Supporting Information for

Deciphering Isoprene Variability Across Dozen of Chinese and Overseas Cities Using Deep Transfer Learning

Song Liu[1], Xiaopu Lyu[2]*, Fumo Yang[1], Zongbo Shi[3], Xin Huang[4], Tengyu Liu[4], Hongli Wang[5], Mei Li[6], Jian Gao[7], Nan Chen[8], Guoliang Shi[9], Yu Zou[10], Chenglei Pei[11], Chengxu Tong[3], Xinyi Liu[1], Li Zhou[1], Alex B. Guenther[13], and Nan Wang[1]*

[1]College of carbon Neutrality Future Technology, Sichuan University, Chengdu 610065, China.

[2]Department of Geography, Hong Kong Baptist University, Hong Kong 000000, China.

[3]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham m B15 2TT, UK.

[4]School of Atmospheric Sciences, Nanjing University, Nanjing 210023, China.

[5]State Environmental Protection Key Laboratory of Formation and Prevention of Urban Air Pollution Complex, Shanghai Academy of Environmental Sciences, Shanghai, 200233, China.

[6]College of Environment and Climate, Institute of Mass Spectrometry and Atmospheric Environment, Guangdong Provincial Engineering Research Center for On-line Source Apportionment System of Air Pollution, Jinan University.

[7]Chinese Research Academy of Environmental Sciences, Beijing 100012, China.

[8]Research Centre for Complex Air Pollution of Hubei Province, Wuhan 430078, China.

[9]State Environmental Protection Key Laboratory of Urban Ambient Air Particulate Matter Pollution Prevention and Control, Tianjin Key Laboratory of Urban Transport Emission Research, College of Environmental Science and Engineering, Nankai University, Tianjin 300350, P. R. China.

[10]Institute of Tropical and Marine Meteorology, China Meteorological Administration, Guangzhou, China.

[11]Guangzhou Sub-branch of Guangdong Ecological and Environmental Monitoring Center, Guangzhou 510006, China.

[12]Department of Earth System Science, University of California, Irvine, California, USA.

Contents of this file

Text S1.

LAI and NDVI are effective proxies for urban vegetation cover and photosynthetic biomass, allowing for the monitoring of changes in vegetation structure and productivity over time (Chen and Black, 1992; Forzieri et al., 2020). To provide a more comprehensive representation of urban vegetation density and coverage, we introduced the metrics VI, which was derived from NDVI and LAI using principal component analysis (PCA). The NDVI was derived from corrected measurements of the Advanced Very High Resolution Radiometer, with a spatial resolution of 0.0833° and global coverage from 1990 to 2022 (Pinzon and Tucker, 2014). The LAI data for 2000 – 2021 was obtained from the Global Land Surface Satellite (GLASS) version 6 (LAI V6) with a resolution of 0.05°, while LAI for 1990 – 1999 was sourced from GLASS version 5 (LAI V5). Compared to the LAI V5, the LAI V6 retrieved by the Bi-LSTM deep learning model was more resistant to the noises or missing values and avoided the reconstruction of surface reflectance data (Ma and Liang, 2022). Therefore, in order to obtain more accurate LAI values, a random forest model was employed to correct the values of LAI V5 during 1990 – 1999. LAI V5, NDVI, and time variables (year and month) were used as independent variables to predict LAI V6. The RF model was trained on data from 2005 to 2018 and tested on data from 2000 to 2004. With the R^2 of 0.66 – 0.97, the good performance on the test datasets suggested that the model was effective in correcting the values of LAI V5 and accurately capturing the historical trend of LAI. Additionally, the NDVI and LAI were downscaled to a 0.25° × 0.25° grid resolution, with the sampling sites at the center, to assess vegetation cover changes at the city scale. Through the PCA analysis, the principal component 1 with an explained variance ratio of 0.98 across all the sites was assigned as VI.

Text S2.

The study has the following limitations. First, although the machine learning model we developed showed its data imputation capability at the data-sparse sites, this approach requires site-specific observational data for optimal performance, limiting its immediate global applicability. Future research should explore data-efficient strategies such as semi-supervised learning to overcome this constraint.

Second, our study focuses on ambient isoprene concentrations rather than emissions. Therefore, the results may not directly guide emission-based numerical simulations. However, the predicted concentrations and their drivers, particularly temperature, radiation, and vegetation indices, provide valuable insights into biogenic emission patterns. The pronounced increase in isoprene concentrations observed at the suburban sites in both London and Hong Kong after 2012 served as a compelling evidence of climate warming's impact on biogenic emissions. In Hong Kong, the sustained upward trend in isoprene concentrations over recent decades likely reflected enhanced emissions driven by urban greenspace expansion. The contrasting importance of vegetation indices between these two cities further underscored how regional differences in vegetation composition and emission characteristics influence local air quality. These findings contribute to our understanding of biogenic isoprene emissions under changing climatic and urban conditions, providing crucial insights for sustainable city development in a warming world.

Third, chemical loss of isoprene was not considered with specific proxies in the model. Isoprene is primarily consumed by reacting with hydroxyl radical (OH) in the daytime. Since the availability of OH data is limited, O_3 is generally used as an OH proxy. We attempted to use O_3 as an input feature, but the model showed a positive isoprene-O_3 relationship, due to the similar diurnal patterns between them, contributions of isoprene to O_3, and their common sensitivities to temperature. It is also difficult to obtain the data of indicative oxidation products of isoprene, such as methyl vinyl ketone. In fact, OH concentration is closely related to meteorological parameters, especially radiation and temperature. By adopting these parameters as input features, we believe that the chemical loss of isoprene was considered by the model. Despite this, the positive responses of isoprene to radiation and temperature suggest that the effect of emissions overwhelmed that of chemical loss. Indeed, this was confirmed by the diurnal pattern of the observed isoprene concentrations across various sites (Figure S7).

Fourth, we assume the concentrations and compositions of many air pollutants, except isoprene and NO_x, unchanged in the simulation of future O_3. This probably led to an overestimate of O_3. However, the conclusions regarding the effects of temperature rise, isoprene increase and NO_x reduction should still hold true.

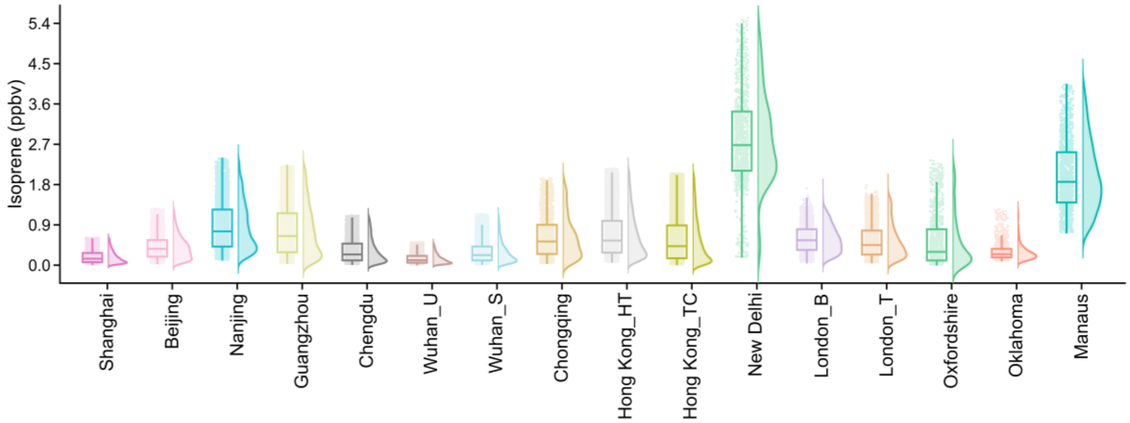

Figure S1. Box plot and distribution of isoprene concentrations at each site. The upper and lower edges of the box denote the third and first quartiles, respectively, while the solid line within the box represents the median. The whiskers extend to 1.5 times the interquartile range.

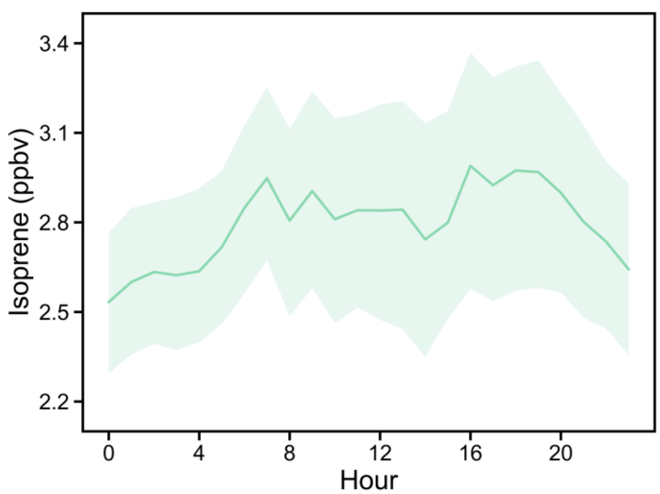

Figure S2. Diurnal variations of isoprene concentrations at the New Delhi site. The bands represent 95% confidence intervals.

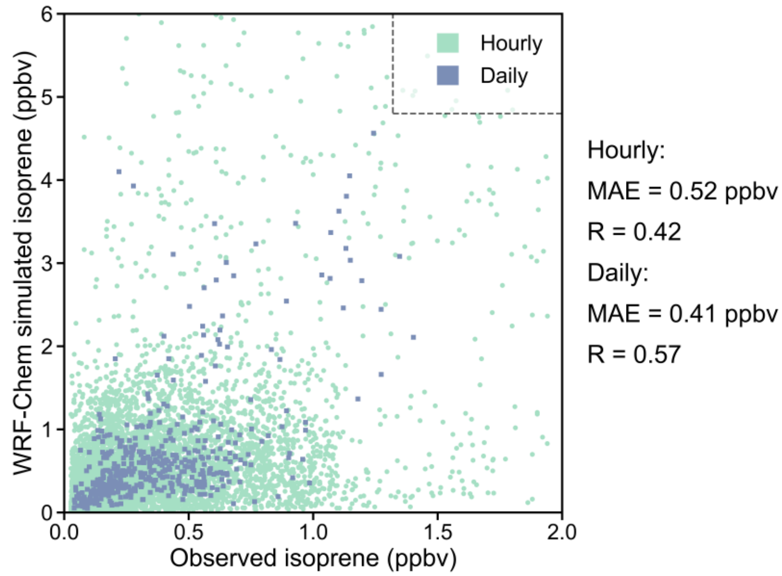

Figure S3. Comparisons of WRF-Chem simulated and measured isoprene concentrations.

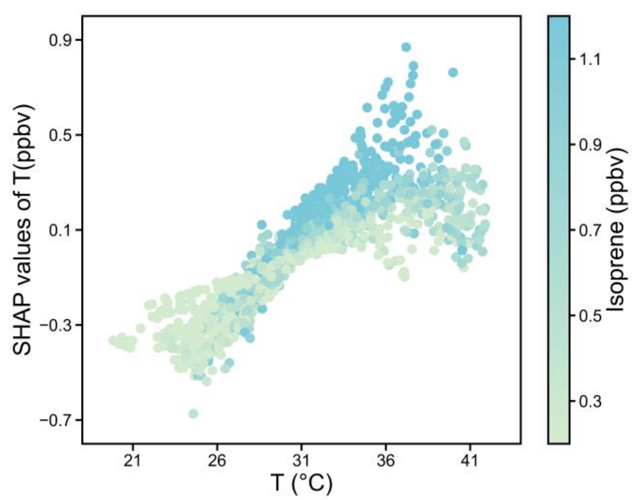

Figure S4. The SHAP dependence plot of temperature at the Chongqing site.

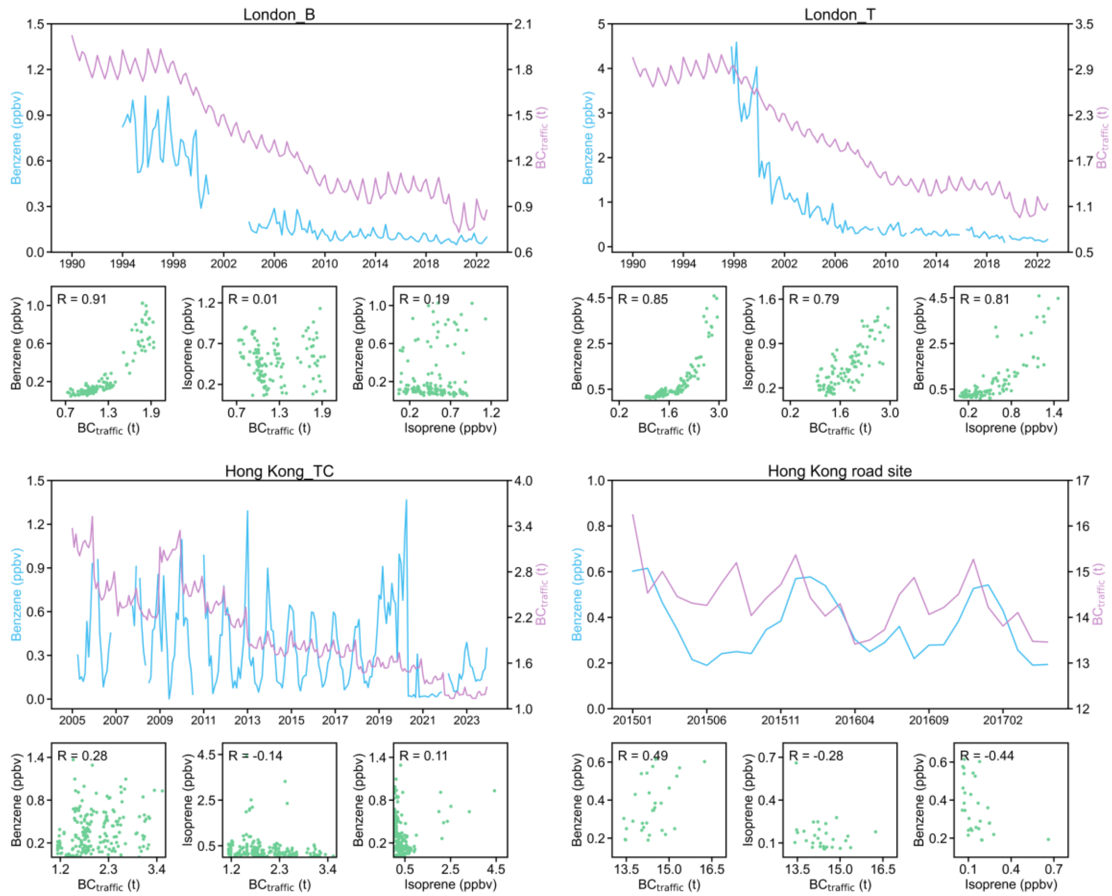

Figure S5. Correlation analysis of monthly isoprene concentrations with benzene and $BC_{traffic}$ in Hong Kong and London.

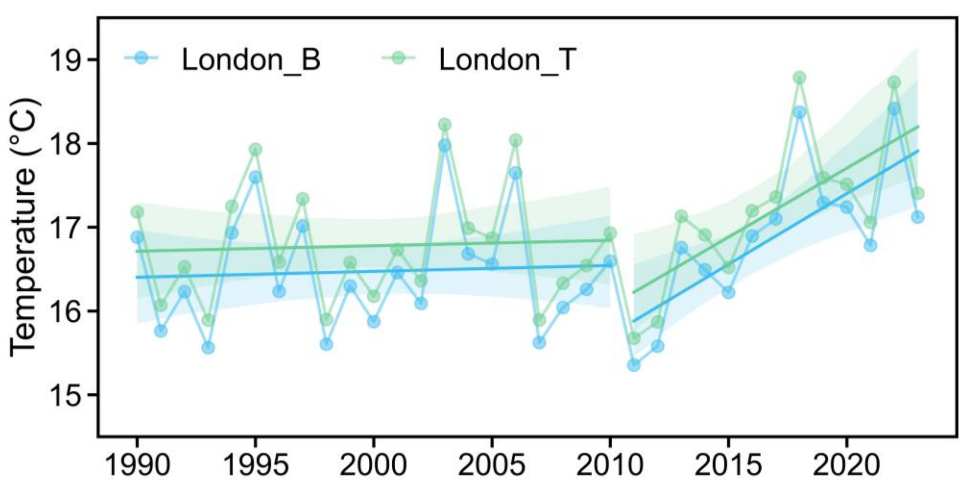

Figure S6. Variations of average summer temperature at the London_B and London_T sites from 1990 to 2023.

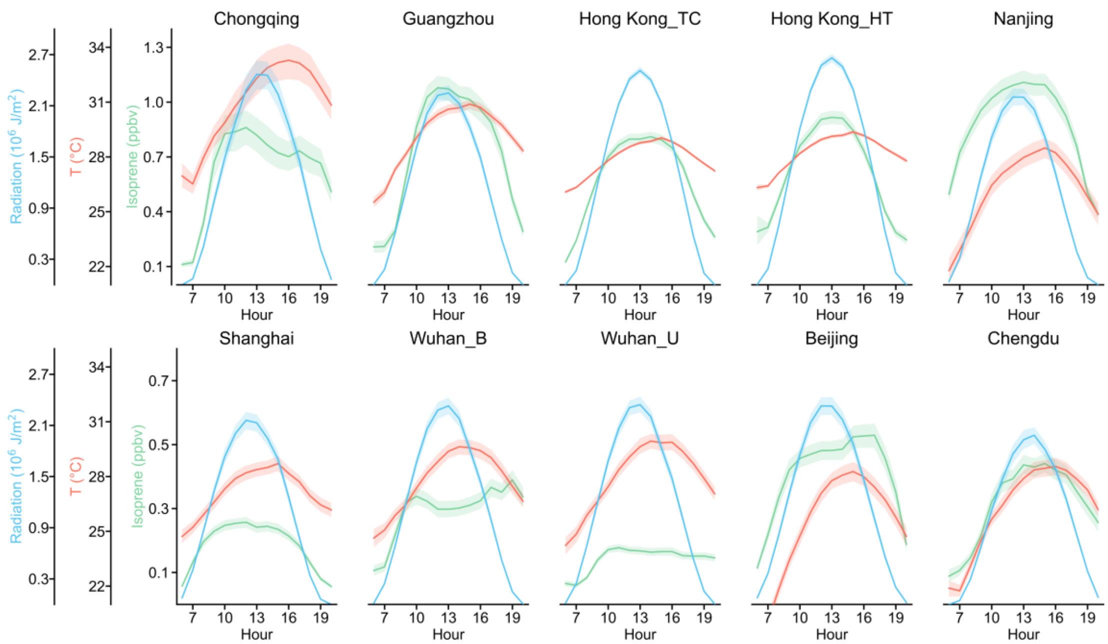

Figure S7. Diurnal variations in isoprene concentrations, temperature, and solar radiation across different sites. The bands represent 95% confidence intervals.

Site	Time coverage	Latitude	Longitude	Number of valid hourly data	Temporal resolution	Site category	Instrument
Beijing	May to September in 2021 and 2022	40.05°	116.42°	3464	hourly	Urban site	GC–FID/MS
Chengdu	July to September from 2019 to 2022	30.66°	104.04°	4574	hourly	Urban site	Synspec GC955-611/811
Chongqing	July to August in 2021 and 2022	29.62°	106.51°	1503	hourly	Urban site	Synspec GC955-611/811
Guangzhou	May to September in 2019 and 2021	23.08°	113.37°	4111	hourly	Urban site	AC-GCMS1000
Hong Kong_TC	May to September from 2005 to 2020	22.29°	113.94°	20775	hourly	Suburban site	GC-PID
Hong Kong_HT	May to September from 2013 to 2023	22.22°	114.26°	9900	hourly	Urban site	GC-PID
Nanjing	June to October in 2017, 2018, 2022, and 2023	32.12°	118.96°	4683	hourly	Urban site	GC-MS/FID
Shanghai	June to September from 2021 to 2023	31.17°	121.43°	4692	hourly	Urban site	GC-FID
Wuhan_U	May to September from 2021 to 2023	30.53°	114.37°	5161	hourly	Urban site	GC-FID/MS

Wuhan_S	May to September from 2021 to 2023	30.60°	114.28°	4974	hourly	Urban site	GC-FID/MS
London_T	May to September; 1994 to 2022	51.45°	0.07°	2061	daily	Traffic site	Perkin Elmer Ozone Precursor Analysers
London_B	May to September; 1999 to 2022	51.52°	0.16°	2063	daily	Suburban site	Perkin Elmer Ozone Precursor Analysers
Oklahoma	April to September; 2016	36.60°	-97.49°	1064	hourly	Rural site	PTR-MS
Manaus	February to April; 2016	-3.10°	-59.99°	1194	hourly	Urban site	PTR-MS
Oxfordshire	June to September; 2018	51.46°	-1.20°	1025	hourly	Forest site	GC-PID
New Delhi	January to March; 2018	28.45°	77.28°	968	hourly	Suburban site	PTR-TOF-MS 8000

Table S1. Detailed information of isoprene observational data at each site.

Predictor variables	Abbreviations	Temporal coverage	Temporal resolution	Spatial coverage	Spatial resolution
Vegetation index	VI	1990-2023	8 days	global	0.25°
Traffic emissions of black carbon	BCtraffic	1990-2023	monthly	global	0.1°
2m Temperature	T	1990-2023	hourly	global	0.1°
Surface solar radiation downwards	SSRD	1990-2023	hourly	global	0.25°
Soil moisture	SWV	1990-2023	hourly	global	0.1°
Relative humidity	RH	1990-2023	hourly	global	0.1°
Surface pressure	SP	1990-2023	hourly	global	0.1°
10-meter Zonal wind component	u10	1990-2023	hourly	global	0.1°
10-meter Meridional wind component	v10	1990-2023	hourly	global	0.1°
Evaporation from vegetation transpiration	EVAVT	1990-2023	hourly	global	0.1°
Boundary layer height	BLH	1990-2023	hourly	global	0.25°
Total precipitation	TP	1990-2023	hourly	global	0.1°

Table S2. Detailed information of variables used for isoprene concentrations prediction.

Site name	Training strategy	Site type	Pre-training dataset	Fine-tuning/retraining dataset
Chongqing	T-training	Pre-training	Data from pre-training sites except Chongqing	Training data from Chongqing
	NT-training		/	Training data from Chongqing
	MIX-training		/	Data from pre-training sites except Chongqing + Training data from Chongqing
Chengdu	T-training	Pre-training	Data from pre-training sites except Chengdu	Training data from Chengdu
	NT-training		/	Training data from Chengdu
	MIX-training		/	Data from pre-training sites except Chengdu + Training data from Chengdu
Wuhan_U	T-training	Pre-training	Data from pre-training sites except Wuhan_U	Training data from Wuhan_U
	NT-training		/	Training data from Wuhan_U
	MIX-training		/	Data from pre-training sites except Wuhan_U + Training data from Wuhan_U
Wuhan_S	T-training	Pre-training	Data from pre-training sites except Wuhan_S	Training data from Wuhan_S
	NT-training		/	Training data from Wuhan_S
	MIX-training		/	Data from pre-training sites except Wuhan_S + Training data from Wuhan_S
Shanghai	T-training	Pre-training	Data from pre-training sites except Shanghai	Training data from Shanghai
	NT-training		/	Training data from Shanghai

	MIX-training		/	Data from pre-training sites except Shanghai + Training data from Shanghai
Nanjing	T-training	Pre-training	Data from pre-training sites except Nanjing	Training data from Nanjing
	NT-training		/	Training data from Nanjing
	MIX-training		/	Data from pre-training sites except Nanjing + Training data from Nanjing
Beijing	T-training	Pre-training	Data from pre-training sites except Beijing	Training data from Beijing
	NT-training		/	Training data from Beijing
	MIX-training		/	Data from pre-training sites except Beijing + Training data from Beijing
Hong Kong_TC	T-training	Pre-training	Data from pre-training sites except Hong Kong_TC	Training data from Hong Kong_TC
	NT-training		/	Training data from Hong Kong_TC
	MIX-training		/	Data from pre-training sites except Hong Kong_TC + Training data from Hong Kong_TC
Hong Kong_HT	T-training	Pre-training	Data from pre-training sites except Hong Kong_HT	Training data from Hong Kong_HT
	NT-training		/	Training data from Hong Kong_HT
	MIX-training		/	Data from pre-training sites except Hong Kong_HT + Training data from Hong Kong_HT
Guangzhou	T-training	Pre-training	Data from pre-training sites except Guangzhou	Training data from Guangzhou
	NT-training		/	Training data from Guangzhou

	MIX-training		/	Data from pre-training sites except Guangzhou + Training data from Guangzhou
London_T	PINN-ResMLP$_T$	Validation	All pre-training sites	Training data from London_T
	other models		/	Training data from London_T
London_B	PINN-ResMLP$_T$	Validation	All pre-training sites	Training data from London_B
	other models		/	Training data from London_B
New Delhi	PINN-ResMLP$_T$	Validation	All pre-training sites	Training data from New Delhi
	other models		/	Training data from New Delhi
Manaus	PINN-ResMLP$_T$	Validation	All pre-training sites	Training data from Manaus
	other models		/	Training data from Manaus
Oklahoma	PINN-ResMLP$_T$	Validation	All pre-training sites	Training data from Oklahoma
	other models		/	Training data from Oklahoma
Oxfordshire	PINN-ResMLP$_T$	Validation	All pre-training sites	Training data from Oxfordshire
	other models		/	Training data from Oxfordshire

Table S3. Pre-training and fine-tuning datasets for different training strategies at each site.

Machine learning algorithm	Hyperparameters	Number of models
Extreme gradient boosting (XGB)	n_estimators: 100, 200, 300 max_depth: 20, 30 learning_rate: 0.2, 0.5, 0.8, 1 colsample_bytree: 0.8, 1.0	48
Random forest (RF)	n_estimators: 100, 200, 300 min_samples_split: 5, 10, 15, 20 max_depth: 10, 20	24
Gradient boosting decision tree (GBDT)	n_estimators: 100, 200, 300 learning_rate: 0.1, 0.3, 0.6, 0.8, 1	15
Support vector machine (SVM)	C: 1, 5, 10, 100, 1000 kernel: linear, poly, rbf	15
Linear regression (LR)	default	1

Table S4. Hyperparameters used for different machine learning algorithms.

Reference

Chen, J. M. and Black, T. A.: Defining leaf area index for non-flat leaves, Plant, Cell Environ., 15, 421-429, https://doi.org/10.1111/j.1365-3040.1992.tb00992.x, 1992.

Forzieri, G., Miralles, D. G., Ciais, P., et al.: Increased control of vegetation on global terrestrial energy fluxes, Nat. Clim. Change, 10, 356-362, 10.1038/s41558-020-0717-0, 2020.

Ma, H. and Liang, S.: Development of the GLASS 250-m leaf area index product (version 6) from MODIS data using the bidirectional LSTM deep learning model, Remote Sens. Environ., 273, 112985, https://doi.org/10.1016/j.rse.2022.112985, 2022.