MONASH University

DEEP ENSEMBLE MACHINE LEARNING IN $PM_{2.5}$ ESTIMATION

深度集成机器学习在$PM_{2.5}$浓度估算的应用

Wenhua Yu, 余文华

2023.04.21

Content

Part 1	Environmental exposure assessment
Part 2	Basic Machine Learning
Part 3	Deep Ensemble Machine Learning (DEML)
Part 4	Global Daily PM2.5 estimation
Part 5	DEML in time series data forecasting
Part 6	Using DEML in R
Part 7	Q & A

Every year, around
7 MILLION DEATHS
are due to exposure from both outdoor and household air pollution.

Air pollution is a major environmental risk to health.

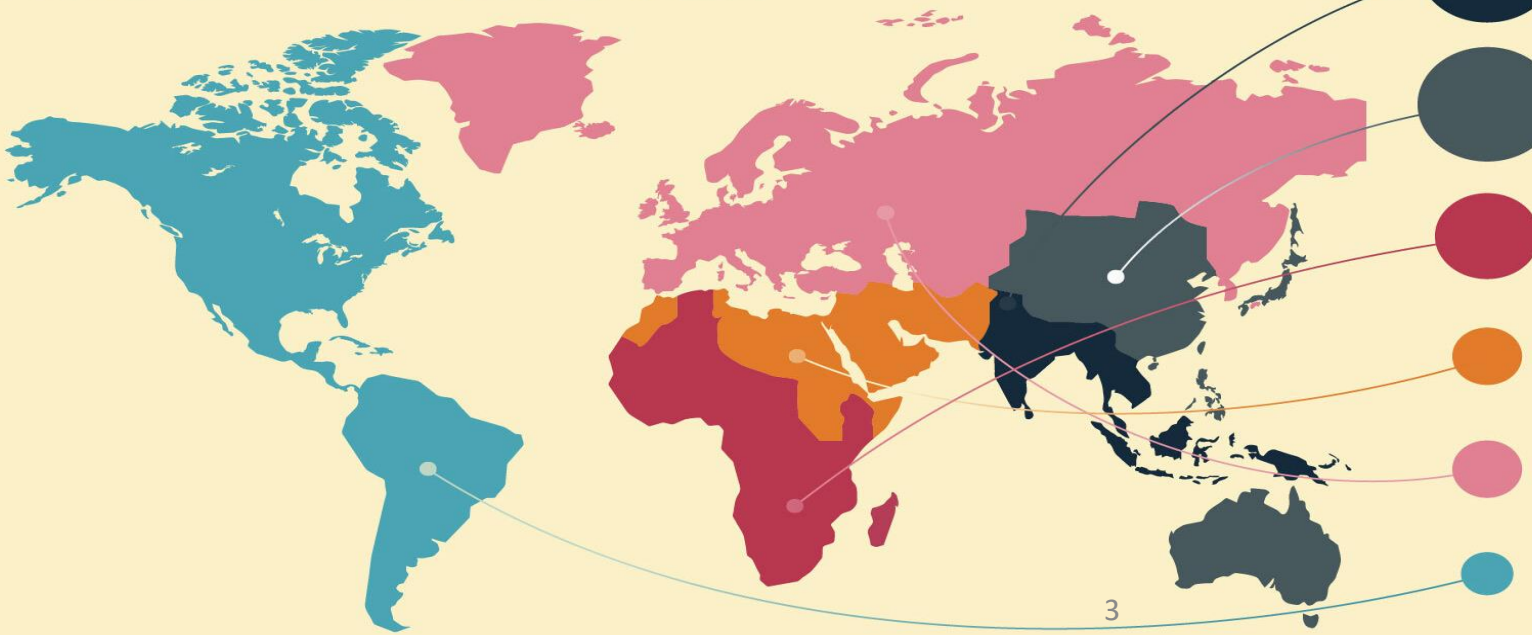
By reducing air pollution levels, countries can reduce:

Stroke

Heart disease

Lung cancer, chronic obstructive pulmonary disease, pneumonia and asthma

REGIONAL ESTIMATES ACCORDING TO WHO REGIONAL GROUPINGS:



More than 2 million
in South-East Asia Region

More than 2 million
in Western Pacific Region

1 million
in Africa Region

500 000
deaths in Eastern Mediterranean Region

500 000
deaths in European Region

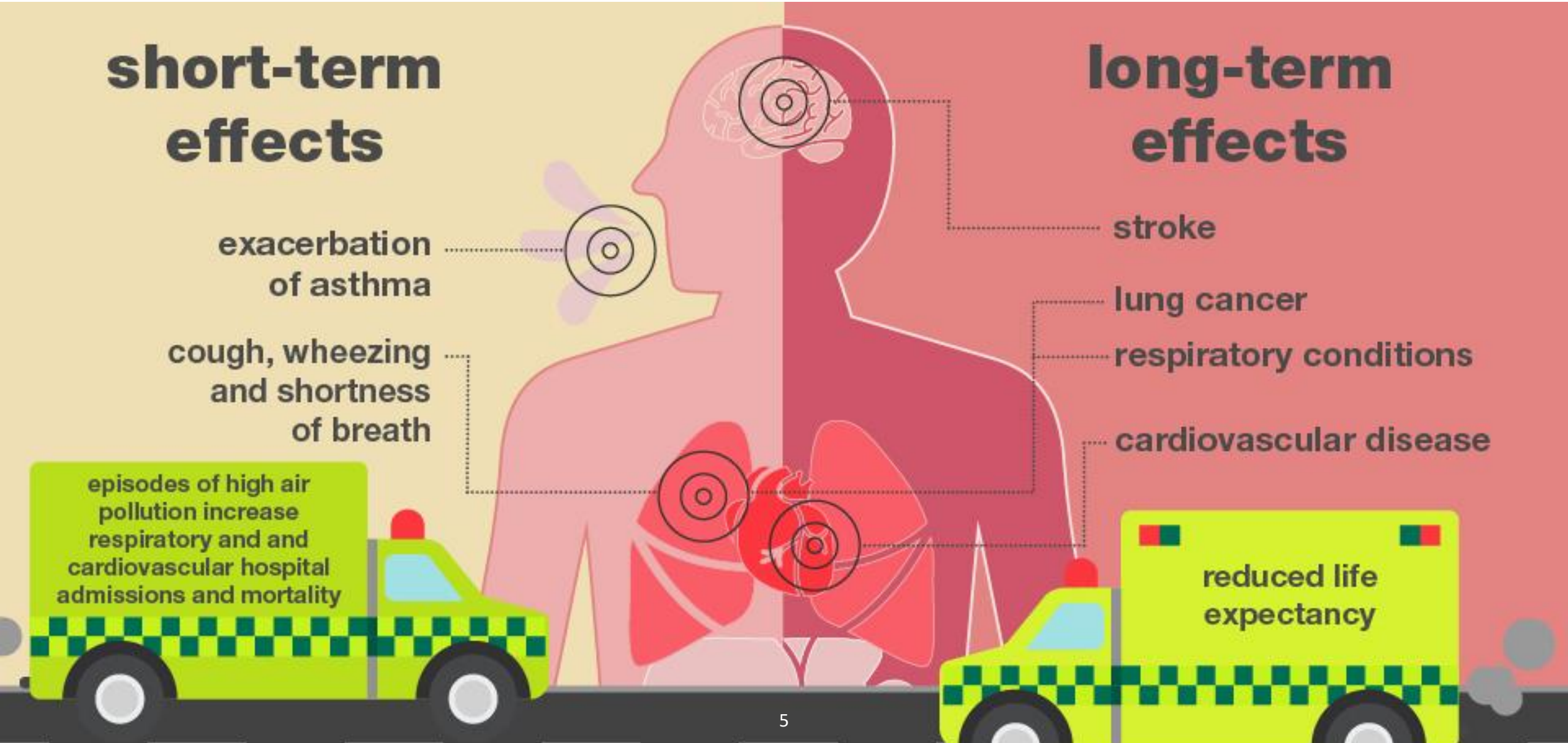
More than 300 000
in the Region of the Americas

Particulate Matter and health

- The WHO estimates “PM affects more people than any other pollutant.”
- $PM_{2.5}$ (particle less than 2.5 microns in diameter) is the more health-damaging particle

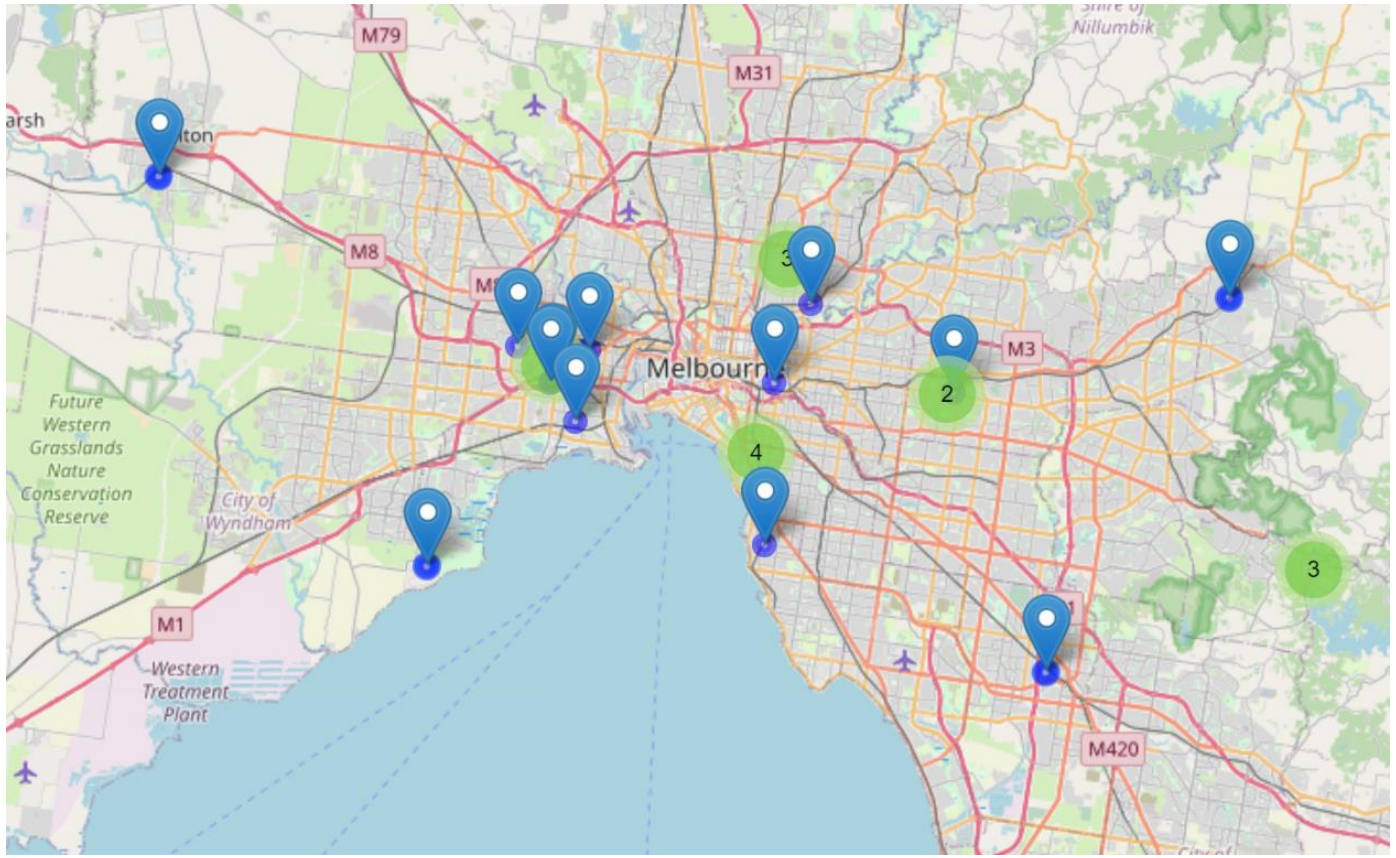
Images from website sources: https://www.urbandemographics.org/post/particulate-matter-size/

Fine particulate matter ($PM_{2.5}$) and health



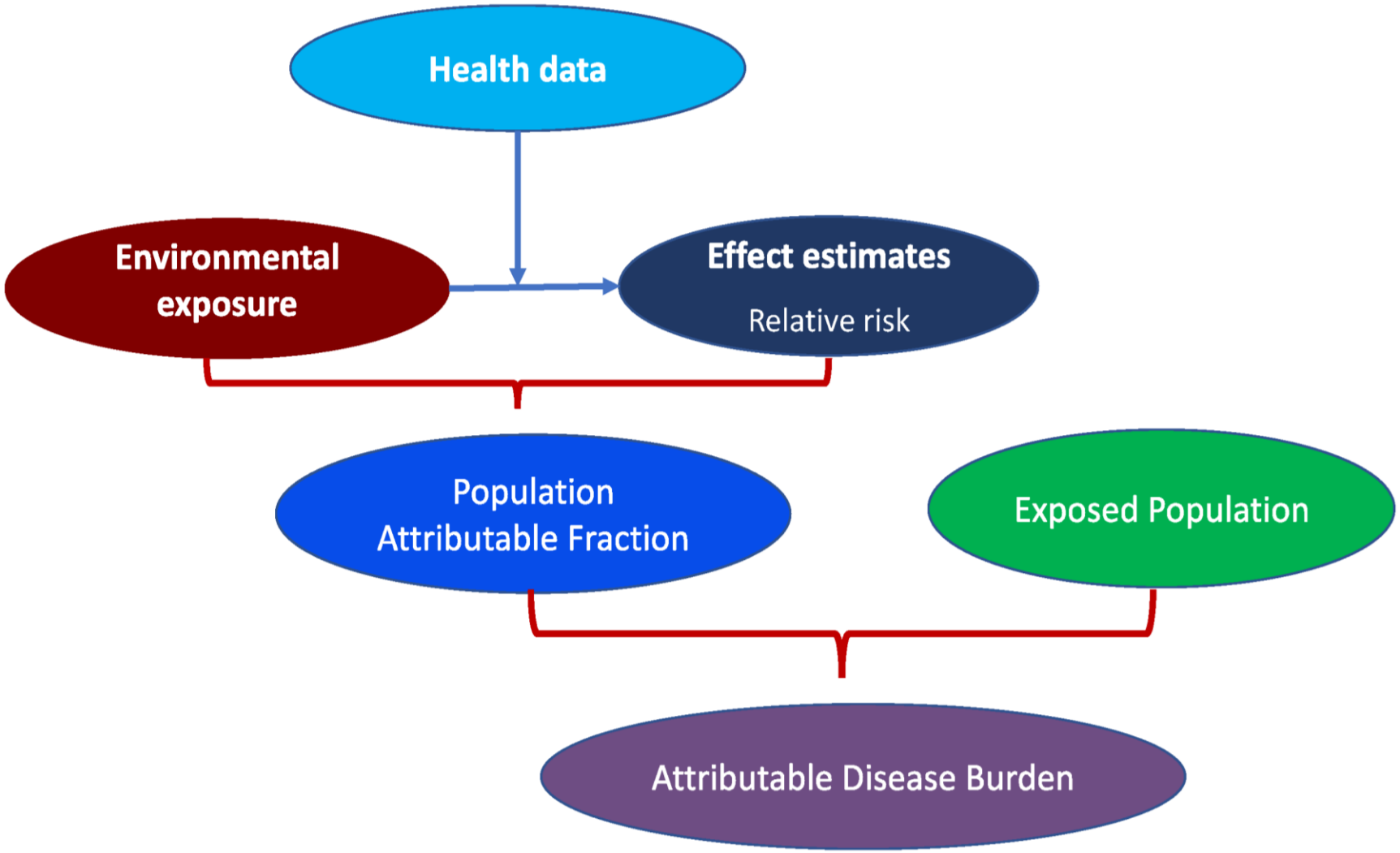
Why perform environmental exposure assessment

- Limited exposure data for both space scale and time scale
- Air quality monitoring networks are mainly distributed in urban areas



MONASH University

The role of environmental exposure assessment



MONASH University

The usually involved variables:

- Ground air quality station measurements
- Satellite remote sensing products
  - Aerosol optical depth(AOD), satellite-based reanalysis data, et al
- Meteorological data
  - Temperature, Relative humidity, precipitation, air pressure, wind speed, et al
- Land use information
  - Land cover, greenness, road density, urban cover, et al
- Spatiotemporal correlations
  - Lat, lon, year, month, days, day of the week, et al
- Others
  - Wildfire, population density, elevation, et al

The commonly used methods:

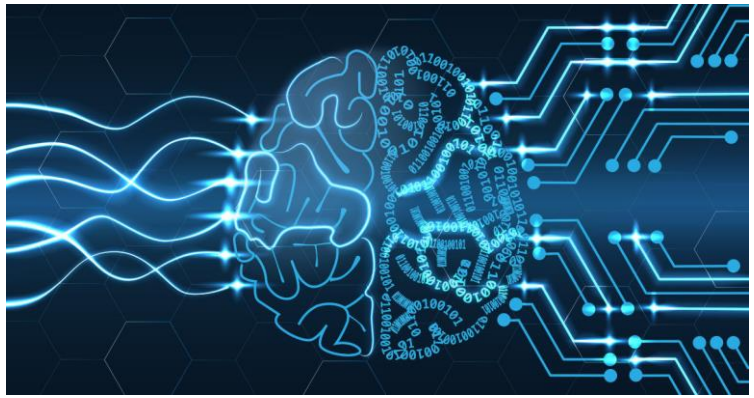
- Generalized Linear regression
  - GLM_model <- glm (PM ~ AOD + WS, family, data = modeling_dataset)
- Generalized addition model
  - GAM_model <- gam (PM ~ s(Temp) + s (WS), family, data = modeling_dataset)
- Mixed effect model
  - LME_model <- lme (fixed = PM ~ AOD + WS, random = list (DOY = ~1 + AOD + WS), data = modeling_dataset)
- Bayesian spatiotemporal model
  - SpatioTemporal, bmstdr packages et al
- Kriging
- Chemical transport model
- Machine learning (ML) & deep learning (DL)

Opportunities for environmental exposure assessment

Available Big data
(大数据)

Machine Learning/ deep learning provides better prediction performances
（人工智能）

High-performance computer and cloud calculation
（高性能云计算）



Challenges for environmental exposure assessment

- How to select ML models? Which one is the best?
- How to apply ML easily?
- How to integrate multiple models to perform better than a single ML model?
- How to capture the spatiotemporal variations and improve the generalization

Machine Learning (ML) has been widely used

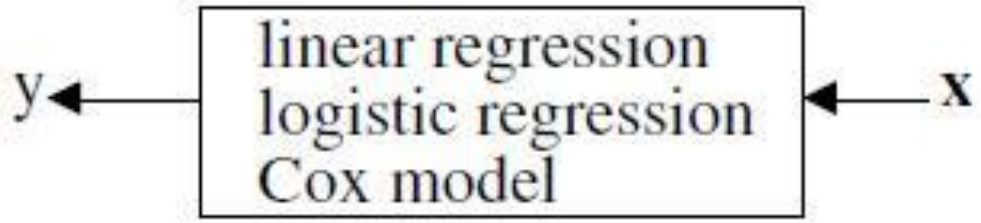
MONASH University

What is Machine Learning

MONASH University

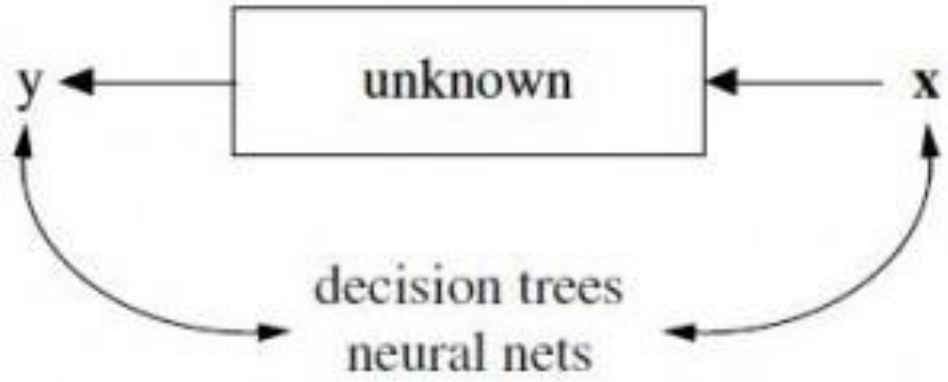
1.1 Machine Learning VS. Statistics

机器学习是给定一些训练样本（x，y），让计算机自动寻找一个决策函数f(.)来建立x与y的关系。



Statistics

寻找一个函数f(x)，用x做输入来预测y，依赖模型预测精度。

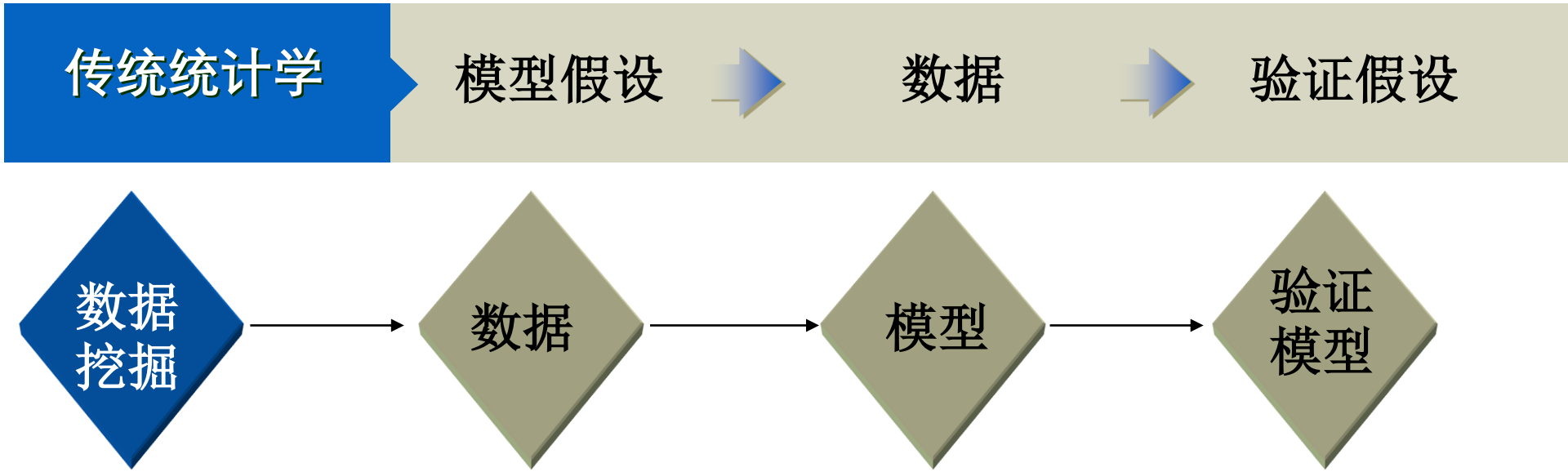


Machine Learning

1.1 Machine Learning VS. Statistics

数理统计：从数据中学习概率统计模型和分布，然后利用模型对新数据进行分析和预测

机器学习：从问题出发，寻找合适的算法形成模型，然后利用模型对新数据进行分析和预测



- 传统统计学：基于数学理论和演绎推理过程。
- 机器学习：基于“实验”和归纳推理过程。

1.2 ML Overview

- David Hand在其《principal of Data Mining》中把机器学习算法解构为4个组件:

- 1) 模型结构（函数形式，如线性模型）

- 2) 评分函数（评估模型拟合数据的质量，如似然函数、误差平方和等）

- 3) 优化和搜索方法（评分函数的优化和模型参数的求解）

- 4) 数据管理策略（优化和搜索时对数据的高效访问，如并行化访问等）

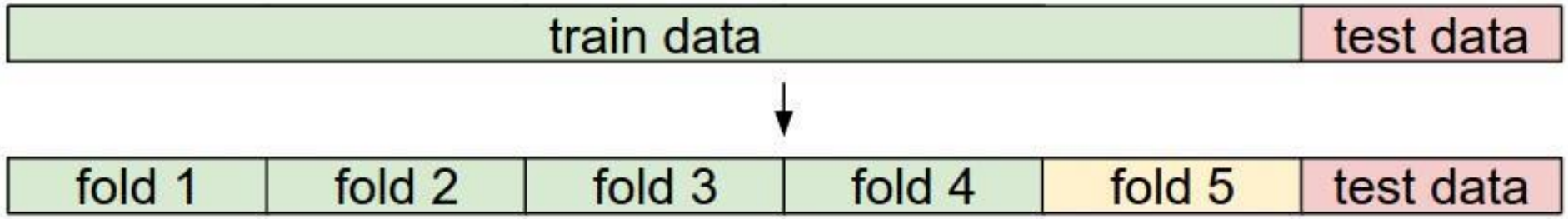
1.2 ML Overview

机器学习分类：

1.2 ML Overview

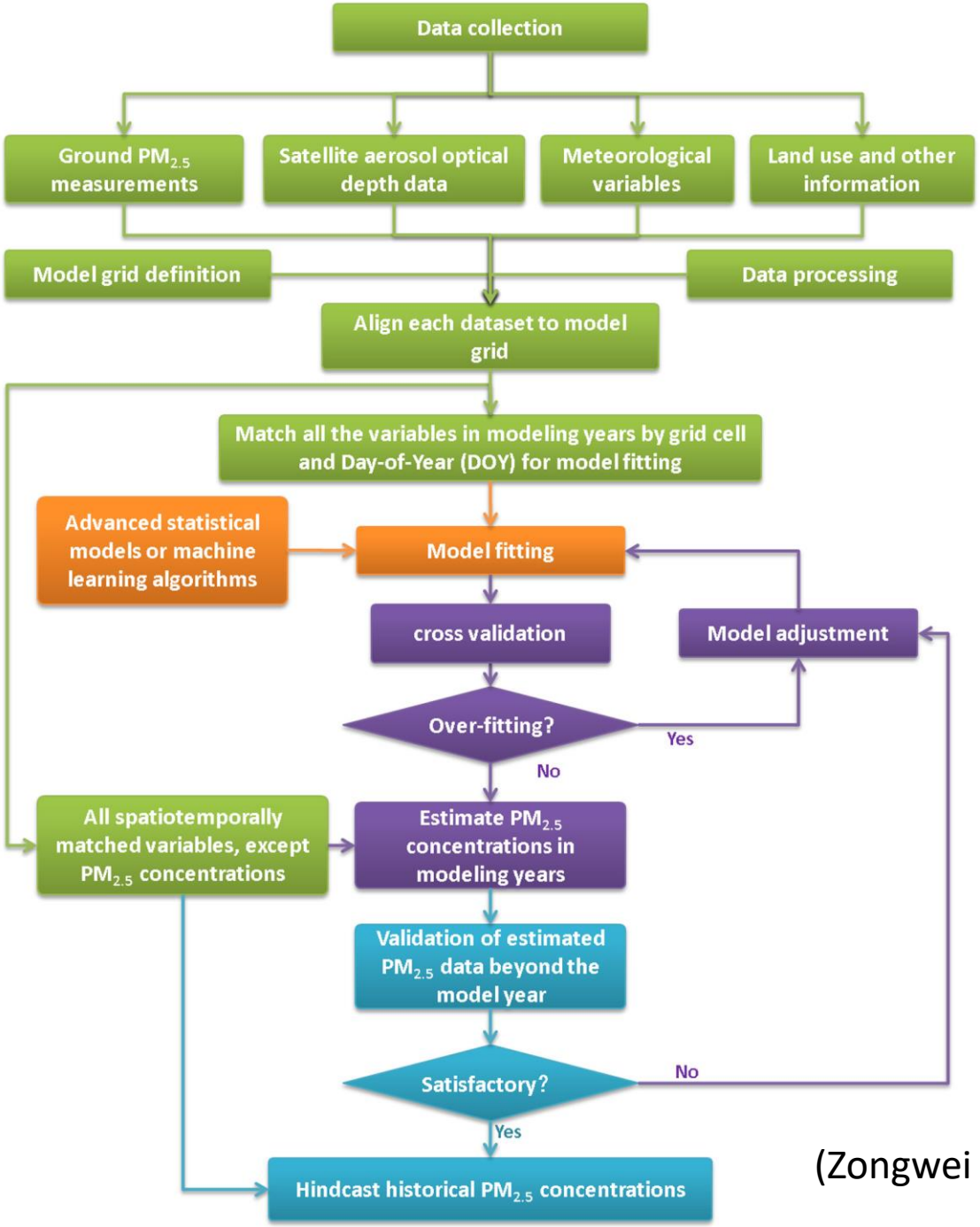
k折交叉验证（K-fold cross validation）

- 随机地将数据切分为K个互不相交的大小相同子集，利用k-1个子集的数据训练模型，利用余下的子集测试模型。重复进行以上k种模型，选出k次测评中平均测试误差最小的模型。特殊情况k=n时，称留一交叉验证（leave-one-out CV）往往在数据缺乏时使用。



- k越小，bias越大，但会减少方差；留一法使unbiased，但会有高方差。

1.2 ML Overview



(Zongwei Ma, et al, 2022)

1.2 ML Overview

Model Performance measurements

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

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

1.2 ML Overview

Model Selection

考虑因素：

1. 适用问题
2. 适用数据类型
3. 缺失值
4. 异常值
5. 数据处理(归一化等)
6. 惩罚项
7. 模型特点 (优缺点和适应性)
8. 模型实现(R & python)

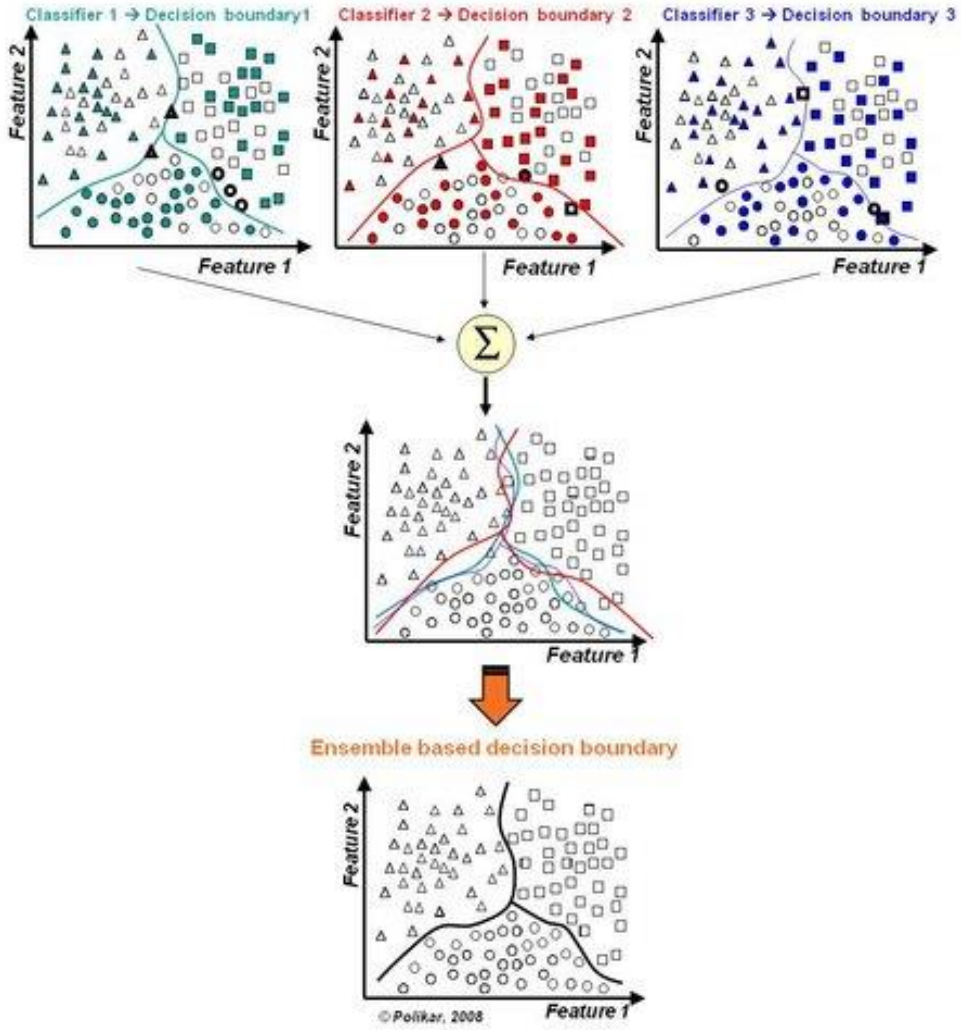
Deep Ensemble Machine Learning (DEML)

MONASH University

2. What is the ensemble approach?

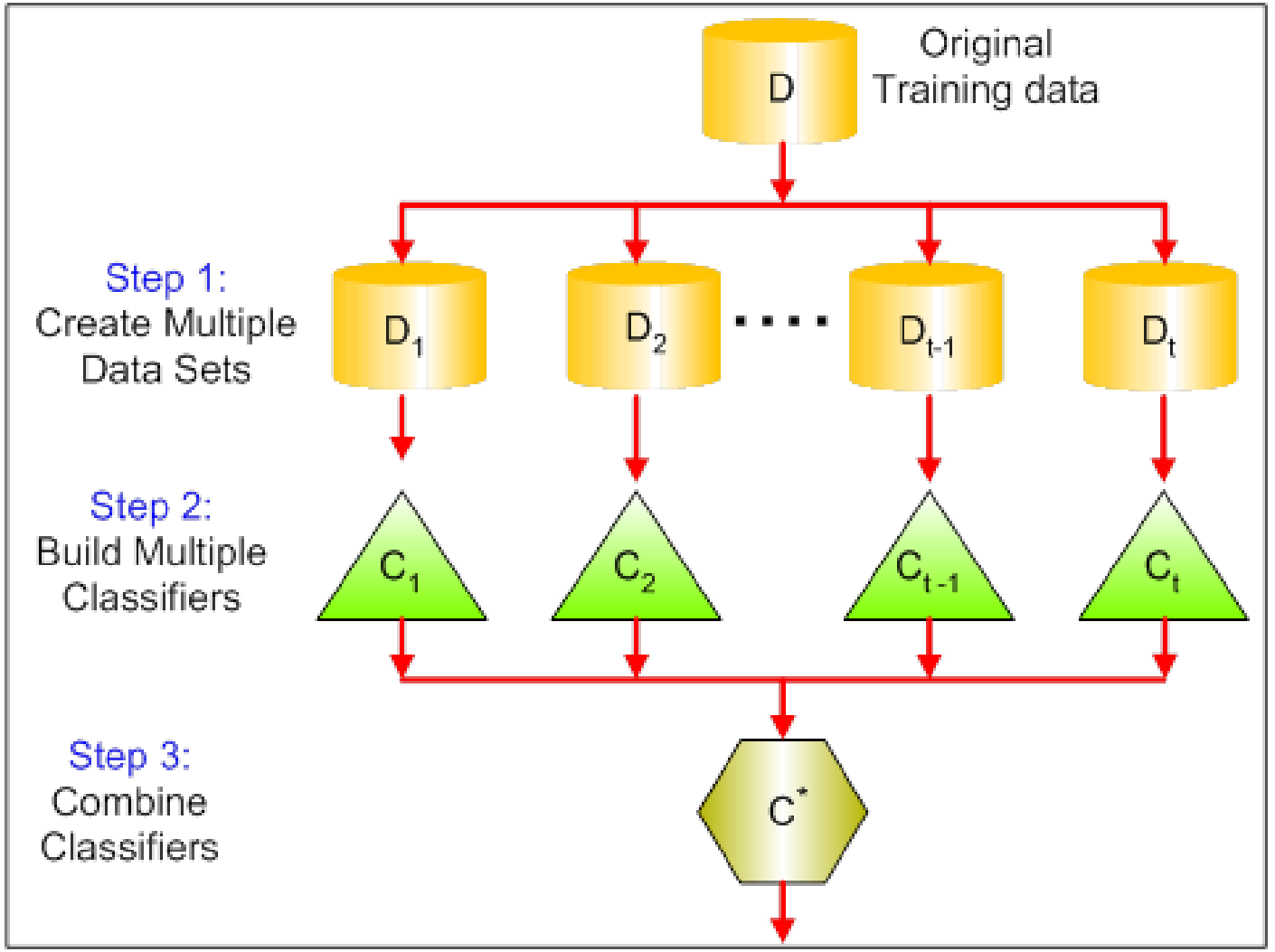
Ensemble learning：是组合许多弱模型(预测效果一般的模型) 以得到一个强模型(预测效果好的模型)。

Ensemble中组合的模型可以是同一类的模型，也可以是不同类型的模型。主要的集成方法有：Bagging（Bootstrap aggregation），Boosting 和Stacking。



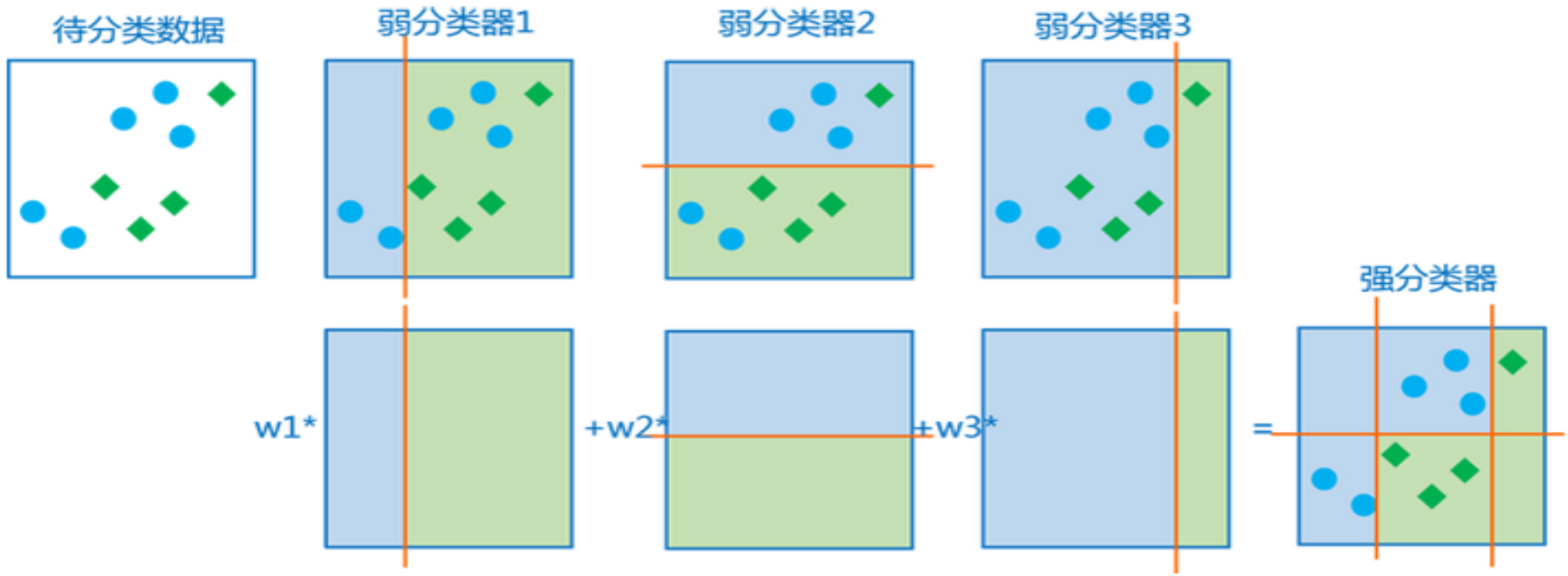
2.1.Bagging（random forest)

- 通过Bootstrap 进行有放回抽样，这样就能获得 D个训练集。
- Bagging对回归树来说，直接加起来求平均即可。对分类树来说，采用的是少数服从多数的投票的策略。
- Bagging里的每棵树都是High Variance, Low bias,平均起来之后成功的避免了High variance。
- 随机森林算法是典型的bagging算法。



2.2 Boosting

- Boosting提升方法就是从弱学习算法出发，反复学习，得到一系列弱分类器，然后组合这些弱分类器，构成一个强分类器。XGBoost, GBM,GBDT, Adaptive Boosting（自适应增强）都应用该思路进行模型提升。



3. Deep Ensemble Machine Learning (DEML)

Research

A Section 508–conformant HTML version of this article is available at https://doi.org/10.1289/EHP9752.

Deep Ensemble Machine Learning Framework for the Estimation of $PM_{2.5}$ Concentrations

Wenhua Yu,[1] Shanshan Li,[1] Tingting Ye,[1] Rongbin Xu,[1] Jiangning Song,[2] and Yuming Guo[1]

[1]Climate, Air Quality Research Unit, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
[2]Monash Biomedicine Discovery Institute, Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia

Background: Accurate estimation of historical $PM_{2.5}$ (particle matter with an aerodynamic diameter of less than 2.5 μm) is critical and essential for environmental health risk assessment.

Objectives: The aim of this study was to develop a multiple-level stacked ensemble machine learning framework for improving the estimation of the daily ground-level $PM_{2.5}$ concentrations.

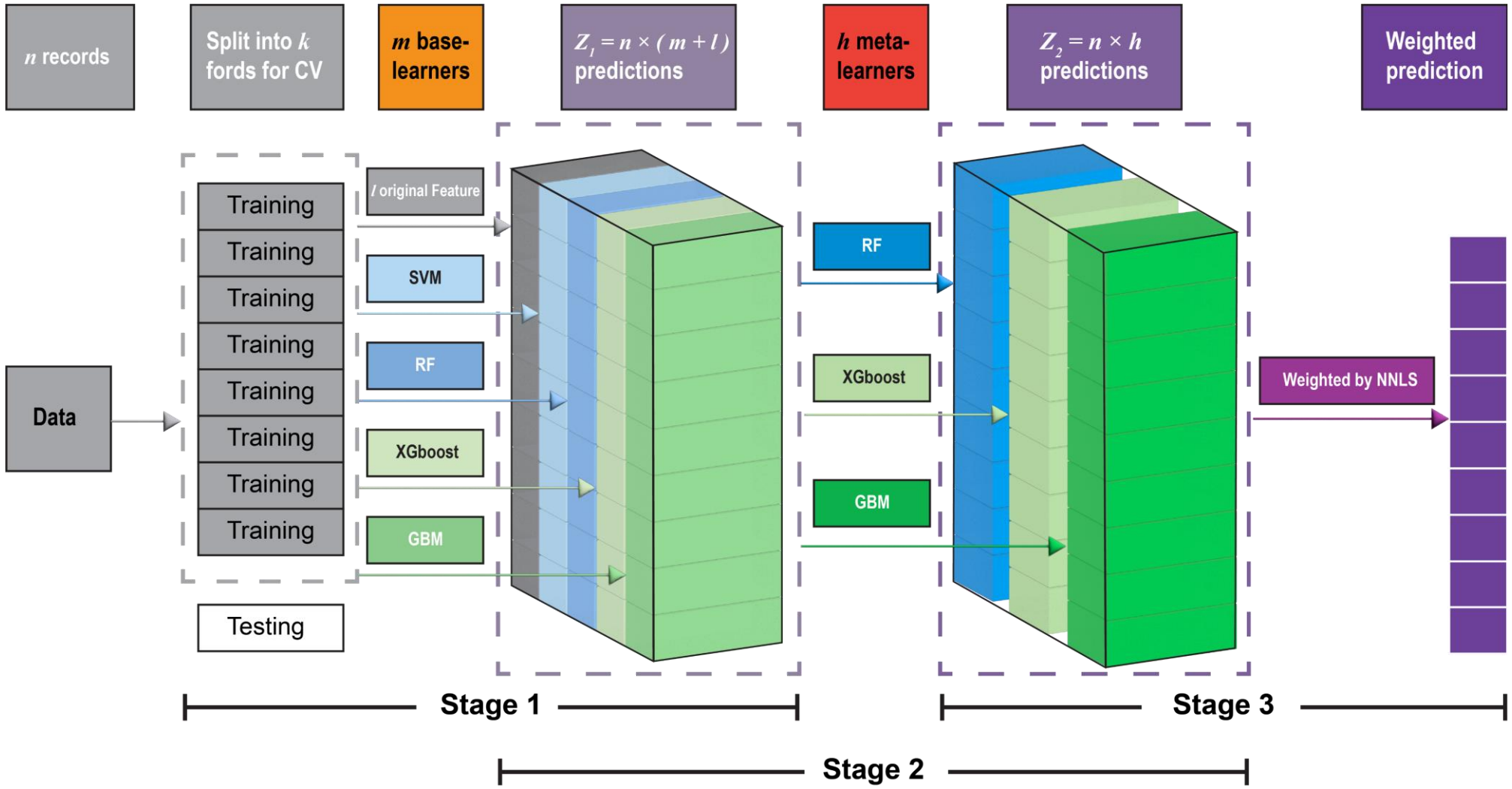
Methods: An innovative deep ensemble machine learning framework (DEML) was developed to estimate the daily $PM_{2.5}$ concentrations. The framework has a three-stage structure: At the first stage, four base models [gradient boosting machine (GBM), support vector machine (SVM), random forest (RF), and eXtreme gradient boosting (XGBoost)] were used to generate a new data set of $PM_{2.5}$ concentrations for training the next-stage learners. At the second stage, three meta-models [RF, XGBoost, and Generalized Linear Model (GLM)] were used to estimate $PM_{2.5}$ concentrations using a combination of the original data set and the predictions from the first-stage models. At the third stage, a nonnegative least squares (NNLS) algorithm was employed to obtain the optimal weights for $PM_{2.5}$ estimation. We took the data from 133 monitoring stations in Italy as an example to implement the DEML to predict daily $PM_{2.5}$ at each 1 km × 1 km grid cell from 2015 to 2019 across Italy. We evaluated the model performance by performing 10-fold cross-validation (CV) and compared it with five benchmark algorithms [GBM, SVM, RF, XGBoost, and Super Learner (SL)].

Results: The results revealed that the $PM_{2.5}$ prediction performance of DEML [coefficients of determination $(R^2) = 0.87$ and root mean square error (RMSE) = 5.38 μg/m^3] was superior to any benchmark models (with R^2 of 0.51, 0.76, 0.83, 0.70, and 0.83 for GBM, SVM, RF, XGBoost, and SL approach, respectively). DEML displayed reliable performance in capturing the spatiotemporal variations of $PM_{2.5}$ in Italy.

Discussion: The proposed DEML framework achieved an outstanding performance in $PM_{2.5}$ estimation, which could be used as a tool for more accurate environmental exposure assessment. https://doi.org/10.1289/EHP9752

MONASH University

3.1 Deep ensemble ML (DEML) framework



Note: SVM: Support Vector Machine; RF: Random Forest; XGBoost: Extreme Gradient Boosting; GBM: Gradient Boosting Machine; NNLS: Non-negative least squares

Fig 1. The overview of DEML framework

MONASH University

3.2 DEML advantages 优势

1. 属于multiple-level stacked ensemble model，应用cross validation 避免过拟合

2. 可自由调整多样的层级结构 (Diverse hierarchy structure)

3. 最小化经验模型选择 (Minimize the extent of the empirical model selection)

4. 自适应调整模型权重 (Automatically provide an optimal set of weights)

5. 理论上, DEML结果会优于单个base模型

3.2 DEML LIMITATIONS 劣势

1. 无法处理带有缺失值的数据，需要提前进行数据处理

2. 由于多个模型叠加，增加了运算量

3. 需要一定的数据样本，否则会造成过拟合

4. Global Daily PM2.5 estimation

Global estimates of daily ambient fine particulate matter concentrations and unequal spatiotemporal distribution of population exposure: a machine learning modelling study

Wenhua Yu, Tingting Ye, Yiwen Zhang, Rongbin Xu, Yadong Lei, Zhuying Chen, Zhengyu Yang, Yuxi Zhang, Jiangning Song, Xu Yue, Shanshan Li, Yuming Guo

Summary

Background Short-term exposure to ambient $PM_{2.5}$ is a leading contributor to the global burden of diseases and mortality. However, few studies have provided the global spatiotemporal variations of daily $PM_{2.5}$ concentrations over recent decades.

Methods In this modelling study, we implemented deep ensemble machine learning (DEML) to estimate global daily ambient $PM_{2.5}$ concentrations at $0{\cdot}1° \times 0{\cdot}1°$ spatial resolution between Jan 1, 2000, and Dec 31, 2019. In the DEML framework, ground-based $PM_{2.5}$ measurements from 5446 monitoring stations in 65 countries worldwide were combined with GEOS-Chem chemical transport model simulations of $PM_{2.5}$ concentration, meteorological data, and geographical features. At the global and regional levels, we investigated annual population-weighted $PM_{2.5}$ concentrations and annual population-weighted exposed days to $PM_{2.5}$ concentrations higher than 15 μg/m³ (2021 WHO daily limit) to assess spatiotemporal exposure in 2000, 2010, and 2019. Land area and population exposures to $PM_{2.5}$ above 5 μg/m³ (2021 WHO annual limit) were also assessed for the year 2019. $PM_{2.5}$ concentrations for each calendar month were averaged across the 20-year period to investigate global seasonal patterns.

Lancet Planet Health 2023; 7: e209–18

Climate, Air Quality Research Unit, School of Public Health and Preventive Medicine (W Yu MPH, T Ye MSc, Yi Zhang MSc, R Xu PhD, Z Yang MPH, Yu Zhang PhD, S Li PhD, Prof Y Guo PhD), Turner Institute for Brain and Mental Health, School of Psychological Sciences (Z Chen PhD), and Monash Biomedicine Discovery Institute, Department of Biochemistry and Molecular Biology (J Song PhD), Monash University, Melbourne, VIC,

4.1 Method: Data collection

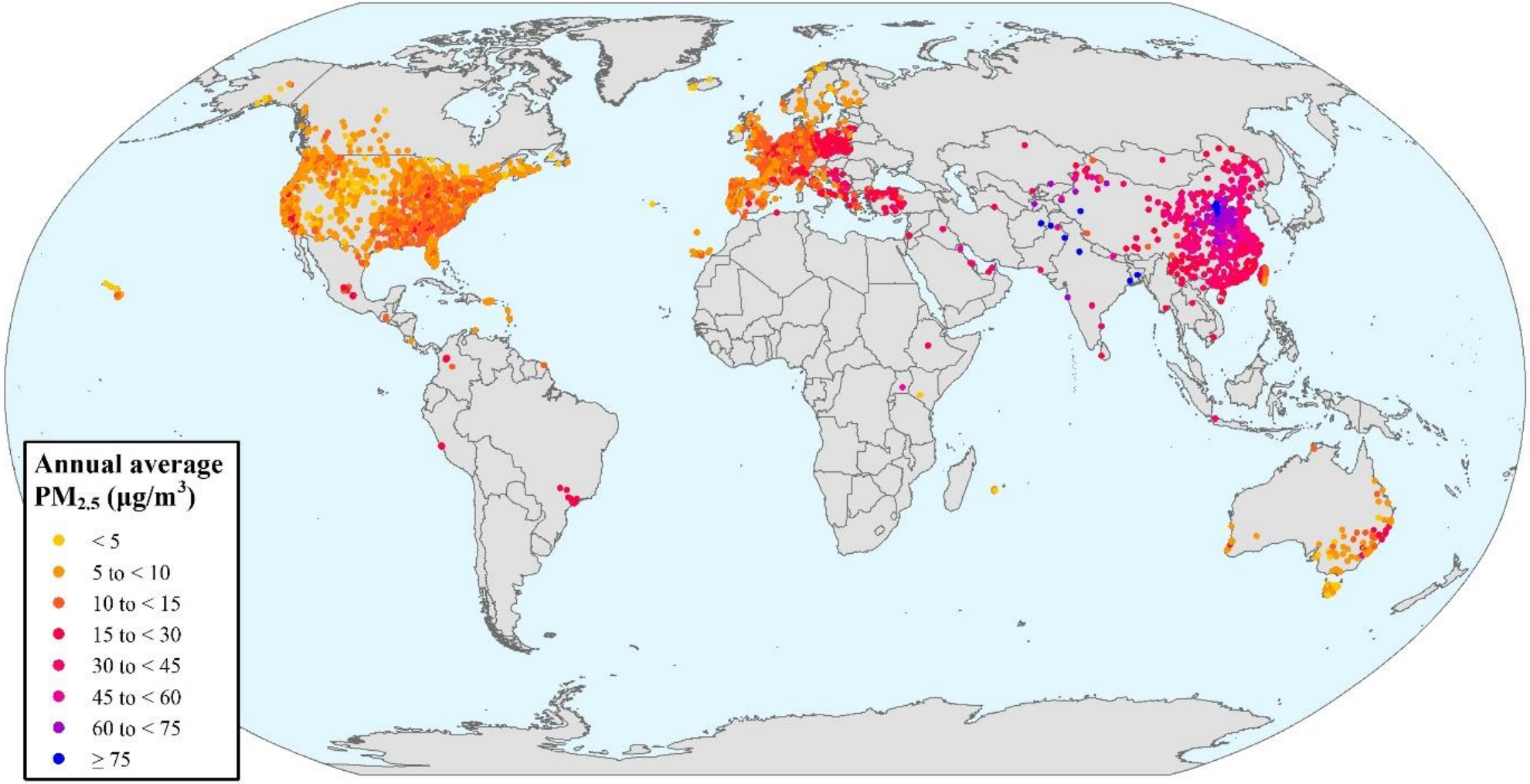


Fig 2. Ground-based observation of $PM_{2.5}$ from 5,446 monitor stations in 65 countries and regions in 2000-2019

MONASH University

4.1 Method

Model establishment

Ground-based $PM_{2.5}$

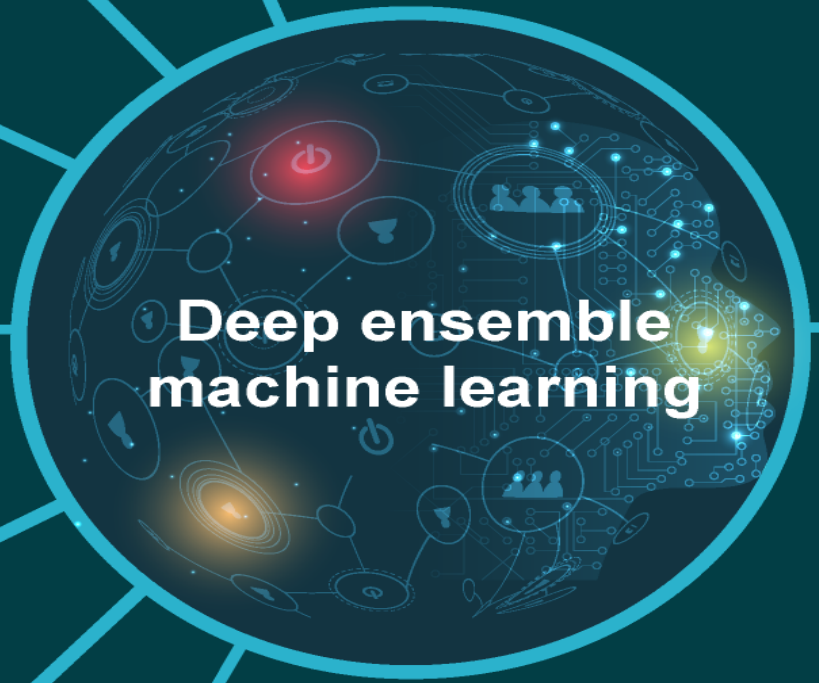
Satellite remote sensing products

Meteorology (temperature, humidity, wind speed, rainfall, air pressure, sunshine duration)

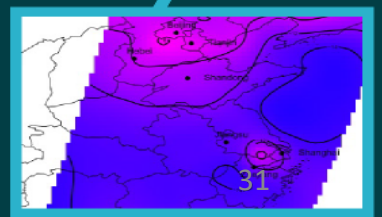
Land use information (land cover, vegetation, road types)

Others (elevation, fire, population density)

Pollution emission inventory



Predicted daily concentrations of air pollutants



Chemical transport model

4.1 Method

Model structure

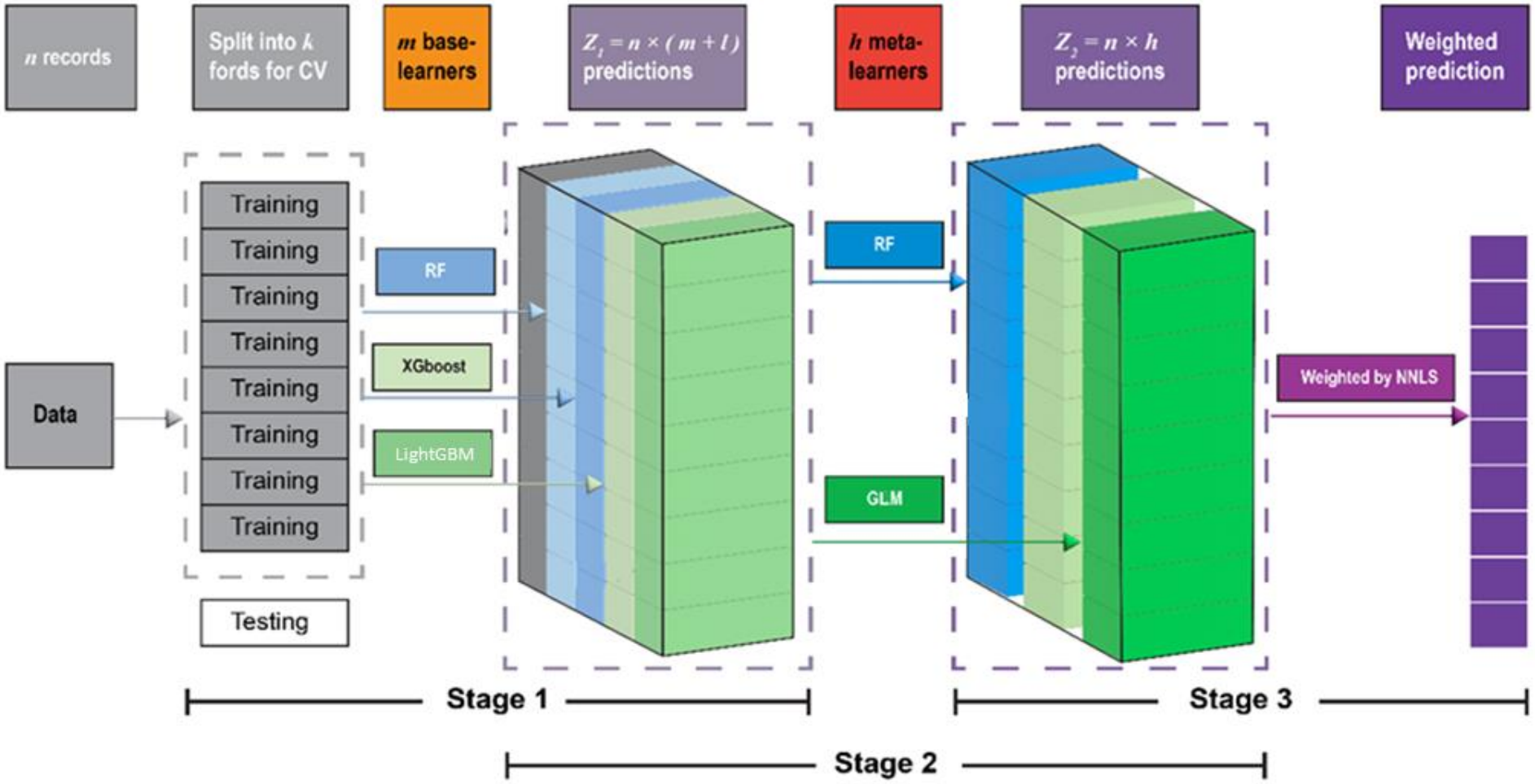


Fig 3. The framework of DEML model for global daily $PM_{2.5}$ estimation

MONASH University

4.1 Method

Exposure measurement

1) Annual average Population-weighted $PM_{2.5}$:

$$PWD = \sum (\frac{p_i}{P} \times C_i)$$

2) Annual accumulate Population-weighted exposed days:

$$PED_{>15\mu g/m^3} = \sum_{j=1}^{365} \sum \left(\frac{p_i}{P} \times D_{ij} \right)$$

where C_i and p_i denote the daily average $PM_{2.5}$ concentration and annual average population in a specific grid cell i, respectively; $P = \sum p_i$, which is the total population of grid cells in a certain region; D_{ij} is a Boolean value, indicating whether daily average $PM_{2.5}$ in a specific j day of a year in a grid cell i was above 15 μg/m³ (D_{ij}=1) or not (D_{ij}=0).

4.2 Results

DEML model performance

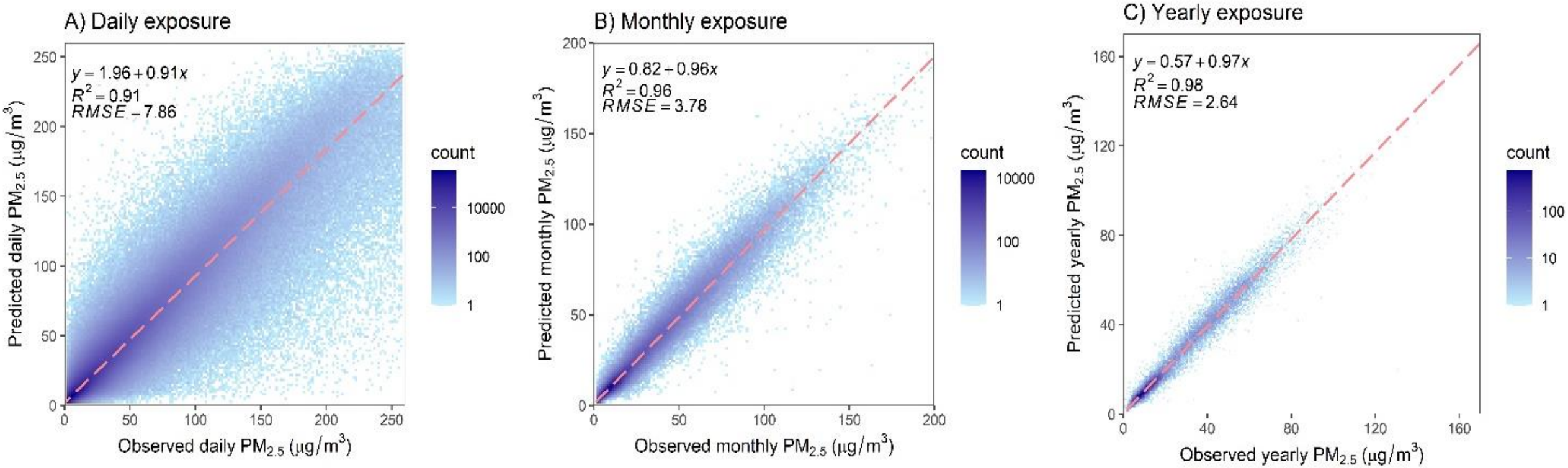


Fig 4. Model performance for global $PM_{2.5}$ prediction in 2000-2019

MONASH University

4.2 Results

$PM_{2.5}$ spatiotemporal distribution

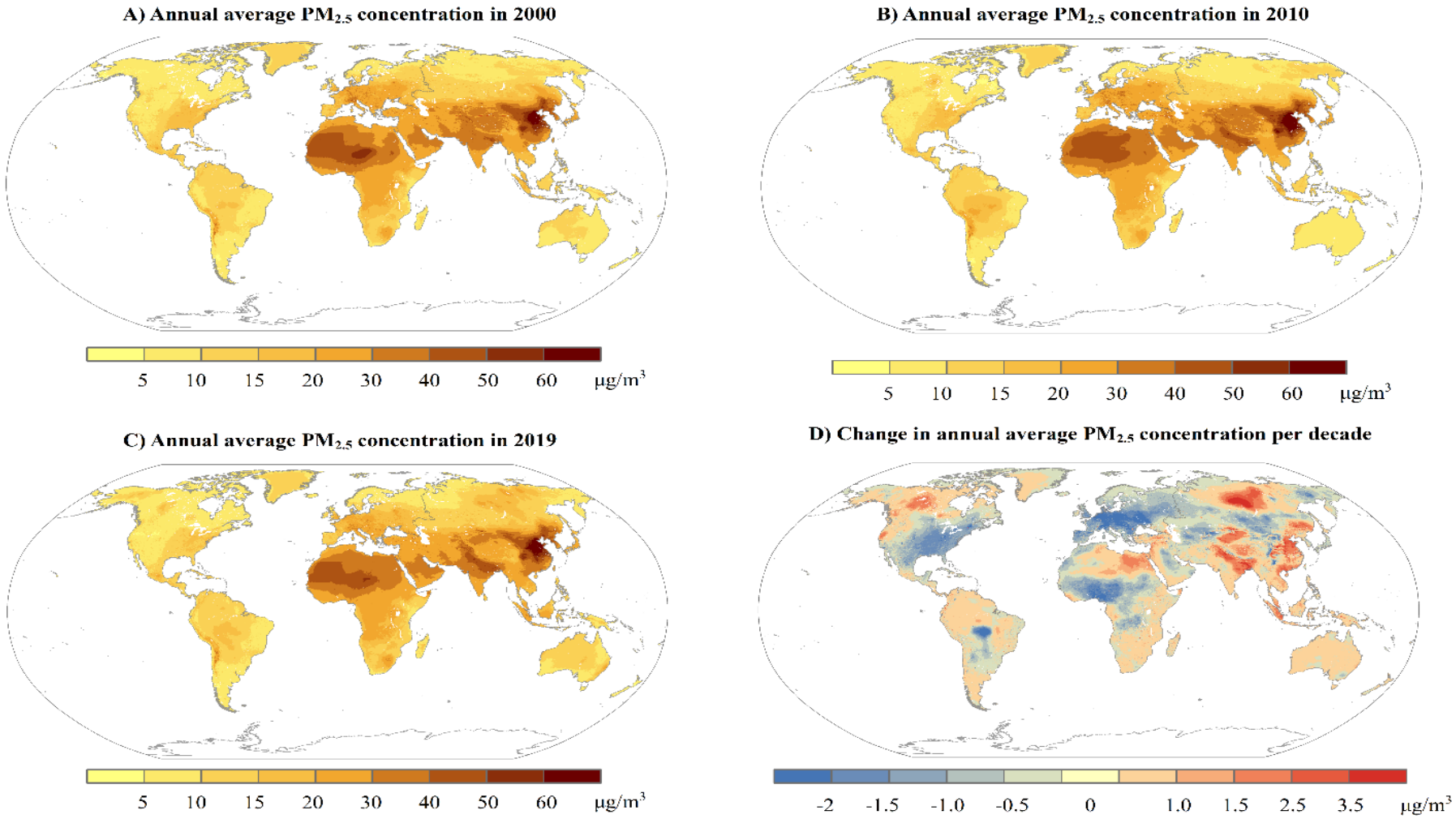
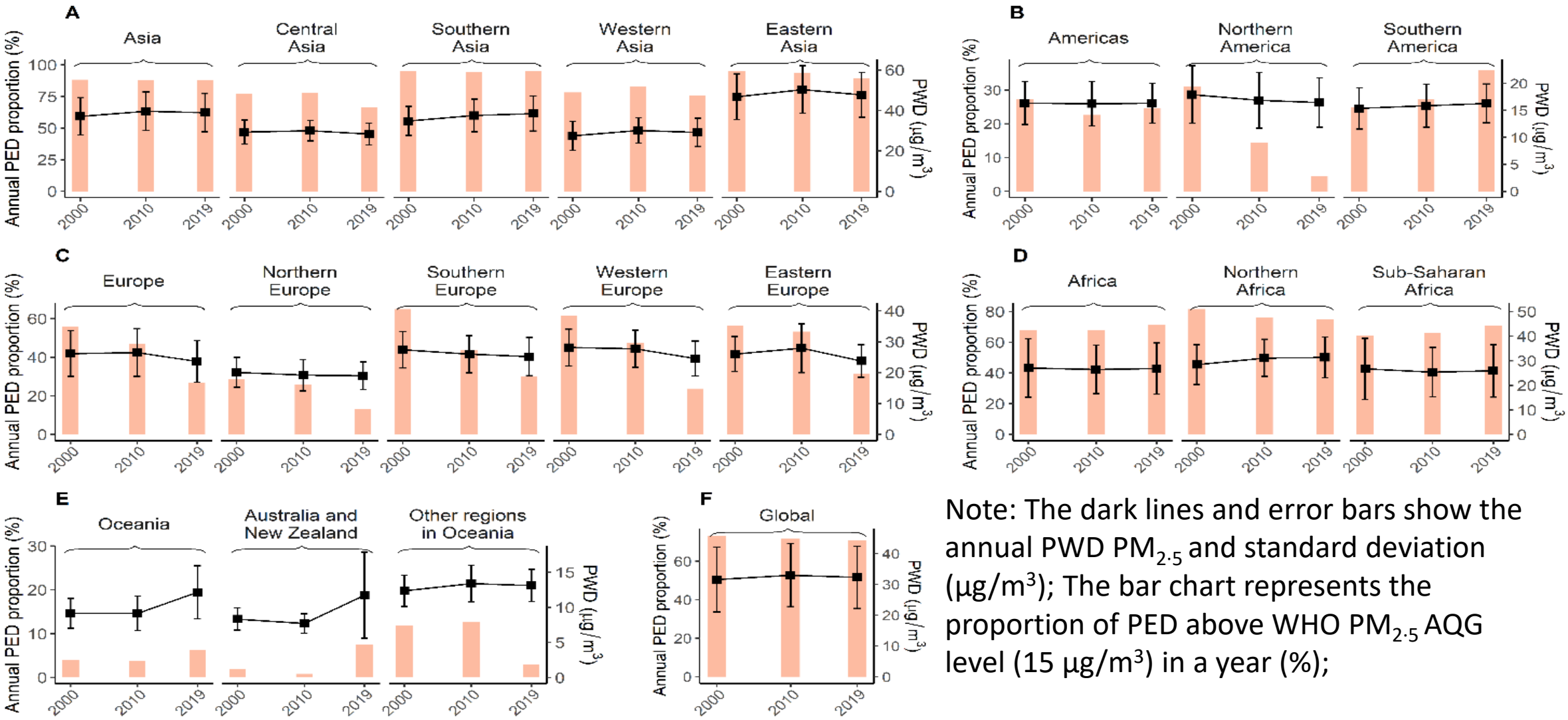


Fig 5. Annual average $PM_{2.5}$ and changes per decade in 2000-2019

MONASH University

4.2 Results

Regional $PM_{2.5}$ changes



Note: The dark lines and error bars show the annual PWD $PM_{2.5}$ and standard deviation ($\mu g/m^3$); The bar chart represents the proportion of PED above WHO $PM_{2.5}$ AQG level (15 $\mu g/m^3$) in a year (%);

Fig 6. Changes in annual $PM_{2.5}$ exposure in 2000, 2010, and 2019 by region

MONASH University

4.2 Results

Global $PM_{2.5}$ seasonal patterns

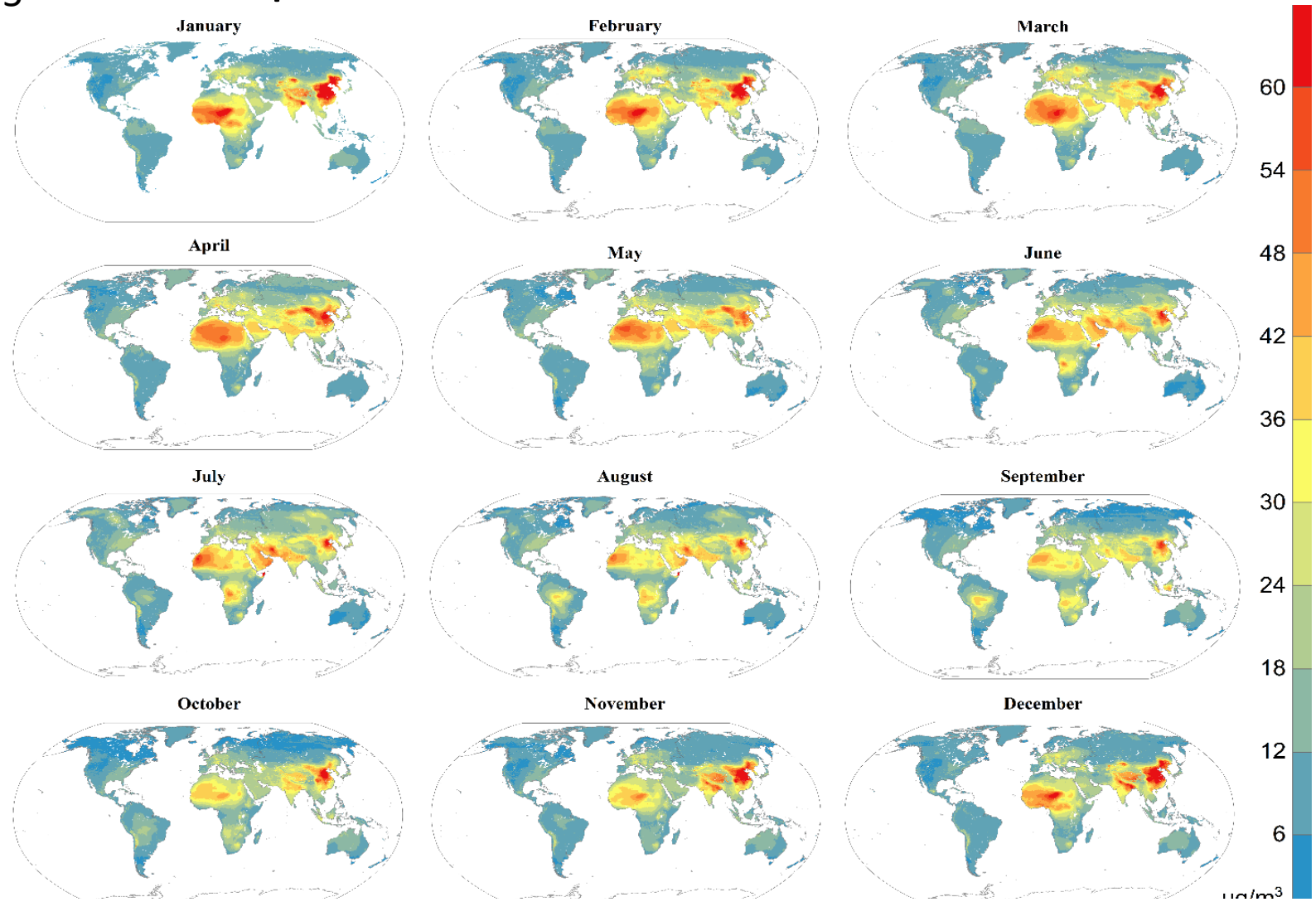
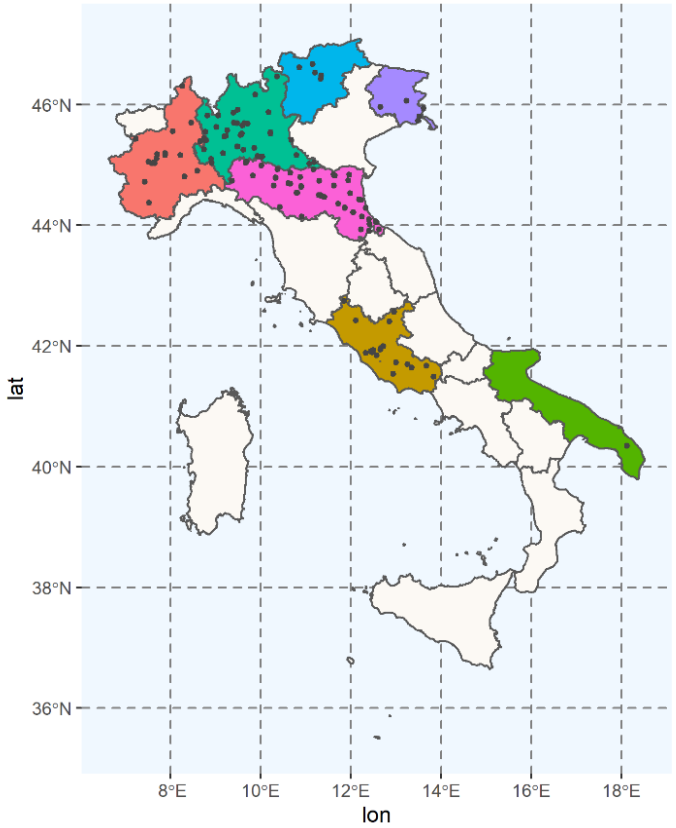


Fig 7. The global monthly average of $PM_{2.5}$ in 2000-2019

MONASH University

5. Daily PM2.5 estimation in Italy

Collected 133 monitor stations daily PM2.5 data in Italy from 2015-2019



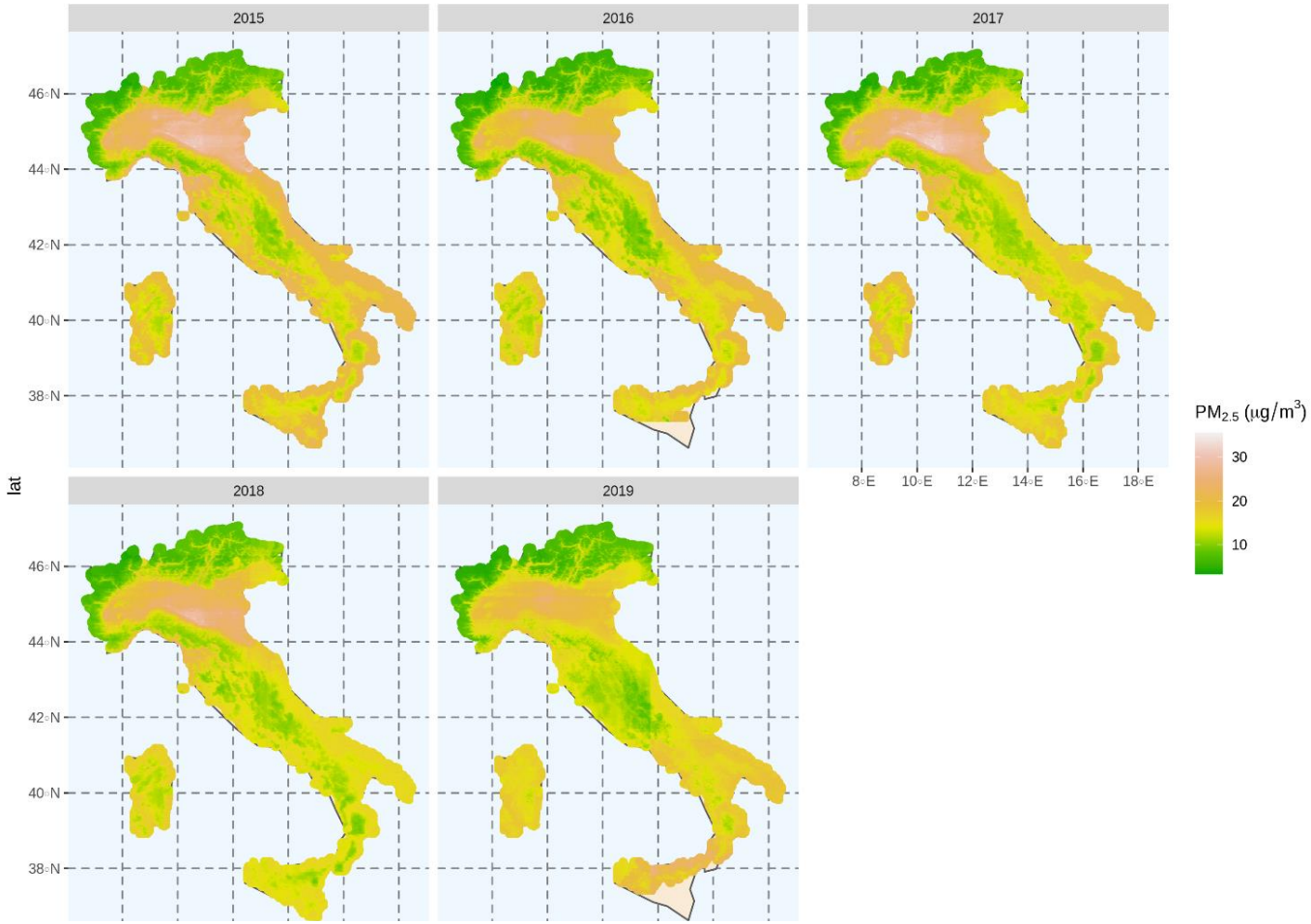


Fig 8. The annual average of $PM_{2.5}$ in Italy in 2015-2019

MONASH University

5.1 Data collection

- Aerosol optical depth (AOD)
- Climate data: Temperature, Relative humidity, precipitation, pressure…
- Land cover information
- Elevation
- Population density and others

5.2 Results

- DEML was superior to other benchmark models

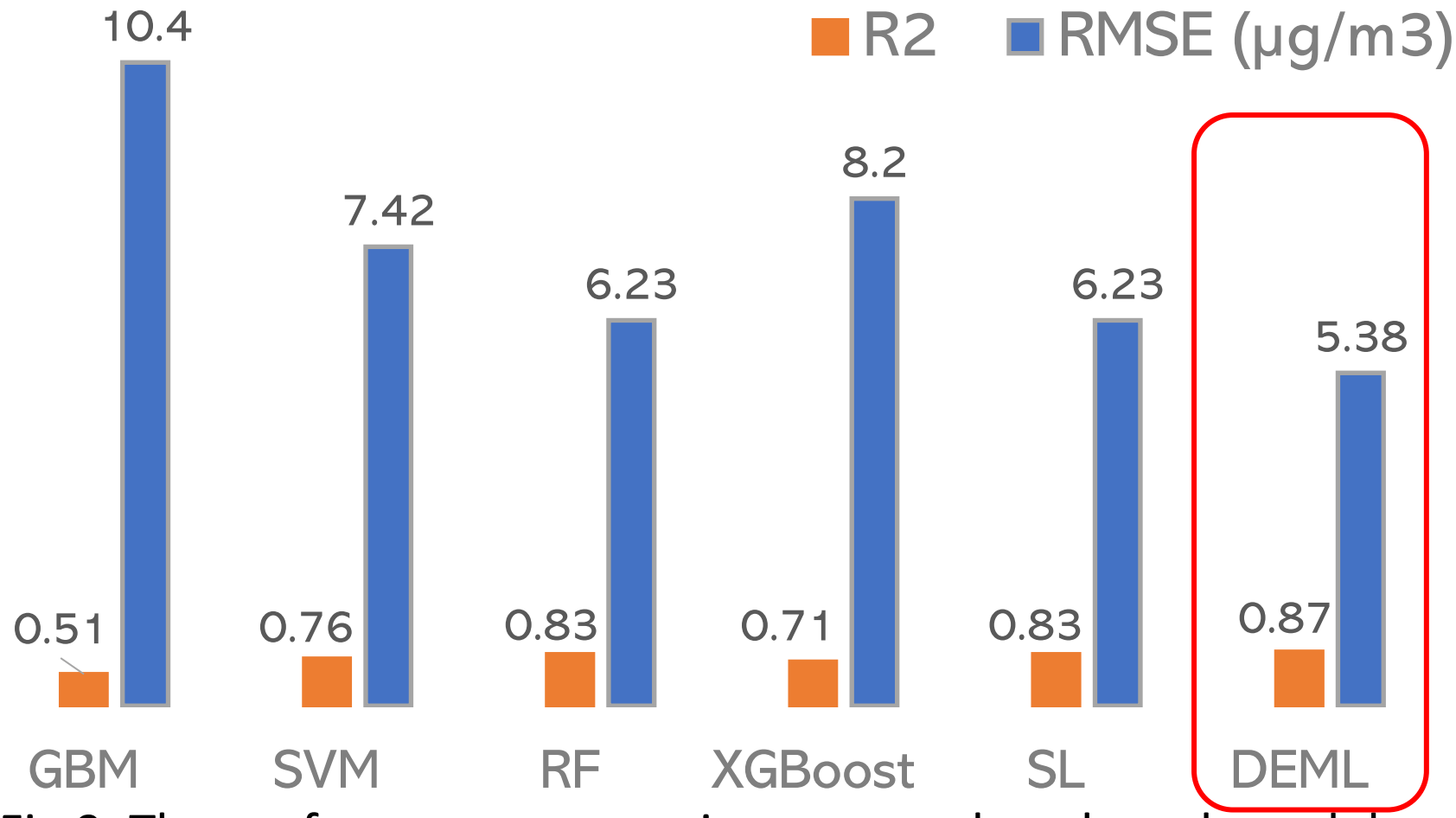


Fig 9. The performance comparison among benchmark models

MONASH University

5.2 Results

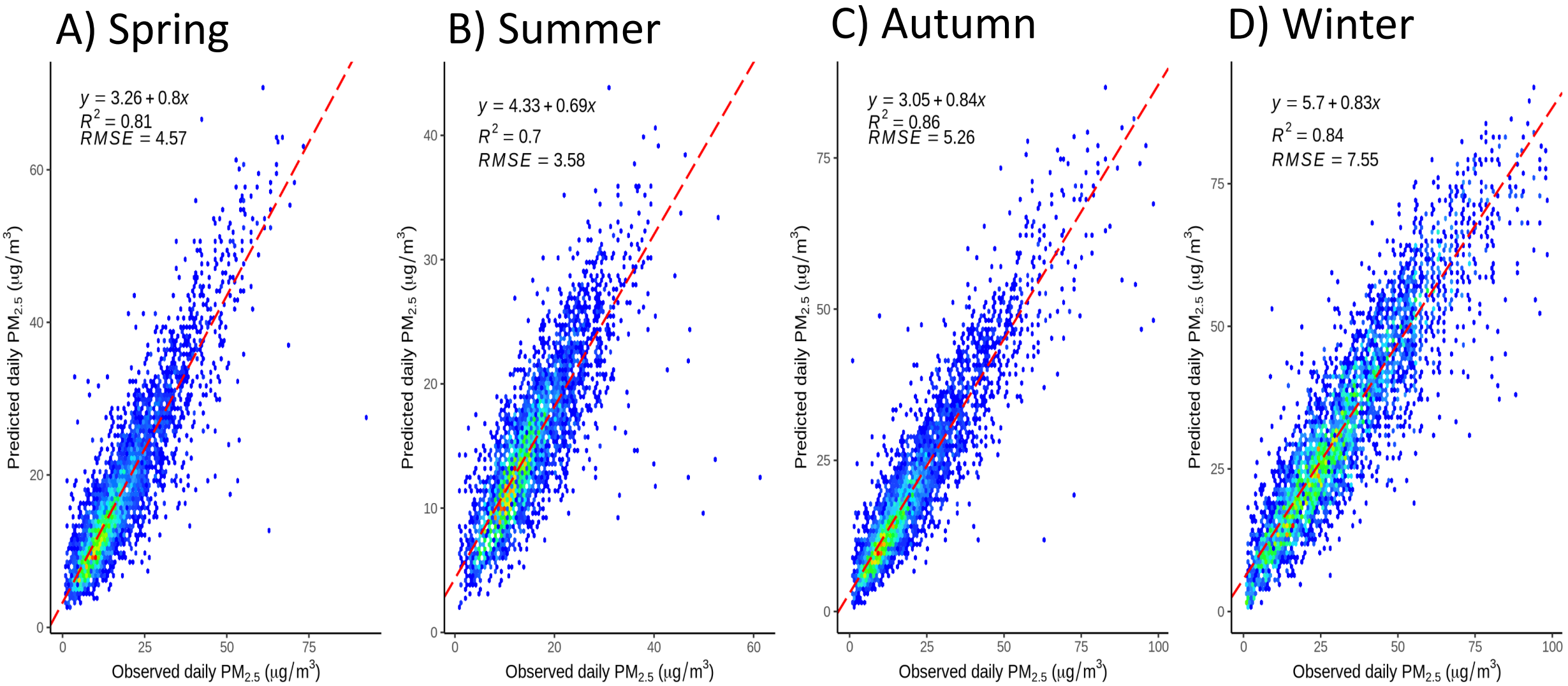


Fig 10. The performance of DEML in different seasons in Italy

MONASH University

6. DEML in time-series data forecasting

Sensor-based indoor air temperature prediction using DEML in Australia

- Aims: Forecasting hourly indoor temperature
- Adjusted DEML model architecture with several machine learning (ML) and deep learning (DL) approaches in time-series data
- Collected 96 indoor climate sensors across 25 buildings in 8 Australian cities from 2019 to 2022.

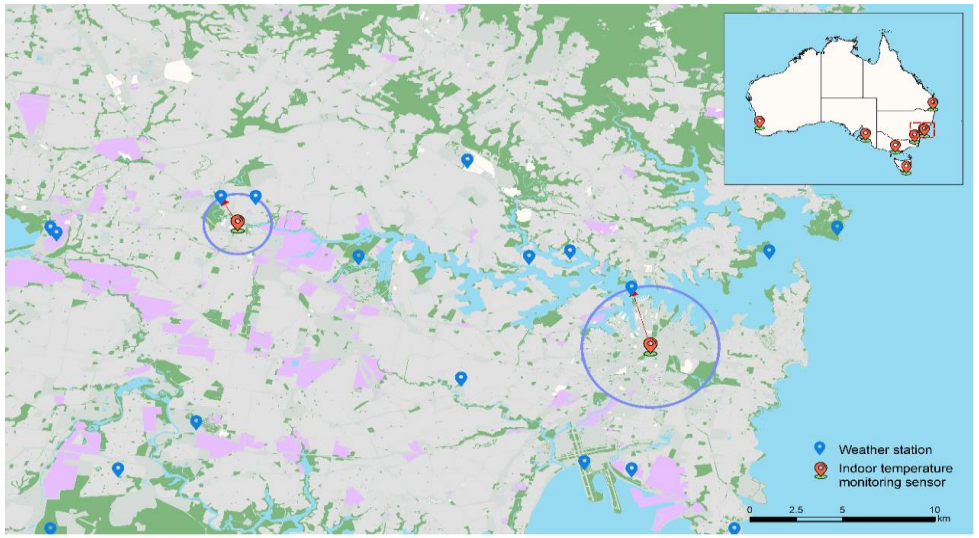


Figure 11. The distribution of indoor and outdoor temperature monitoring sensors in Sydney, Australia

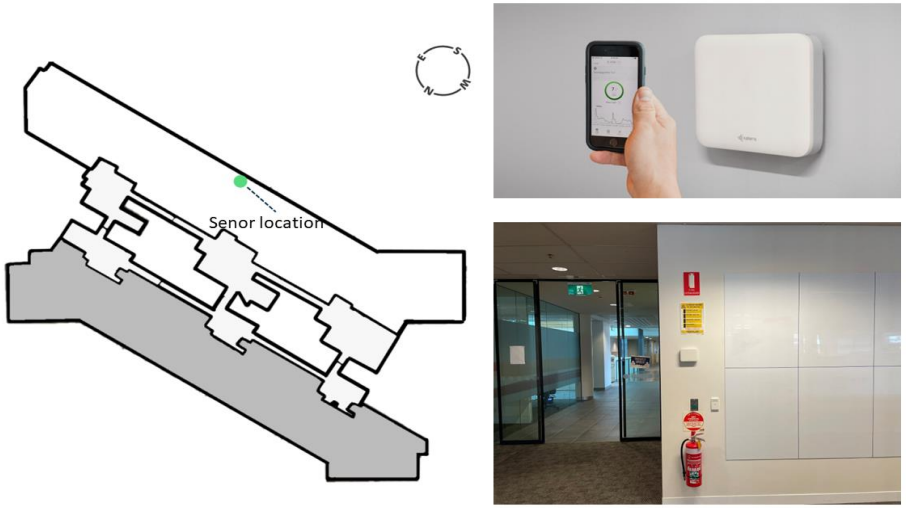


Figure 12. Indoor climate sensor installation location demonstration

MONASH University

6.1 DEML Model Structure

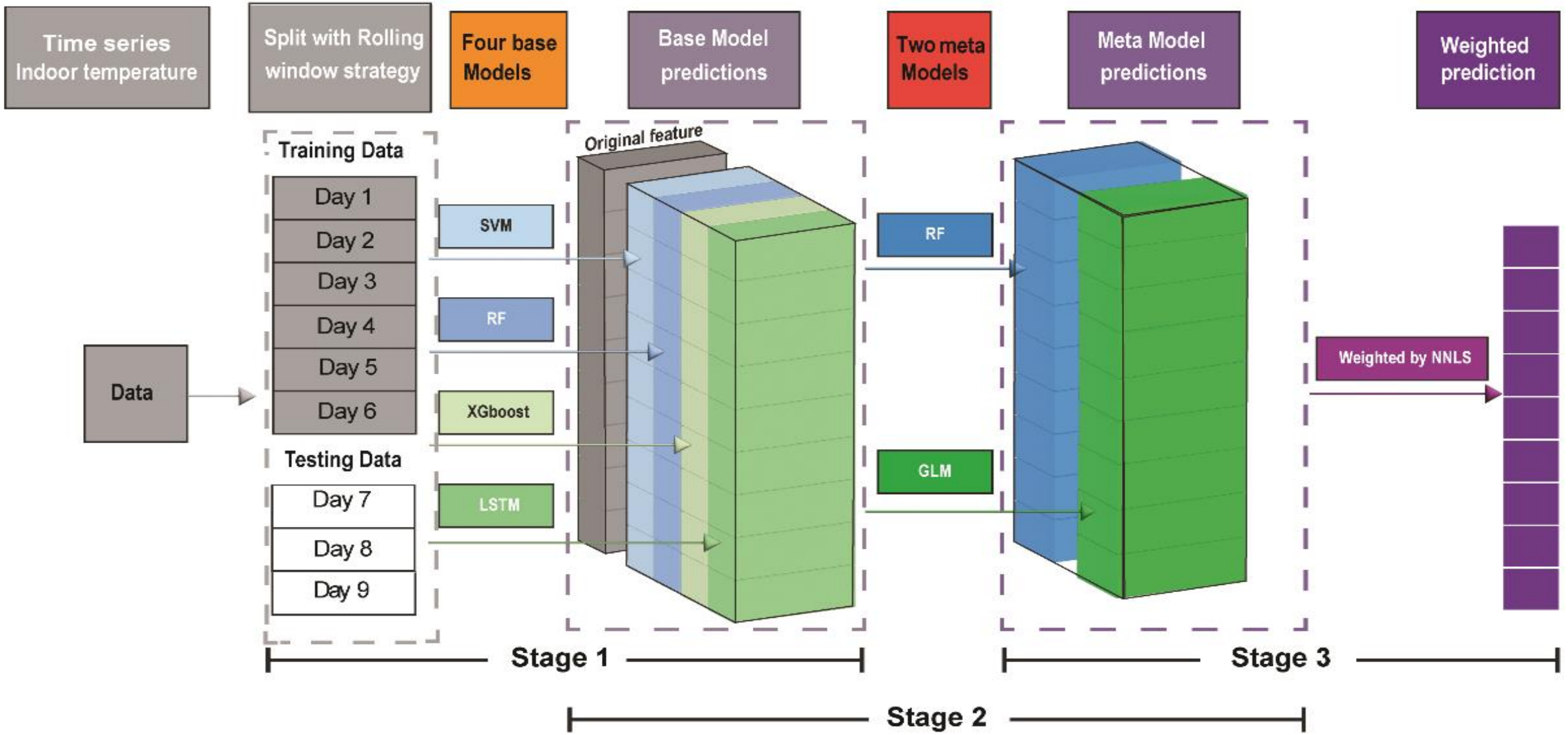


Fig 13. The framework of the DEML for indoor temperature forecasting

MONASH University

6.1 Slide moving window

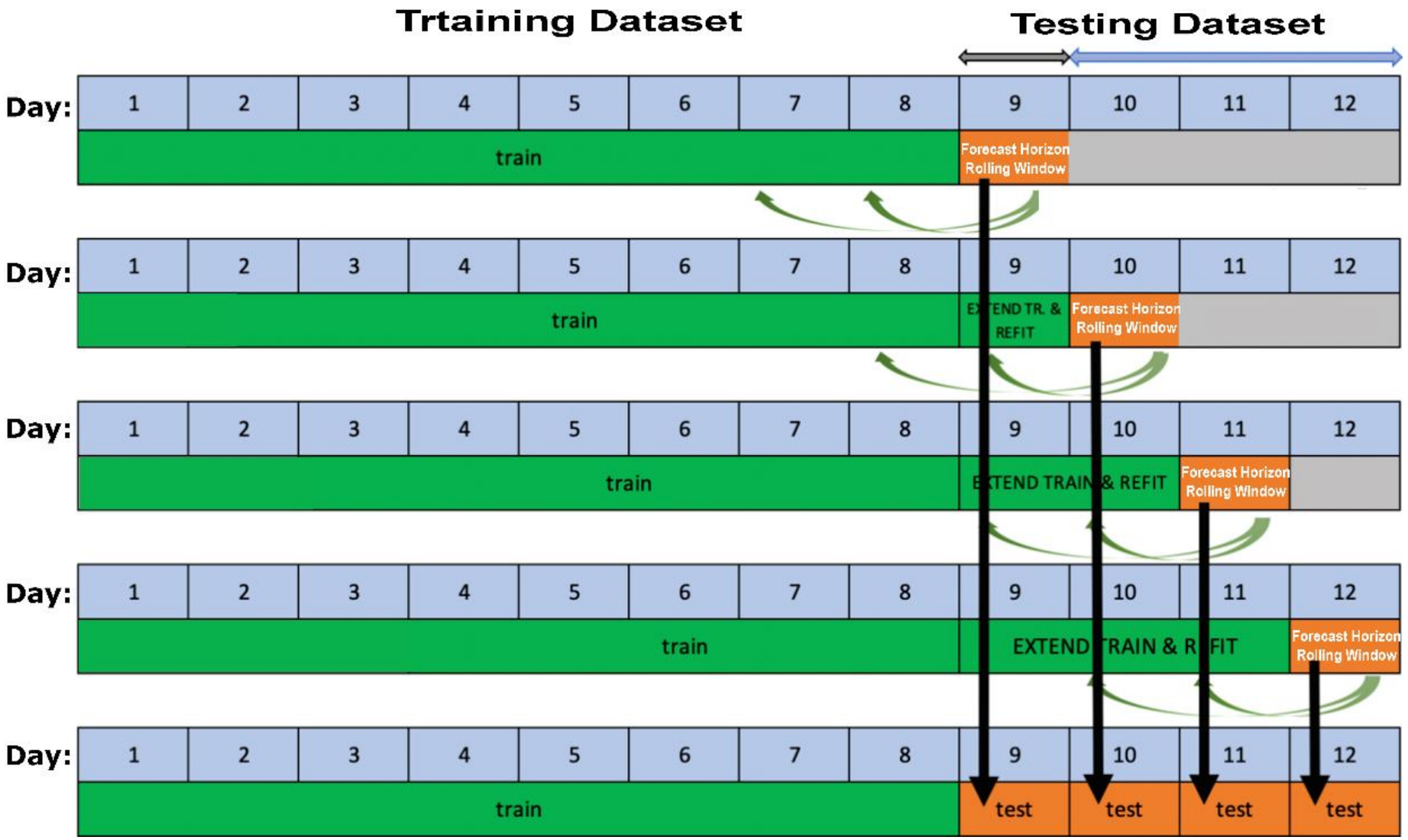


Figure 14. the illustration of the models training and testing for time series data prediction using a forecast horizon rolling window refitting process

MONASH University

6.2 Results

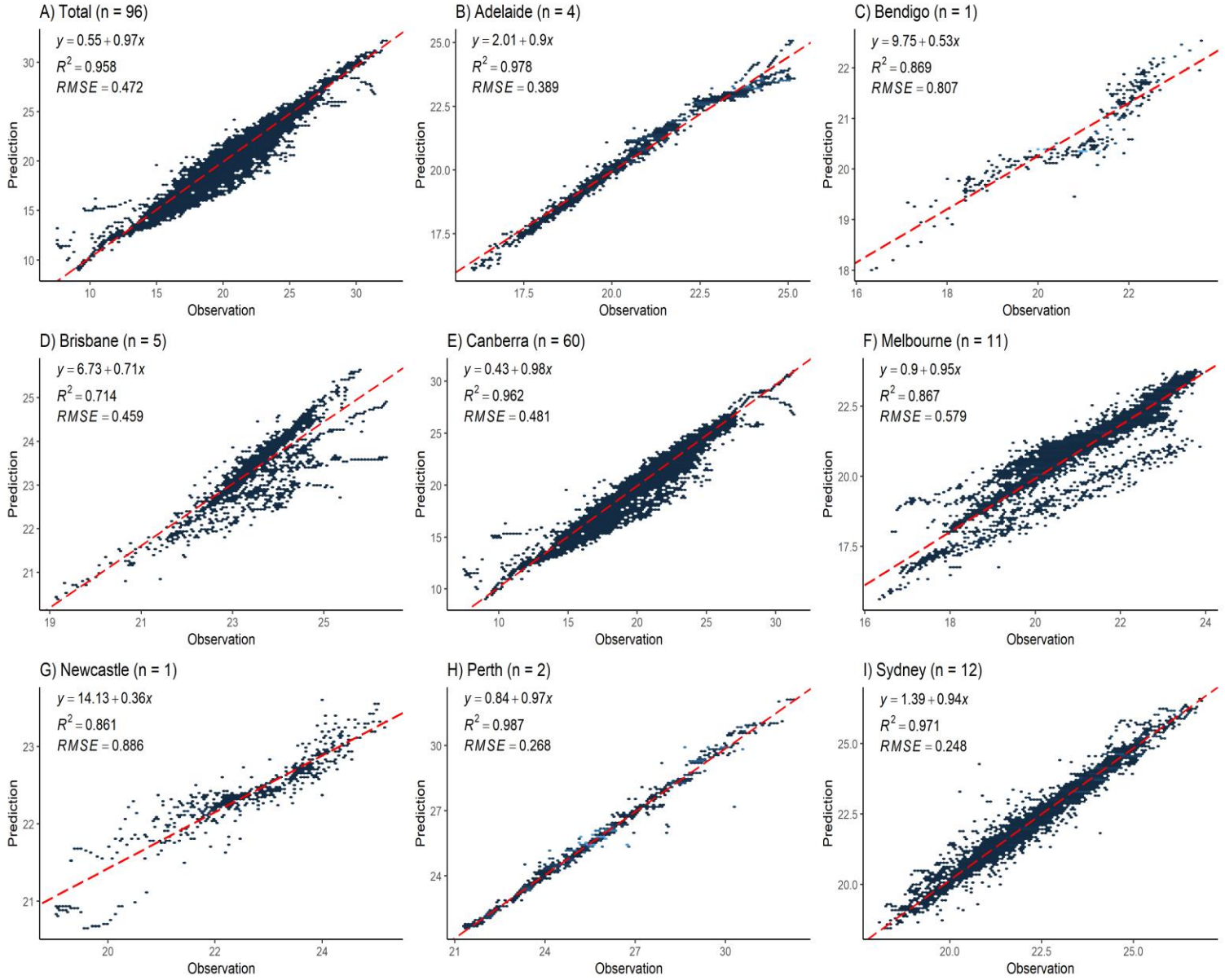


Figure 15. The comparison between observed and DEML-predicted hourly indoor temperature (℃) in eight cities in Australia in the unseen dataset

MONASH University

6.2 Results

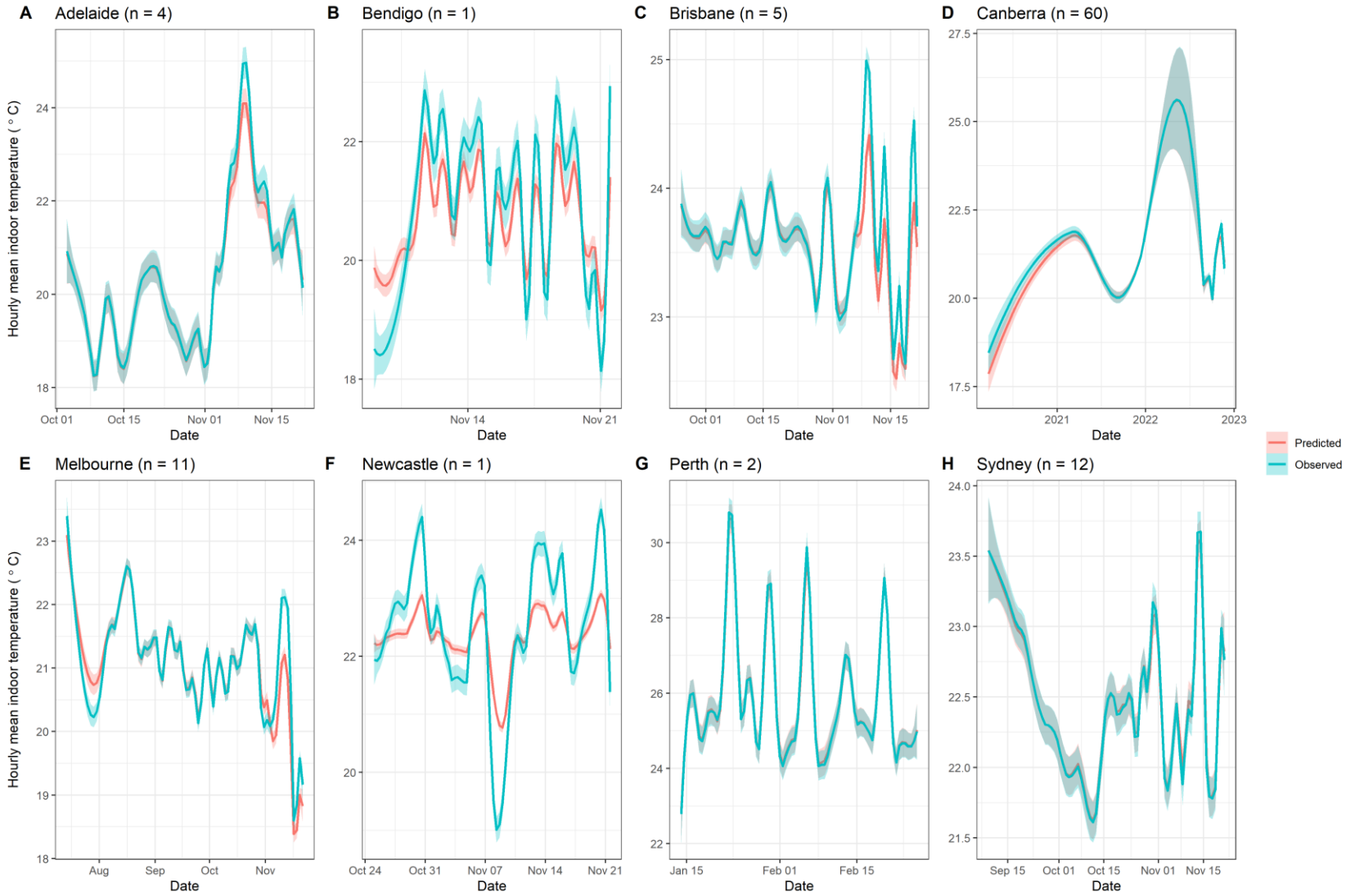


Figure 16. The comparison of the temporal trends between observed and DEML-predicted hourly indoor temperature (℃) in eight cities in Australia in the unseen dataset

MONASH University

Using Deep Ensemble Machine Learning in R

MONASH University

R ‘deeper’ Installation

Please make sure:

- using R (\>= 3.5.0)
- install certain dependent R packages: devtools, SuperLearner(\>= 2.0-28)
- install other suggested R packages: caret, skimr, CAST, ranger, gbm, xgboost

Install ‘deeper’ through following syntax:

```
library(devtools)
install_github("Alven8816/deeper")
```

Algorithms selection

Deeper include 35 algorithms which are based on 'SuperLearner' R package

parameter	algorithm	required package	types
SL.bayesglm	Bayesian generalized linear regression	arm	R
SL.biglasso	Extending Lasso Model Fitting to Big Data	biglasso	R
SL.caret	random Forest as default	caret	R
SL.caret.rpart	decision trees as default	caret	R
SL.cforest	Breiman's random forests	party	R
SL.earth	Multivariate Adaptive Regression Splines	earth	R
SL.gam	generalized additive models	gam	N
SL.gbm	generalized boosting algorithm	gbm	R
SL.glm	generalized linear models	NA	R
SL.glm.interaction	generalized linear models	NA	R
SL.ipredbagg	Bootstrap aggregation (bagging)	ipred	R
SL.kernelKnn	Kernel k Nearest Neighbors	kernelknn	C
SL.ksvm	Kernlab's SVM Algorithm	kernlab	R
SL.lda	Linear discriminant analysis, used for classification	MASS	C
SL.lm	OLS via lm(),be faster than glm()	NA	R
SL.loess	Local Regression is a non-parametric approach that fits multiple regressions in local neighborhood	loess	N
SL.logreg	Logic Regression	LogicReg	N
SL.mean	mean value	NA	R
SL.nnet	Feed-Forward Neural Networks and Multinomial Log-Linear Models	nnet	N
SL.nnls	Non-negative least squares algorithm	nnls	N
SL.polymars	Polynomial Spline Routines	polspline	R
SL.qda	Quadratic discriminant analysis, used for classification	MASS	C
SL.randomForest	random Forest	randomForest	R
SL.ranger	a fast implementation of Random Forest	ranger	R

Basic steps for DEML

- Step 1. Data preparation
- Step 2. Establish base models
- Step 3. Stacking meta models
- Step 4. Prediction based on new data set

Step 1. Data preparation

- Data collection
- Data Clean
  - missing values, extreme values, Data transforming (normalization/standardization, eg. scale, centralize, log-transform …)
- Data split (training Vs. testing)
- Identify the model structure
- Tuning parameters

Step 2. Establish base models

- To establish the base models
  - ➢ predictModel()
  - ➢ predictModel_parallel()
- tuning the parameters
  - ➢ tuningModel()

Step 3. Stacking meta models

- To establish the Meta models
  - stack_ensemble()
  - stack_ensemble.fit()
  - stack_ensemble_parallel()

Step 4. Prediction based on a new data set

- To predict the unseen data set
  - ➢ deeper::predict()
- return the point scatter plot
  - ➢ assess.plot()

Wenhua.yu@monash.edu

Presentation Materials and R code can be downloaded here:
https://github.com/Alven8816/DEML_PM2.5_estimation

MONASH University

Model hyper-parameters and computing cost information for DEML

Model	Main Hyper-parameters	R package	Time cost for 10,000 cases
GLM	alpha = 1, nlambda = 100	glmnet	44.1 secs
GBM	n.trees = 100, interaction.depth = 1,n.minobsinnode = 10, shrinkage = 0.1, bag.fraction = 0.5,	gbm	14.15 mins
SVM	gamma = 0.1, tolerance=0.001, epsilon =0.1	e1071	2.97 mins
RF	mtry=4, num.trees = 500	ranger	30.4 secs
XGBoost	ntrees = 1000, max_depth = 4, shrinkage = 0.1, minobspernode = 10	XGBoost	5.00 mins
SL	The same as GBM,SVM, RF, and XGBoost	SuperLearner	6.35 mins
DEML	The same as GLM, GBM,SVM, RF, and XGBoost	deeper	11.98 mins

The CPU time cost was calculated using Intel(R) Core(TM) i7-1165G7 @ 2.80GHz with 4 physical cores paralleling computing.

MONASH University

Random Forest

- 在Bagging中,我们基于Boostrap的抽样训练了B棵决策树,在每棵决策树里，树的每一次分裂都考虑了所有的Feature。
- 随机森林里，树的分裂不是考虑所有的Feature,而是只考虑部分Feature。
  - 采样:
    - 行采样，采用有放回的方式，也就是在采样得到的样本集合中，可能有重复的样本。假设输入样本为N个,那么采样的样本也为N个。
    - 列采样，从M个feature中,选择m个(m « M)，有放回采样
  - 完全分裂

    不需要树的剪枝。决策树的某一个叶子节点要么是无法继续分裂的，要么里面所有样本都指向同一个分类。
  - 由于之前的两个随机采样的过程保证了随机性,所以就算不剪枝,也不会出现over-fitting。

Random Forest 优缺点

- 优点:
  - 可以进行特征重要性排序
  - 对数据缺失值、异常值不敏感，适用性广泛
  - 对数据类型不敏感，可用于分类和回归分析
  - 可处理大型高维数据，可以方便进行并行化计算
  - 可应用与不平衡数据
  - 可用于非监督学习中的聚类分析、异常值检测等
  - 减少了方差，模型不易过拟合
- 缺点:
  - 模型不易解释，存在黑箱理论
  - 对于回归问题，模型预测值不能超处训练数据范围

Bagging 与Boosting的区别

- 取样方式不同
  - Bagging采用均匀取样，训练集随机选择，各轮训练集之间相互独立;
  - Boosting根据错误率来取样，各轮训练集的选择与前面各轮的学习结果有关，Boosting的分类精度要优于Bagging。
- 预测函数不同
  - Bagging的各个预测函数没有权重;
  - Boosting是有权重的;
- 学习效率不同
  - Bagging的各个预测函数可以并行生成，并行训练节省大量时间开销
  - 而Boosting的各个预测函数只能顺序生成。bagging和boosting都可以有效地提高分类的准确性。在大多数数据集中，boosting的准确性比bagging高。在有些数据集中，boosting会引起退化— Overfit。

Reference

1. Yu, W., Li, S., Ye, T., Xu, R., Song, J., & Guo, Y. (2022). Deep Ensemble Machine Learning Framework for the Estimation of PM 2.5 Concentrations. Environmental health perspectives, 130(3), 037004.
2. Ma, Z., et al. (2022). "A review of statistical methods used for developing large-scale and long-term PM2. 5 models from satellite data." Remote Sensing of Environment 269: 112827.
3. Yu, W., et al. (2023). "Global estimates of daily ambient fine particulate matter concentrations and unequal spatiotemporal distribution of population exposure: a machine learning modelling study." The Lancet Planetary Health 7(3): e209-e218.
4. 李航，《统计学习方法》，清华大学出版社，2012
5. Statistics Versus Machine Learning – Larry Wasserman：
6. https://www.52ml.net/14463.html
7. 从统计学角度看待机器学习： https://www.52ml.net/14472.html
8. 机器学习与统计学的区别：http://www.52ml.net/14518.html

MONASH University