Supplementary information for **Retrieving Ground-Level PM$_{2.5}$ Concentrations in China (2013-2021) with a Numerical Model-Informed Testbed to Mitigate Sample Imbalance-Induced Biases**

Siwei Li[1,3,4,*] , Yu Ding[1], Jia Xing[2], Joshua S. Fu[2]

[1] Hubei Key Laboratory of Quantitative Remote Sensing of Land and Atmosphere, School of Remote Sensing and Information Engineering, Wuhan University, Hubei, 430000, China

[2] Department of Civil and Environmental Engineering, the University of Tennessee, Knoxville, TN 37996, USA

[3] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

[4] Hubei Luojia Laboratory, Wuhan University, Wuhan 430079, China

Correspondence to: Siwei Li (siwei.li@whu.edu.cn); Jia Xing (jxing3@utk.edu)
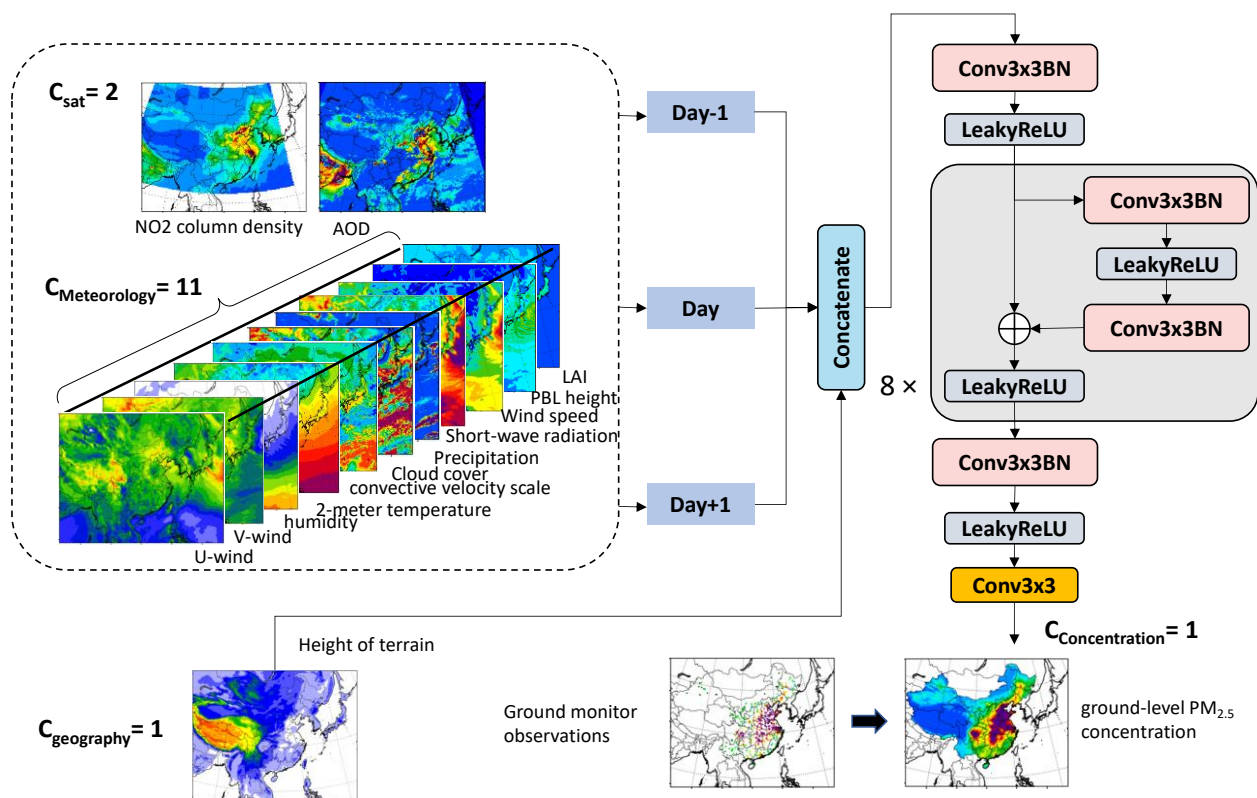
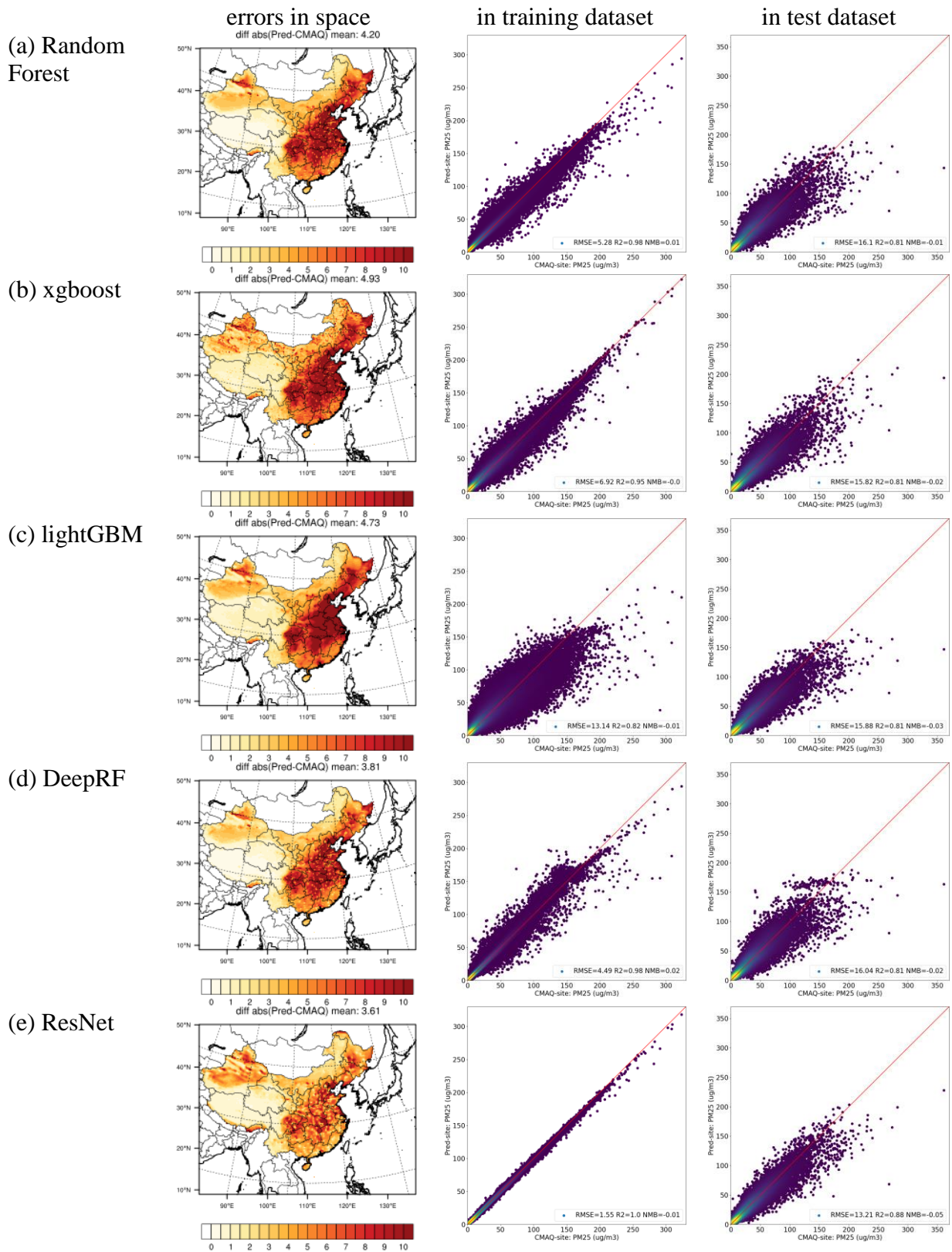**Figure S1.** ResNet model structure

Figure S2. Errors in predicting surface PM$_{2.5}$ with monitor-located grids training models

model	difference in error	Difference in performance on the sites

(a) Random Forest

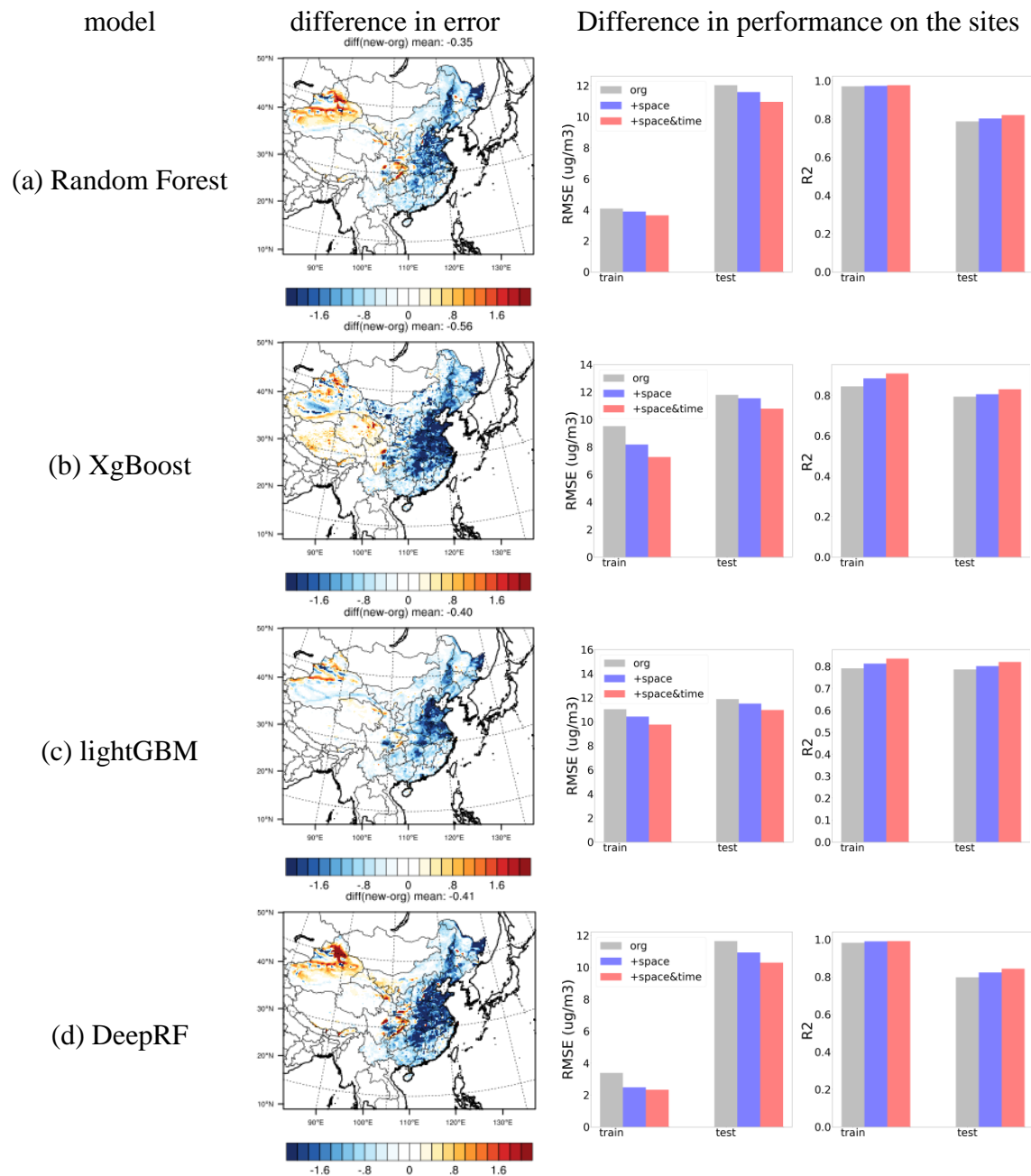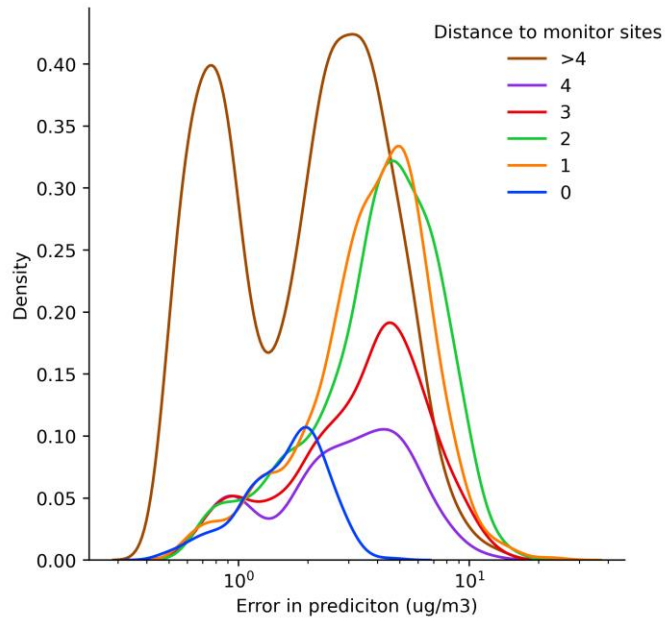(b) XgBoost
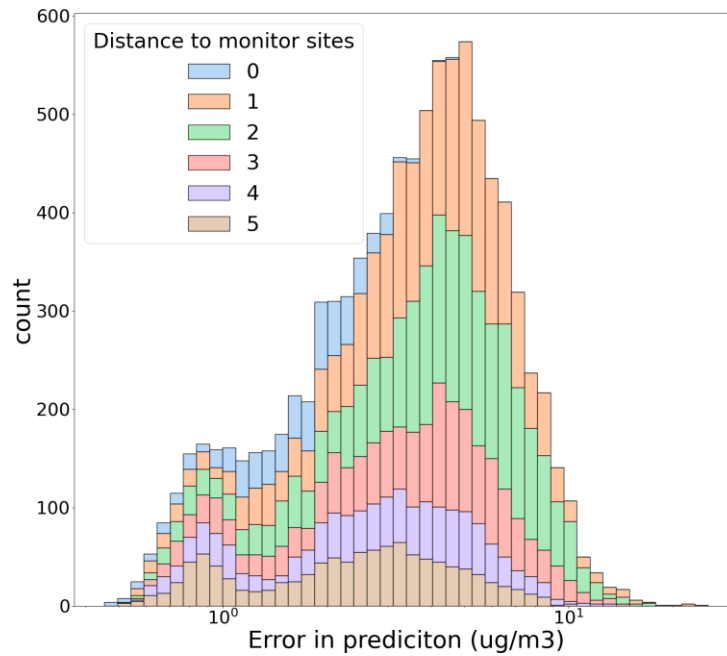
(c) lightGBM

(d) DeepRF

Figure S3. Improvement after implementing the features in surrounding grid cells (compared to each baseline model without spatiotemporal-neighbourhood features)

(a) by density

(b) by count

Figure S4. Error distribution across the distance to monitor sites (D-site) based on ResNet-time model
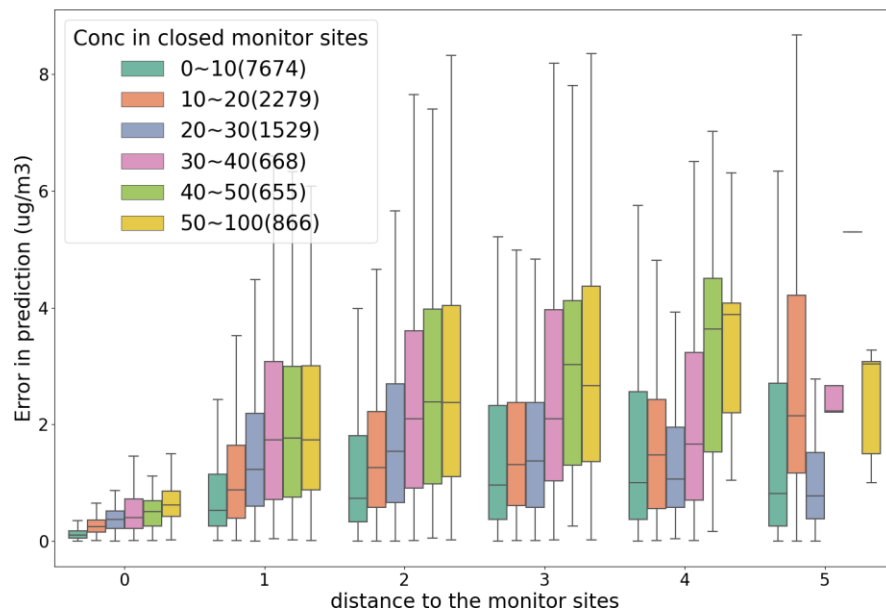
Figure S5. Error distribution across the monitor concentrations (B-conc) based on ResNet-time model

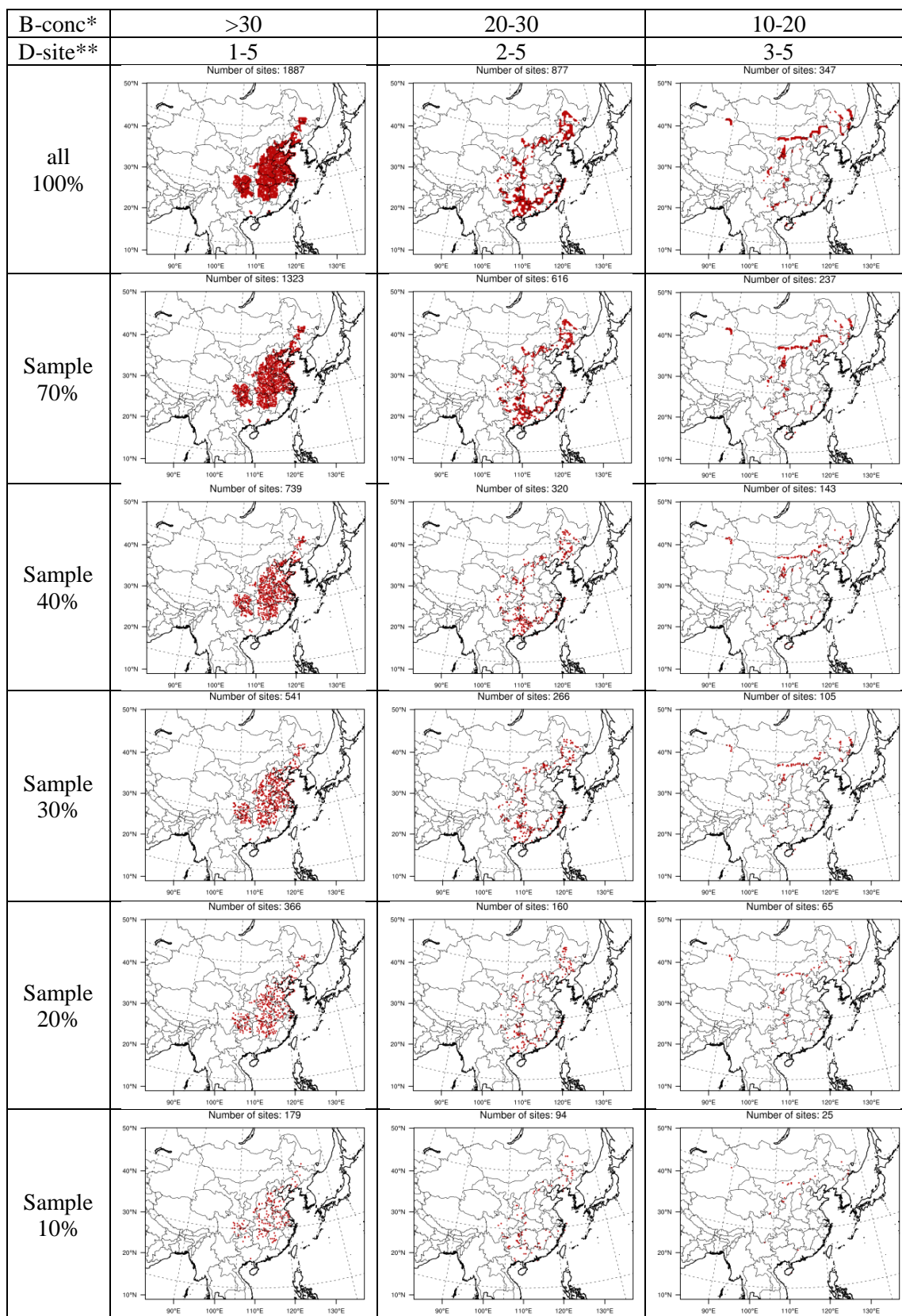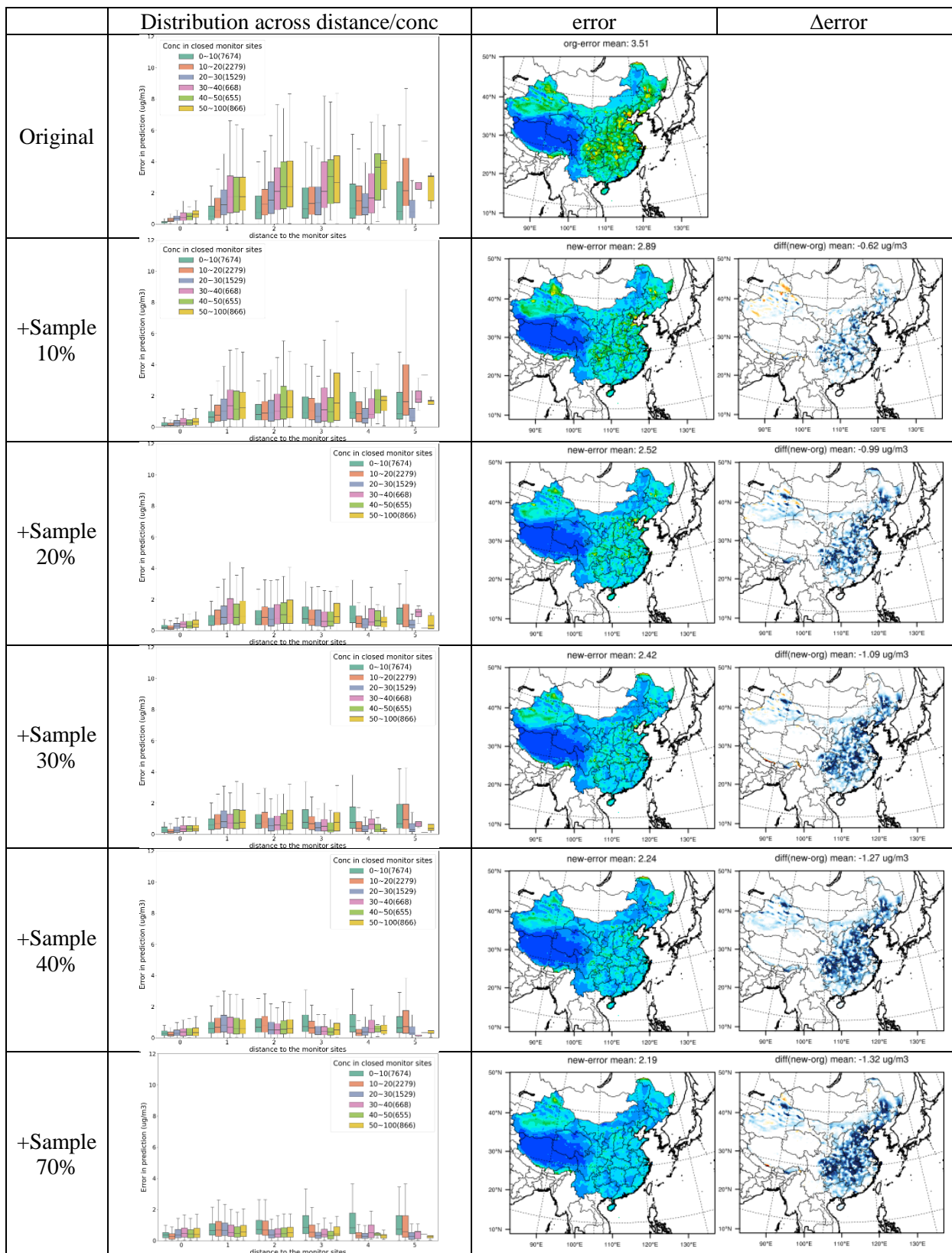| B-conc*	>30	20-30	10-20
D-site**	1-5	2-5	3-5
all 100%	Number of sites: 1887	Number of sites: 877	Number of sites: 347
Sample 70%	Number of sites: 1323	Number of sites: 616	Number of sites: 237
Sample 40%	Number of sites: 739	Number of sites: 320	Number of sites: 143
Sample 30%	Number of sites: 541	Number of sites: 266	Number of sites: 105
Sample 20%	Number of sites: 366	Number of sites: 160	Number of sites: 65
Sample 10%	Number of sites: 179	Number of sites: 94	Number of sites: 25

Figure S6. Spatial distribution of selected adding sites with certain levels of sampling (B-conc:
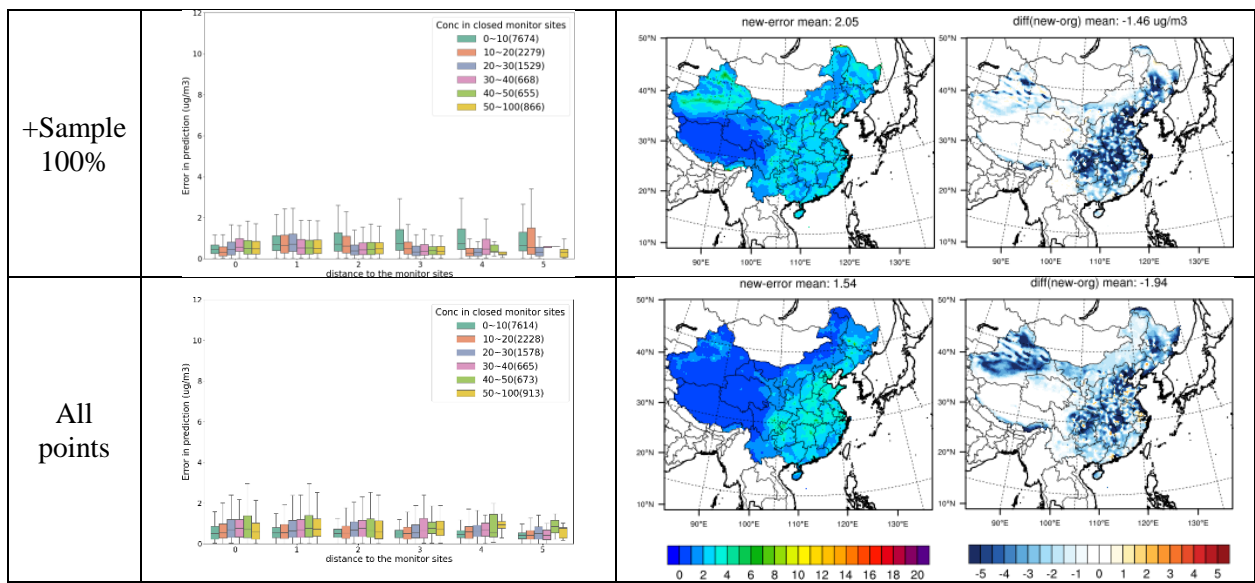conc in closed monitor sites; D-site: distance from monitor sites)

	Distribution across distance/conc	error	Δerror
Original		org-error mean: 3.51	
+Sample 10%		new-error mean: 2.89	diff(new-org) mean: -0.62 ug/m3
+Sample 20%		new-error mean: 2.52	diff(new-org) mean: -0.99 ug/m3
+Sample 30%		new-error mean: 2.42	diff(new-org) mean: -1.09 ug/m3
+Sample 40%		new-error mean: 2.24	diff(new-org) mean: -1.27 ug/m3
+Sample 70%		new-error mean: 2.19	diff(new-org) mean: -1.32 ug/m3

Figure S7. Improvement with selected adding sites with certain levels of sampling

	(a) predicted surface PM$_{2.5}$ concentration			(b) performance on the sites	
	new	org	Δ (new - org)	RMSE	R^2
Random Forest					
XgBoost					
lightGBM					
deepRF					

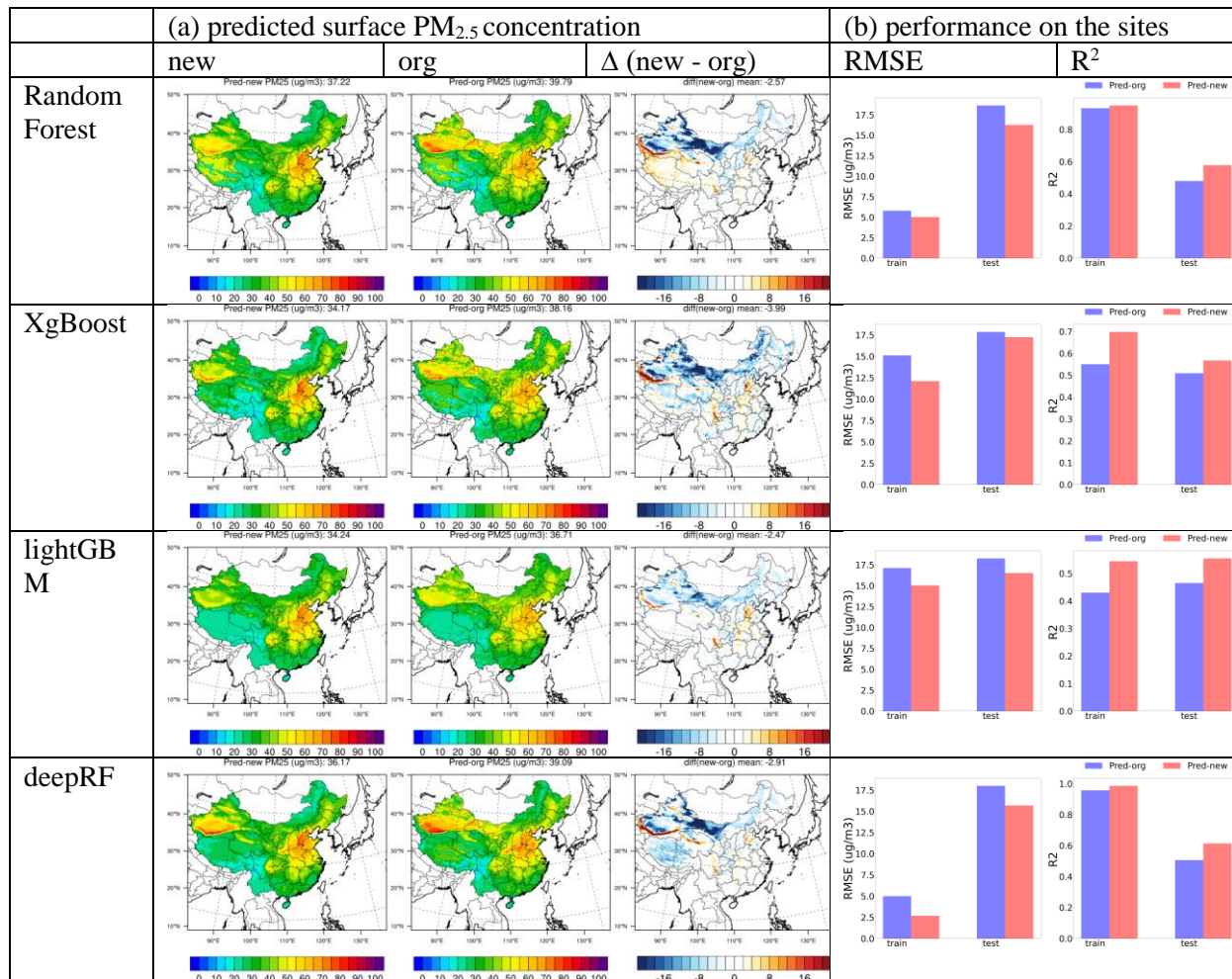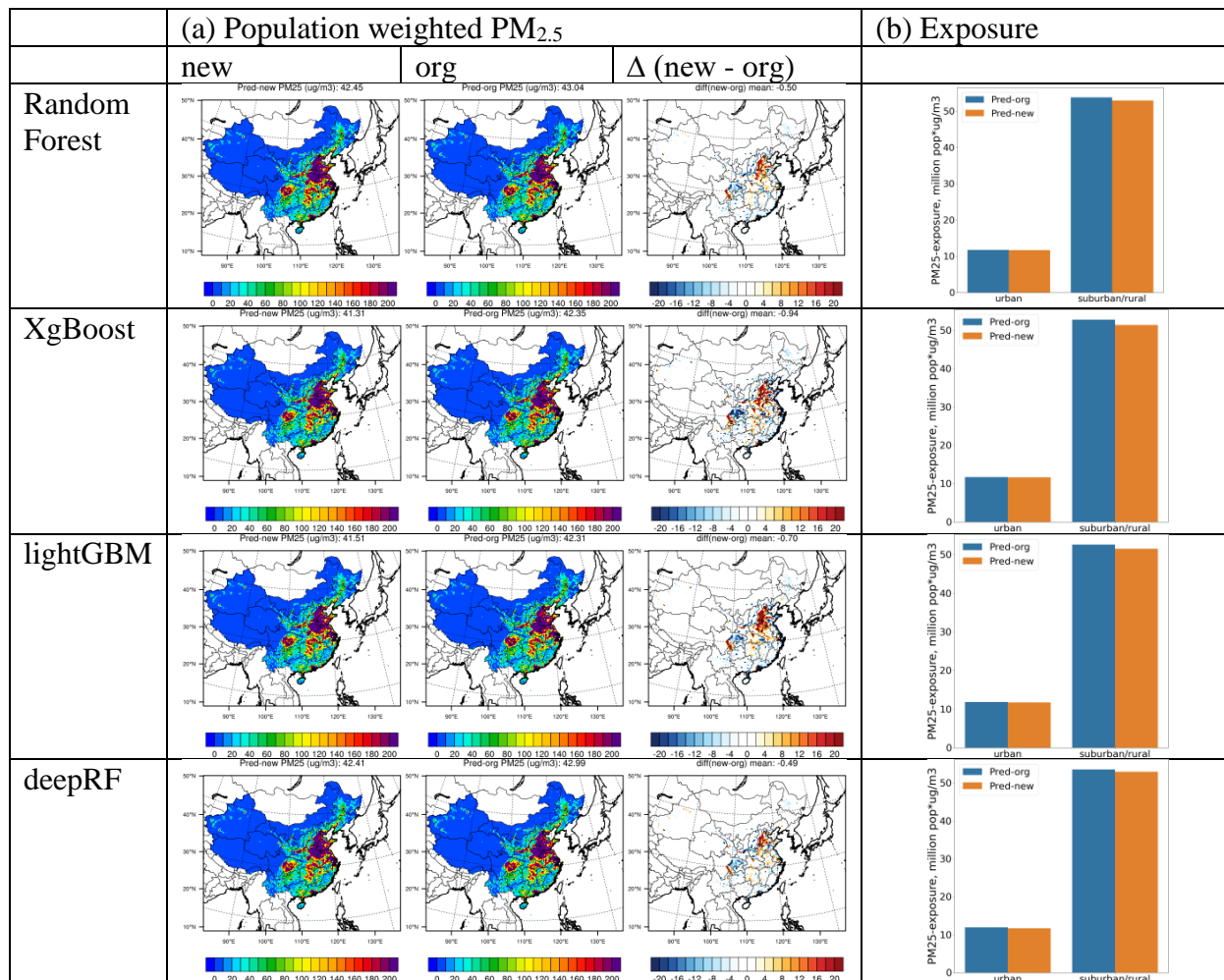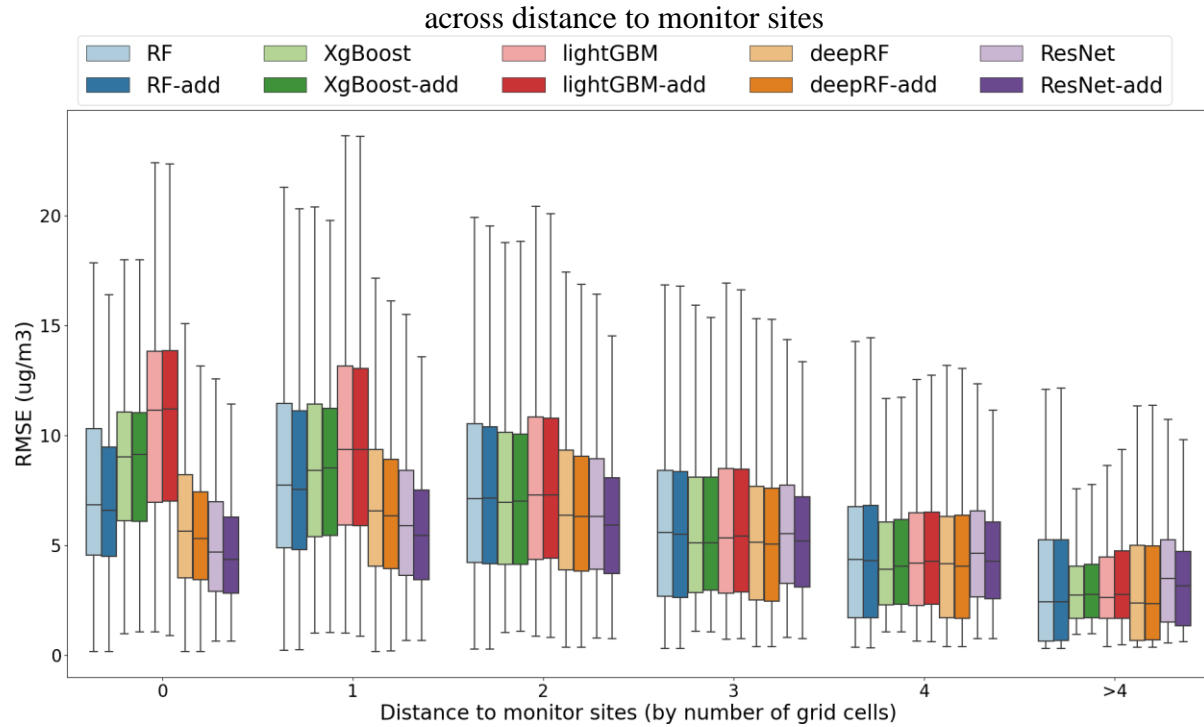**Figure S8.** Improved performance with inclusion of spatiotemporal-neighbourhood features trained with real measurement dataset

	(a) Population weighted PM$_{2.5}$			(b) Exposure
	new	org	Δ (new - org)	
Random Forest				
XgBoost				
lightGBM				
deepRF				

*Note: within 5 grid cells

Figure S9. Uncertainties in estimation of PM2.5-related exposure across China

(a) performance in scenarios with adding points
across distance to monitor sites

(b) +sample new sites during 2017-2021

Location of sites

*black dot: original monitor sites (619)
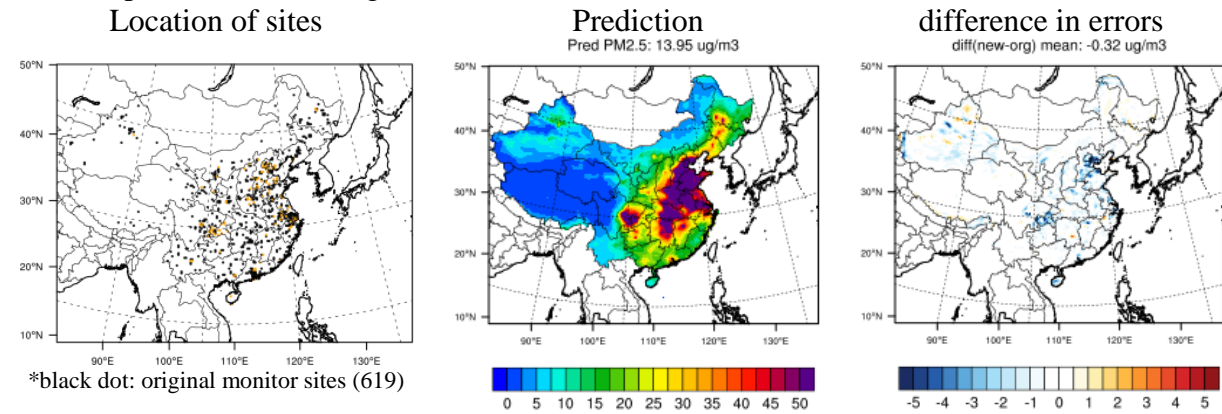
Prediction

difference in errors

Figure S10. Improvement with the inclusion of new sites after 2017 in testing with CMAQ simulations