Supporting information

LGHAP v2: A global gap-free aerosol optical depth and $PM_{2.5}$ concentration dataset since 2000 derived via big earth data analytics

Kaixu Bai[1,2], Ke Li[1], Liuqing Shao[1], Xinran Li[1], Chaoshun Liu[1], Zhengqiang Li[3], Mingliang Ma[4], Di Han[1], Yibing Sun[1], Zhe Zheng[1], Ruijie Li[1], Ni-Bin Chang[5], Jianping Guo[6]

[1]Key Laboratory of Geographic Information Science (Ministry of Education), School of Geographic Sciences, East China Normal University, Shanghai 200241, China

[2]Institute of Eco-Chongming, 20 Cuiniao Rd., Chongming, Shanghai 202162, China

[3]State Environmental Protection Key Laboratory of Satellite Remote Sensing, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China

[4]School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan 250101, China

[5]Department of Civil, Environmental, and Construction Engineering, University of Central Florida, Orlando, FL, USA

[6]State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China

Correspondence to: Kaixu Bai (kxbai@geo.ecnu.edu.cn) and Jianping Guo (jpguocams@gmail.com)

Text S1. Multisource data homogenization

Given excessive missing values in satellite-based AOD retrievals, it is promising to improve the gap-filling accuracy by increasing data abundance via an integration of external observations. Benefiting from the powerful approximation capacity of machine learning algorithms (i.e., random forest in our study), a set of machine-learned regression models were established to generate MODIS-like AOD estimates from diverse data sources, aiming at providing critical prior information to facilitate AOD gap-filling, especially over regions with massive data voids. AOD_{Terra} observations were hereby deemed the response variable while AOD data from other satellites, MERRA-2 AOD simulations, even in-situ air quality measurements, were used respectively as the critical predictor other than meteorological and geographic factors for AOD prediction. The data homogenization models can be expressed as follows.

$$AOD_{Terra} \sim RF(AQ, MET, AER, LULC, DEM, NDVI, POP, month) \quad (1)$$

where AQ refers to AOD data other than AOD_{Terra} and in situ air quality measurements such as atmospheric visibility and concentrations of major air pollutants that are indicative of regional air quality. MET, AER, $LULC$, DEM, $NDVI$, POP, and $month$ refer to meteorological variables, numerical aerosol simulations, land use and land cover, elevation, vegetation cover, population, and month identifier respectively.

By taking advantage of these data-specific machine learning models, gridded AOD products from other satellites and numeric models were harmonized to resemble AOD_{Terra} by correcting for both the scaling effect (varied spatial resolution) and cross-sensor biases. More importantly, virtual AOD observations were derived from in situ air quality measurements, providing additional AOD prior information to facilitate AOD gap-filling, especially over regions without satellite-based AOD retrievals. This homogenization approach greatly favors the assimilation of multisensory AODs and heterogenous air quality data (Bai et al., 2022a; Li et al., 2022a).

Text S2. Scene-aware ensemble learning graph attention network (SeGAT) for global $PM_{2.5}$ mapping

To accommodate global big earth observation data and to account for spatial representativeness issue of model extrapolation, we developed a novel scene-aware ensemble learning graph attention network (SeGAT) model to fulfill global $PM_{2.5}$ concentration mapping. The workflow of this method was illustrated in Figure S4. Differing from previous data-driven models which were established using either all available data (global model) or regional observations (regional model), the SeGAT model was dedicated to solving the scaling problem in large scale modeling practices (e.g., global $PM_{2.5}$ modeling in this study), avoiding to answer the open questions at what scale the data-driven model should be established and/or how to determine the boundary size (city, province, national, and global) for selecting proper training samples. In the following, we briefly introduced the technical flows of the SeGAT model.

Firstly, we proposed to establish $PM_{2.5}$ estimation models at each individual monitoring site using random forest given its good approximation capacity. Specifically, ground measured $PM_{2.5}$ concentrations were used as the learning target while the collocated AOD from the LGHAP v2 dataset were used as the proxy variable along with a set of explainable variables. The site-specific $PM_{2.5}$ estimation models can be formulated as:

$$PM_{2.5} \sim RF(AOD, MET, AER, LULC, DEM, NDVI, POP, month) \quad (1)$$

where AOD refers to the gap-free AOD grids from LGHAP v2 dataset. MET, AER, $LULC$, DEM, $NDVI$, POP, and $month$ denote meteorological variables, numerical aerosol diagnostics, land use and land cover, elevation, vegetation index, population, and month of the year, respectively. Therefore, tens of thousands of regional $PM_{2.5}$ concentration estimation models were established at the local scale across the globe.

Secondly, an adjacency matrix was calculated between each footprint of gap-filled AOD_{Terra} and monitoring sites overlaid grids in reference to nine distinct features indicating the scene attribute of each grid cell, i.e., *latitude*, *longitude*, *AOD*, *relative humidity*, *air temperature*, *NDVI*, *elevation*, *population*, and *land use and land cover ratio*. Specifically, the high-dimension Euclidian distance was calculated between grids on the basis of these nine features after normalization. The assumption is that the nonlinear interactions between AOD and $PM_{2.5}$ may comply with a similar relationship over scenes with comparable ambient environment. Therefore, $PM_{2.5}$ concentration over one grid could be estimated from models trained over sites with scene features similar to this given grid.

Thirdly, a graph attention network was then employed to integrate multiple $PM_{2.5}$ estimates derived from a set of site-specific models with similar scene features. Specifically, $PM_{2.5}$ estimates from 32 models with similar scene features were used as the learning input, while the normalized attribute differences were used as the weights in the adjacency matrix and the testing accuracy of each random forest model was used as the node bias. During the graph network training, the model utilized attention operations to discern crucial associations between scene attributes, and the model was continuously optimized by adjusting graph structures and incorporating residual connections. A global pooling layer was then employed to amalgamate contextual data from all nodes.

Distinct from other learning models, as a hybrid model, the proposed SeGAT model not only takes advantage of powerful approximation capacity of random forest but also accounts for spatial representativeness of each data-driven model. More importantly, the SeGAT model is capable of predicting $PM_{2.5}$ concentration even over regions without monitoring sites.

Table S1. Data accuracy of raw AOD datasets used for generating global gap-free LGHAP v2 AOD dataset by comparing against AOD observations from AERONET during 2000–2021.

Dataset	Region	Mean AOD	Number of monitors	Number of samples	R	RMSE	Bias	Below EE (%)	Within EE (%)	Above EE (%)
MCD19A2 (Aqua)	Global	0.17	1335	341254	0.88	0.11	0.01	12.11	75.45	12.44
	North America	0.11	433	94531	0.87	0.07	-0.01	3.72	82.54	13.74
	South America	0.11	81	20537	0.93	0.07	0.00	9.46	77.61	12.93
	Europe	0.11	208	83773	0.81	0.06	0.02	10.69	83.42	5.90
	Asia	0.32	321	79146	0.90	0.14	0.00	15.53	67.80	16.67
	Africa	0.21	110	40867	0.78	0.19	0.05	29.20	56.75	14.05
	Australia	0.09	28	10272	0.79	0.06	-0.02	4.81	76.71	18.48
VIIRS/NPP	Global	0.19	1335	204573	0.90	0.11	-0.01	9.68	75.58	14.73
	North America	0.12	433	69371	0.86	0.12	-0.01	6.76	81.61	11.63
	South America	0.08	81	15326	0.81	0.07	0.03	18.93	75.61	5.46
	Europe	0.13	208	45874	0.82	0.06	-0.01	4.42	83.62	11.96
	Asia	0.38	321	42570	0.91	0.15	-0.02	11.88	67.43	20.69
	Africa	0.23	110	25183	0.89	0.13	0.00	17.00	61.47	21.53
	Australia	0.11	28	4409	0.58	0.11	-0.04	3.38	65.28	31.34
MISR/Terra	Global	0.19	1335	79125	0.87	0.11	0.00	5.24	81.72	13.04
	North America	0.13	433	20839	0.79	0.09	-0.02	1.76	82.12	16.13
	South America	0.13	81	4526	0.89	0.12	0.00	4.20	87.38	8.42
	Europe	0.14	208	18630	0.87	0.05	0.00	2.59	90.85	6.56
	Asia	0.31	321	15792	0.85	0.18	0.02	12.61	72.44	14.96
	Africa	0.25	110	10003	0.87	0.14	0.00	7.56	73.78	18.66
	Australia	0.11	28	2241	0.76	0.07	-0.03	1.56	73.05	25.39
PARASOL/ POLDER	Global	0.30	1335	72120	0.86	0.18	-0.08	4.02	54.12	41.87
	North America	0.21	433	15849	0.68	0.16	-0.10	1.54	45.09	53.37
	South America	0.25	81	3235	0.95	0.16	-0.08	1.58	54.37	44.05
	Europe	0.20	208	19960	0.72	0.12	-0.05	3.47	63.65	32.88
	Asia	0.51	321	17651	0.85	0.24	-0.11	6.10	46.07	47.83
	Africa	0.39	110	8108	0.83	0.20	-0.07	7.24	56.18	36.58
	Australia	0.10	28	2171	0.69	0.07	-0.03	1.89	71.44	26.67
AATSR/ Envisat	Global	0.19	1335	30870	0.83	0.11	0.00	10.05	76.91	13.04
	North America	0.12	433	7828	0.87	0.06	0.00	5.81	86.89	7.29
	South America	0.12	81	1578	0.75	0.10	0.01	14.32	71.55	14.13
	Europe	0.14	208	8139	0.84	0.06	0.01	9.23	85.17	5.60
	Asia	0.31	321	5358	0.79	0.15	0.00	14.73	64.11	21.16
	Africa	0.30	110	3672	0.79	0.20	0.00	18.57	61.55	19.88
	Australia	0.13	28	997	0.36	0.13	-0.05	4.01	61.79	34.20
SeaWiFS/ OrbView-2	Global	0.21	1335	21643	0.88	0.12	0.00	12.34	70.64	17.02
	North America	0.12	433	4885	0.73	0.08	-0.02	5.69	75.78	18.53
	South America	0.17	81	1158	0.93	0.13	0.03	25.47	68.83	5.70
	Europe	0.16	208	3949	0.79	0.07	0.00	8.38	77.67	13.95
	Asia	0.32	321	3972	0.77	0.15	0.00	20.62	58.26	21.12
	Africa	0.34	110	3230	0.90	0.16	0.03	22.57	59.66	17.77
	Australia	0.07	28	717	0.34	0.09	0.00	11.30	73.92	14.78
AOD estimates derived from air quality indicators	Global	0.19	1335	203153	0.84	0.14	0.01	14.08	70.57	15.35
	North America	0.13	433	39913	0.78	0.11	-0.01	6.27	76.09	17.64
	South America	0.11	81	18282	0.81	0.10	0.02	17.27	71.44	11.29
	Europe	0.12	208	61389	0.71	0.07	0.01	10.47	81.32	8.21
	Asia	0.33	321	62283	0.84	0.19	0.03	20.41	61.96	17.62
	Africa	0.23	110	19041	0.72	0.19	0.00	19.67	50.11	30.21
	Australia	0.09	28	2245	0.67	0.07	-0.01	2.90	83.61	13.50

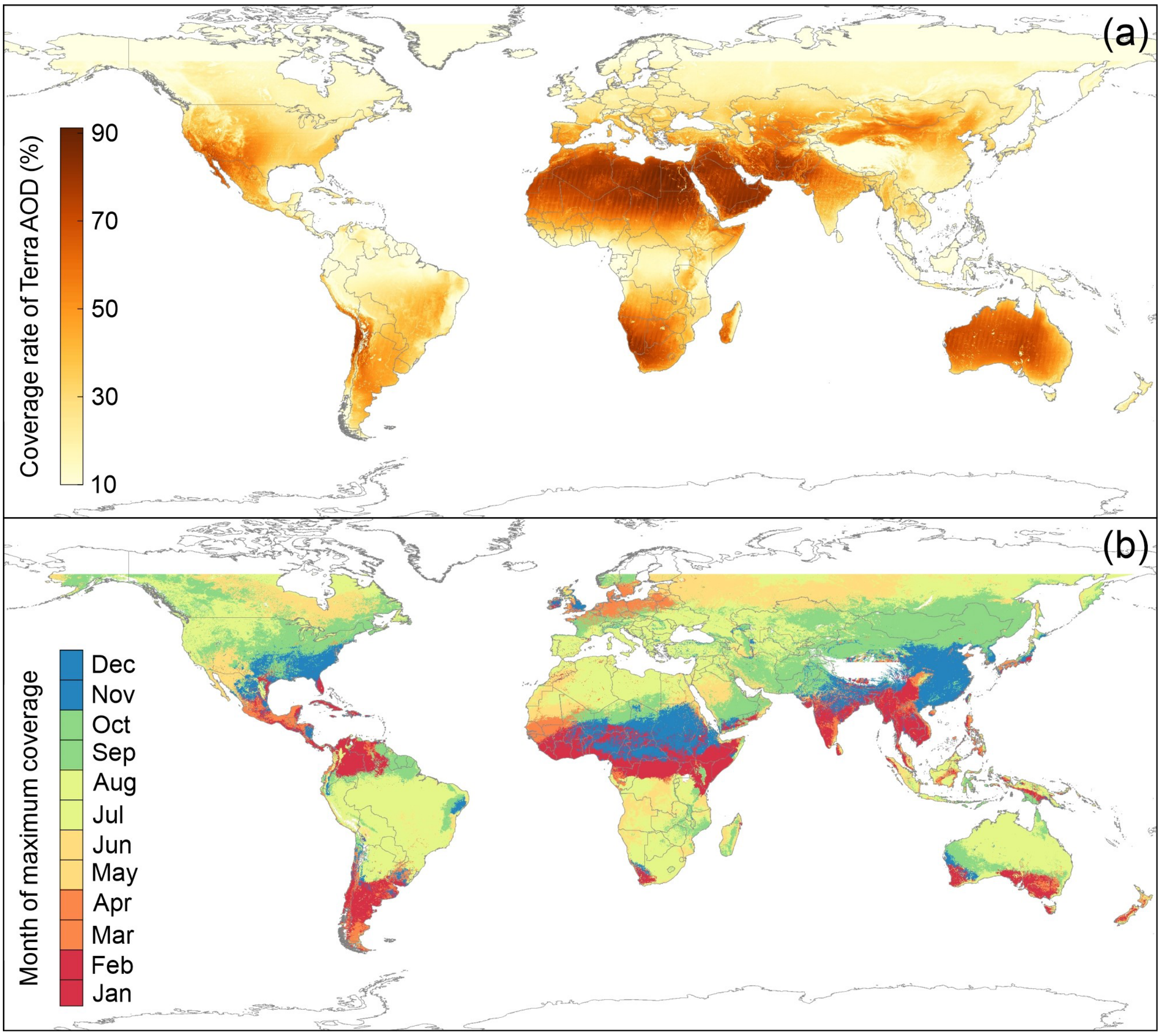

Figure S1. Spatial and temporal variations in AOD data coverage from Terra across the globe during 2000 to 2020.

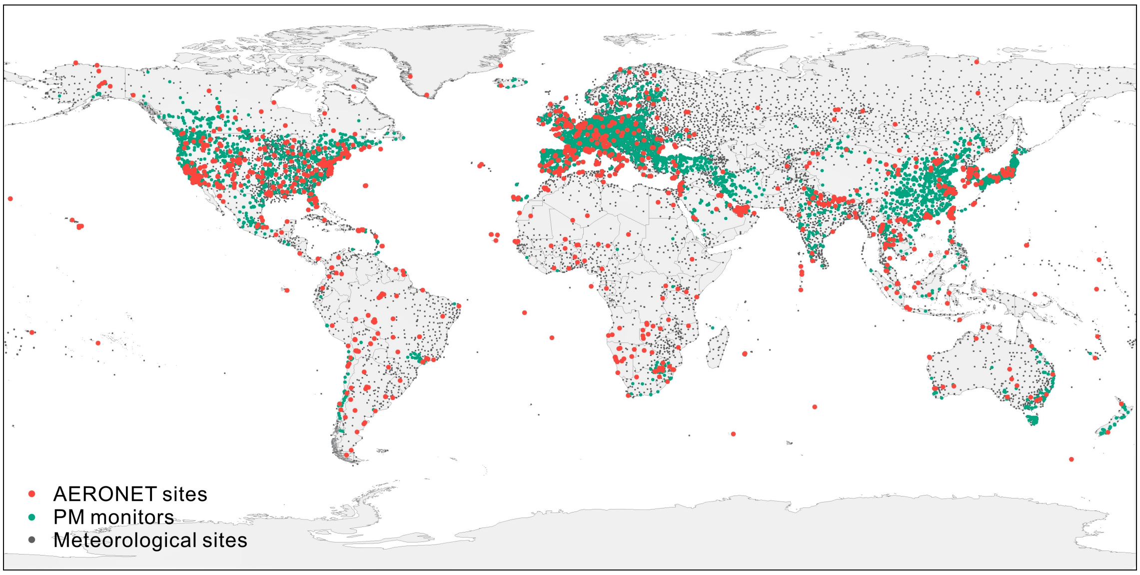

Figure S2. Spatial distribution of ground monitors providing AOD, PM, and atmospheric visibility used in this study across the globe.

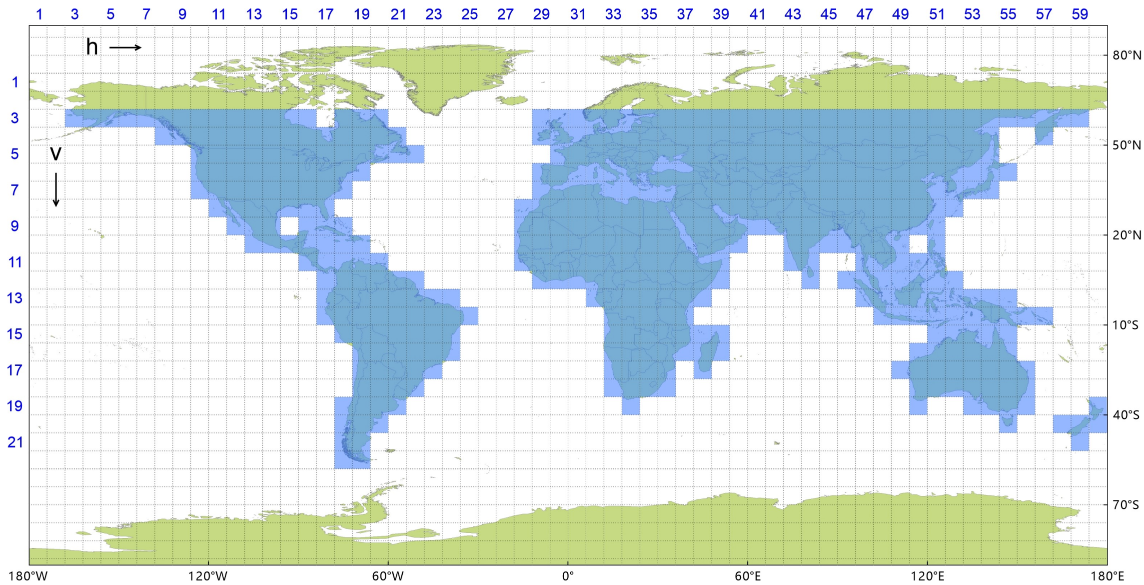

Figure S3. Spatial distribution of data tiles used for global-scale AOD gap-filling.

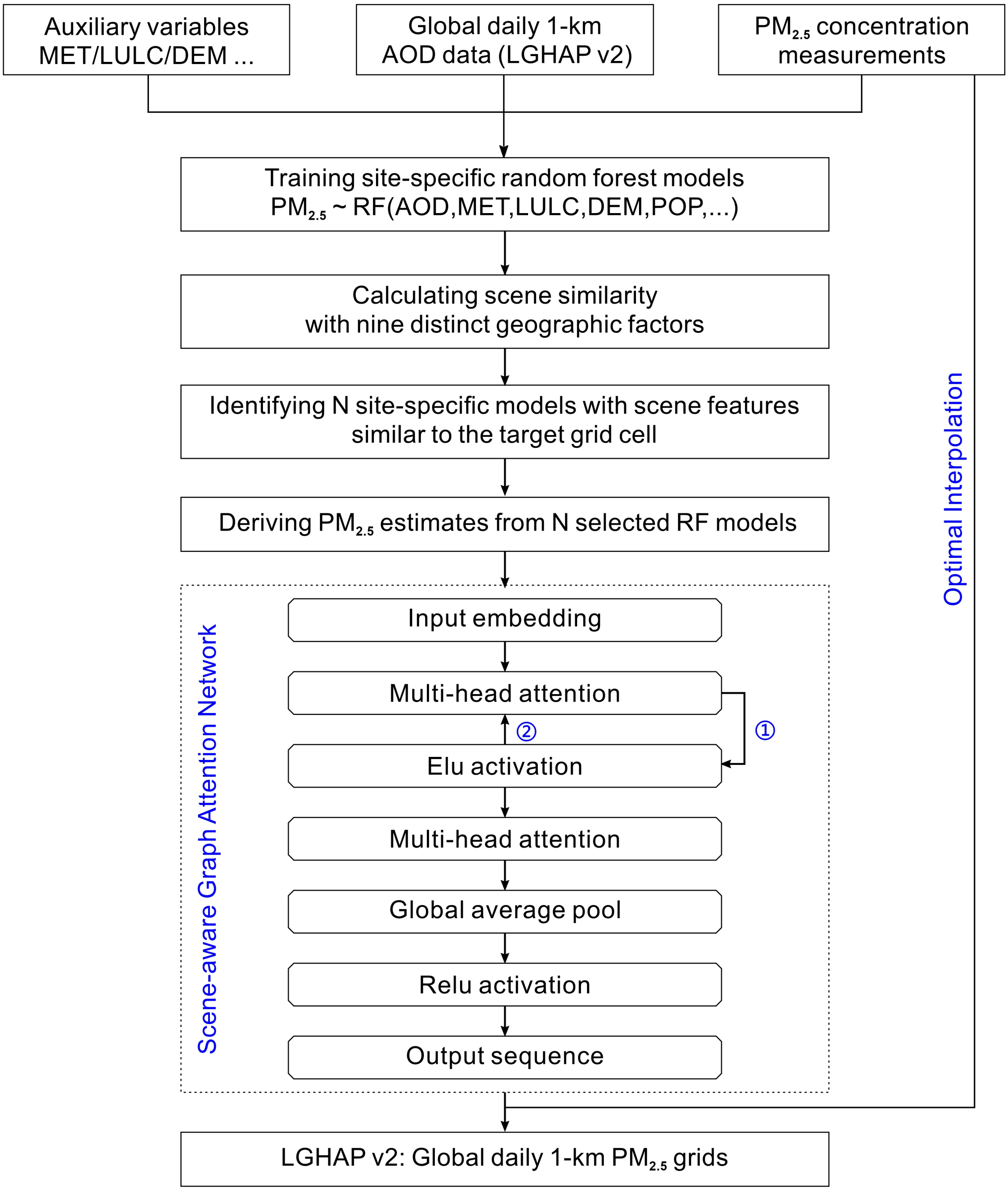

Figure S4. The flow chart of the scene-aware ensemble learning graph attention network (SCAGAT) model.

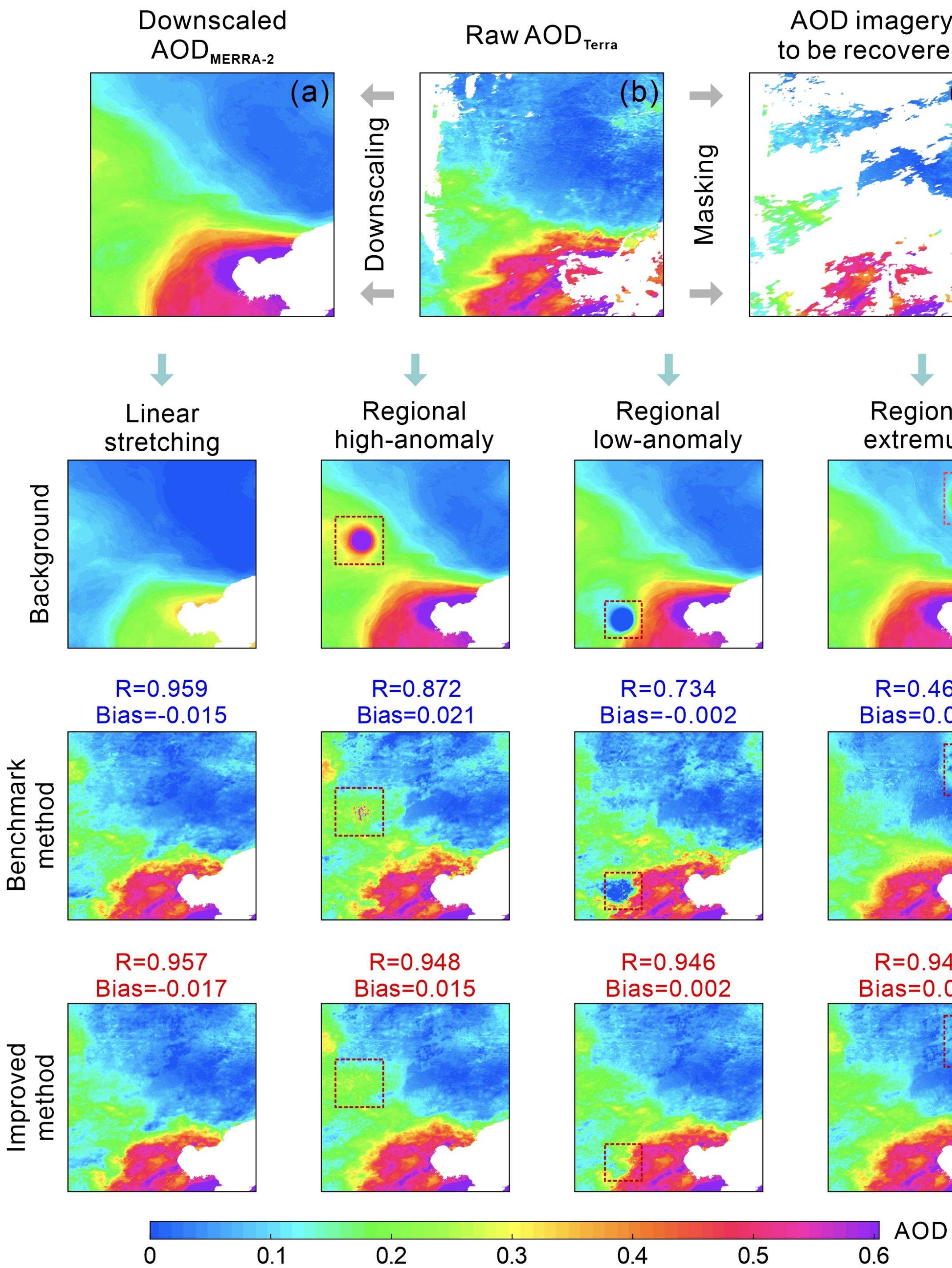

Figure S5. Performance evaluation of the adaptive background information updating module on improving AOD reconstruction patterns. Intercomparisons were conducted between the benchmark method (the method developed in Bai et al. (2022) to generate LGHAP dataset in China) and the one embedding adaptive background information updating module.

References

Bai, K., Li, K., Ma, M., Li, K., Li, Z., Guo, J., Chang, N.-B., Tan, Z., and Han, D.: LGHAP: the Long-term Gap-free High-resolution Air Pollutant concentration dataset, derived via tensor-flow-based multimodal data fusion, Earth Syst Sci Data, 14, 907–927, https://doi.org/10.5194/essd-14-907-2022, 2022.

Li, K., Bai, K., Ma, M., Guo, J., Li, Z., Wang, G., and Chang, N.-B.: Spatially gap free analysis of aerosol type grids in China: First retrieval via satellite remote sensing and big data analytics, ISPRS Journal of Photogrammetry and Remote Sensing, 193, 45–59, https://doi.org/10.1016/j.isprsjprs.2022.09.001, 2022.