Machine learning for spatiotemporal $PM_{2.5}$ estimates across the western US, 2008-2014

Melissa May Maestas, PhD
Research Associate (Post-Doctoral)
CU Boulder, CIRES, Earth Lab

Colleen E. Reid, PhD, MPH
Assistant Professor of Geography
CU Boulder

Ellen Considine
BS-candidate

Gina Li
MA-candidate

International Workshop on Air Quality Forecasting Research
November 9, 2018
Boulder, CO

Image: https://www.nasa.gov/image-feature/nasas-science-aircraft-flies-over-thomas-fire-in-california

Why wildfires?

Globally and regionally, wildfire risk is projected to increase under various potential future climate scenarios.

The percent of our air pollution due to wildfires will likely increase, not just from climatic changes, but also because of declines in other sources of air pollution

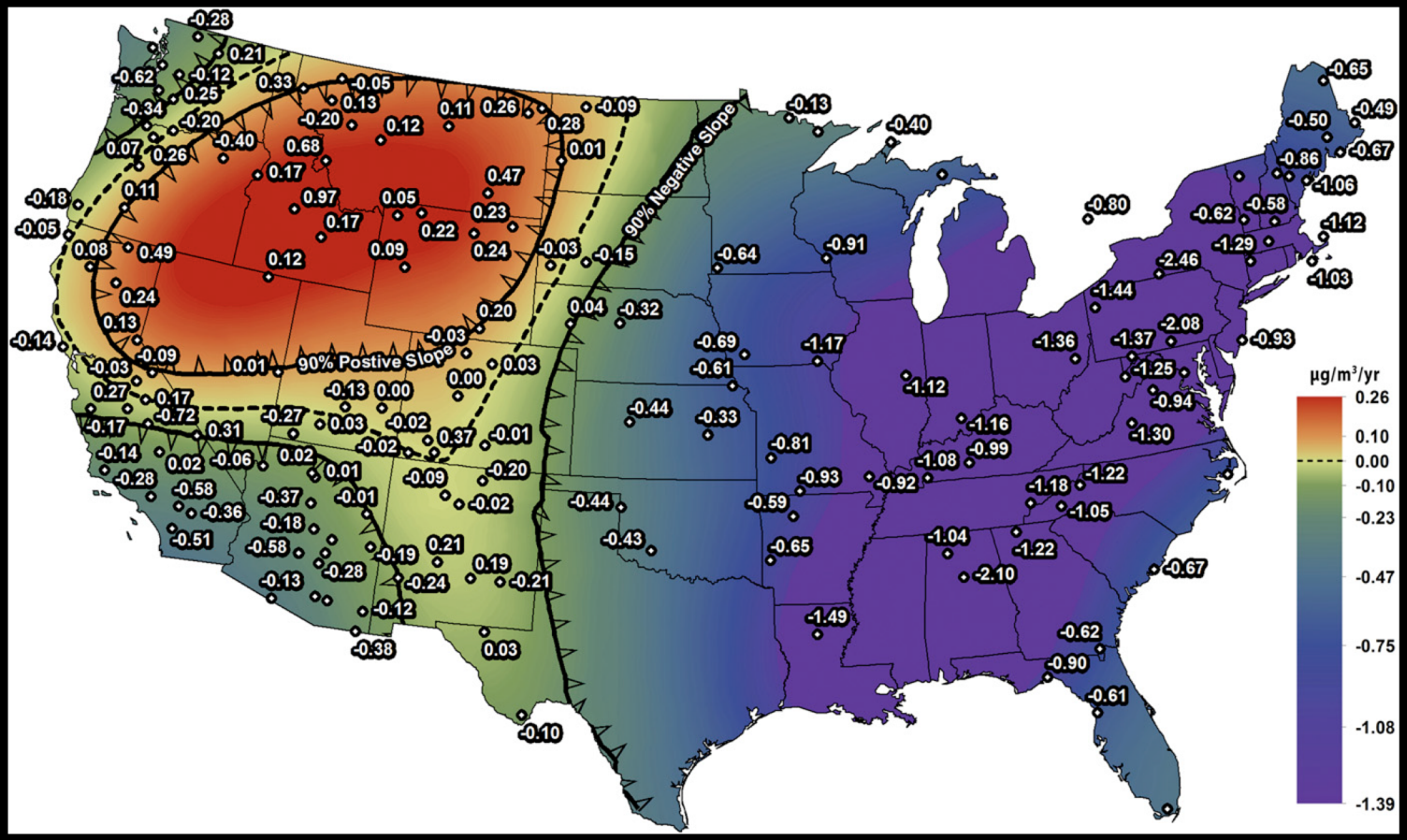


Image: McClure and Jaffe, 2018. www.pnas.org/cgi/doi/10.1073/pnas.1804353115

What are the health effects from exposure to wildfire smoke?

- Clear evidence of respiratory health effects
  - Particularly for exacerbation of asthma
- Growing evidence of increased risk of all-cause mortality
- Mixed findings for cardiovascular outcomes
- A few studies have found associations with adverse birth outcomes
- More research needed
  - Health outcomes not yet studied or understudied
  - Further evaluate cardiovascular outcomes
  - Better understand effects of duration and intensity of wildfire smoke exposures
  - Identify vulnerable subpopulations

Source: Reid et al. 2016. Critical review of health impacts of wildfire smoke exposure. *Environmental Health Perspectives*
Image: National Fire Interagency Fire Center

Exposure assessment difficulties

- Sparse air pollutant monitoring network
- Many $PM_{2.5}$ monitors only measure every sixth or third day
- Leads to spatial and temporal averaging of exposure measurements
- But, smoke plumes migrate quickly, changing exposures over smaller spatial and temporal scales

Image: National Interagency Fire Center, Fox Fire East, Glacier 2000

Previous and Current ML work

- Dr. Colleen E. Reid modeled Northern CA 2008 wildfire season for her dissertation
  - Reid et al., Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning. Environ. Sci. Technol. 2015, 49, 3887−3896
- Current Project – Scaling up:
  - Multi-year, 11 western states
  - Similar modeling (e.g., Di et al., 2016) don't perform well in western US and fires were left out
  - More than EPA monitors (Forest Service, TEOM, states, field campaigns, etc.)
    - ❋ If you have $PM_{2.5}$ monitor data – we would like to include it in our study ❋
  - Plan to do ensemble of different machine learning algorithms

Di et al. Environ. Sci. Technol. 2016, 50, 4712−4721

Methods – adapt land use regression modeling with machine learning

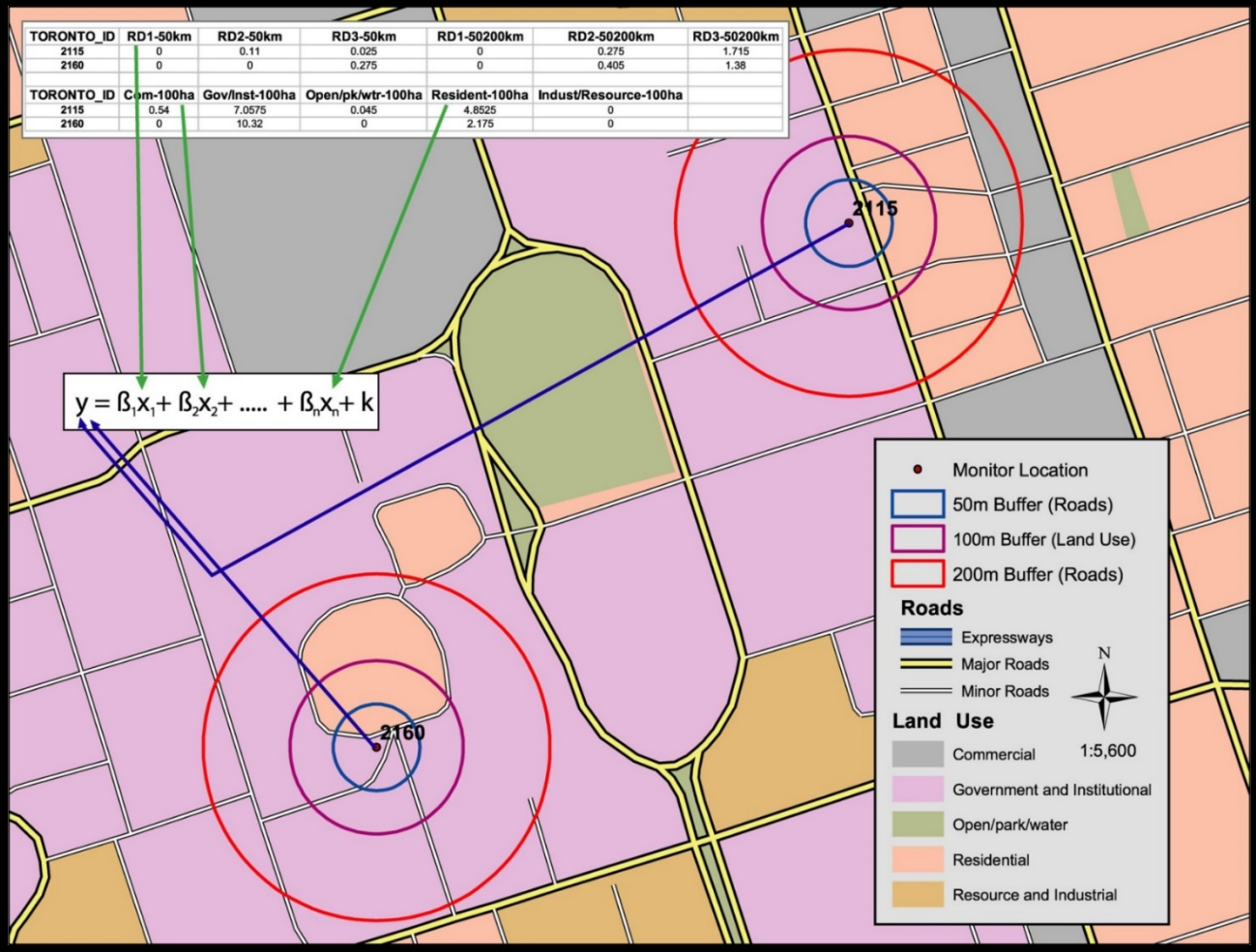
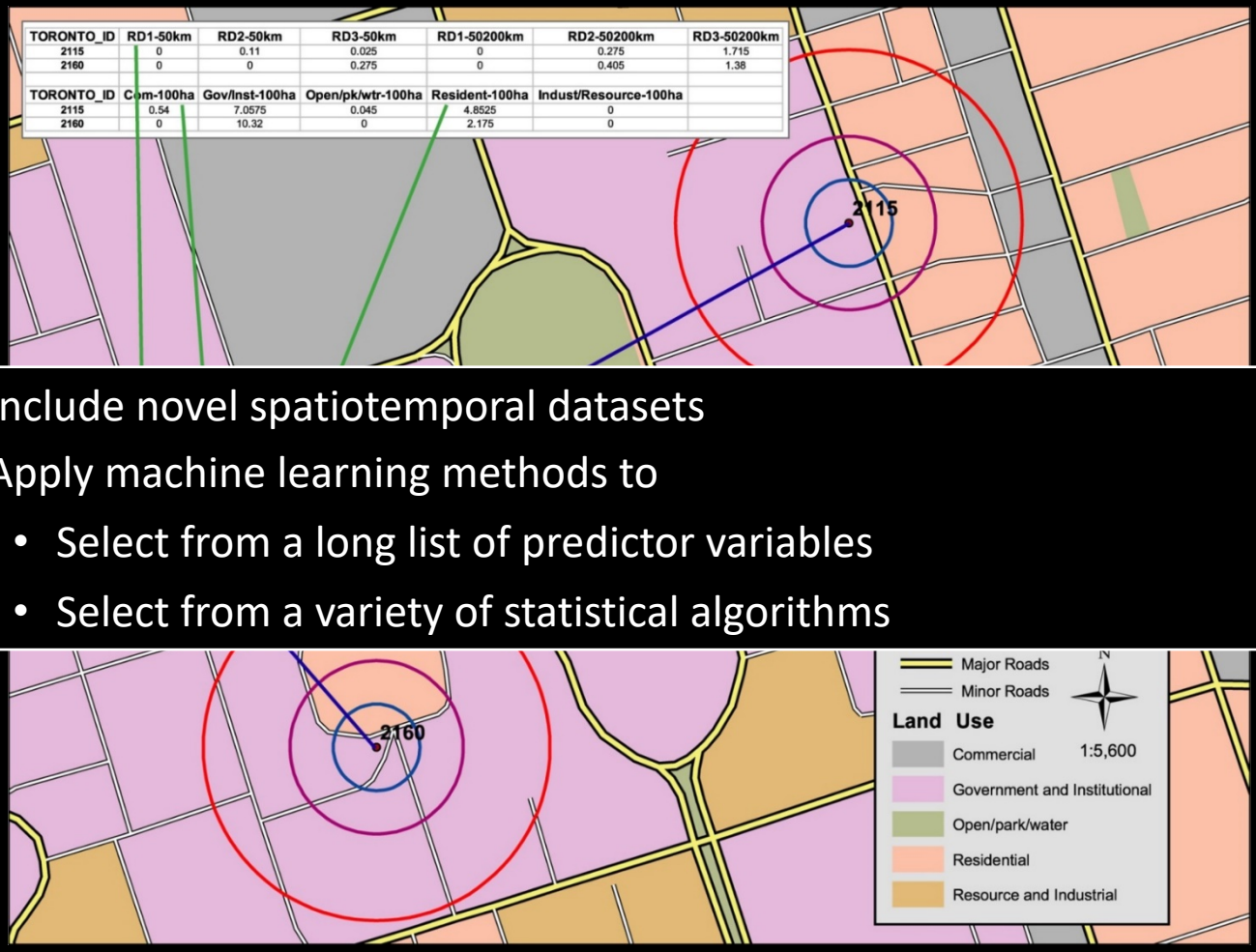


Image: courtesy of Mike Jerrett

Methods – adapt land use regression modeling with machine learning

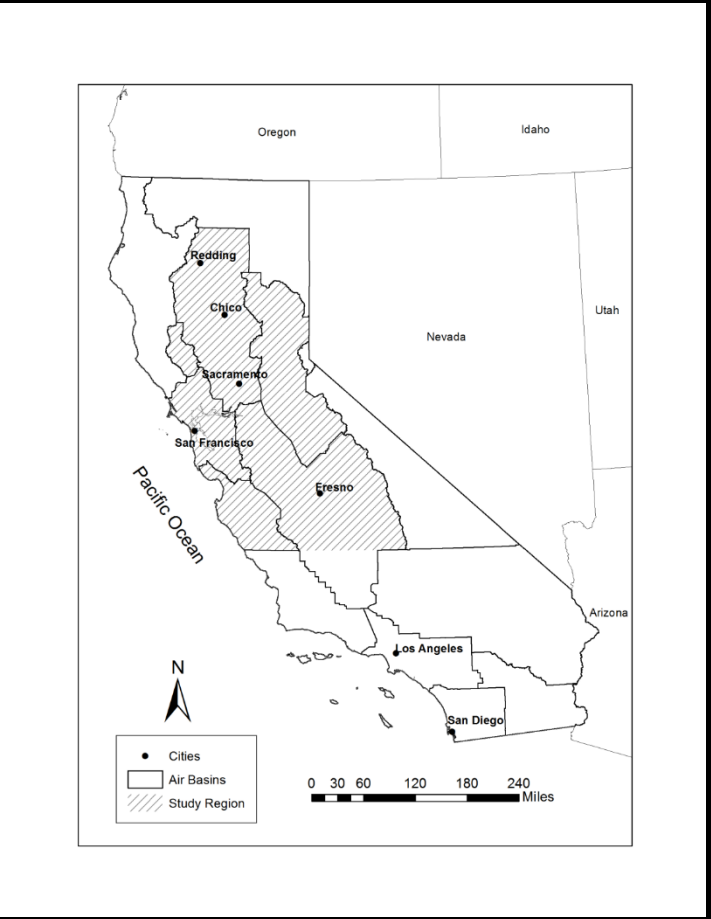


- Include novel spatiotemporal datasets
- Apply machine learning methods to
  - Select from a long list of predictor variables
  - Select from a variety of statistical algorithms

Image: courtesy of Mike Jerrett

2008 northern California wildfires

- Lightning storm on June 20-21, 2008
- Over 6000 lightning strikes
- Thousands of fires
- Smoke covered large population areas for weeks (est. 10-12 million people exposed)
- Potential for adverse health effects on large population



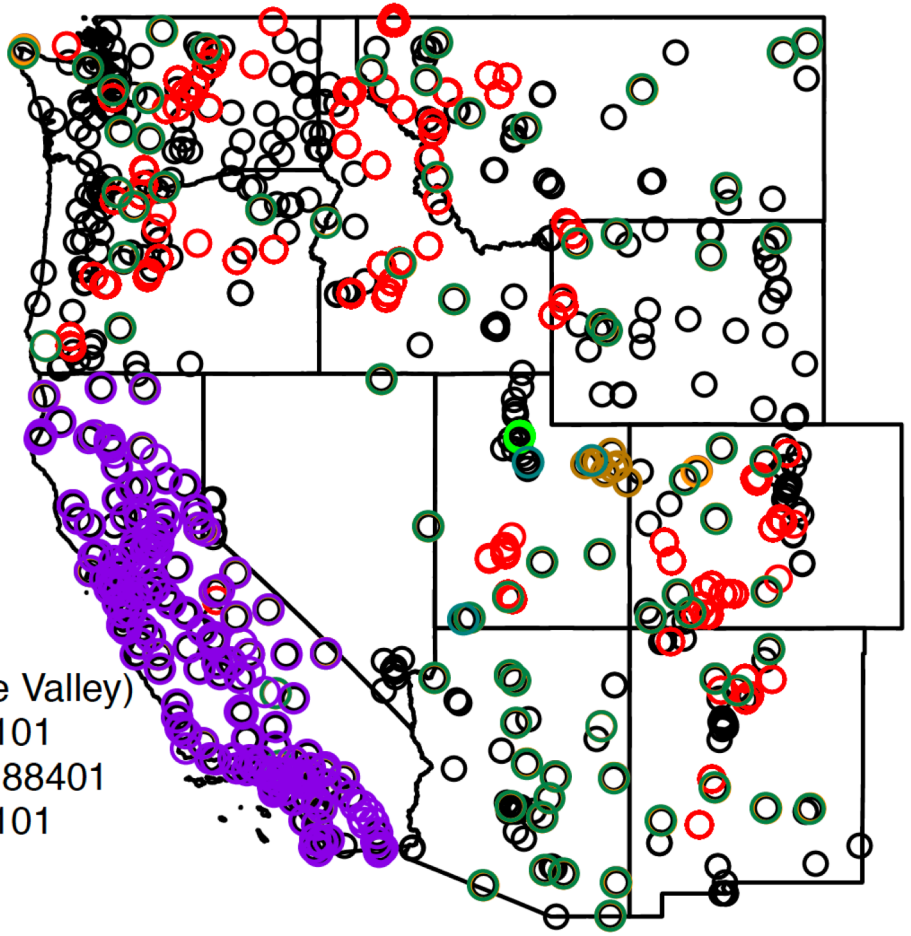
Reid et al., Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning Environ. Sci. Technol. 2015, 49, 3887−3896

2008 northern California wildfires

- Lightning storm on June 20-21, 2008
- Over 6000 lightning strikes
- Thousands of fires
- Smoke covered large population areas for weeks (est. 10-12 million people exposed)
- Potential for adverse health effects on large population

Reid et al., Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning Environ. Sci. Technol. 2015, 49, 3887−3896

All PM2.5 Observation Locations

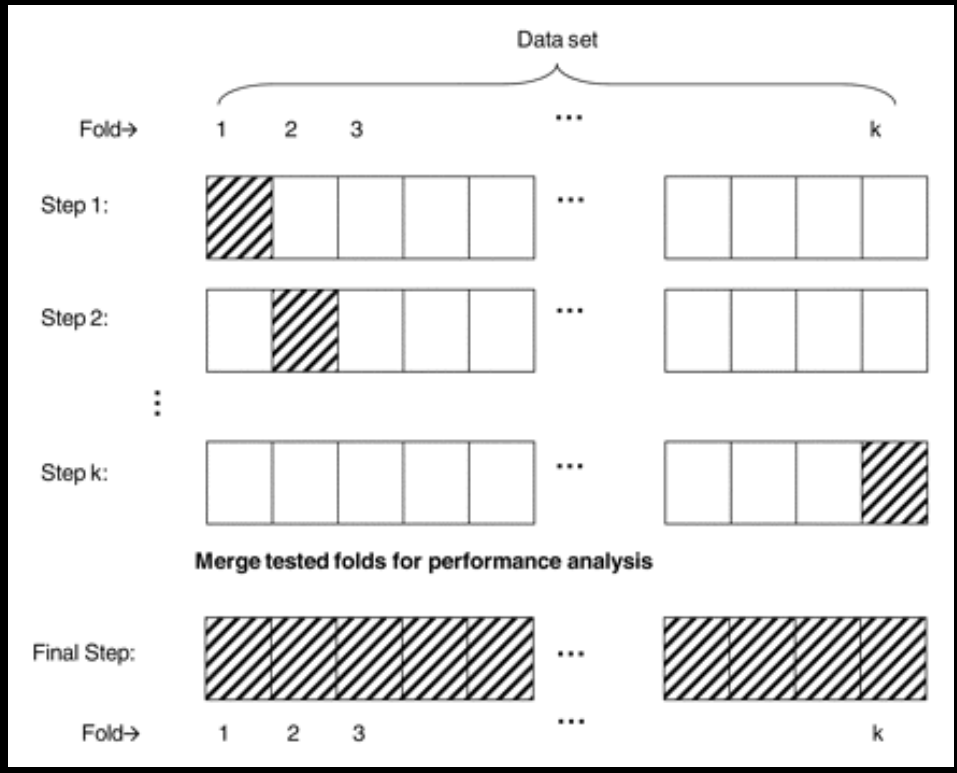


# locations (with lots of overlap)	Data Source
587	EPA PM2.5
3439	Fire Cache (DRI)
9	Uintah Basin
7	PCAPS (Salt Lake Valley)
95	IMPRHR2 MF 88101
95	IMPRHR2 RCFM 88401
91	IMPRHR3 MF 88101
190	CARB
3	Utah DEQ

Variables	Data Source	Temporal Resolution	Spatial Resolution	Buffer Size
Dependent Variable				
PM2.5 from monitoring stations	US EPA, states Federal Land Manager Environmental Database, Fire Cache Smoke Monitor Archive, IMPROVE Network, academic research groups	Daily or hourly	point	
Spatiotemporal Variables				
GOES Aerosol and Smoke Product (GASP) AOD	NOAA	Hourly	4 km	
Multi-Angle Implementation of Atmospheric Correction (MAIAC) AOD	NASA	Daily	1 km	
MODIS Active Fire Detection	NASA	Daily	1 km	
VIIRS Fire Occurence	NASA	Daily	375 m	
Hazard Mapping System (HMS) Smoke and Fire Product	NOAA	Daily	4km	25km, 50, 100km, 500km, 1000km, 2000km
MODIS Normalized Difference Vegetation Index (NDVI)	NASA	Monthly	1 km	
14 Meteorological Variables	NOAA: North American Mesoscale (NAM) Forecast System	6-Hourly	12 km	
Spatial Variables				
Elevation (m)	USGS	Nominal 2-month cycle	1 arc-second	
Percentages of land cover types	National Land Cover Database 2011	Every 5 years	30 m	1km
Kilometers of highway within buffer zones	National Highways Planning Network, US DOT			100, 250, 500, and 1000 m
Temporal Variables				
Julian Date		Daily		
Weekend		Daily		

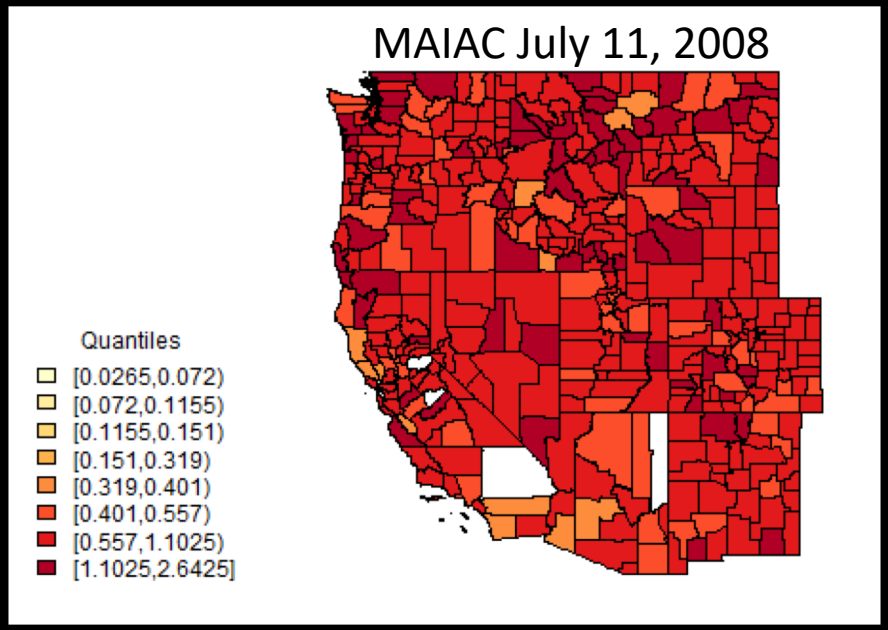
Statistical Methods – Machine Learning

- 10-fold cross validation within each algorithm
  - To choose the covariates
  - To choose the tuning parameters of the algorithm
- Minimize the CV-RMSE
- Caret package in R
- May do spatial, temporal CV
  - leave out airshed or year

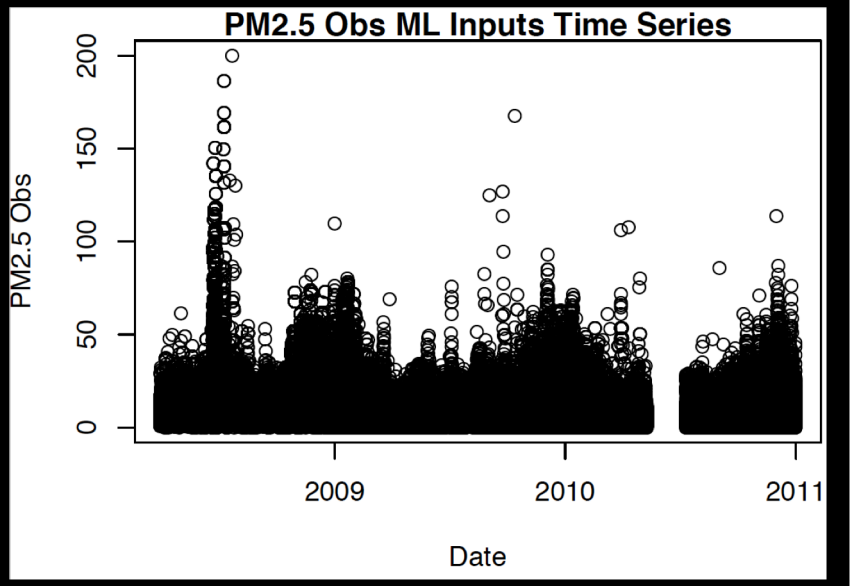


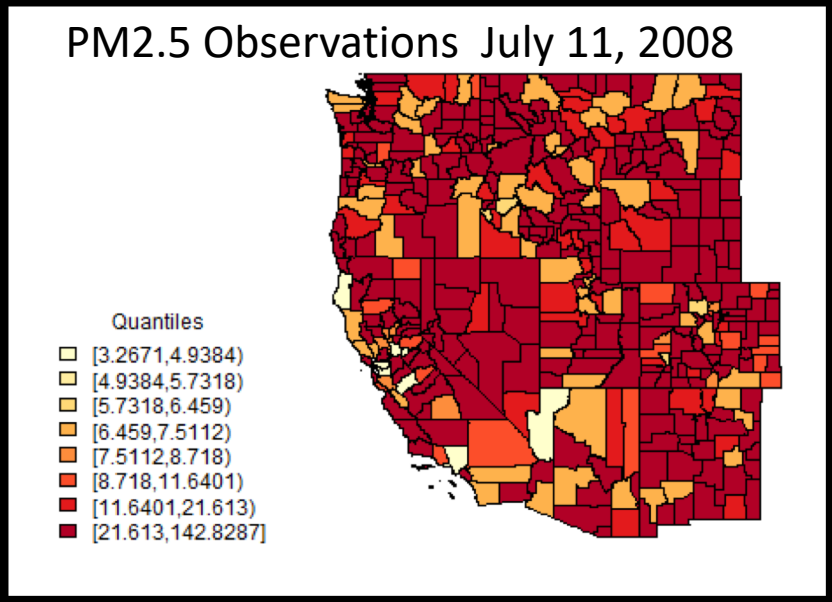
Status of Project

- Assembling data
  - GASP AOD 2008-2011
  - MAIAC AOD 2008-2014
  - Highways
  - NED (elevation)
  - NLCD (land cover)
  - NAM (weather) 18Z run 2008-2011 (with gap)
  - Others in Progress
- Starting to run ML models

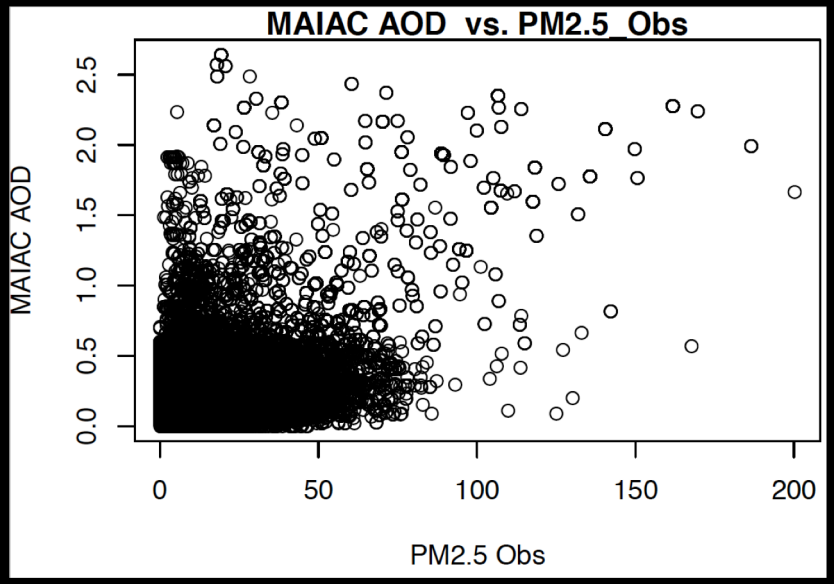


$PM_{2.5}$ Time series



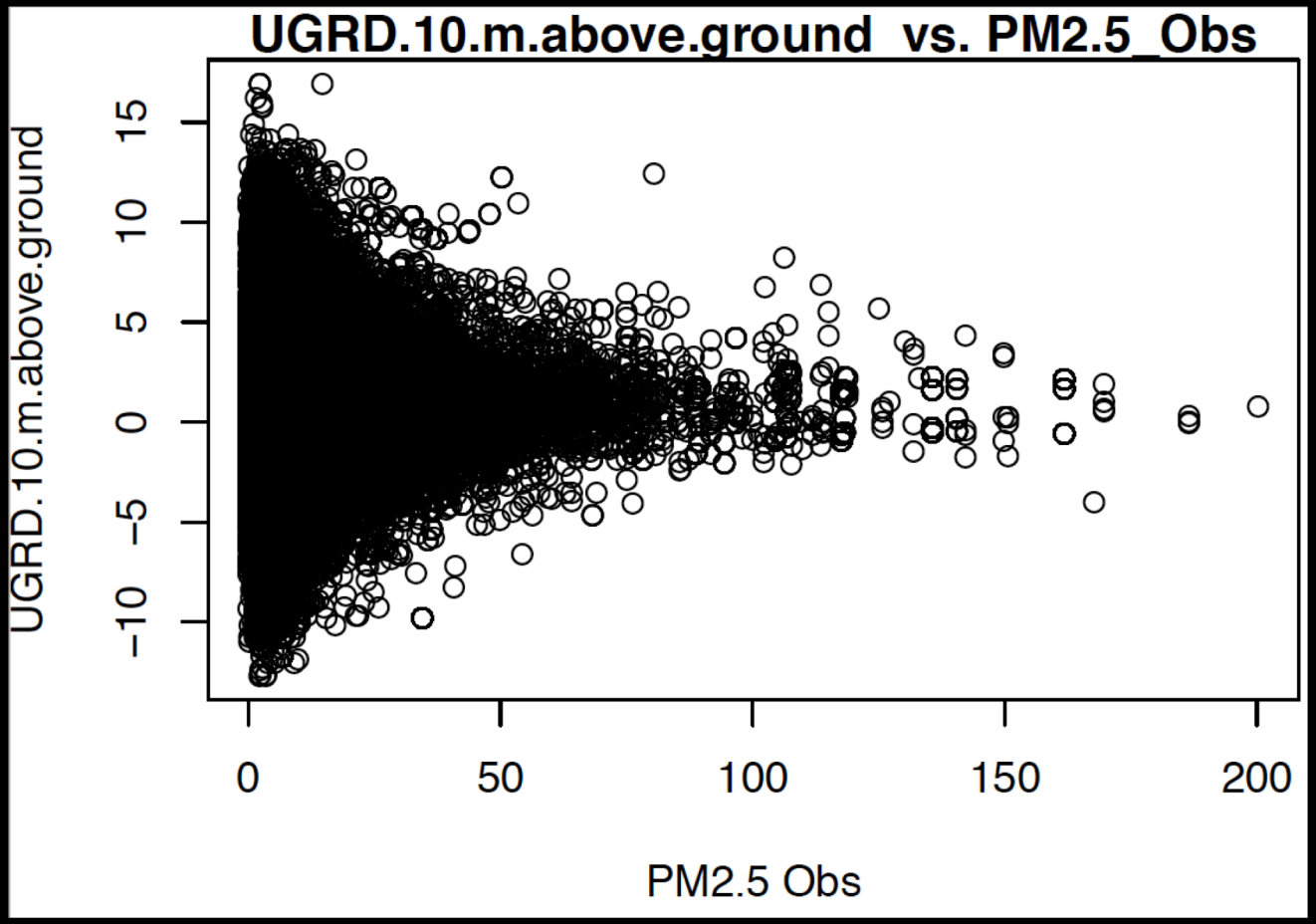


MAIAC AOD

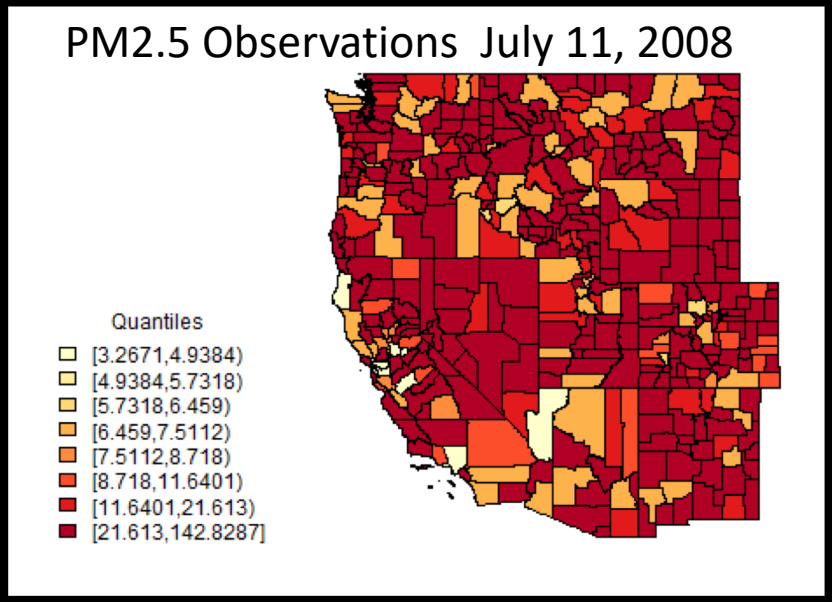


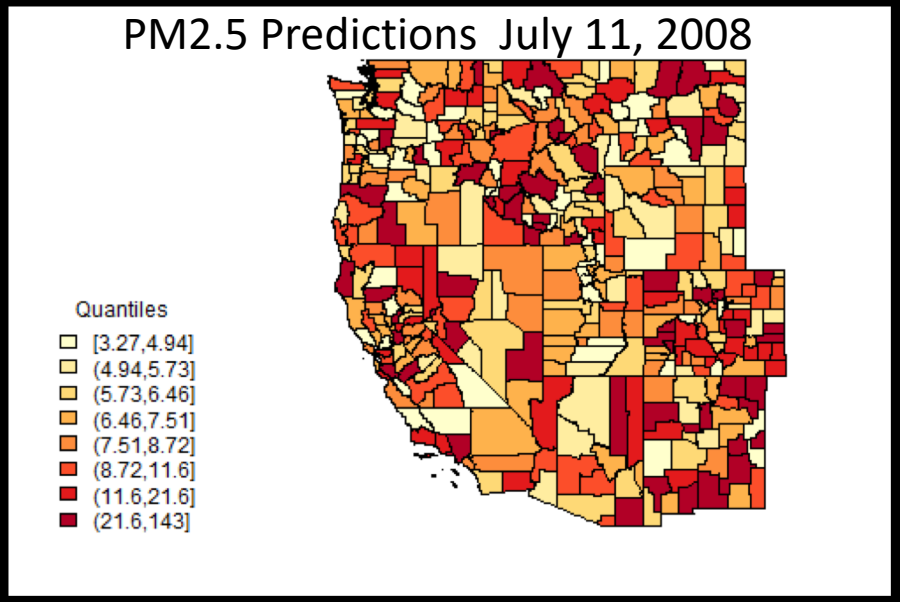
- Multi-Angle Implementation of Atmospheric Correction Aerosol Optical Depth
- Aqua-Terra MODIS Satellites
- Daily, 1 km resolution
- Available 2/26/2000 to present

A check on the data: Higher $PM_{2.5}$ with calm winds



Extremely Preliminary Results





Next Steps and Future Plans

- Finish data assimilation
  - ❋ Looking for more $PM_{2.5}$ data ❋
- ML modeling
  - Ensemble modeling
    - Random Forest, others
  - Cross validation by region/time instead of 10-fold
- Model by Region:
  - Predictor relationships could be different
- Pull in CTM model output
  - ❋ Looking for available data ❋
- Connect ML estimates to medical data for epidemiological analysis
- Flexible Geographic Resolution
  - County centroid
  - Population-weighted county centroid
  - ZIP code
  - Geo-coded address

Image: National Interagency Fire Center, Castle Rock Fire, Ketchum ID, 2007

Thank You!!

Questions?

Melissa May Maestas, PhD
Melissa.Maestas@Colorado.edu

Colleen E. Reid, PhD, MPH
Colleen.Reid@Colorado.edu

Gina Li, MA-candidate
Ellen Considine, BS-candidate

Funding: CU Boulder Grand Challenge to Earth Lab

Image: National Interagency Fire Center, NorCal Fires, 2008

Extra Slides

NAM Meteorological Variables

Variable Number	VariableName	VariableCode	AtmosLevelName	AtmosLevelCode	Units	file_type	time frame	24-hr summary
384	Planetary Boundary Layer Height	HPBL	surface	surface	m	grib2	anl	max
321	Temperature	TMP	2 m above ground	2 m above ground	K	grib2	anl	max
321	Temperature	TMP	2 m above ground	2 m above ground	K	grib2	anl	mean
324	Relative Humidity	RH	2 m above ground	2 m above ground	%	grib2	anl	mean
323	Dew Point Temperature	DPT	2 m above ground	2 m above ground	K	grib2	anl	mean
326	Total Precipitation	APCP	surface	surface	kg/m^2	grib2	0-0 day acc fcst	sum
316	Water Equivalent of Accumulated Snow Depth	WEASD	surface	surface	kg/m^2	grib2	anl	sum
317	Snow Cover	SNOWC	surface	surface	%	grib2	anl	max
325.1	U-Component of Wind	UGRD	10 m above ground	10 m above ground	m/s	grib2	anl	mean
325.2	V-Component of Wind	VGRD	10 m above ground	10 m above ground	m/s	grib2	anl	mean
1	Pressure Reduced to MSL	PRMSL	mean sea level	mean sea level	Pa	grib2	anl	mean
297	Pressure	PRES	surface	surface	Pa	grib2	anl	mean
246	Vertical Velocity (Geometric)	DZDT	850 mb	850 mb	m/s	grib2	anl	mean
203	Vertical Velocity (Geometric)	DZDT	700 mb	700 mb	m/s	grib2	anl	mean

Table 2. CV-RMSE and CV-R^2 Values for the Best Model Across the 11 Algorithms

	model with smallest CV-RMSE for subsets of variables			model with fewer variables whose CV-RMSE was within 1.5% of the smallest CV-RMSE		
	CV-RMSE ($\mu g/m^3$)	CV-R^2	no. of variables selected	CV-RMSE ($\mu g/m^3$)	CV-R^2	no. of variables selected
random forest	1.513	0.796	20	1.521	0.790	14
bagged trees	1.687	0.672	27	1.696	0.665	15
generalized boosting model	1.489	0.803	29	1.495	0.799	13
elastic net regression	1.848	0.538	28	1.852	0.535	27
multivariate adaptive regression splines	1.642	0.701	28	1.648	0.696	26
lasso regression	1.821	0.558	28	1.834	0.548	23
support vector machines	1.556	0.761	16	1.561	0.758	15
gaussian processes	1.580	0.746	16	1.591	0.739	14
generalized linear model	1.821	0.558	29	1.834	0.549	23
K-nearest neighbors	2.030	0.387	2	2.044	0.374	1
generalized additive model	1.607	0.725	26	1.609	0.724	25

Source: Reid et al. 2015. Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning. *Environmental Science & Technology.*

Predictions for the top two models agreed with visible imagery

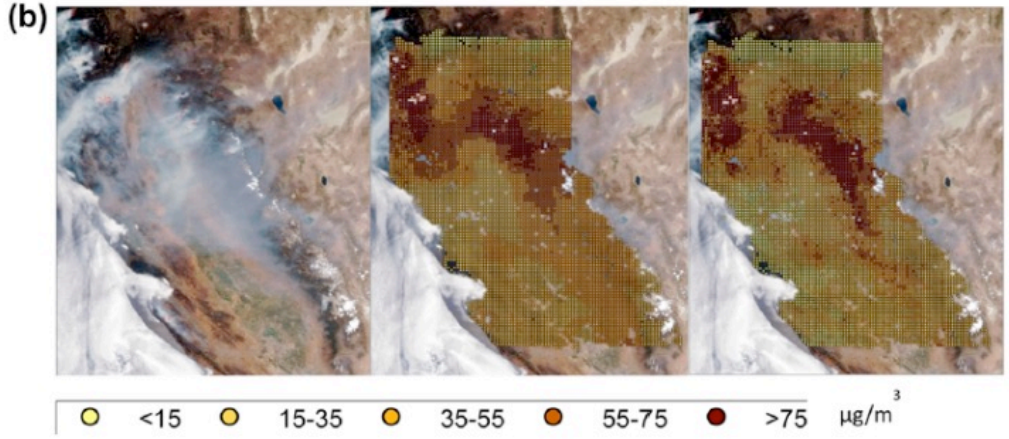


Figure 2. (a) Satellite image, predicted grid from RF and from GBM on June 29, 2008. (b) Satellite image and predicted grids from RF and GBM on July 11, 2008.

Source: Reid et al. 2015. Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning. *Environmental Science & Technology.*

Average Variable Importance	
GASP AOD	**1088.40**
distance to the nearest fire cluster	327.99
WRF-Chem	266.49
Julian date	263.31
Surface pressure	201.03
Local AOD	195.17
sea level pressure	182.81
Relative humidity	175.75
V component of wind speed	157.06
U component of wind speed	135.87
X-coordinate	134.86
MODIS AOD	130.63
temperature at 2 meters	116.02
Y-coordinate	122.27
dew point temperature	96.16
fire counts in nearest cluster / distance	95.87
elevation	75.52
planetary boundary layer height	84.49
% urban land use within 1 km	72.40
sum of traffic counts within 1 km	71.98
% vegetated land use within 1 km	37.64

Reid et al. 2015. *Environmental Science & Technology*

Table 3. CV-R^2 and CV-RMSE for GBM Models with Different Subsets of Variables

	all variables	GASP AOD plus[a]	WRF-Chem plus[a]	emissions[b] plus[a]	just plus[a]	MODIS AOD plus[a]	local AOD plus[a]	universal variables[c]
CV-RMSE	1.489	1.495	1.531	1.556	1.542	1.548	1.520	1.542
CV-R^2	0.803	0.800	0.774	0.757	0.768	0.764	0.784	0.770
no. of variables chosen	29 out of 29	25 out of 26	25 out of 26	18 out of 26	19 out of 25	22 out of 26	16 out of 26	19 out of 20

Reid et al. 2015. *Environmental Science & Technology*

Conclusions – Exposure Assessment

- Combining multiple spatiotemporal datasets including satellite data and/or chemical transport model data could **predict ground-level $PM_{2.5}$ well** during a wildfire
- **Tree-based algorithms** predicted ground-level $PM_{2.5}$ better than linear or other commonly used statistical models (e.g. GAM)
- Data that could apply to any fire, **"universal" variables, could predict ground-level $PM_{2.5}$ almost as well as the full model**

Statistical Algorithms

- Random Forest
- Tree bagging
- Generalized Boosting Models (GBM)
- Generalized Linear Models (GLM)
- GLM with penalized maximum likelihood (glmnet)
- Multivariate adaptive regression splines (Earth)
- Lasso regression
- Ridge regression
- Support Vector Machines
- Gaussian Processes
- Generalized Additive Models (GAM)
- K nearest neighbors regression

Reid et al. 2015. *Environmental Science & Technology*

Spatiotemporal exposure data sources

- Aerosol optical depth (AOD) satellite data
  - Benefits – covers the whole land area and has temporal coverage of more than once a day
  - Drawbacks – full column aerosol loading not just where people breathe, just measures during daylight hours
  - Incomplete data due to cloud cover, missing retrevals, resolution (improving)

Chemical transport models

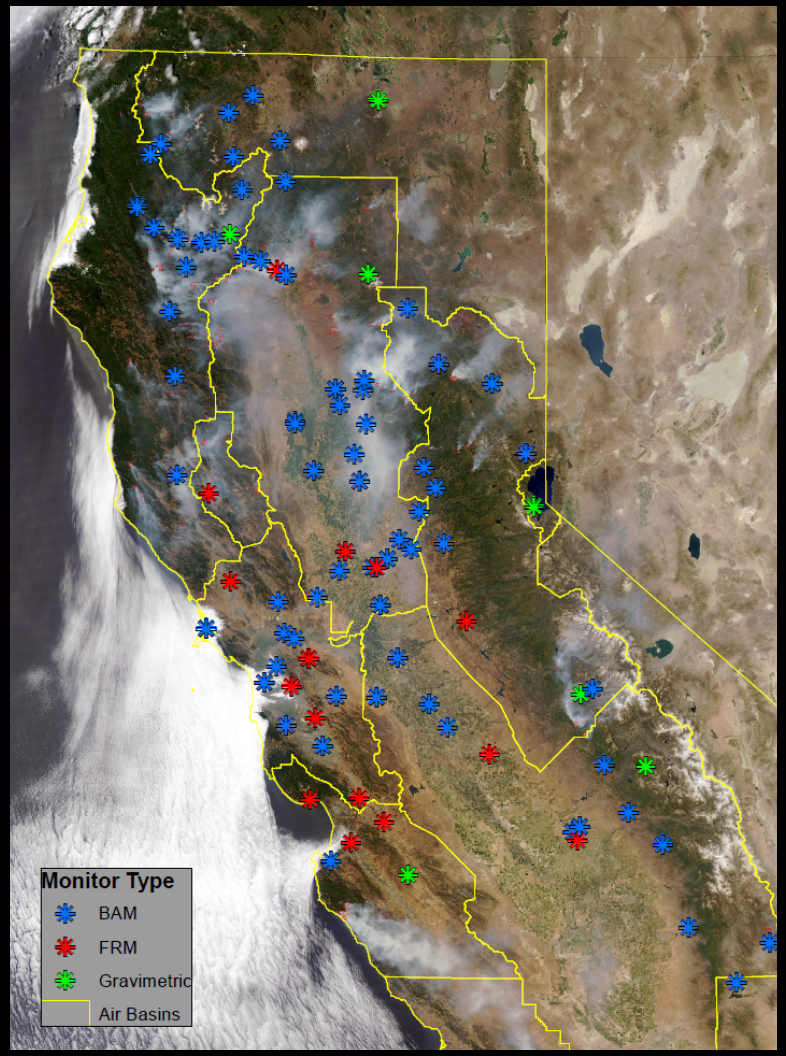
  - Benefits – can extract pollution estimates at ground level and covers whole land area
  - Drawbacks – modeled rather than measured data, uncertainties in the inputs, particularly the emissions estimates from wildfires

Variables	Data Source	Temporal Resolution	Spatial Resolution
Dependent Variable			
$PM_{2.5}$ from monitoring stations (N=121)	US EPA, California Air Resources Board, Air Districts, and US Forest Service	Daily or hourly	
Spatiotemporal Variables			
GASP AOD	National Oceanic and Atmospheric Administration	Half-hourly, daylight	4 km
MODIS AOD	NASA	Twice daily	10 km
Local AOD	Sonoma Technology, Inc.	Daily	0.5 km
WRF-Chem $PM_{2.5}$ ($\mu g/m^3$)	National Center for Atmospheric Research	Hourly	12 km
Distance to nearest cluster of active fires (m) Counts of fires in nearest cluster / distance	Derived from USDA Forest Service Remote Sensing Applications Center	Daily	
Relative Humidity (%) Sea level pressure (Pa) Surface pressure (Pa) Planetary boundary layer height (m) U-component of wind speed (m/s) V-component of wind speed (m/s) Dew point temperature (K) Temperature at 2 m (K)	Rapid Update Cycle	Daily	13 km
Spatial Variables			
X-coordinate (m) Y-coordinate (m)	U.S. Environmental Protection Agency Air Quality System		
Counts of traffic within 1 km	Dynamap 2000, TeleAtlas	Annual	1 km
% of urban land use within 1km % of agricultural land use within 1km % of vegetation land use within 1km Any High intensity land use within 1km	2006 National Land Cover Database		1 km
Elevation (m)	National Elevation Dataset 2010		
Binary indicator variables for air basin	California Air Resources Board		Air Basin
Population Density	U.S. Census 2000		Block Group
Temporal Variables			
Julian Date Weekend		Daily	

Reid et al. 2015. *Environmental Science & Technology*

$PM_{2.5}$ Monitoring Data

- 121 $PM_{2.5}$ Monitors
  - EPA, CARB, USFS
  - 38 FRM
  - 16 other gravimetric
  - 67 BAMs
- Co-located FEM monitors agree with FRM (Pearson r values 0.94 – 1.00).



Statistical Methods – Machine Learning

- 29 predictor variables
- 121 monitoring locations
- 11 statistical algorithms
- 10-fold cross validation within each algorithm
  - To choose the covariates
  - To choose the tuning parameters of the algorithm
- Minimize the CV-RMSE
- Caret package in R
  - Functions rfe and train