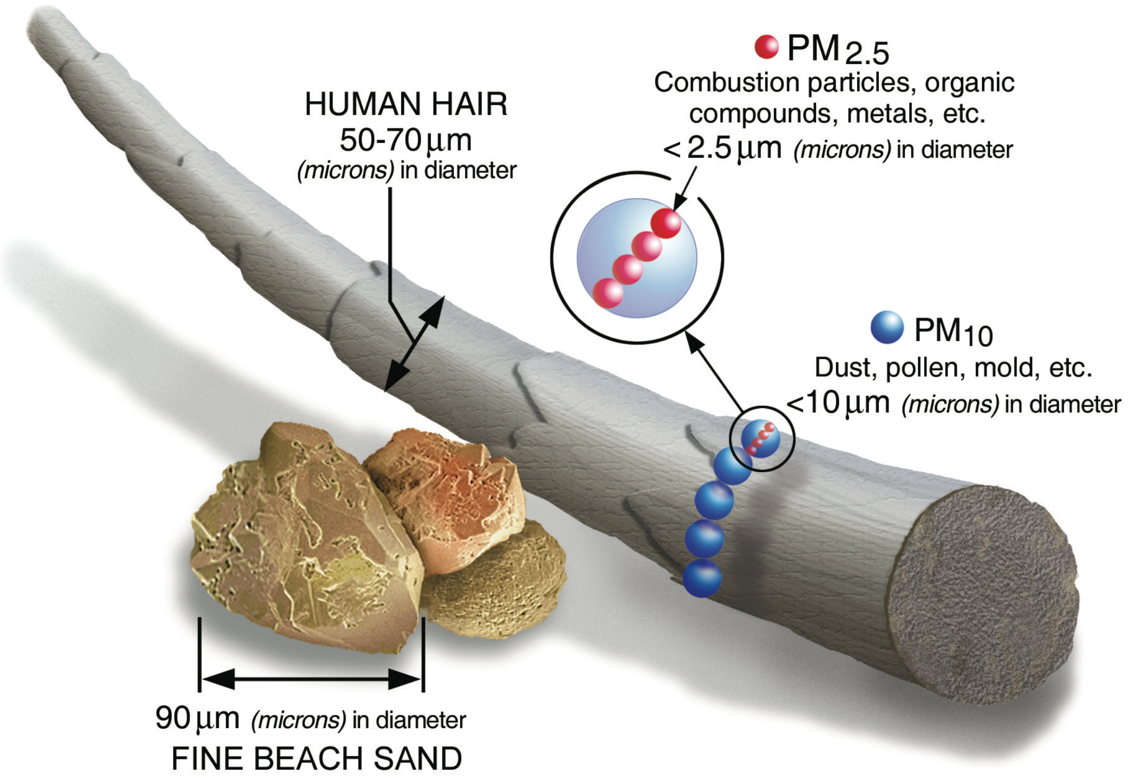
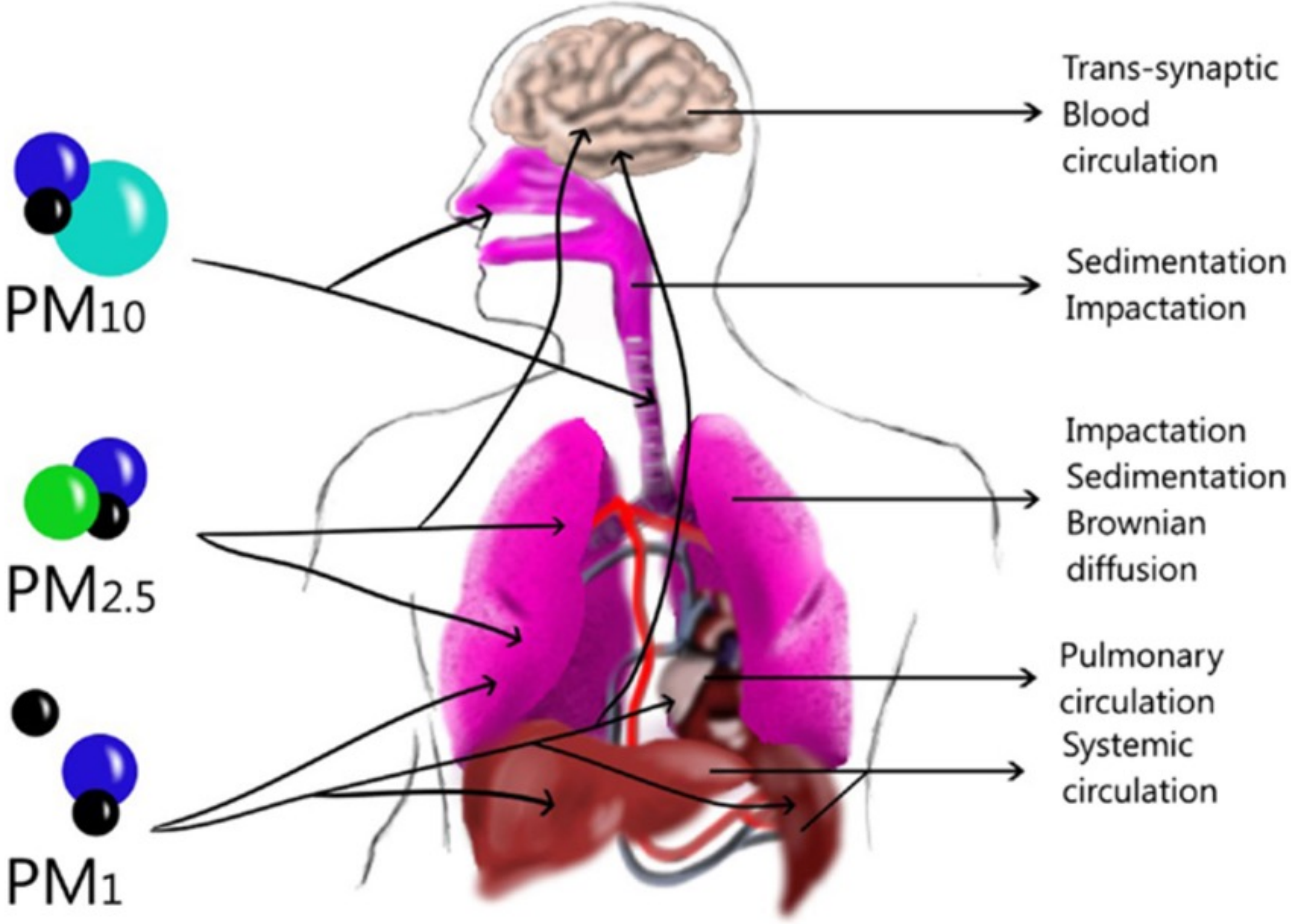
Using machine learning to map fine particulate matter air quality in East Asia

Drew Pendergrass | AGU Fall Meeting 2021

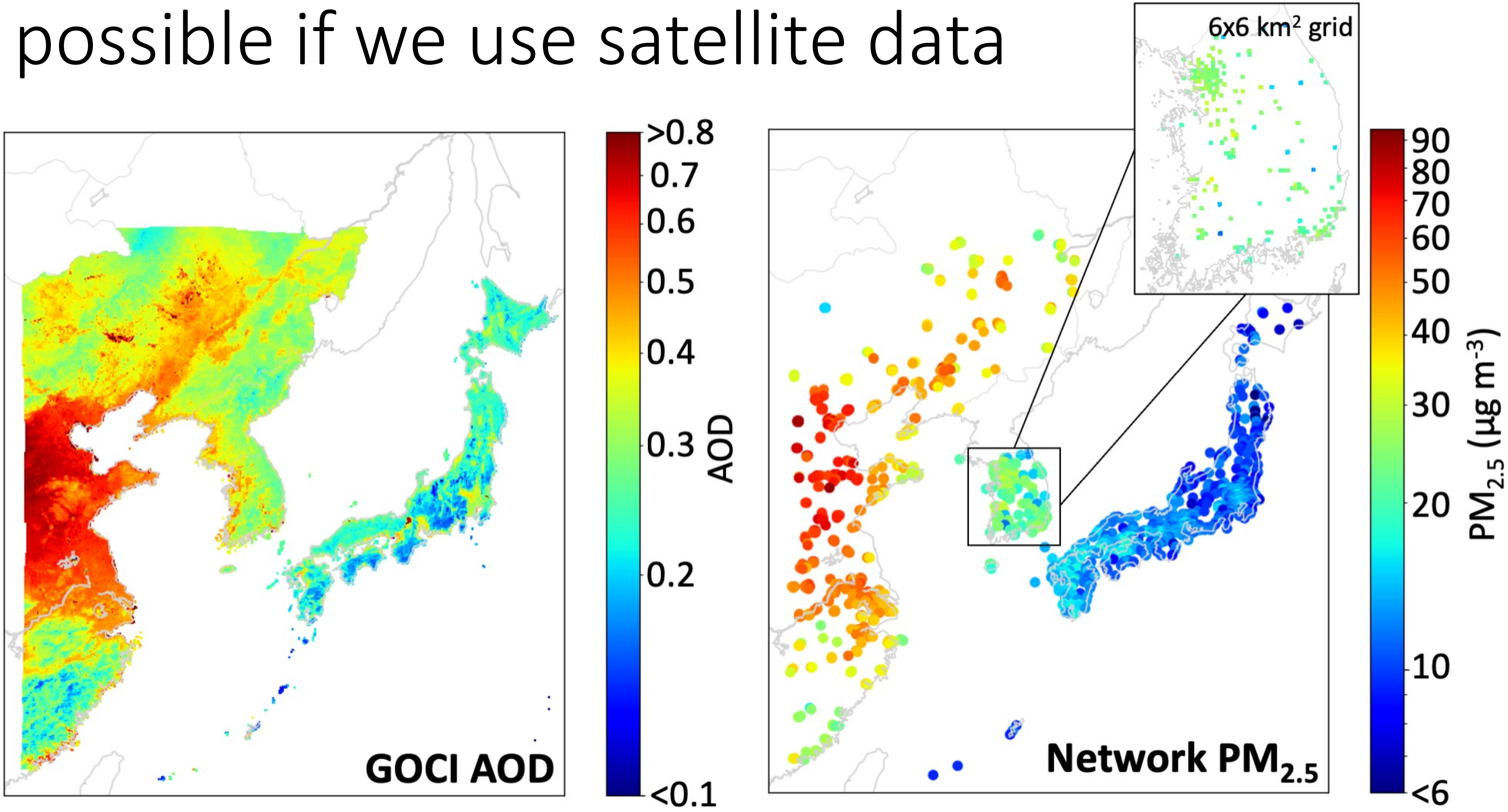
with Daniel Jacob, Shixian Zhai, Jhoon Kim, Ja-Ho Koo, Seoyoung Lee, Minah Bae, and Soontae Kim

Why care about fine particulate matter?





A huge increase in spatial coverage of $PM_{2.5}$ is possible if we use satellite data

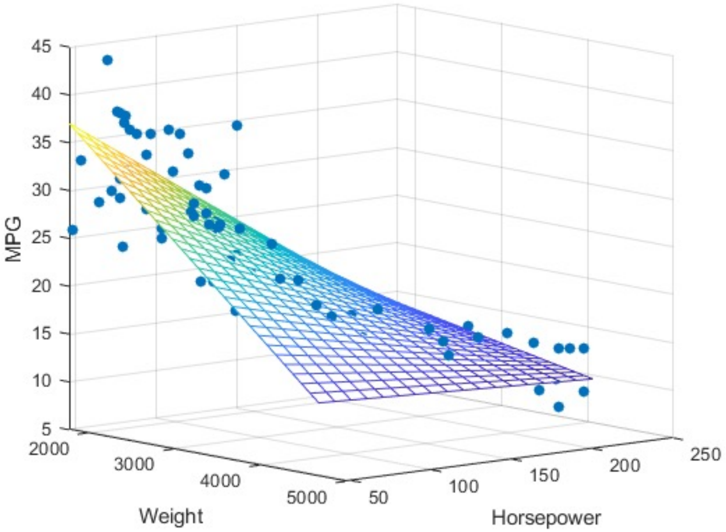


Linking satellite AOD to surface $PM_{2.5}$ is challenging

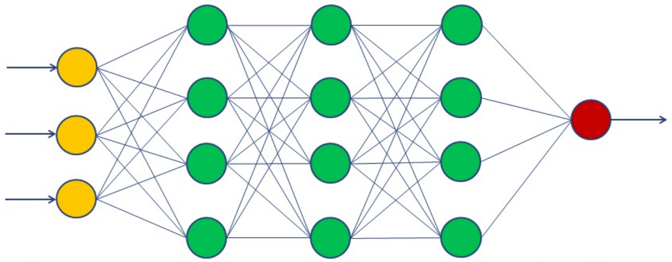
Machine learning

Chemical transport models

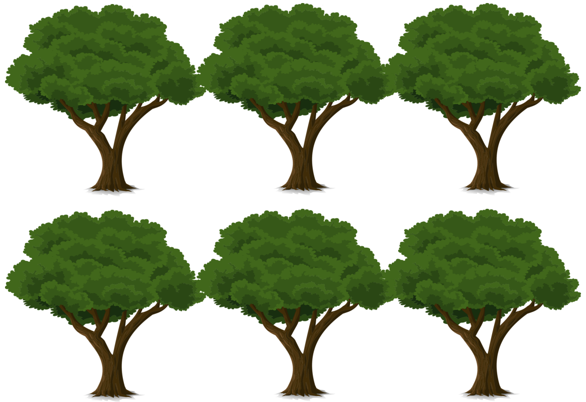
Multi-linear regression



Artificial neural networks



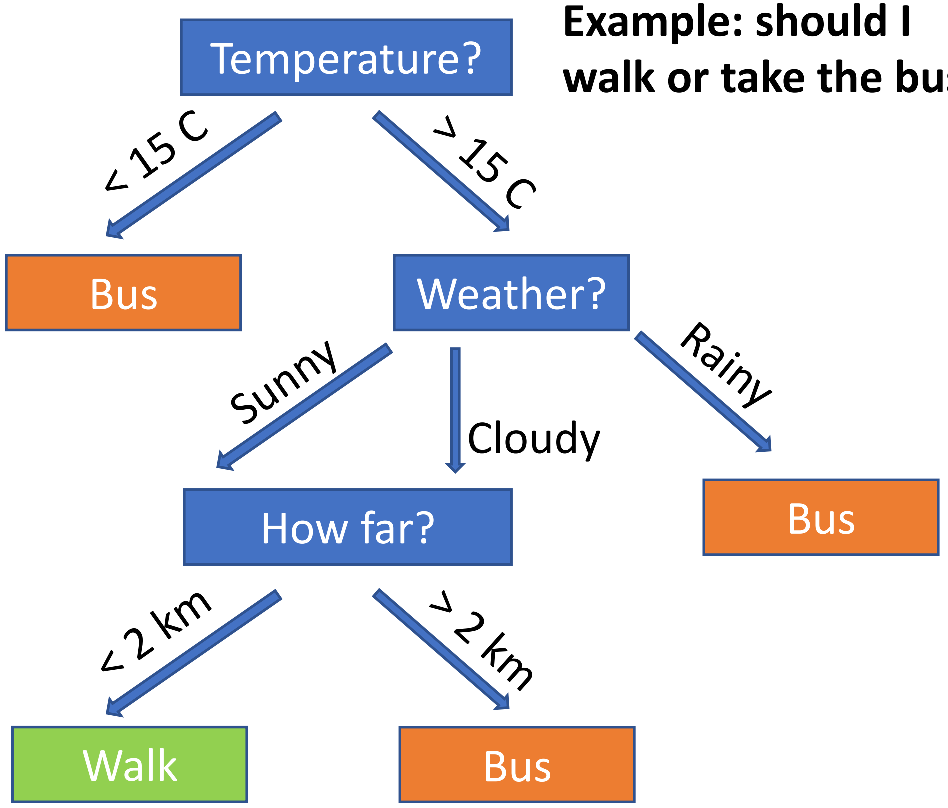
Random forests



Our algorithm choice: random forest machine learning method

A random forest is an ensemble of uncorrelated *decision trees*…

Example: should I walk or take the bus?



…trained on a series of input vectors $\vec{x}$ each with target value y…

$$\begin{bmatrix} 3 \\ 1.4 \\ \vdots \\ 2 \\ 1 \end{bmatrix} \rightarrow 2.3$$

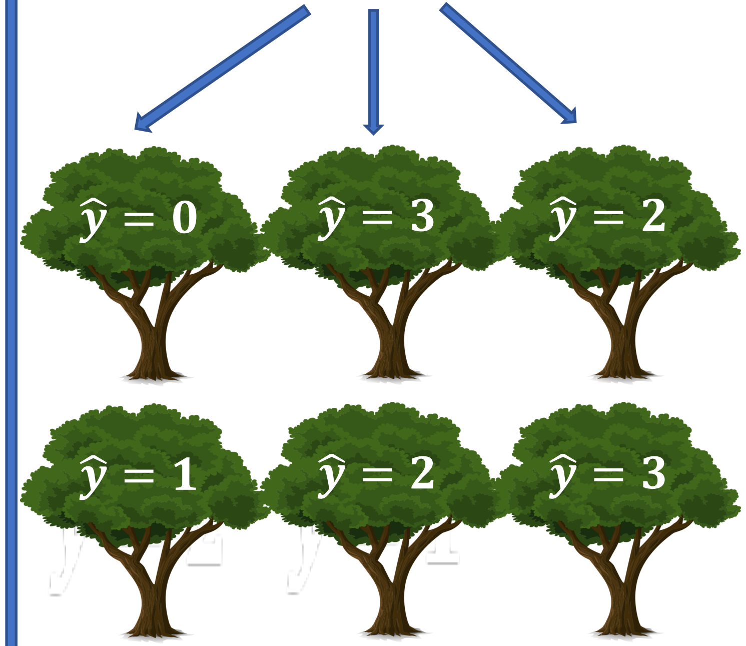
$$\begin{bmatrix} 0 \\ 1 \\ \vdots \\ 2.4 \\ 1 \end{bmatrix} \rightarrow 6.3$$

$$\begin{bmatrix} 2 \\ 7 \\ \vdots \\ 3 \\ 0 \end{bmatrix} \rightarrow 1.2$$

$$\begin{bmatrix} 1 \\ 3.5 \\ \vdots \\ 0.9 \\ 0 \end{bmatrix} \rightarrow 4.7$$

…that average their predicted $\hat{y}$ made from the same input data

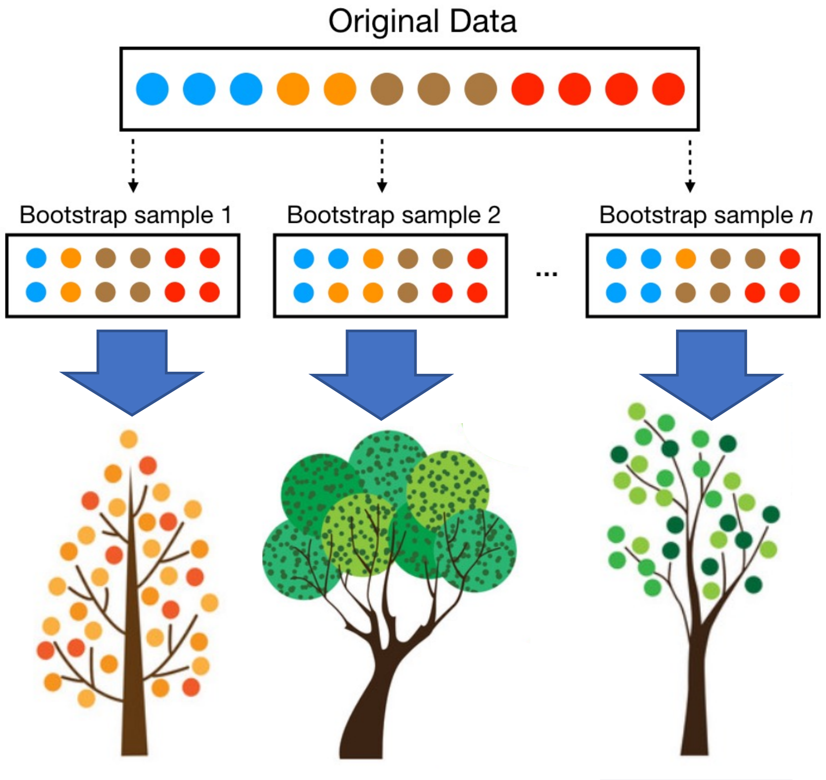
Input: $\vec{x}$



Random forest predicts 1.83

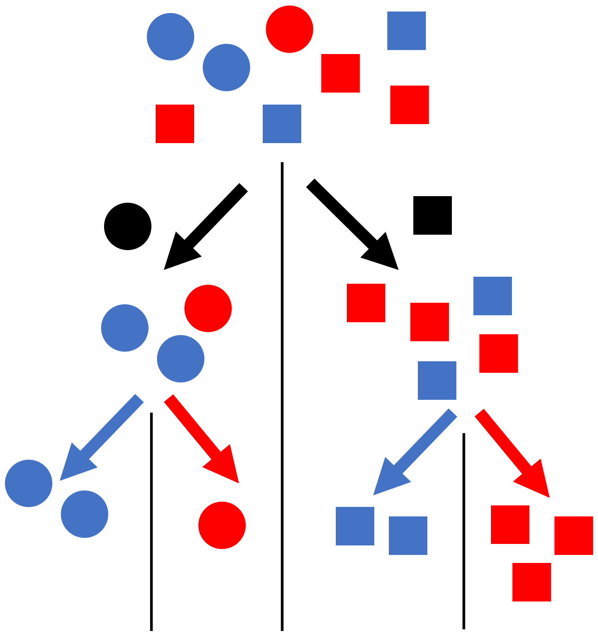
What makes the random forest random? And why does it work?

Step 1: Draw a bootstrap sample with replacement from the training data



Step 2: Grow different decision trees for each bootstrap sample

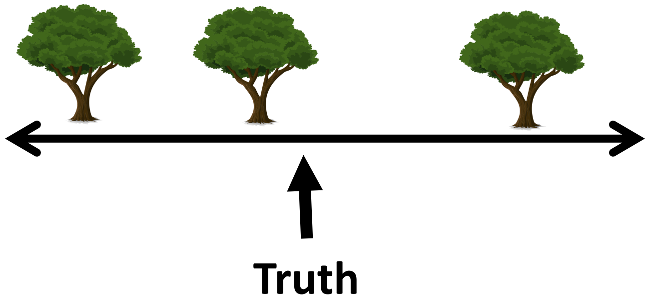
Each tree is trained **recursively**



Splits at each phase in training **minimize error**

Splits chosen are **highly sensitive** to input data

Step 3: Average tree output to make prediction.



Trees make a wide variety of guesses but on average they are **unbiased**.

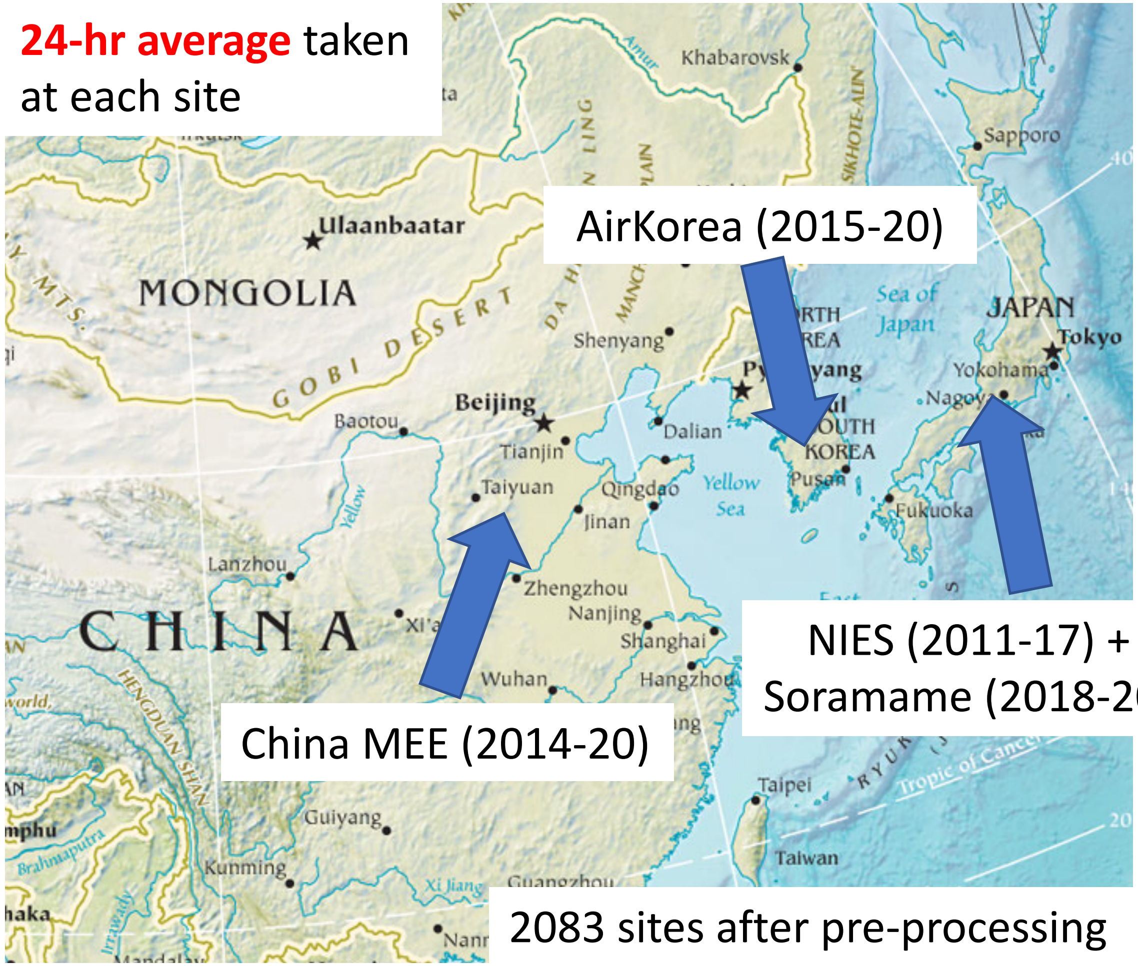
Averaging many trees should give an accurate estimate.

Data sources for training algorithm

Target value y

Ground $PM_{2.5}$ data

24-hr average taken at each site



AirKorea (2015-20)

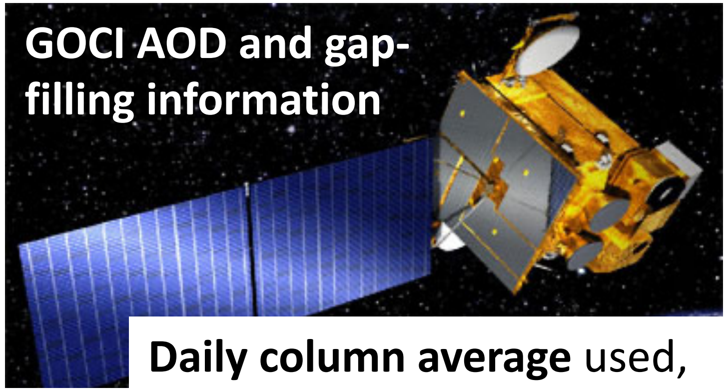
China MEE (2014-20)

NIES (2011-17) + Soramame (2018-20)

2083 sites after pre-processing

Input vector $\vec{x}$

Remote/reanalysis data



Daily column average used, missing data removed

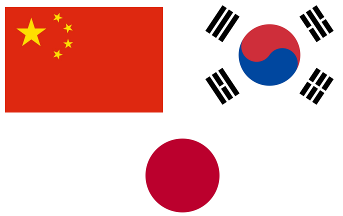
ECMWF

ERA5 quarter degree products:

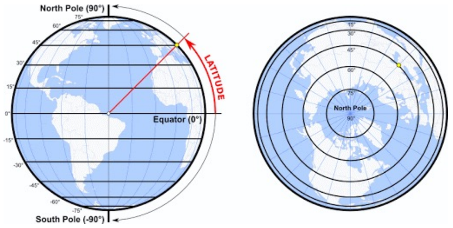
- Relative humidity
- Surface u/v wind
- 2m temperature
- Sea level pressure
- Boundary layer height

Other data

Day of year to capture seasonality

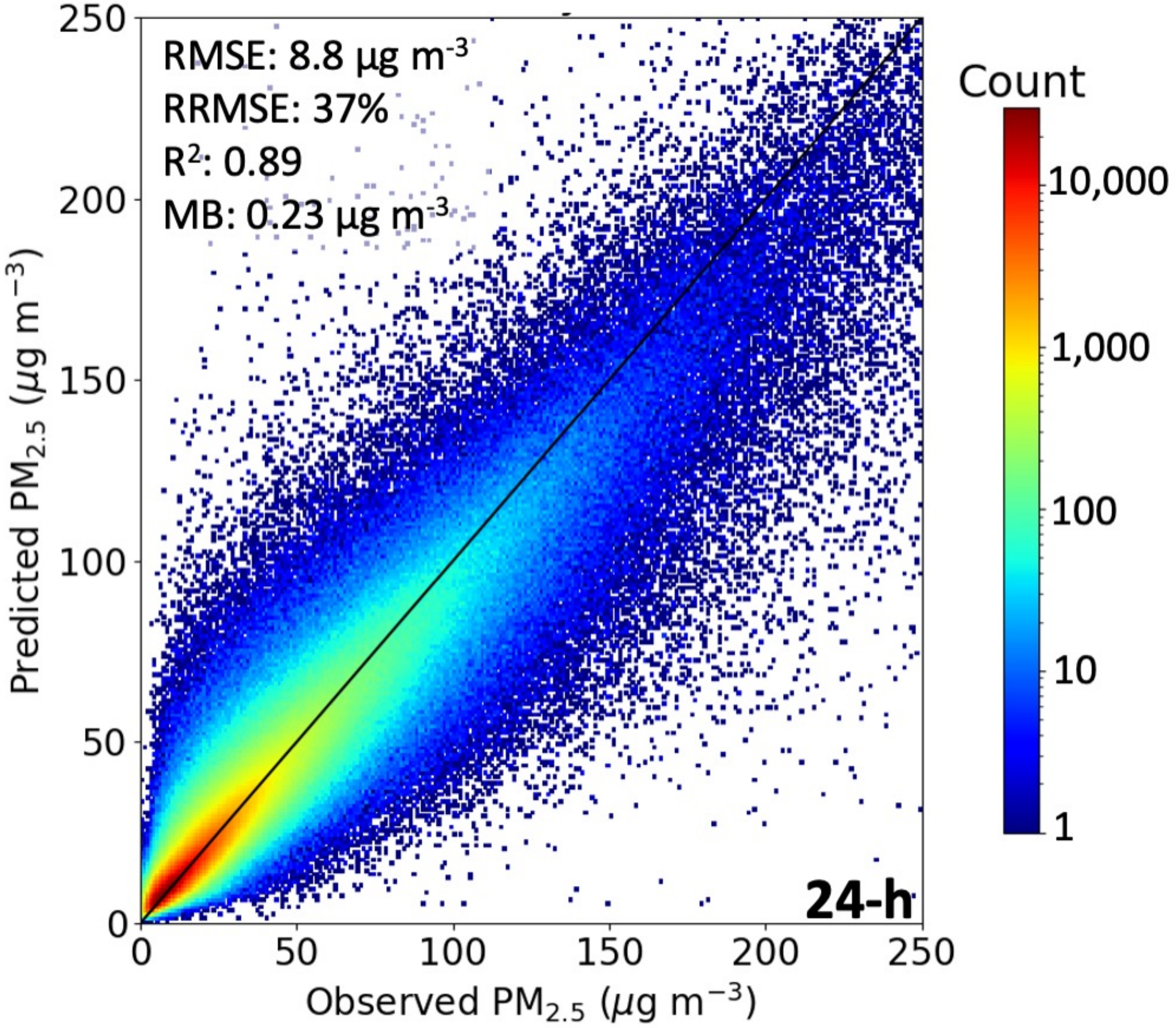


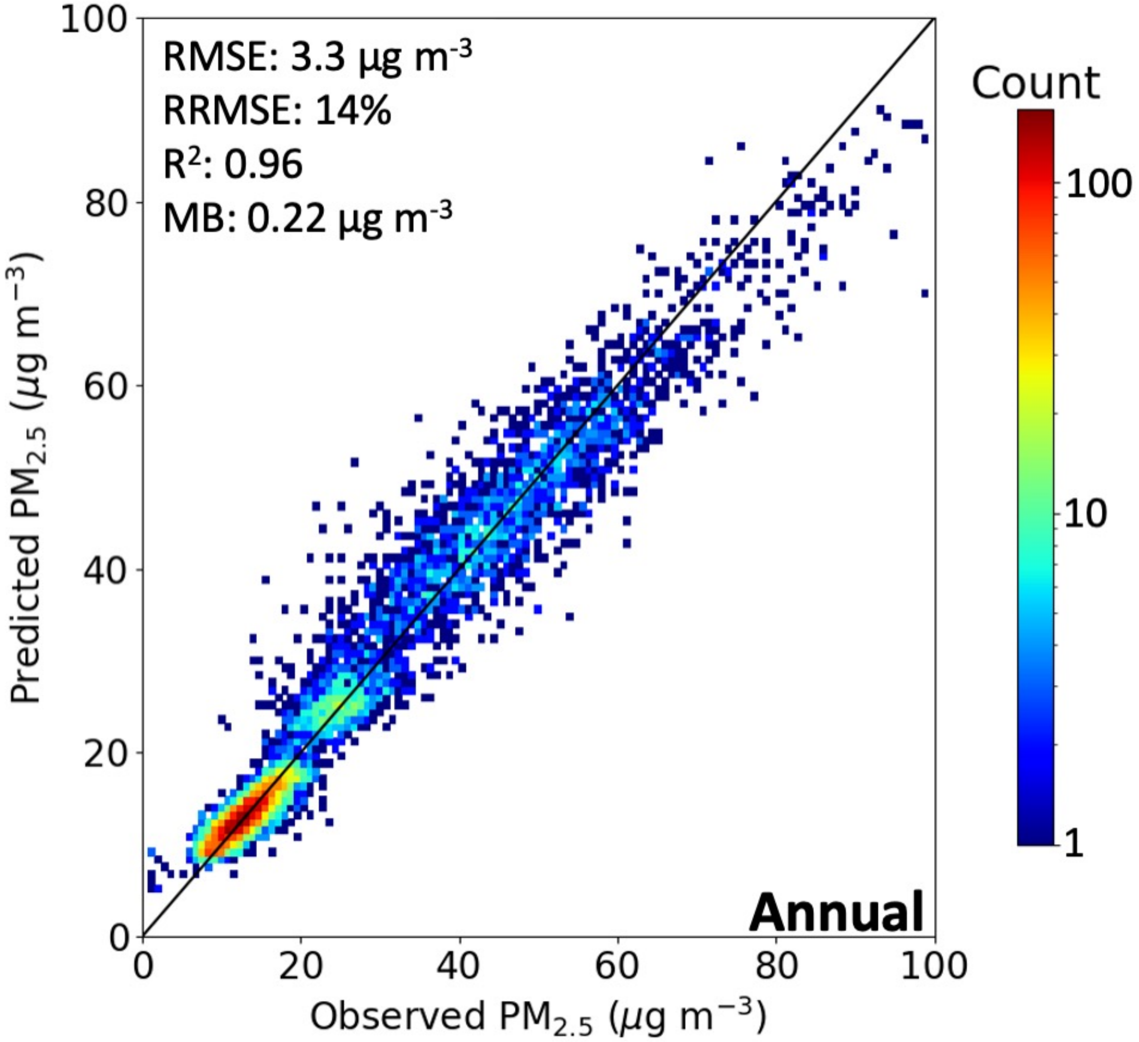
Nation



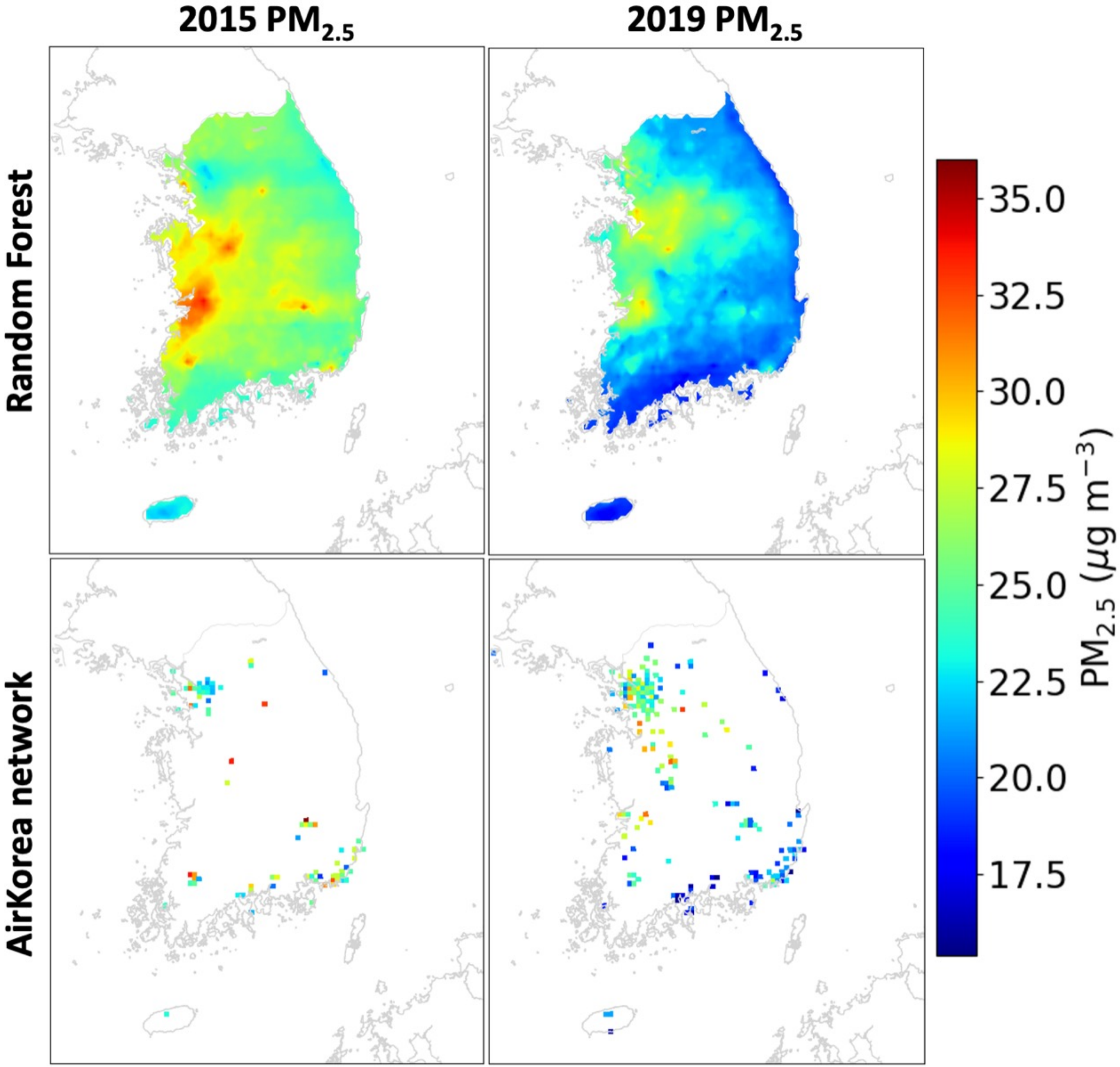
Latitude

Accuracy on both daily and annual resolution compares favorably to the literature

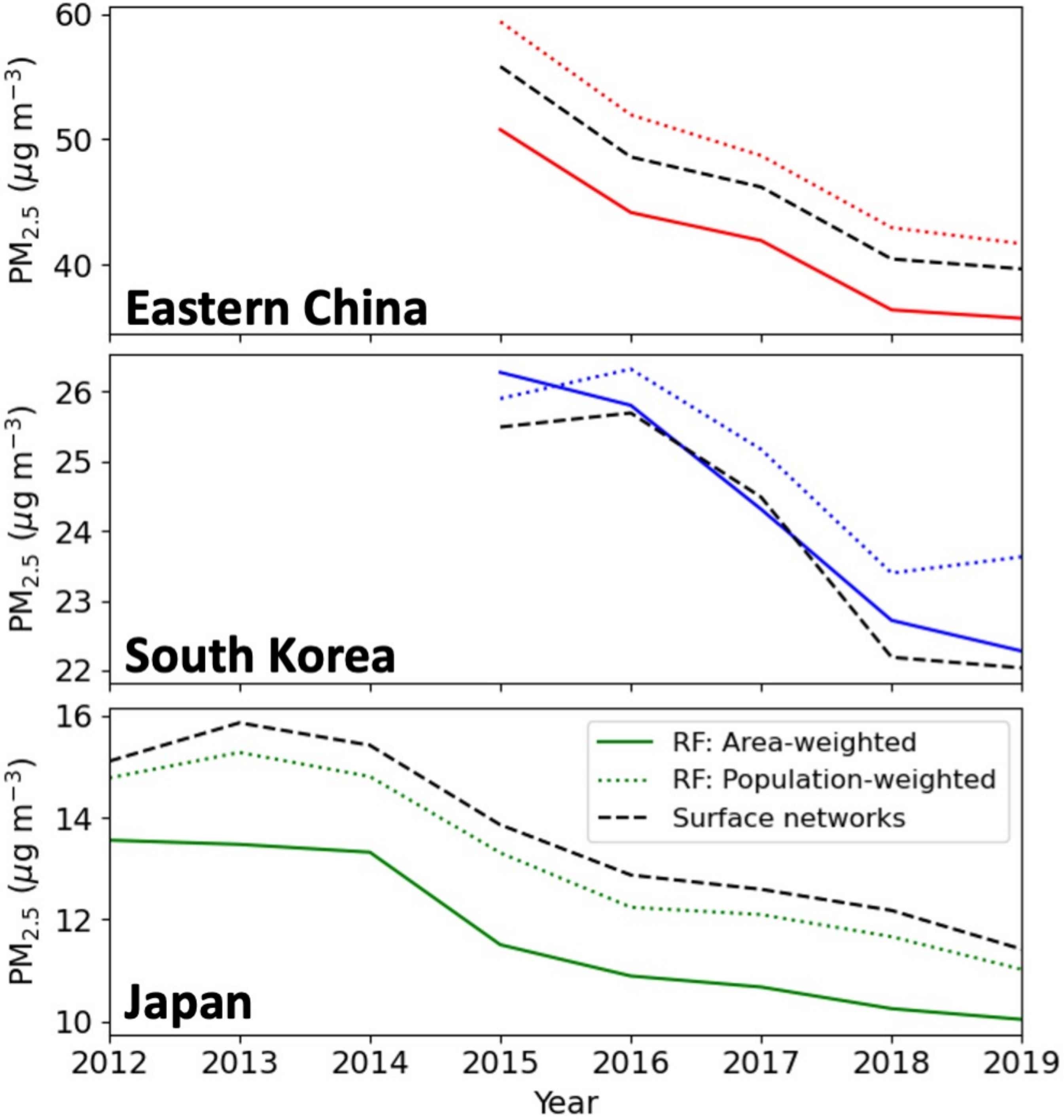




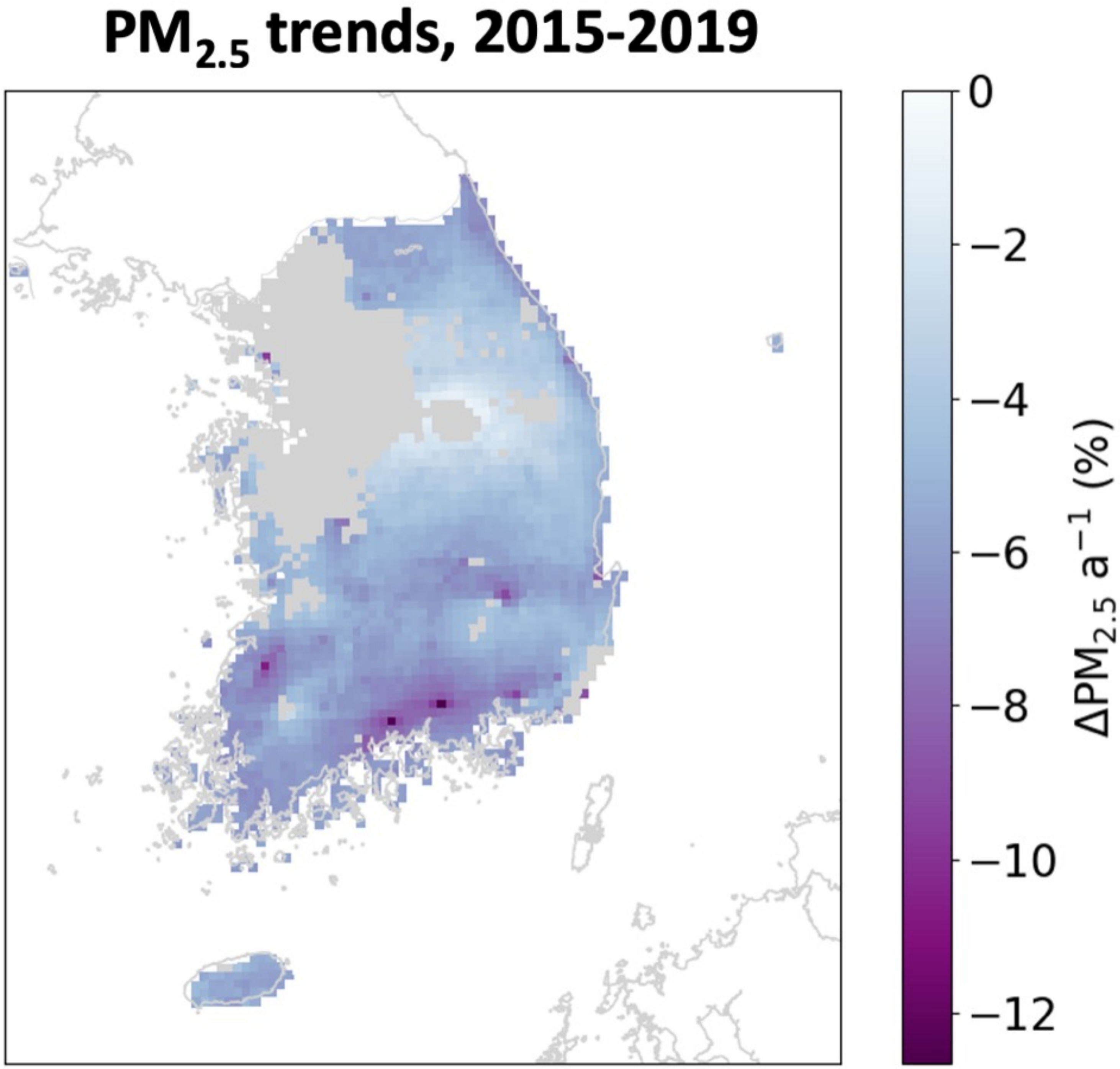
Coverage is much improved and reveals pollution hotspots



Expanding coverage offers new perspective on annual trends



Fine particulate matter decreases throughout South Korea, but no trend in Seoul despite emissions controls

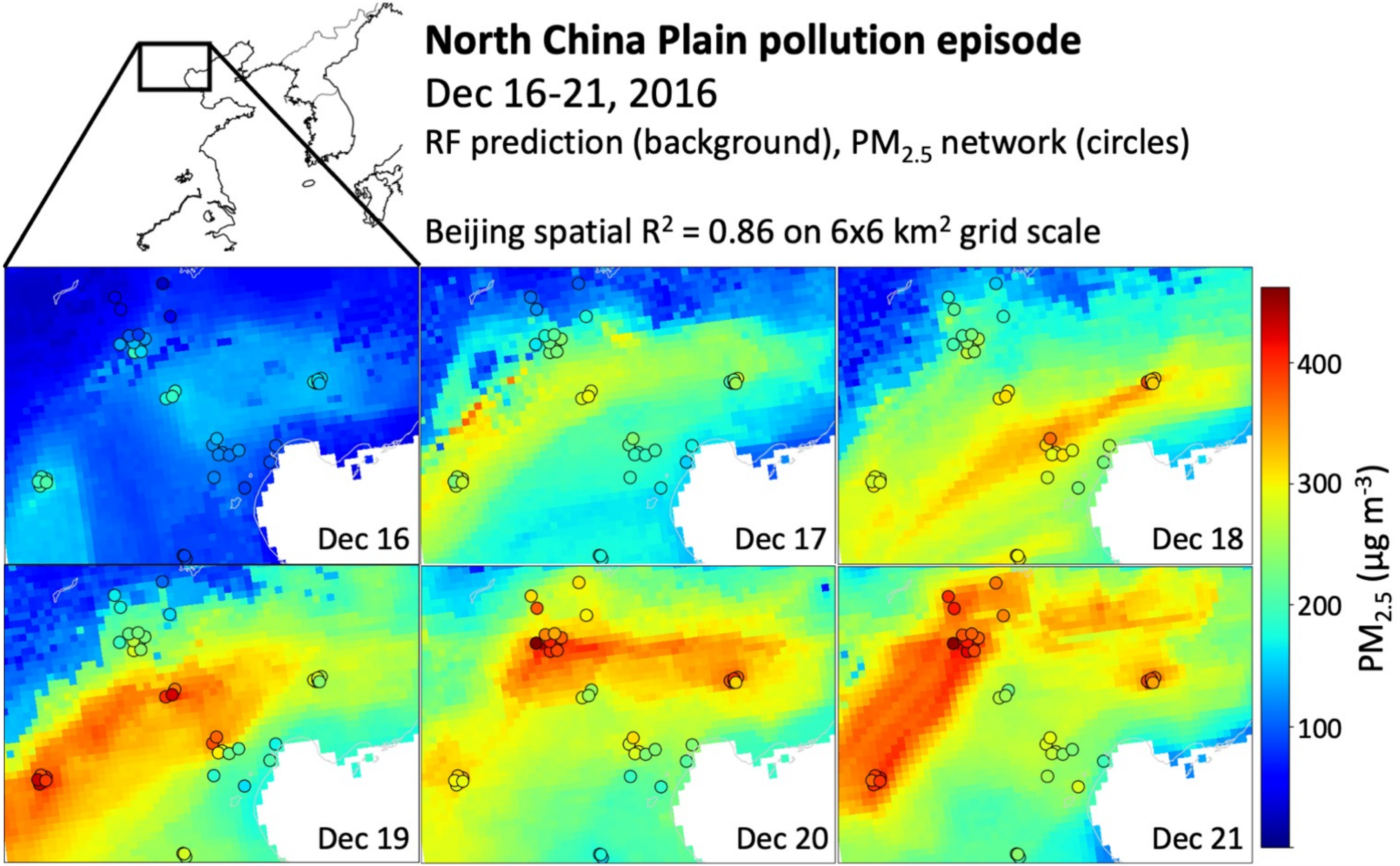


North China Plain pollution episode

Dec 16-21, 2016

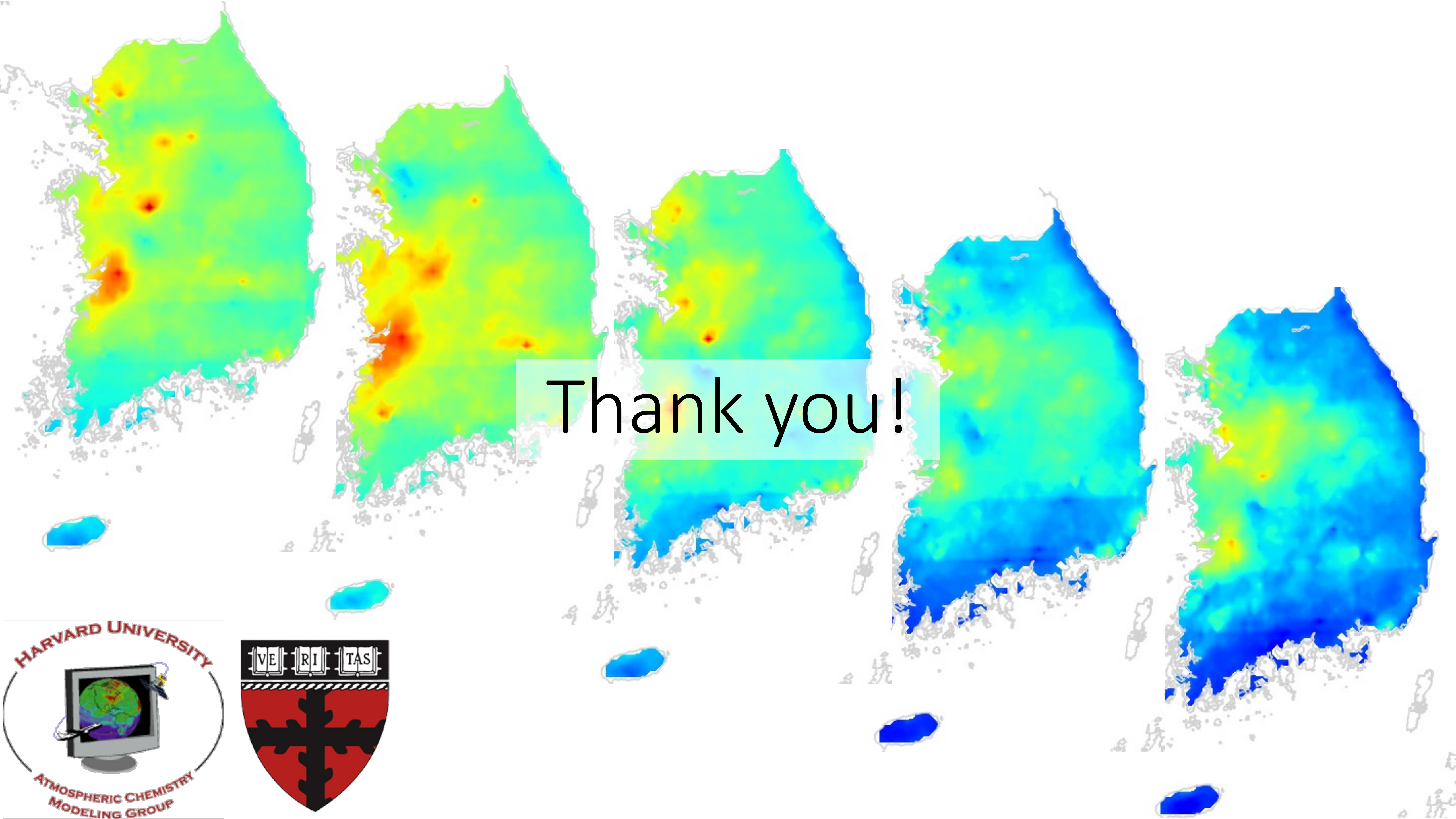
RF prediction (background), $PM_{2.5}$ network (circles)

Beijing spatial $R^2 = 0.86$ on 6x6 km^2 grid scale



Conclusions

- Random forest accurate but has trouble predicting very low and especially very high $PM_{2.5}$ days
  - Further work will be needed to increase resolution and reduce tail bias
- $PM_{2.5}$ concentrations predicted by the RF algorithm for individual countries show steady 2015-2019 declines consistent with surface networks
- Further examination of RF results for South Korea shows general 2015-2019 $PM_{2.5}$ decreases across South Korea except for flat concentrations in Seoul



Thank you!

HARVARD UNIVERSITY
ATMOSPHERIC CHEMISTRY MODELING GROUP

VE RI TAS