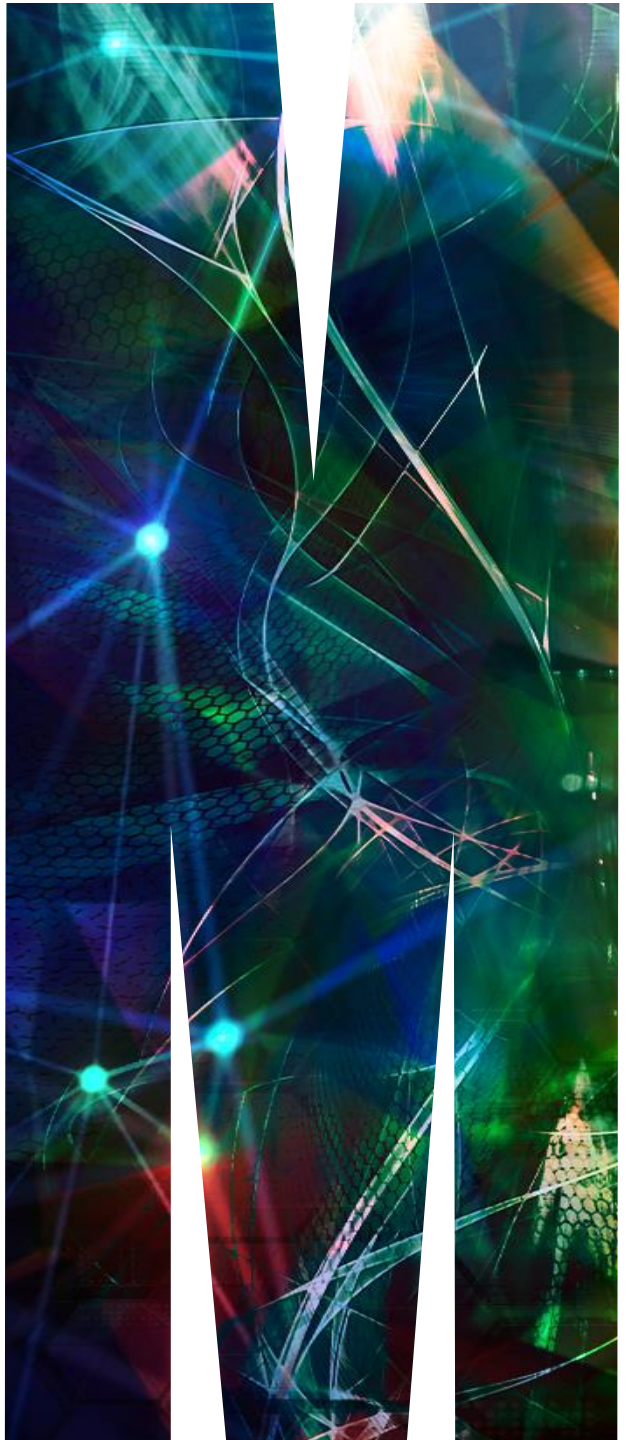
MONASH University

Satellite data for air quality exposure assessment and epidemiology

Yuming Guo, MD, PhD

Professor of Global Environmental Health and Biostatistics

Monash University School of Public Health and Preventive Medicine

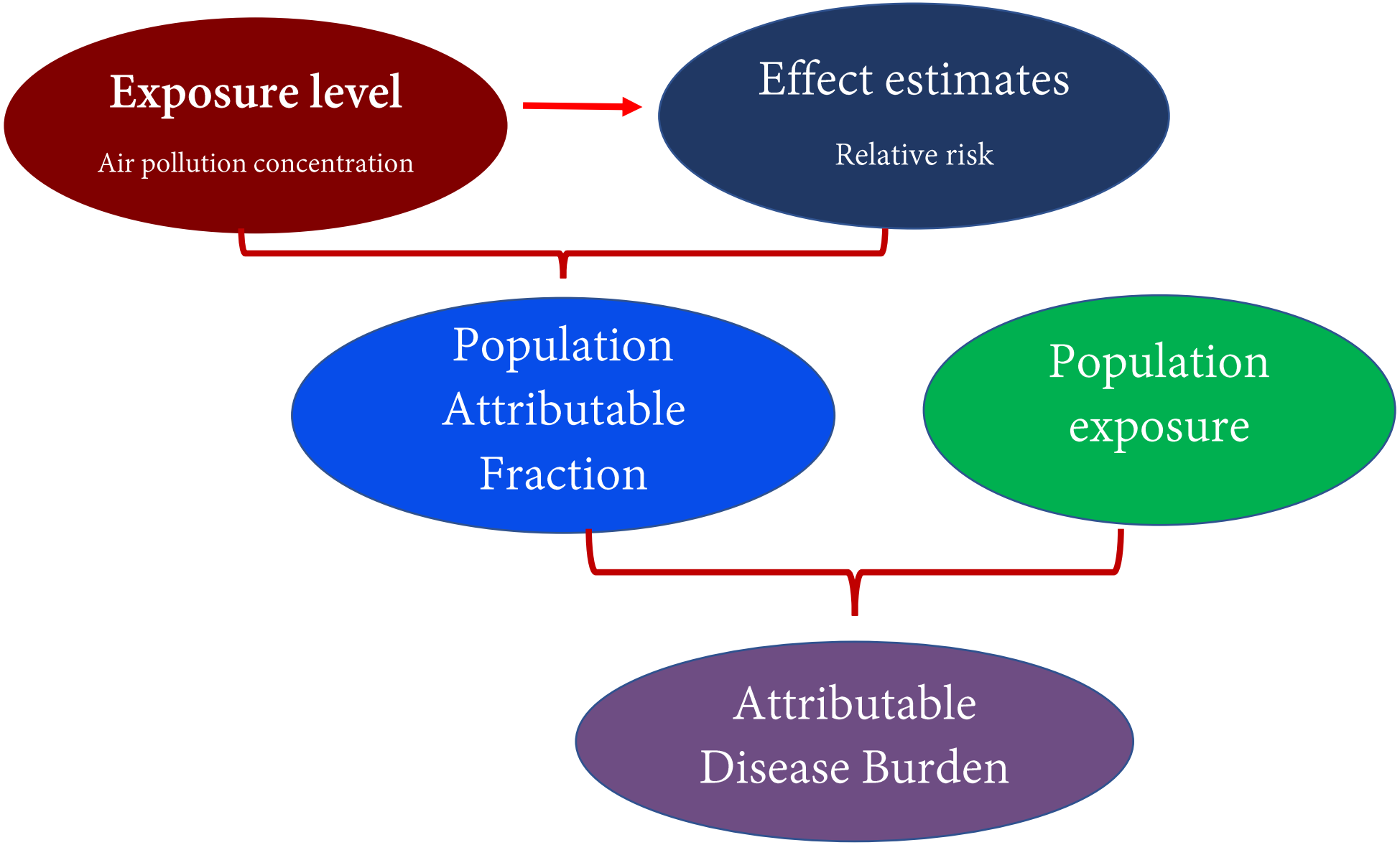


Outline

- Why perform air pollution exposure assessment
- Measurement error
- Challenges
- Machine learning
- The role of satellite data
- Wildfire smoke exposure assessment
- Opportunities

Why perform air pollution exposure assessment?

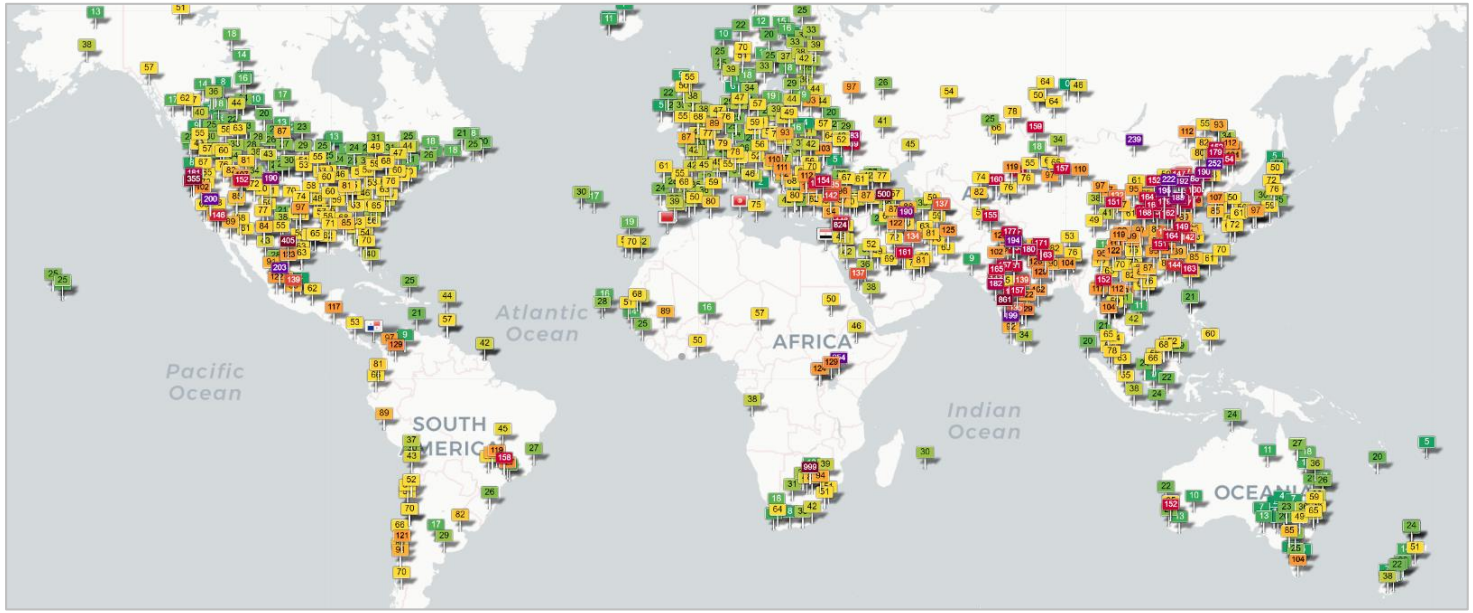
The role of air pollution data in environmental risk assessment

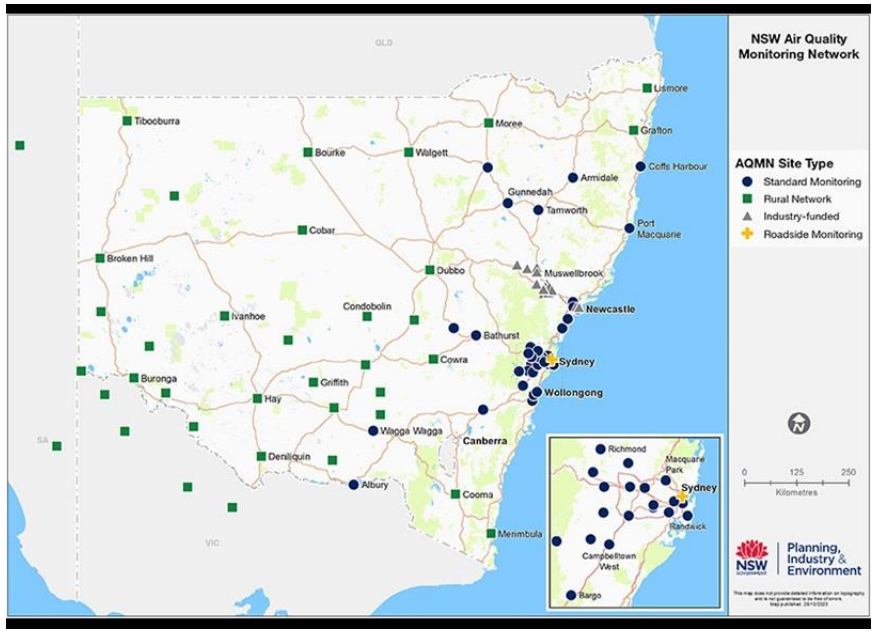


Why perform air pollution exposure assessment?

MONASH University

➢Limited exposure data, for both space scale and time scale.





What variables are involved in exposure assessment?

MONASH University

- Satellite data.
- Land use information (greenness, road density, urban cover..).
- Weather conditions (temperature, humidity, rainfall, wind speed..).
- Spatiotemporal correlations/trends.
- Others (wildfire, population density..).

Measurement error

Measurement error in air pollution exposure assessment

MONASH University

Estimating spatiotemporal distribution of PM_1 concentrations in China with satellite remote sensing, meteorology, and land use information☆

Gongbo Chen [a], Luke D. Knibbs [b], Wenyi Zhang [c], Shanshan Li [a], Wei Cao [d], Jianping Guo [e], Hongyan Ren [d], Boguang Wang [f], Hao Wang [g], Gail Williams [b], N.A.S. Hamm [h], Yuming Guo [a, *]

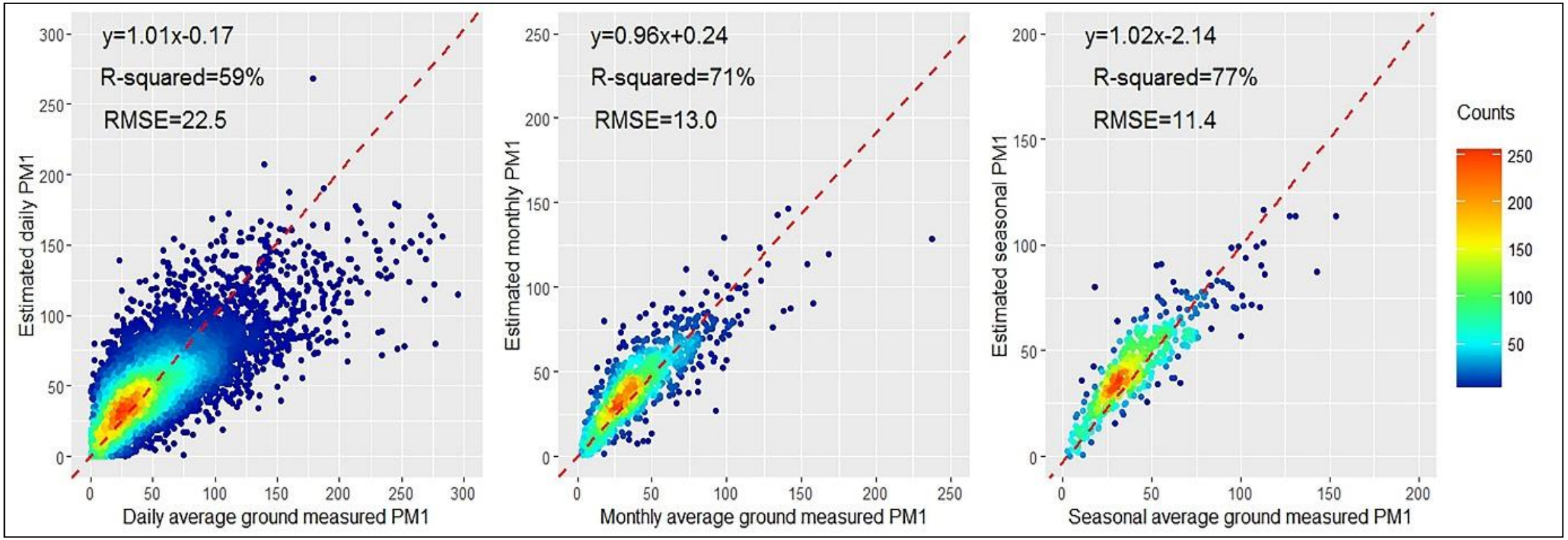
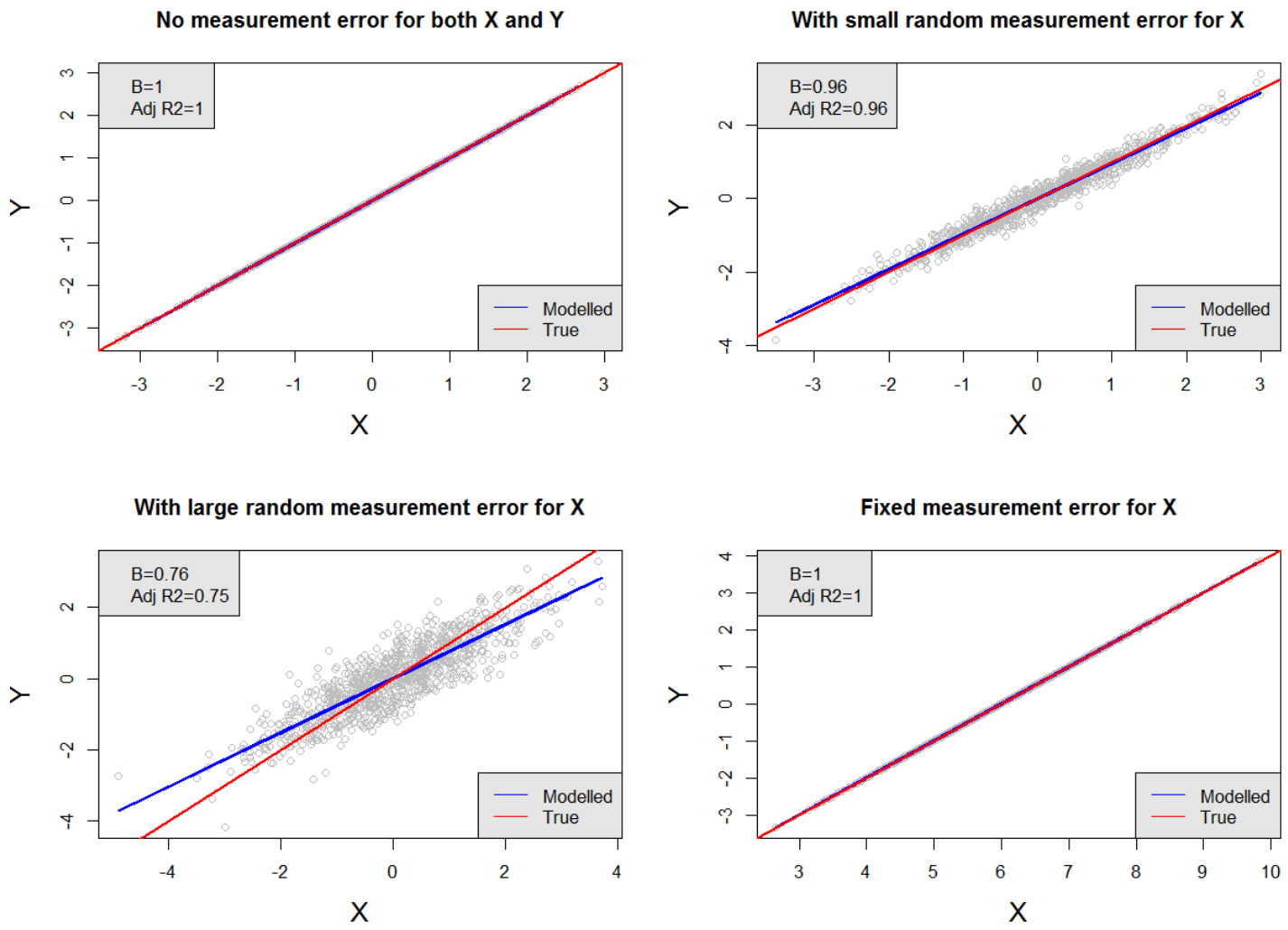


Figure . Scatterplots of 10-fold cross-validation for daily, monthly and seasonal estimation of PM1 concentrations (μg/m3)

Measurement error in environmental risk assessment



The relationship between simulated exposure (X) and response outcome (Y)

- Random measurement error lead to underestimated effect estimates.
- Does this mean random measurement error lead to underestimated burden of disease attributable to the exposure?

MONASH University

The more accurate exposure assessment, the more accurate effect estimates and the more accurate estimation of disease burden !

Challenges

Challenges for air pollution exposure assessment

MONASH University

- Missing values in satellite data.

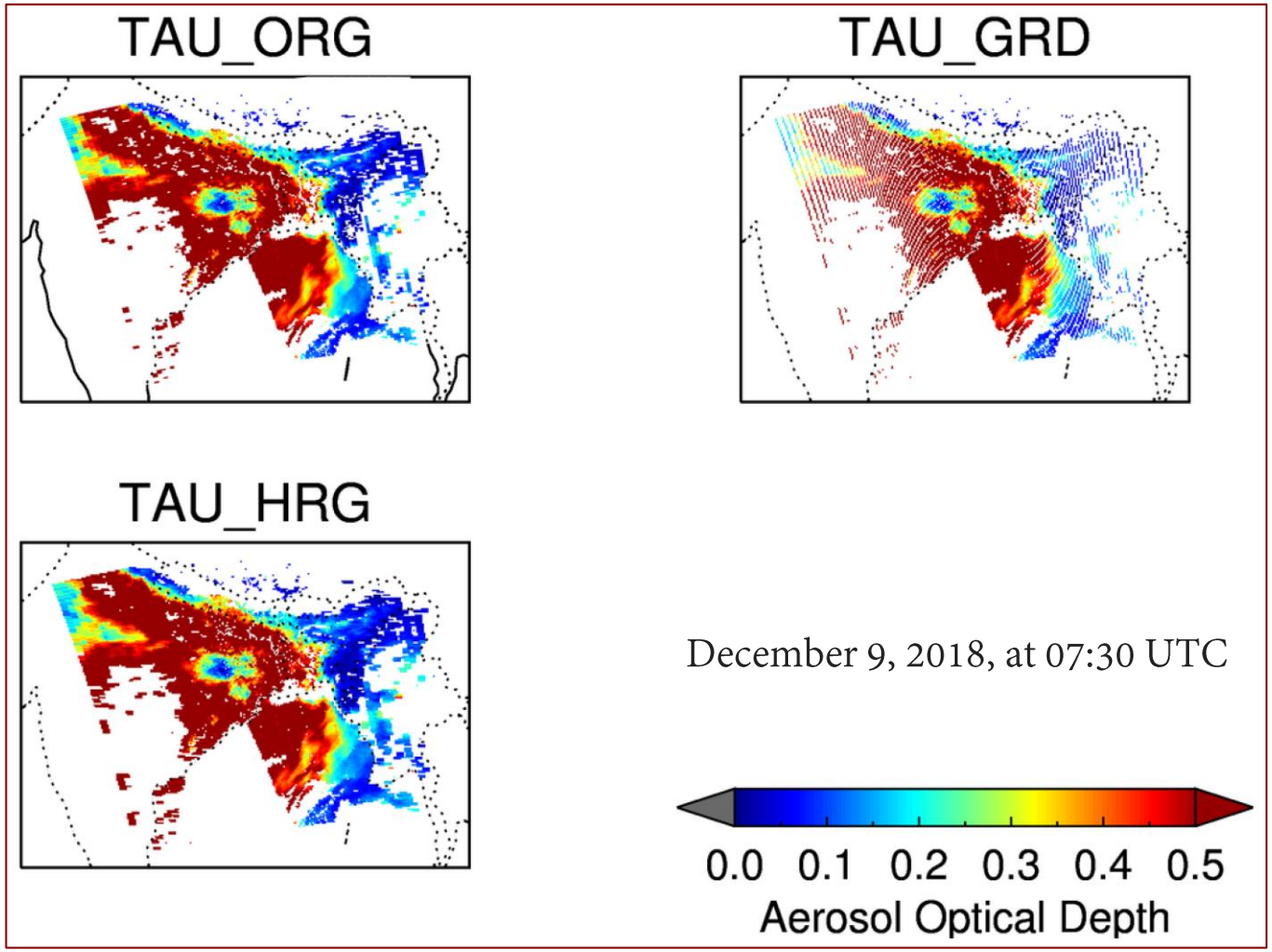


Figure . Aerosol optical depth (550 nm) from Moderate Resolution Imaging Spectroradiometer (MODIS)-Aqua over Asia.

The aerosol optical depth (AOD) data are mapped to show: TAU_ORG: original data considering varying pixel size; TAU_GRD: high resolution gridded data with no filling of empty grids; and TAU_HRG: high resolution gridded data with spatial filling at the edge of the swath.

Gupta et al. High-Resolution Gridded Level 3 Aerosol Optical Depth Data from MODIS. Remote Sensing.

Challenges for air pollution exposure assessment

MONASH University

➢ Filling missing values of satellite data.

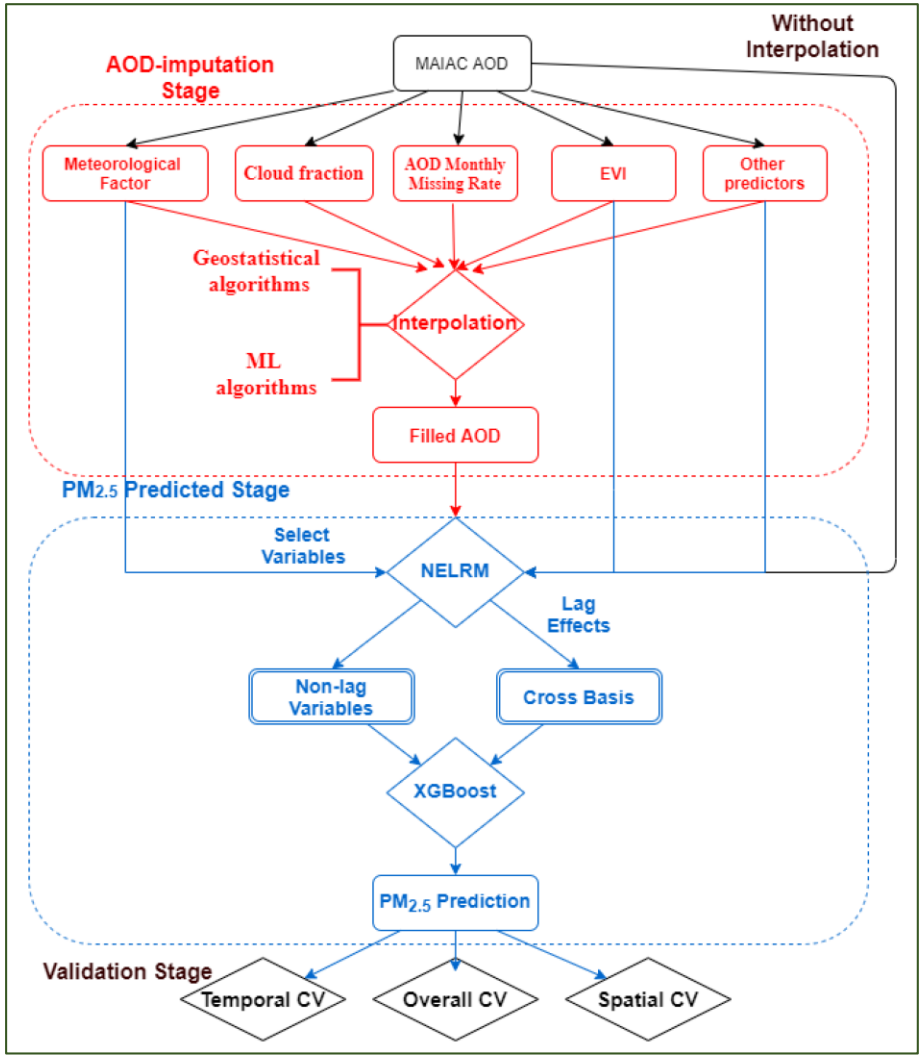


Table 2. Comparison of geostatistical and machine learning (ML) interpolation for MAIAC AOD product.

Interpolation Method	Coverage (%)	Computation Time [a]	CV RMSE	CV R^2	CV MAPE (%)
Before Interpolation	15.46			\	
Geostatistical Algorithms					
TS	87.31	55:38:56.56	0.17	0.75	20.56
ST Kriging	67.73	128:56:43.56	0.17	0.78	20.36
IDW	45.22	45:35:28.39	0.18	0.65	21.35
Kriging	42.37	88:46:37.24	0.17	0.66	20.89
NN	21.43	15:39:27.93	0.19	0.49	25.38
ML Algorithms					
RF	98.64	120:55:28.65	0.15	0.89	18.00
XG	98.64	18:00:38.20	0.15	0.85	19.06
SVM	98.64	19:04:47.64	0.17	0.72	19.41
BRNN	98.64	18:45:36.22	0.17	0.70	22.39
GBM	98.64	06:35:47.65	0.18	0.69	25.17
GAM	98.64	01:05:38.20	0.17	0.62	21.88
LASSO	98.64	01:18:30.23	0.19	0.49	28.03

TS (two-step interpolation), IDW (inverse distance weighting), ST Kriging (spatio-temporal kriging), NN (nearest neighbors with 30 km buffer), RF (Conditional Inference Random Forest), XG (extreme gradient boosting, XGBoost), SVM (support vector machine), BRNN (Bayesian regularized neural network), GBM (gradient boost model), GAM (generalized additive model), and LASSO (least absolute shrinkage and selection operator); [a] tested by Windows 10 system in 3.4 GHz computer with 16 GB of RAM.

Chen et al. Comparison of Different Missing-Imputation Methods for MAIAC (Multiangle Implementation of Atmospheric Correction) AOD in Estimating Daily PM2.5 Levels. Remote Sensing.

Machine learning

Which models are usually used for exposure assessment?

MONASH University

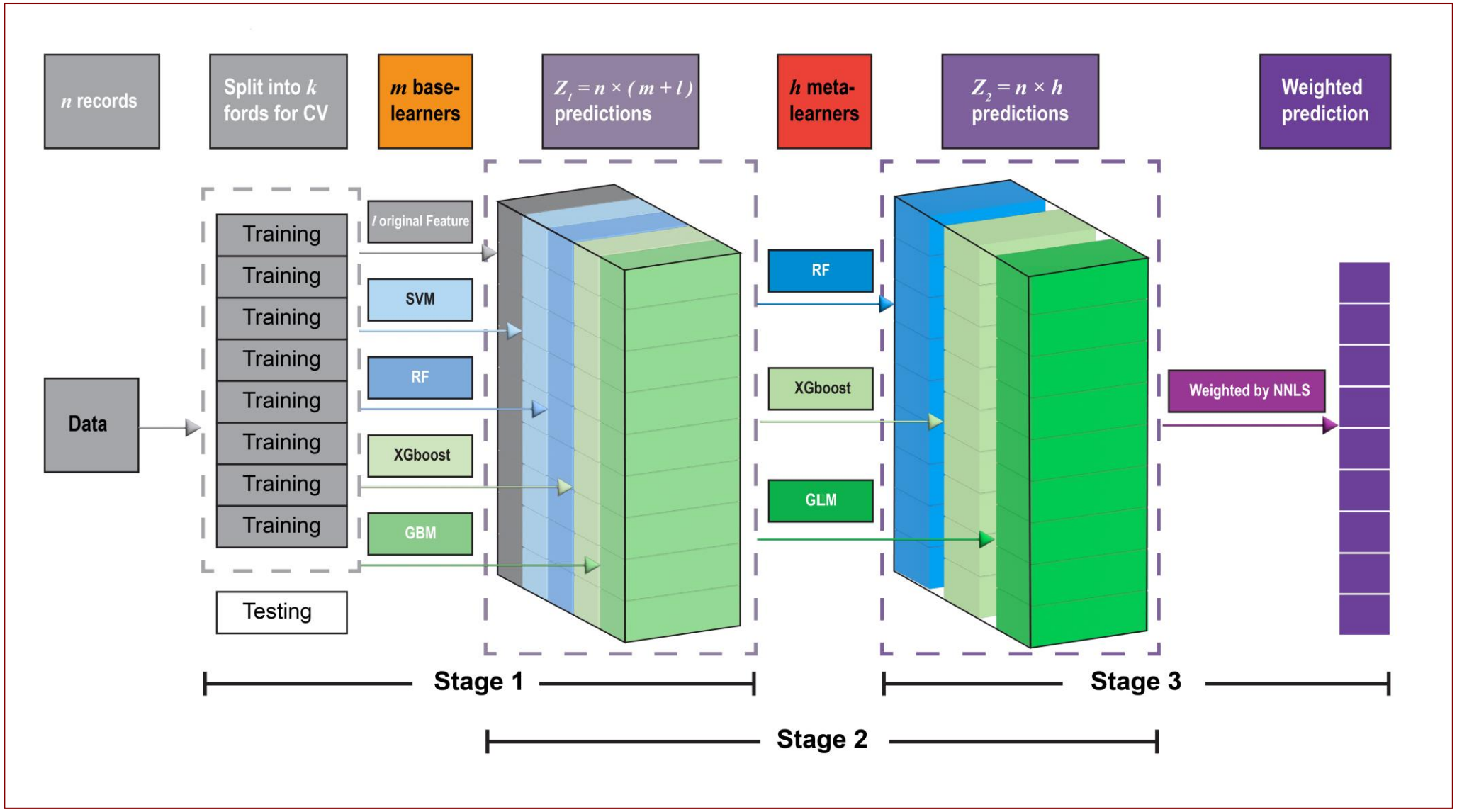
- Linear regression.
- Generalized linear regression or generalized additive regression.
- Mixed effect model.
- Bayesian spatiotemporal model.
- Geospatial model, e.g., Kriging, inverse distance weighting.
- Chemical transport model.
- Machine learning and deep learning

Machine learning for air pollution exposure assessment

MONASH University

- How to choose machine learning models? Which one is best?
- How to apply machine learning in environmental exposure assessment easily?
- Do ensemble machine learning models perform better than single machine learning model?

Deep Ensemble Machine Learning Model



Yu et al. Deep Ensemble Machine Learning Framework for the Estimation of PM2.5 Concentrations. Environmental Health Perspectives.

Deep Ensemble Machine Learning Model

MONASH University

Table 2. $PM_{2.5}$ prediction performances of DEML model and five benchmark models from 2015 to 2019 in Italy.

Year	Measurement	GBM	SVM	RF	XGBoost	SL[a]	DEML[b]
2015	R^2	0.69	0.79	0.85	0.81	0.85	0.89
	RMS E ($\mu g/m^3$)	9.25	6.42	6.49	7.23	6.47	5.54
2016	R^2	0.72	0.80	0.84	0.81	0.84	0.87
	RMSE ($\mu g/m^3$)	7.74	6.51	5.84	6.33	5.82	5.18
2017	R^2	0.74	0.81	0.85	0.81	0.85	0.89
	RMSE ($\mu g/m^3$)	8.20	7.19	6.41	7.09	6.38	5.37
2018	R^2	0.70	0.78	0.86	0.82	0.86	0.89
	RMSE ($\mu g/m^3$)	7.44	6.22	5.18	5.69	5.13	4.43
2019	R^2	0.68	0.76	0.84	0.79	0.84	0.87
	RMSE ($\mu g/m^3$)	7.34	6.42	5.13	5.78	5.12	4.55
Total	R^2	0.51	0.76	0.83	0.70	0.83	0.87
	RMSE ($\mu g/m^3$)	10.4	7.42	6.23	8.20	6.23	5.38

Yu et al. Deep Ensemble Machine Learning Framework for the Estimation of PM2.5 Concentrations. Environmental Health Perspectives.

Deep Ensemble Machine Learning Model

MONASH University

- The advantages:
  1. Have high prediction performance
  2. Avoid over-fitting through cross-validation analysis
  3. Set the optimal non-negative weight for each base-model/meta-model
  4. Minimizes the extent to the empirical experience in select models
  5. Assessed and compared models' results directly

- The disadvantage:
  1. Be cautious to select features
  2. Be sensitive to missing values
  3. Need more time to run big data

The role of satellite data

The role of satellite data in air pollution exposure assessment

MONASH University

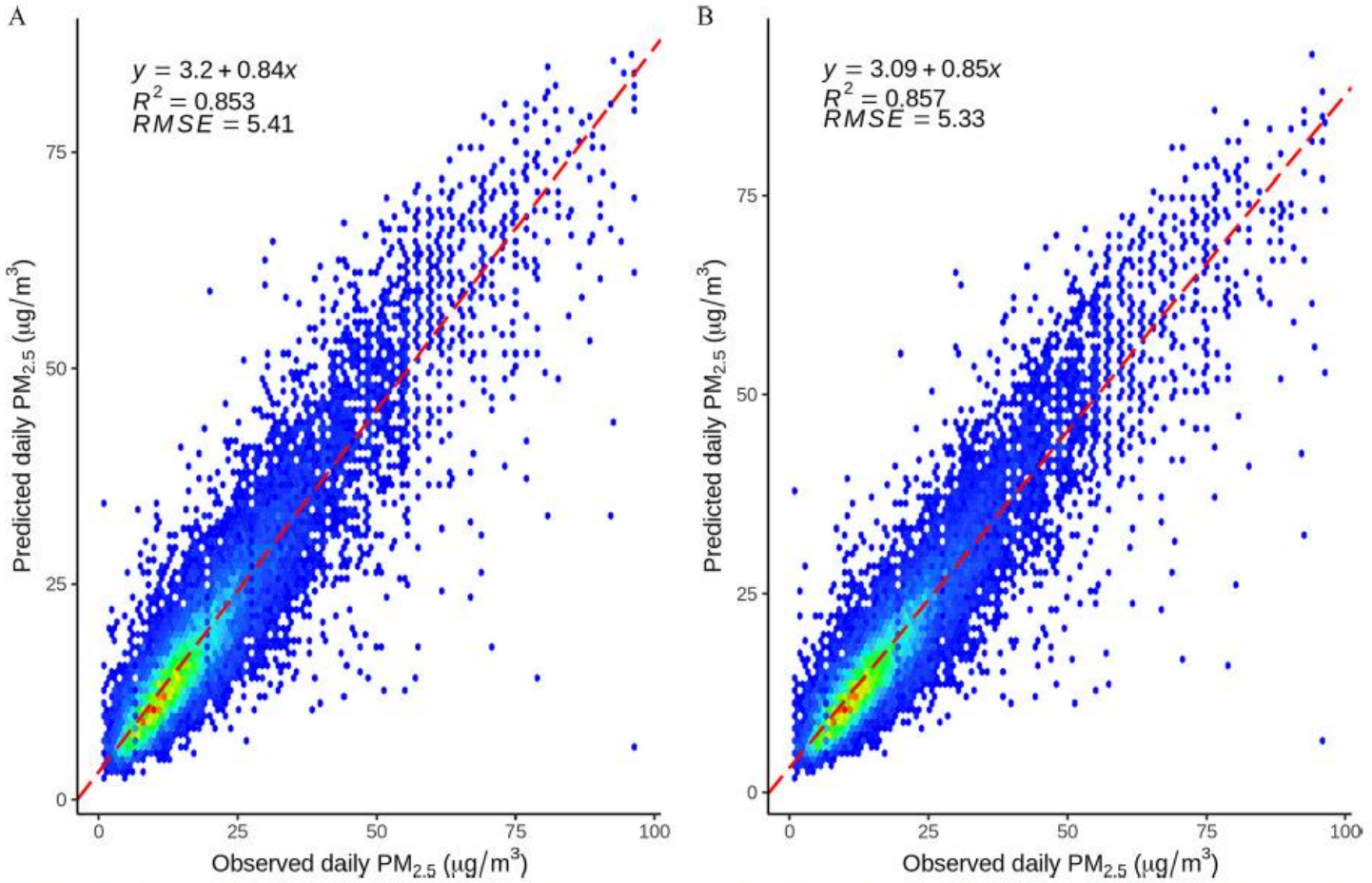


Figure 4. The $PM_{2.5}$ prediction performance of the DEML models with and without AOD as a predictor from 2015–2019 in Italy. The *x*-axis indicates the observed daily $PM_{2.5}$ in the monitor stations; the *y*-axis indicates the estimated $PM_{2.5}$ by the DEML model. The points represent the corresponding $PM_{2.5}$ for both observed and predicted values. The solid line represents a regression line for the observed and predicted $PM_{2.5}$ by using the simple linear regression. R^2 is the coefficients of determination for the unseen independent data. (A) The DEML prediction model including AOD. (B) The DEML prediction model without AOD. Note: DEML, the three-stage stacked deep ensemble machine learning method; $PM_{2.5}$, particulate matter with aerodynamic diameter <2.5 μm; RMSE, the root mean square error (micrograms per cubic meter).

Yu et al. Deep Ensemble Machine Learning Framework for the Estimation of PM2.5 Concentrations. Environmental Health Perspectives.

The role of satellite data in air pollution exposure assessment

MONASH University

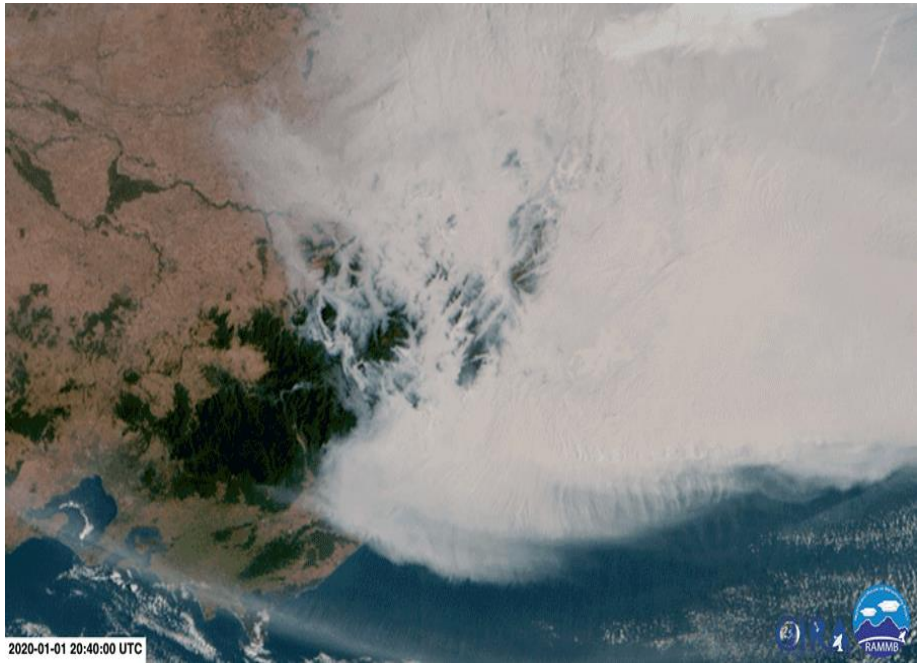
- In following scenario, we might not need satellite data:
  1. Have enough observed air pollution data and predictors (correlated with satellite data) in a specific region; and
  2. Don't predict air pollution in the locations far away from the training region; and
  3. Don't predict air pollution in the period outside the training period.

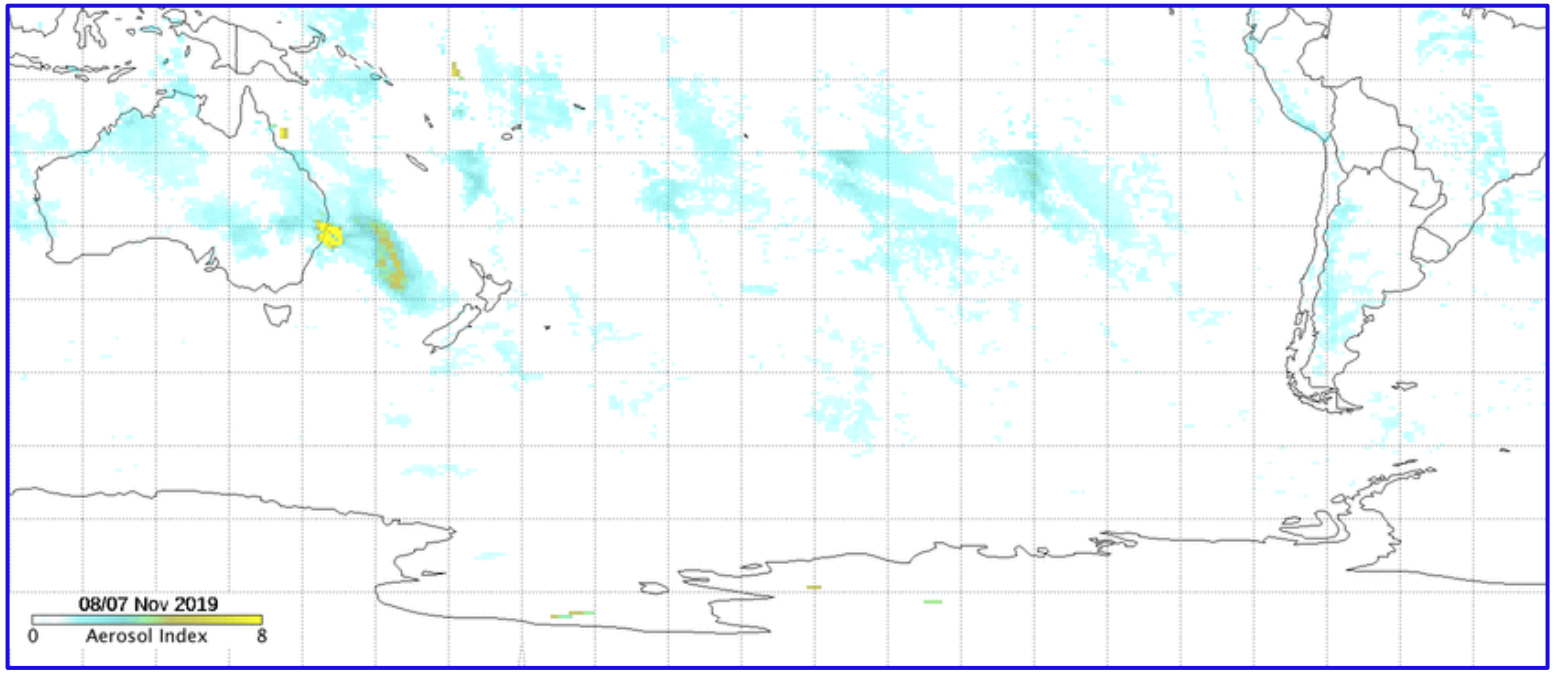
- Correspondingly, in following scenarios, we need satellite data:
  1. Have limited observed air pollution data and predictors in a specific region; or
  2. Predict air pollution in the locations far away from the training region; or
  3. Predict air pollution in the period outside the training period.

Wildfire smoke exposure assessment

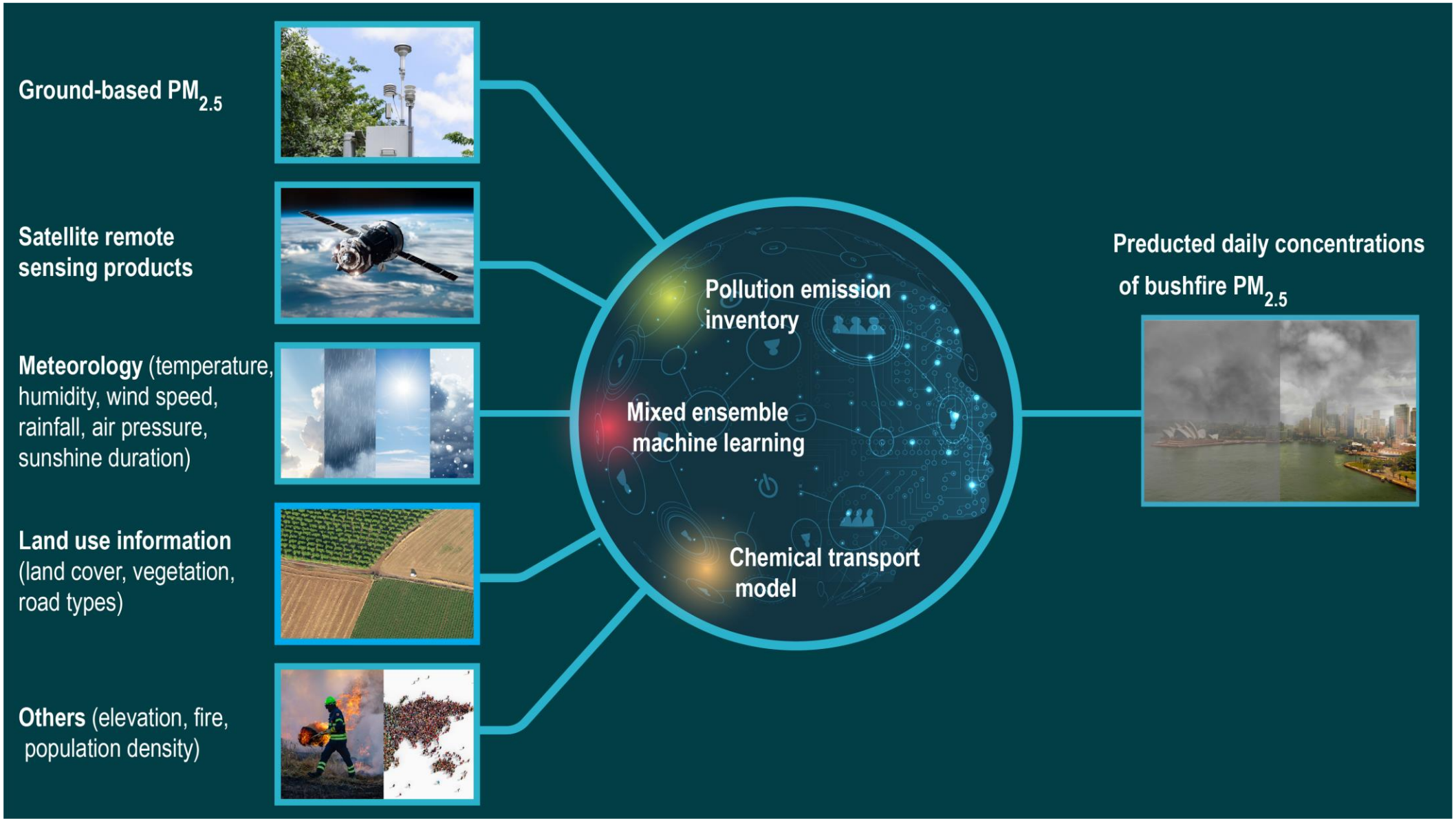
Wildfire smoke exposure assessment

MONASH University

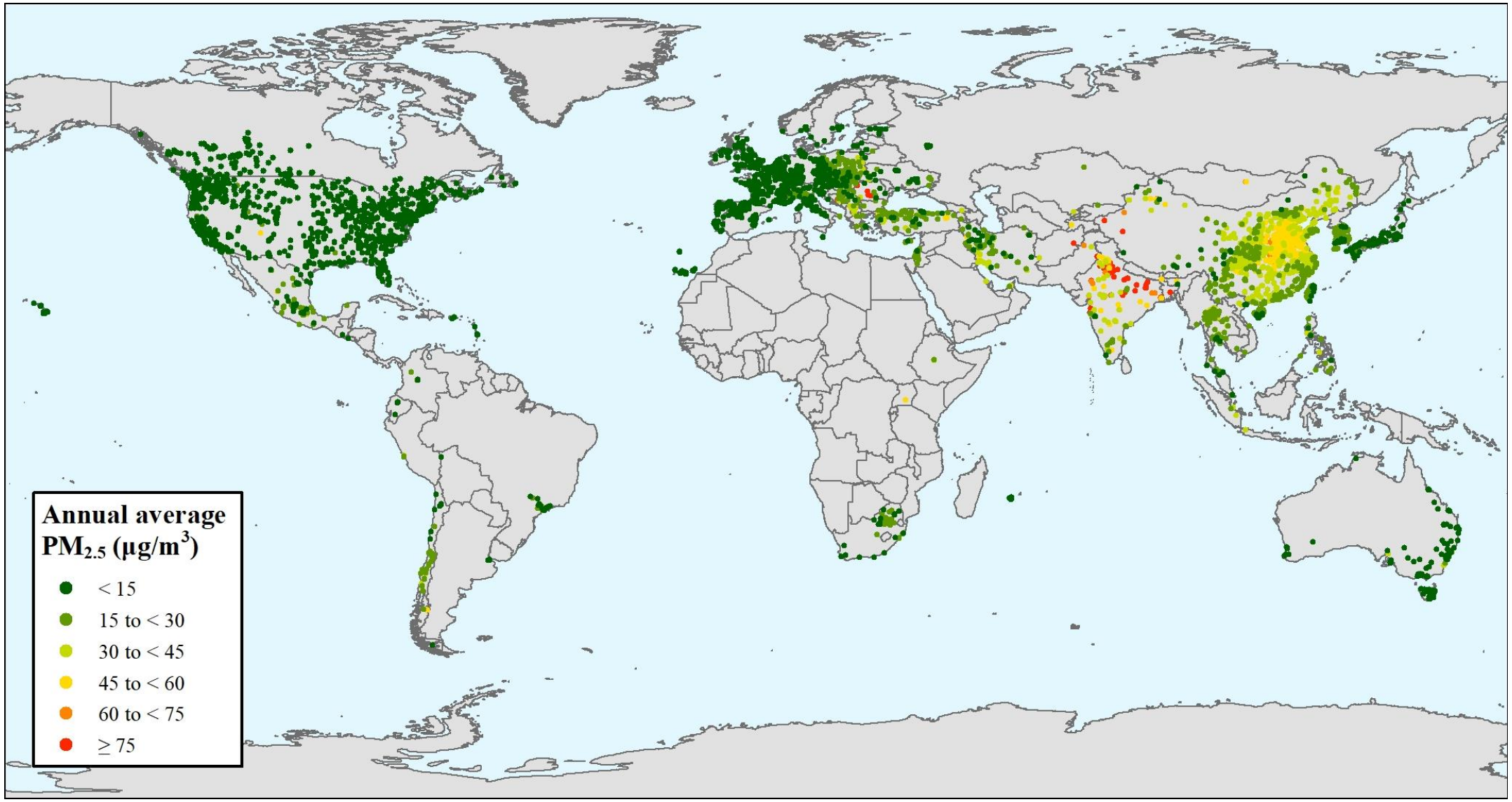




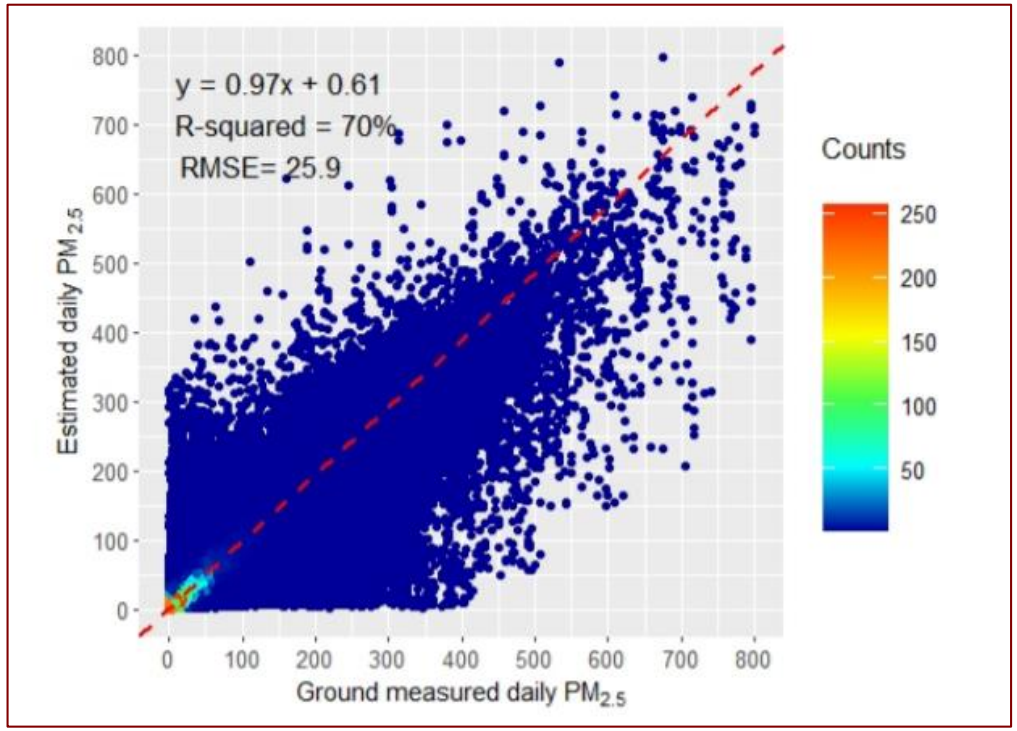
Wildfire smoke exposure assessment

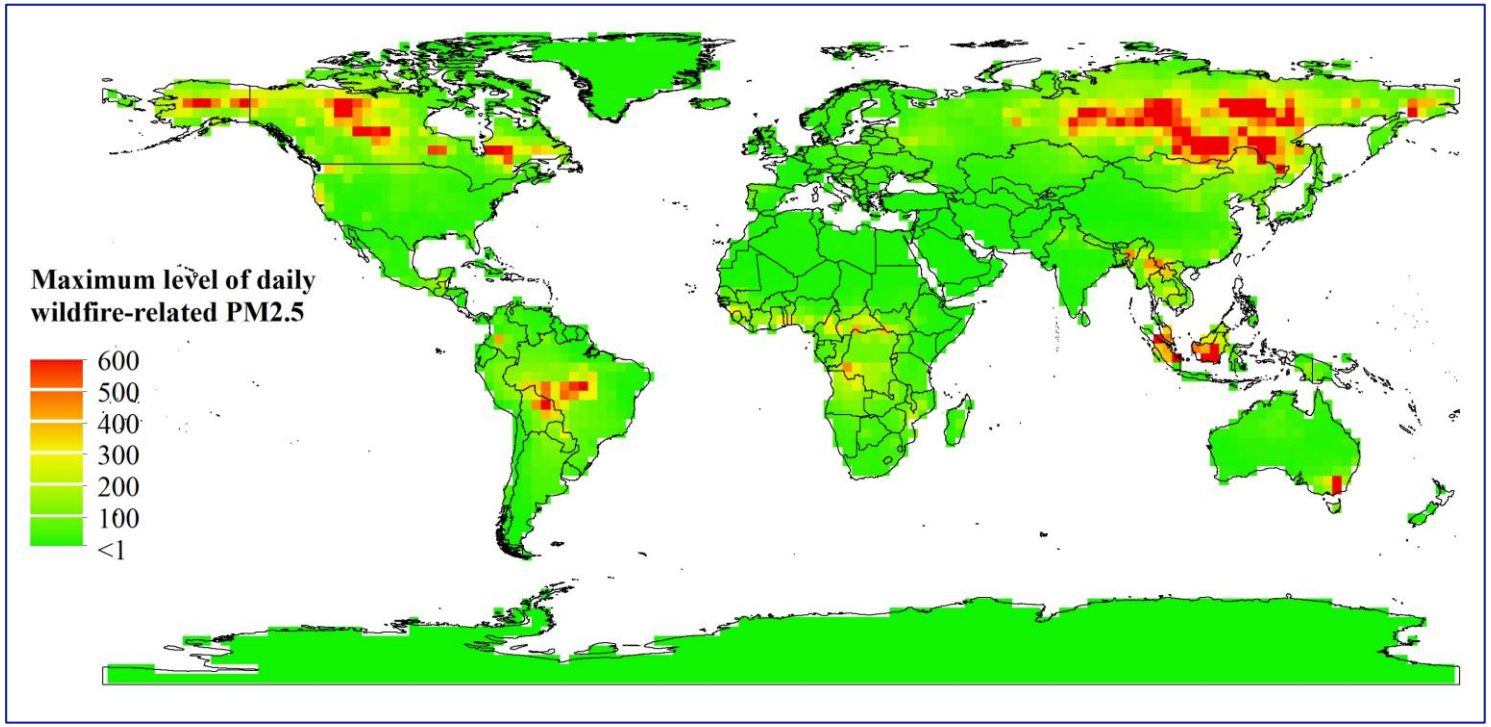


Wildfire smoke exposure assessment



Estimating daily wildfire-related $PM_{2.5}$ (µg/m3) 2000-2016

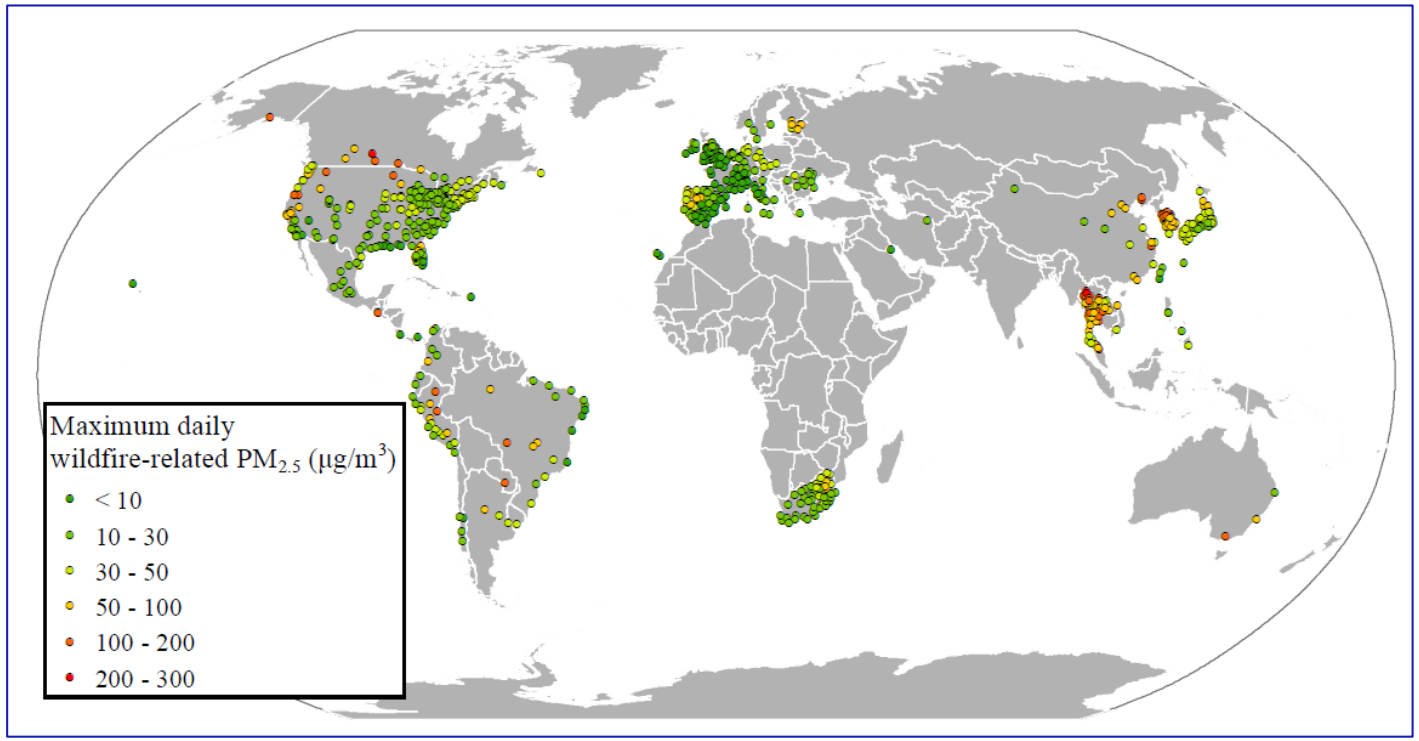


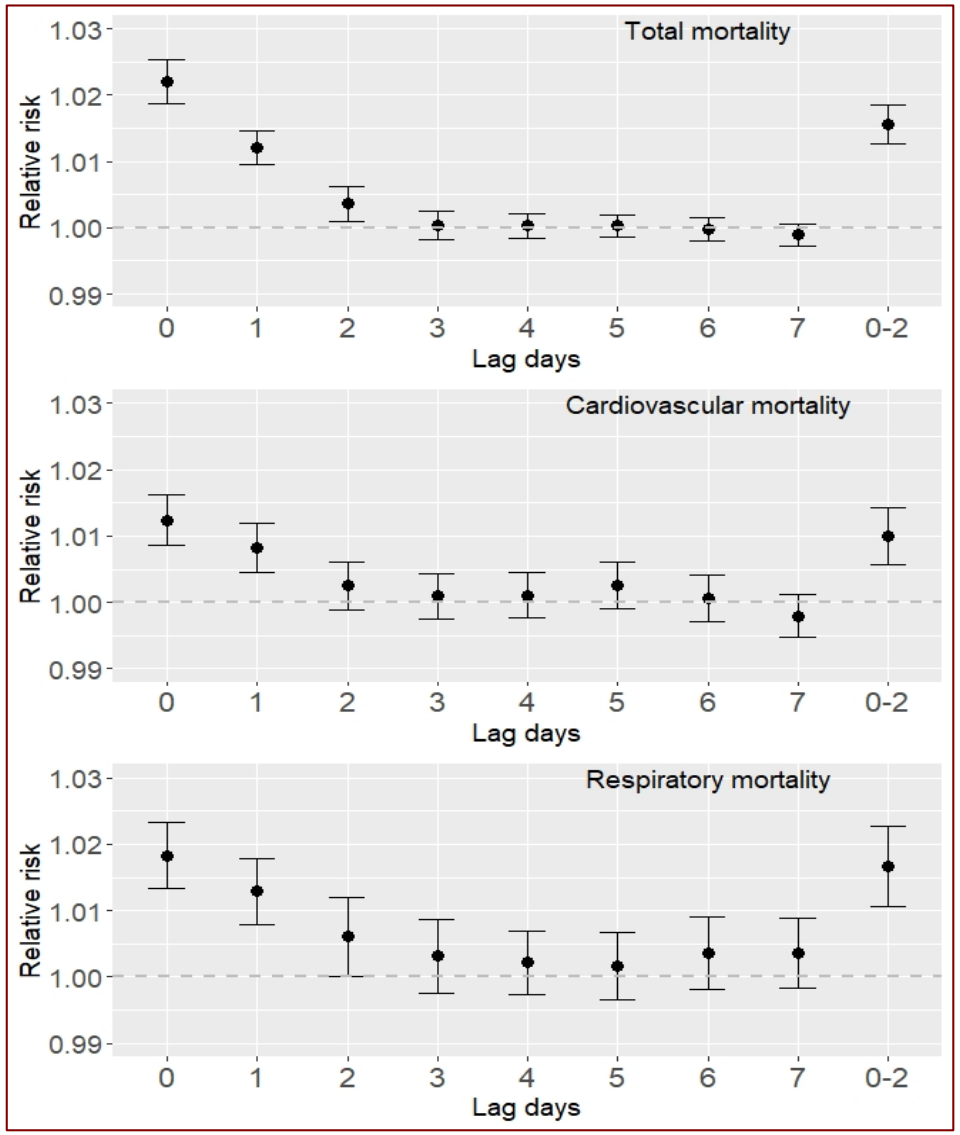


Wildfire-specific $PM_{2.5}$

MONASH University

Wildfire-related PM2.5 and mortality risks





Chen et al. Mortality risk attributable to wildfire-related PM2.5 pollution: a global time series study in 749 locations. The Lancet Planetary Health.

Challenges in estimating wildfire smoke concentrations

MONASH University

1. Many studies just simply use total concentrations of air pollutants during fire period as wildfire-related concentrations.

2. Low spatial resolution.

3. It is hard to directly validate wildfire smoke concentration with the measured air pollution data.

Opportunities

Opportunities for air pollution exposure assessment

MONASH University

- Available big data (observed data including those from low cost sensors, remote sensing, weather data), makes it possible to perform accurate prediction.
- Machine learning /deep learning technologies provides better predication performance than traditional models.
- High performance computer / and cloud analysis are available to perform big data analysis.

Thank you !


Improving the health of populations
in a changing and inequitable world