Supplementary of Bridging the Data Gap: Enhancing the Spatiotemporal Accuracy of Hourly $PM_{2.5}$ Concentration through the Fusion of Satellite-derived Estimations and Station Observations

Wenhao Chu [1], Chunxiao Zhang [1,2,*] and Heng Li [2]

[1] School of Information Engineering, China University of Geosciences in Beijing, No. 29, Xueyuan Road, Haidian District, Beijing 100083, China; whchu@ email.cugb.edu.cn (W.C.); 3004210009@email.cugb.edu.cn (H.L.)
[2] Observation and Research Station of Beijing Fangshan Comprehensive Exploration, Ministry of Natural Resources, Beijing 100083, China; e-mail@e-mail.com
* Correspondence: zcx@cugb.edu.cn; Tel.: +86-186-1821-9549

S1. The autogeoi-stacking method

As a new GeoAI technique, the autogeoi-stacking method is based on automated feature engineering (abbreviated as autofeat) and stacking approaches with tuned hyperparameters [1]. Compared with the baseline models (e.g., Random Forest), the three parts improved by 8%, wherein the autofeat, stacking, and hyperparameter tuning resulted in 2%, 3%, and 3%, respectively. The other parts will be introduced successively in the following sections.

Firstly, autofeat can enhance the information of the dataset in machine learning and thus is an essential part of automated machine learning (AutoML) [2–4]. The approach usually contains two parts: automatic feature synthesis and automatic feature selection. The first part will create features automatically from a candidate dataset based on multiple mathematic operations, such as logarithmic, sine, and multiplication (from n features to n+m features), and the second part will select several optimal ones from them (from n+m features to n+m-j features). This study implements this method using the autofeat library (version 2.0.10) in Python.

Secondly, the stacking method is one of the ensemble learning methods (e.g., boosting, bagging) that integrates the results of different models. It is concerned with combining multiple outputs by using different machine learning algorithms (L1, …, LN) on the same dataset. In the first phase, a set of base-level outputs (O1, …, ON) is generated through different learners (L1, …, LN). In the second phase, a second-level algorithm (also called meta-learner) that combines the outputs of the base-level methods is used [5]. To obtain superior accuracy, the base-level learners should perform well, and the diversity of these learners should be as high as possible. In this study, the stacking method comprises seven tree-based models, including random forest, extremely randomized trees, gradient boosting decision tree, extreme gradient boosting, light gradient boosting machine, histogram-based gradient boosting, and Catboost.

Most state-of-the-art machine learning methods need to be tailored to specific tasks by selecting an appropriate set of hyperparameters. In the example of random forests, the number of estimators, the number of features per estimator, or the minimal number of samples per leaf have to be tuned. The typical procedure to tune hyper-parameter sets is as

follows: The dataset is split into a training, a validation, and a test set. Different hyperparameter sets are trained on the training data and tested on the validation data. The best-performing model is used as the final model, retrained on the training and validation sets, and tested on the test set. In this study, the hyperparameters of machine learning methods are tuned using the Optuna library (version 2.9.1) in Python. The optimized parameters and value ranges can be found in our previous work [1].

S2. Long-term gap-free high-resolution air pollutant

The Long-term Gap-free High-resolution Air Pollutants concentration dataset (abbreviated as LGHAP) is of great significance for environmental management and earth system science analysis [6]. The current release of the LGHAP aerosol dataset (LGHAP.v1) provides a 21-year-long (2000-2020) AOD product with a daily 1-km resolution that covers China's land area and is free of gaps.

This dataset was generated using a seamless integration of tensor flow-based multimodal data fusion and ensemble learning-based knowledge transfer in statistical data mining. The proposed method involved transforming a set of data tensors, including AOD and other related datasets, such as air pollutant concentrations and atmospheric visibility, acquired from diverse sensors or platforms through integrative efforts of spatial pattern recognition.

The daily resolution AOD, $PM_{2.5}$, and PM_{10} datasets are publicly available at https://doi.org/10.5281/zenodo.5652257, https://doi.org/10.5281/zenodo.5652265, and https://doi.org/10.5281/zenodo.5652263, respectively. Monthly and annual datasets can be acquired from https://doi.org/10.5281/zenodo.5655797 and https://doi.org/10.5281/zenodo.5655807, respectively. The daily AOD data agrees with the Aerosol Robotic Network (AERONET) with a correlation coefficient (R) of 0.91 and RMSE equaling 0.21. Meanwhile, $PM_{2.5}$ and PM_{10} estimations also agreed well with ground measurements, with R values of 0.95 and 0.94 and root mean squared error (RMSE) values of 12.03 and 19.56 $\mu g/m^3$.

S3. Landscan population

In this paper, we utilized the LandscanTM High-Resolution Global Population Dataset to represent the population of the Beijing–Tianjin–Hebei (BTH) region in 2018.

This dataset, which Oak Ridge National Laboratory developed, provides an ambient (average over 24 h) global population distribution at approximately 1 km spatial resolution by modeling population distribution with the best available demographic and geographic data and remote sensing imagery analysis techniques [7].

S4. Population-weighted exposure

Population-weighted exposure (PWE) is a method to estimate the average exposure level of a population to air pollutants [8,9]. It is calculated by multiplying the population size and the pollutant concentration in different areas and dividing it by the total population. This metric can better reflect the actual exposure level of people because it gives more weight to the areas where more people live. The formula of PWE is depicted below:

$$PWE = \frac{1}{P}\sum_i (P_i \cdot C_i) \qquad (S1).$$

In Equation (S1), i represents each grid in the study region. Pi and Ci are the population and $PM_{2.5}$ concentrations at grid i, respectively. Moreover, P indicates the total population of the study region.

S5. Regional exposure risk (RER)

The regional exposure risk (RER) of air pollution refers to the time when the air pollutant concentration exceeds the limitation in a region, which reflects the length or proportion of time people in this region are exposed to air pollution. The exceedance frequency was used to estimate the annual and seasonal RER of $PM_{2.5}$, and the calculation formula is:

$$RER_j = \frac{count(C_j > S)}{n} \quad (S2),$$

where RERj is the RER of $PM_{2.5}$ in grid j, Cj is the $PM_{2.5}$ concentration in grid j, and S is the restrictions of annual (35 μg/m^3) and daily $PM_{2.5}$ (75 μg/m^3) concentrations of GB 3095-2012 (see Table S2). The annual and seasonal RERs were calculated based on the annual $PM_{2.5}$ concentration and daily $PM_{2.5}$ concentration, respectively. n represents the time span, which is years for annual RER and days of each season for seasonal RER. The value of RER ranges from 0 to 1. A value of 0 indicates that there is no event exceeding the pollution standard of the restriction, while the larger the value of RER is, the greater the proportion of pollution events, and a value of 1 means that the $PM_{2.5}$ concentration exceeds the standard of GB 3095-2012 throughout the whole period.

Table S1. Populations of each subregion of the BTH region (the unit of population is ten thousand people).

	0-14	15-64	>65	Overall
Beijing	226.4	1706.7	237.6	2170.7
Tianjin	158.65	1240.55	157.67	1556.87
Hebei	1402.39	5271.18	845.95	7519.52

Table S2. Summary of datasets and sources used in this study.

Categories	**Content**	**Unit**	**Spatial resolution**	**Data Source**
Ground Truth	$PM_{2.5}$	$\mu g/m^3$	Point	CNEMC
Satellite Retrieval	$PM_{2.5}$	$\mu g/m^3$	0.05° x 0.05°	CHAP
Meteorological	10m u-component of wind	m/s	0.25° x 0.25°	ERA5
	10m v-component of wind	-	-	-
	100mu-component of wind	-	-	-
	100m v-component of wind	-	-	-
	2m temperature	K	-	-
	2m dewpoint temperature	-	-	-
	Relative humidity	%	-	-
	Surface pressure	Pa	-	-
	Boundary-layer height	m	-	-
	Total precipitation	-	-	-
	K Index	K	-	-
Aerosols	$PM_{2.5}$	$\mu g/m^3$	0.5° x 0.625°	MERRA2
	Black carbon aerosol	-	-	-
	Organic carbon aerosol	-	-	-
	Dust aerosol	-	-	-
	Sulfate aerosol	-	-	-
	Sea salt aerosol	-	-	-
	Carbon monoxide	-	0.136° x0.136°	CAQRA
	Ozone	-	-	-
Topographic	Surface elevation	m	90m	SRTM

Table S3. National ambient air quality of annual and 24-hour mean $PM_{2.5}$ concentration.

	Level 1	Level 2
Annual	15	35
24-h	35	75

Table S4. Comparison of generated data in this study: China High Air Pollutant (CHAP), LGHAP, and the second Modern-Era Retrospective analysis for Research and Applications (MERRA-2). From left to right in the table below, the sub-titles are spatial and temporal resolutions, pairs of images, gap-free (yes or no), coverage ratios of available samples, and annual mean values of these $PM_{2.5}$ datasets.

	Spatial	Temporal	Pairs	Gap-free	Coverage ratios	Annual mean
This study	0.05° x 0.05°	Hourly	8760	Yes	100%	44.97±14.19
CHAP	-	Hourly	3998	No	32.62%	43.91±12.94
LGHAP	0.5° x 0.625°	Daily	365	Yes	100%	43.67±16.59
MERRA-2	0.5° x 0.625°	Hourly	8760	Yes	100%	33.19±12.30

Table S5. Summary of mean values of each subregion across the BTH region during each hour.

City	Mean	City	Mean
Baoding	53.26	Langfang	54.53
Beijing	44.19	Qinhuangdao	43.13
Cangzhou	55.54	Shijiazhuang	60.19
Chengde	29.06	Tangshan	49.64
Handan	64.93	Tianjin	52.71
Hengshui	58.62	Xingtai	62.39
Zhangjiakou	29.07	BTH	44.97

Table S6. Summary of comparison with relevant studies. Bold indicates the value with the best metric (R^2 and RMSE).

	Methods	Annual	monthly	daily	hourly
This study		**0.93, 3.58**	**0.96, 5.21**	**0.92,** 12.24	**0.82**-0.88, 16.12-22.06
Xiao et al. (2017)	MI[1]+LME[2]+GAM[3]	0.73-0.81, 18-25	-	0.71-0.82, 17-24	-
Hua et al. (2019)	IVW[4]+GAM	0.63-0.86,13.83-27.48	-	-	-
Jing et al. (2023)	LME+GWR[5]	**-**	0.92, 5.72	-	-
Xue et al. (2019)	ML[6]+GAM	0.75, 9.9	0.68, 18.1	0.61, 27.8	-
Liu et al. (2022)	RF[7]	-	0.90,18.7	0.89,**12.0**	0.84, **15.9**
Geng et al. (2021)	RF	0.80-0.88, 13.9-22.1	-	-	-

1 MI: multiple imputation

2 LME: linear mixed-effects model.

3 GAM: generalized additive model.

4 IVW: inversed variance weights.

5 GWR: geographical weighted regression.

6 ML: high-dimensional expansion (HD-expansion) + elastic-net regression.

7 RF: random forest.

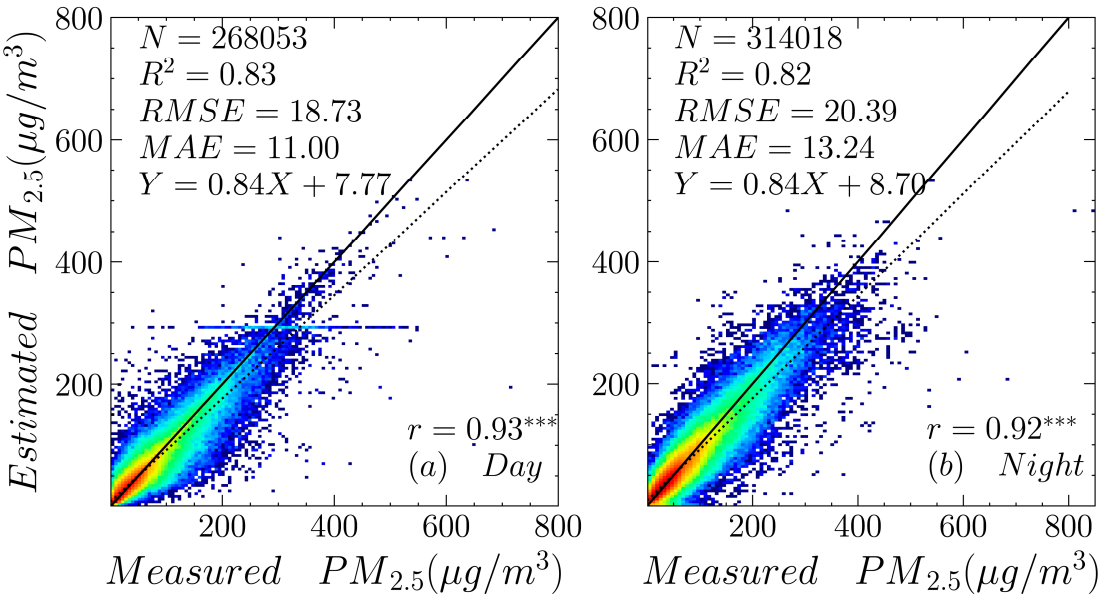

Figure S1. Density scatterplots of results of $PM_{2.5}$ estimates (μg/m³) during the (a) day and (b) night.

The dashed and solid lines denote 1:1 and best-fit lines from linear regression, respectively.

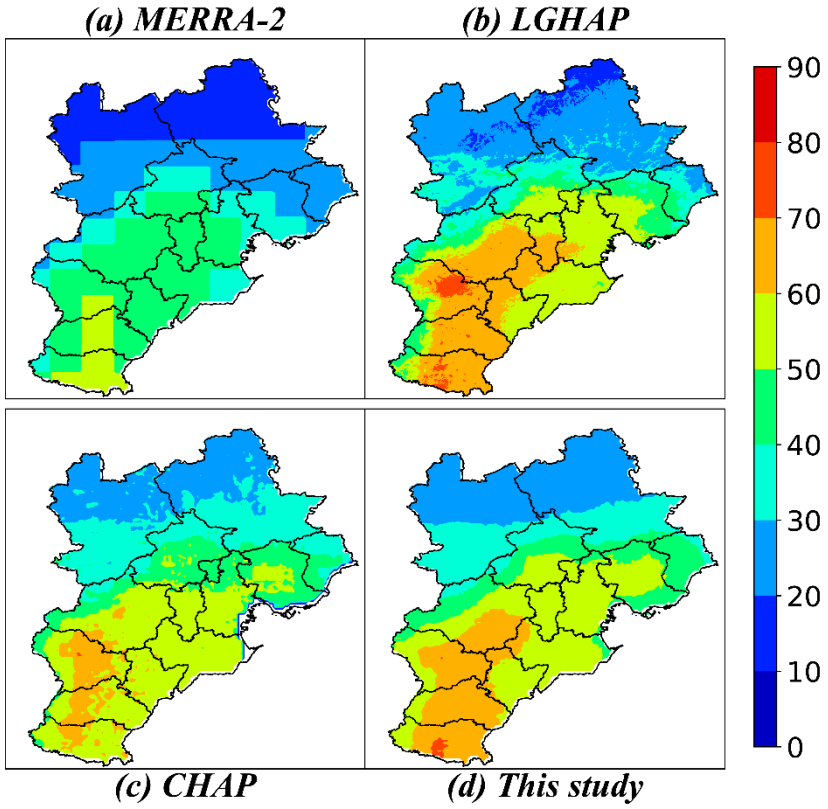

Figure S2. Annual $PM_{2.5}$ distributions of MERRA-2, CHAP, LGHAP, and this study.

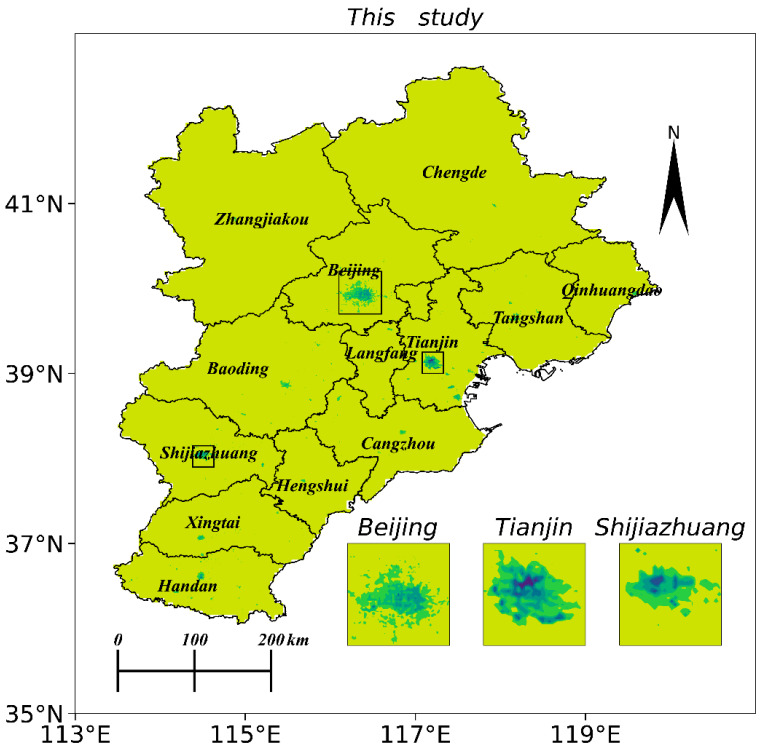

Figure S3. Distributions of calculated PWE using the generated PWE in this study.

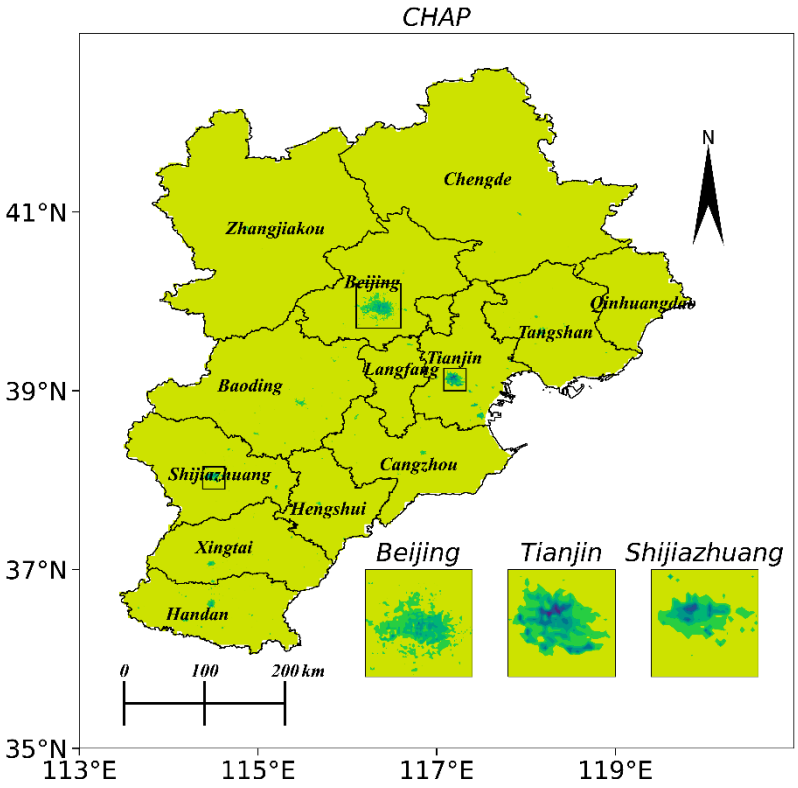

Figure S4. Distributions of calculated PWE using the CHAP data.

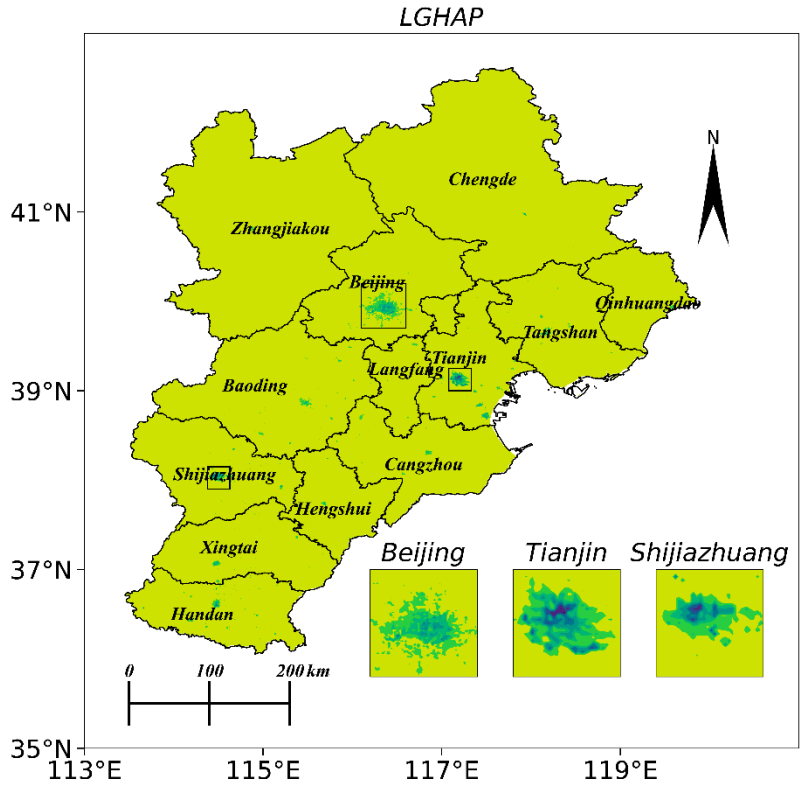

Figure S5. Distributions of calculated PWE using the LGHAP data.

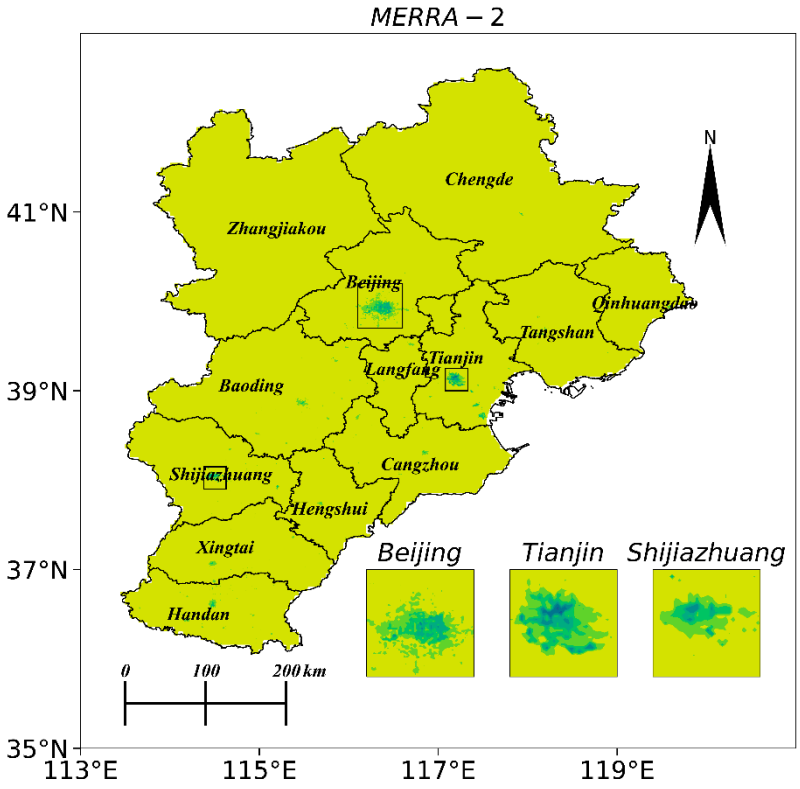

Figure S6. Distributions of calculated PWE using the LGHAP data.

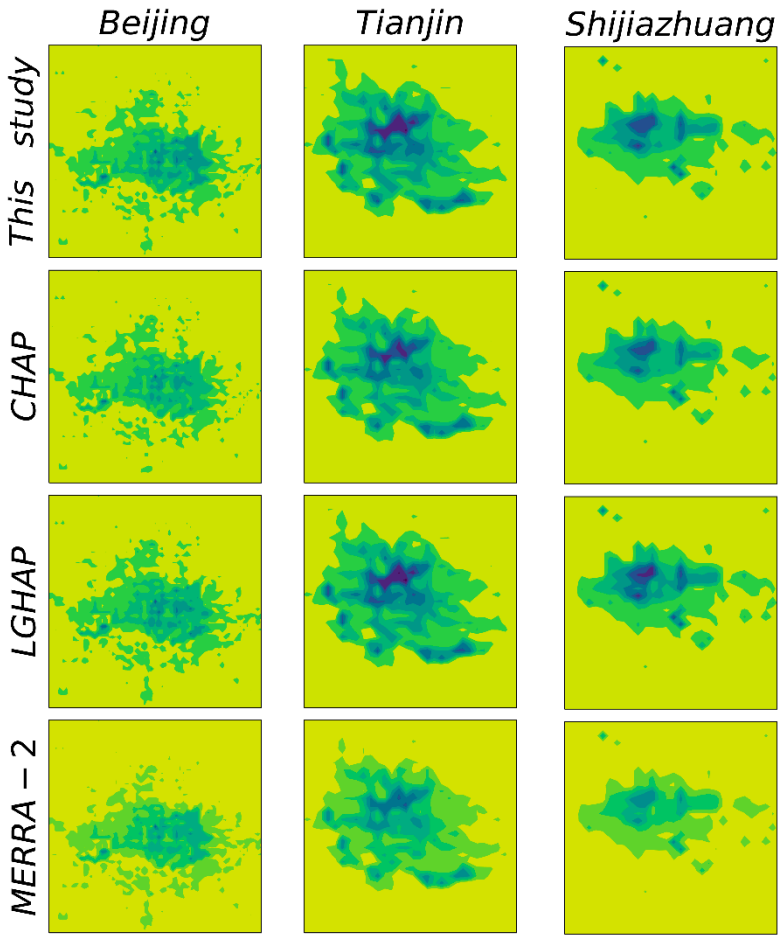

Figure S7. Distributions of calculated PWE of Beijing, Tianjin, and Shijiazhuang using generated data in this study, CHAP, LGHAP, and MERRA-2.

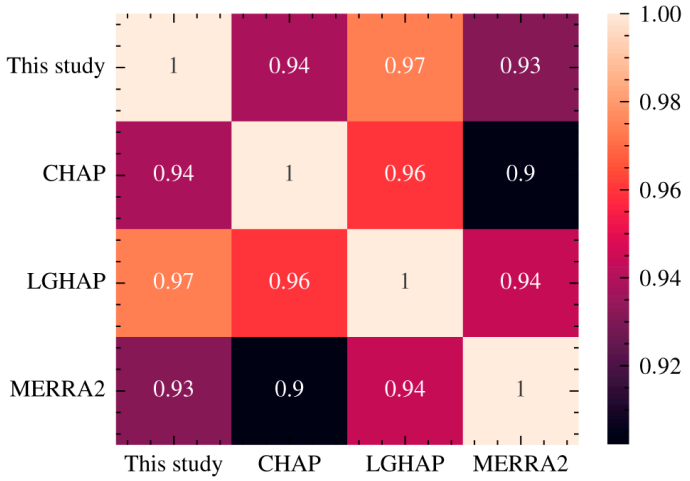

Figure S8. Heatmap of correlation coefficient (r) of PWE among all datasets.

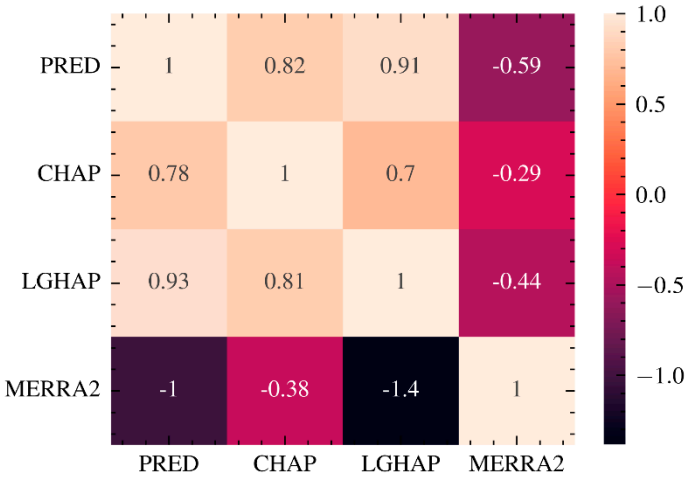

Figure S9. Heatmap of coefficient determination (R2) of PWE among all datasets.

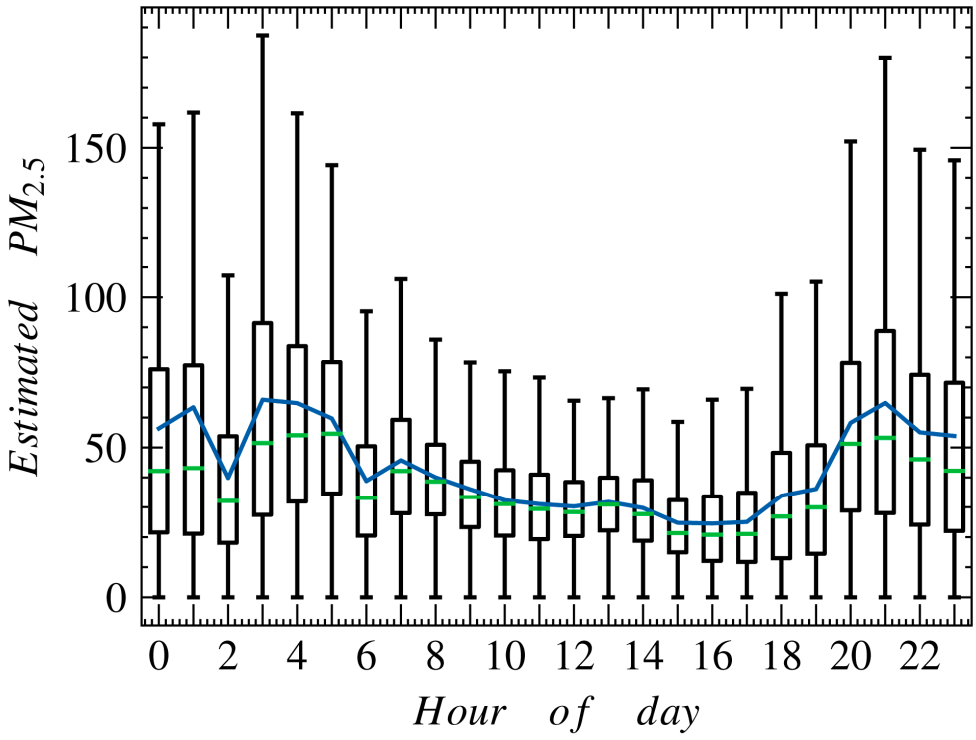

Figure S10. Boxplots of variations of estimated $PM_{2.5}$ of the BTH region during each hour. The blue solid line represents the mean value of each hour.

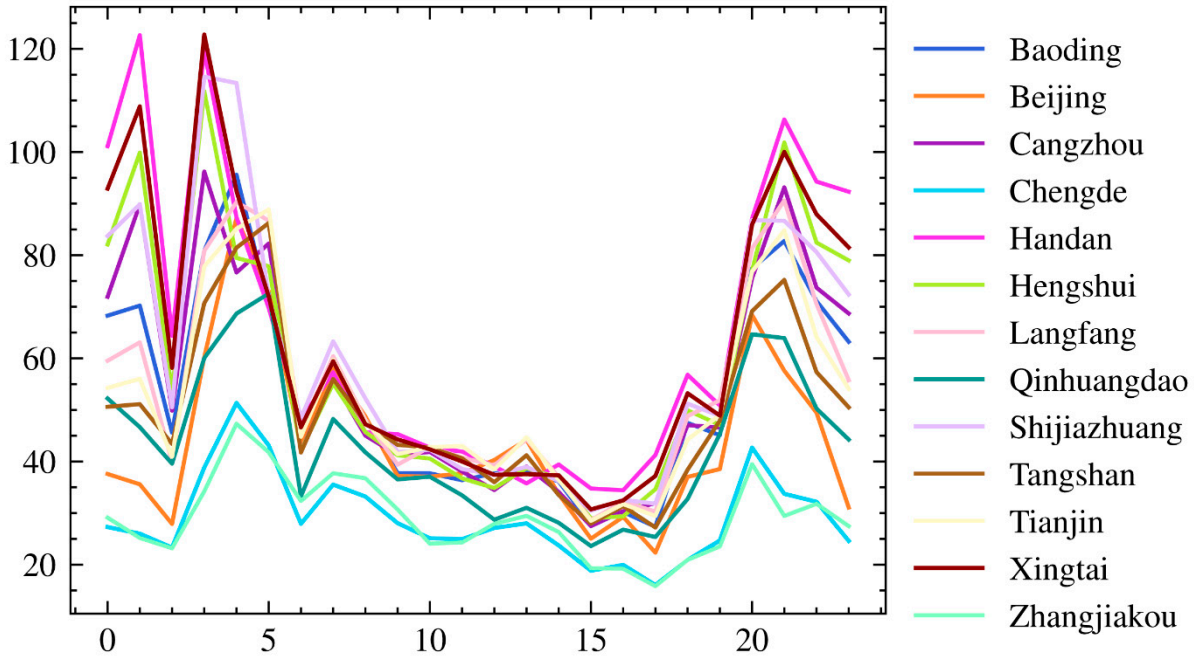

Figure S11. Mean values of each subregion of the BTH region for each hour.

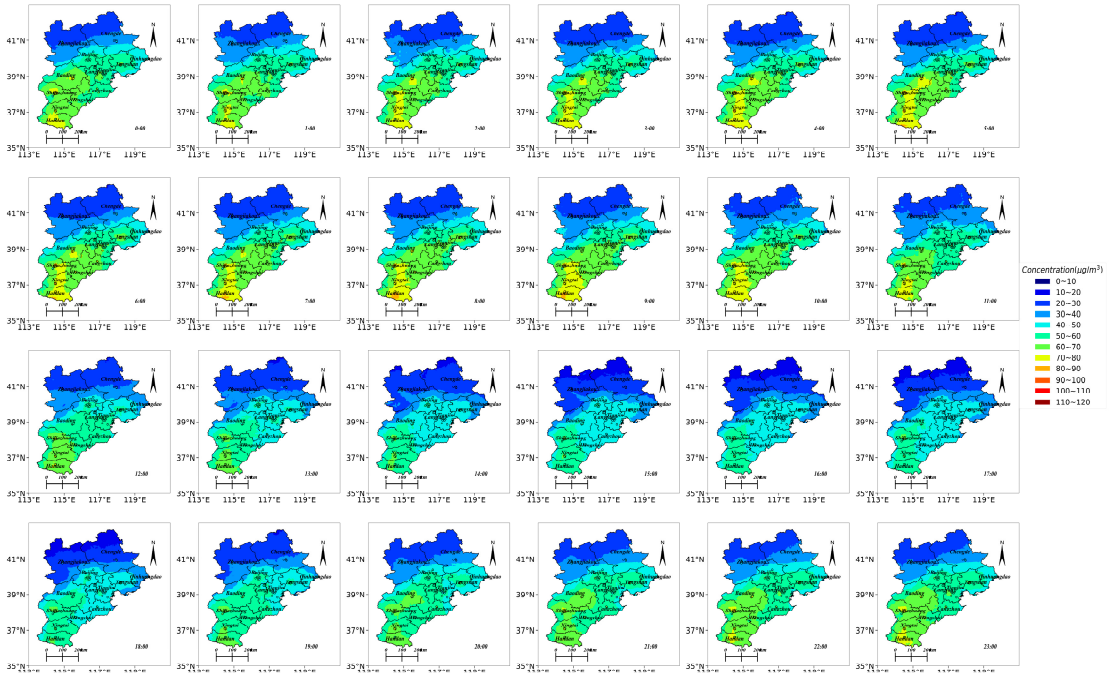

Figure S12. Distributions of estimated $PM_{2.5}$ in the BTH region over 24 h.

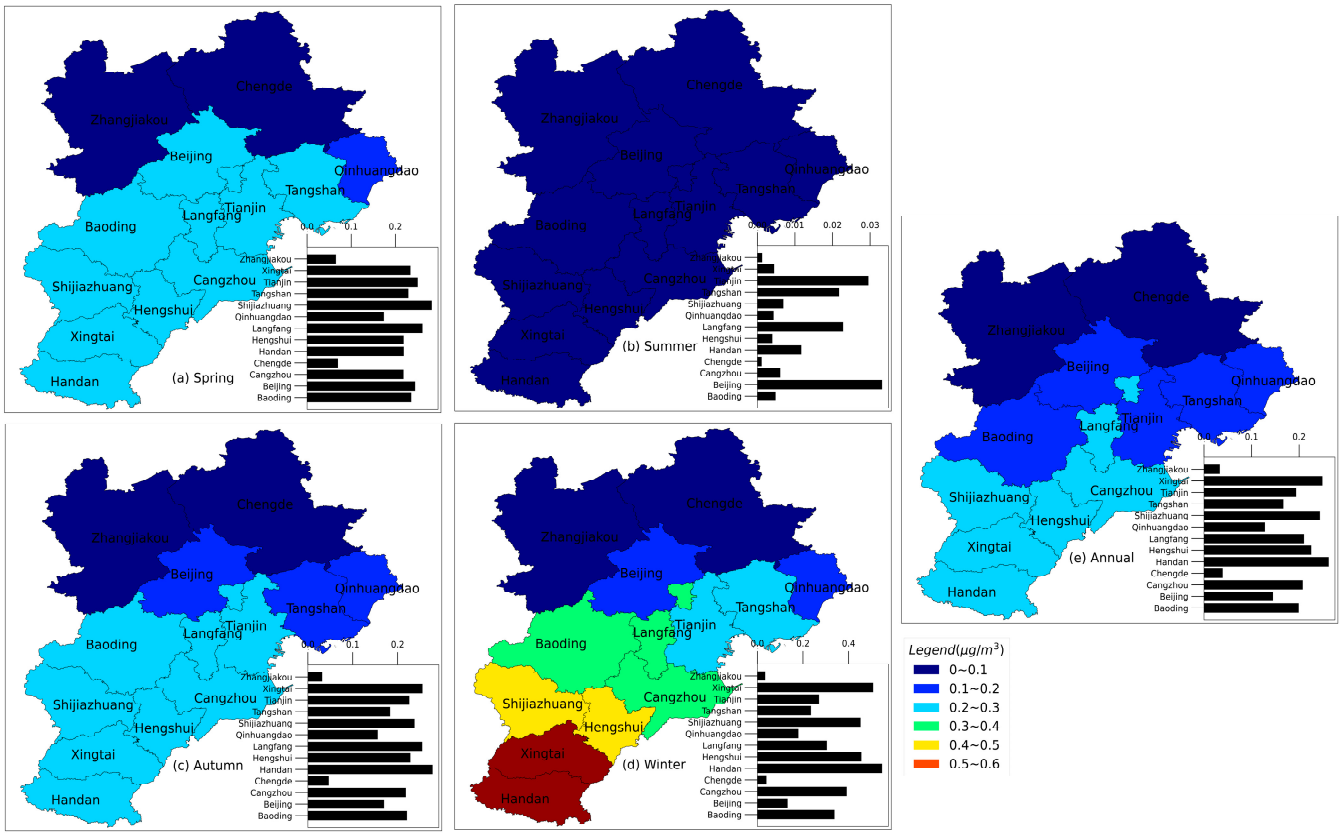

Figure S13. Distributions and bar plots of RER of each subregion in the BTH region: (a) Spring, (b) Summer, (c) Autumn, (d)Winter, and (e) Annual.

References

1. Chu, W.; Zhang, C.; Zhao, Y.; Li, R.; Wu, P. Spatiotemporally Continuous Reconstruction of Retrieved $PM_{2.5}$ Data Using an Autogeoi-Stacking Model in the Beijing-Tianjin-Hebei Region, China. *Remote Sensing* **2022**, *14*, 4432, doi:10.3390/rs14184432.
2. Kaul, A.; Maheshwary, S.; Pudi, V. AutoLearn — Automated Feature Generation and Selection. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM); IEEE: New Orleans, LA, November 2017; pp. 217–226.
3. Wu, D.; Liu, W.; Fang, B.; Chen, L.; Zang, Y.; Zhao, L.; Wang, S.; Wang, C.; Marcato, J.; Li, J.; et al. Automated Feature Engineering Improves Prediction of Protein–Protein Interactions. *Amino Acids* **2019**, *51*, 1187–1200, doi:10.1007/s00726-019-02756-9.
4. Zheng, Z.; Fiore, A.M.; Westervelt, D.M.; Milly, G.P.; Goldsmith, J.; Karambelas, A.; Curci, G.; Randles, C.A.; Paiva, A.R.; Wang, C.; et al. Automated Machine Learning to Evaluate the Information Content of Tropospheric Trace Gas Columns for Fine Particle Estimates Over India: A Modeling Testbed. *J Adv Model Earth Syst* **2023**, *15*, doi:10.1029/2022MS003099.
5. Džeroski, S.; Ženko, B. Is Combining Classifiers with Stacking Better than

Selecting the Best One? *Machine Learning* **2004**, *54*, 255–273, doi:10.1023/B:MACH.0000015881.36452.6e.

6. Bai, K.; Li, K.; Ma, M.; Li, K.; Li, Z.; Guo, J.; Chang, N.-B.; Tan, Z.; 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* **2022**, *14*, 907–927, doi:10.5194/essd-14-907-2022.
7. LandScan: A Global Population Database for Estimating Populations at Risk. In *Remotely-Sensed Cities*; Mesev, V., Ed.; CRC Press, 2003; pp. 301–314 ISBN 978-0-429-18116-0.
8. Abdul Shakor, A.S.; Pahrol, M.A.; Mazeli, M.I. Effects of Population Weighting on PM $_{10}$ Concentration Estimation. *Journal of Environmental and Public Health* **2020**, *2020*, 1–11, doi:10.1155/2020/1561823.
9. Aunan, K.; Ma, Q.; Lund, M.T.; Wang, S. Population-Weighted Exposure to $PM_{2.5}$ Pollution in China: An Integrated Approach. *Environment International* **2018**, *120*, 111–120, doi:10.1016/j.envint.2018.07.042.
10. Xiao, Q.; Wang, Y.; Chang, H.H.; Meng, X.; Geng, G.; Lyapustin, A.; Liu, Y. Full-Coverage High-Resolution Daily $PM_{2.5}$ Estimation Using MAIAC AOD in the Yangtze River Delta of China. *Remote Sensing of Environment* **2017**, *199*, 437–446, doi:10.1016/j.rse.2017.07.023.
11. Hua, Z.; Sun, W.; Yang, G.; Du, Q. A Full-Coverage Daily Average $PM_{2.5}$ Retrieval Method with Two-Stage IVW Fused MODIS C6 AOD and Two-Stage GAM Model. *Remote Sensing* **2019**, *11*, 1558, doi:10.3390/rs11131558.
12. Jing, Y.; Pan, L.; Sun, Y. Estimating PM $_{2.5}$ Concentrations in a Central Region of China Using a Three-Stage Model. *International Journal of Digital Earth* **2023**, *16*, 578–592, doi:10.1080/17538947.2023.2175499.
13. Xue, T.; Zheng, Y.; Tong, D.; Zheng, B.; Li, X.; Zhu, T.; Zhang, Q. Spatiotemporal Continuous Estimates of $PM_{2.5}$ Concentrations in China, 2000–2016: A Machine Learning Method with Inputs from Satellites, Chemical Transport Model, and Ground Observations. *Environment International* **2019**, *123*, 345–357, doi:10.1016/j.envint.2018.11.075.
14. Geng, G.; Xiao, Q.; Liu, S.; Liu, X.; Cheng, J.; Zheng, Y.; Xue, T.; Tong, D.; Zheng, B.; Peng, Y.; et al. Tracking Air Pollution in China: Near Real-Time PM $_{2.5}$ Retrievals from Multisource Data Fusion. *Environ. Sci. Technol.* **2021**, *55*, 12106–12115, doi:10.1021/acs.est.1c01863.
15. Liu, Y.; Li, C.; Liu, D.; Tang, Y.; Seyler, B.C.; Zhou, Z.; Hu, X.; Yang, F.; Zhan, Y. Deriving Hourly Full-Coverage $PM_{2.5}$ Concentrations across China's Sichuan Basin by Fusing Multisource Satellite Retrievals: A Machine-Learning Approach. *Atmospheric Environment* **2022**, *271*, 118930, doi:10.1016/j.atmosenv.2021.118930.