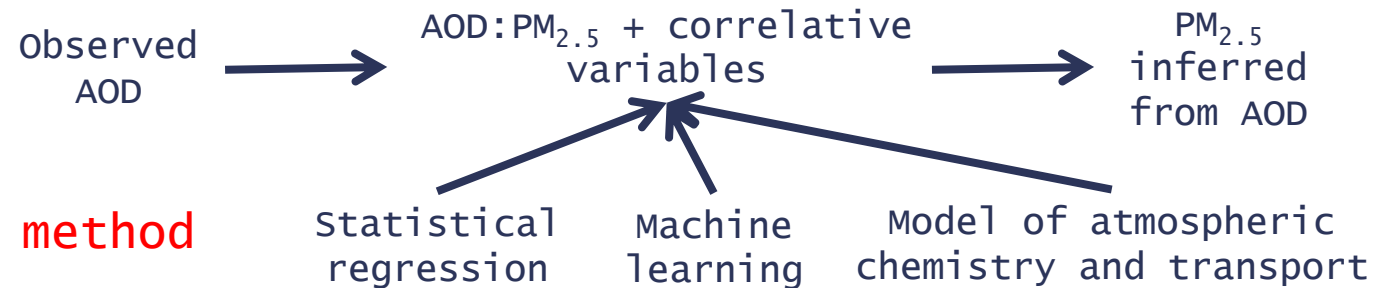# A model framework to reduce bias in ground-level PM$_{2.5}$ concentrations from satellite-retrieved AOD

Fei Yao[1] (fei.yao@ed.ac.uk), Paul I. Palmer[1,2]

[1]School of GeoSciences, University of Edinburgh, UK;
[2]NCEO, University of Edinburgh, UK

Observed AOD $\rightarrow$ AOD:PM$_{2.5}$ + correlative variables $\rightarrow$ PM$_{2.5}$ inferred from AOD

**Original method** (in red)

Statistical regression | Machine learning | Model of atmospheric chemistry and transport

**Revised method** (in green)

**Data clustering**

- GEOS-Chem model AOD for individual chemical components sampled at Chinese PM$_{2.5}$ monitoring locations.

- Clustering algorithm to identify locations where PM$_{2.5}$:AOD varies coherently.

**Data suitability**

- Within identified monthly data clusters calculate AOD$_{PBL}$:AOD$_{TOTAL}$.

- Identify threshold below which data are discarded.
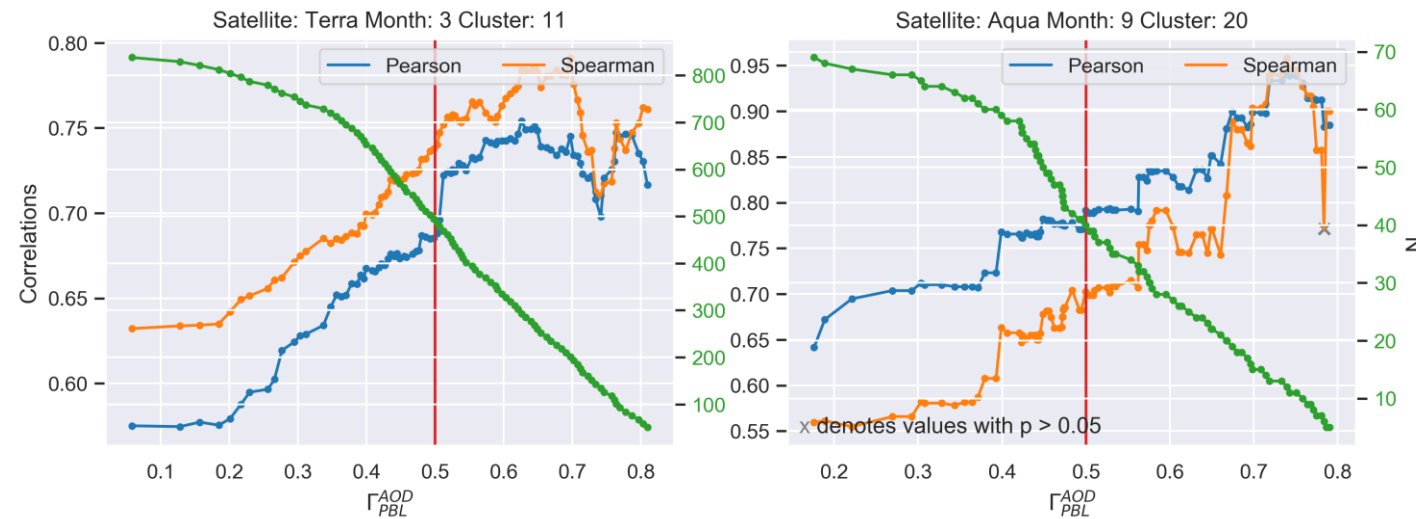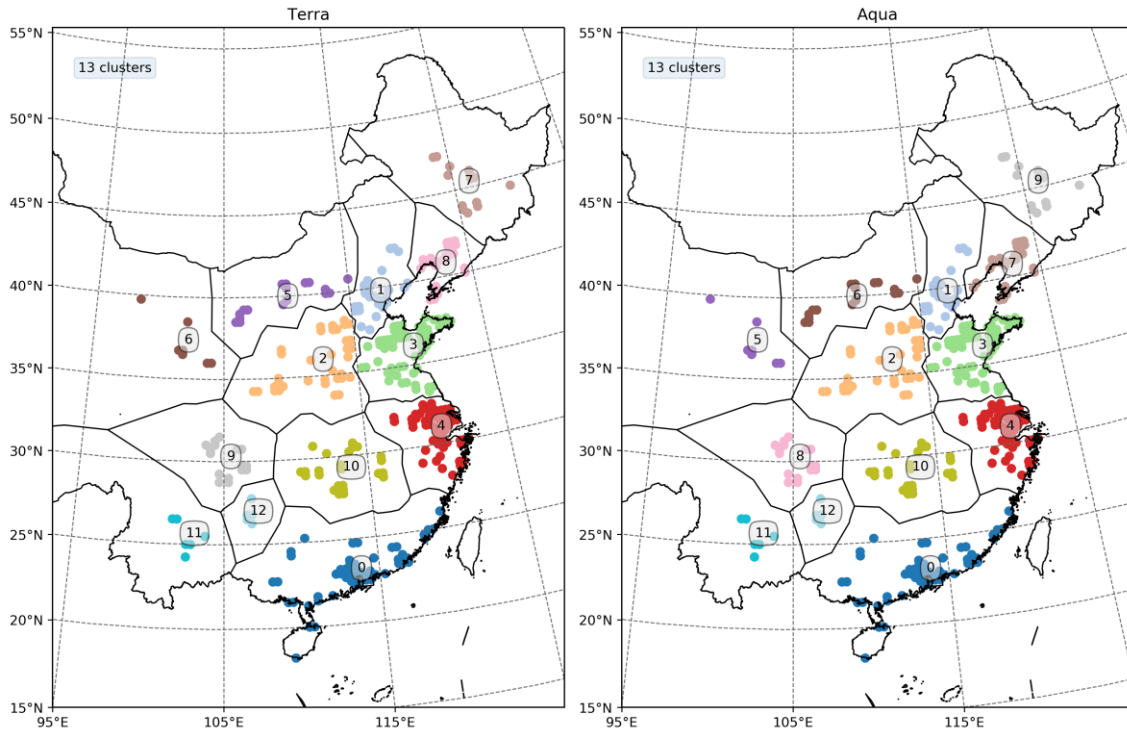
**Data-driven PM$_{2.5}$:AOD model development**

- Fit PM$_{2.5}$:AOD data using (2x) statistical models and (2x) machine learning models.

- Use Monte Carlo method to determine improvement in this approach with traditional approach.

**Mapping PM$_{2.5}$ from AOD**

- Map PM$_{2.5}$ inferred from AOD.

# Results of data clustering and suitability



- We determine a total of 13 spatial clusters with similar extent across China. Among them the majority correspond to urban agglomerations.

- We define $\Gamma_{PBL}^{AOD} = \dfrac{AOD_{PBL}}{AOD_{TOTAL}}$ and determine 0.5 as the threshold, above which we retain the data to develop physically-meaningful $PM_{2.5}$:AOD relationships.

# Results of data-driven model development

Benefiting from the improved representiveness of AOD for ground-level $PM_{2.5}$, the revised method:

1. reduces bias in inferred estimates of ground-level $PM_{2.5}$ by 9-15%;

2. captures more variations in ground-level $PM_{2.5}$ by up to 8%.

Model structure: $PM_{2.5}{}_g^d = f(AOD_g^d + PBLH_g^d + RH_{PBL}{}_g^d + TS_g^d + PRECTOT_g^d + U10M_g^d + V10M_g^d + SLP_g^d + DOY_g^d)$

| Satellite | Model | $N$ | $N'$ | $R^2$ | $R^{2'}$ | $R^2_p$ | $MPE$ | $MPE'$ | $MPE_p$ |
|---|---|---|---|---|---|---|---|---|---|
| Terra | PooledOLS | 57819.0 | 36692.0 | 0.36 | 0.39 | 0.0 | -0.48 | -0.41 | 0.0 |
|  | PanelOLS | 57819.0 | 36692.0 | 0.58 | 0.58 | 0.0 | -0.27 | -0.24 | 0.0 |
|  | RF1 | 57819.0 | 36692.0 | 0.63 | 0.63 | 0.0 | -0.32 | -0.28 | 0.0 |
|  | RF2 | 57819.0 | 36692.0 | 0.68 | 0.66 | 0.0 | -0.29 | -0.26 | 0.0 |
| Aqua | PooledOLS | 55939.0 | 46961.0 | 0.43 | 0.45 | 0.0 | -0.45 | -0.41 | 0.0 |
|  | PanelOLS | 55939.0 | 46961.0 | 0.64 | 0.66 | 0.0 | -0.26 | -0.23 | 0.0 |
|  | RF1 | 55939.0 | 46961.0 | 0.67 | 0.69 | 0.0 | -0.31 | -0.28 | 0.0 |
|  | RF2 | 55939.0 | 46961.0 | 0.73 | 0.73 | 0.0 | -0.28 | -0.25 | 0.0 |

- PBLH: Planetary boundary layer (PBL) height; RH_PBL: mean relative humidity in PBL; TS: surface temperature; PRECTOT: total precipitation; U10M: 10-metre eastward wind; V10M: 10-metre northward wind; SLP: sea level pressure; DOY: day of year (only included in PanelOLS and RF2).
- X and X' denote statistics trained by the full (ignoring the step of data suitability) and suitable data. $X_p$ denotes the possibility of achieving the performance no worse than ours by chance determined from a Monte Carlo simulation.
- RF: Random Forest; MPE: mean percentage error.

# Results of ground-level PM$_{2.5}$ mapping

Accordingly, we improve the seasonal ground-level PM$_{2.5}$ maps, e.g. the bias of the autumn (winter) mean of ground-level PM$_{2.5}$ estimates over Qinghai and Gansu (Shaaxi, Shanxi, and Henan) provinces reduces from -8% to -5% (11% to 6%).