24.5	100	100	1295	100	99.8	99	90.9	79.4	65	51.1	41	29.5	16.5	9.5
1089	China	2005	44.1	25	100	100	1302.077	100	99.9	98.6	92	81.5	67.5	53.8	45.3	35.8	25.9	18
1090	China	2006	48.2	28	100	100	1309.099	100	99.9	99.3	94.1	88	76.7	62.2	50.4	40.2	31.8	24.7
1091	China	2007	48.9	28.3	100	100	1316.122	100	99.9	99.2	94.4	88.5	78.5	65.8	52.8	42.9	33	25.1
1092	China	2008	48.1	29	100	100	1323.144	100	99.9	99.5	94.9	88	78.5	66.5	55	46.3	36.4	25.6
1093	China	2009	45.9	26.5	100	100	1330.166	100	99.9	99.1	93.1	83.8	70.6	59.3	50.4	41.6	31	21.3
1094	China	2010	45.5	26.9	100	100	1337.189	100	99.9	98.9	91.2	80.6	67.1	54.6	47.5	40.6	32.8	24.5
1095	China	2011	50.4	29.4	100	100	1344.15	100	99.9	99.3	95.3	89	80.4	66.3	55.4	46.1	36.6	29.1
1096	China	2012	46.7	28.4	100	100	1351.11	100	100	99.2	93.5	84.2	73.1	60.2	49.4	38.4	30.1	22.4
1097	China	2013	49.4	27.7	100	100	1358.071	100	99.8	99.1	93.1	84.5	72.3	61.3	53.2	44.6	37.1	29.5
1098	China	2014	52.2	29.6	100	100	1365.032	100	100	99.5	95.2	89.9	80.6	70.6	60.5	51	42.2	34.1
1099	China	2015	50.2	31.5	100	100	1371.993	100	100	99.7	95.8	88.1	76	64.1	56.5	48.1	38	28.4
1100	China	2016	45.6	29.7	100	100	1377.299	100	100	99.6	93	82	68.6	57.2	46.5	36.1	27	20.3
1101	China	2017	43.4	26.1	100	100	1382.604	100	99.9	99.2	92.4	81.3	67.5	53.8	42.7	31.4	23	15.6
1102	China	2018	39.1	24.8	100	100	1387.91	100	100	98.9	88.6	72.5	55.9	41.8	32.3	22.7	14.2	6
1103	China	2019	38.3	24.2	100	100	1393.215	100	99.7	98.8	85.8	68.7	53.8	41.2	29.7	20.2	13.9	7
1104	China	2020	34.4	23.3	100	100	1398.52	100	99.7	98.5	75.2	56.9	43	30.2	21.3	12.6	4.2	1.1
1105	China	2021	32.3	22.6	100	100	1398.52	100	99.9	98.9	74.4	52.8	37.3	23.9	11.5	2.3	0.8	0.6
1106	China	2022	32.2	22.6	100	100	1398.52	100	99.8	98.4	70.5	53.5	36.7	23	10.8	5.1	2.7	1
1107	China	2023	33.3	23.1	100	100	1398.52	100	99.9	98.7	75.9	59.2	43.4	25.2	12	4.6	2	0.8
1108	China	2024	32.6	23.1	100	100	1398.52	100	99.9	99	73.5	57.6	40.3	24.5	9.6	3.5	1.5	0.8
1109	CÃ´te d'Ivoire	1998	19.7	21.5	100	100	16.473	100	100	97.2	8.5	0.1	0	0	0	0	0	0



via Box Folders
and via AWS Registry of Open Data

Processed Datasets Provide Simplified Summaries



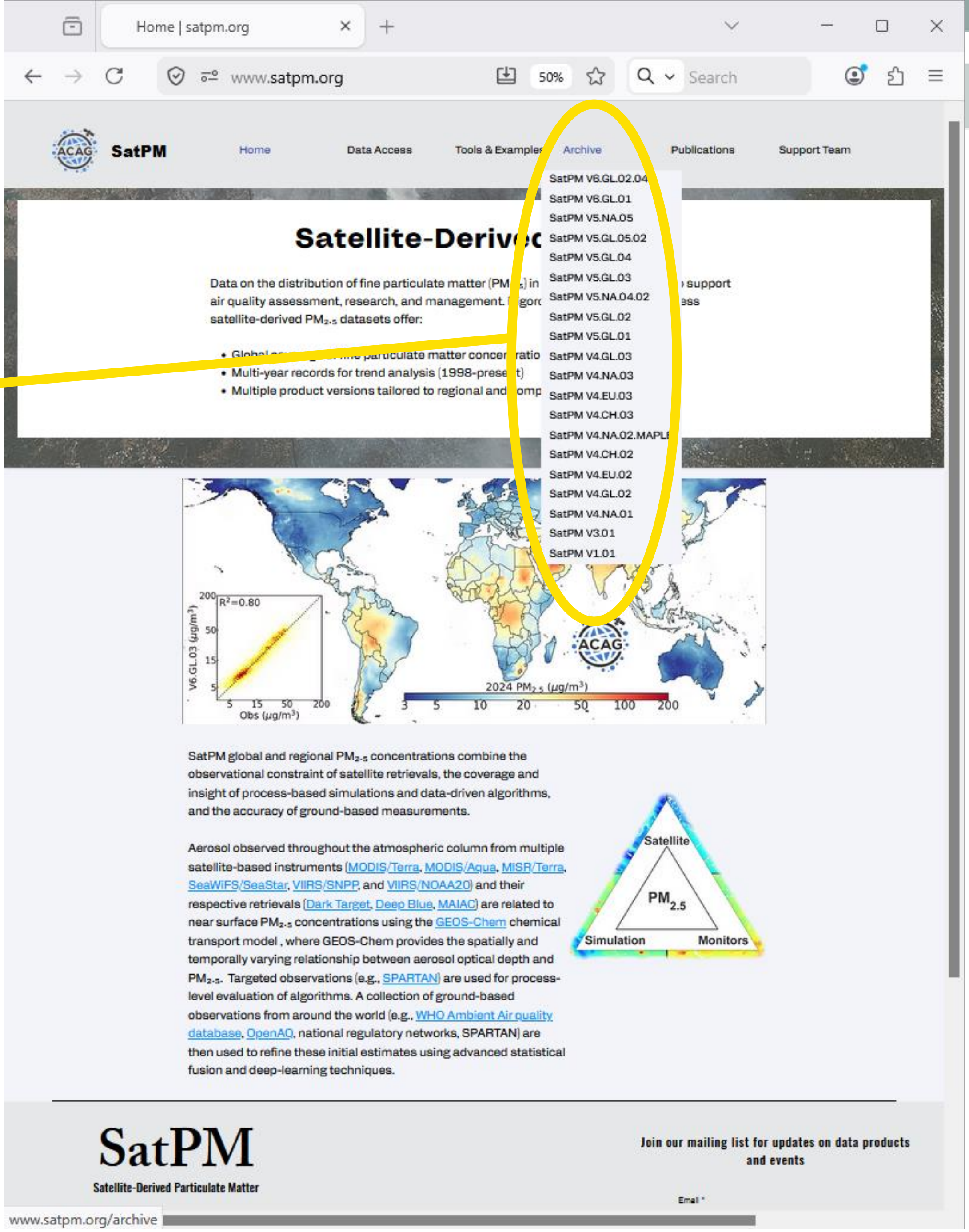


includes dataset/access instructions
via Box folders
and via AWS Registry of Open Data

Processed datasets provide simplified summaries

Dedicated Website (www.satpm.org) Provides Access to Latest SatPM$_{2.5}$ Versions

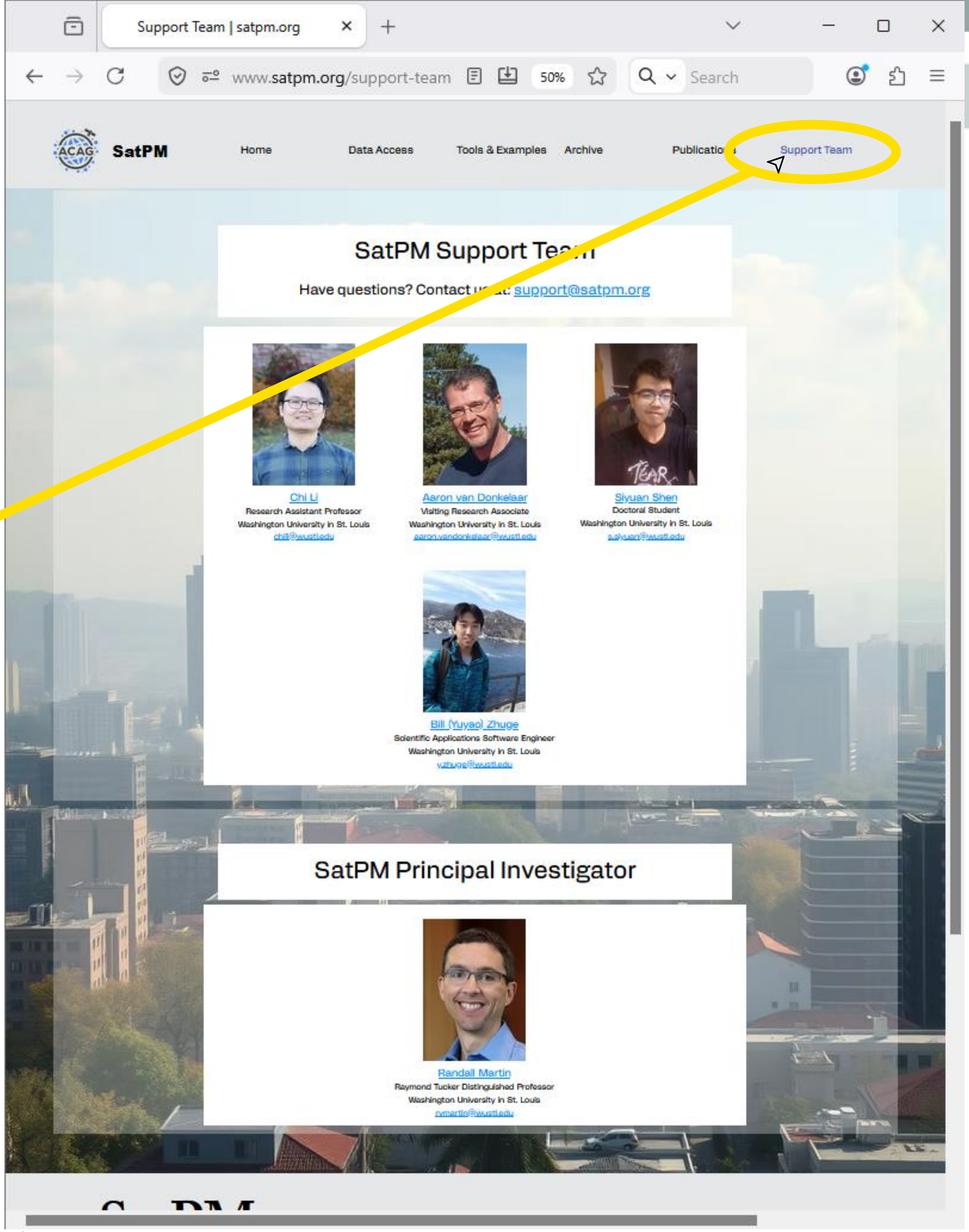
Previous Versions are Archived for Subsequent Access



Dedicated Website (www.satpm.org) Provides Access to Latest SatPM$_{2.5}$ Versions

Previous Versions are Archived for Subsequent Access

Dedicated Support Team Available for Assistance

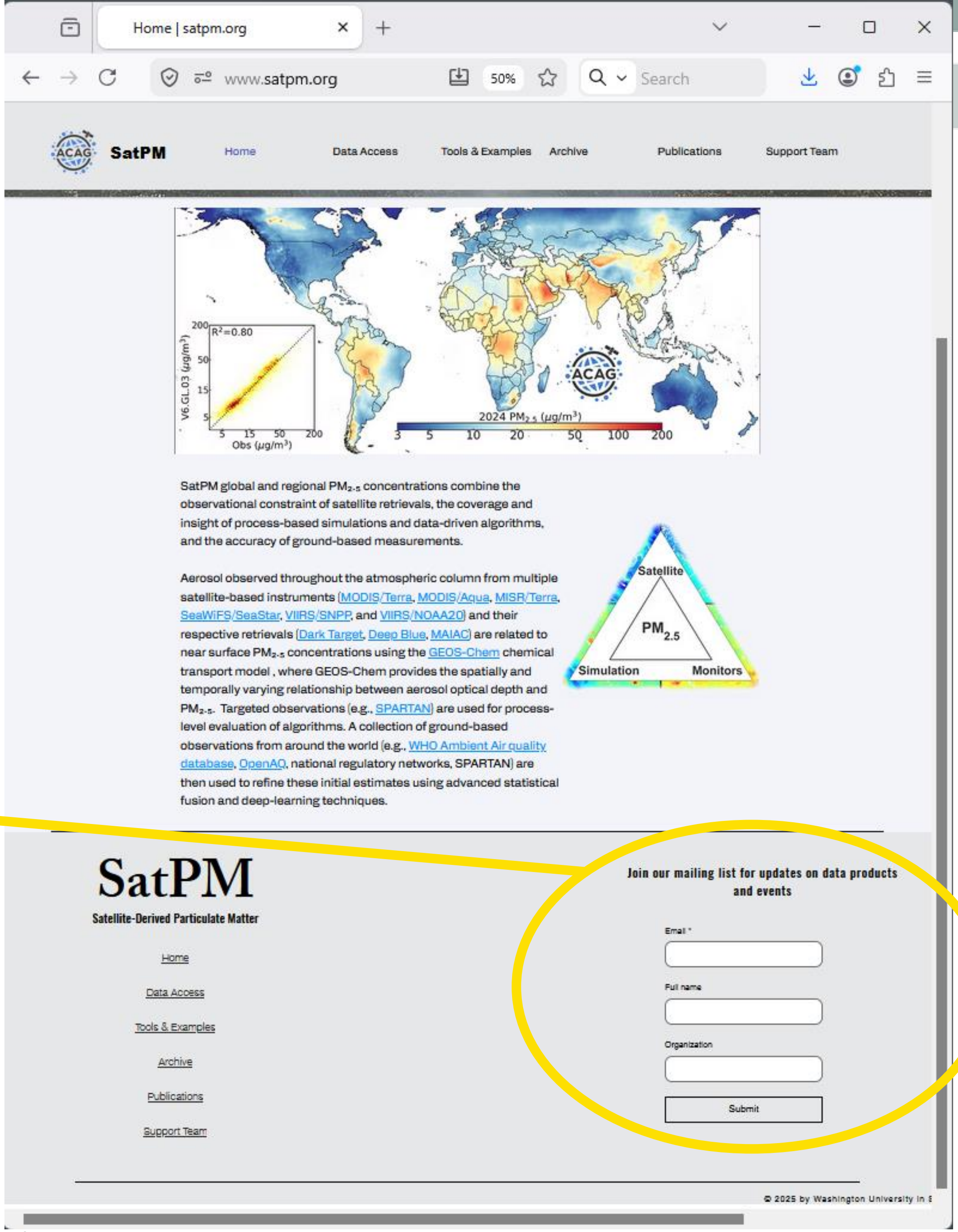


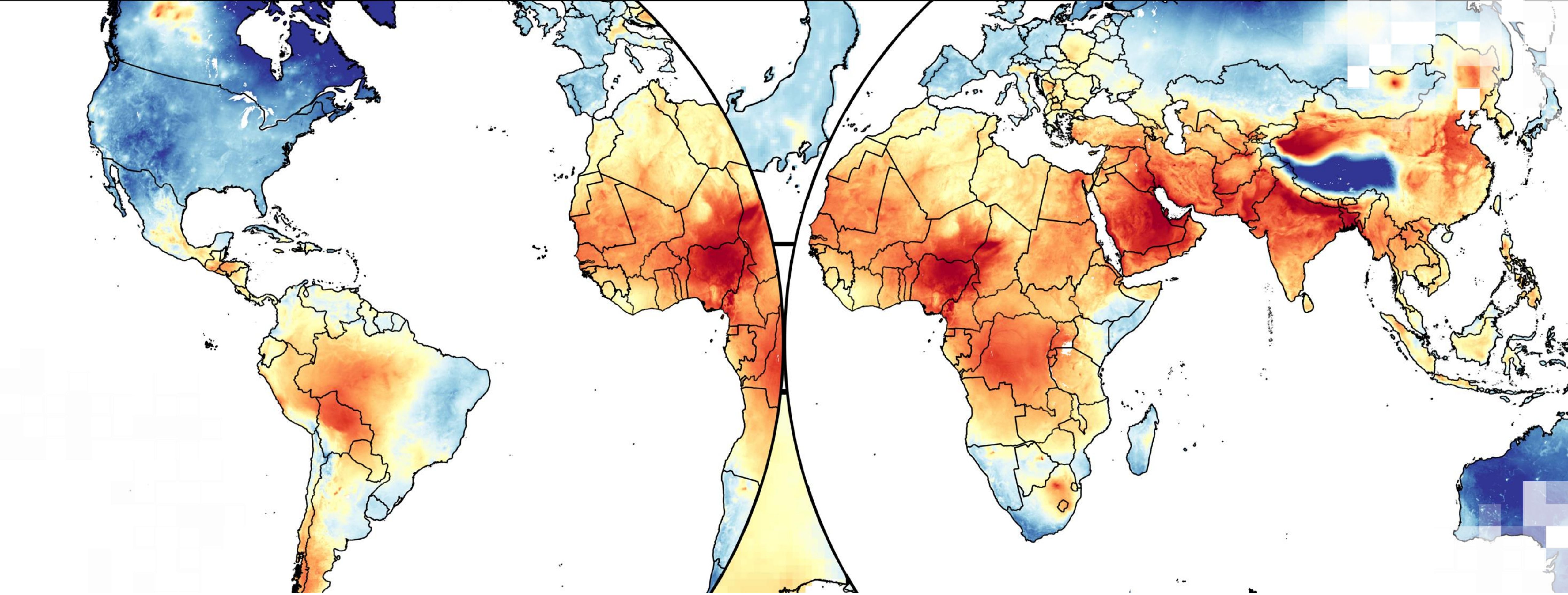
Dedicated Website (www.satpm.org) Provides Access to Latest SatPM$_{2.5}$ Versions

Previous Versions are Archived for Subsequent Access

Dedicated Support Team Available for Assistance

Sign up to receive data product updates!





Direct Machine Learning for $PM_{2.5}$

Direct Machine Learning Approaches for Surface $PM_{2.5}$ Estimation

- A Machine Learning (ML) model learns the mapping from satellite/reanalysis inputs to surface $PM_{2.5}$ without a chemical transport model in the loop.

Geophysical	Hybrid	Direct ML
Section 1	*Sections 1–2*	*This Section*
Chemical transport model (CTM) provides the AOD → $PM_{2.5}$ conversion physics. Satellite AOD constrains the result.	Start from a geophysical estimate; then calibrate to ground monitors using GWR, statistics, or ML methods like CNN.	ML model learns the mapping end-to-end from observations alone — no CTM scaffold.
e.g., MERRA-2, GEOS-Chem-based products	*e.g., WashU SatPM$_{2.5}$*	*e.g., GHAP, MERRA2_CNN_HAQAST_PM25*

Why Direct ML Grew: Satellite + reanalysis archives, ground-monitor networks (e.g., OpenAQ), and GPU computing made data-driven approaches tractable – and often more flexible in regions where CTM emissions inventories are weak.

ML-Based $PM_{2.5}$ Applications Across Different Time Horizons

PAST | NOW | FUTURE

Historical

"What was $PM_{2.5}$ in 2010?"

DEFINING CONSTRAINT

Full retrospective archive: satellite + reanalysis + ground monitors

TYPICAL USE

Reconstruct decades-long records for exposure studies, trend analysis

Near-Real-Time

"What is $PM_{2.5}$ right now?"

DEFINING CONSTRAINT

Streaming geostationary AOD + NRT meteorology forecasts

TYPICAL USE

Operational monitoring, wildfire alerts, public health advisories

Forecasting

"What will $PM_{2.5}$ be tomorrow?"

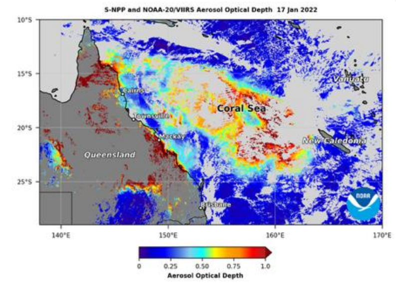
DEFINING CONSTRAINT

Forecast meteorology only; no current satellite observations

TYPICAL USE

Public alerts, planning, advance exposure mitigation

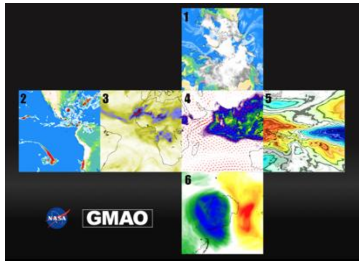
Typical Workflow of ML-Based $PM_{2.5}$ Estimation



1 Primary Input

Satellite AOD, Reanalysis Aerosols

MODIS, VIIRS, MISR, SeaWiFS, GOES-R/Geo, MERRA-2, CAMS



2 Auxiliary Fields

Mostly from Reanalysis (ERA-5, MERRA-2)

Humidity, Boundary layer height, Temperature, Winds, Surface pressure, Land cover

3 Ground Truth

Target Variable — In-Situ $PM_{2.5}$

EPA AirNow, OpenAQ, Regional Networks

ML MODEL

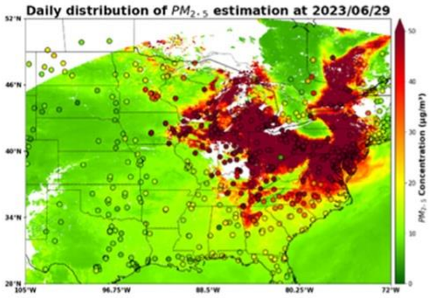
Random Forest
XGBoost
CNN, DNN, LSTM,
GNN, Transformer

OUTPUT

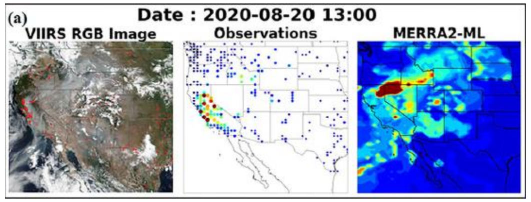
Surface $PM_{2.5}$

($\mu g/m^3$)

On Satellite/Reanalysis Grid

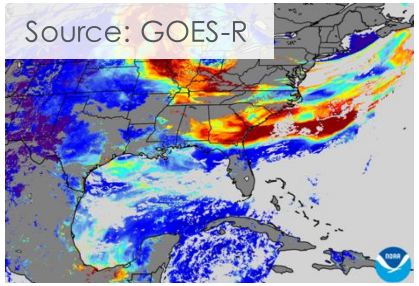


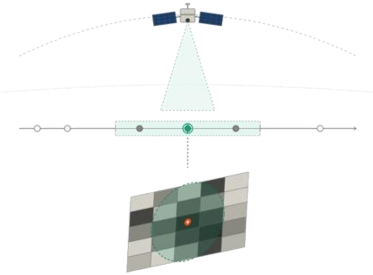
Park et al., (2025)

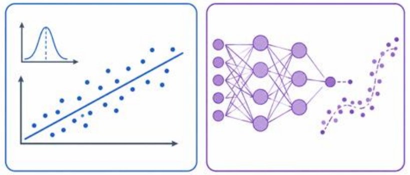


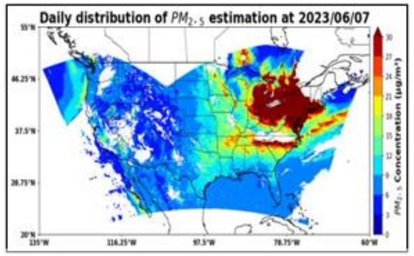
Sayeed et al., (2022)

From Satellite Signal to Surface PM$_{2.5}$ – ML Method Pipeline

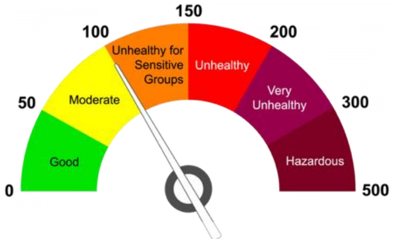

Source: GOES-R







Source: Park et al., (2025)



Source: US EPA

STEP 01

Acquire Data

Obtain AOD from satellite retrievals and PM$_{2.5}$ from ground monitoring stations.

- Satellite AOD
- Ground PM$_{2.5}$

STEP 02

Collocate

Match in time (closest to overpass) and space (nearest pixel or local average).

- Temporal match
- Spatial match

STEP 03

Build Model

Apply regression or ML models to predict PM$_{2.5}$ from satellite AOD.

- AOD ↔ PM$_{2.5}$
- Statistical & AI/ML

STEP 04

Estimate PM$_{2.5}$

Derive surface PM$_{2.5}$ using column AOD.

- Regression
- Machine Learning

STEP 05 (OPTIONAL)

Convert to AQI

Translate PM$_{2.5}$ concentrations into the Air Quality Index (AQI).

- Concentration
- AQI category

Examples of ML-Based $PM_{2.5}$ Datasets Across Temporal Modes

	⏪ Historical/Retrospective	▶ Near-Real-ime	⏭ Forecasting
Examples	**Di et al. (2019)** — CONUS, ensemble (NN + RF + GB) **Park et al. (2020)** — East Asia, RF + GOCI AOD **Wei et al. (2023)** — Global, big-data ML (GHAP)	**Pendergrass et al. (2022)** — East Asia, RF + GOCI **Sayeed et al. (2025)** — CONUS, DNN + GOES-R **Zhang & Kondragunta (2024)** — CONUS, DNN + geostationary	**Sayeed et al. (2021)** — CONUS, CNN + CMAQ **Lee et al. (2022)** — East Asia, ML + DA + GOCI **Seo & Gupta (2026)** — Global, TCN + MoE + GEOS-FP
Inputs	MODIS/MAIAC AOD, GOCI, modeled $PM_{2.5}$, reanalysis met, land cover, ground monitors	Geostationary AOD (GOES-R / GOCI), radiance, HRRR / NRT met, AirNow / AirKorea	Forecast meteorology (CMAQ, GEOS-FP, WRF), ground observations - no current AOD
Methods	Ensemble (NN + RF + GB), Random Forest, big-data ML with gap-filling	RF/LGBM/DNN bias-correction of operational products; gap-filled AOD	CNN/TCN bias-correction of CTM forecasts; data assimilation + ML; Mixture-of-Experts
Common ML pattern	Mostly **direct estimation** – rich data archives support full ML mapping	Often **bias correction** of an operational product (GOES-GWR, CAMS)	Predominantly **bias correction** of CTM forecasts
Highlight	**Global daily 1 km gapless (GHAP); CONUS 1 km ensemble widely used in health studies (Di et al. 2019)**	**RMSE ↓ 8.72 → 6.55 μg/m³ (Sayeed et al. 2025); DNN beats RF/LGBM by up to ~45% IOA**	**Reliable 2-day forecast; DA + ML improves $PM_{2.5}$ analysis; R^2=0.88, captures wildfire peaks**

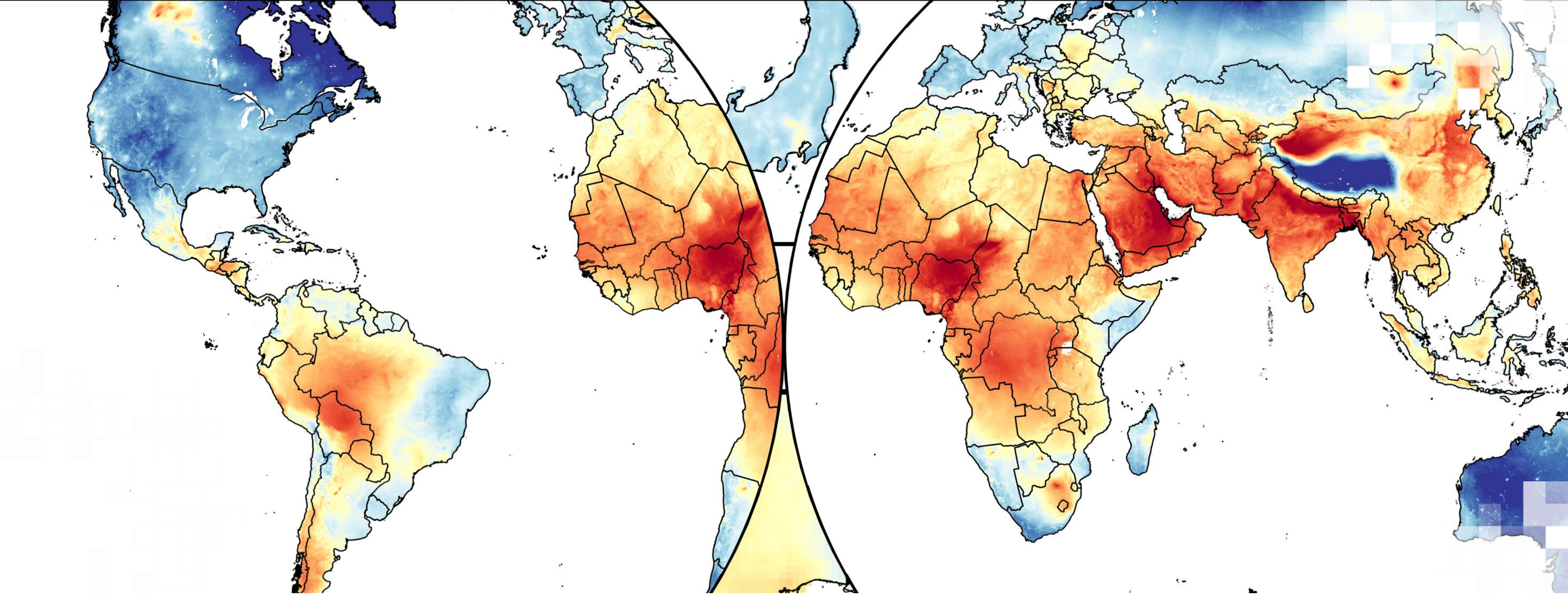
Strengths & Limitations of ML-Based $PM_{2.5}$

✓ Strengths

- Captures nonlinear AOD ↔ $PM_{2.5}$ relationships
- Flexible in regions with sparse CTM emissions inventories
- Can incorporate diverse predictors (met, land cover, time)
- Reduces systematic biases of physics-based estimates
- Fast inference once trained

✗ Limitations

- Depends heavily on density and quality of ground monitors; spatial CV is important!
- Risk of poor performance on extreme events (wildfires) if under-represented in training
- Less mechanistic interpretability than CTM-based methods
- Satellite AOD gaps (clouds, snow) propagate to $PM_{2.5}$
- Generalization across regions and time periods is not guaranteed



Example: Bias-Corrected MERRA-2

Comparative Overview of Global $PM_{2.5}$ Products

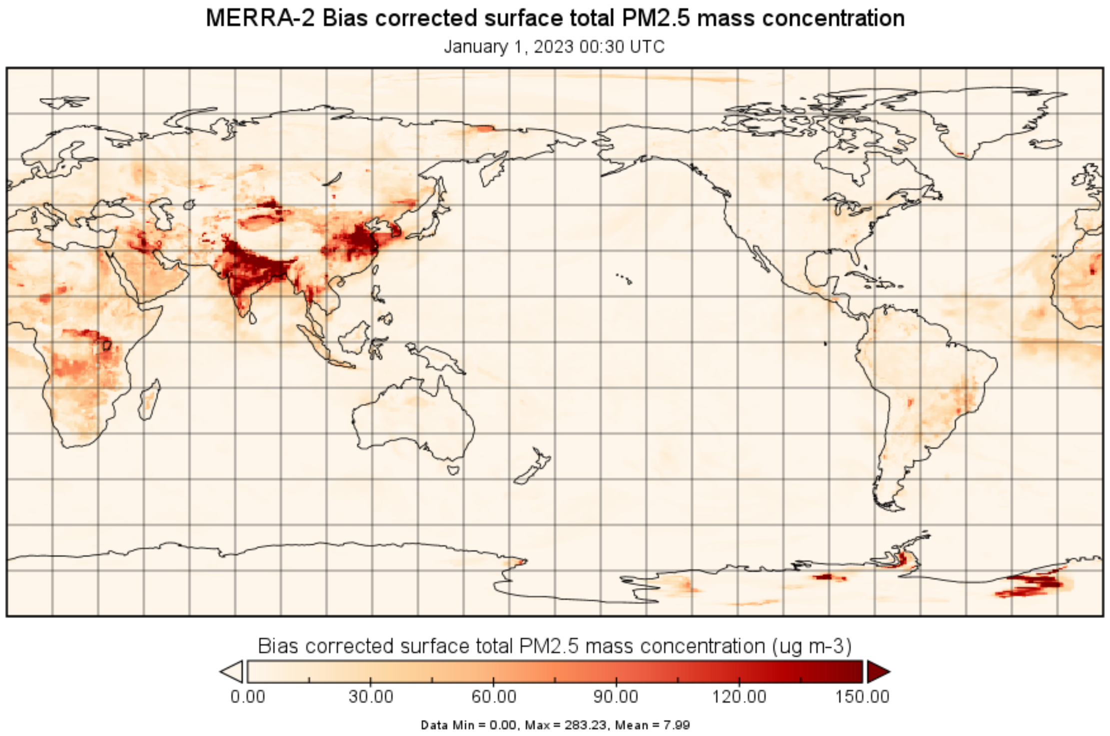
Dataset	Temporal Coverage	Resolution	Source/Method
DeepCAMS	2003–2020	0.25°, hourly	CAMS + Ground + Deep Learning
Global High Air Pollutants (GHAP)	2017–2022	1 km, daily	Satellite + Ground + Deep Learning
Berkeley Earth Real-time AQ	Real-time	0.10°, data gaps	Ground Measurements + Interpolation
WashU SatPM	**1998–2023**	**1 km, monthly/yearly**	**Satellite + Models + Ground**
MERRA2_CNN_HAQAST_PM25	**2000–2024**	**0.5° × 0.625°, hourly**	**MERRA-2 + CNN Ensemble + OpenAQ**

MERRA2_CNN_HAQAST_PM25 Niche:
Long-term (25 years), globally complete, hourly, and bias-corrected — filling the gap left by daily/monthly products and short-record hourly products.

MERRA2_CNN_HAQAST_PM25 Product Overview

Temporal Coverage	2000–2024 (periodically updated)
Temporal Resolution	Hourly (00:30–23:30 UTC)
Spatial Resolution	0.5° latitude × 0.625° longitude
Spatial Coverage	Global (90°S–90°N, 180°W–180°E)
File Format	NetCDF, ~14 MB per daily file
Variables	MERRA2_CNN_Surface_PM25 (μg/m³), QFLAG (1–4)

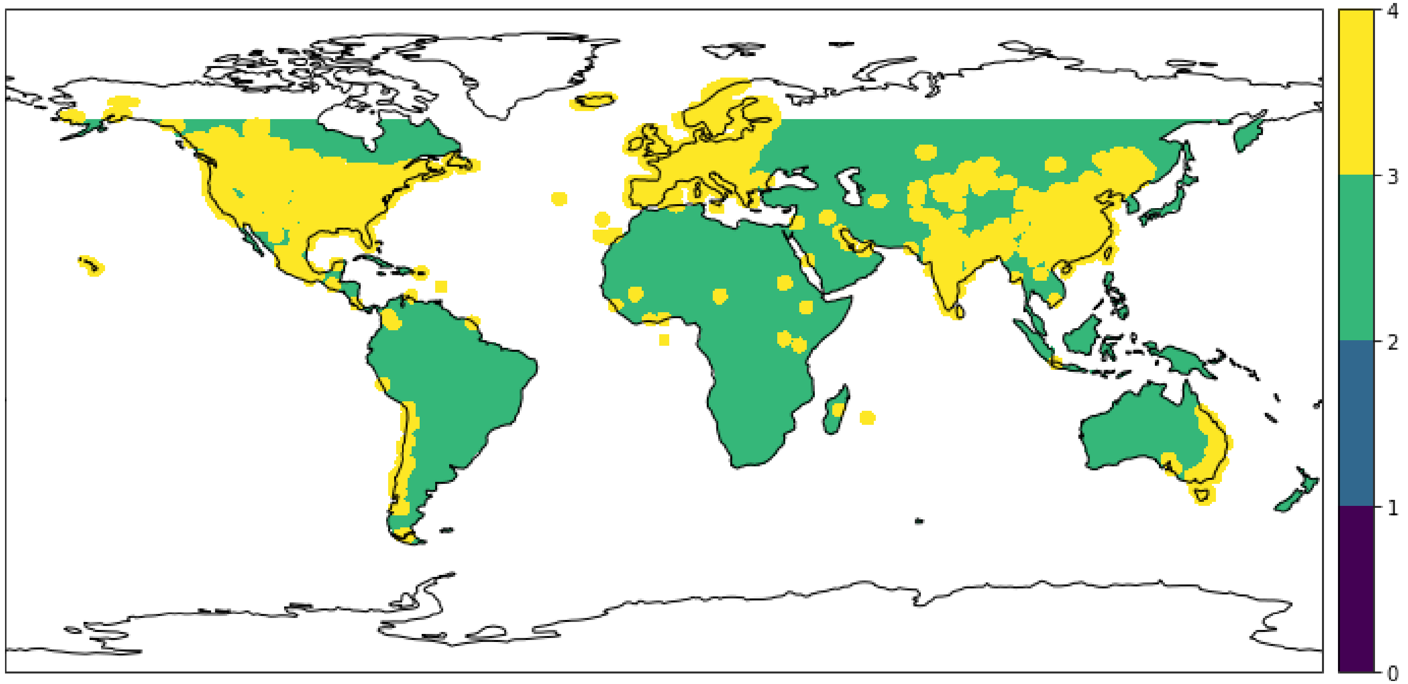
DOI: Gupta and Sayeed (2026)
10.5067/OCKK5HCFW5N3



Bias-Corrected Surface Total $PM_{2.5}$ — 1 January 2023, 00:30 UTC.

MERRA2_CNN_HAQAST_PM25 Quality Control

- Quality Flag (QFLAG) provides qualitative indicator of data quality
- QFLAG determined based on proximity to surface monitors, land cover type, latitude
- QFLAG Values 1-4
  - QFLAG = 1: Low Data Quality
  - QFLAG = 2: Medium Data Quality
  - QFLAG = 3: High Data Quality
  - QFLAG = 4: Best Data Quality
- **Use QFLAG ≥ 3 for quantitative analysis.**



Bias-Corrected Surface Total $PM_{2.5}$ — 8 June 2023, 00:30 UTC.
Inset map shows the QFLAG quality flag at each grid point (1 = low, 4 = best).

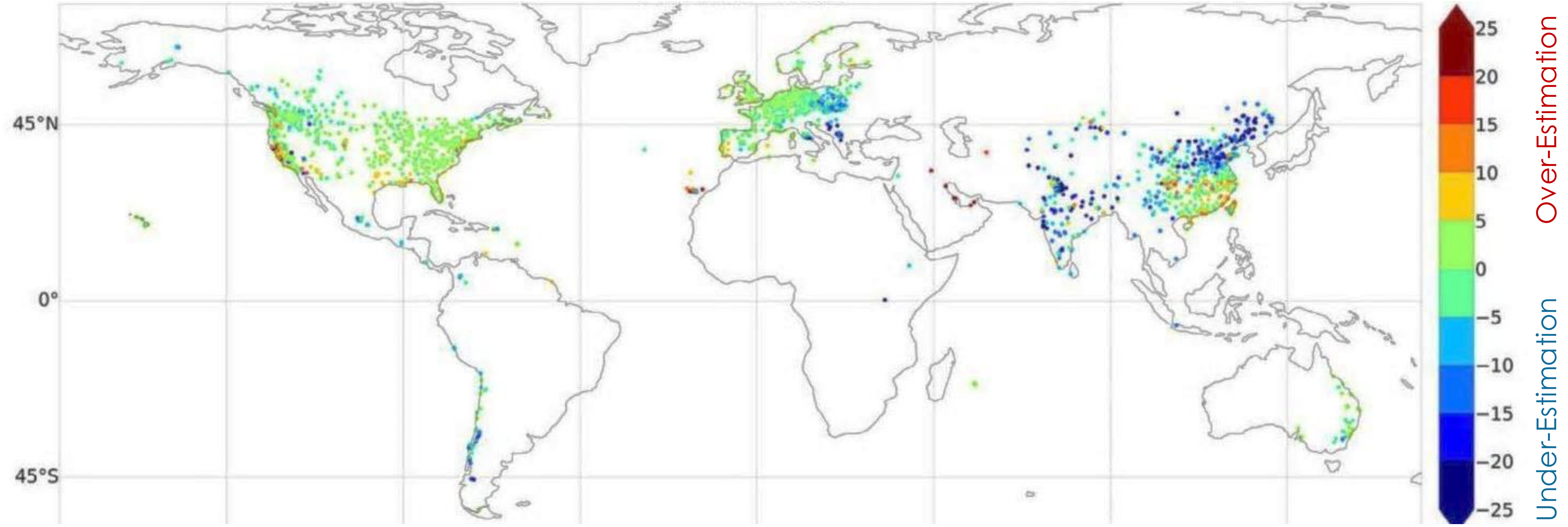
Motivation for MERRA-2 $PM_{2.5}$ Bias Correction

- Raw MERRA-2 $PM_{2.5}$ is computed empirically:

$$PM_{2.5} = DUST_{2.5} + SS_{2.5} + BC + OC + 1.375 \times SO_4$$

  - **Missing Species:** No nitrate, no ammonium — both can contribute substantially in many regions
  - **Grid-Box Averaging:** ~50 × 65 km cells smooth over local sources and gradients
  - **Emission & Meteorology Assumptions:** Bias varies regionally and seasonally
  - **Aerosol Size & Composition:** Semi-volatiles and hygroscopic growth not captured

Mean Bias of Raw MERRA-2 $PM_{2.5}$ vs. Ground Monitors in 2020



- Bias patterns vary by region and aerosol regime
- ML approach can correct these biases

Result: Regionally Variable Bias

Over-Estimation: California, Eastern China, Middle East (> +20 μg/m³)

Under-Estimation: Indo-Gangetic Plain, Southern South America (< −15 μg/m³)

Inputs: MERRA-2 Aerosols and Meteorology

- Selected based on their geophysical connection to surface $PM_{2.5}$ formation, transport, and removal.

Aerosol Diagnostics (M2T1NXAER)	
BCSMASS	Black Carbon Surface Mass
OCSMASS	Organic Carbon Surface Mass
DUSMASS25	Dust Surface Mass ($PM_{2.5}$ fraction)
SO2SMASS	SO_2 Surface Mass
SO4SMASS	Sulfate Surface Mass
TOTEXTTAU	Total Aerosol Extinction AOT (550 nm)

Meteorology (M2T1NXSLV)	
PS	Surface Pressure
QV10m, Q500	Specific Humidity (10 m, 500 hPa)
T10m, T500, T850	Temperature (10 m, 500, 850 hPa)
U10, V10	Wind Components at 10 m

All inputs are at MERRA-2 native resolution (0.5° × 0.625°, hourly) — ensuring the output product inherits the same grid and timing.

Ground Truth: OpenAQ Reference-Grade Observations

- Reference-grade $PM_{2.5}$ observations obtained from OpenAQ

- How OpenAQ Data were used:
  - Each station is spatially collocated with its nearest MERRA-2 grid point
  - Hourly $PM_{2.5}$ values temporally are matched with MERRA-2 inputs
  - Quality control is applied to remove outliers and invalid records
  - Split: 2018–2019 training, 2020 independent validation

5,187

Monitoring Stations Worldwide

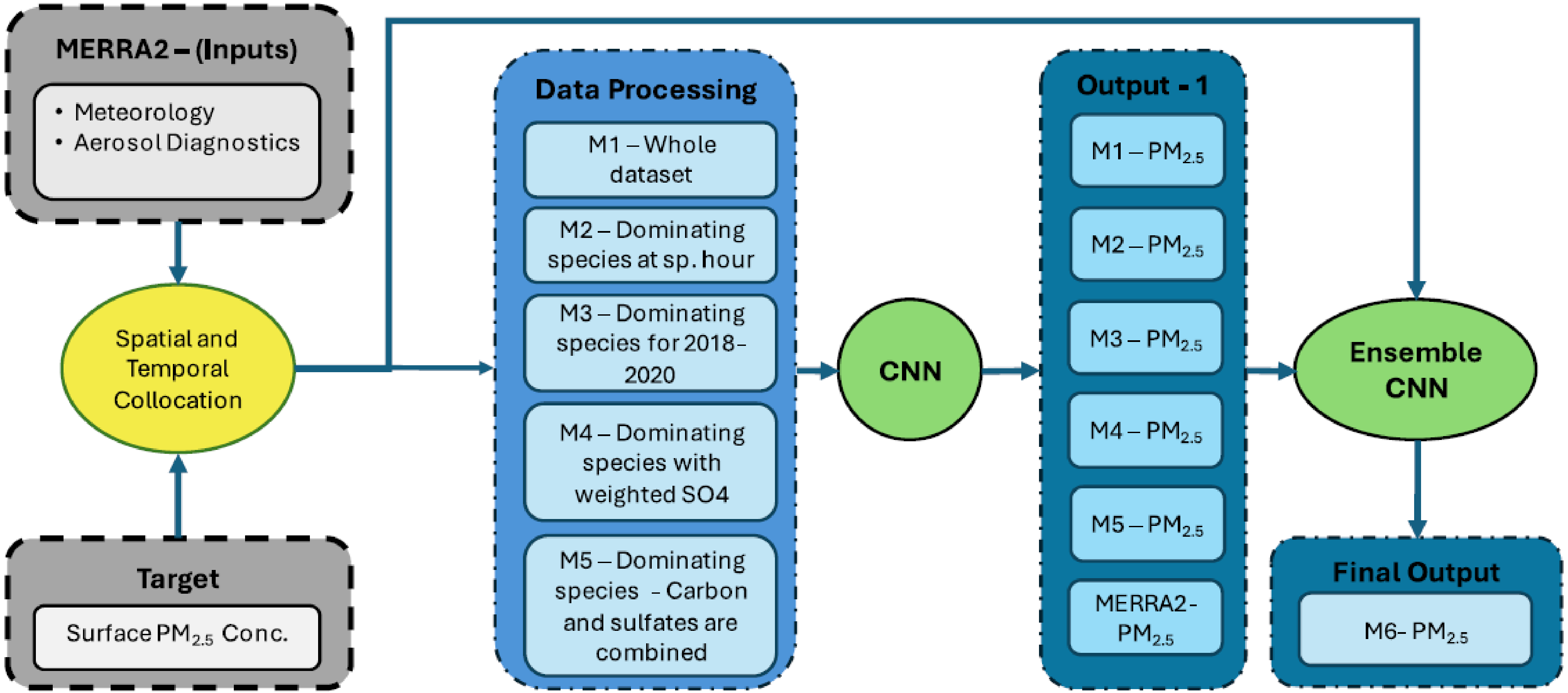
30.4M

Valid Hourly Collocated Observations

2018–2020

Data Collection Period

Methodology: From MERRA-2 Inputs to Bias-Corrected $PM_{2.5}$



Gupta & Sayeed, 2026

1. MERRA-2 inputs spatially & temporally collocated with OpenAQ surface $PM_{2.5}$
2. Five aerosol-aware data-processing strategies (M1–M5) each train an independent CNN
3. Ensemble CNN (M6) combines M1–M5 outputs → single bias-corrected $PM_{2.5}$ value per grid per hour

Five Aerosol-Aware Strategies → One Ensemble

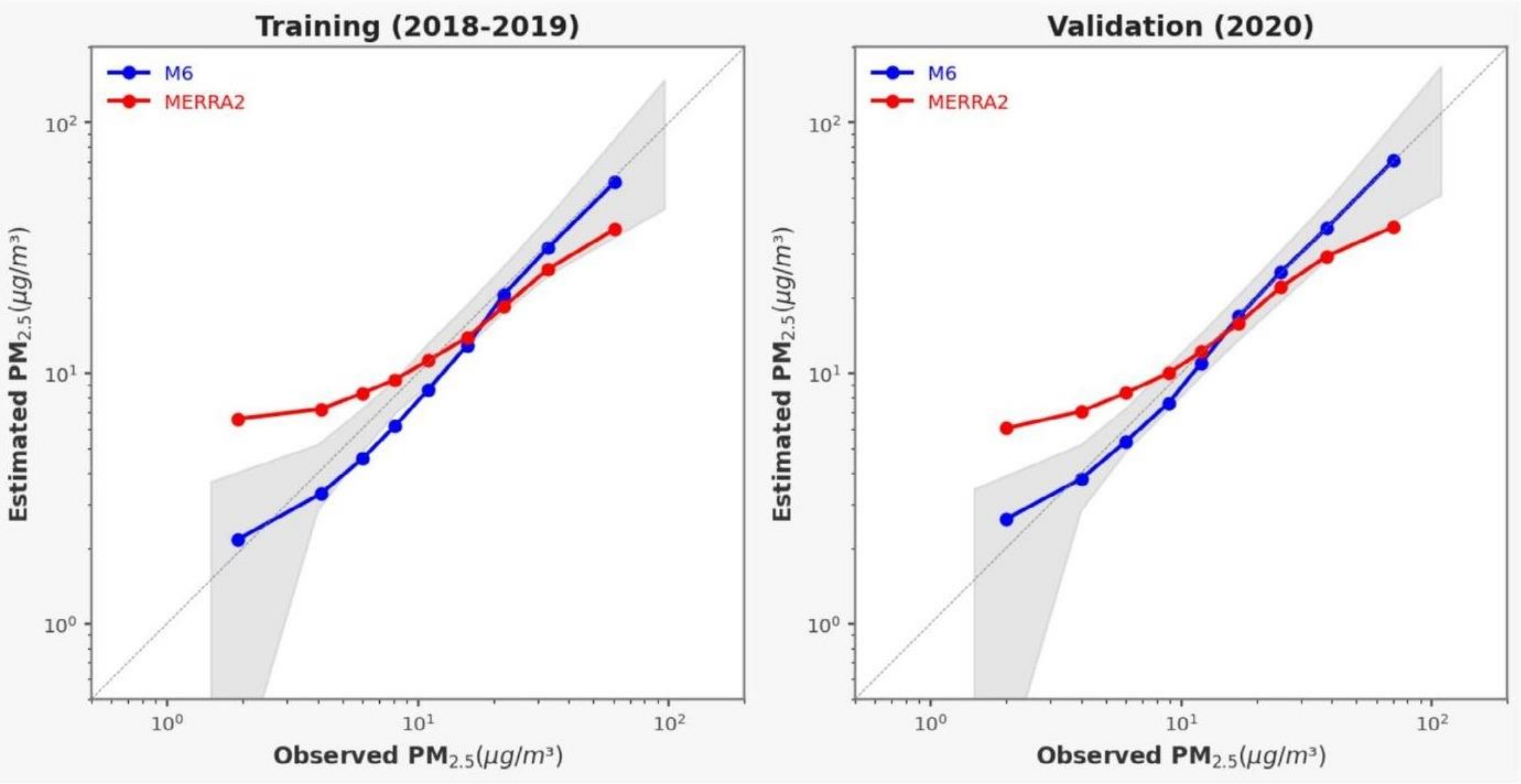
- Each strategy trains an independent CNN; their predictions are combined by a final ensemble model (M6)

M1	M2	M3	M4	M5
Baseline Whole dataset, no stratification.	**By Hour** Dominant aerosol species at each hour, from MERRA-2.	**By Station** Each station assigned its average dominant aerosol type (2018–2019).	**By Grid Cell** Each MERRA-2 grid cell assigned its dominant species (2018–2020).	**M4 Refined** M4 with sulfate + carbonaceous aerosols merged into one class.

M6 Ensemble CNN

Combines M1–M5 outputs → single bias-corrected $PM_{2.5}$ per grid per hour

Validation: M6 reduces concentration-dependent bias in MERRA-2.



Setup

Observed $PM_{2.5}$ binned into 10 quantiles (1st–99th percentile).

M6 (Blue)

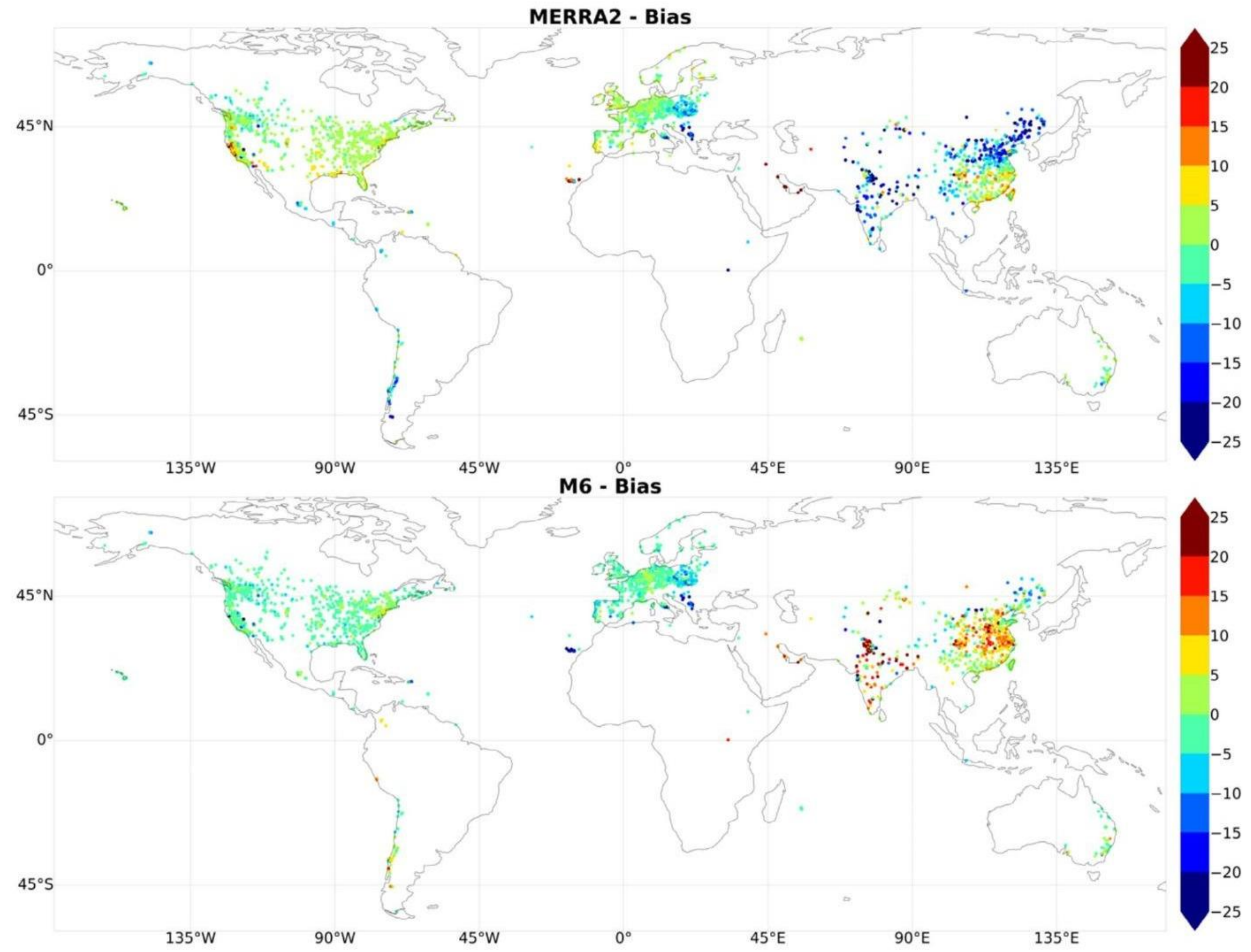
Tracks the 1:1 line across the full concentration range and stays inside the $\pm 2\sigma$ band.

MERRA-2 (Red)

Systematically overestimates at low $PM_{2.5}$; underestimates at high $PM_{2.5}$. Both effects visible in training and validation.

Gupta & Sayeed, 2026

Validation: M6 substantially reduces bias across regions in MERRA-2.



Key Findings:

MERRA-2 (Top)

Wide Error Range: > +20 μg/m³ in California, Eastern China, Middle East; < −15 μg/m³ in Indo-Gangetic Plain, Southern S. America.

M6 (Bottom)

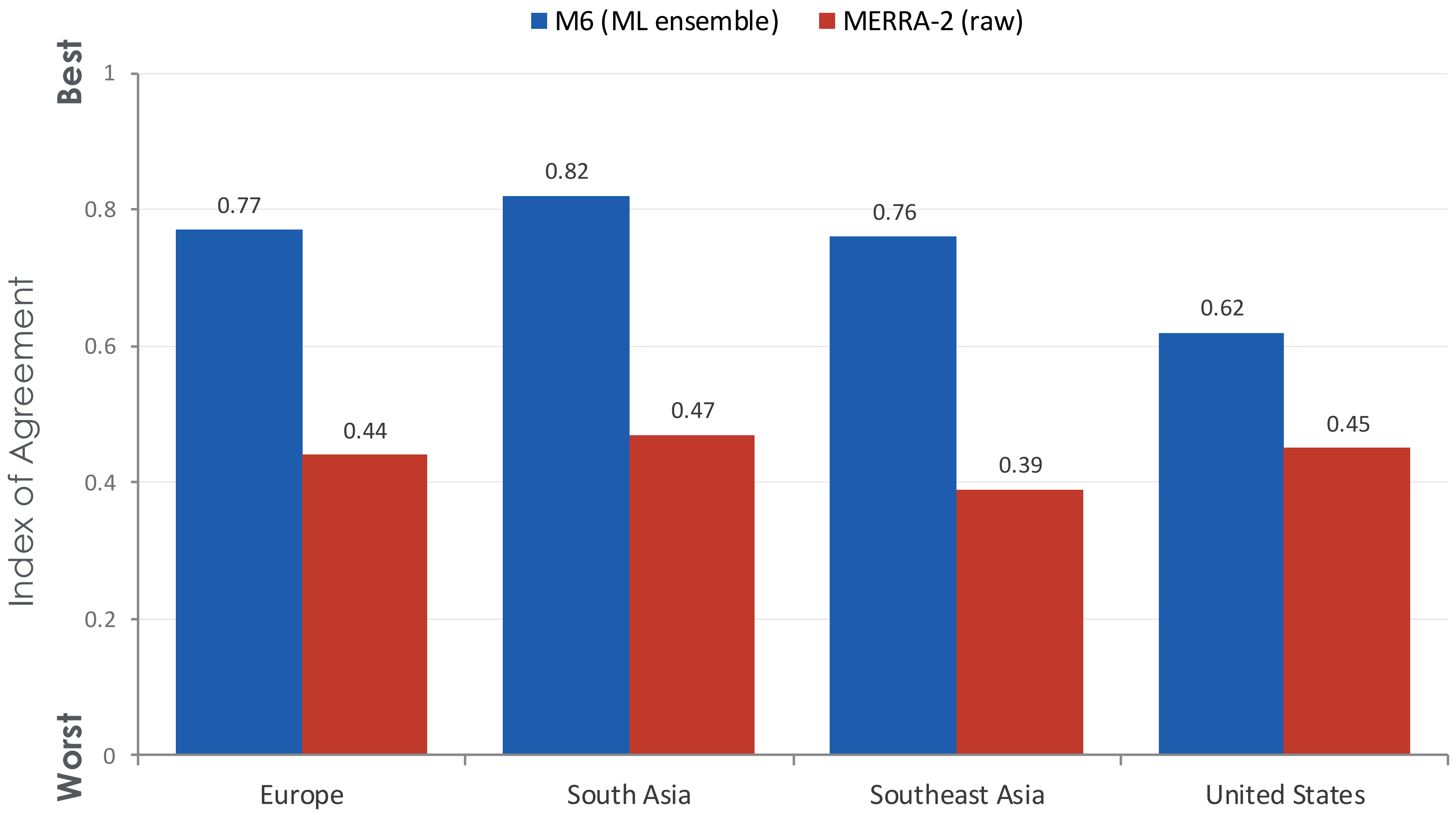
Substantially reduced bias, more localized error patterns, and better calibration across diverse aerosol regimes.

Honest Caveat

Some M6 over-prediction remains in E/SE Asia at high concentrations – though IOA still exceeds MERRA-2 there.

Regional Performance: M6 Outperforms MERRA-2 Everywhere

Median Index of Agreement (IOA) by region — higher is better, max = 1.



What IOA Measures

Index of Agreement scores how well the predictions reproduce both magnitude and pattern. Closer to 1 is better.

Almost 2× Better

In Southeast Asia (0.76 vs. 0.39) and South Asia (0.82 vs. 0.47).

Take-Home

M6 better reproduces temporal & spatial variability of $PM_{2.5}$.

Summary for MERRA2_CNN_HAQAST_PM25 Bias-Corrected MERRA-2 $PM_{2.5}$

✓ Long-Term, Global, Hourly

2000–2024 coverage at MERRA-2 native 0.5°×0.625° resolution.

✓ Bias-Corrected via CNN Ensemble

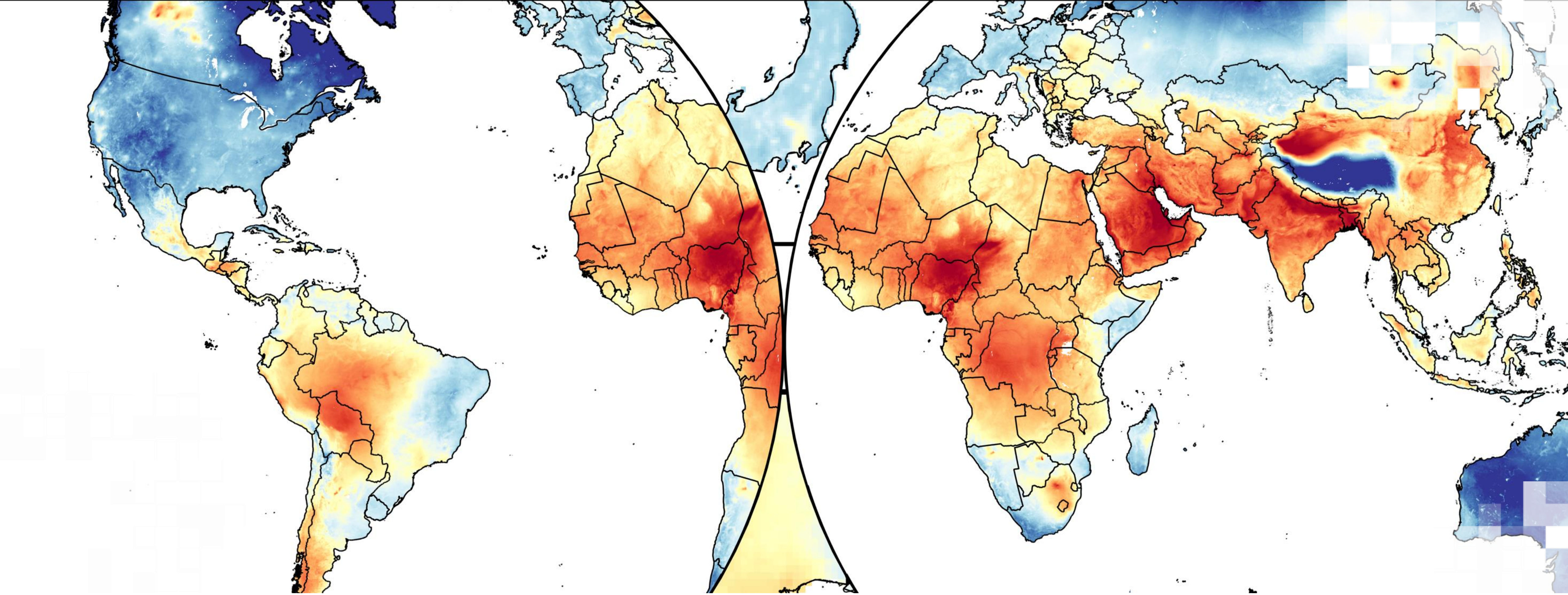
Aerosol-aware stratification (M1–M5) combined into an ensemble (M6).

✓ Improves $PM_{2.5}$ over MERRA-2

Higher IOA, reduced spatial bias, robust to unseen years and regions.

✓ Quality-Aware

QFLAG enables filtering;
use ≥ 3 for quantitative work.

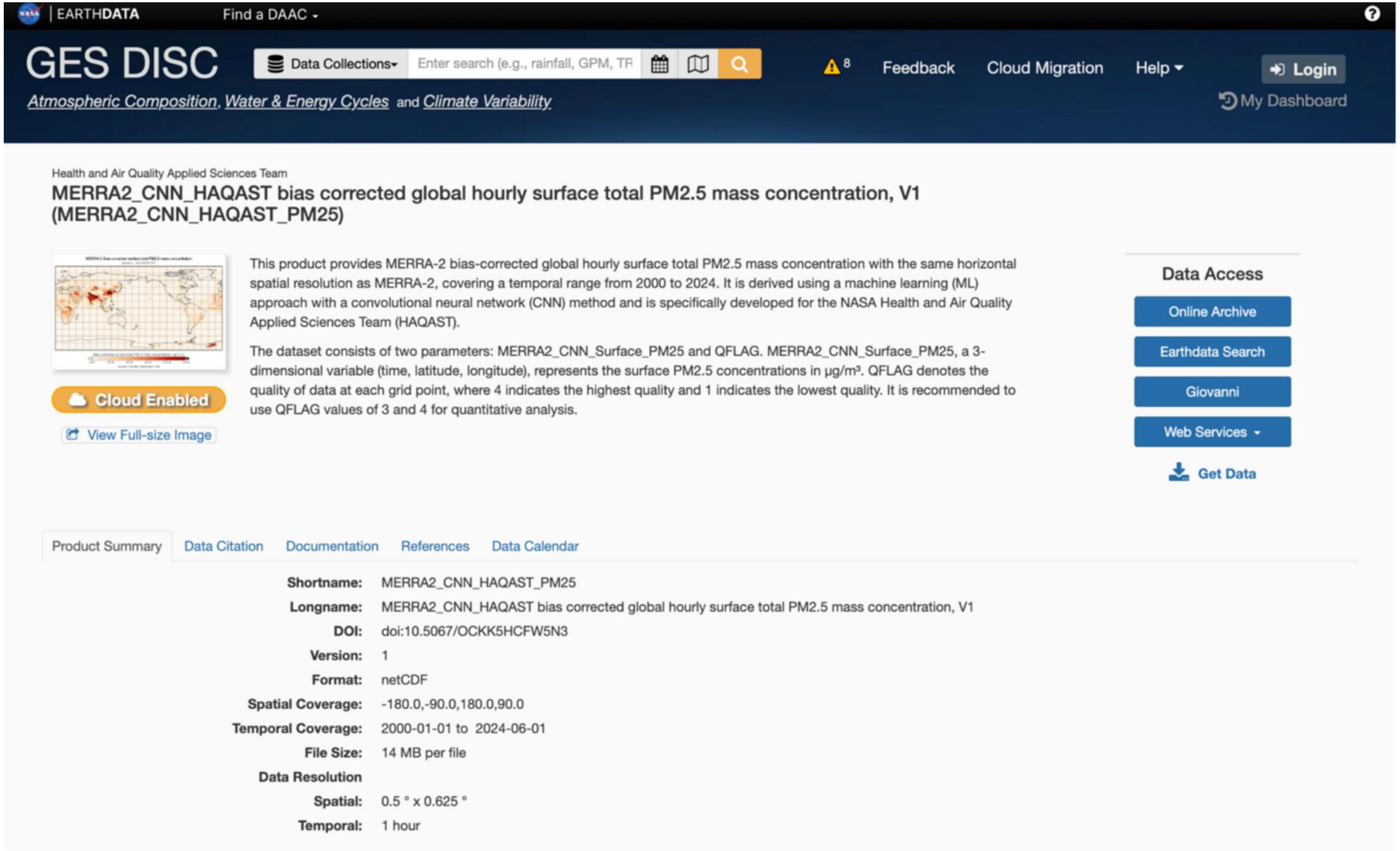


Bias-Corrected MERRA-2 Data Access

Data Access Options

Resource	Description
Data Landing Page	Primary source for information about the product, including access options and links to other resources
Readme Document	Detailed description of the dataset and the format of data files
Online Archive	Download individual files
Earthdata Search	Search for and download sets of files for a specific time interval
AWS	Access data directly on in Amazon S3 cloud
OPeNDAP	Remotely access files using common programming languages and tools (e.g., Python, Matlab, R, Panoply)
Giovanni	Online tool to subset, average, and plot data (e.g., map, time series)

Data Landing Page at NASA GES DISC



EARTHDATA Find a DAAC

GES DISC

Data Collections | Enter search (e.g., rainfall, GPM, TR

Feedback Cloud Migration Help Login

My Dashboard

Atmospheric Composition, Water & Energy Cycles and Climate Variability

Health and Air Quality Applied Sciences Team

MERRA2_CNN_HAQAST bias corrected global hourly surface total PM2.5 mass concentration, V1 (MERRA2_CNN_HAQAST_PM25)

This product provides MERRA-2 bias-corrected global hourly surface total PM2.5 mass concentration with the same horizontal spatial resolution as MERRA-2, covering a temporal range from 2000 to 2024. It is derived using a machine learning (ML) approach with a convolutional neural network (CNN) method and is specifically developed for the NASA Health and Air Quality Applied Sciences Team (HAQAST).

The dataset consists of two parameters: MERRA2_CNN_Surface_PM25 and QFLAG. MERRA2_CNN_Surface_PM25, a 3-dimensional variable (time, latitude, longitude), represents the surface PM2.5 concentrations in μg/m³. QFLAG denotes the quality of data at each grid point, where 4 indicates the highest quality and 1 indicates the lowest quality. It is recommended to use QFLAG values of 3 and 4 for quantitative analysis.

Cloud Enabled

View Full-size Image

Data Access

Online Archive

Earthdata Search

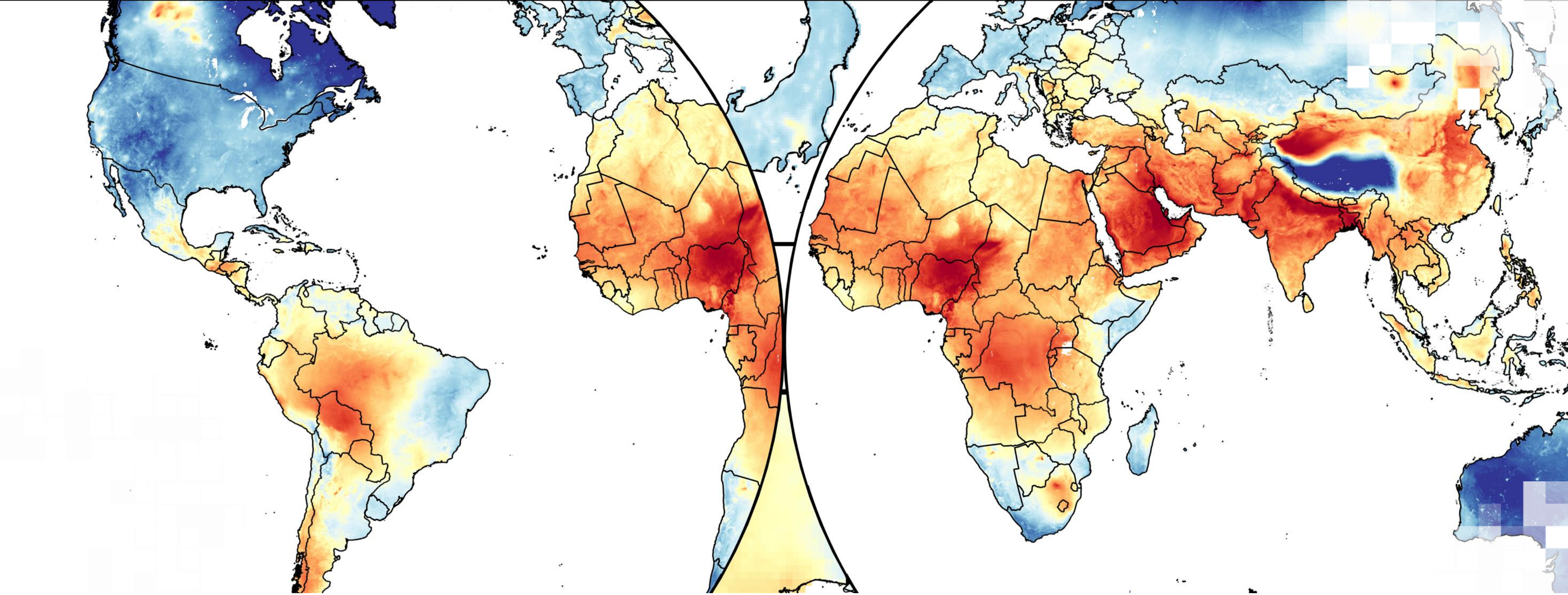
Giovanni

Web Services

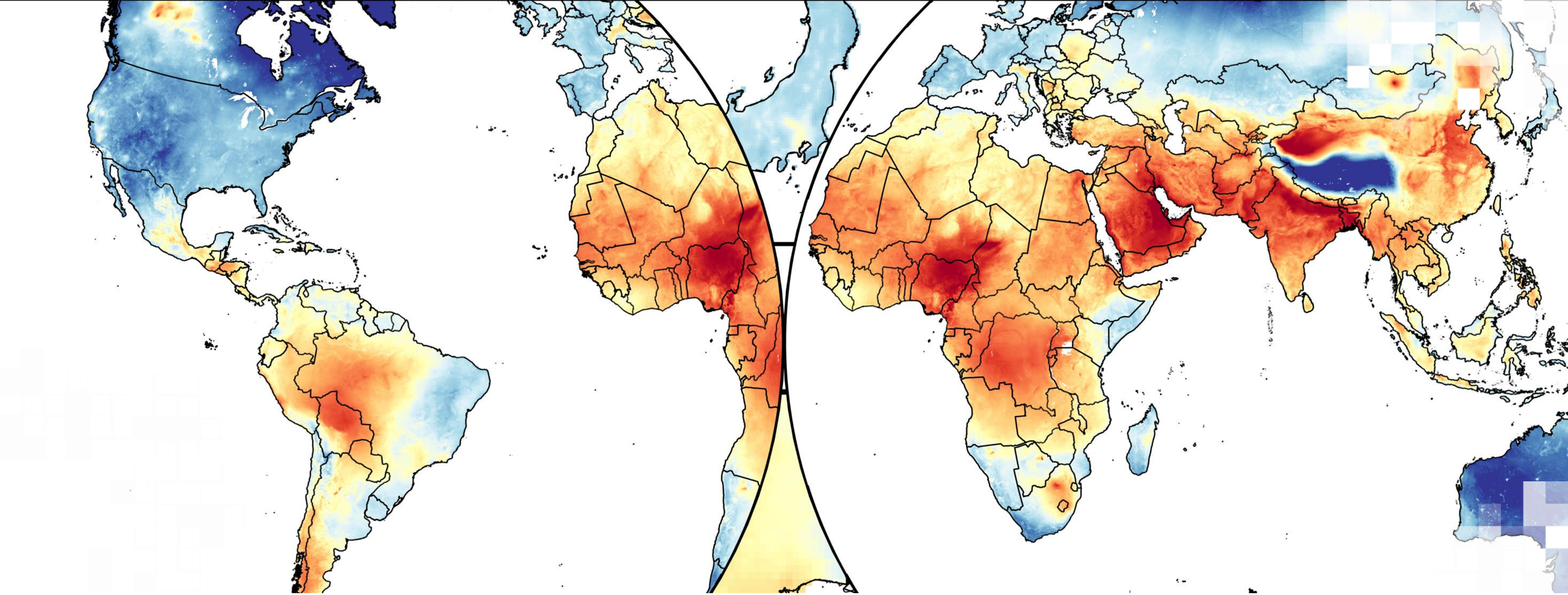
Get Data

Product Summary | Data Citation | Documentation | References | Data Calendar

Shortname:	MERRA2_CNN_HAQAST_PM25
Longname:	MERRA2_CNN_HAQAST bias corrected global hourly surface total PM2.5 mass concentration, V1
DOI:	doi:10.5067/OCKK5HCFW5N3
Version:	1
Format:	netCDF
Spatial Coverage:	-180.0,-90.0,180.0,90.0
Temporal Coverage:	2000-01-01 to 2024-06-01
File Size:	14 MB per file
Data Resolution	
Spatial:	0.5 ° x 0.625 °
Temporal:	1 hour



Visualizing Bias-Corrected MERRA-2 $PM_{2.5}$ with NASA Giovanni



Part 2:
Summary

Summary

Washington University in St. Louis SatPM$_{2.5}$

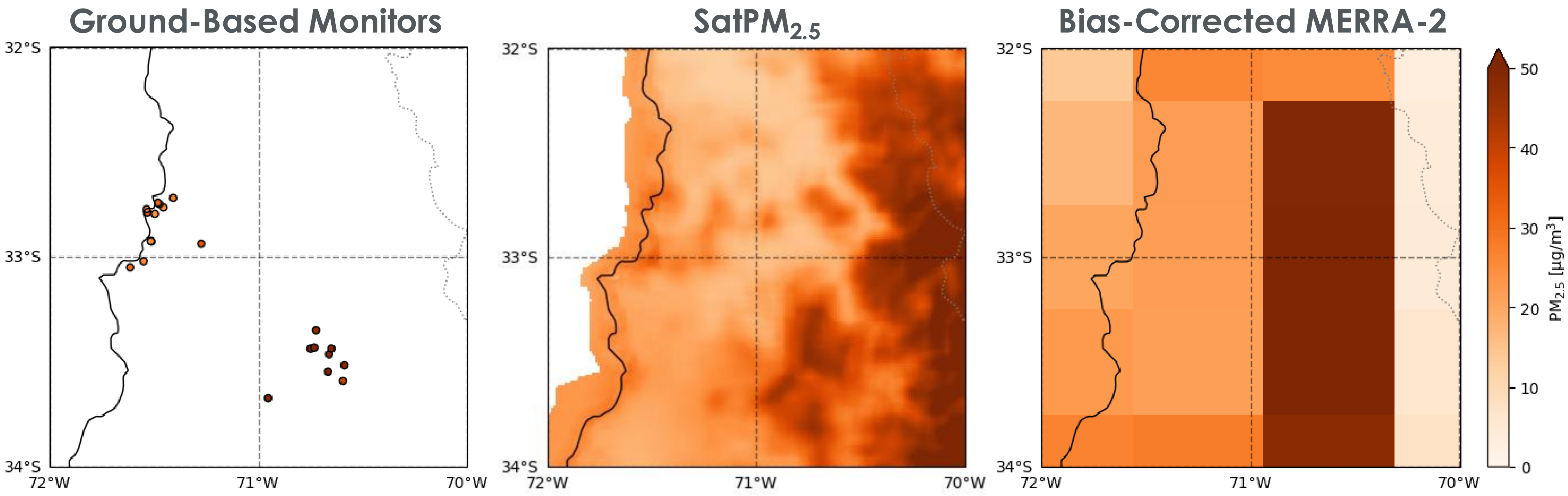
- Methodology:
  - Hybrid-Geophysical
- Strengths:
  - Fine spatial resolution (0.01°×0.01°)
  - Long record (1998-2024)
  - Consistent performance even far from ground-based monitors
- Limitations:
  - Monthly temporal resolution
- Data Access:
  - Free & open
  - http://satpm.org

NASA Bias-Corrected MERRA-2 PM$_{2.5}$

- Methodology:
  - Ensemble CNN
- Strengths:
  - Hourly temporal resolution
  - Long record (2000-2024)
  - Better represents PM$_{2.5}$ spatial and temporal variability compared to MERRA-2
- Limitations:
  - Coarse spatial resolution (0.5° × 0.625°)
- Data Access:
  - Free & open
  - NASA Earthdata (GES DISC)
    - e.g., online analysis with Giovanni

Looking Ahead to Part 3

In Part 3, we will present a case study comparing SatPM$_{2.5}$ and Bias-Corrected MERRA-2 PM$_{2.5}$ to ground monitor data, demonstrating how to access & evaluate these data in practical applications.



Preparing for Part 3

To prepare for part 3, please ensure that:

- You have access to Google Colab, using a free Google or Gmail account.
- You have a free NASA Earthdata account and know your username and password.
- You have a free OpenAQ account and know your OpenAQ API Key.

Homework and Certificates

- **Homework:**
  - One homework assignment
  - Opens on July 22, 2026
  - Access from the training webpage
  - Answers must be submitted via Google Forms
  - **Due by August 5, 2026**

- **Certificate of Completion:**
  - Attend all three live webinars (attendance is recorded automatically)
  - Complete the homework assignment by the deadline
  - You will receive a certificate via email approximately two months after completion of the course.

Contact Information

Trainers:

- Carl Malings
  - carl.a.malings@nasa.gov
- Pawan Gupta
  - pawan.gupta@nasa.gov
- Aaron van Donkelaar
  - aaron.vandonkelaar@wustl.edu
- Junhyeon Seo
  - junhyeon.seo@nasa.gov

- ARSET Website
- ARSET YouTube

For questions, comments, or to share how you have applied our trainings to your work or studies, email nasa.arset@gmail.com.

Join our quarterly newsletter to stay up-to-date on our latest trainings:

1. Send an email with no subject line to arset-join@lists.nasa.gov.
2. Follow the instructions sent in response.

References

- Martin, R. V., M. Brauer, A. van Donkelaar, G. Shaddick, U. Narain, S. Dey (2019). "No one knows which city has the highest concentration of fine particulate matter", *Atmos. Environ.*, doi:10.1016/j.aeaoa.2019.100040.
- van Donkelaar, A., R. V. Martin, R. J. D. Spurr and R. T. Burnett (2015). "High-resolution satellite-derived $PM_{2.5}$ from optimal estimation and geographically weighted regression over North America", *Environ. Sci. and Tech.*, doi:10.1021/acs.est.5b02076.
- Shen, S., Li, C. van Donkelaar, A. Jacobs, N. Wang, C. Martin, R. V. (2024). "Enhancing Global Estimation of Fine Particulate Matter Concentrations by Including Geophysical a Priori Information in Deep Learning", *ACS ES&T Air*, doi:10.1021/acsestair.3c00054.
- Shen, S., van Donkelaar, A., Jacobs, N., Li, C., and Martin, R. V. (2026). "Enhancing Estimation of Fine Particulate Matter Chemical Composition across North America by Including Geophysical A Priori Information in Deep Learning with Uncertainty Quantification", *ACS ES&T Air*, doi:10.1021/acsestair.5c0025.
- Hammer, M. S., A. van Donkelaar, C. Li, A. Lyapustin, A. M. Sayer, N. C. Hsu, R. C., Levy, M. Garay, O. Kalashnikova, R. A. Kahn, M. Brauer, J. S. Apte, D. K. Henze, L. Zhang, Q. Zhang, B. Ford, J. R. Pierce, R. V. Martin (2020). "Global Estimates and Long-Term Trends of Fine Particulate Matter Concentrations (1998-2018)", *Environ. Sci. Technol.*, doi:10.1021/acs.est.oco1764.
- van Donkelaar, A., Hammer, M. S., Bindle, L., Brauer, M., Brook, J. R., Garay, M. J., Hsu, N. C., Kalashnikova, O. V., Kahn, R. A., Lee, C., Levy, R. C., Lyapustin, A., Sayer, A. M., and Martin, R. V. (2021). "Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty", *Environ. Sci. Technol.*, doi:10.1021/acs.est.1c05309.
- van Donkelaar, A., R.V. Martin, and R.J. Park (2006). "Estimating ground-level $PM_{2.5}$ using aerosol optical depth determined from satellite remote sensing", *J. Geophys. Res.*, 111, D21201, doi:10.1029/2005JD006996.
- Hoff & Christopher (2009). "Remote Sensing of Particulate Pollution from Space: Have we reached the promised land?" *Journal of the Air & Waste Management Association*; 59, 642-675, doi:10.3155/1047-3289.59.6.645.

References

- Gupta, P., and A. Sayeed (2026). "MERRA2_CNN_HAQAST_PM25: Hourly Bias-Corrected PM2.5 Datasets for Global Air Quality Assessment." Geoscience Data Journal13, no. 2: e70070. doi:10.1002/gdj3.70070.
- Di, Q., et al. (2019). An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution. *Environment International, 130, 104909.* doi:10.1016/j.envint.2019.104909
- Wei, J., et al. (2023). First close insight into global daily gapless 1 km $PM_{2.5}$ pollution, variability, and health impact. *Nature Communications, 14, 8195. [GHAP]* doi:10.1038/s41467-023-43862-3
- Park, S., et al. (2020). Estimation of spatially continuous daytime particulate matter concentrations under all sky conditions through the synergistic use of satellite-based AOD and numerical models. *Science of the Total Environment, 713, 136516.* doi:10.1016/j.scitotenv.2020.136516
- Pendergrass, D. C., et al. (2022). Continuous mapping of fine particulate matter ($PM_{2.5}$) air quality in East Asia at daily 6 × 6 km^2 resolution by application of a random forest algorithm to 2011–2019 GOCI geostationary satellite data. *Atmospheric Measurement Techniques, 15, 1075–1091.* doi:10.5194/amt-15-1075-2022
- Sayeed, A., et al. (2025). GOES-R $PM_{2.5}$ Evaluation and Bias Correction: A Deep Learning Approach. *Earth and Space Science, 12, e2024EA004012.* doi:10.1029/2024EA004012
- Zhang, H. & Kondragunta, S. (2024). Hour by Hour $PM_{2.5}$ Mapping Using Geostationary Satellites. *ACS ES&T Air.* doi:10.1021/acsestair.4c00365
- Sayeed, A., et al. (2021). Bias correcting and extending the PM forecast by CMAQ up to 7 days using deep convolutional neural networks. *Atmospheric Environment, 253, 118376.* doi:10.1016/j.atmosenv.2021.118376

References

- Lee, S., et al. (2022). Air Quality Forecasts Improved by Combining Data Assimilation and Machine Learning With Satellite AOD. *Geophysical Research Letters, 49, e2021GL096066.* doi:10.1029/2021GL096066
- Seo, J. & Gupta, P. (2026). Adaptive expert-guided deep imbalanced regression for global $PM_{2.5}$ forecasting with temporal convolutional networks and GEOS-FP inputs. *npj Climate and Atmospheric Science.* doi:10.1038/s41612-026-01437-1

Thank You!