Proceeding of ICSPHS 2026 Symposium: Urban Industrial Innovation and Resilience-oriented Regional Transformation
DOI: 10.54254/2753-7048/2026.HZ30778

Machine Learning-based Downscaling of Sentinel-5P CO Over Beijing Using LightGBM, CatBoost and Ridge

Haoyi Wu

Stony Brook Institute, Anhui University, Hefei, China
r32314010@stu.ahu.edu.cn

Abstract. Urban health studies require near-surface carbon monoxide (CO) at high spatial resolution, but satellite products mainly provide total-column information. This study develops a machine learning-based downscaling framework to estimate daily mean near-surface CO over Beijing on a 3 km grid by combining Sentinel-5P/TROPOMI Level-2 CO, ERA5 reanalysis and ChinaHighCO ground observations. Grid-day samples (134,282 in 2023) are described by 50 predictors, including TROPOMI column statistics, dynamically aligned boundary-layer meteorology, and temporal indicators. Three regression models are compared—LightGBM, CatBoost and Ridge regression—under a rolling monthly cross-validation scheme in which each validation month is predicted using only past data. Model skill is assessed using out-of-fold coefficients of determination (R^2), root mean square error (RMSE), month-wise R^2 and spatial error patterns. The best configuration, a LightGBM model using a compact "core ERA5 + hS5P" predictor set, achieves a global R^2 of 0.56 and RMSE of 0.13 mg m^{-3}, with month-wise R^2 remaining between about 0.40 and 0.62. Adding the full ERA5 predictor set further increases global R^2 to about 0.66. Feature-importance analysis highlights the dominant role of TROPOMI columns, spatial coordinates, wind direction and seasonal indicators, demonstrating the potential of physically informed tree-ensemble models for urban CO downscaling.

Keywords: Carbon monoxide, Sentinel-5P, LightGBM, CatBoost, Ridge

1. Introduction

Accurate mapping of urban air pollution remains a core challenge for environmental health and air-quality management. Outdoor air pollution is one of the most important environmental health risks worldwide, and recent assessments attribute several million premature deaths each year to ambient air pollution exposure, with particularly heavy burdens in rapidly urbanizing regions of East and South-East Asia [1]. Carbon monoxide (CO) is not only a toxic gas harmful to human health, but also an important tracer of combustion activity and atmospheric transport. Resolving its spatial distribution at the city scale is therefore valuable for identifying emission hot spots, evaluating control policies, and informing public-health interventions.

Satellite remote sensing has greatly expanded the ability to observe atmospheric composition over large areas. Among current instruments with CO sensitivity, Sentinel-5P/TROPOMI provides global total-column CO measurements with broad coverage and relatively high spatial resolution,

DOI: 10.54254/2753-7048/2026.HZ30778

which is particularly useful in regions where ground observations are sparse [2]. However, these column-integrated products do not directly correspond to the near-surface concentrations that matter most for human exposure and city-level air-quality management. Additional processing is required to translate column information into ground-level fields that can be used in practice, for example through chemical-transport modeling or statistical and machine-learning approaches.

A large body of recent work has therefore used data-fusion and downscaling techniques to estimate surface pollutant fields from heterogeneous data sources. Many studies have focused on particulate matter ($PM_{2.5}$) and nitrogen dioxide (NO_2), combining ground monitoring, multiple satellite products, reanalysis, and chemical-transport model outputs with advanced machine-learning algorithms such as gradient-boosted trees and deep neural networks to produce seamless gridded products at kilometre-scale resolution [3-5]. In comparison, dedicated applications to CO are fewer. On the global and regional scale, several studies assimilate TROPOMI CO into chemistry–transport models, such as those underpinning the Copernicus Atmosphere Monitoring Service, to improve large-scale CO distributions and vertical profiles [6]. At the national scale, big-data fusion frameworks exemplified by ChinaHighAirPollutants and related ChinaHighCO products combine satellite, reanalysis and in-situ data to generate seamless grids of surface CO and other pollutants over China [7]. These developments demonstrate the potential of satellite-based CO monitoring, but they remain either model-centric at relatively coarse resolution or tailored to national-scale products.

Taken together, existing approaches have substantially advanced the monitoring of atmospheric pollutants, yet a gap remains between global or national products and the practical needs of city-level stakeholders. Large-scale products still have limited ability to resolve intra-urban gradients, while city-scale applications that specifically target CO, adopt strict time-series validation, systematically compare different families of machine-learning models, and provide a transparent, easily transferable pipeline are still relatively rare.

The present study addresses this gap by developing a modular machine-learning framework that downscales Sentinel-5P/TROPOMI Level-2 total-column CO, together with ERA5 meteorological reanalysis and routine near-surface CO measurements, to 3 km near-surface CO fields over a representative Chinese megacity. Gradient-boosted decision trees (LightGBM), categorical boosting (CatBoost), and ℓ_2-regularized linear models (Ridge regression) are trained under a rolling monthly time-series cross-validation design that avoids information leakage from future months and yields out-of-fold predictions suitable for honest evaluation. The framework is designed to rely only on routinely available inputs and simple configuration files so that it can be transferred to other cities with minimal changes. Overall, this work provides a city-scale CO downscaling pipeline and an empirical comparison of several widely used learning algorithms, aiming to narrow the scale gap between global Earth-observation products and local air-quality decision-making.

2. Method

2.1. Dataset preparation

This study focuses on the Beijing metropolitan area and constructs a gridded dataset of daily mean CO at approximately 3 km spatial resolution. The analysis domain is defined by an urban-area polygon and discretised into a regular latitude–longitude grid in the WGS84 coordinate system. Each record in the dataset corresponds to one grid cell on one day (a "grid–day"). Three data sources are combined: Sentinel-5P/TROPOMI Level-2 CO total-column retrievals, ERA5 meteorological reanalysis, and ground-level CO from the ChinaHighCO (CHCO) product derived from the national monitoring network. Daily CHCO values are aggregated onto the 3 km grid and used as the

DOI: 10.54254/2753-7048/2026.HZ30778

regression target, expressed in milligrams per cubic metre $\mathrm{mg\,m^{-3}}$. After merging all data sources and discarding records with missing or non-finite labels, the final dataset contains 134,282 grid–day samples and 50 predictor variables.

Predictors are grouped into three categories. First, satellite-based predictors are derived from TROPOMI Level-2 CO retrievals after applying the recommended quality-assurance flag and a basic cloud filter. Remaining pixels are reprojected from the swath geometry onto the 3 km grid using bilinear interpolation in latitude and longitude. For each grid–day this paper computes simple statistics of the CO total column, including the mean, standard deviation, number of valid pixels, and a three-day running mean, which together characterise the column burden and short-term variability.

Second, meteorological predictors are taken from the ERA5 reanalysis. This paper focuses on a compact subset of near-surface and boundary-layer variables that are physically related to transport and dilution of pollutants, including 10 m wind components, 2 m air and dew-point temperature, surface pressure, boundary-layer height, and total column water vapour. For the core configuration analysed in this paper (bj_core_hS5P), these fields are temporally aligned with the local Sentinel-5P overpass time over Beijing. From the basic variables this paper derives diagnostics such as wind speed, wind direction encoded by sine and cosine terms, the inverse and logarithm of boundary-layer height, and a temperature–dew-point difference as a proxy for near-surface humidity.

Third, temporal and spatial descriptors are added to encode large-scale structure in the data. These include year and month, day-of-year, day-of-week, their sine and cosine to represent cyclical behaviour, and the longitude and latitude of each grid cell. All predictors are taken directly from the pre-computed feature table used in the experiment code. Apart from dropping rows with missing values, no additional normalisation or rescaling is applied before training, because for monotonic transformations of individual features, the tree-based models used in the project are invariant. For the Ridge model, standardisation of predictors is handled within the scikit-learn pipeline described in the configuration file.

Because neighbouring days are strongly autocorrelated, a random train–test split would yield overly optimistic estimates of model skill. Instead, this paper adopts a rolling monthly time-series cross-validation scheme over January–December 2023. In each fold, one calendar month between May and December is held out for validation, while all preceding months serve as training data. This design mimics a forecasting setting in which only past information is available when predicting a future month and produces out-of-fold predictions for every grid–day in the eight validation months. All performance metrics reported in this study are computed from these out-of-fold predictions.

2.2. Machine learning models

The downscaling task is formulated as supervised regression from the 50-dimensional predictor vector to the daily mean near-surface CO value at each grid cell. This paper considers three model families that span a range of functional complexity: gradient-boosted decision trees implemented with LightGBM [8], gradient boosting with symmetric trees implemented with CatBoost [9], and an l_2 -regularised linear regression model: Ridge [10]. Each model is trained under several input configurations that differ in how satellite and meteorological information is combined, but always using the same feature table and cross-validation splits described above. Implementation relies on Python, the scikit-learn ecosystem, and the official LightGBM and CatBoost libraries.

All models are trained to minimise squared error between predicted and observed CO concentrations. Within each fold of the rolling cross-validation, the model is fitted on the training months and then applied to the held-out validation month to produce out-of-fold (OOF) predictions.

From these OOF predictions this study computes the coefficient of determination (R^2) and root mean square error (RMSE) pooled across all validation months, as well as month-wise R^2 values that quantify temporal robustness.

Hyperparameters for each model family are chosen conservatively based on a small number of exploratory runs and then fixed for all folds and configurations. Because LightGBM is used as the main reference model, its key settings are reported explicitly here: the learning rate is set to 0.05 with up to 6000 boosting iterations, each tree is allowed up to 95 leaves (num_leaves = 95), and at least 40 samples are required per leaf (min_data_in_leaf = 40). To limit overfitting, this paper uses feature and bagging fractions of 0.9 (feature_fraction = 0.9, bagging_fraction = 0.9, bagging_freq = 1) and an l_2 regularisation strength of lambda_l2 = 0.5. CatBoost and Ridge are configured using similarly conservative hyperparameters following common practice for regression, and no per-month retuning is performed; together with the fixed cross-validation design, this ensures that model comparisons reflect differences in functional form and input configuration rather than extensive hyperparameter optimisation.

2.2.1. LightGBM

LightGBM is a gradient-boosted decision tree algorithm designed for efficiency on large tabular datasets [8]. The model is grown incrementally by appending small regression trees, each of which is trained to explain the remaining errors of all previously added trees. In contrast to traditional level-wise boosting algorithms, LightGBM grows trees leaf-wise: at each step it selects the split on a single leaf that yields the largest reduction in loss. When combined with appropriate regularisation, this strategy can achieve high accuracy with relatively shallow trees and a modest number of boosting iterations. LightGBM also uses histogram-based binning of numeric features and histogram differences to accelerate split finding, which is advantageous for the high-resolution gridded dataset used here.

In this study, LightGBM is configured for regression with a squared-error objective and a shrinkage learning rate. Tree depth, number of leaves and minimum samples per leaf are restricted to limit model complexity, and subsampling of rows and features is used to reduce correlation between individual trees and mitigate overfitting. An l_2 regularisation term on leaf weights further stabilises the ensemble. The same hyperparameters are used for all input configurations so that differences in performance can be attributed primarily to the predictor sets rather than to changes in model capacity.

2.2.2. CatBoost

CatBoost is another gradient-boosted tree method, originally developed to handle categorical predictors through target-based encodings and to mitigate prediction shift. It employs symmetric (oblivious) trees, where the same split is applied at each level of the tree, leading to highly regular structures that are efficient to store and evaluate. CatBoost also uses an ordered boosting procedure that simulates online learning to reduce target leakage when computing target statistics. In the present application, all predictors are continuous, so CatBoost is used primarily as an alternative tree ensemble with a different tree topology and optimisation scheme compared with LightGBM.

This paper applies the regression version of CatBoost with squared-error loss, moderate tree depth and learning rate, and a fixed upper limit on the number of boosting iterations. The remaining hyperparameters are kept close to their library defaults. By comparing CatBoost and LightGBM

DOI: 10.54254/2753-7048/2026.HZ30778

under identical feature sets and cross-validation splits, it is possible to assess how sensitive the CO downscaling task is to the specific implementation of gradient-boosted trees.

2.2.3. Ridge regression

Ridge regression provides a linear baseline against which the benefits of non-linear tree ensembles can be evaluated. The model assumes that near-surface CO is an approximately linear function of the predictors and introduces an l_2 penalty on the regression coefficients to control multicollinearity and reduce variance. The penalty strength determines the degree of shrinkage applied to the coefficients: larger values favour smaller, smoother coefficients at the cost of increased bias. Ridge regression is implemented using the scikit-learn library and is fitted separately for each input configuration using the same training and validation splits as the tree-based models.

All 50 predictors from the feature table are included as linear terms without manual feature selection, so that the comparison with LightGBM and CatBoost reflects genuine differences in functional form rather than differences in input information. Although Ridge regression cannot capture non-linear interactions between satellite-derived columns, meteorology and time, it offers transparent coefficients that are easy to interpret and serves as a conservative reference model for urban CO downscaling.

3. Results and discussion

3.1. Overall predictive skill of the core configuration

Table 1 summarises the cross-validated performance of the three model families under the core ERA5 + hS5P configuration (bj_core_hS5P). When out-of-fold predictions from all eight validation months are pooled together, the LightGBM model achieves a global coefficient of determination of $R^2 = 0.562$ with a root mean square error of RMSE = 0.127 mg m^{-3} (Figure 1). CatBoost attains lower but still positive skill (global $R^2 \approx 0.44$, RMSE ≈ 0.11 mg m^{-3}), whereas the linear Ridge baseline yields only marginal explanatory power with a global R^2 close to zero despite a comparable RMSE. These results already suggest that non-linear tree ensembles extract substantially more information from the same 50-dimensional predictor set than a regularised linear model.

Table 1. Overall out-of-fold performance for the core ERA5 + hS5P configuration (bj_core_hS5P)

Model	Global R^2	RMSE (mg m^{-3})
LightGBM	0.562	0.127
CatBoost	0.439	0.112
Ridge	0.026	0.106

DOI: 10.54254/2753-7048/2026.HZ30778

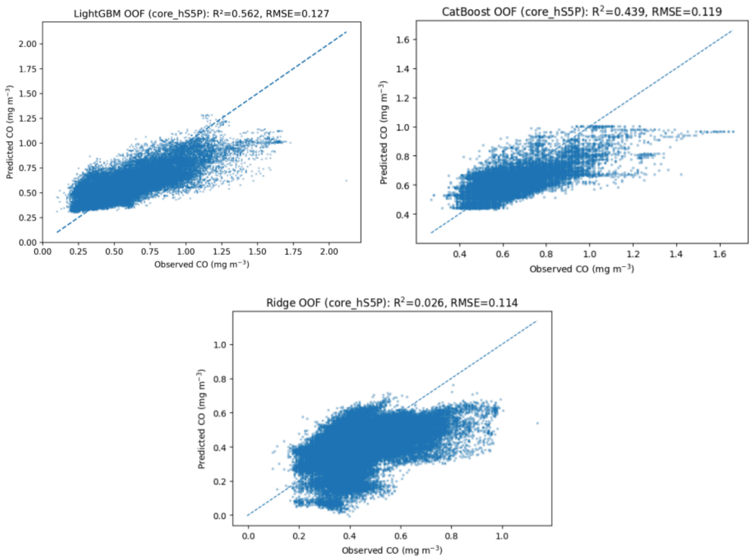

Figure 1. Scatter plots of observed versus predicted near-surface CO for the three models under the bj_core_hS5P configuration. The panels show LightGBM (top left; R^2 = 0.562, RMSE = 0.127 mg m^{-3}), CatBoost (top right; R^2 = 0.439, RMSE = 0.119mg m^{-3}), and Ridge regression (bottom; R^2 = 0.026, RMSE = 0.114mg m^{-3}) (picture credit: original)

The multi-panel scatter plot between observed and predicted CO in Figure 1 provides a more detailed view of the error structure across the three models. For LightGBM, the point cloud follows the 1:1 line reasonably well up to about 1.0–1.2 mg m^{-3}, confirming that the model captures most of the day-to-day variability in near-surface CO. At higher concentrations, the cloud bends slightly below the 1:1 line, indicating a tendency to underestimate the most polluted days, while at very low concentrations the model slightly overestimates CO. CatBoost shows a somewhat tighter cluster at intermediate concentrations but a clearer saturation at high CO values, which is consistent with its lower overall R^2 and suggests that boosted symmetric trees smooth out peak events more aggressively. In contrast, the Ridge regression panel exhibits a much more diffuse cloud with a strong bias: predictions are compressed into a narrow range around 0.4–0.7 mg m^{-3} and deviate markedly from the 1:1 line at both ends of the distribution. These patterns are typical when linear models and tree ensembles are applied to air-quality downscaling: all three approaches tend to pull extreme values towards the mean, but the non-linear ensembles retain substantially more structure in the observed variability than the linear baseline.

3.2. Month-wise robustness and comparison of model families

Figure 2 shows the month-wise R^2 from May to December 2023 for the three model families, with the y-axis truncated to the range (-1,-0.8) for visual clarity. LightGBM exhibits consistently positive skill in all months, with R^2 values ranging from about 0.40 to 0.62. Performance peaks in June and October ($R^2 \approx 0.62$), while the weakest months (July and August) still retain moderate skill ($R^2 \approx$ 0.40–0.46). This behaviour indicates that the boosted-tree model generalises well across seasons and that the core predictor set captures most of the drivers of month-to-month CO variability.

DOI: 10.54254/2753-7048/2026.HZ30778

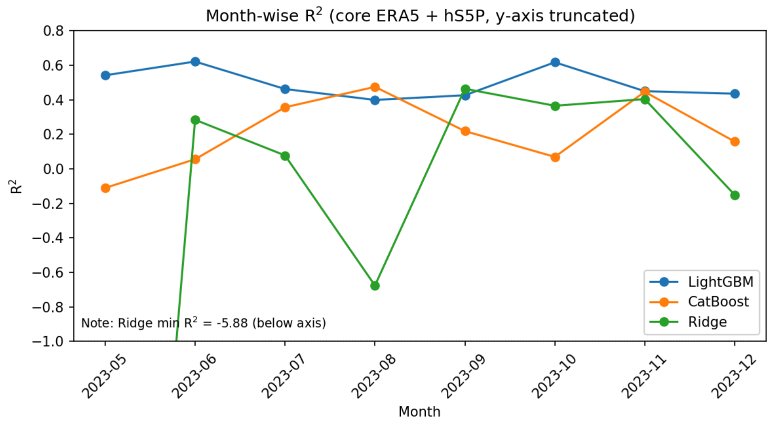

Figure 2. Month-wise coefficient of determination (R^2) for the three models under the core ERA5 + hS5P setting (bj_core_hS5P). The y-axis is truncated to the range (-1,-0.8) for clarity (picture credit: original)

CatBoost shows broadly similar behaviour but with larger month-to-month fluctuations. Its R^2 is slightly negative in May, improves to 0.35–0.48 in mid-summer (July–September), and drops again to around 0.07–0.16 in October and December. The comparison between LightGBM and CatBoost suggests that, for this particular task and dataset, the leaf-wise growth strategy of LightGBM leads to somewhat more stable performance than the symmetric-tree design of CatBoost, even when both models use conservative hyperparameters.

The Ridge regression baseline is markedly less robust. In May and August its month-wise R^2 falls well below zero (approximately −5.88 and −0.68, respectively; the May value lies outside the plotted range), indicating that a simple linear function of the predictors can perform substantially worse than a constant mean predictor in some months. In other months, Ridge recovers to modestly positive R^2 values, up to about 0.45 in September–November, but it never matches the boosted-tree models. The combination of Figure 1 and Table 1 therefore highlights that non-linear tree ensembles are essential for achieving both high average accuracy and acceptable temporal robustness in this CO downscaling problem.

3.3. Spatial distribution of errors

The spatial distribution of LightGBM prediction errors for December 2023 is shown in Figure 3. Each dot represents one grid cell, coloured by the difference between predicted and observed daily mean CO ($\hat{y}$ - y). Errors mostly fall within ±0.3 mg m^{-3}, and there is no obvious city-wide bias pattern: both positive and negative errors are scattered throughout the domain. Some small clusters of underestimation (purple colours) and overestimation (yellow colours) can be seen, for example along certain north–south corridors, but their magnitude remains moderate. This suggests that the model does not systematically over- or under-predict CO in specific parts of the city, and that remaining errors are driven more by local processes or representativeness differences between the gridded predictors and the point-like ground observations.

DOI: 10.54254/2753-7048/2026.HZ30778

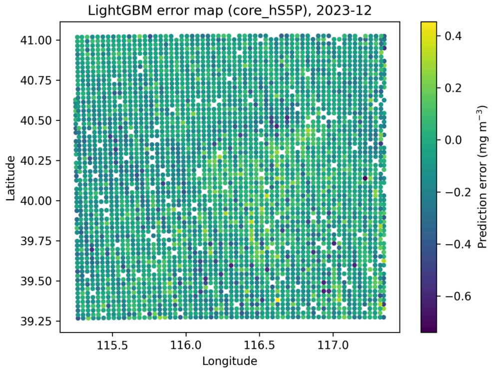

Figure 3. Spatial distribution of LightGBM prediction errors (ŷ - y) for December 2023 under the core ERA5 + hS5P configuration (picture credit: original)

The lack of a strong large-scale bias is encouraging for applications that rely on spatial contrasts, such as exposure assessments across urban neighbourhoods. However, the presence of localised error clusters also indicates that additional information, such as higher-resolution emission inventories or land-use descriptors, could further improve the spatial fidelity of the downscaled fields, especially near strong emission hotspots and along major transport corridors.

3.4. Dominant predictors and physical interpretation

Figure 4 presents the top 15 predictors ranked by LightGBM split importance. The most influential features are the latitude and longitude of the grid cell, followed by the TROPOMI total-column CO mean and its three-day running mean. This confirms that the model relies heavily on large-scale spatial gradients and on the satellite-derived column burden to reconstruct near-surface CO. Wind direction at the analysis time, represented by sine and cosine terms of the 10 m wind components, also plays a prominent role, as do temporal descriptors such as day-of-year and day-of-week and their harmonic transforms.

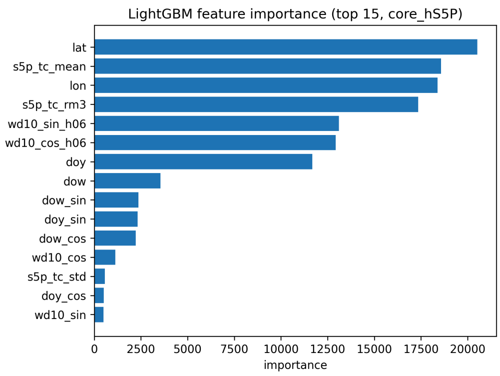

Figure 4. Top 15 predictor variables ranked by LightGBM split importance under the core ERA5 + hS5P configuration (picture credit: original)

The relatively high importance of temporal indicators suggests that recurring seasonal and weekly patterns in emissions and meteorology contribute substantially to predictable CO variability. In contrast, many of the remaining ERA5 variables appear with much lower importance, indicating that, once wind direction and a small set of core boundary-layer variables are included, adding more

DOI: 10.54254/2753-7048/2026.HZ30778

meteorological predictors yields diminishing returns. This finding is consistent with the design of the bj_core_hS5P configuration, which intentionally focuses on a compact, physically motivated predictor subset.

Taken together, the results from Figures 1–4 show that combining TROPOMI CO columns with a small number of dynamically aligned meteorological and temporal predictors enables LightGBM to achieve cross-validated R^2 values around 0.5 and to provide spatially detailed yet temporally robust estimates of near-surface CO over Beijing. The comparison with CatBoost and Ridge underlines both the advantages of flexible tree-based ensembles and the importance of carefully designed, physically informed input features.

4. Conclusion

This work sets out to bridge the gap between satellite-derived CO columns and the near-surface concentrations needed for urban exposure assessment. By fusing Sentinel-5P/TROPOMI, ERA5 and ChinaHighCO on a 3 km grid and training LightGBM, CatBoost and Ridge models with rolling monthly cross-validation, this paper showed that non-linear tree ensembles can recover a substantial fraction of daily CO variability. The core LightGBM configuration reached cross-validated R^2 values around 0.5–0.6 and produced spatially detailed yet temporally robust fields, while the linear Ridge baseline performed poorly, underscoring the importance of flexible model structures and physically motivated predictors. The study is limited to one city and one year, and does not explicitly quantify predictive uncertainty. Future work will extend the framework to multiple cities and years, incorporate land-use and emission information, and explore probabilistic or spatio-temporal models for uncertainty-aware air-quality applications.

References

[1] World Health Organization Regional Office for South-East Asia. (2016). Air pollution. Retrieved from the WHO South-East Asia Regional Office website.

[2] Borsdorff, T., aan de Brugh, J., Pandey, S., Hasekamp, O., Aben, I., & Landgraf, J. (2018). Measuring carbon monoxide with TROPOMI: First results. Geophysical Research Letters, 45(6), 2826–2832. https: //doi.org/10.1002/2018GL077045 AGU Publications+1

[3] Schneider, R., Vicedo-Cabrera, A. M., Sera, F., Masselot, P., Stafoggia, M., de Hoogh, K., Kloog, I., Reis, S., Vieno, M., & Gasparrini, A. (2020). A satellite-based spatio-temporal machine learning model to reconstruct daily $PM_{2.5}$ concentrations across Great Britain. Remote Sensing, 12(22), 3803. https: //doi.org/10.3390/rs12223803 MDPI

[4] Li, X., Liu, L., Chen, X., Li, C., & Xu, Y. (2019). Predicting ground-level $PM_{2.5}$ concentrations in the Beijing–Tianjin–Hebei region: A hybrid remote sensing and machine learning approach. Environmental Pollution, 249, 735–749. https: //doi.org/10.1016/j.envpol.2019.03.055

[5] Li, L. (2020). A robust deep learning approach for spatiotemporal estimation of satellite AOD and $PM_{2.5}$. Remote Sensing, 12(2), 264. https: //doi.org/10.3390/rs12020264 MDPI+1

[6] Inness, A., Chabrillat, S., Huijnen, V., Flemming, J., & Peuch, V.-H. (2022). Assimilation of S5P/TROPOMI carbon monoxide data with the global CAMS near-real-time system. Atmospheric Chemistry and Physics, 22, 14, 355–14, 381. https: //doi.org/10.5194/acp-22-14355-2022 acp.copernicus.org+1

[7] Wei, J., Li, Z., Wang, J., Li, C., Zhang, Y., Min, M., … Guo, J. (2023). Ground-level gaseous pollutants (NO_2, SO_2, and CO) in China: Daily seamless mapping and spatiotemporal variations. Atmospheric Chemistry and Physics, 23, 1511–1532. https: //doi.org/10.5194/acp-23-1511-2023

[8] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems, 30, 3146-3154.

[9] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. (2018) CatBoost: Unbiased Boosting with Categorical Features. Advances in Neural Information Processing Systems, 31, 6638-6648.

[10] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12, 55-67.