MONASH University

Deep ensemble machine learning for estimating environmental exposure and beyond

Yuming Guo, MD, PhD

Professor of Global Environmental Health and Biostatistics

Monash University School of Public Health and Preventive Medicine

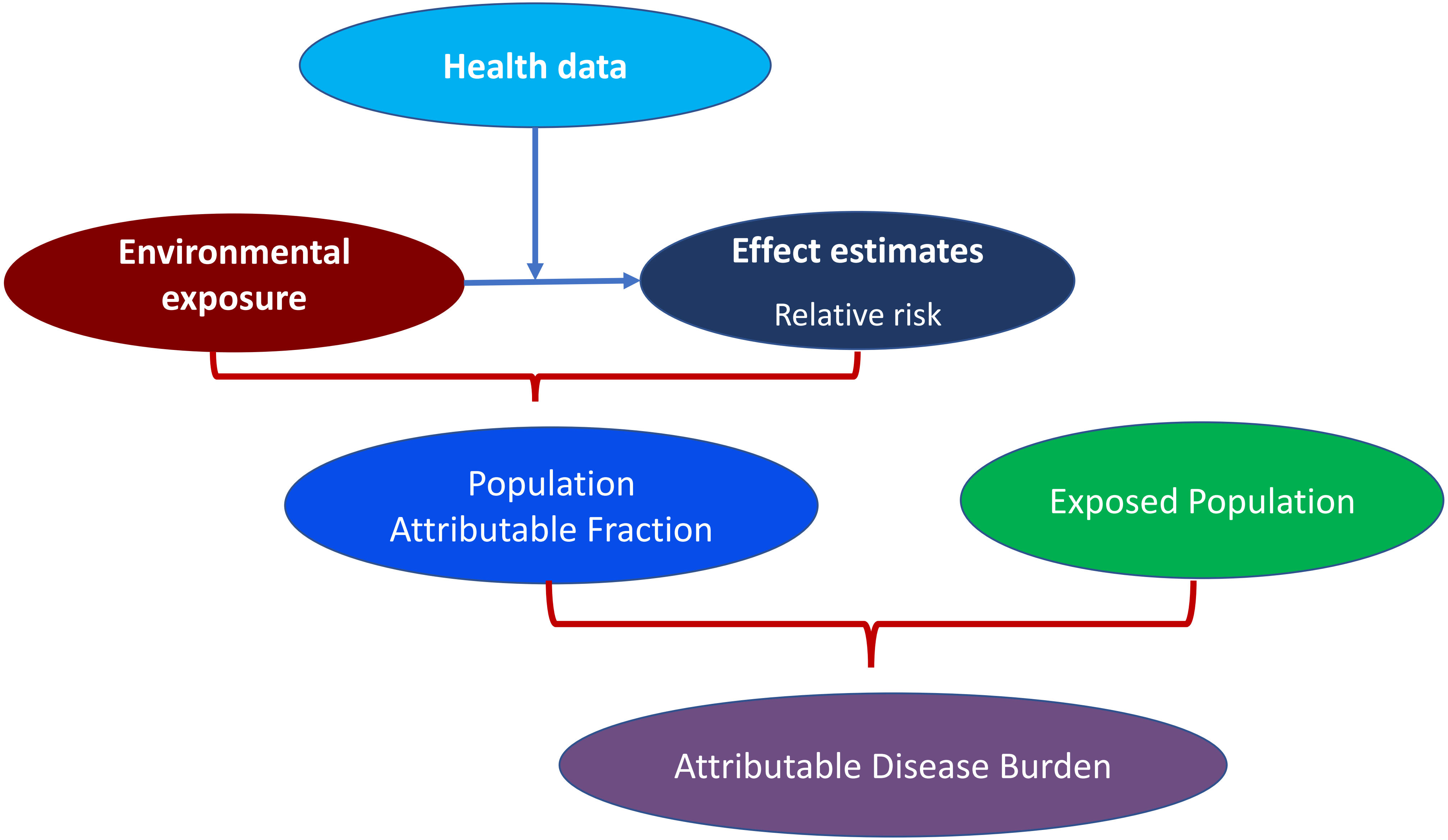
Outline

MONASH University

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

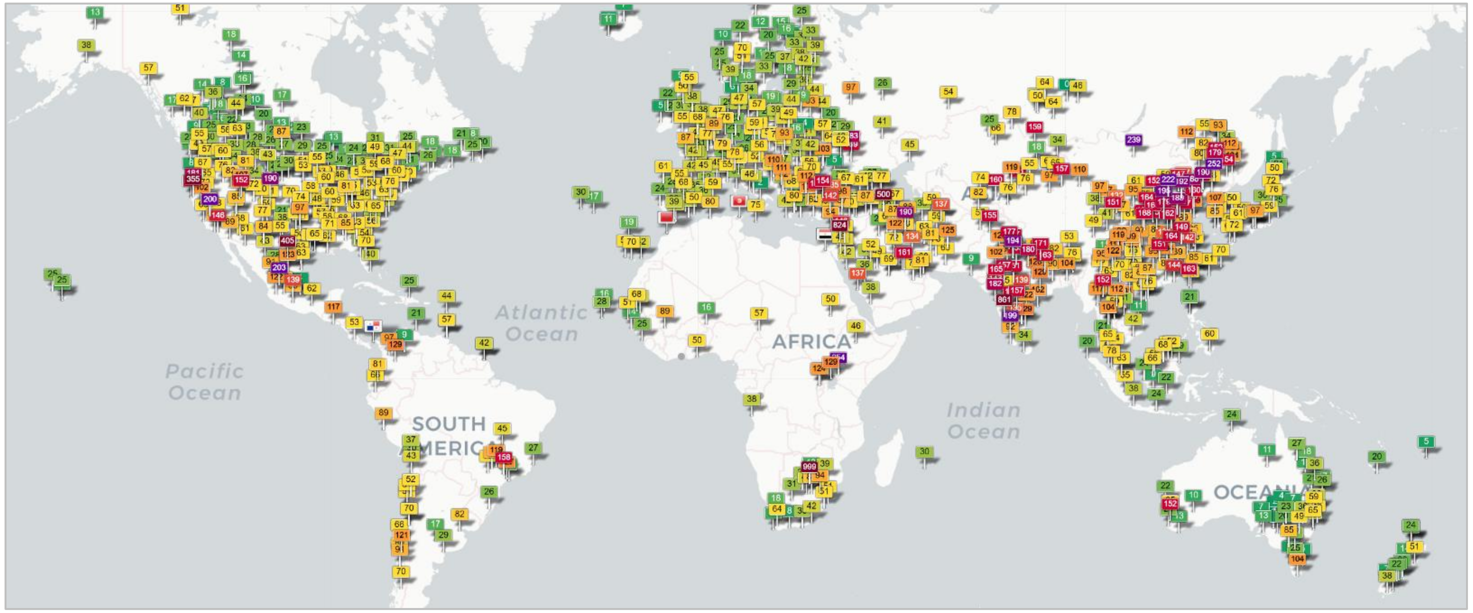
Why perform environmental exposure assessment?

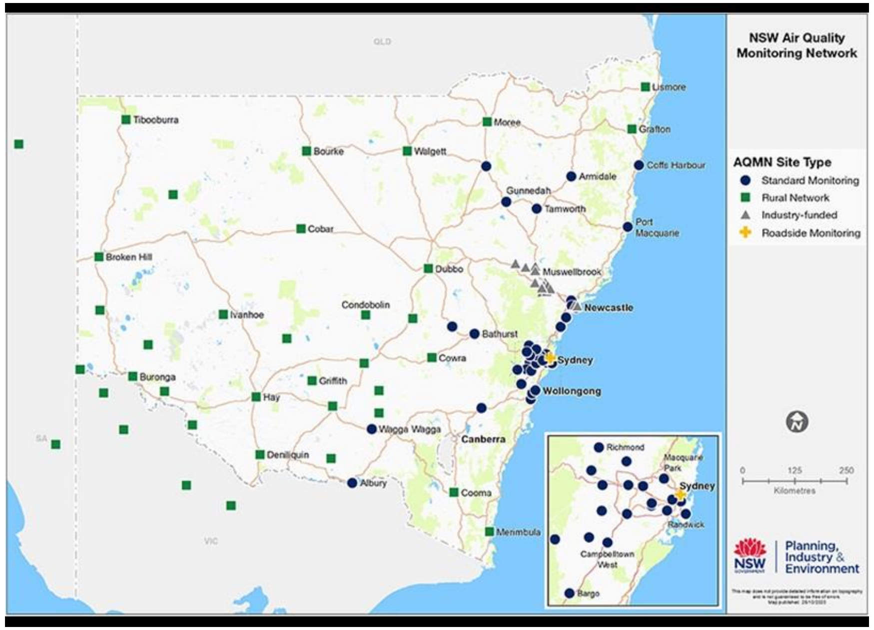
The role of environmental data in environmental risk assessment



Why perform exposure assessment?

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





What variables are involved in exposure assessment?

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

Which models are usually used for exposure assessment?

- 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 (e.g., random forest, xGBoost) and deep learning

The theories underline the prediction model

Regression model (including GLM, GAM, Bayesian)

- Y = X1 + X2 +X3 +X4………….. (Linear model)
- Y = s(X1) + s(X2) + s(X3) + s (X4) …….. (Non-linear or mixed model)
- Y = X1*X2 + X1*X3 + X1*X4 + X2 + X3 + X4………. (Linear Interaction)
- Y= s(X1)*s(X2) + s(X1)*s(X3) + X1*X4 + X2 + X3 + X4 ……………….. (non-linear interaction)

Geospatial model

- Y= s(latitude, longitude). s can be a function of weighting, spline, et c.

Chemical transport model

- Based on emission inventory, chemical species, meteorological factors, and circulation models.

Machine learning model

- Y=f(X1, X2, X3, X4, …….)

Model validation

Why perform model validation?

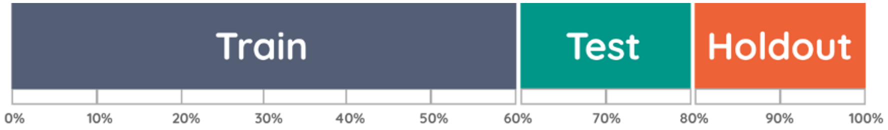
Model validation is the process of evaluating a trained model on test data set. Model validation provides the generalization ability of a trained model.

- Increase generalizability and flexibility
- Enhance the model quality
- Discover more errors
- Prevents overfitting and underfitting.

Model validation strategies

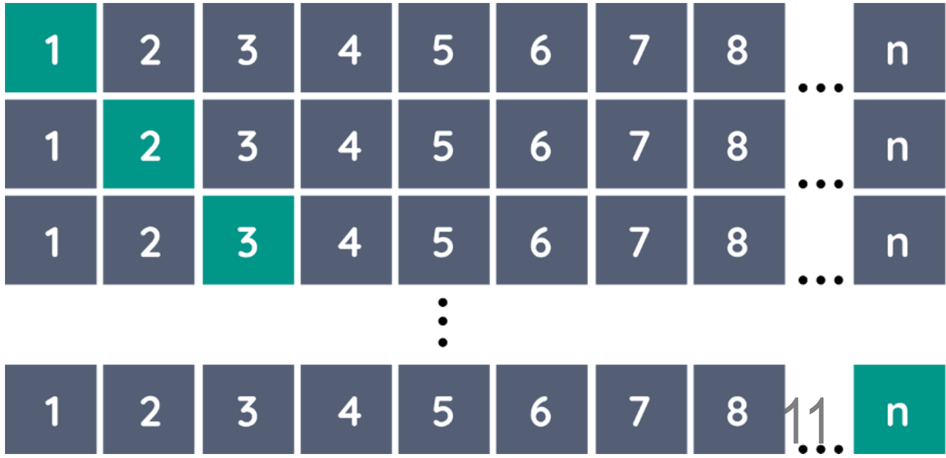
MONASH University

1. **Train/Test Split**



2. **K-fold cross-validation with independent test data set.**

3. **Leave-one-out cross-validation with independent test data set.**



4. **Others**

Recommended model validation strategy

k-fold cross-validation with an independent test data set

	Data				
1.	Validate	Train	Train	Train	Train
2.	Train	Validate	Train	Train	Train
3.	Train	Train	Validate	Train	Train
...	Train	Train	Train	Validate	Train
k	Train	Train	Train	Train	Validate

Compare model performance

RMSE: Root Mean Square Error

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Mean Absolute Error

$$MAE = \frac{1}{n} \sum_{j=1}^{n} |y_j - \hat{y}_j|$$

R Square

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

Slope

$$\text{Slope} = m = \frac{rise}{run} = \frac{y_2 - y_1}{x_2 - x_1}$$

Measurement error

Measurement error in exposure assessment

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,*]

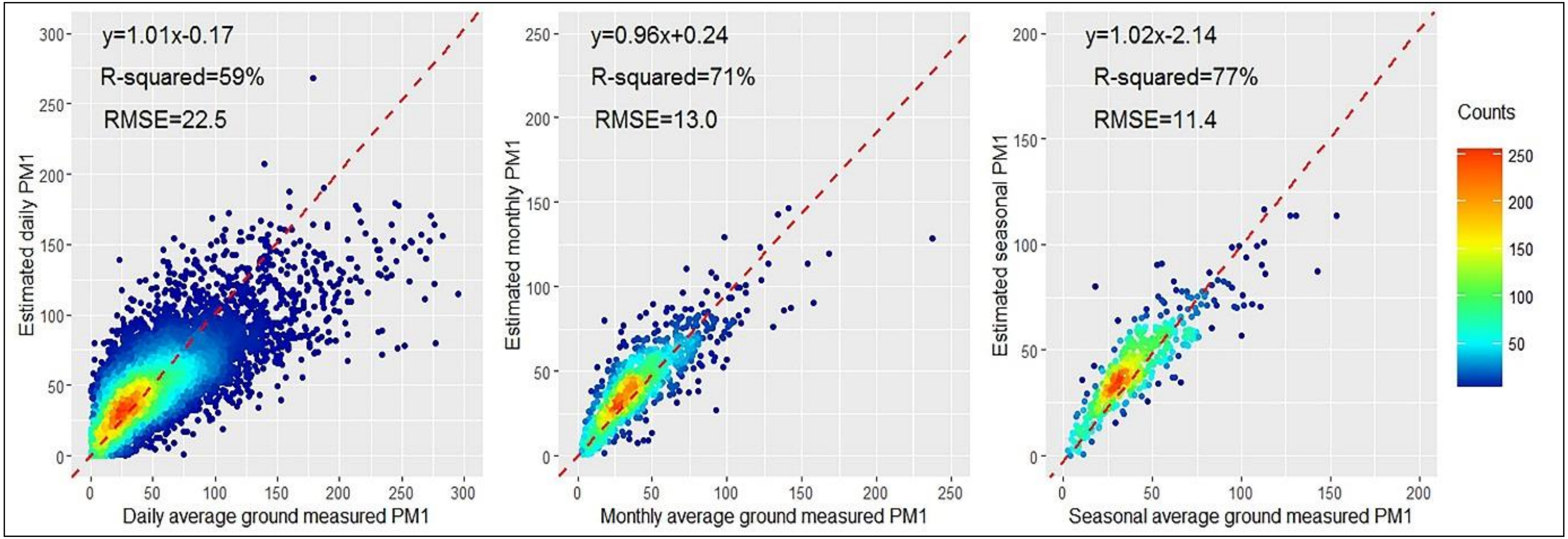
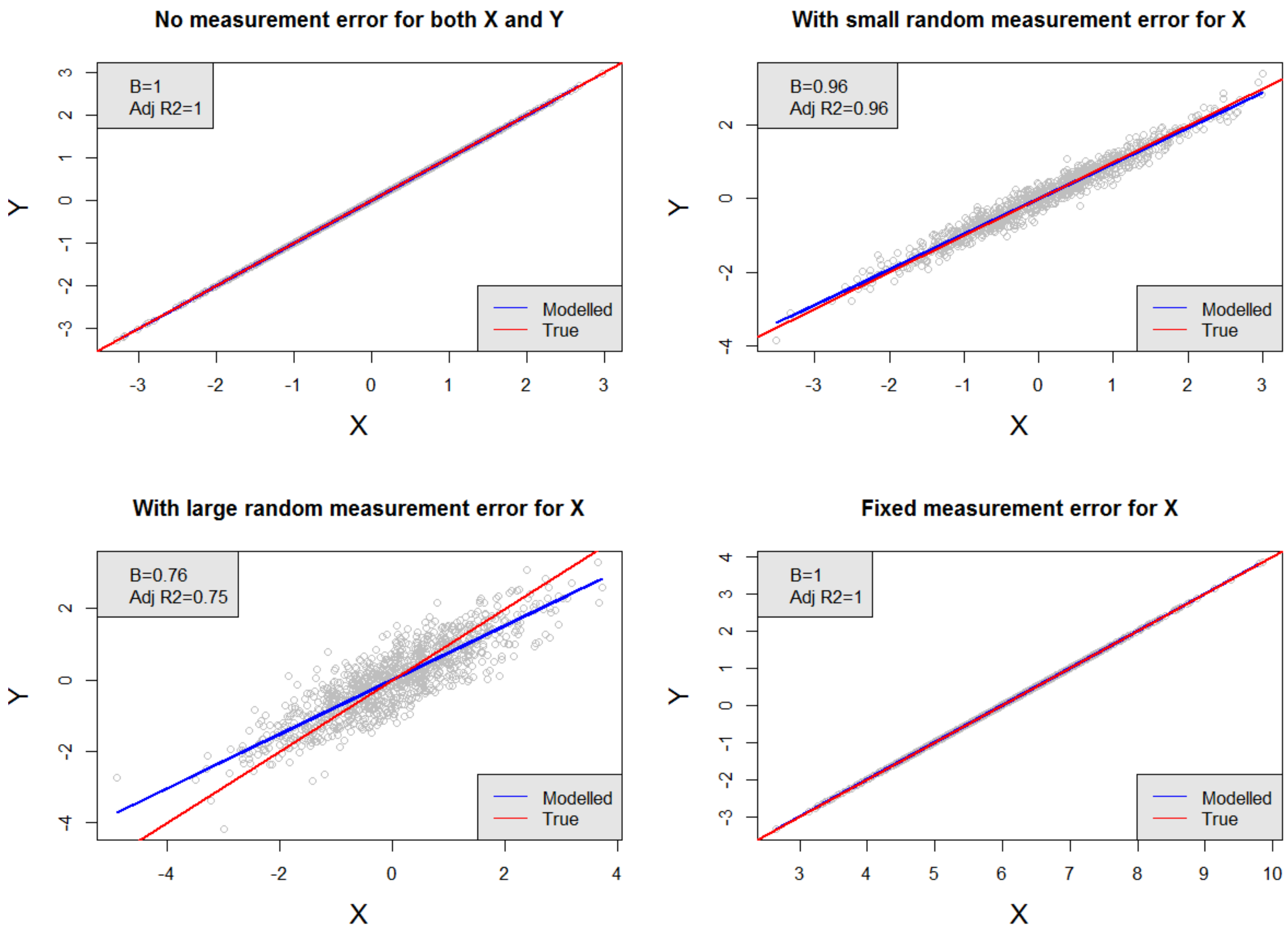


Figure . Scatterplots of 10-fold cross-validation for daily, monthly and seasonal estimation of PM1 concentrations (μg/m3)

Measurement error in environmental risk assessment



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

- Random measurement error lead to underestimated effect estimates.

The more accurate exposure assessment, the more accurate effect estimates!

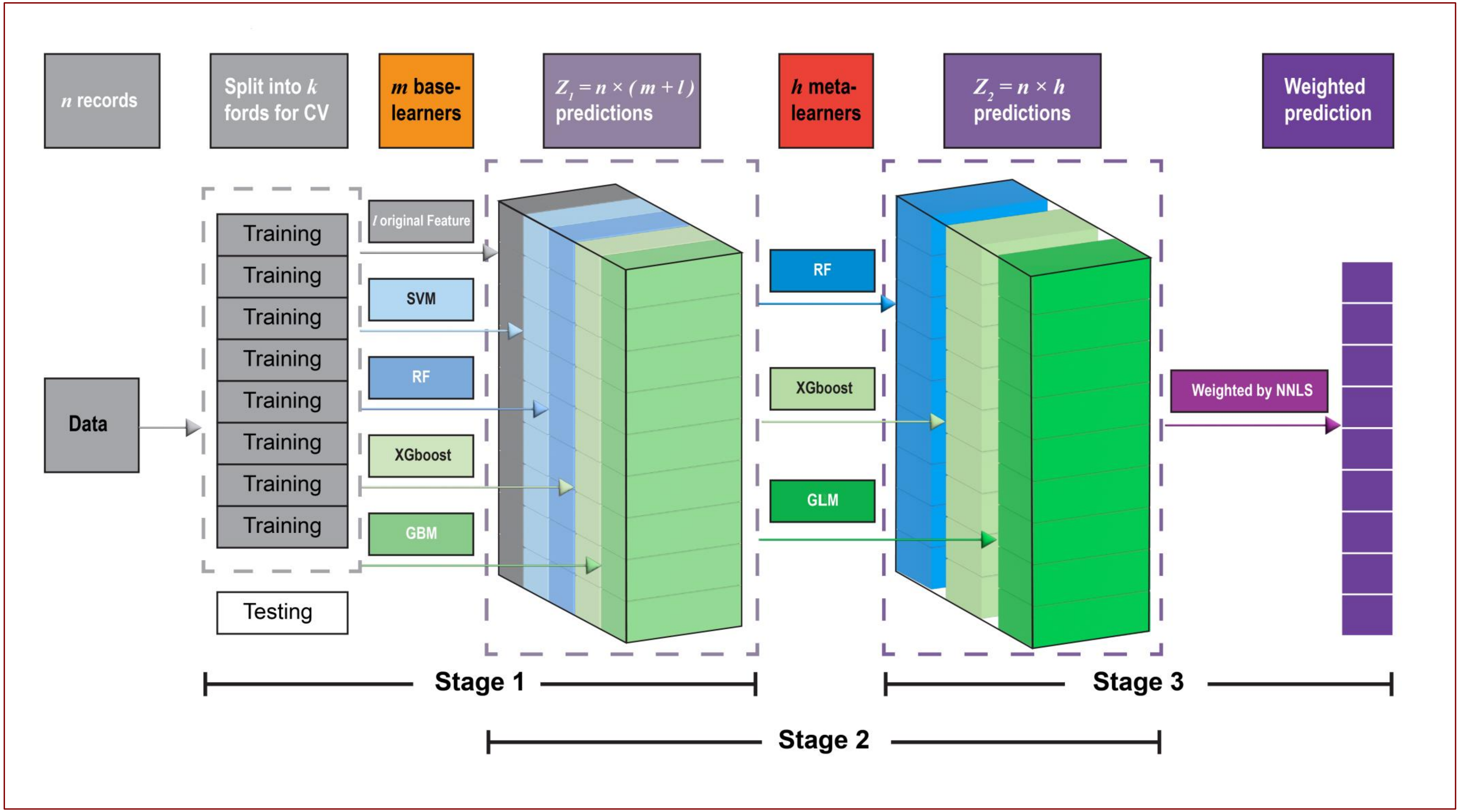
Machine learning

Machine learning for air pollution exposure

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

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

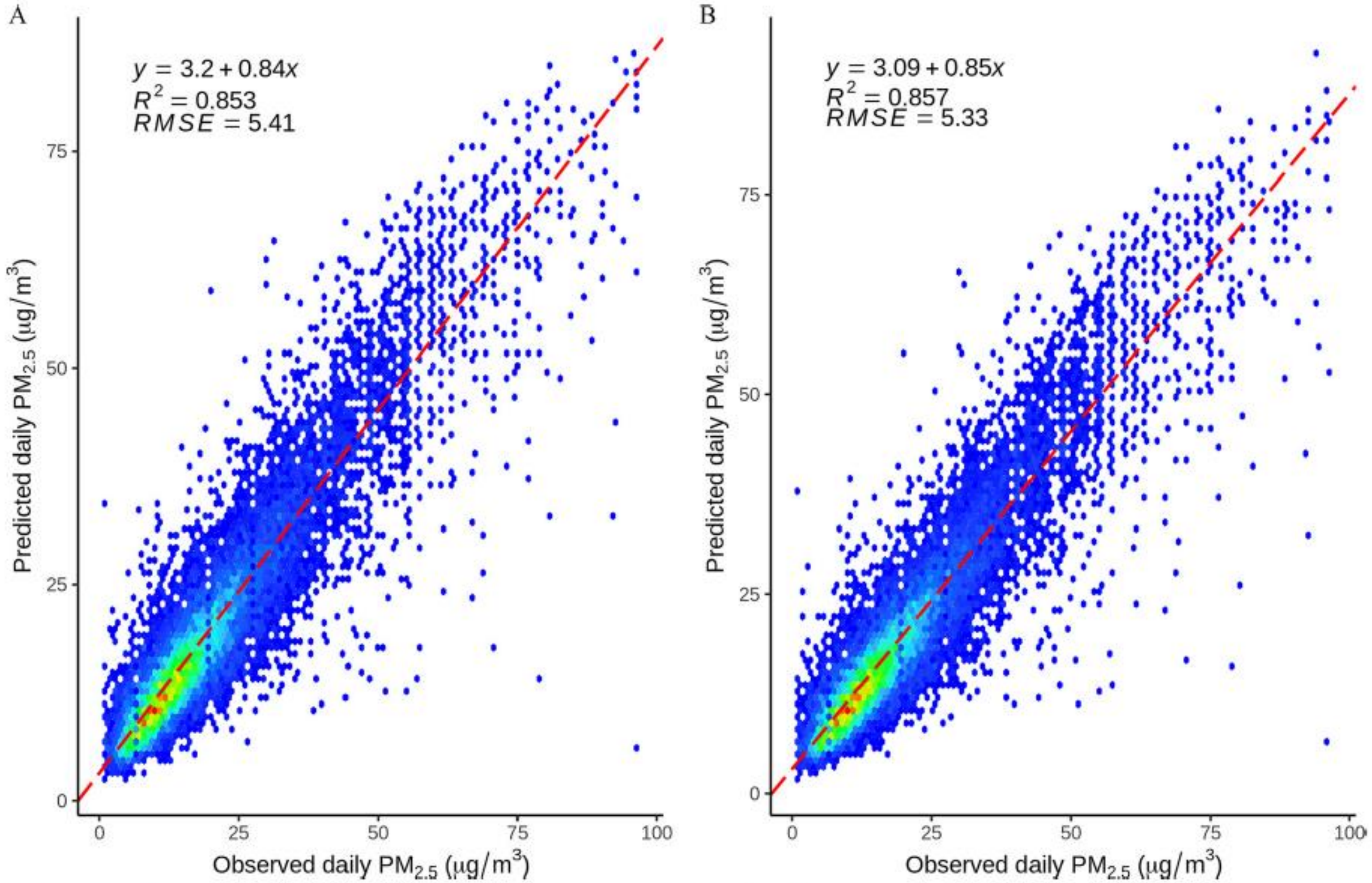


Figure 4. The $PM_{2.5}$ prediction performance of the DEML models with and without AOD as a predictor from 2015–2019 in Italy. The *x*-axis indicates the observed daily $PM_{2.5}$ in the monitor stations; the *y*-axis indicates the estimated $PM_{2.5}$ by the DEML model. The points represent the corresponding $PM_{2.5}$ 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.

Challenges for air pollution exposure assessment

MONASH University

➢ Missing values in satellite data.

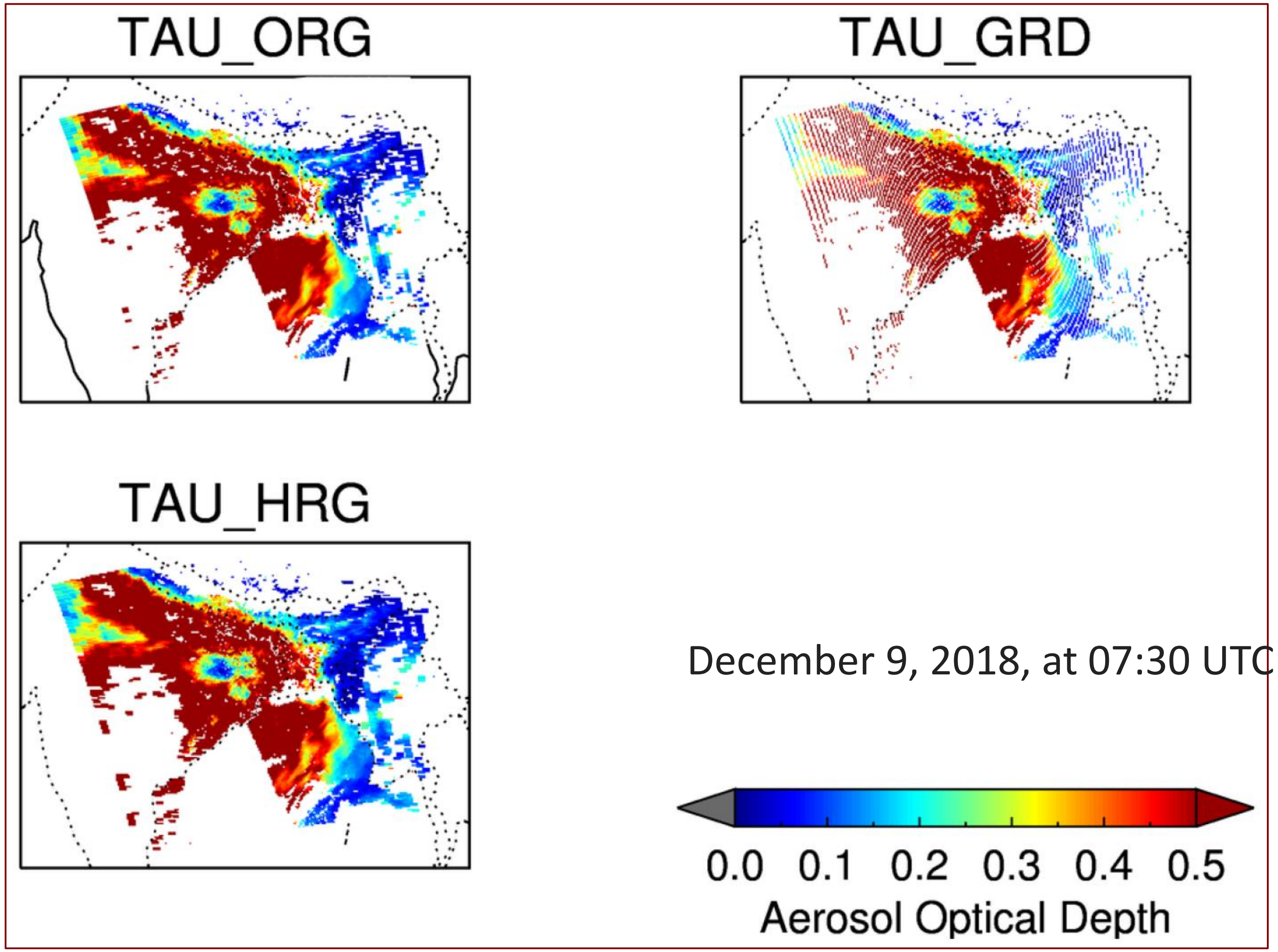


Figure . Aerosol optical depth (550 nm) from Moderate Resolution Imaging Spectroradiometer (MODIS)-Aqua over Asia.

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.

The role of satellite data in air pollution exposure

- 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.

Opportunities

Opportunities for air pollution exposure assessment

- Available big data (observed data including those from low cost sensors, remote sensing, weather data), makes it possible to perform accurate prediction.

- Deep ensemble machine learning, or even 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.

Application of deep ensemble machine learning

MONASH University

- Not only for environmental exposure assessment
- For example, it can be used to predict transmission of infectious diseases

Thank you!


Improving the health of populations
in a changing and inequitable world