Estimating Ground-Level Carbon Monoxide Concentrations Using Machine Learning Techniques: The Metropolitan City of Milan Case Study

Zhongyou Liang[1], Jesus Rodrigo Cedeno Jimenez[1], Vasil Yordanov[1], Maria A. Brovelli[1]

[1] Dept. of Civil and Environmental Engineering, Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133, Milan, Italy
zhongyou.liang@mail.polimi.it, (jesusrodrigo.cedeno, vasil.yordanov, maria.brovelli)@polimi.it

Keywords: Carbon monoxide, Air quality, Sentinel-5P, Machine learning, Deep learning, Data fusion.

Abstract

This work presents a structured data-driven framework for estimating ground-level carbon monoxide (CO) concentrations in the Metropolitan City of Milan (MCM) by integrating Sentinel-5P satellite observations, Copernicus Atmosphere Monitoring Service reanalysis data, and ERA5 meteorological variables with advanced machine learning techniques. The methodology employs unified data preprocessing, systematic feature engineering (e.g., boundary layer height-adjusted CO, lagged meteorological variables), Bayesian optimization for hyperparameter tuning, SHAP-based feature selection, model ensembling, and robust statistical validation. Eight regression models, including a custom Dense Attention Network (DAN), were evaluated across multiple temporal aggregation windows (4–24 hours before 15:00 GMT+1) to identify optimal configurations for CO estimation. Using data from January 2019 to November 2024, the framework identified the 21:00–15:00 GMT+1 window as most effective for capturing atmospheric dynamics such as nighttime accumulation, morning emission peaks, and daytime dilution. The DAN achieved the best performance, with a mean normalized root mean squared error of 0.4879 ± 0.0252 on the test set, outperforming ensemble and traditional regression models, offering a scalable, interpretable, and cost-effective approach to urban CO monitoring in data-scarce environments with potential adaptation to other pollutants and regions.

1. Introduction

Carbon monoxide (CO) is emitted primarily through incomplete combustion of fossil fuels and biomass and its presence in the atmosphere has significant impacts both on public health, causing respiratory problems (World Health Organization, 2024), and on climate, contributing to the deterioration of air quality and climate forcing (von Schneidemesser et al., 2015). Despite the critical need to monitor ground-level CO for environmental and health management, global surface-level monitoring remains highly uneven. As seen in Figure 1, out of the 33,984 CO world air quality monitoring stations, 32,254 stations (approximately 95%) are in Asia, Europe, or North America, while only 996 (approximately 3%) of these stations are located in low or lower middle-income countries (LMIC) (Smith et al., 2025), resulting in data scarcity that limits scientific understanding and policy formulation in underrepresented regions.

In response to this gap, we present a modeling framework that integrates satellite-based observations, modeled reanalysis products, and ground-level measurements to estimate surface-level CO concentrations. These datasets have been widely used to capture non-linear dependencies and spatiotemporal variability in atmospheric pollution modeling (Shetty et al., 2024, Fania et al., 2024, Chen et al., 2024, Chen et al., 2025). We implemented Machine Learning (ML) models aimed at overcoming the limitations of satellite measurements and sparse ground-based monitoring, offering a scalable and cost-effective approach to surface-level CO estimation in data-scarce regions.

2. Data and Preprocessing

The area of interest of this study is the Metropolitan City of Milan (MCM) from January 2019 to November 2024. Ground-level measurements are to be estimated as a single daily average at the overpass time of the satellite Sentinel-5P (from 11:00 to 15:00 UTC+1). Results are validated directly with measurements from the ground monitoring network from Lombardy ARPA (Regional Environment Protection Agency) (`https://www.dati.lombardia.it/stories/s/auv9-c2sj`).

2.1 Data Sources

2.1.1 Sentinel-5P data In this study, we used the Level 3 Sentinel-5P CO data (s5p_CO), which is already preprocessed and ingested by Google in the Earth Engine Data Catalogue. To ensure data quality, Google Earth Engine (GEE) applies predefined filtering thresholds based on the quality assurance (QA) values provided in the original Level 2 products. For CO, pixels with QA values below 50% are discarded. The Level 3 grid uses a pixel size smaller than the native one, resulting in a spatial resolution of approximately 1.1 km × 1.1 km (Google Earth Engine, 2025).

2.1.2 CAMS data This study employs the European Centre for Medium-Range Weather Forecasts Atmospheric Composition Reanalysis 4 (EAC4), a product of Copernicus Atmosphere Monitoring Service (CAMS). EAC4 is the fourth-generation global reanalysis of atmospheric composition, produced by combining model simulations with a vast array of satellite and in situ observations using a data assimilation system based on the Integrated Forecasting System (IFS). EAC4 spans from 2003 onward and provides global estimates of atmospheric variables, including surface-level CO, with a temporal resolution of 3 hours and a horizontal resolution of approximately 0.75° × 0.75° (Inness et al., 2019).

We downloaded the CO (cams_CO) and relative humidity (r) data from CAMS Atmosphere Data Store (`https://ads.atmosphere.copernicus.eu/datasets/`

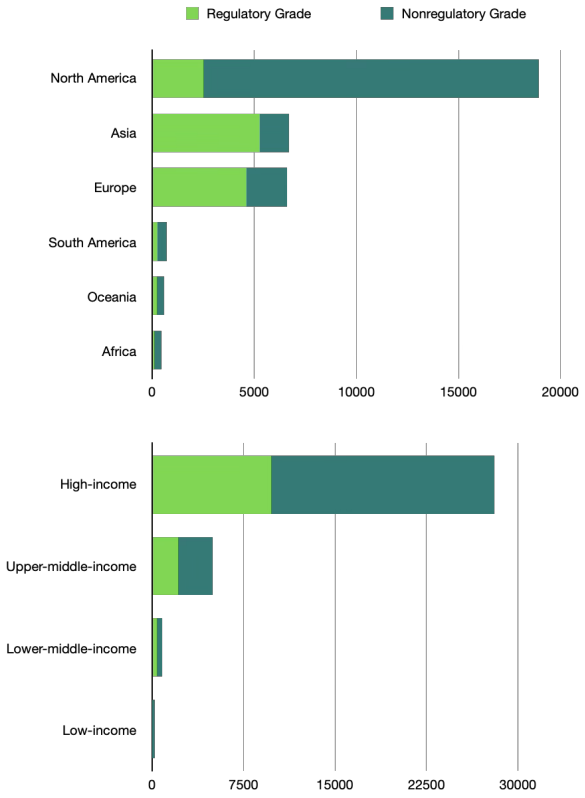

Figure 1. Worldwide distribution of on-the-ground monitoring stations reporting in April 2024. Regulatory grade sensors in light green and non-regulatory grade monitors in dark green. Source: (Smith et al., 2025), reproduced under the terms of the Creative Commons Attribution License (CC BY 4.0).

`cams-global-reanalysis-eac4?tab=download`). According to the instructions of the data store, to obtain surface values for CO, we select model level 60, which is the level of the Earth's surface. For relative humidity, we select pressure level 1000 hPa, which is close to standard atmospheric pressure.

CAMS CO assimilates surface-level CO measurements, which helps to address the mismatch between CO total column densities of Sentinel-5P and ground truth CO values, and then to enhance the accuracy and reliability of our models. Thus, we included it as a feature. As for relative humidity, it affects the hygroscopic growth of particles, influencing their deposition rates and radiative properties. High humidity may also enhance secondary aerosol formation, indirectly affecting pollutant levels (Niyogi and Raman, 2001). Including relative humidity as a feature helps our models to learn the physical and chemical relationships that affect CO dispersion.

2.1.3 ERA5 data ERA5, ECMWF's fifth-generation global atmospheric reanalysis, provides hourly estimates of atmospheric, ocean-wave, and land-surface variables from 1940 to the present, updated daily with a 5-day latency. It combines model data with global observations using data assimilation to produce a physically consistent dataset. The data is provided on a regular 0.25° latitude-longitude grid (0.5° for ocean waves) and includes both single-level and pressure-level products in hourly and monthly-mean formats. To quantify uncertainty, ERA5 incorporates a 10-member ensemble (sampled at 3-hourly intervals, with pre-computed mean and spread), reflecting the evolving observing system's information content. This regridded subset offers readily accessible data suitable for most common climate and weather applications (Hersbach et al., 2025).

From Copernicus Climate Change Service, Climate Data Store (`https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=download`), we collected boundary layer height (blh), 10m u-component of wind ($u10$), 10m v-component of wind ($v10$), surface net solar radiation (ssr), surface net thermal radiation (str), 2m temperature (t2m), surface pressure (sp), total precipitation (tp), leaf area index, high vegetation (lai_hv) and leaf area index, low vegetation (lai_lv) for this study. These variables contribute uniquely to atmospheric processes such as mixing, transport, and chemical transformation. By integrating these factors, our model is better equipped to learn the underlying physical and chemical relationships that affect CO dispersion, thereby enhancing the estimation of ground-level CO concentrations. The following provides a detailed account of how each variable influences these processes and contributes to the overall dynamics of pollutant behavior in the atmosphere (Barlow, 2009, McNider and Pour-Biazar, 2020, Li et al., 2017, Yang et al., 2016, Sorbjan, 2003, Arya, 2001, Pleim and Ran, 2011).

2.1.4 ARPA data In-situ CO measurements were obtained from ARPA Lombardy (located in the North of Italy) monitoring stations, which are publicly accessible via ARPA's online portal (`https://dati.lombardia.it/`), offering validated hourly to annual datasets, interactive maps, and downloadable reports. These data are fundamental for atmospheric modeling, emission inventories, and policy assessment.

2.2 Data Harmonization

2.2.1 Temporal Harmonization Given the varying temporal resolutions of our datasets, we need to perform temporal alignment. Considering the daily overpass time of the Sentinel-5P satellite over the MCM (12:00 to 15:00 GMT+1) (Cedeno Jimenez and Brovelli, 2023, Veefkind et al., 2012) and the time required for CO to be transported from the surface to the atmosphere and detected by the TROPOMI instrument on board Sentinel-5P, we first conducted the study in the time window from 11:00 to 15:00 GMT+1.

Additionally, we selected several distinct periods prior to 15:00 GMT+1 (i.e., the preceding 6, 12, 18, and 24 hours) to explore how the results vary with the temporal aggregation window. This approach allows us to assess the sensitivity of the relationship between surface-level variables and satellite-detected CO concentrations to different timescales of atmospheric processes, such as pollutant accumulation, transport, and vertical mixing. By comparing results across these timeframes, we aim to identify the most representative temporal window for capturing the surface-to-atmosphere dynamics relevant to the Sentinel-5P overpass measurements. To enable this comparison across different time windows, it is necessary to perform temporal harmonization of the data according to the specified time periods.

After initial loading and inspection, we convert all datasets to *pandas* DataFrames. For Sentinel-5P data, no time alignment is required as it contains only one measurement per day in the MCM. For the other datasets (CAMS, ERA5 and ARPA datasets), we performed temporal filtering, selecting data within a specified time window. Within each window, we computed the

mean of the values to create a consistent daily measurements. This approach ensured coherent temporal coverage across all datasets and enabled accurate integration for downstream analysis.

2.2.2 Spatial Harmonization To ensure all datasets share the same spatial resolution, we first reprojected all data to a common coordinate reference system, UTM Zone 32N, using the *PyProj* library (`https://pypi.org/project/pyproj/`). This projection was selected because it provides consistent distance measurements in meters, which is essential for accurate spatial interpolation and grid alignment, particularly in northern Italy.

To determine a suitable spatial resolution we used the minimum distance between ARPA CO monitoring stations, which was 2.96 km. Since ARPA sensors provide ground-truth observations, this resolution serves as a reliable reference for comparing and integrating other datasets. The *meshgrid* function from *NumPy* library (`https://numpy.org/doc/stable/reference/generated/numpy.meshgrid.html`) and *GeoPandas* library (`https://geopandas.org/en/stable/`) were used to create a uniform spatial grid at this resolution. These functions generate coordinate matrices from coordinate vectors, enabling us to create a regular grid for interpolation and assignment of data points.

For datasets with sparse or coarse resolution, such as CAMS and ERA5, we applied bilinear interpolation using the *LinearNDInterpolator* function from *SciPy* library (`https://scipy.org/`). This interpolation method is appropriate for scattered data, as it estimates values based on the weighted average of the surrounding data points, providing smooth approximations for the grid.

In contrast, for datasets with finer spatial resolution or discrete point measurements, such as ARPA CO and Sentinel-5P CO, we assigned data to the nearest grid point using *cKDTree*function from *SciPy* library. The *cKDTree* is highly efficient for nearest-neighbor queries, making it ideal for quickly finding the closest grid point to each data point, especially when dealing with large datasets.

2.3 Feature Engineering

In addition to the raw variables, several derived features were computed to enhance model performance.

2.3.1 Wind speed and wind direction Wind speed and wind direction were computed from *u* and *v* components of wind at 10m above the ground (*u10* and *v10*). Since machine learning models cannot inherently interpret directional or angular vector data such as wind direction, it is necessary to apply an appropriate encoding technique. We used cyclical encoding (Pranonsatit et al., 2025) to transform the wind direction, which ranges from 0° to 360°, into its sine and cosine components (wd_sin and wd_cos). This approach preserves the cyclical nature of the data.

2.3.2 Normalized Carbon Monoxide Using normalized CO (CO_per_blh), calculated as the ratio of Sentinel-5P CO (s5p_CO) to boundary layer height (blh), offers a more physically meaningful and context-aware feature for estimating ground-level CO concentrations compared to using raw *s5p_CO* values alone. This normalization accounts for the vertical mixing capacity of the atmosphere, which directly influences how satellite-observed column concentrations relate to surface-level pollution.

2.3.3 Solar thermal contrast Solar thermal contrast, defined as the normalized difference between surface net solar radiation (ssr) and surface net thermal radiation (str). This feature captures the balance between incoming solar energy and outgoing longwave radiation, which influences atmospheric stability and vertical mixing processes(Li et al., 2017, Yang et al., 2016).

2.3.4 Previous meteorological factors As mentioned earlier, we selected several distinct periods prior to 15:00 GMT+1: we tested among the preceding 4, 6, 12, 18, and 24 hours to explore how the results varies with the timeframe. For each time window, we calculated the average value of the meteorological factors over a specific period. This period starts at 15:00 GMT+1 the previous day and ends at the start of the current time window. We then added this calculated average value to the model's input data as a new feature. To identify these new features, we appended "_pre" to abbreviations of the original meteorological factor names. This helps to better describe the physical and chemical dynamics of the ground.

2.4 Dataset Splitting

We divided 60% of the dataset into a training set, 20% into a validation set, and 20% into a test set. We tested two data-splitting strategies and compare the results of them.

- **Shuffle split**: We used *train_test_split* function from *scikit-learn* library (`https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`) to split the data. We set the random state to control reproducibility and use random shuffle to put the data in a random order. In this case, models capture the relationship exclude the time trend of data.

- **Chronological split**: We sorted all unique dates from the index of the *pandas* DataFrame. The dates were then divided into three parts: training, validation, and testing. We created each dataset by filtering the DataFrame to contain only rows matching the dates in each part. In this way, the data is split in chronological order to maintain the temporal integrity of the time series. Models can also capture the time variation of data.

3. Methodology

3.1 Machine Learning Framework

The machine learning framework includes data download, data preprocessing and training pipeline, as shown in Figure 2. Data download and preprocessing were detailed in Chapter 3. In the following sections, we describe the training pipeline and the associated workflow in detail.

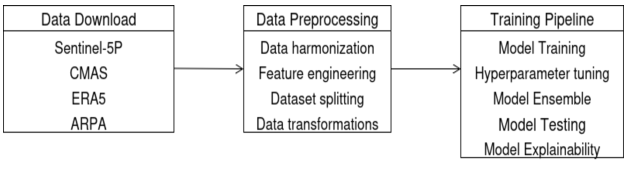

Figure 2. Machine learning framework.

3.2 Training Pipeline

We developed this work using a training pipeline based on previous work (Cedeno Jimenez and Brovelli, 2023). We referred to this pipeline as it has been deriving good performance, with an average NRMSE of 55.92% and R^2 of 0.76 when estimating ground-level NO_2 concentrations in MCM during the Sentinel-5P satellite pass. The algorithm was then adapted for estimating CO concentrations.

3.2.1 Model Training In this study, we implemented a series of models:

- Support Vector Regression (SVR)
- Decision Tree (DT)
- Random Forest (RF)
- Gradient Boosting (GB)
- XGBoost (XGB)
- Multilayer Perceptron (MLP)
- Dense Attention Network (DAN)
- Long Short-Term Memory (LSTM)

DAN and LSTM are two models adapted by us using *TensorFlow* (`https://www.tensorflow.org/`). DAN is a feed-forward neural network, featuring an attention mechanism that dynamically weights input features to emphasize informative signals and suppress noise. LSTM is a specialized recurrent neural network (RNN) with memory cells to capture long-range dependencies in sequential/temporal data (Hochreiter and Schmidhuber, 1997). The rest of the models were imported from *scikit-learn* (`https://scikit-learn.org/stable/`)library.

After data preprocessing, all data sources were integrated into a single pandas Dataframe. As mentioned earlier, we divided 60% of the DataFrame into a training set, 20% into a validation set, and 20% into a test set. Each set of data was divided into *X* and *y*. *X* was used as the input feature and *y* was used as the target (ground-level CO concentration).

3.2.2 Hyperparameter tuning For hyperparameter tuning, we used Bayesian optimization method via *BayesSearchCV* function (`https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html`). Bayesian optimization is an effective method to tune model's hyperparameters when evaluation is computationally expensive (Wang et al., 2024, Snoek et al., 2012). We used the validation set for hyperparameter tuning, after that we obtained the optimal set of parameters and then used them to retrain the models.

3.2.3 Model Ensemble In addition to the models previously described, we selected the two best-performing models from those imported from *scikit-learn*, based on their validation performance. We then used the *VotingRegressor* function (`https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingRegressor.html`) to develop an ensemble model from these two models.

The *Voting Regressor* is an ensemble learning method that combines predictions from multiple base regression models to produce a final output. Rather than training a meta-model, the *Voting Regressor* performs a simple averaging (either weighted or uniform) of the individual predictions from each base regressor. This technique helps to reduce variance and improve prediction robustness by using the strengths of diverse models (Breiman, 1996, Zhou, 2012). This method is particularly effective when combining diverse models as it tends to cancel out individual model biases and variances. Thus, we used this method to further improve our results.

3.2.4 Model Testing During testing, we retrained all models on the combined set of training and validation sets and then test them on the test set. To evaluate the performance of the model, we used the following metrics:

- Root Mean Square Error (RMSE): Measures the square root of the average squared differences between predicted and actual values.
- Mean Absolute Error (MAE): Measures the average absolute differences between predicted and actual values.
- Coefficient of Determination (R^2): Represents the proportion of variance in the dependent variable explained by the model.
- Normalized Root Mean Square Error (NRMSE): Normalizes RMSE using the standard deviation of actual values.

By using these metrics, we can perform a comprehensive evaluation of our model. But in this study, we mainly focus on NRMSE. NRMSE is a scale-independent metric that allows for intuitive and meaningful comparison of model performance across different datasets. It measures how prediction errors compare to the natural variability in the data: values below 1 indicate that the model performs better than simply using the mean, with lower values reflecting excellent performance. This normalization provides a clear sense of relative error and is especially useful for comparing models when data scales vary. This will allow us to compare the results with other studies in related fields in the future.

3.2.5 Model Explainability To interpret the estimations of our models, we employed SHAP (SHapley Additive exPlanations), a unified framework for interpreting model outputs based on cooperative game theory (Lundberg and Lee, 2017). SHAP assigns each feature an importance value for a particular prediction by computing Shapley values, which represent a fair distribution of the model's output among the input features. SHAP values were calculated to assess both global feature importance across the dataset and local explanations for individual predictions. In this study, the method provided insight into how variables influenced the predicted ground-level CO concentrations, enhancing interpretability and trust in the model outputs.

3.3 Workflow

To optimize the estimation of ground-level CO concentrations in the Metropolitan City of Milan (MCM), we applied a structured machine learning framework, focusing initially on the 11:00–15:00 GMT+1 window, aligned with Sentinel-5P overpass and CO detection timing. We compared models with and without hyperparameter tuning and evaluated various data splitting strategies. The analysis was then extended to additional

time windows (6, 12, 18, and 24 hours before 15:00 GMT+1) to identify the most effective period. Using SHAP values, we refined the feature set by removing less impactful variables, selected the best-performing model, and analyzed residuals to assess error patterns. To ensure robustness, we ran 20 shuffle-split validations, addressed extreme values in the ARPA dataset using the IQR method, and confirmed the statistical significance of improvements through Wilcoxon signed-rank tests.

4. Results and Discussion

4.1 Model Training with and without hyperparameter tuning

For the first set of tests, we used shuffle split to divide the dataset, trained models in the time window 11:00 to 15:00 GMT+1 without hyperparameter tuning. The resulting best model was RF. The NRMSE on the test set was 0.6363, as shown in Table 1. After implementing hyperparameter tuning, the best model was the ensemble model of MLP and GB using voting regressor. The NRMSE on the test set was 0.6216, as shown in Table 2. The performance improvement was not significant, but we mitigated overfitting. From Figure 3, we can see that the gap between the training and validation curves has narrowed before and after hyperparameter tuning.

Model	MAE	RMSE	R^2	NRMSE
RF	0.1195	0.1669	0.5952	0.6363
RF+GB	0.1198	0.1673	0.5932	0.6378
SVR	0.1205	0.1702	0.5794	0.6485
MLP	0.1219	0.1714	0.5735	0.6531
XGB	0.1239	0.1720	0.5703	0.6555
GB	0.1257	0.1743	0.5588	0.6642
DT	0.1769	0.2552	0.0544	0.9724

Table 1. Results of model training without hyperparameter tuning. Unit of MAE and RMSE is mg/m^3.

Model	MAE	RMSE	R^2	NRMSE
MLP+GB	0.1183	0.1631	0.6136	0.6216
MLP	0.1199	0.1640	0.6094	0.6250
XGB	0.1199	0.1644	0.6074	0.6266
SVR	0.1190	0.1663	0.5985	0.6336
GB	0.1220	0.1686	0.5873	0.6424
RF	0.1245	0.1701	0.5796	0.6484
DT	0.1394	0.1935	0.4564	0.7373

Table 2. Results of model training with hyperparameter tuning. Unit of MAE and RMSE is mg/m^3.

4.2 Comparing chronological split and shuffle split

Model	MAE	RMSE	R^2	NRMSE
LSTM	0.1844	0.2173	0.2782	0.8496
RF	0.1881	0.2178	0.2608	0.8598
XGB	0.1861	0.2179	0.2605	0.8600
SVR	0.1887	0.2203	0.2442	0.8694
SVR+GB	0.1922	0.2239	0.2188	0.8838
DT	0.1874	0.2274	0.1943	0.8976
GB	0.1988	0.2332	0.1525	0.9206
MLP	0.2019	0.2376	0.1204	0.9379

Table 3. Results of Chronological Split. Unit of MAE and RMSE is mg/m^3.

After the hyperparameter tuning was completed, we then tested chronological split with same features, models parameters and

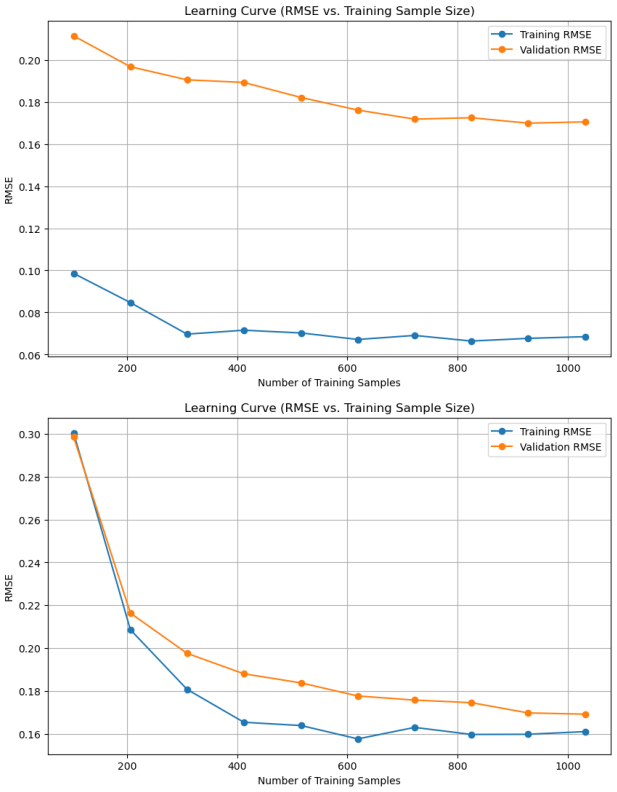

Figure 3. Learning curves of the best model before (top) and after (bottom) hyperparameter tuning.

time window. In this stage, the LSTM model performed best with a test set NRMSE of 0.8496, as shown in Table 3, which was 36.68% worse than the best result of the shuffle split. Since shuffle split works better in our case, we continued to use shuffle split in the following steps.

4.3 Additional features

After adding blh, blh_pre, CO_per_blh, stc, cams_CO, lai_lv, lai_hv, lai_lv_pre, lai_hv_pre, r and r_pre as features, the best model was ensemble model SVR + GB, with a test NRMSE of 0.5819, as shown in Table 4, which was 6.38% lower than the best result without adding them.

Model	MAE	RMSE	R^2	NRMSE
SVR+GB	0.1204	0.1619	0.6614	0.5819
GB	0.1214	0.1631	0.6563	0.5862
RF	0.1212	0.1650	0.6485	0.5929
SVR	0.1221	0.1655	0.6464	0.5947
XGB	0.1254	0.1681	0.6349	0.6042
MLP	0.1282	0.1698	0.6276	0.6102
DT	0.1365	0.1933	0.5177	0.6945

Table 4. Results of adding other features. Unit of MAE and RMSE is mg/m^3.

4.4 Comparing different time windows

We evaluated different time windows and found that the 21:00–15:00 GMT+1 period yielded the best performance (Table 5), with the SVR+GB ensemble model achieving a test NRMSE of 0.4794. This optimal performance likely arises from the window's alignment with key diurnal CO dynamics in MCM: nighttime accumulation, morning emission peaks, and

Time Window	Best Model	RMSE	NRMSE
11:00-15:00 GMT+1	SVR+GB	0.1619	0.5819
09:00-15:00 GMT+1	SVR+GB	0.1593	0.5432
03:00-15:00 GMT+1	SVR+GB	0.1487	0.5101
21:00-15:00 GMT+1	SVR+GB	0.1460	0.4794
15:00-15:00 GMT+1	SVR	0.1429	0.5132

Table 5. Results of comparing different time windows. Unit of RMSE is mg/m^3.

daytime dilution (Barlow, 2009, Turner, 2020). In the following studies, we continued to conduct research within the 21:00–15:00 GMT+1 time window.

4.5 Dropping unimportant features

Figures 4 and 5 illustrate SHAP-based feature importance, where the y-axis ranks features by their overall influence, and the x-axis shows the direction and magnitude of each feature's effect on individual predictions, with red points indicating higher feature values. In Figure 4, features such as cams_CO, blh_pre, CO_per_blh, lai_lv, lai_lv_pre, r, and r_pre show high importance, confirming their positive contribution to model performance. In contrast, features like lai_hv_pre, wind_dir_sin, and lai_hv had minimal impact and were progressively removed. The final model retained 16 features, as shown in Figure 5, leading to the best result with the DAN model, which achieved a test NRMSE of 0.4730 (Table 6), representing a 1.34% improvement over using all features. This configuration, using the 21:00–15:00 GMT+1 time window, was found to be optimal. To further assess model performance, Figure 6 presents a residual histogram with KDE (left) and a Q-Q plot (right); residuals appear approximately normally distributed with slight positive skew (skewness = 0.28), indicating mild underestimation and suggesting a generally good model fit with consistent variance (homoscedasticity).

Model	MAE	RMSE	R^2	NRMSE
DAN	0.1085	0.1440	0.7763	0.4730
SVR+MLP	0.1112	0.1459	0.7705	0.4791
SVR	0.1120	0.1483	0.7629	0.4869
MLP	0.1140	0.1488	0.7611	0.4888
GB	0.1145	0.1526	0.7489	0.5011
XGB	0.1170	0.1548	0.7415	0.5084
RF	0.1174	0.1570	0.7343	0.5155
DT	0.1401	0.1913	0.6054	0.6282

Table 6. Results of dropping unimportant features in time window 21:00-15:00 GMT+1. Unit of MAE and RMSE is mg/m^3.

4.6 20 independent shuffle-split experiments

In order to make the results more reliable. We set 20 different random states to perform 20 different shuffle splits in time window 21:00-15:00 GMT+1, and trained all models to make estimations 20 times.

From Table 7, we can see that the best model is still DAN, with Mean ± STD of the NRMSE being 0.4879 ± 0.0252, which is 0.4692% higher than the second-ranked ensemble model SVR_MLP.

4.7 Statistical tests on results

To assess whether the performance difference between DAN and SVR+MLP was statistically significant, we first computed

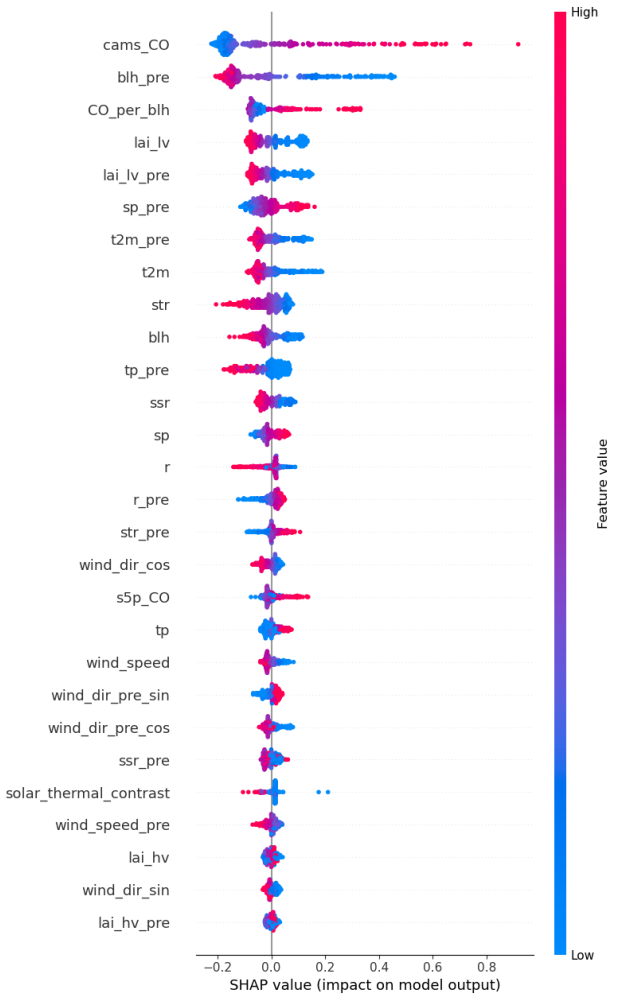

Figure 4. Distribution of SHAP values for CO prediction of best model in the 21:00-15:00 GMT+1 time window using all features.

Model	Mean ± STD of NRMSE
DAN	0.4879 ± 0.0252
SVR+MLP	0.4902 ± 0.0280
SVR	0.4938 ± 0.0264
GB	0.4959 ± 0.0272
MLP	0.4994 ± 0.0281
XGB	0.5104 ± 0.0288
RF	0.5164 ± 0.0290
DT	0.6026 ± 0.0379

Table 7. Results of 20 independent shuffle-split experiments without removing extreme values in time window 21:00-15:00 GMT+1.

the squared errors of their estimates in the 21:00–15:00 GMT+1 time window and observed a non-normal distribution of their differences (skewness = -1.15). This led us to use the Wilcoxon signed-rank test instead of a paired t-test. The result yielded a p-value of 0.02, indicating a significant difference in performance between DAN and SVR+MLP in that particular split. However, when we repeated the evaluation using 20 independent shuffle-split experiments and again applied the Wilcoxon test at a 95% confidence level, the performance difference between DAN and SVR+MLP was no longer statistically significant. This contrast is due to the single-split test being sensitive to data partition

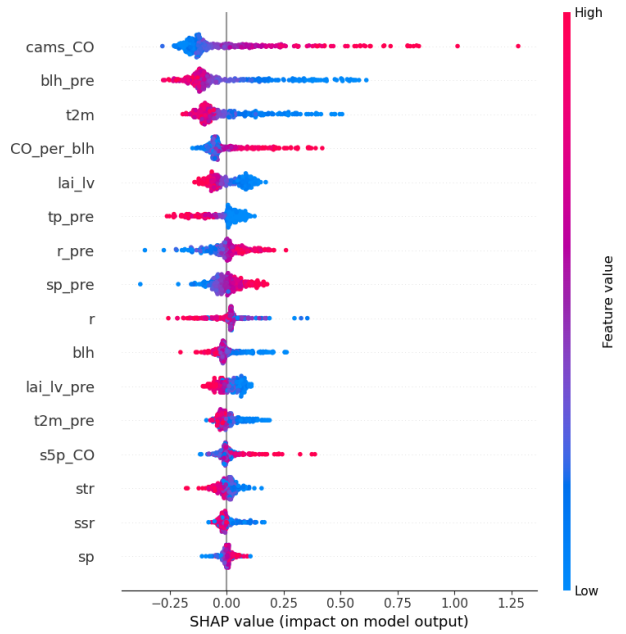

Figure 5. Distribution of SHAP values for CO prediction of best model in the 21:00-15:00 GMT+1 time window after removing unimportant features.

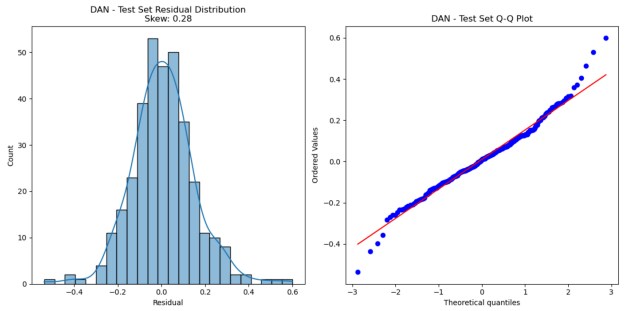

Figure 6. Histogram and Quantile-Quantile plot of residuals of ground-level CO concentration estimated by DAN in time window 21:00-15:00 GMT+1.

randomness, whereas the shuffle-split approach averages out variability, offering a more robust comparison. Across all pairwise comparisons, only a few, including DAN vs. SVR+MLP, did not show significant differences. Nevertheless, both DAN and SVR+MLP significantly outperformed all other models and were the top performers overall. Given that SVR_MLP relies on two separate models (SVR and MLP) and an ensemble strategy, while DAN is a standalone model, DAN remains the more practical and flexible choice despite their similar performance.

5. Conclusions and Future Work

This study developed a robust, data-driven framework for estimating ground-level CO concentrations in the Milan Central Metropolitan area (MCM) by integrating Sentinel-5P, CAMS reanalysis, and ERA5 meteorological data with advanced machine learning models. Through careful temporal analysis, we identified the 21:00–15:00 GMT+1 window as optimal for capturing CO accumulation and dispersion dynamics. Feature engineering efforts, particularly the inclusion of CO_per_blh and lagged meteorological variables, significantly enhanced model performance. The DAN model emerged as the top-performing architecture, outperforming others in terms of test NRMSE and residual behavior. Although an ensemble model (SVR+MLP) showed comparable performance across multiple shuffle splits, DAN was favored due to its simplicity and comparable accuracy. Statistical tests confirmed that DAN and SVR+MLP significantly outperformed other models, while the robustness of the final model was validated through 20 independent shuffle-split experiments.

Looking ahead, this framework could be extended in several important directions. First, adapting the approach for multi-pollutant modeling would enable more comprehensive air quality assessments, especially by including pollutants such as NO_2, O_3, and $PM_{2.5}$. Second, testing the framework across diverse geographic regions with varying emission profiles and meteorological conditions would demonstrate its transferability, potentially supported by domain adaptation or transfer learning methods. Lastly, integrating complementary data sources, such as traffic patterns, urban morphology, or mobile sensor networks, could further improve the model's ability to resolve fine-scale spatial variability and capture localized pollution events.

In conclusion, this study offers a scalable and cost-effective methodology for ground-level CO estimation in data-scarce environments, reinforcing the value of fusing remote sensing and reanalysis data with deep learning. The identification of optimal temporal windows and critical predictive features contributes valuable insights to pollutant-specific modeling strategies. By demonstrating the effectiveness of DAN and highlighting future enhancements, this work lays the foundation for improved urban air quality monitoring and supports evidence-based decision-making in environmental and public health policy.

Acknowledgements

This study was carried out within the Space It Up project funded by the Italian Space Agency, ASI, and the Ministry of University and Research, MUR, under contract n. 2024-5-E.0 - CUP n. I53D24000060005.

References

Arya, P. S., 2001. *Introduction to Micrometeorology*. 79, Elsevier.

Barlow, J., 2009. Boundary layer meteorology and dispersion. *Atmospheric Science for Environmental Scientists*, 218.

Breiman, L., 1996. Bagging Predictors. *Machine Learning*, 24(2), 123–140. doi: 10.1007/BF00058655.

Cedeno Jimenez, J. R., Brovelli, M. A., 2023. NO2 Concentration Estimation at Urban Ground Level by Integrating Sentinel 5P Data and ERA5 Using Machine Learning: The Milan (Italy) Case Study. *Remote Sensing*, 15(22), 5400. doi: 10.3390/rs15225400.

Chen, B., Hu, J., Wang, Y., Feng, T., Sun, W., Feng, Z., Yang, G., Wang, H., 2025. An interpretable physics-informed deep learning model for estimating multiple air pollutants. *GIScience & Remote Sensing*, 62(1), 2482272. doi: 10.1080/15481603.2025.2482272.

Chen, B., Zheng, Q., Sun, W., Yang, G., Feng, T., Wang, Y., 2024. Geo-STO3Net: A Deep Neural Network Integrating Geographical Spatiotemporal Information for Surface Ozone Estimation. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–14. doi: 10.1109/TGRS.2024.3358397.

Fania, A., Monaco, A., Pantaleo, E., Maggipinto, T., Bellantuono, L., Cilli, R., Lacalamita, A., La Rocca, M., Tangaro, S., Amoroso, N., Bellotti, R., 2024. Estimation of Daily Ground Level Air Pollution in Italian Municipalities with Machine Learning Models Using Sentinel-5P and ERA5 Data. *Remote Sensing*, 16(7), 1206. doi: 10.3390/rs16071206.

Google Earth Engine, 2025. Earth Engine Data Catalog - Sentinel-5P OFFL CO: Offline Carbon Monoxide. `https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_CO` (25 March 2025).

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J., 2025. (2023): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). doi: 10.24381/cds.adbb2d47.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V., Razinger, M., Remy, S., Schulz, M., Suttie, M., 2019. The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*, 19(6), 3515–3556. doi: 10.5194/acp-19-3515-2019.

Li, Z., Guo, J., Ding, A., Liao, H., Liu, J., Sun, Y., Wang, T., Xue, H., Zhang, H., Zhu, B., 2017. Aerosol and boundary-layer interactions and impact on air quality. *National Science Review*, 4(6), 810–833. doi: 10.1093/nsr/nwx117.

Lundberg, S. M., Lee, S., 2017. A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777. doi: 10.5555/3295222.3295230.

McNider, R. T., Pour-Biazar, A., 2020. Meteorological modeling relevant to mesoscale and regional air quality applications: a review. *Journal of the Air & Waste Management Association*, 70(1), 2–43. doi: 10.1080/10962247.2019.1694602.

Niyogi, D. S., Raman, S., 2001. Numerical modeling of gas deposition and bi-directional surface-atmosphere exchanges in mesoscale air pollution systems. *Mesoscale Atmospheric Dispersion*, 9, 424–484.

Pleim, J., Ran, L., 2011. Surface Flux Modeling for Air Quality Applications. *Atmosphere*, 2(3), 271–302. doi: 10.3390/atmos2030271.

Pranonsatit, J., Wongwailikhit, K., Painmanakul, P., Vateekul, P., 2025. Enhancing PM2.5 forecasting using video-based spatiotemporal models and cyclical encoding. *Recent Challenges in Intelligent Information and Database Systems*, Springer Nature Singapore, Singapore, 357–372. doi: 10.1007/978-981-96-5884-8$_2$6.

Shetty, S., Schneider, P., Stebel, K., Hamer, P. D., Kylling, A., Berntsen, T. K., 2024. Estimating surface NO2 concentrations over Europe using Sentinel-5P TROPOMI observations and Machine Learning. *Remote Sensing of Environment*, 312, 114321. doi: 10.1016/j.rse.2024.114321.

Smith, E. K., de Lauriere, C. F., Henninger, E., 2025. Persistent inequalities in global air quality monitoring should not delay pollution mitigation. *Proceedings of the National Academy of Sciences*, 122(18), e2423259122. doi: 10.1073/pnas.2423259122.

Snoek, J., Larochelle, H., Adams, R. P., 2012. Practical bayesian optimization of machine learning algorithms. *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, Curran Associates Inc., Red Hook, NY, USA, 2951–2959.

Sorbjan, Z., 2003. Air-pollution meteorology. P. Zannetti (ed.), *Air Quality Modeling - Theories, Methodologies, Computational Techniques, and Available Databases and Software. Vol. I - Fundamentals*, The EnviroComp Institute and the Air & Waste Management Association, chapter 4.

Turner, D. B., 2020. *Workbook of Atmospheric Dispersion Estimates: An Introduction to Dispersion Modeling*. CRC Press.

Veefkind, J. P., Aben, I., McMullan, K., Förster, H., de Vries, J., Otter, G., Claas, J., Eskes, H. J., de Haan, J. F., Kleipool, Q., van Weele, M., Hasekamp, O., Hoogeveen, R., Landgraf, J., Snel, R., Tol, P., Ingmann, P., Voors, R., Kruizinga, B., Vink, R., Visser, H., Levelt, P. F., 2012. TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sensing of Environment*, 120, 70–83. doi: 10.1016/j.rse.2011.09.027.

von Schneidemesser, E., Monks, P. S., Allan, J. D., Bruhwiler, L., Forster, P., Fowler, D., Lauer, A., Morgan, W. T., Paasonen, P., Righi, M., Sindelarova, K., Sutton, M. A., 2015. Chemistry and the Linkages between Air Quality and Climate Change. *Chemical Reviews*, 115(10), 3856–3897. doi: 10.1021/acs.chemrev.5b00089.

Wang, Z., Dahl, G. E., Swersky, K., Lee, C., Nado, Z., Gilmer, J., Snoek, J., Ghahramani, Z., 2024. Pre-trained Gaussian Processes for Bayesian Optimization. *Journal of Machine Learning Research*, 25(212), 1–83. http://jmlr.org/papers/v25/23-0269.html.

World Health Organization, 2024. Ambient (outdoor) air quality and health. `https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health` (23 May 2025).

Yang, X., Zhao, C., Zhou, L., Wang, Y., Liu, X., 2016. Distinct impact of different types of aerosols on surface solar radiation in China. *Journal of Geophysical Research: Atmospheres*, 121(11), 6459–6471. doi: 10.1002/2016JD024938.

Zhou, Z., 2012. *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC. doi: 10.1201/b12207.