remote sensing

Supplementary Material

For

Emission-based machine learning approach for Large-scale estimates of black carbon in China

Ying, Li [a,b,c], Sijin Liu [a], Reza Bashiri Khuzestani [a], Kai Huang [a], Fangwen Bao [a]*

[a] Department of Ocean Sciences and Engineering, Southern University of Science and Technology, Shenzhen, China.

[b] Center for the Oceanic and Atmospheric Science at SUSTech (COAST), Southern University of Science and Technology, Shenzhen, China.

[c] Guangdong-Hong Kong-Macao Joint Laboratory for Data-Driven Fluid Mechanics and Engineering Applications, Southern University of Science and Technology, Shenzhen, China.

* Correspondence to: **Ying LI** (Email: ***liy66@sustech.edu.cn***)

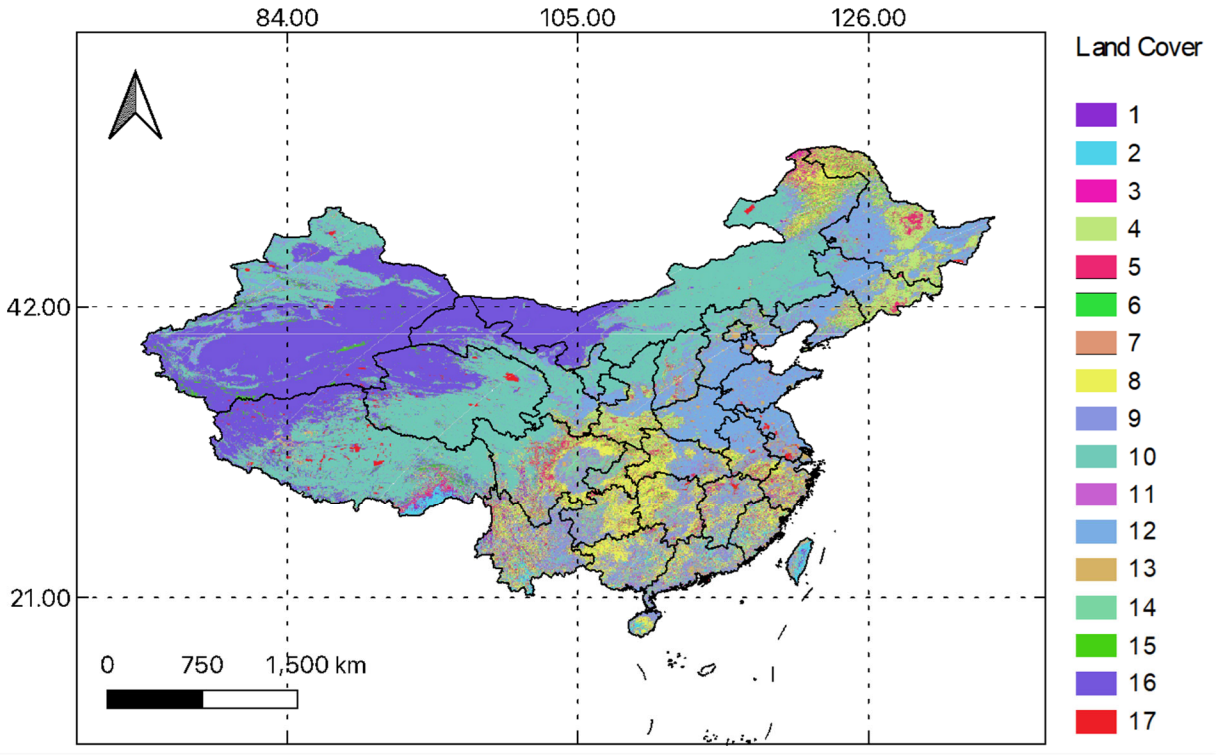

Figure S1. IGBD classification of the land cover type based on MODIS satellite database.

Table S1. Detailed information, including the name and the classification of the land cover types obtained in Figure S2 based on the MODIS satellite database.

Name	Type	Description
Evergreen Needleleaf Forests	1	Dominated by evergreen conifer trees (canopy >2m). Tree cover >60%.
Evergreen Broadleaf Forests	2	Dominated by evergreen broadleaf and palmate trees (canopy >2m). Tree cover >60%.
Deciduous Needleleaf Forests	3	Dominated by deciduous needle leaf (larch) trees (canopy >2m). Tree cover >60%.
Deciduous Broadleaf Forests	4	Dominated by deciduous broadleaf trees (canopy >2m). Tree cover >60%.
Mixed Forests	5	Dominated by neither deciduous nor evergreen (40-60% of each) tree type (canopy >2m). Tree cover >60%.
Closed Shrublands	6	Dominated by woody perennials (1-2m height) >60% cover.
Open Shrublands	7	Dominated by woody perennials (1-2m height) 10-60% cover.
Woody Savannas	8	Tree cover 30-60% (canopy >2m).
Savannas	9	Tree cover 10-30% (canopy >2m).
Grasslands	10	Dominated by herbaceous annuals (<2m).
Permanent Wetlands	11	Permanently inundated lands with 30-60% water cover and >10% vegetated cover.
Croplands	12	At least 60% of area is cultivated cropland.
Urban and Built-up Lands	13	At least 30% impervious surface area including building materials, asphalt, and vehicles.
Cropland/Natural Vegetation Mosaics	14	Mosaics of small-scale cultivation 40-60% with natural tree, shrub, or herbaceous vegetation.
Permanent Snow and Ice	15	At least 60% of area is covered by snow and ice for at least 10 months of the year.
Barren	16	At least 60% of area is non-vegetated barren (sand, rock, soil) areas with less than 10% vegetation.
Water Bodies	17	At least 60% of area is covered by permanent water bodies.

$$E_1 = e^{53.67957-6743.769/T2m-4.8451\times ln\,T2m} \quad \text{(S1)}$$

$$e_1 = e^{53.67957-6743.769/D2m-4.8451\times ln\,D2m} \quad \text{(S2)}$$

$$RH = {}^{e_1}/_{E_1} \quad \text{(S3)}$$

$$WS = \sqrt{U10^2 + V10^2} \quad \text{(S4)}$$

$$WD = tan^{-1}(V10/U10) \times 180/\pi \quad (\text{U10} > 0\ \&\ \text{V10} > 0) \quad \text{(S5)}$$

$$WD = tan^{-1}(V10/U10) \times 180/\pi + 180 \quad (\text{U10} < 0\ \&\ \text{V10} > 0) \quad \text{(S6)}$$

$$WD = tan^{-1}(V10/U10) \times 180/\pi + 360 \quad (\text{U10} > 0\ \&\ \text{V10} < 0) \quad \text{(S7)}$$

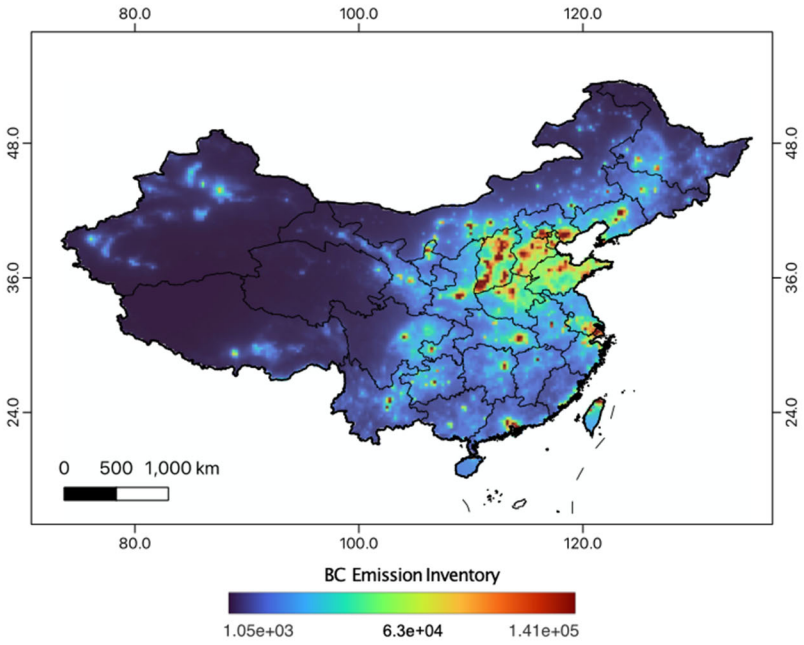

Figure S2. Spatial distribution of the annual averages BC emission inventories developed by Peking University over China .

Section S2: Haversine equation

For some variables with a rough spatial resolution (0.25° * 0.25°), meteorological, radiation, albedo, and other data are selected as the closest value to the site. The distance between the BC stations and these variable's grid points is calculated by the Haversine formula based on the latitude and longitude of the two points. All predictors in the spatial distribution data set used to estimate the concentration of BC over the Chinese mainland will be resampled to the same spatial (0.25°×0.25°) and temporal resolution (daily average) because of the sparse distribution of black carbon monitoring sites, the spatial resolution of black carbon concentration estimated by the model in mainland China is 0.25°. Variables with high spatial resolution, such as black carbon emission source data, AOD, and land use data, will be averaged according to the resolution grid.

$$hav(\theta) = ((1 - cos\ (\theta))) \ / \ 2 \tag{S8}$$

$$\theta = d/r \tag{S9}$$

$$hav(\theta) = hav(lat1 - lat2) + cos\ (lat1)cos\ (lat2)hav(lon1 - lon2) \tag{S10}$$

Eq.S10 is the haversine formula. Assuming that θ = d/r, d is the distance between two stations, and r is the radius of the earth.

Table S2. Final selected input variables included in the Random Forest Model, including their classifications and data information.

Dataset	Variable	Full name	Spatial resolution	Temporal resolution	Source
BC monitoring	BC ground-base Monitoring	Ground-based BC	In-situ 42 sites	5-minute	CMA
Meteorology (C1)	pt	Potential Temperature	0.25°0.25°	1h	ERA-Interim
	RH	Relative Humidity			
	e	Evaporation			
	blh	Boundary Layer Height			
	bld	Boundary Layer Dissipation			
	fsr	Forecast Surface Roughness			
	sp	Surface Pressure			
	ws	Wind Speed			
	wd	Wind Direction			
	mbld	Mean Boundary Layer Dissipation			
Albedo (C2)	Fal	Forecast Albedo	0.25°0.25°	1h	ERA-Interim
	aluvp	UV visible albedo for direct radiation			
	aluvd	UV visible albedo for direct radiation			
Land Use Cover (C3)	LUC (Type 1-17)	Land use cover	500m500m	annual	MCD12Q1
LAI (C4)	Lai-lv	Leaf Area Index-low vegetation	0.25°0.25°	1h	ERA-Interim
	Lai-hv	Leaf Area Index-high vegetation			

Gridded BC emissions (C5)	BC emission	Black carbon emission	0.1°0.1°	Monthly	PKU
	Energy Data	Emission Inventories from Energy Sectors			
	Industrial data	Emission Inventories from Industrial Sectors			
	Residential data	Emission Inventories from Residential Sectors			
	Vehicle data	Emission Inventories from Vehicle Sectors			
	Wildfire data	Emission Inventories from Wildfire Sectors			
	Other data	Emission Inventories from Other Sectors			
Extra	Day of year	Day of year	In-situ 42 sites	5-minute	CMA
	Station lat	Latitude of the BC monitoring			
	Station lon	Longitude of the BC monitoring			
	AOD	Aerosol Optical depth	1km x 1km	1day	MODIS

Table S3. Hyperparameter tuning of the models employed in this research.

Models	**Optimal Hyperparameters**
Boosted Trees (BT)	mstop = 150, maxdepth = 3 and nu = 0.1
Stochastic Gradient Boosting (SGB)	n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10
Extreme Gradient Boosting (XGBoost)	nrounds = 150, lambda = 0.1, alpha = 0.1, eta = 0.3, and gamma = 0
Random Forest (RF)	ntree = 200, and mtry = 12

Note: **nrounds** refers to the maximum number of iterations allowed to prevent excessively deep iterations. **eta** represents the step size shrinkage used in the update to avoid overfitting. **Gamma** is the minimum loss reduction required to partition a leaf node of the tree further. **Lambda** is the L2 regularization, and alpha is the L1 regularization. **mtry** is the number of drawn candidate variables in each split. **ntree** is the number of trees in the RF model.

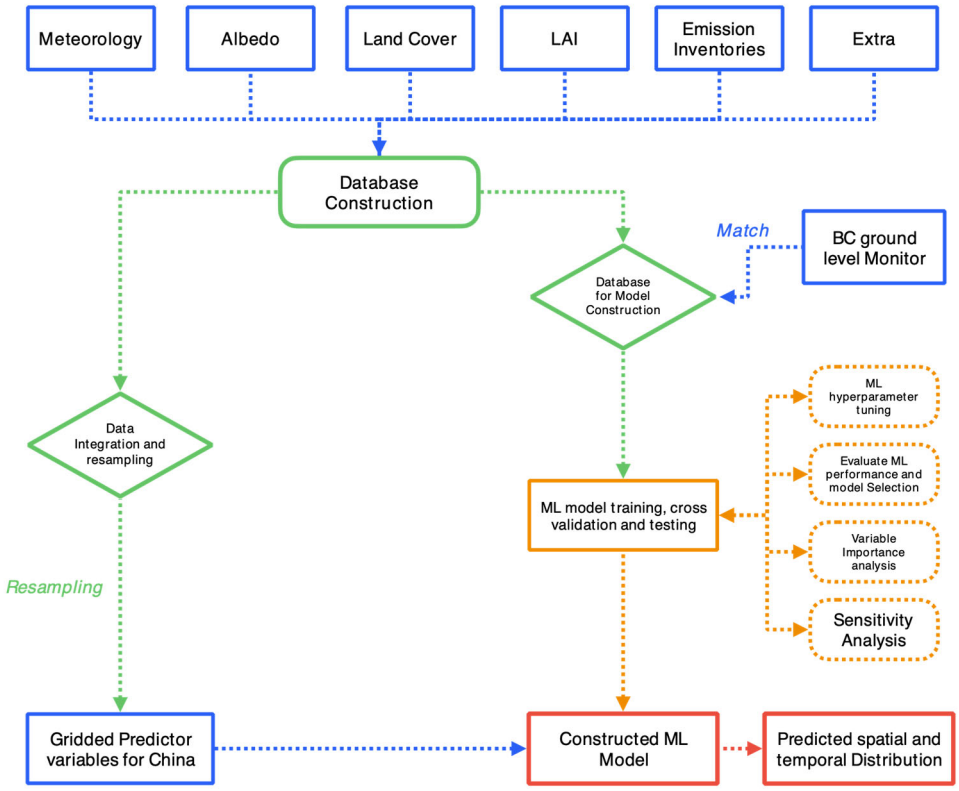

Figure S3. Flow chart of the construction of the machine learning (ML) algorithm construction predictive model.

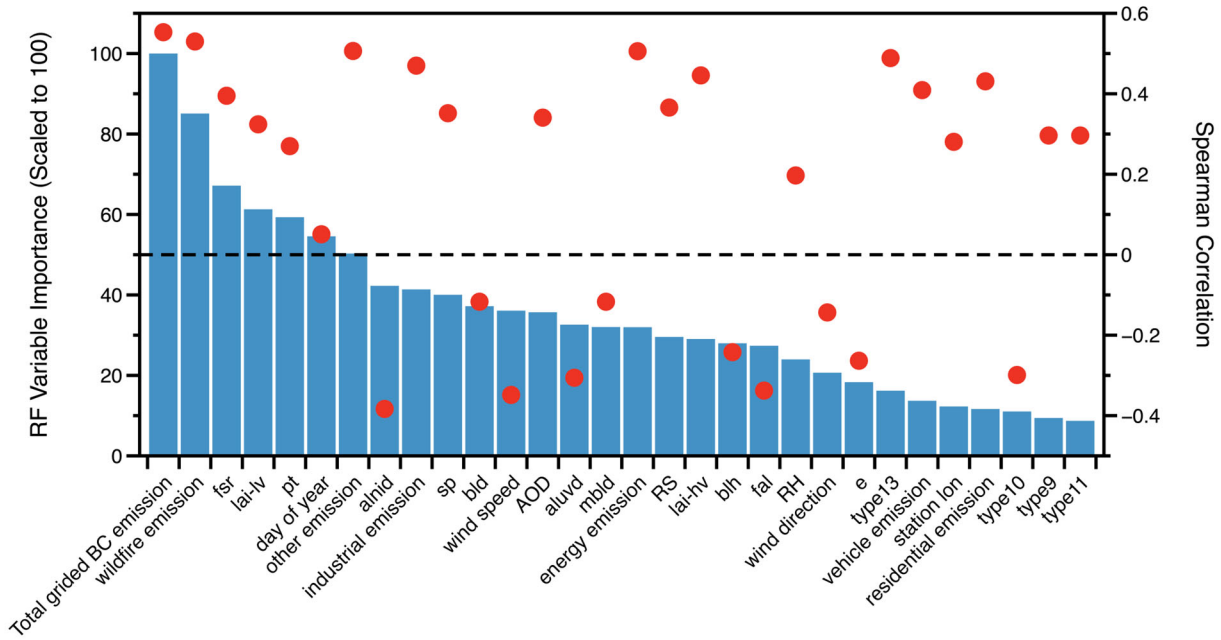

Figure S4. Variable importance measures derived from the RF algorithms. The importance values were scaled to 100 for better illustration.

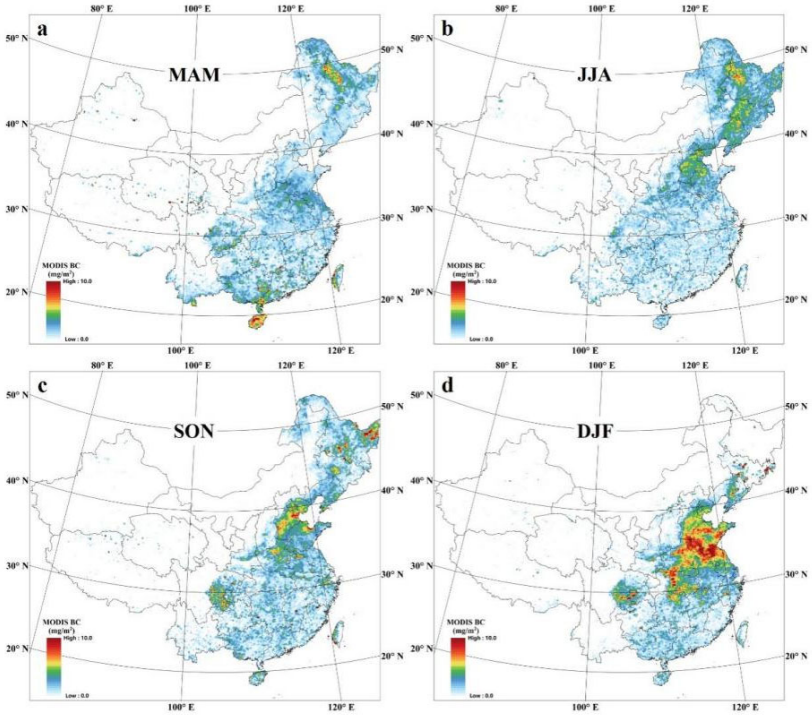

Figure S5. Spatial distribution patterns of seasonal-averages BC concentrations estimated by our model across China (retrieved from Bao et al., 2020).

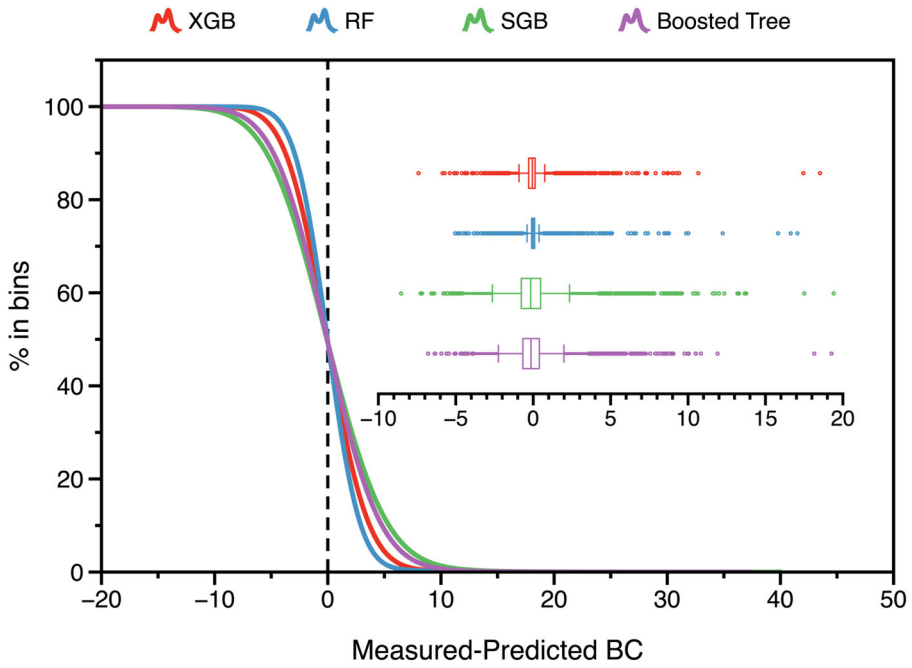

Figure S6. Distributions of the residuals of BC predictive models in different quantiles. Boxplots represent the scatter residual distributions of different predictive models.

Table S4. Permutation-based variable importance for all the input features included in our ML algorithms. The numbers represent RMSE loss after permutations, as illustrated in Figure 6.

Variable List	RF	XGB	SGB	BT
alnid	1.05	1.42	1.76	1.39
aluvd	0.95	0.85	1.78	1.45
AOD	0.98	1.00	2.01	1.64
Total grided BC emission	1.44	1.12	2.06	1.73
bld	0.97	0.97	1.81	1.57
blh	0.90	0.94	1.82	1.48
Day Of Year	1.07	1.08	1.92	1.55
e	0.82	0.72	1.74	1.42
energy emission	0.96	0.26	1.78	1.40
fal	0.97	0.45	1.74	1.39
fsr	1.15	1.34	2.35	1.92
industrial emission	1.07	1.12	1.83	1.45
lai_hv	0.87	0.39	1.77	1.42
lai_lv	1.28	1.80	1.90	1.91
mbld	0.95	0.14	1.73	1.38
other emission	1.17	0.88	1.78	1.41
pt	1.27	1.08	1.99	1.75
residential emission	0.79	0.17	1.75	1.41
RH	0.85	0.73	1.79	1.48
Remote sensing BC	0.89	0.55	1.79	1.47
sp	1.03	1.15	1.76	1.50
station_lat	0.76	0.83	1.73	1.39
station_lon	0.81	0.34	1.73	1.39
type1	0.71	0.15	1.74	1.39
type10	0.79	0.29	1.74	1.40
type11	0.74	0.26	1.75	1.41
type12	0.78	0.35	1.77	1.40
type13	0.85	0.88	1.84	1.46
type14	0.73	0.28	1.78	1.39
type16	0.72	0.14	1.74	1.40
type17	0.74	0.15	1.73	1.39
type2	0.72	0.15	1.73	1.38
type4	0.74	0.23	1.73	1.38
type5	0.75	0.16	1.73	1.38
type6	0.71	0.14	1.73	1.38
type8	0.72	0.14	1.73	1.38
type9	0.75	0.15	1.75	1.40
vehicle emission	0.82	0.33	1.83	1.48
wildfire emissopn	1.54	2.82	1.95	2.17
wind_direction	0.85	0.64	1.78	1.47
wind_speed	0.96	0.81	1.90	1.52

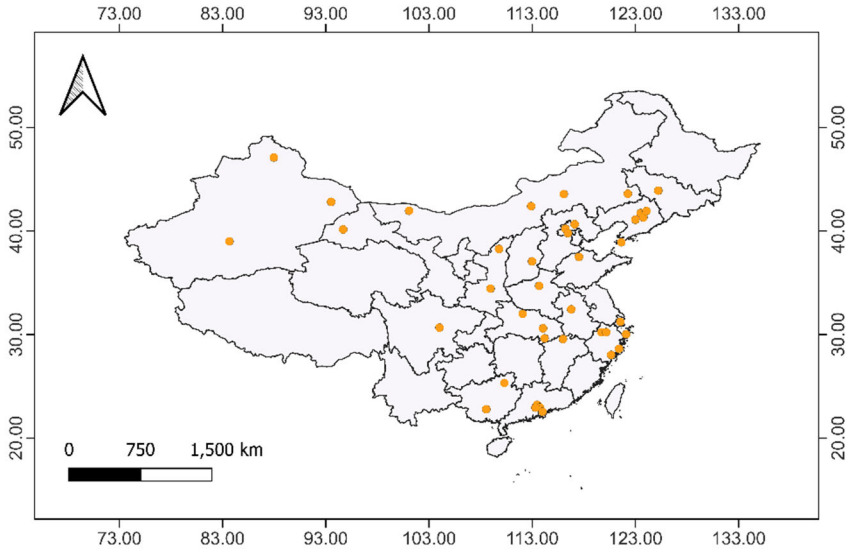

Figure S7. Schematic map of the spatial distribution of BC monitoring sites over China.