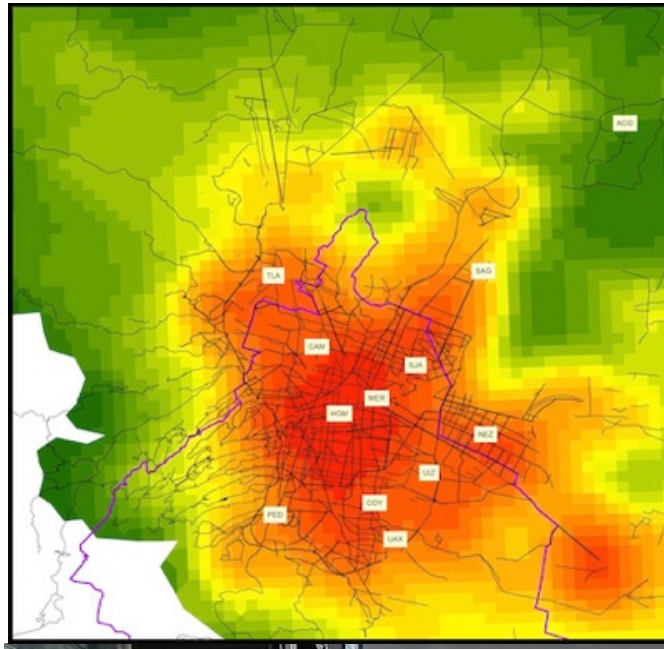
Leveraging MODIS for epidemiologic studies of fine particulate air pollution and health

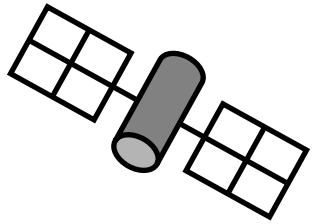


Allan C. Just PhD

Assistant Professor

Department of Environmental Medicine and Public Health

Icahn School of Medicine at Mount Sinai

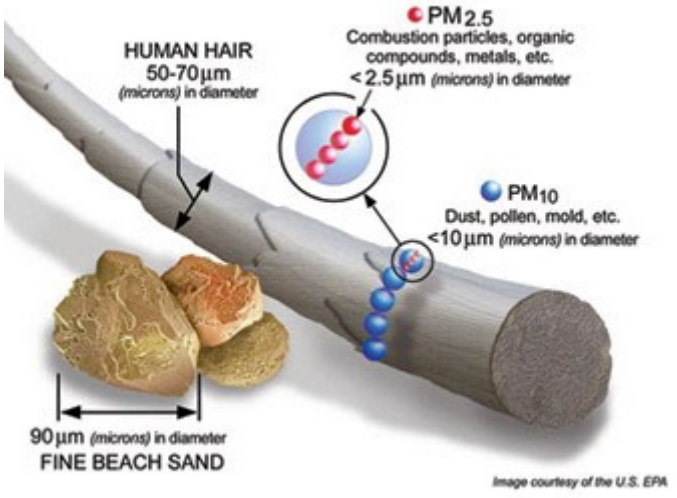


Outline

- How we use MODIS AOD to estimate $PM_{2.5}$
- Machine learning for refining AOD
- Epidemiologic approaches using satellite data
- Future directions

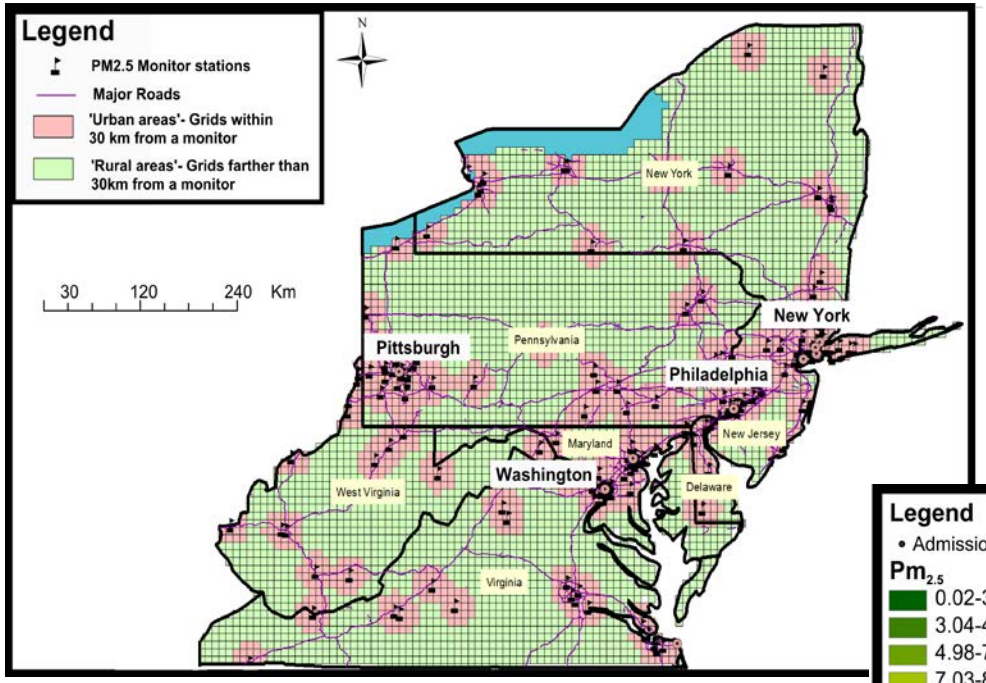
Fine particulate air pollution ($PM_{2.5}$) and cardiovascular-cerebrovascular mortality

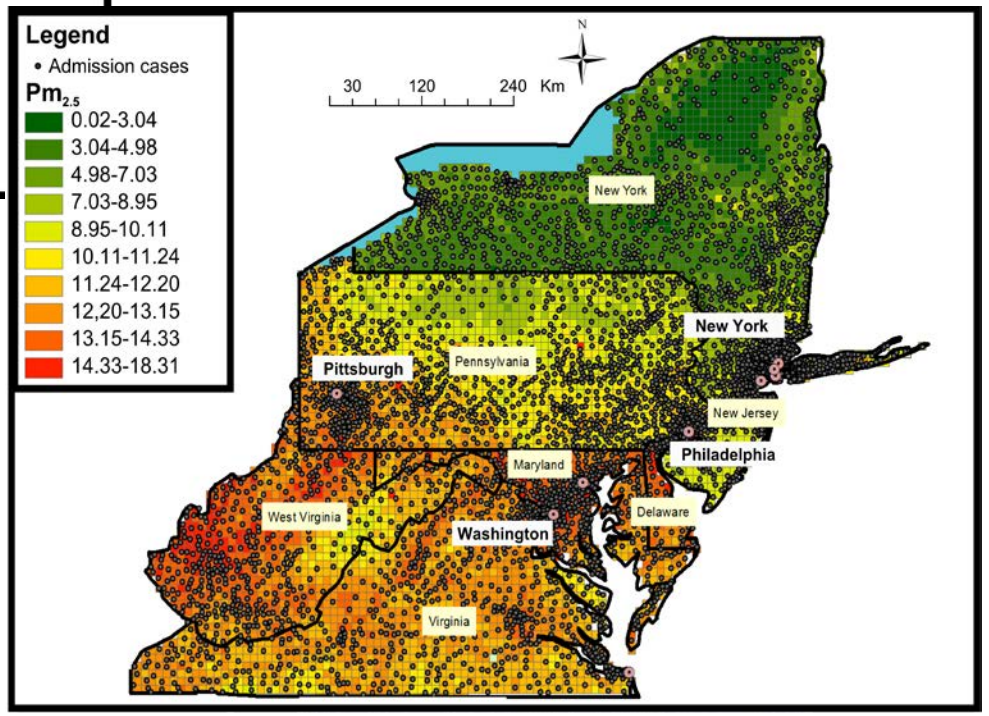
- Air pollution: main environmental risk factor for health
- ~3.7 million deaths in 2012 from outdoor air pollution globally
  - 80% due to cardiovascular and cerebrovascular diseases
  - Majority of impacts in low and middle income countries
- $PM_{2.5}$ is the largest contributor



(WHO 2014)

Beyond nearest monitor exposure assessment approaches



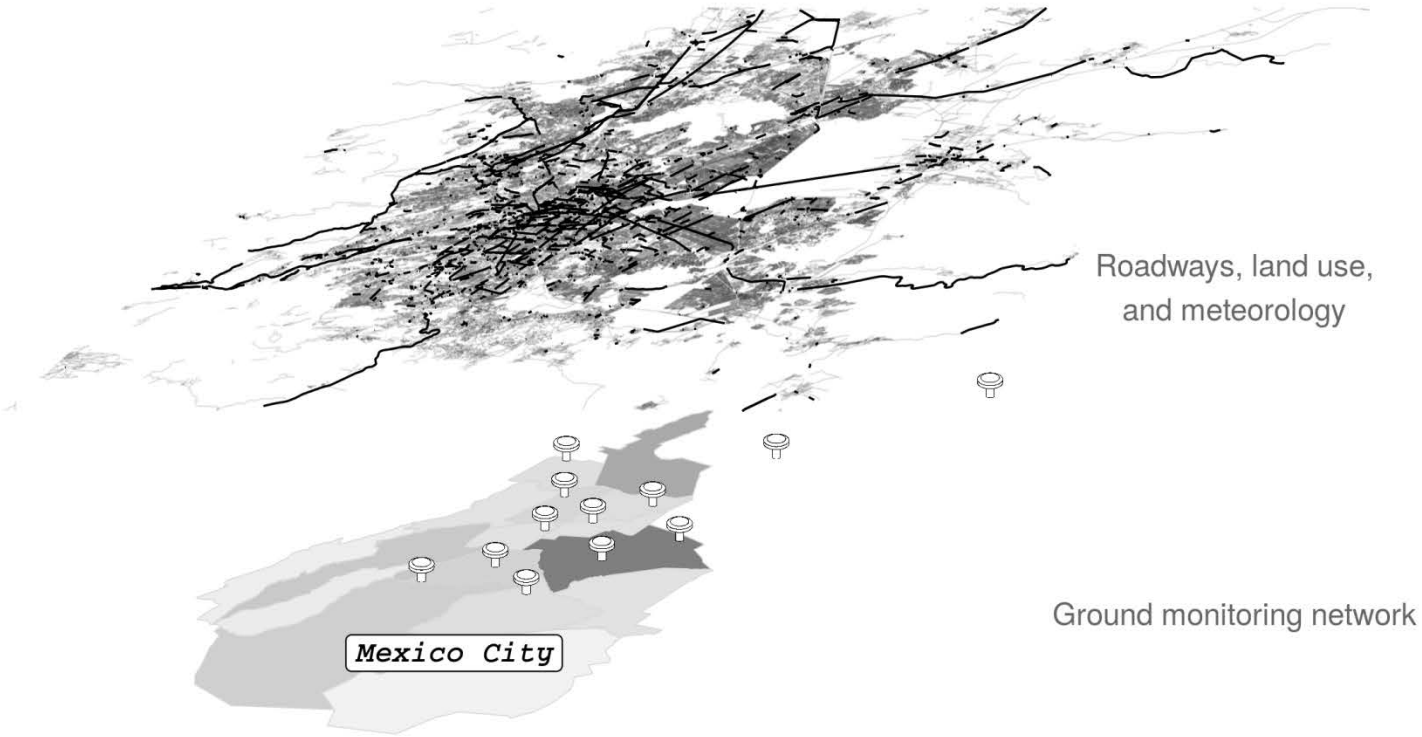


Kloog et al 2014

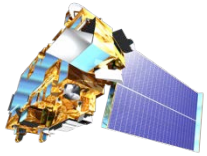
Layering information to estimate exposures

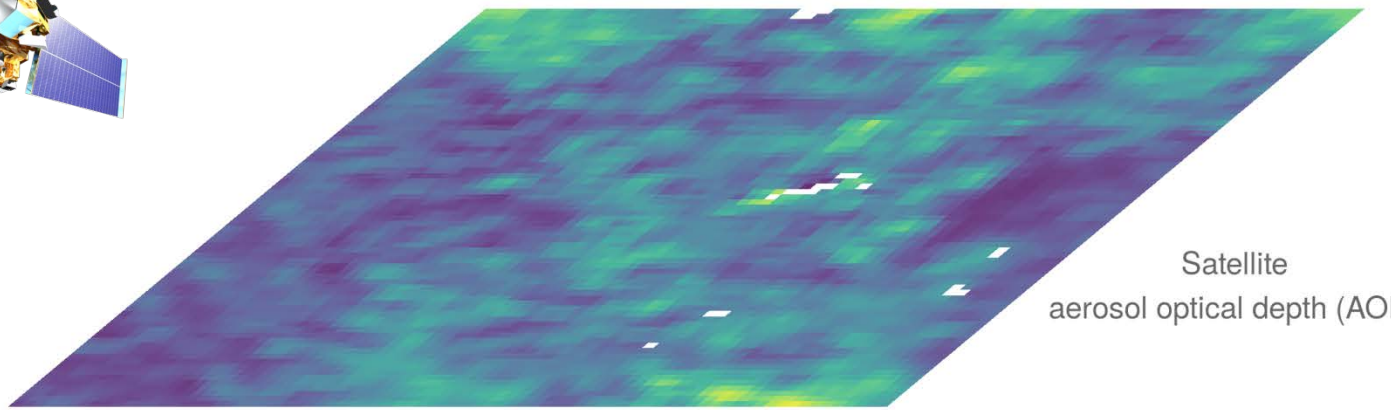
Ground monitoring network

Layering information to estimate exposures



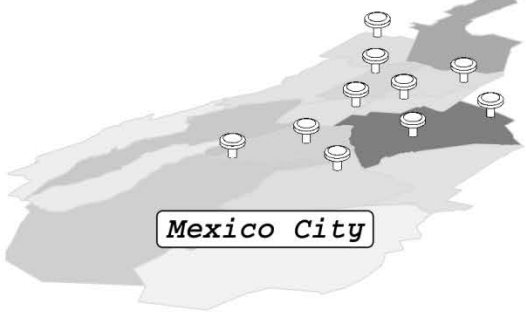
Layering information to estimate exposures





Satellite
aerosol optical depth (AOD)

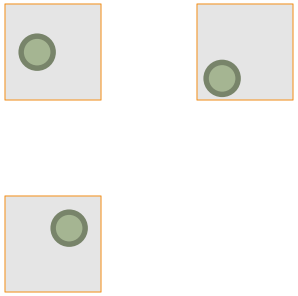
Roadways, land use,
and meteorology



Ground monitoring network

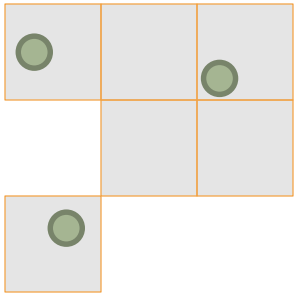
To estimate $PM_{2.5}$ concentrations in each grid cell on each day:

Fit daily calibration at monitor sites:
$PM_{2.5}$ ~ AOD + other predictors
(fit with mixed effects model)



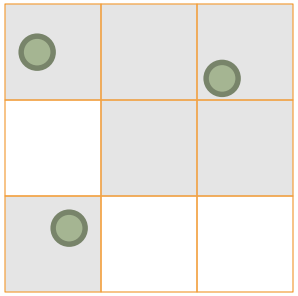
Stage 1

Use this stage 1 model to predict $PM_{2.5}$
in grid cells with AOD but without monitors



Stage 2

Estimate $PM_{2.5}$ in cells with *no available AOD data* using spatial smoothing of nearby AOD and daily regional patterns



Stage 3

Environmental Science & Technology

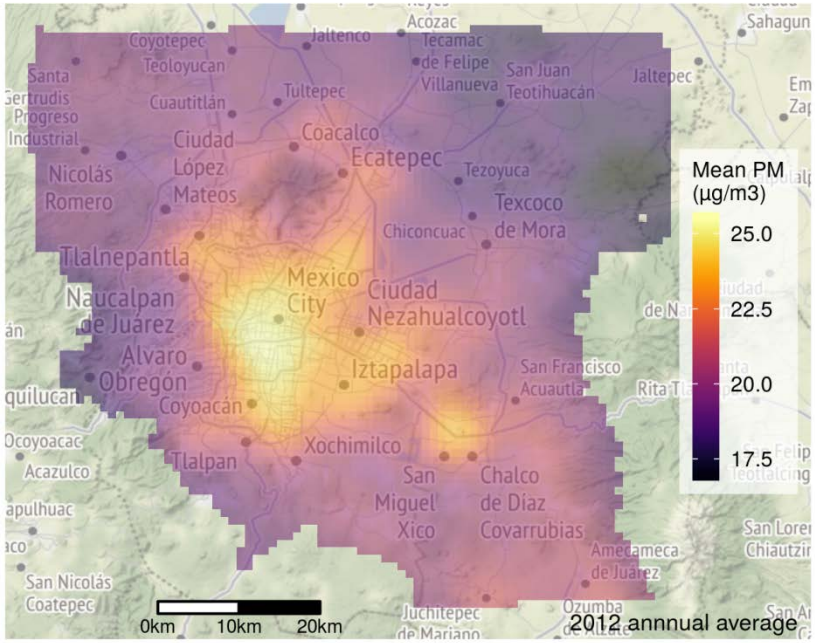
Using High-Resolution Satellite Aerosol Optical Depth To Estimate Daily $PM_{2.5}$ Geographical Distribution in Mexico City

Allan C. Just,*,† Robert O. Wright,‡ Joel Schwartz,† Brent A. Coull,§ Andrea A. Baccarelli,† Martha María Tellez-Rojo,∥ Emily Moody,⊥ Yujie Wang,# Alexei Lyapustin,∇ and Itai Kloog¶

Key features

- Daily estimates
- On a 1km * 1km grid
- Spanning 2004-2014
- Cross-validated R^2 of 0.72

Just et al. *Environ Sci Technol*. 2015

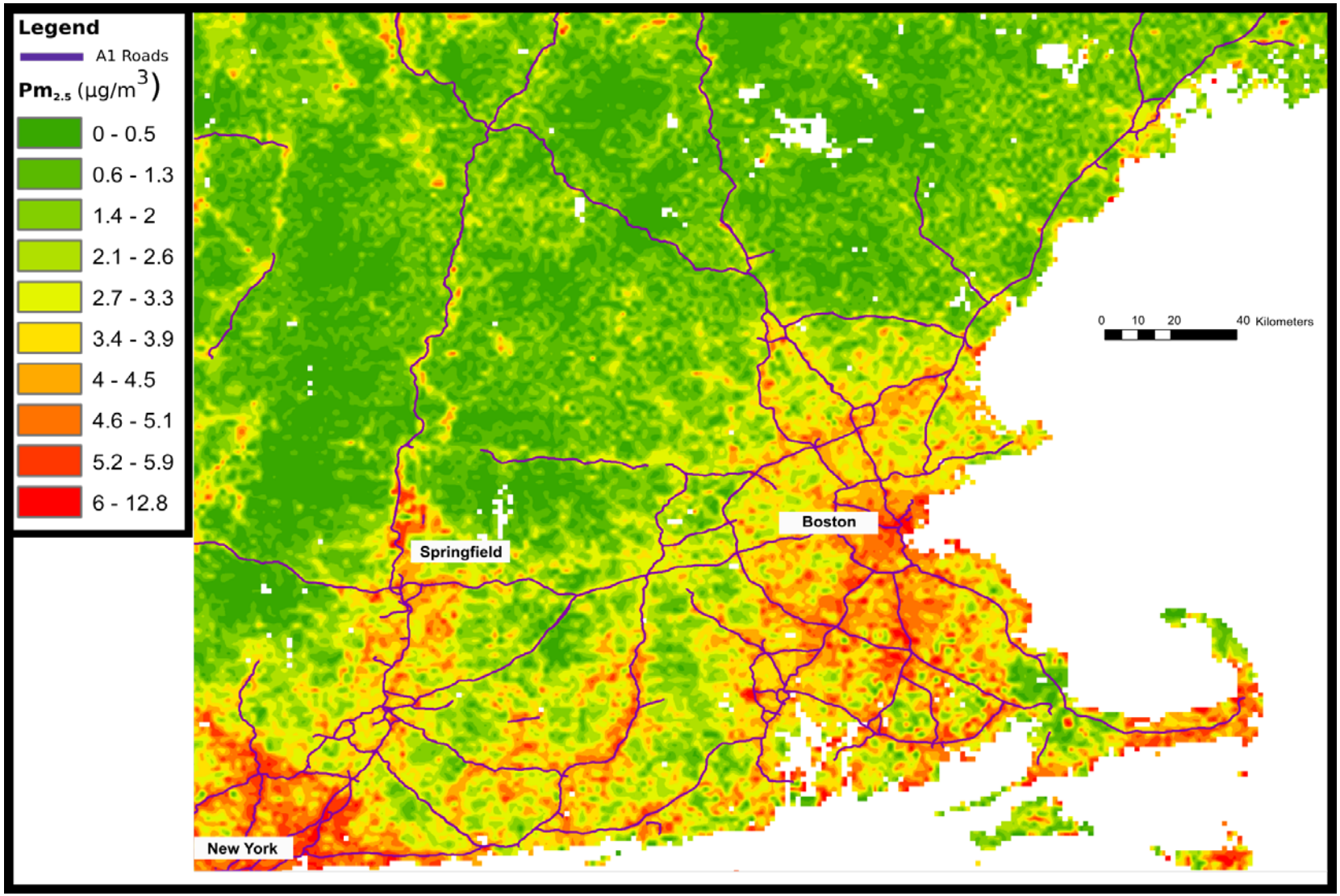


Cardiovascular and Cerebrovascular Mortality Associated With Acute Exposure to $PM_{2.5}$ in Mexico City

Stroke

Iván Gutiérrez-Avila, MSc; Leonora Rojas-Bracho, ScD; Horacio Riojas-Rodríguez, PhD; Itai Kloog, PhD; Allan C. Just, PhD; Stephen J. Rothenberg, PhD

10 ug/m^3 higher $PM_{2.5}$ in lag 0-1 days associated with 3.43% (0.10-6.28) higher cerebrovascular mortality



Mean $PM_{2.5}$ concentrations in each 1x1 km grid cell in the Boston region during 2003 predicted by the hybrid satellite-LUR models

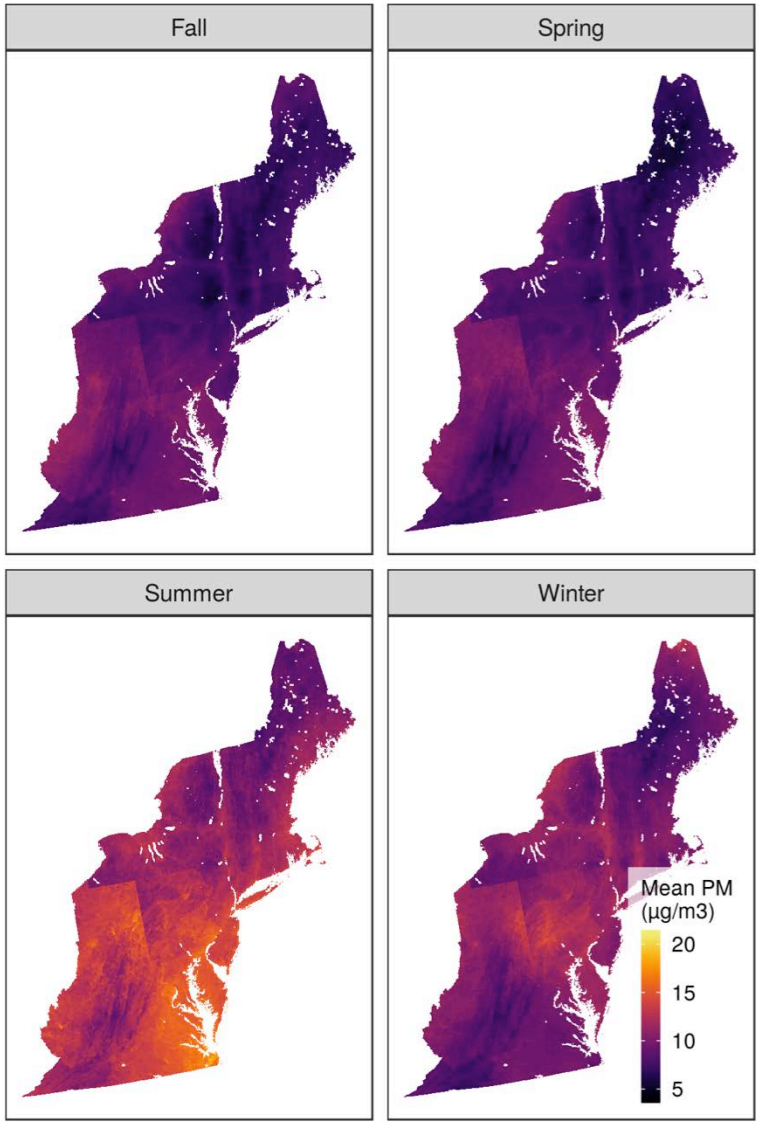
Kloog I, Chudnovsky AA, **Just AC**, Nordio F, Koutrakis P, Coull BA, Lyapustin A, Wang Y, Schwartz J. A new hybrid spatio-temporal model for estimating daily multi-year PM2.5 concentrations across northeastern USA using high resolution aerosol optical depth data. *Atmospheric Environment*. 2014;95(0):581-90.

Updating our New England & Mid-Atlantic $PM_{2.5}$ Model

- ~3 billion point-day predictions

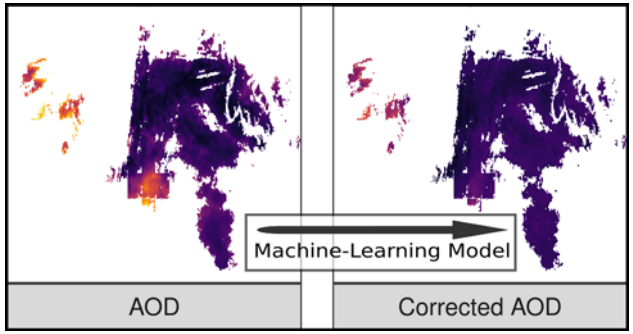
(1 x 1km; 600,000 per day; 2000-2017)

- Using Machine-Learning algorithms to make our approaches scalable
- When AOD is available:
  - **Mixing layer height**, **Terra AOD**, and **Aqua AOD** the most important variables
  - But AOD is usually missing; GOES-16 can help fill in



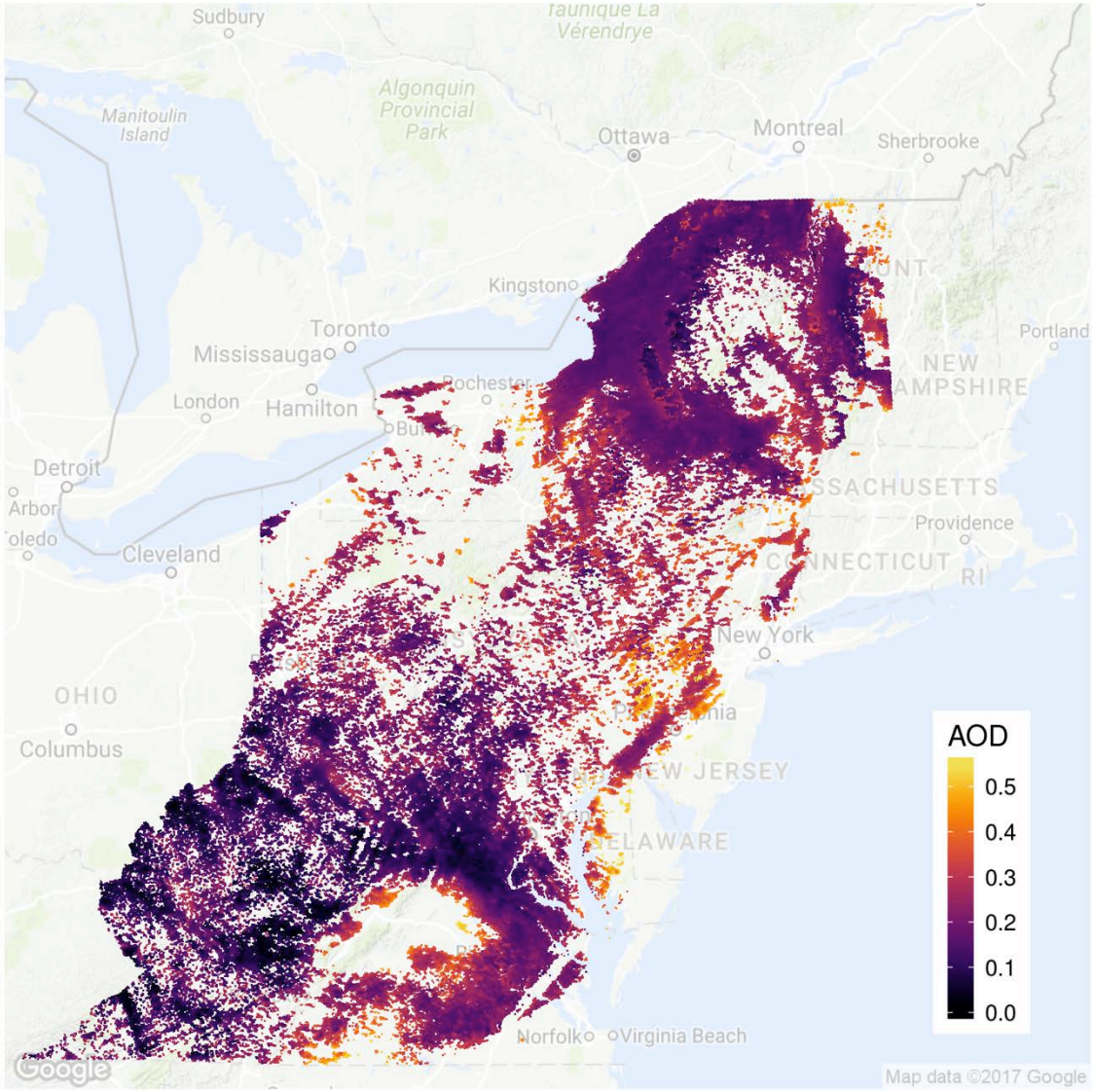
2008 Seasonal average $PM_{2.5}$

remote sensing



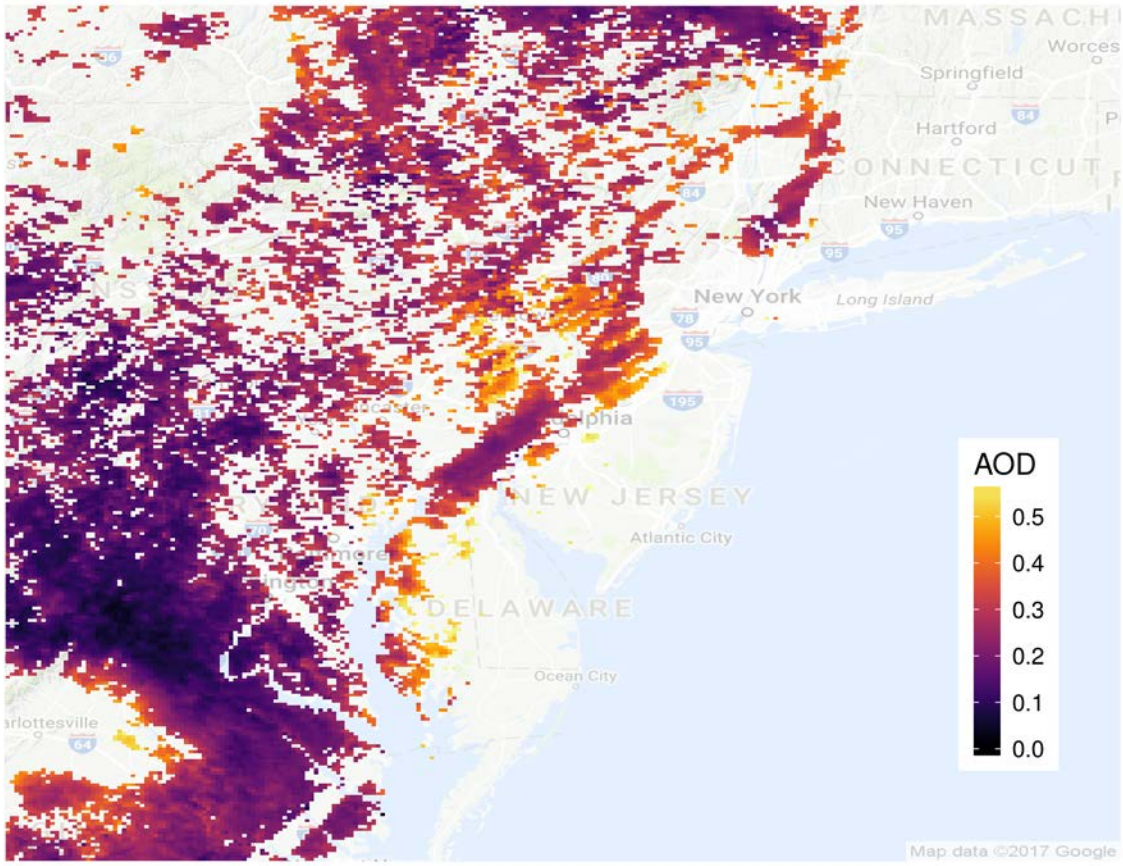
Remote Sens. 2018, 10(5), 803;
https://doi.org/10.3390/rs10050803

Observation: MAIAC AOD retrieval has spatial patterns



Complex retrieval algorithms have to account for surface brightness and atmospheric physics

Potential for spatial artifacts (due to cloud edges, bright surfaces, etc)



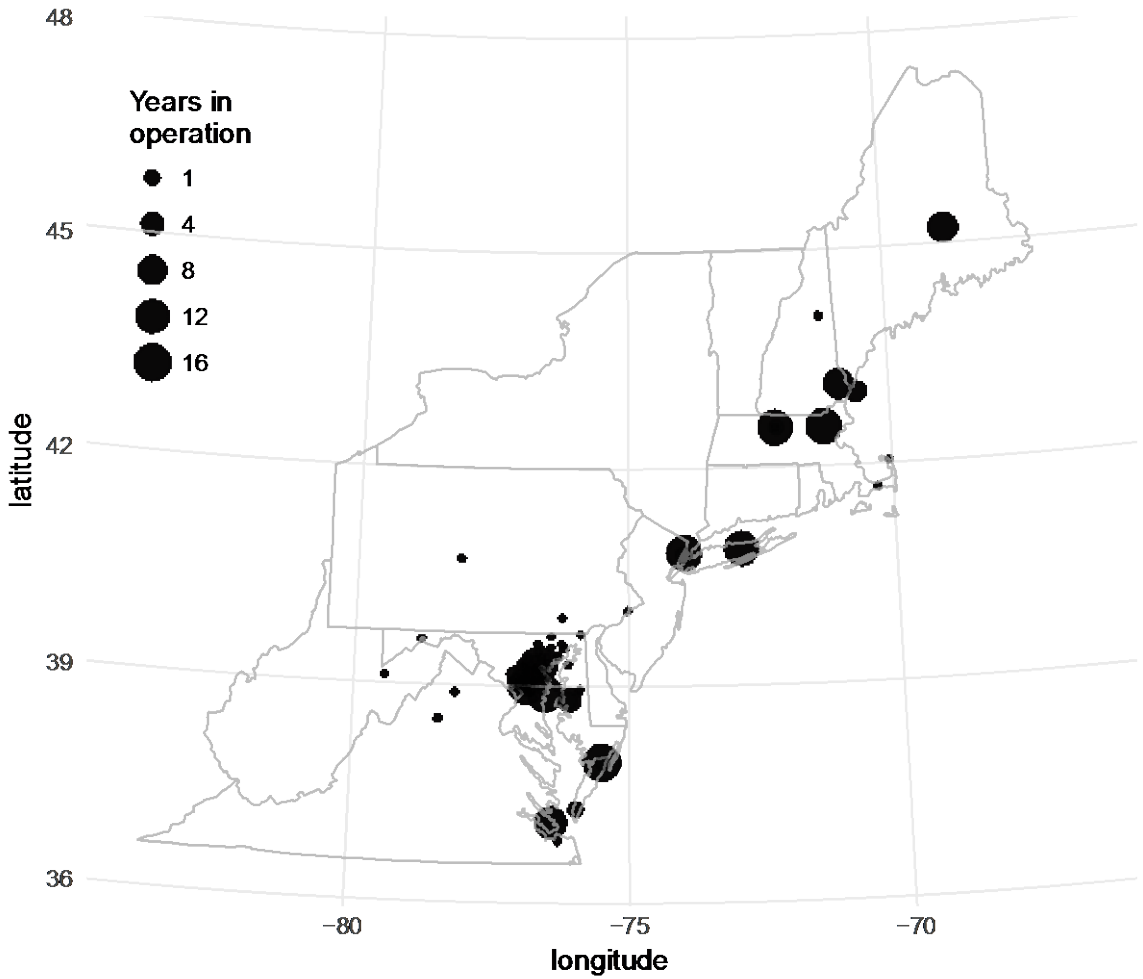
Goal: Compare and correct satellite retrieval from the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm with ground-based Aerosol Robotic Network (AERONET) quality-controlled measures

- Careful calibration and QA/QC (e.g. cloud screening)
- Direct measure (points at the sun, not impacted by surface brightness)
- But – a point measurement, not a regional measure

Closest thing we have to a "Gold Standard" – and commonly used as such

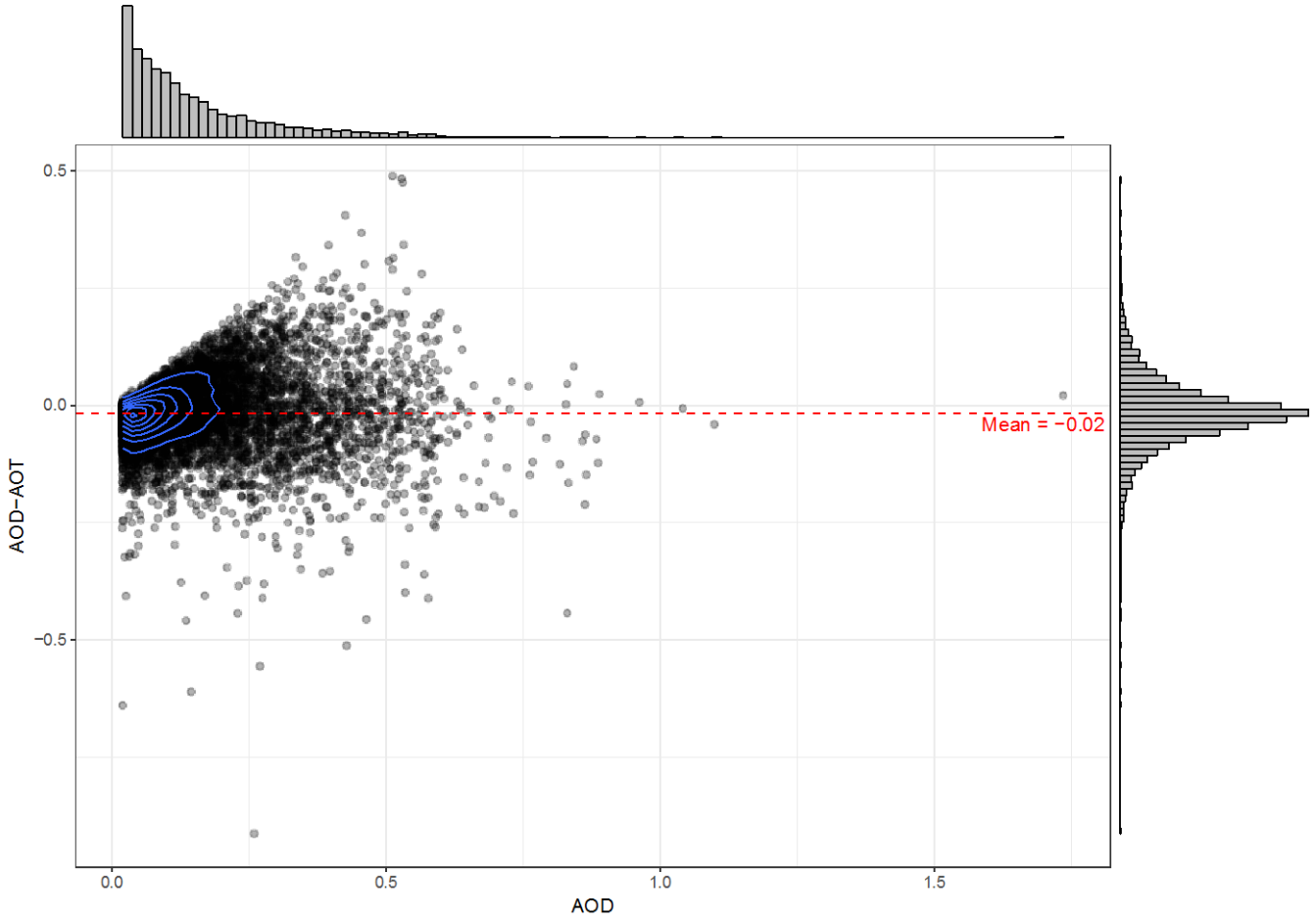
https://www.nasa.gov/centers/langley/science/aeronet.html

Study region – 13 States in the Northeastern and Mid-Atlantic USA



Study region in Northeastern and Mid-Atlantic USA with 79 unique AERONET stations showing the years of coverage for use in measurement error modeling.

Define our target parameter as the difference between MAIAC AOD and AERONET AOT (goal is zero)



MAIAC AOD versus (AOD minus AERONET AOT) in collocated observations in the Aqua measurement error dataset (n=8,531). Since both AOD and AOT are strictly positive, the apparent lack of points in the upper left of the Bland-Altman plot is expected. Marginal histograms show AOD is skewed but the difference of AOD - AOT which is an estimate of measurement error is more normally distributed.

Machine-Learning Methods used with MAIAC and AERONET

Three Tree-based Models

Bagging (aggregate across independent bootstraps)

Random Forest (RF) – parameters mtry (1/3 of k) and ntrees (10,000)

Boosting (iteratively learn by fitting trees on residual error)

Gradient Boosting Machine (GBM) – 10k trees, interaction depth of 6, learning rate 0.002

Extreme Gradient Boosting (XGBoost) – 10k trees, depth of 5, learning rate 0.01, subsample .5

Training/Testing

Because AOD-to-ground relationship varies daily; split out testing dataset by withholding randomly selected days (~15% of all observations)

Hyper-parameter tuning with cross-validation within training dataset (minimize RMSE)

Feature Engineering: How do we capture what might describe retrieval biases?

(aod - aot) ~ temperature + evap + planetary boundary layer height + surface pressure + precipitable water + specific humidity + u wind at 10m + v wind at 10m + visibility + elevation + proportion water in 1km + proportion forest in 1km + AOT uncertainty + column water vapor + relative azimuth angle + cloud mask + adjacency mask + aerosol model + distance to an edge + n non-missing (3km window) + std deviation (3km window) + *21 variables* for mean, difference, and n non-missing across 7 moving windows (30, 50, 110, 210, 310, 410, 510 km side lengths) + percentile of political region + percentile of eco region + mean of political region + mean of eco region + difference from political region + difference from eco region + size of cluster + mean of cluster + integer date + two month season indicator

The color coding of the variables indicates their meaning/origin:

AERONET aerosol optical thickness

MAIAC variables

meteorological and land use variables

distance to an edge

focal variables

regional variables

cluster variables

temporal variables

Comparative performance in testing set

Table 2. Performance predicting AOD-AOT on a test set

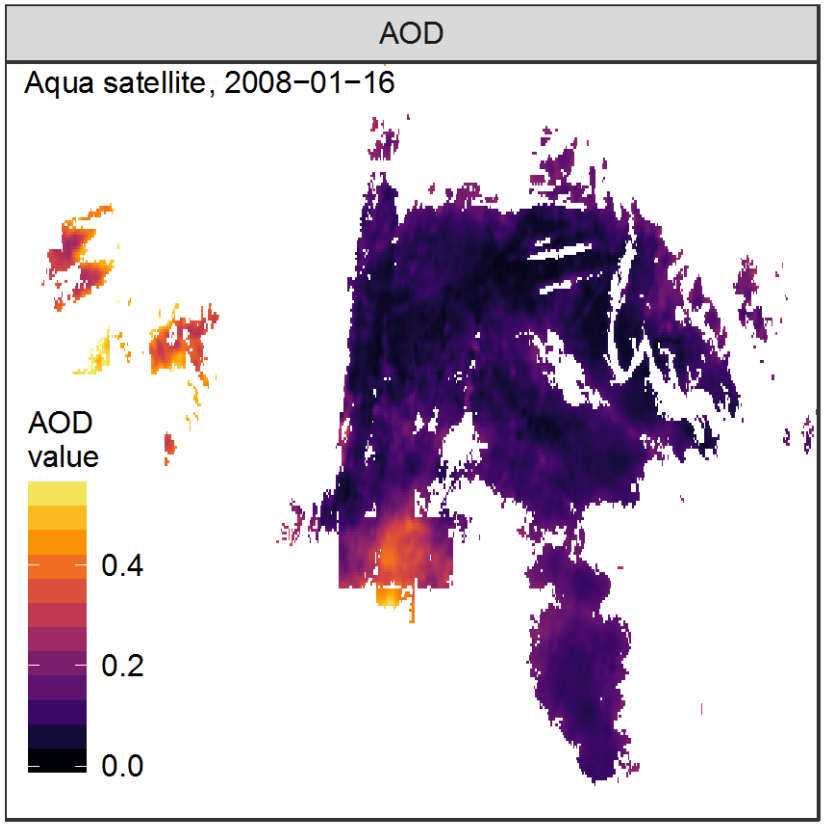
Model	Aqua (n=1,251)		Terra (n=1,478)	
	RMSPE	**R^2**	**RMSPE**	**R^2**
Raw data (aod vs aot)*	0.074	--	0.079	--
RF	0.047	0.59	0.049	0.62
GBM	0.044	0.64	0.047	0.65
XGBoost	0.042	0.67	0.044	0.68

*Raw data reports the comparable root mean square difference between raw AOD and AOT.

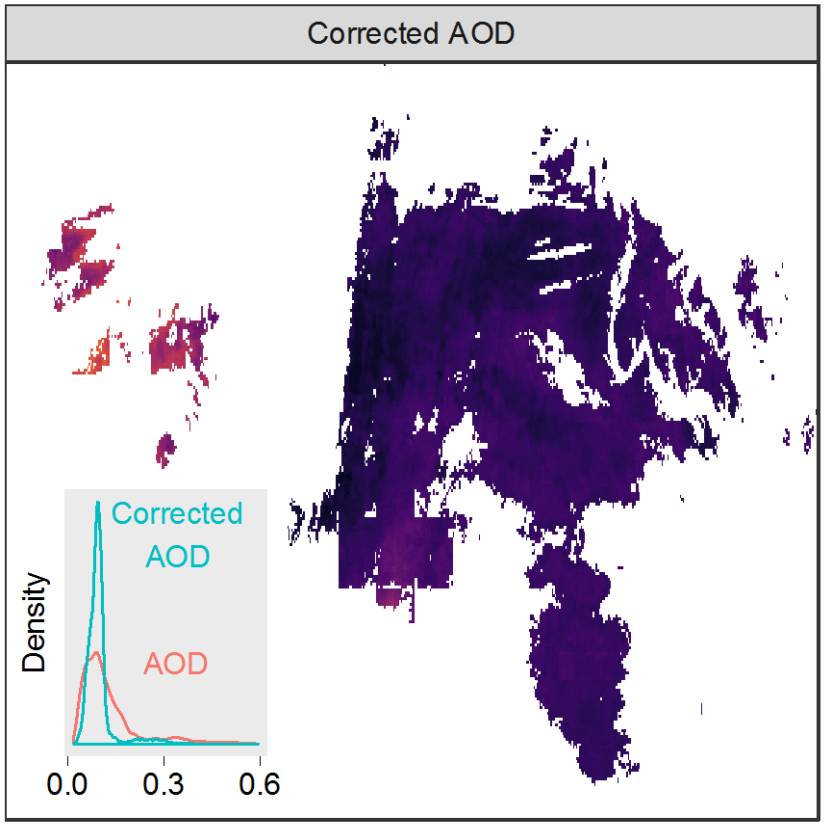
43% (Aqua) and 44% (Terra) reduction in RMSE on test data!

An example of one day (from the testing set)

Before:



After:

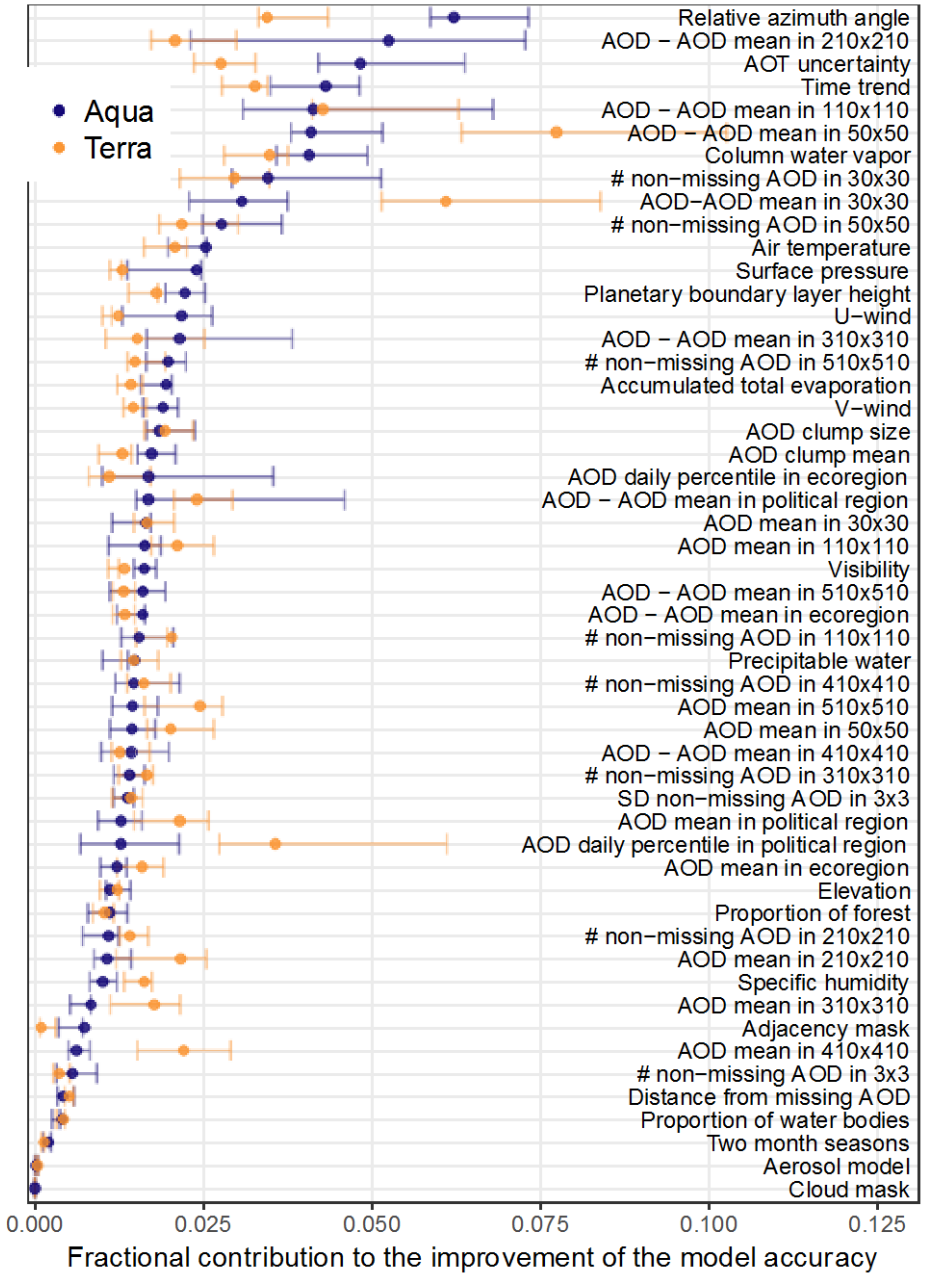


Maps of MAIAC AOD for 2008-01-16 before and after correction with our XGBoost measurement error prediction model. The inset shows the density of AOD within this scene.

Interpreting variable importance

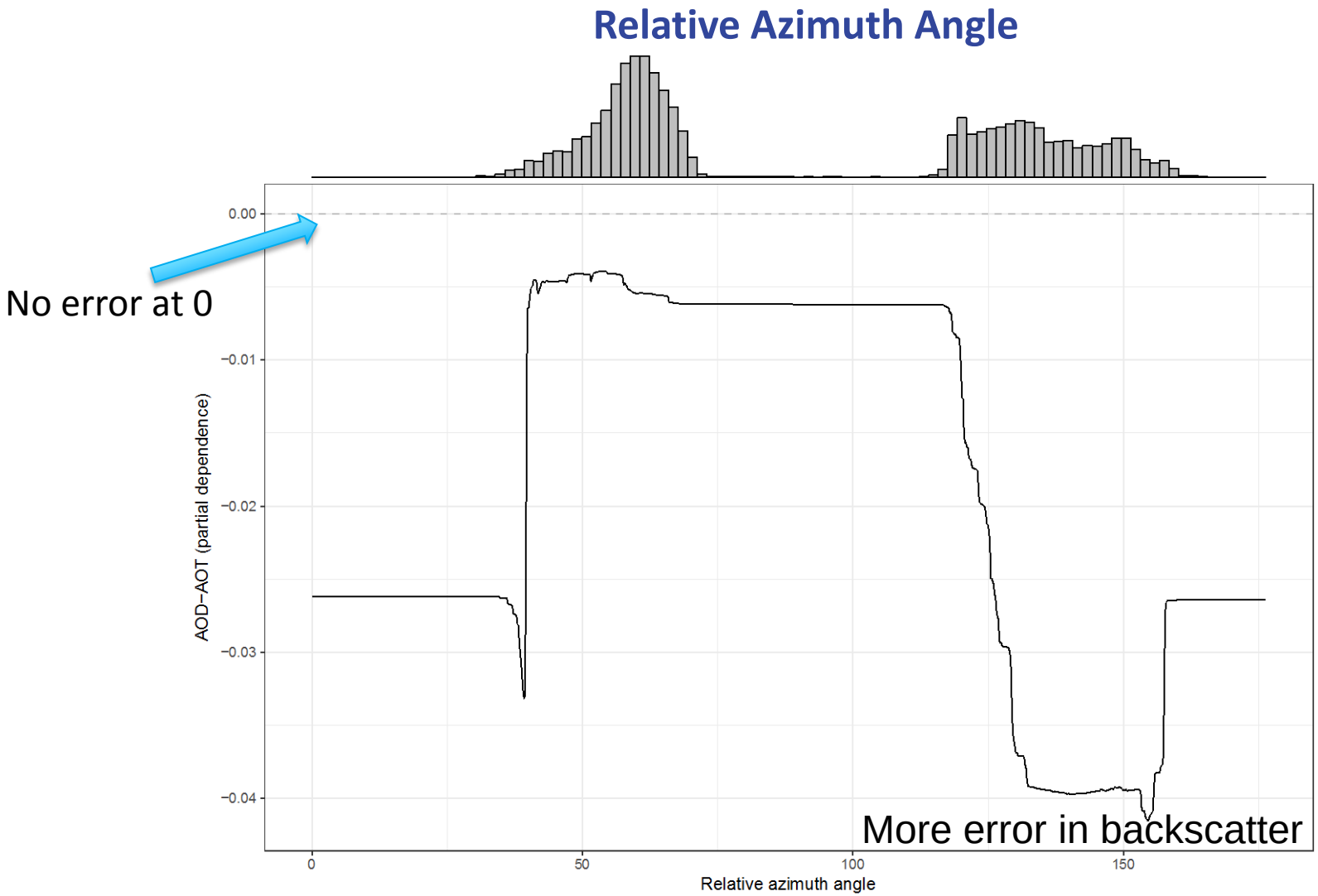
Top variables:

- relative azimuth
- AOD uncertainty (surface brightness in blue band)
- AOD difference in 30-210km moving windows



Variable importance predicting measurement error by node impurity from XGBoost for the Aqua and Terra dataset with intervals showing the range of variable importance measures across ten bootstrap-resampling fits of the training dataset.

Partial Dependence Plots: marginal relationships of individual features with predictions to understand complex relations



Partial dependence plot of measurement error as a function of relative azimuth for the Aqua training set (n=7,280) from the GBM approach. The marginal histogram shows the distribution of relative azimuth, with larger errors (further from zero) seen for the second mode with angle >120° in backscattering conditions.

What about $PM_{2.5}$ on the ground?

Does this algorithmic correction improve correlation of AOD and $PM_{2.5}$?

Yes, by 10 percentage points

Table 3. Correlations between $PM_{2.5}$ and the predicted value of AOT

Model	**Aqua** (n=105,798)	**Terra** (n=131,788)
Raw data	0.473	0.557
RF	0.548	0.633
GBM	0.567	0.645
XGB	0.572	0.649

Collocated dataset included 362 and 381 daily $PM_{2.5}$ monitors with n=105,798 and n=131,788 observations with concurrent AOD for Aqua and Terra, respectively

Just et al. *Remote Sens*. 2018, 10(5), 803; https://doi.org/10.3390/rs10050803

Epidemiologic modeling to link exposure estimates with health effects

$PM_{2.5}$ associated with:	**Respiratory disease** (Kloog 2012, *PLOS One;* Rosa 2017, *Ann Allergy Asthma Immunol*)
Hospital Admissions (Kloog 2012, *PLoS One*)	**Myocardial infarction (heart attack)** (Madrigano 2012, *Env Health Perspectives*)
Mortality (Kloog 2013, *Epidemiology*)	**Cardiovascular disease** (Kloog 2012, *PLOS One*)
Cerebrovascular mortality (Gutierrez-Avila 2018, *Stroke*)	**Reduced birth weight** (Kloog 2015, *Env Health Perspectives*; Rosa 2017 *Environ Int*)

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JUNE 29, 2017 VOL. 376 NO. 26

Air Pollution and Mortality in the Medicare Population

Qian Di, M.S., Yan Wang, M.S., Antonella Zanobetti, Ph.D., Yun Wang, Ph.D., Petros Koutrakis, Ph.D., Christine Choirat, Ph.D., Francesca Dominici, Ph.D., and Joel D. Schwartz, Ph.D.

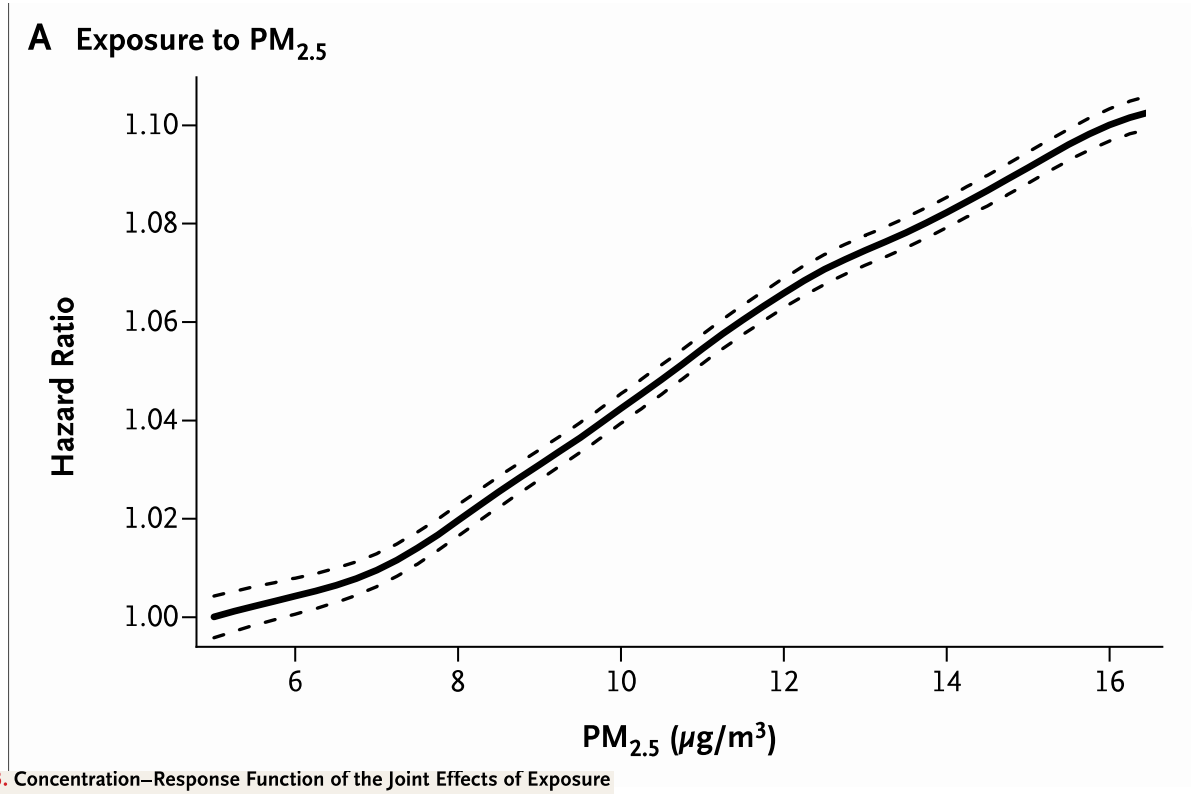


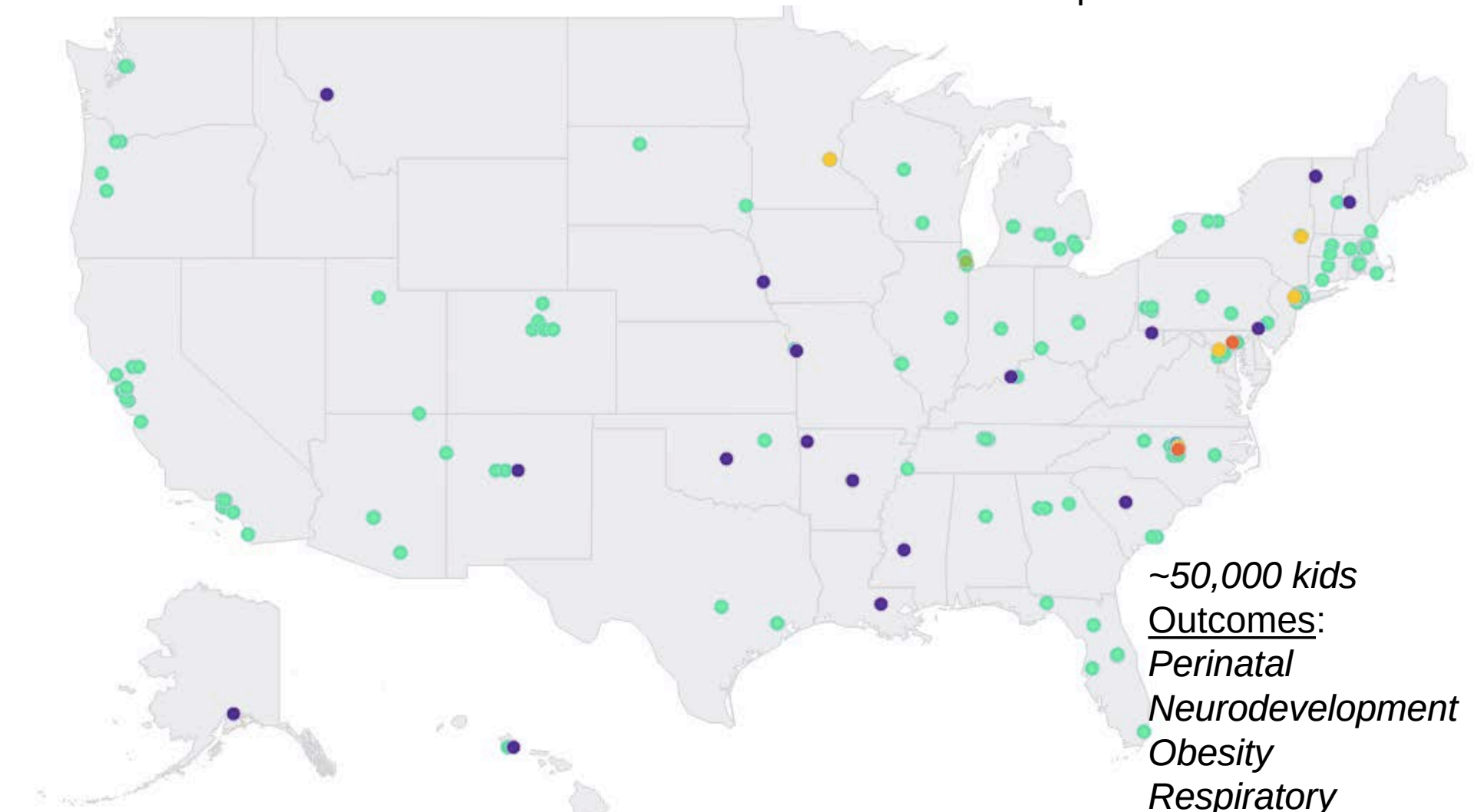
Figure 3. Concentration–Response Function of the Joint Effects of Exposure to $PM_{2.5}$ and Ozone on All-Cause Mortality.

ECHO

Environmental influences on Child Health Outcomes

A program supported by the NIH

New Grant: awarded April 2018
Opportunities and Infrastructure Fund (OIF)
"ECHO-wide platform for studying air pollution, temperature, and greenness using satellite remote sensing with daily high-resolution national exposure estimates"



~50,000 kids
Outcomes:
Perinatal
Neurodevelopment
Obesity
Respiratory

Epidemiologic modeling to link exposure estimates with health effects

Identifying relevant etiologic windows:

Temperature variability, air pollution, and birth outcomes in a changing climate: epidemiological innovation contrasting populations in Massachusetts USA and Southern Israel

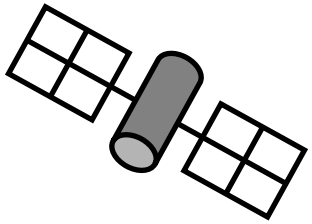
- Distributed lag modeling in >1 million birth records

Studying acute effects of temperature:

Hourly temperature dynamics from satellites and risk of cardiovascular events

- Examining >2 million cardiovascular hospitalizations in NYS in relation to extreme temperatures and temperature variation

Acknowledgments



Main Collaborators

Ben Gurion University, Israel

Itai Kloog

Icahn School of Medicine at Mount Sinai

Bob Wright

Roz Wright

Columbia Mailman School of Public Health

Andrea Baccarelli

Marianthi Kioumourtzoglou

Harvard Chan School of Public Health

Joel Schwartz

Brent Coull

Instituto Nacional de Salud Publica (INSP), Mexico

Martha Tellez-Rojo

NASA Goddard Space Flight Center

Alexei Lyapustin

Yujie Wang

Funding from the National Institutes of Health:
R00ES023450, P30ES023515, UH3OD023337,
R01ES013744, R24ES028522, U2CES026561
and the Binational Science Foundation Grant No. 2017277

allan.just@mssm.edu