

**Ensemble-Based Deep Learning for Estimating PM_{2.5} over
California with Multisource Big Data Including Wildfire Smoke**

Environment International

Supplementary Data

Lianfa Li ^{1,2,*}, Mariam Girguis ¹, Frederick Lurmann ³, Nathan Pavlovic ³, Crystal McClure ³, Meredith Franklin ¹, Jun Wu ⁴, Luke D. Oman ⁵, Carrie Breton ¹, Frank Gilliland ¹, Rima Habre ^{1,*}

- 1. Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA*
- 2. State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources, Chinese Academy of Sciences, Beijing, China*
- 3. Sonoma Technology, Inc., Petaluma, CA, USA*
- 4. Program in Public Health, Susan and Henry Samueli College of Health Sciences, University of California, Irvine, CA, USA*
- 5. Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, MD, USA*

Number of pages: 28

Number of figures: 18

Number of tables: 5

Contents

Table S1. A summary of acronyms and their brief descriptions.....	S1-S2
Table S2. Correlation between MAIAC AOD and AERONET AOD for the different projections of the 2008-2016 dataset and the 2017 dataset.....	S3
Table S3. Fixed and sampled predictors for the PM _{2.5} models.....	S4
Table S4. Performance statistics of 100 trained models for the samples of simple block bootstrap	S5
Table S5. Comparison of different models	S6
Fig. S1. Histograms (a and c) and boxplots (b and d) of observed PM _{2.5} (a and b) and MAIAC AOD (already converted) (c and d).....	S7
Fig. S2. Four temporal basis functions for routine PM _{2.5} measurements.....	S8
Fig. S3. Bar plots of correlation for the features.....	S9
Fig. S4. Boxplot plots of R ² (a) and RMSE (b) for the samples of simple block bootstrap.....	S10
Fig. S5. Scatter plots of observed vs. predicted PM _{2.5} (a) and observed vs. residual PM _{2.5} (b) for the samples of simple block bootstrap.....	S11
Fig. S6. ICE for instance and PDP for selected features with high importance score.....	S12
Fig. S7. ICE and PDP of several representative features.....	S13-S15
Fig. S8. Bar plots of feature importance of each feature by SHAP for a single model (Model 1).....	S16
Fig. S9. Bar plots of feature importance of each feature by SHAP (averages of 100 trained models).....	S17
Fig. S10. Histogram of the observed (a) and predicted (b) PM _{2.5}	S18
Fig. S11. Boxplots of observed (a) vs. predicted (b) PM _{2.5} (a).....	S19
Fig. S12. Histogram (a) and boxplot (b) of the residuals of the ensemble predictions for test samples.....	S20
Fig. S13. ACF plot of the residuals of ensemble predictions.....	S21
Fig. S14. Bar plot of mean coefficients of variation for predicted PM _{2.5} across different land-use.....	S22
Fig. S15. Distributions of predicted PM _{2.5} surfaces (a, b and c) vs. HYSPLIT modeled wildfire smoke PM _{2.5} (d, e and f) for the week before (a and d) and after (c and f) the 2012 wildfire week (b and e).....	S23-S24

Fig. S16. Distributions of predicted $PM_{2.5}$ surfaces (a, b and c) vs. HYSPLIT modeled wildfire smoke $PM_{2.5}$ (d, e and f) for the week before (a and d) and after (c and f) the 2015 wildfire week (b and e).....	S25-S26
Fig. S17. Comparison of the time series of predicted $PM_{2.5}$, observed $PM_{2.5}$, and HYSPLIT modeled wildfire smoke $PM_{2.5}$ at three AQS monitoring sites surrounding the wildfire region for the 2012 wildfire week.....	S27
Fig. S18. Comparison of the time series of predicted $PM_{2.5}$, observed $PM_{2.5}$, and HYSPLIT modeled wildfire smoke $PM_{2.5}$ at three AQS monitoring sites surrounding the wildfire region for the 2015 wildfire week.....	S28

Table S1 A summary of acronyms and their brief descriptions

Acronyms	Description
PM _{2.5}	Fine particulate matter with aerodynamic diameter smaller than 2.5 μ m
CA	California
US	The United States
MODIS	The Moderate Resolution Imaging Spectroradiometer
TERRA	Originally known as EOS AM-1
AQUA	Originally known as EOS PM-1
AOD	Aerosol optical depth
DT	The Dark Target algorithm to retrieve AOD
DB	The Deep Blue algorithm to retrieve AOD
MAIAC	The Multiangle Implementation of Atmospheric Correction algorithm to retrieve AOD
M2GMI	Modern-Era Retrospective analysis for Research and Applications Global Modeling Initiative Replay Simulation, or MERRA-2 GMI Replay Simulation
AERONET	AErosol RObotic NETwork
LUR	Land-Use Regression
BME	Bayesian Maximum Entropy
CV	Cross Validation
GWR	Geographically Weighted Regressions
ML	Machine Learning
GAM	Generalized Additive Model
RMSE	Root Mean Square Error
PBLH	Planetary Boundary Layer Height
NDVI	Normalized Difference Vegetation Index
EPA	Environmental Protection Agency
AQS	Air Quality System
ICV2	The University of Southern California Intra-Community Variability study 2

IQR	InterQuartile Range
UTM	Universal Transverse Mercator
NASA	National Aeronautics and Space Administration
GMI	Global Modeling Initiative
FEER.v1	Fire Energetics and Emissions Research version 1.0
HYSPLIT	Hybrid Single-Particle Lagrangian Integrated Trajectory
WRAP	Western Regional Air Partnership
NAM	North American Mesoscale Forecasting System
NOAA	National Oceanic and Atmospheric Administration
NLCD	National Land Cover Database
bagging	bootstrap aggregating
USC	The University of Southern California
SHAP	Shapley Additive exPlanations
ICE	Individual Conditional Expectation
PDP	Partial Dependence Plot
PM	Particular Matter
FRP	Fire Radiative Power
MAP	The NASA Modeling, Analysis and Prediction
PBL	Planetary Boundary Layer
NCCS	NASA Center for Climate Simulation

Table S2 Correlation between MAIAC AOD and AERONET AOD for the different projections of the 2008-2016 dataset and the 2017 dataset

Buffer	Correlation with AERONET AOD (Alberty projection for 2008-2016 data)	Correlation with AERONET AOD (Sinusoidal projection for 2017 data)	R ² with AERONET AOD (Alberty projection for 2008-2016 data)	R ² with AERONET AOD (Sinusoidal projection for 2017 data)
Number of samples ^a	2732	2712	2732	2712
0 m	0.83*	0.77 *	0.67*	0.54*
1000 m	0.84*	0.78*	0.68*	0.55*
3000 m	0.85*	0.81 *	0.71*	0.61*
5000 m	0.87*	0.83*	0.73*	0.64*
7000 m	0.87*	0.84*	0.74*	0.66*
9000 m	0.87*	0.84*	0.74*	0.66*

Note: ^a the samples for 16 AERONET sites; * indicates statistical significance.

Table S3 Fixed and sampled predictors for the PM_{2.5} models

Fixed predictors	Predictors for bootstrap sampling
MAIAC AOD, carbon monoxide, nitrate surface mass concentration - PM _{2.5} , dry deposition of O _x , sea salt surface mass concentration PM _{2.5} , SO ₄ surface mass concentration, 10-meter eastward wind, 2-meter eastward wind, indicator of stagnation, wind sheer/mixing, year, day of year, month index, roadway length for major roads, traffic density, wildfire, impervious layer, NDVI, wind speed, daily mean downward shortwave radiation, longitude, latitude, square of longitude, square of latitude, product of latitude and longitude, local AOD mean, temporal basis function 1, temporal basis function 2	Evaporation land, black carbon surface mass concentration, total precipitation land, organic carbon surface mass concentration, dust surface mass concentration - PM _{2.5} , 10-meter northward wind, pressure, 2-meter northward wind, minimum temperature, specific relative humidity, precipitation, elevation, maximum temperature, mean PBLH, open water, developed, open space, developed-low intensity, developed- medium intensity, developed-high intensity, barren land, deciduous forest, evergreen forest, mixed forest, shrub/scrub, grassland/herbaceous, pasture/hay, cultivated crops, woody wetlands, emergent herbaceous wetlands, temporal basis function 3, temporal basis function 4

Table S4. Performance statistics of 100 trained models for the samples of simple block bootstrap

	Training		Validation ^a		Testing ^b	
	Mean	Range	Mean	Range	Mean	Range
Sample size	17,629		4,407		12,776	
R ² (range)	0.92	(0.71, 0.97)	0.72	(0.63, 0.77)	0.72	(0.63, 0.77)
RMSE (µg/m ³)	1.55	(0.91, 3.26)	3.49	(3.10, 3.92)	3.50	(2.95, 4.29)

Table S5 Comparison of different models

	Training R ²	RMSE ($\mu\text{g}/\text{m}^3$)	Testing R ²	RMSE ($\mu\text{g}/\text{m}^3$)
GAM	0.65	3.76	0.61	3.79
Feed-forward neural network	0.84	2.51	0.77	2.95
Base model (residual deep network)	0.94	1.54	0.82	2.70
Bagging of 100 base models	-	-	0.87	2.30

Note: - indicates no availability for the metrics.

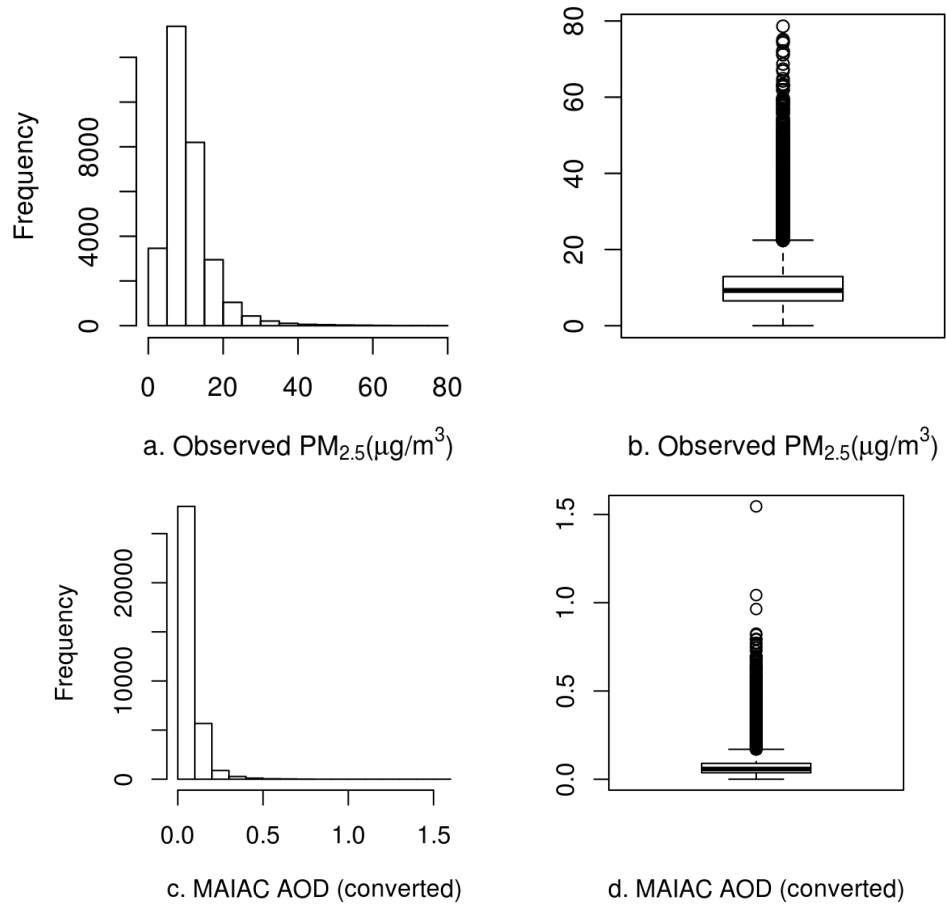


Fig. S1 Histograms (a and c) and boxplots (b and d) of observed PM_{2.5} (a and b) and MAIAC AOD (already converted) (c and d).

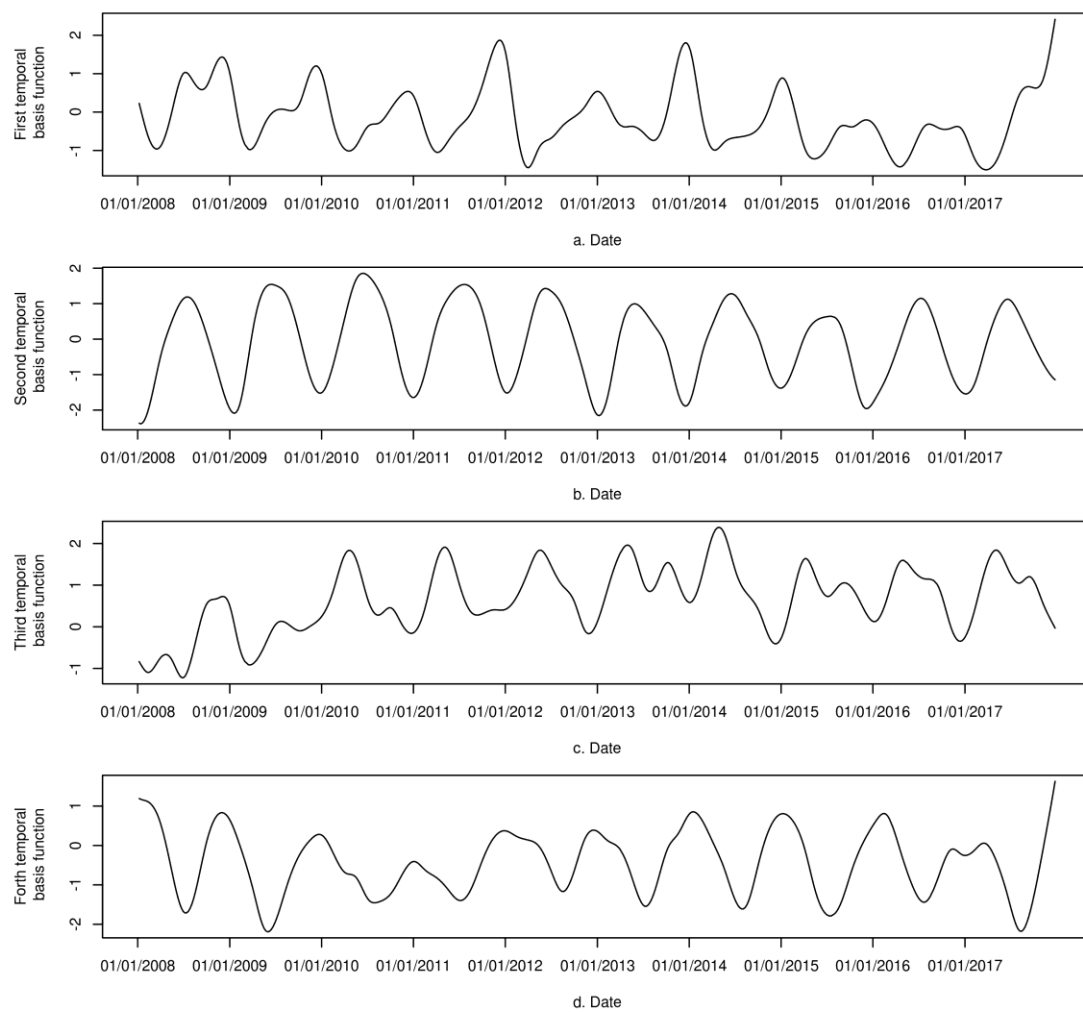


Fig. S2 Four temporal basis functions for routine PM_{2.5} measurements.

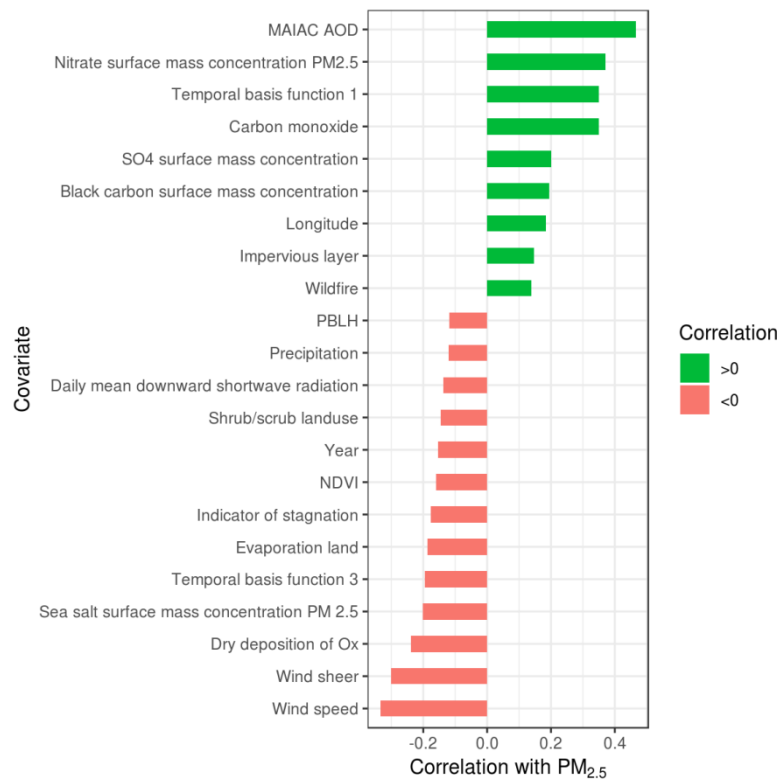


Fig. S3 Bar plots of correlation for the features.

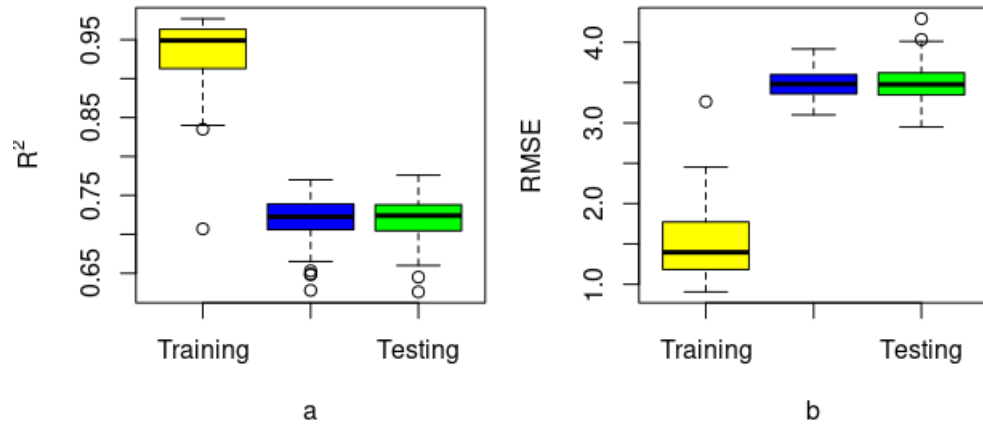


Fig. S4 Boxplot plots of R^2 (a) and RMSE (b) for the samples of simple block bootstrap.

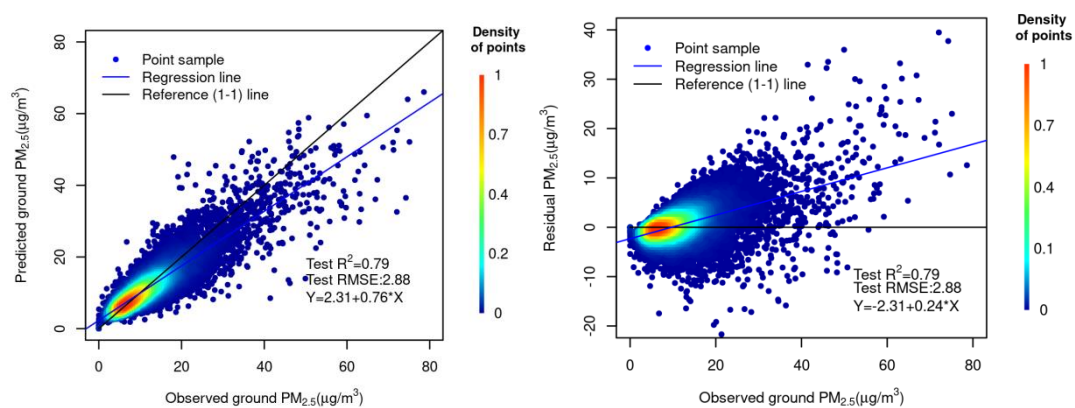


Fig. S5 Scatter plots of observed vs. predicted PM_{2.5} (a) and observed vs. residual PM_{2.5} (b) for the samples of simple block bootstrap.

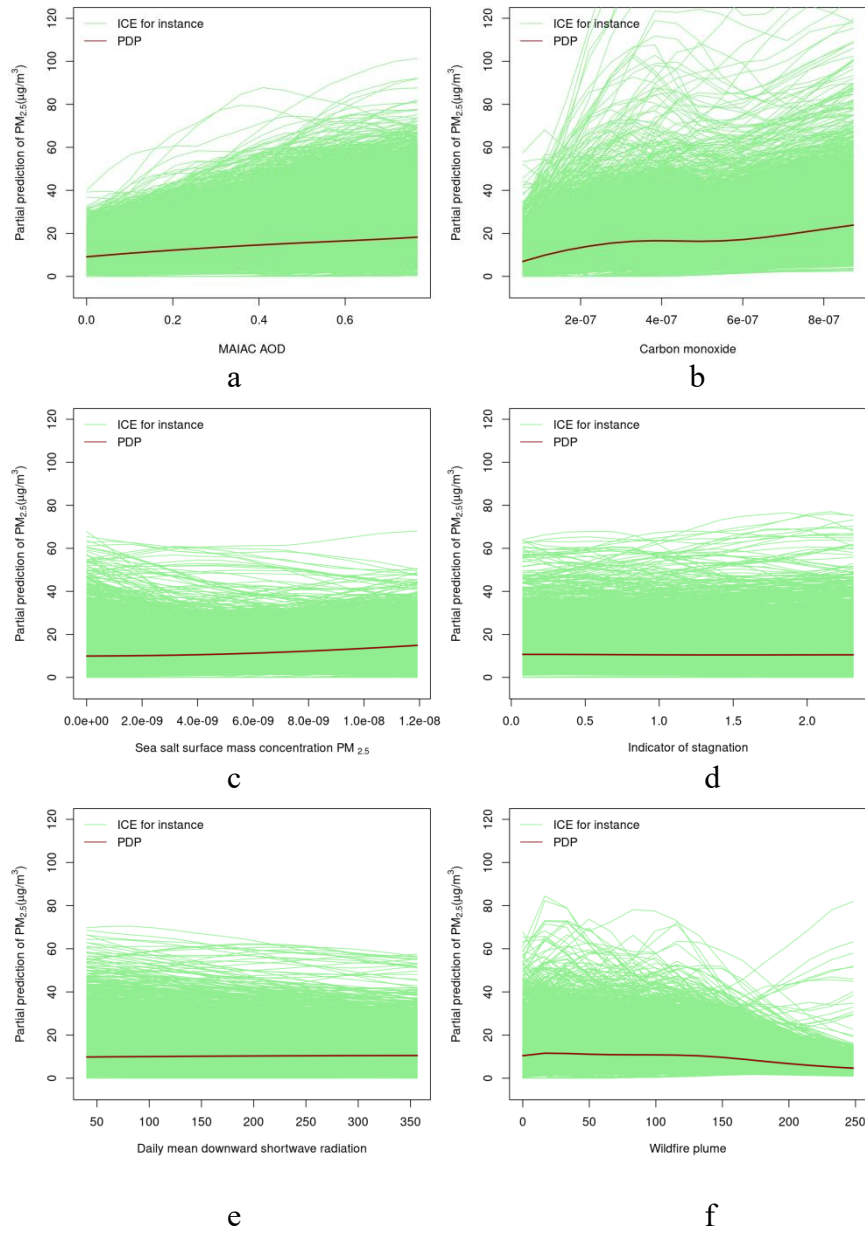
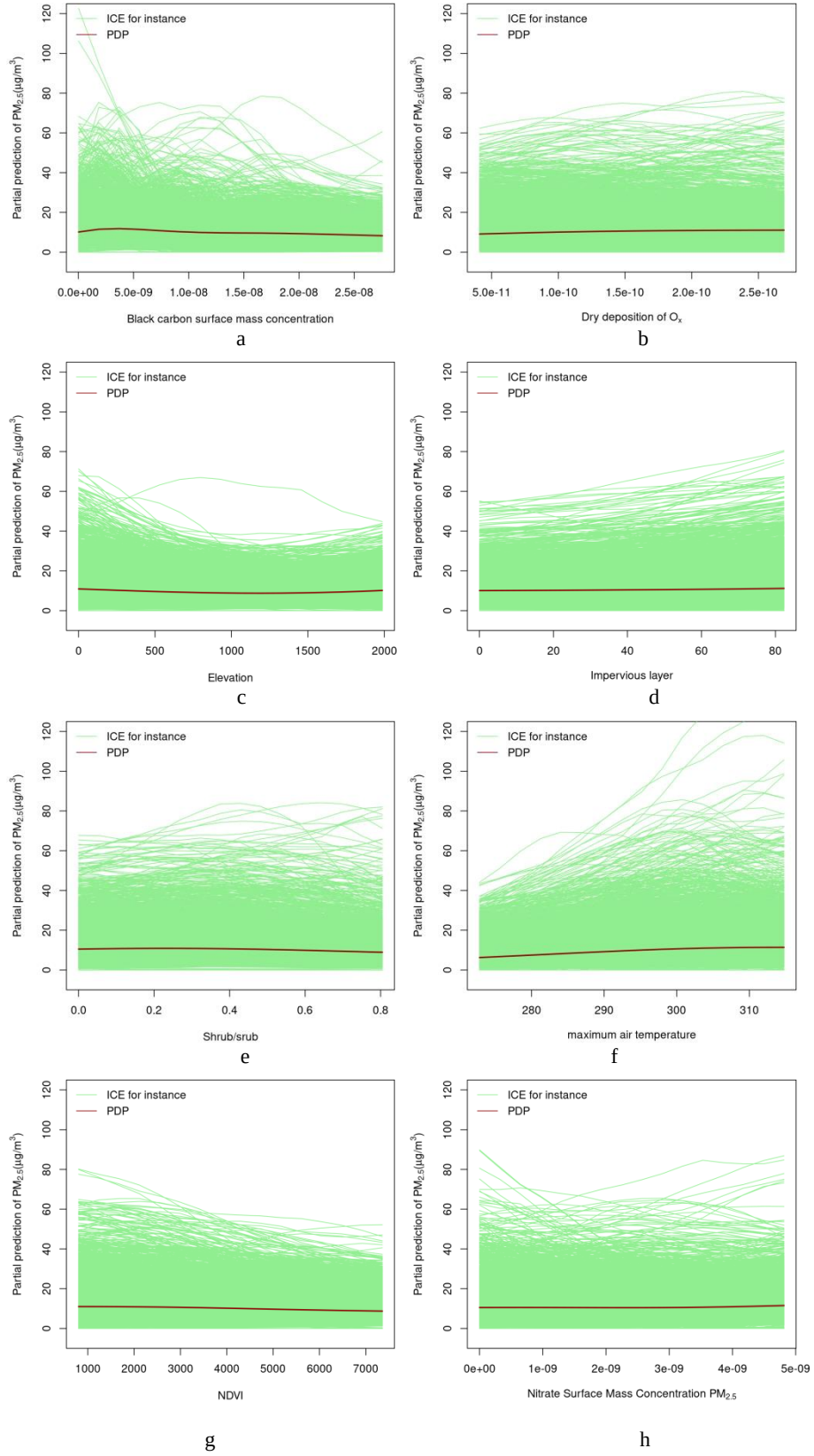
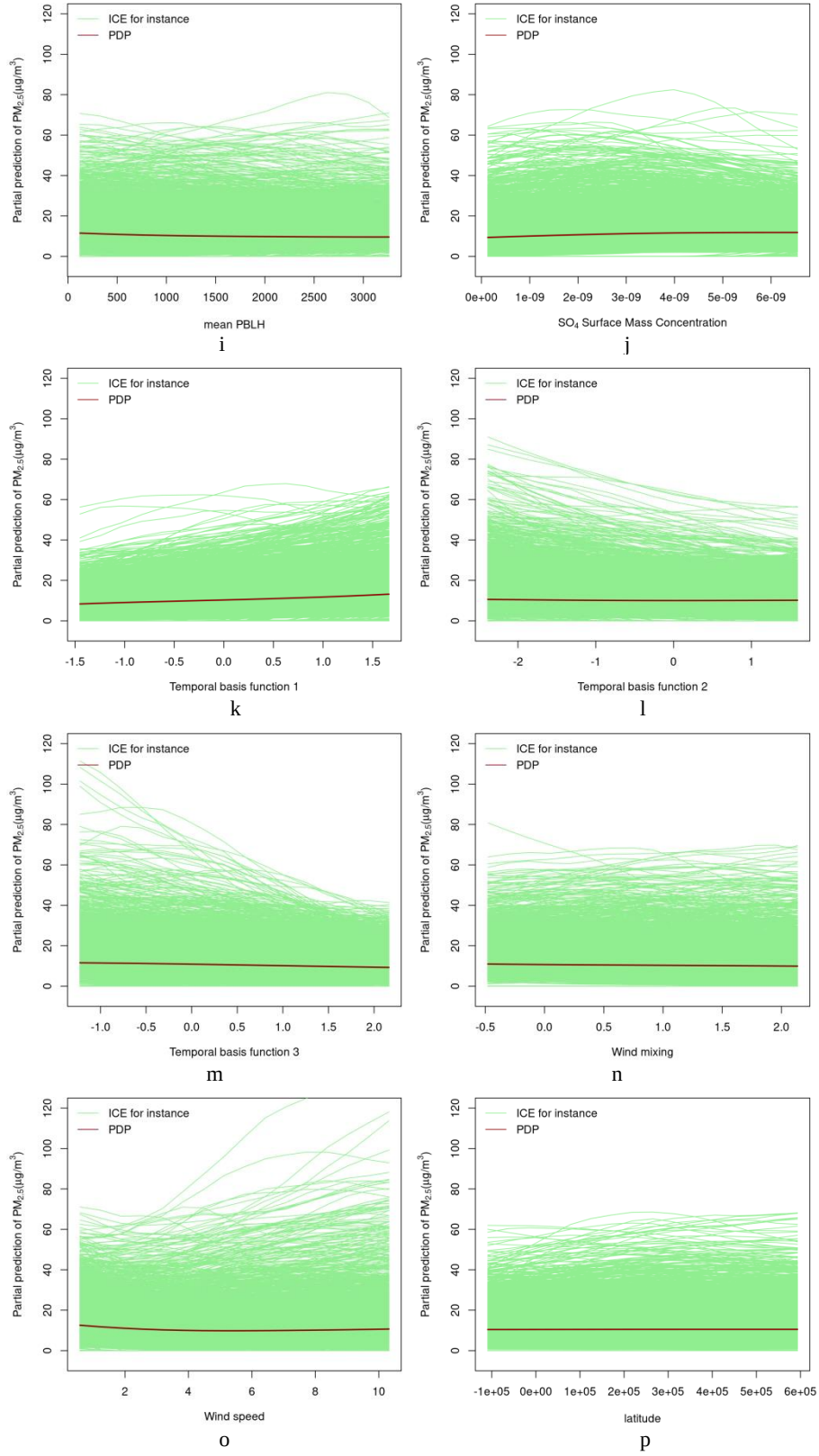


Fig. S6 ICE for instance and PDP for selected features with high importance score.





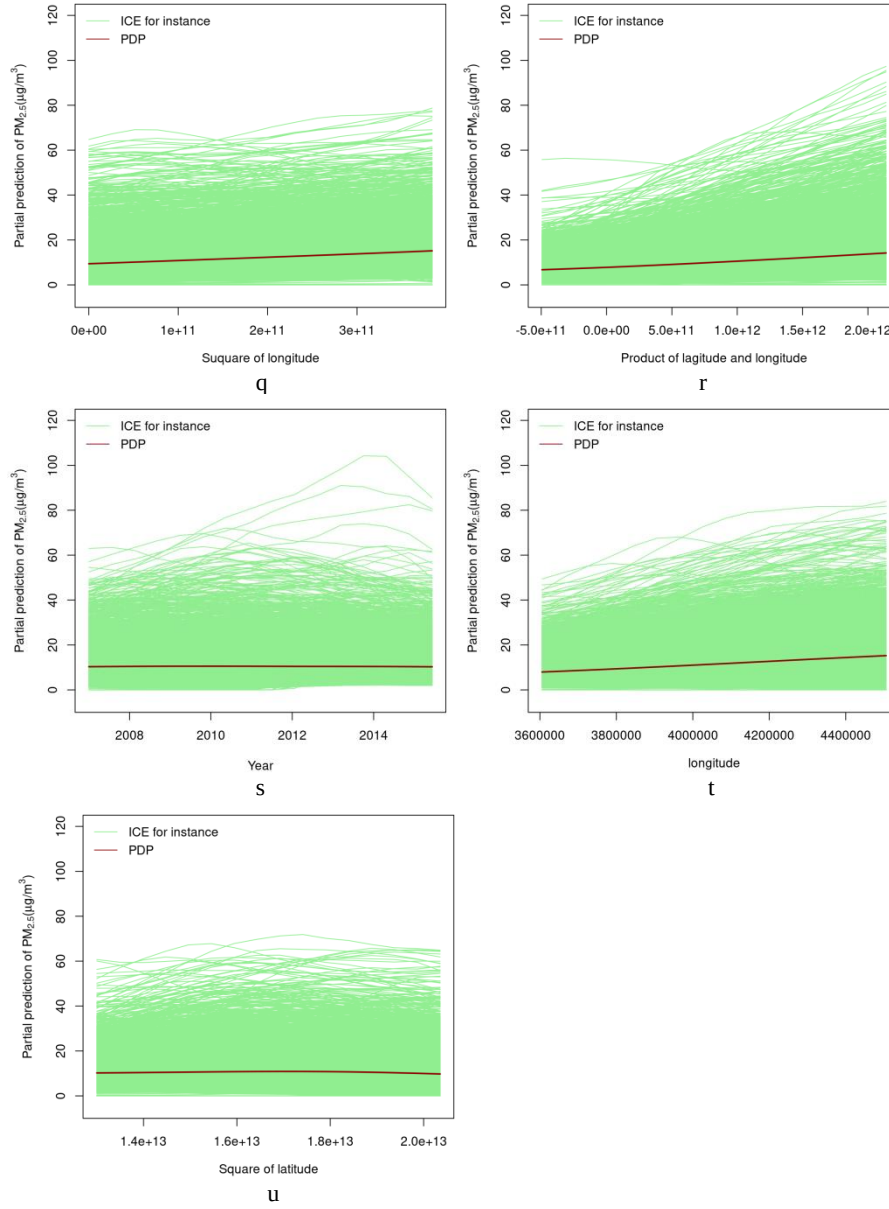


Fig. S7 ICE for instance and PDP for several representative features.

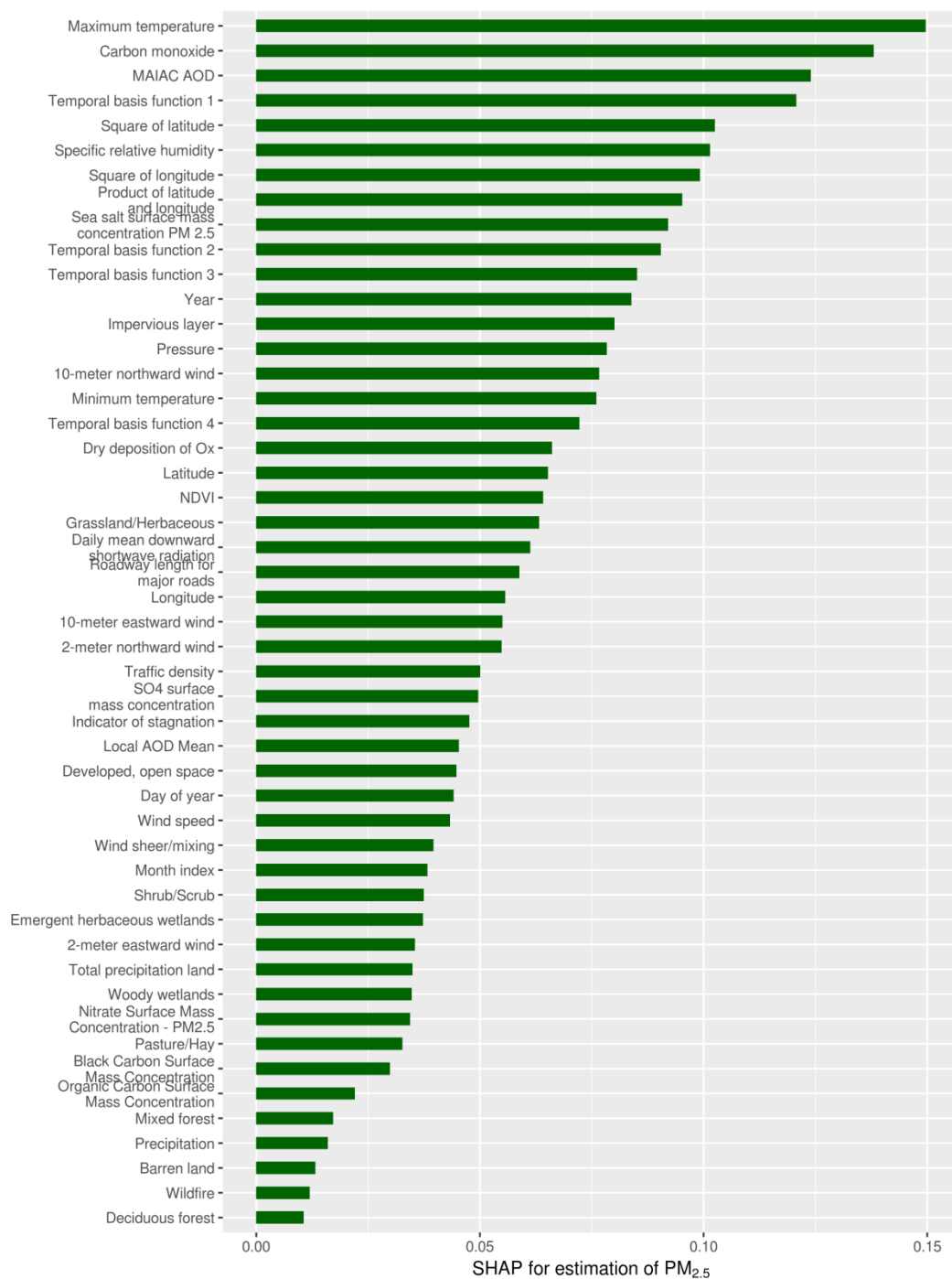


Fig. S8 Bar plots of feature importance of each feature by SHAP for a single model (Model 1).

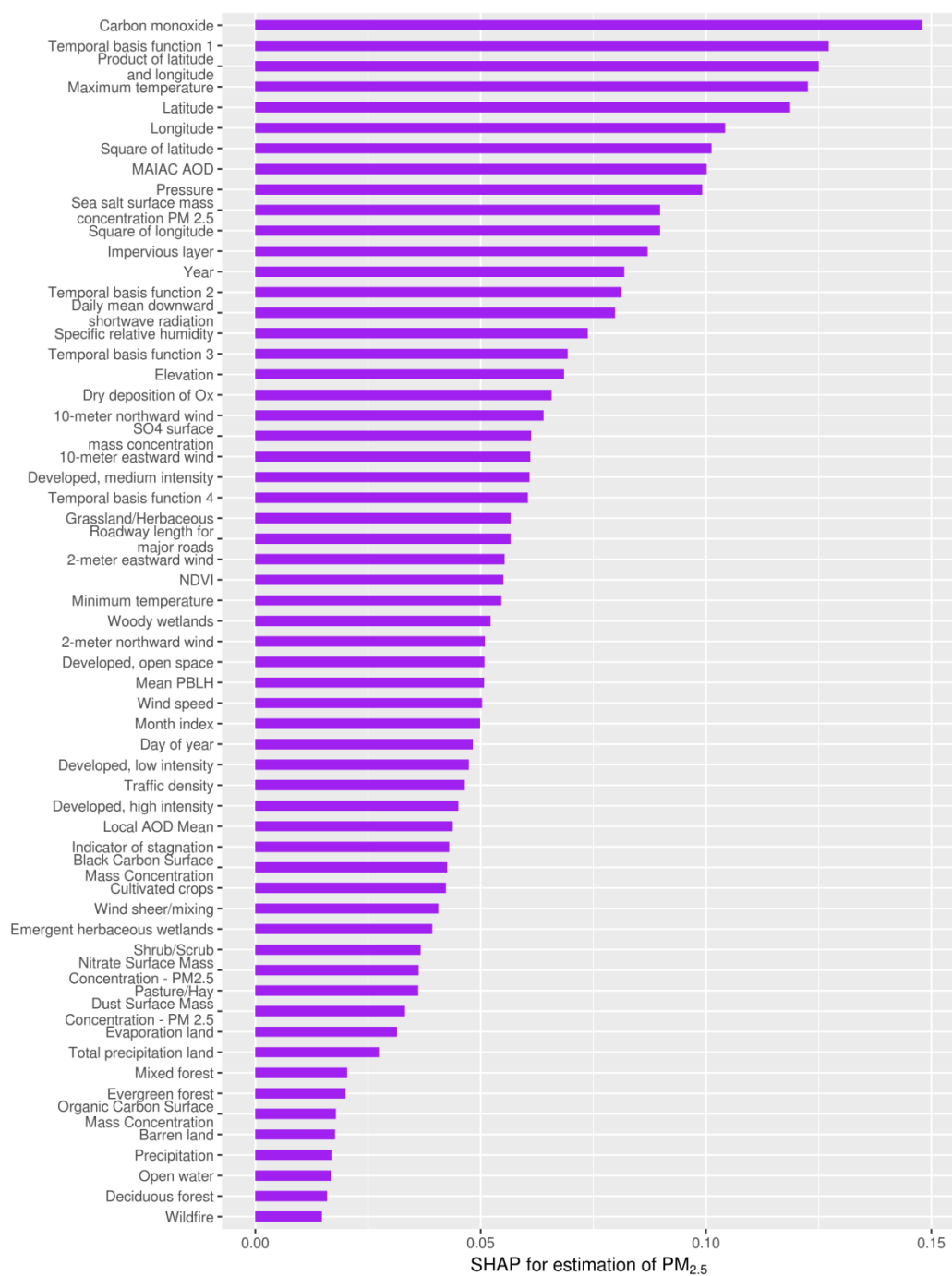


Fig. S9 Bar plots of feature importance of each feature by SHAP (averages of 100 trained models).

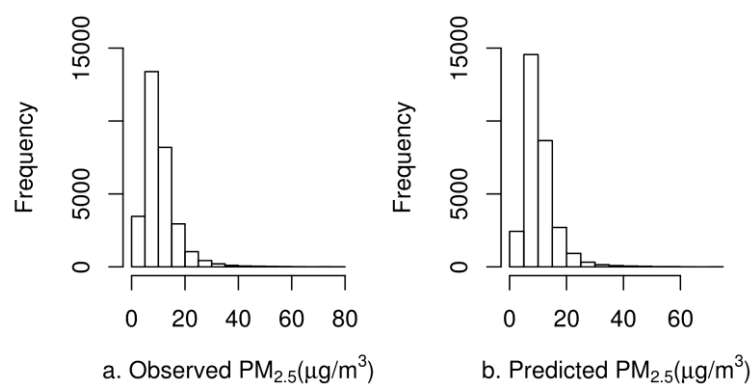


Fig. S10 Histogram of the observed (a) and predicted (b) PM_{2.5}.

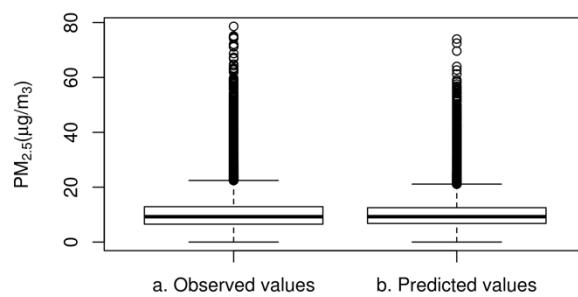


Fig. S11 Boxplots of observed (a) vs. predicted (b) PM_{2.5} (a).

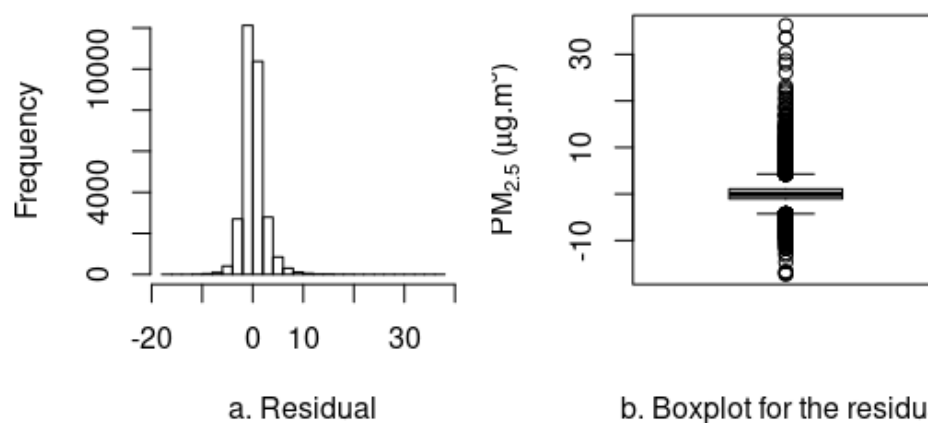


Fig. S12 Histogram (a) and boxplot (b) of the residuals of the ensemble predictions for test samples.

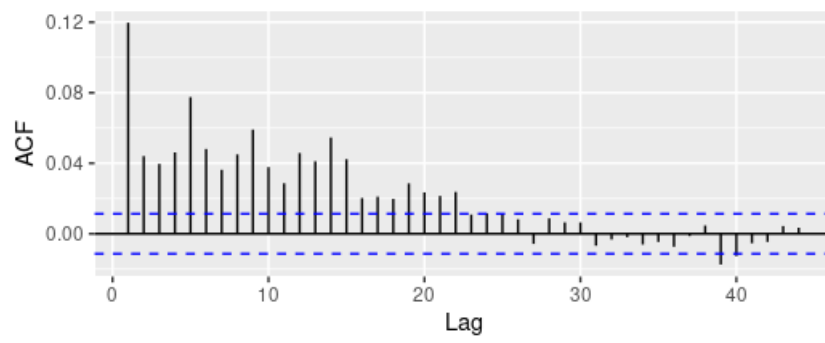


Fig. S13 ACF plot of the residuals of ensemble predictions.

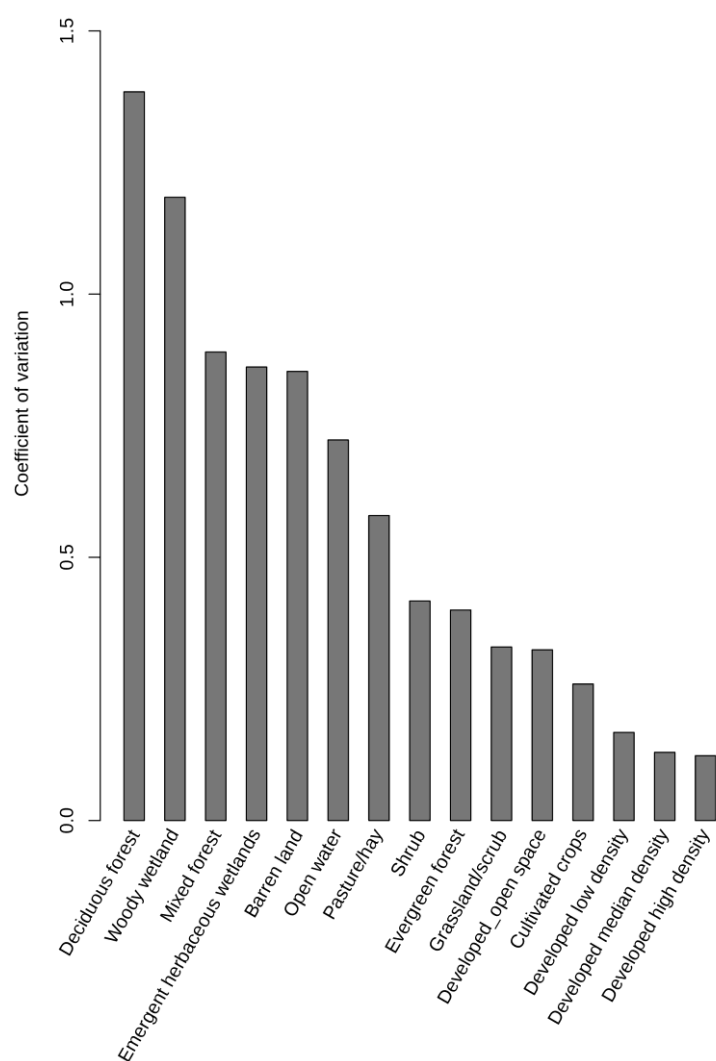
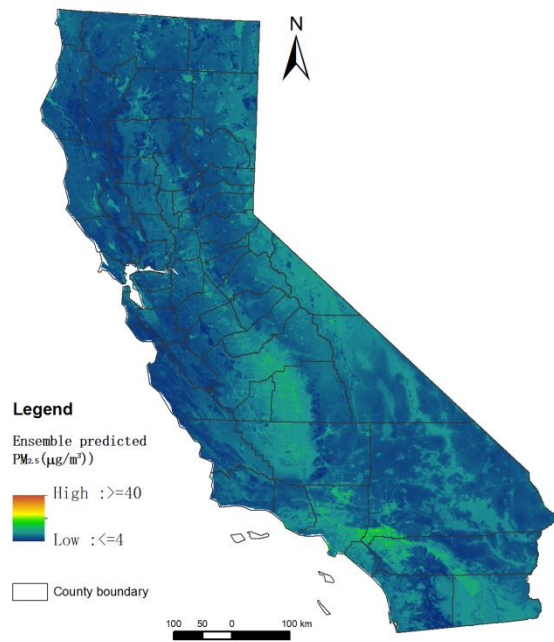
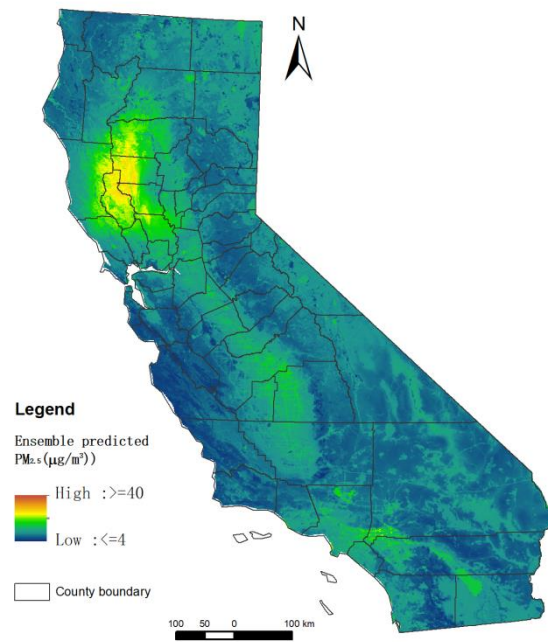


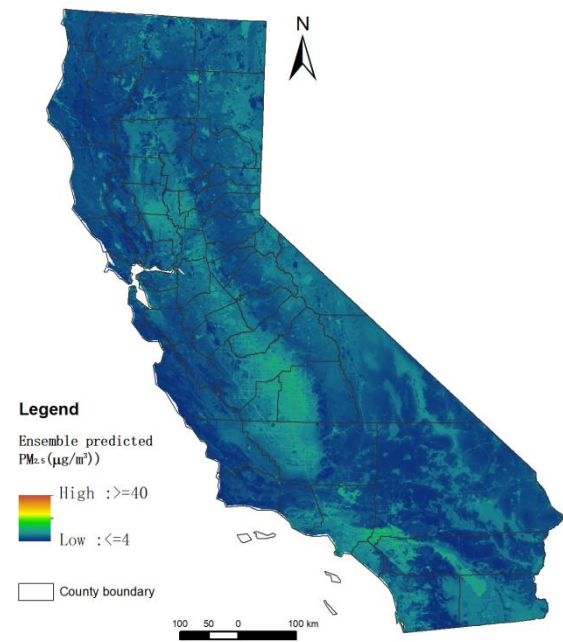
Fig. S14 Bar plot of mean coefficients of variation for predicted PM_{2.5} across different land-use.



a. Ensemble predicted PM_{2.5}
for the summer week of 2012 (Jul. 09 to 15)



b. Ensemble predicted PM_{2.5}
for the summer week of 2012 (Jul. 16 to 22)



c. Ensemble predicted PM_{2.5}
for the summer week of 2012 (Jul. 23 to July 29)

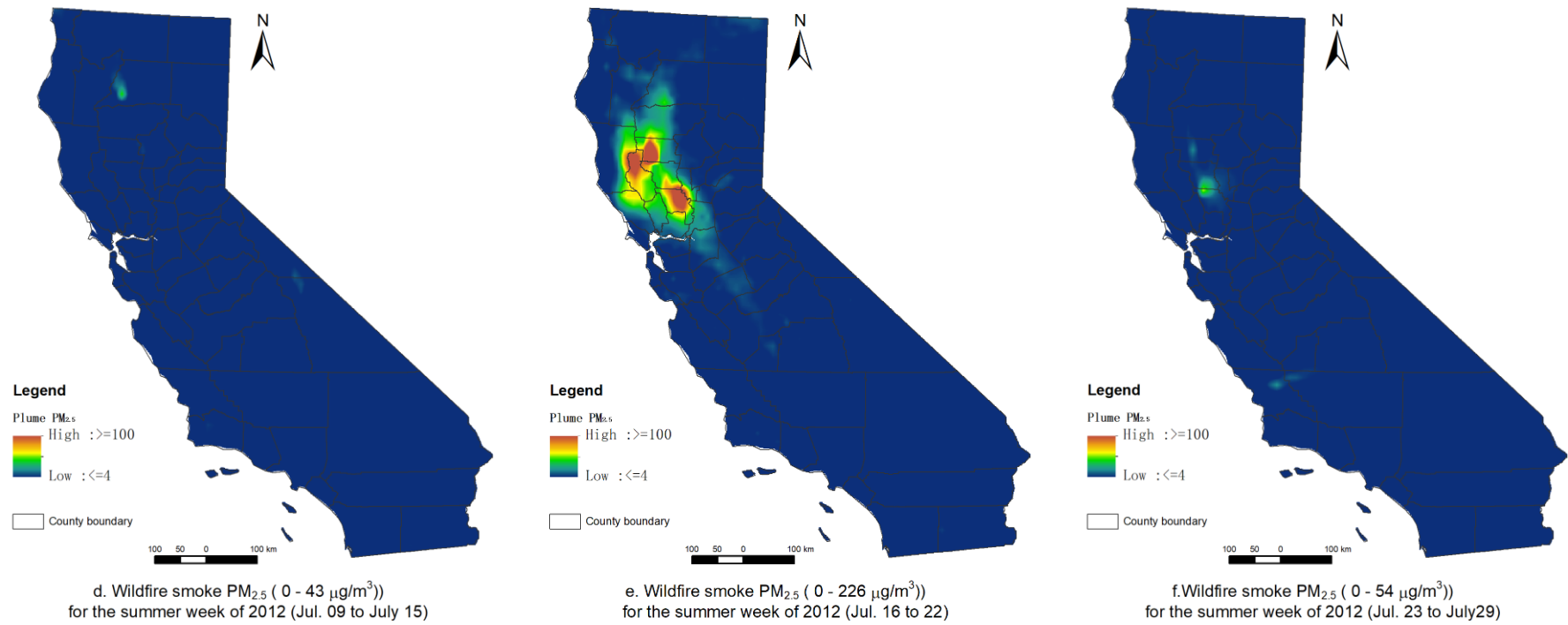
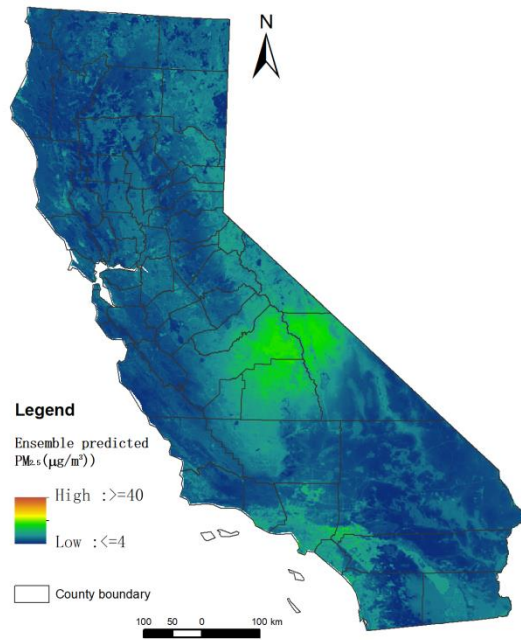
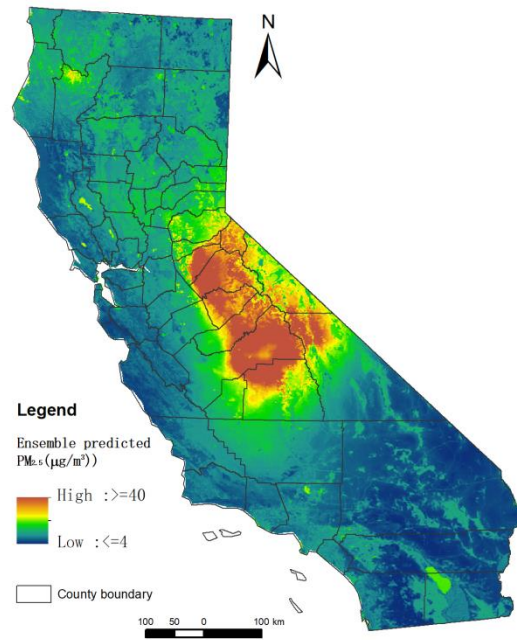


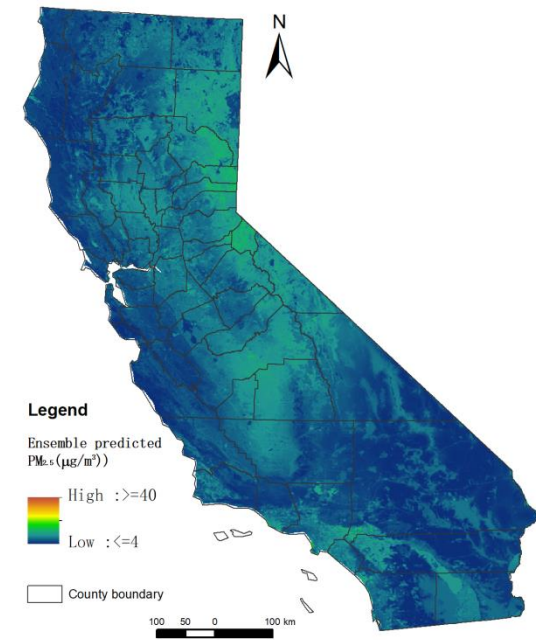
Fig. S15 Distributions of predicted PM_{2.5} surfaces (a, b and c) vs. HYSPLIT modeled wildfire smoke PM_{2.5} (d, e and f) for the week before (a and d) and after (c and f) the 2012 wildfire week (b and e).



a. Ensemble predicted PM_{2.5}
for the autumn week of 2015 (Sep. 07 to 13)



b. Ensemble predicted PM_{2.5}
for the autumn week of 2015 (Sep. 14 to 20)



c. Ensemble predicted PM_{2.5}
for the autumn week of 2015 (Sep. 21 to 27)

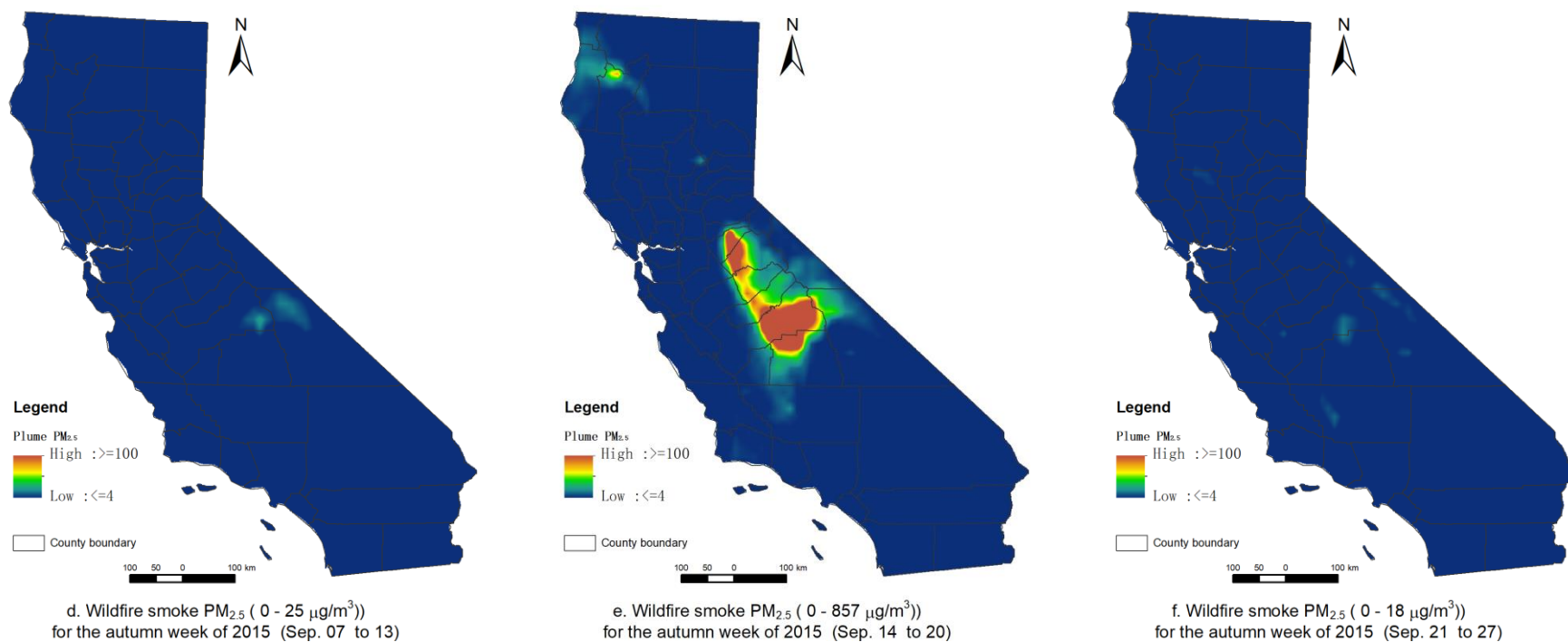


Fig. S16 Distributions of predicted PM_{2.5} surfaces (a, b and c) vs. HYSPLIT modeled wildfire smoke PM_{2.5} (d, e and f) for the week before (a and d) and after (c and f) the 2015 wildfire week (b and e).

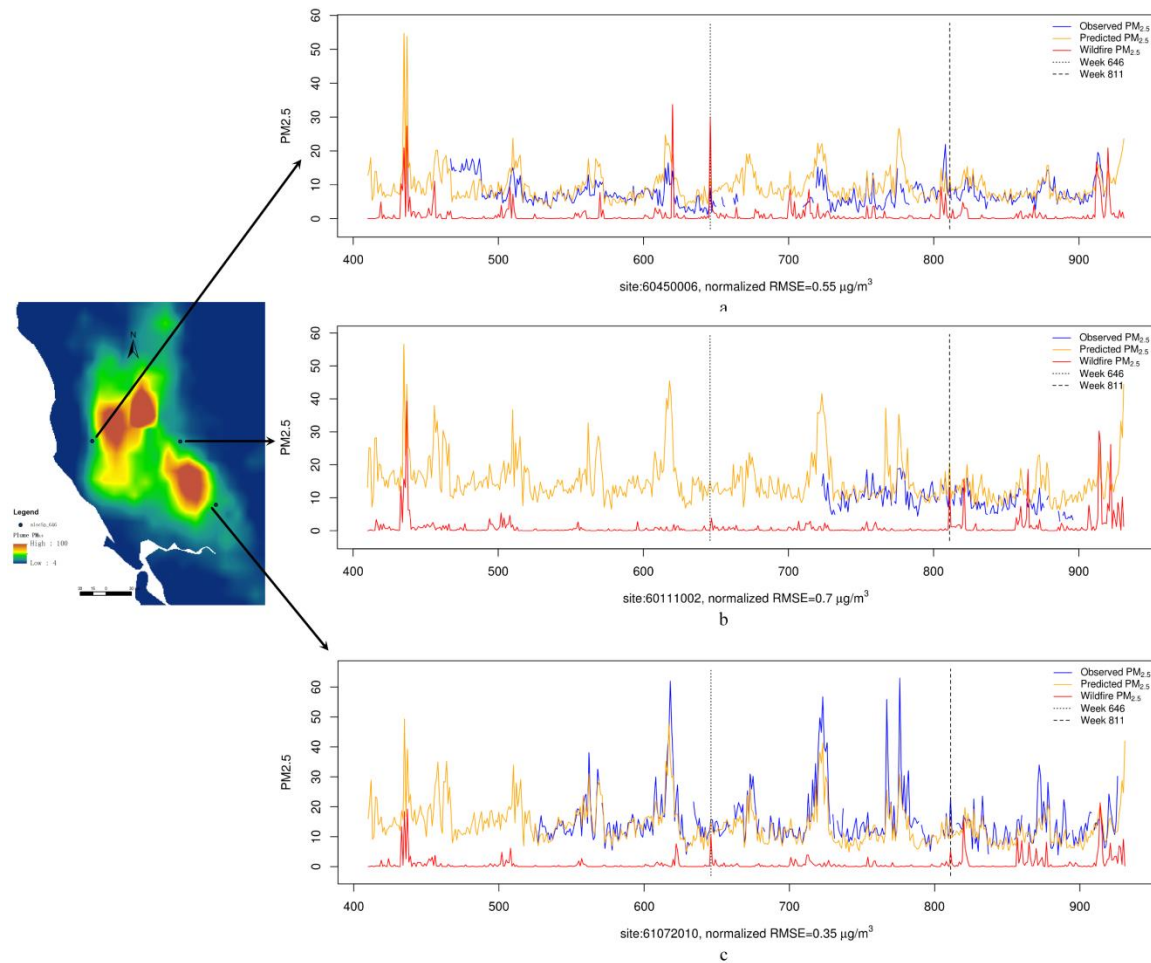


Fig. S17 Comparison of the time series of predicted PM_{2.5}, observed PM_{2.5}, and HYSPLIT modeled wildfire smoke PM_{2.5} at three AQS monitoring sites surrounding the wildfire region for the 2012 wildfire week.

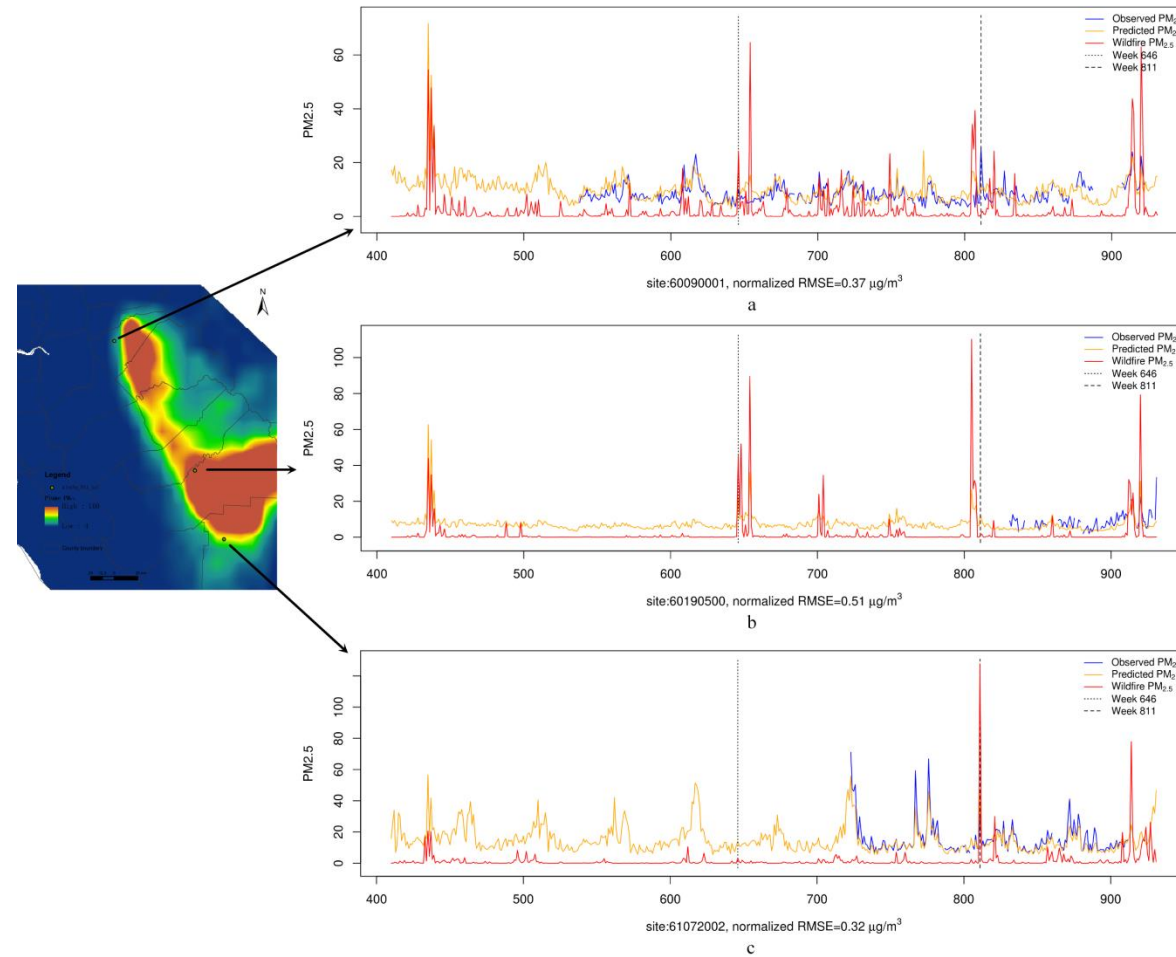


Fig. S18 Comparison of the time series of predicted PM_{2.5}, observed PM_{2.5}, and HYSPLIT modeled wildfire smoke PM_{2.5} at three AQS monitoring sites surrounding the wildfire region for the 2015 wildfire week.