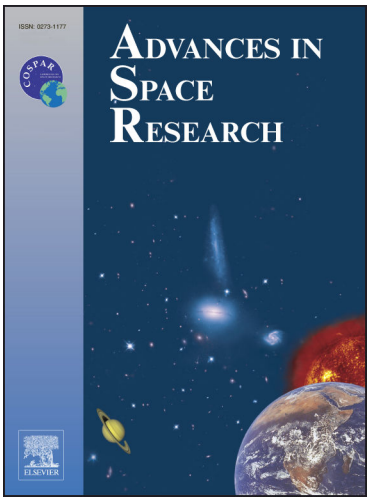
Journal Pre-proofs



Surface NO_2 Estimation in the Madrid Region from Sentinel-5P and ERA5-Land Meteorological Data Using Machine Learning Techniques

Rocco Giosa, Carlos Morillas, Sergio Alvarez, Carmine Serio, Guido Masiello, Giuliano Liuzzi, Sara Martinez

PII: S0273-1177(26)01545-0
DOI: https://doi.org/10.1016/j.asr.2026.11.020
Reference: JASR 20304

To appear in: *Advances in Space Research*

Received Date: 6 July 2026
Revised Date: 12 October 2026
Accepted Date: 5 November 2026

Please cite this article as: Giosa, R., Morillas, C., Alvarez, S., Serio, C., Masiello, G., Liuzzi, G., Martinez, S., Surface NO_2 Estimation in the Madrid Region from Sentinel-5P and ERA5-Land Meteorological Data Using Machine Learning Techniques, *Advances in Space Research* (2026), doi: https://doi.org/10.1016/j.asr.2026.11.020

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2026 The Author(s). Published by Elsevier B.V. on behalf of COSPAR.

Surface NO_2 Estimation in the Madrid Region from Sentinel-5P and ERA5-Land Meteorological Data Using Machine Learning Techniques

Rocco Giosa[a], Carlos Morillas[b]*, Sergio Alvarez[b], Carmine Serio[a], Guido Masiello[a], Giuliano Liuzzi[a] and Sara Martinez[b].

[a]Dipartimento di Ingegneria, Università degli Studi della Basilicata, Via dell'Ateneo Lucano 10, 85100 Potenza (PZ)

[b]Department of Land Morphology and Engineering, Universidad Politécnica de Madrid, 28040 Madrid, Spain

*Corresponding autor (c.morillas@upm.es)

Highlights

- Surface NO_2 estimated from TROPOMI tropospheric NO_2 columns and reanalysis data
- ERA5-Land reanalysis replaces local meteorological station data as model input
- Five machine learning models evaluated on an independent test year (2024)
- Sector-based spatial units gave the lowest errors for all models
- Errors analysed by sector, season and during the COVID-19 lockdown

Surface NO_2 Estimation in the Madrid Region from Sentinel-5P and ERA5-Land Meteorological Data Using Machine Learning Techniques

Abstract

Accurate estimation of surface-level nitrogen dioxide (NO_2) concentrations is critical for air quality assessment in urban regions where ground-based monitoring networks provide limited spatial coverage. This study investigates whether high predictive performance can be achieved by combining Sentinel-5P TROPOMI tropospheric NO_2

Journal Pre-proofs

column observations exclusively with reanalysis meteorological fields (ERA5-Land, complemented by ERA5 boundary-layer height): a globally available, infrastructure-independent data source. Five machine learning approaches were evaluated across the Community of Madrid over a five-year period (2020-2024): Random Forest, Support Vector Machine, a stacking ensemble (Random Forest and XGBoost), a Hybrid XGBoost Optimized model, and an Artificial Neural Network. Models were systematically assessed under two temporal aggregation strategies and four spatial configurations, using a year-based partition (training 2020–2022, validation 2023 and independent test 2024). The inclusion of antecedent atmospheric conditions through an extended temporal window improved predictive performance in 15 of the 20 model-configuration combinations. Environmentally coherent spatial stratification of monitoring stations outperformed both uniform regional aggregation and geometric grid-based schemes, and the sector-based configuration yielded the lowest errors for all five algorithms. The best-performing configuration (Artificial Neural Network combined with sector-based spatial classification) achieved an RMSE of 3.11 μg/m³, an R^2 of 0.86, and a MAE of 2.04 μg/m³ on the test year, closely followed by the Support Vector Machine (RMSE = 3.16 μg/m³). Absolute errors were largest in winter and in the urban core, and relative errors at rural background sites. During the COVID-19 lockdown, excluded from training, the model reproduced the temporal variability (r = 0.81) but overestimated concentrations by 19%. These results indicate that accurate daily NO_2 estimation is achievable using exclusively globally available satellite and reanalysis products, supporting a transferable and scalable framework for air quality monitoring in regions lacking dense observational infrastructure.

Keywords: Nitrogen dioxide; Sentinel-5P TROPOMI; ERA5-Land; Machine learning; Air quality modeling; Air pollution

Journal Pre-proofs

1. INTRODUCTION

Nitrogen dioxide (NO_2) remains among the most concerning air pollutants in large urban environments. Its presence in ambient air has been consistently linked to adverse health outcomes, including increased respiratory morbidity, cardiovascular complications, and premature mortality (Jonson et al., 2017; Shaw and Van Heyst, 2022). Elevated NO_2 concentrations also contribute to ecosystem acidification, reduce agricultural productivity, and accelerate the deterioration of built heritage (Kwon and Cho, 2023). These multiple impacts position NO_2 monitoring as a central pillar of urban environmental management and public health policy.

Anthropogenic NO_2 originates predominantly from combustion processes, with emissions from the transport sector accounting for the largest share in most metropolitan areas. Additional contributions come from industrial activities, domestic heating systems, and power generation facilities (McDuffie et al., 2020). The spatial distribution of these sources creates pronounced concentration gradients across cities, with traffic corridors and industrial zones exhibiting substantially higher pollution levels than residential or green areas. Temporal patterns are equally complex, driven by daily commuting cycles, seasonal heating demands, and meteorological variability that governs atmospheric dispersion and chemical transformation rates.

Madrid exemplifies the air quality challenges facing European cities. As Spain's capital and largest metropolitan area, it concentrates intense economic activity, high population density, and a transportation system historically centered on private vehicles. Multiple epidemiological studies have documented the health burden associated with air pollution in the city, with NO_2 exposure identified as a leading contributor to preventable mortality (Khomenko et al., 2021, 2023). Despite progressive implementation of traffic restriction measures and low-emission zones, achieving compliance with increasingly stringent air quality standards remains a formidable challenge. This difficulty is not only related to emissions. The metropolitan area lies in an elevated basin (approximately 600-700 m a.s.l.) bounded to the north-west by the Sierra de Guadarrama, and winter anticyclonic conditions frequently combine weak winds, shallow mixing layers and persistent thermal inversions that trap traffic emissions near the surface. High-pollution episodes in the city typically develop under such stagnant conditions (Borge et al., 2018). Conversely, the surrounding mountains favour mixing and lower concentrations, which produces strong spatial contrasts across the region (Morillas et al., 2024b).

The regulatory framework governing NO_2 concentrations in Europe has evolved considerably over the past two decades. While current legislation establishes a limit of 40 $\mu g/m^3$ for annual mean concentrations and 200 $\mu g/m^3$ for hourly peaks (European Commission, 2008), recently adopted directives mandate a reduction to 20 $\mu g/m^3$ by 2030 (European Commission, 2024), aligning more closely with World Health Organization (WHO) recommendations of 10 $\mu g/m^3$ (WHO, 2021). Simultaneously, Spanish national legislation compels cities exceeding 50,000 inhabitants to develop and implement comprehensive sustainable mobility strategies (Government of Spain, 2021). These converging regulatory pressures underscore the urgent need for accurate,

spatially resolved air quality information to guide policy decisions and evaluate intervention effectiveness.

Conventional monitoring approaches rely on networks of fixed ground-based stations equipped with reference-grade instruments. While these systems provide high-quality measurements at specific locations, their deployment faces significant practical constraints. Capital costs for fully equipped stations typically exceed €100,000, with substantial additional expenditures for site preparation, ongoing calibration, and maintenance. These economic barriers have resulted in marked spatial inequalities in monitoring coverage across Spain: of the 152 municipalities exceeding 50,000 residents, 99 operate with only a single station or none whatsoever (Government of Spain, 2026). Such limited infrastructure compromises the ability to characterize pollution patterns comprehensively, identify vulnerable populations, and verify regulatory compliance at appropriate spatial scales.

The advent of satellite-based atmospheric monitoring has opened new pathways for air quality assessment. The Sentinel-5 Precursor (Sentinel-5P) mission, operational within the European Copernicus programme since 2018, delivers daily measurements of tropospheric NO_2 columns at unprecedented spatial resolutions. The spectral capabilities of its TROPOspheric Monitoring Instrument (TROPOMI), combined with its sun-synchronous orbit configuration, enable systematic tracking of NO_2 distributions across continental scales. These observations have proven particularly valuable for identifying emission hotspots, characterizing pollution transport patterns, and detecting temporal trends associated with policy interventions or economic disruptions (Cersosimo et al., 2020; Garcia Santos and Parés, 2025; Goldberg et al., 2021; Griffin et al., 2019; Jeong and Hong, 2021; Morillas et al., 2024b; Rudke et al., 2023).

The utility of satellite NO_2 observations for surface-level air quality assessment depends fundamentally on their relationship with ground concentrations. This relationship varies spatially and temporally as a function of atmospheric vertical structure, boundary layer dynamics, and local emission characteristics (Morillas et al., 2025). In Madrid, satellite-derived tropospheric columns have demonstrated strong correspondence with surface measurements (Morillas et al., 2024b), likely attributable to the basin topography that promotes pollutant accumulation and the dominance of surface-level emission sources.

Translating column abundances into surface concentrations therefore requires accounting for meteorology, which influences NO_2 through several pathways. Temperature profiles determine atmospheric stability and vertical mixing intensity, and thus the depth of the boundary layer over which surface emissions are diluted. Wind patterns control horizontal advection and determine whether emissions accumulate locally or disperse regionally. Solar radiation drives photochemical reactions that convert NO_2 to other nitrogen species, creating pronounced diurnal and seasonal cycles. Precipitation events can temporarily reduce concentrations through wet deposition, while atmospheric pressure systems modulate larger-scale circulation patterns (Rahman et al., 2022; Zhang et al., 2025).

Recent methodological advances in statistical learning have enabled increasingly sophisticated integration of satellite and meteorological data for air quality estimation. Machine learning algorithms, encompassing both traditional ensemble methods and deep neural architectures, have demonstrated capacity to capture nonlinear relationships among atmospheric variables and generate accurate surface concentration predictions. While no single modeling approach has emerged as universally superior, comparative studies consistently highlight the importance of algorithm selection and feature engineering in determining predictive performance (Balamurugan et al., 2023; Cao, 2023; Cedeno Jimenez et al., 2023; Chan et al., 2021; Das et al., 2021; Deng et al., 2022; Grzybowski et al., 2023; He et al., 2022; Kim et al., 2021; Naseer et al., 2022; Shetty et al., 2024; Xing et al., 2024).

Usually, these models rely on surface meteorological observations obtained from urban monitoring networks, which provide essential contextual information for interpreting satellite-derived NO_2 estimations. However, the availability of such data is far from uniform. As with air quality monitoring, meteorological station coverage varies widely across regions, often limiting the weather information typically required for model training. This reliance on local infrastructure constrains the scalability and applicability of current approaches, particularly in areas with sparse monitoring networks.

Atmospheric reanalysis products offer a potential solution to this constraint. By assimilating diverse observational datasets into physically consistent numerical weather models, reanalysis systems generate spatially complete, temporally continuous estimates of atmospheric conditions at standardized resolutions. The Copernicus Climate Change Service provides the fifth-generation global atmospheric reanalysis of the European Centre for Medium-Range Weather Forecasts (ECMWF), ERA5 (Hersbach et al., 2020) and its land-surface counterpart ERA5-Land (Muñoz Sabater, 2019), which offer hourly fields over several decades, the latter at approximately 9 km resolution, with consistent coverage even in regions with sparse or absent observational infrastructure.

This study develops and evaluates a methodology for estimating surface NO_2 concentrations by combining Sentinel-5P tropospheric column observations with reanalysis meteorology, and continues our earlier work in the same region. Morillas et al. (2024b) assessed the consistency between TROPOMI columns and surface observations, and Morillas et al. (2024a) used TROPOMI data to quantify the impact of Madrid's low emission zone; neither study estimated surface concentrations. Morillas et al. (2026) introduced the machine learning framework and the spatial stratification strategies adopted here, but relied on meteorological observations from a dense network of ground-based weather stations, which limits its transferability. The present study examines whether comparable predictive performance can be maintained when this local infrastructure is replaced by globally available products. Its specific contributions are: (i) all predictors are globally available (Sentinel-5P, ERA5-Land and ERA5), with no station-based meteorology or surface NO_2 observations among the inputs; (ii) five algorithms are compared under a year-based validation design with an independent test year; (iii) performance is quantified by spatial unit and season; (iv) the sensitivity of the results to feature screening, dimensionality reduction, input scaling and data partitioning is tested; and (v) the models are evaluated out of sample during the COVID-19 lockdown, when emissions departed abruptly from their usual levels.

The analysis encompasses the Community of Madrid over a five-year period (2020-2024). Tree-based ensembles, a support vector machine (SVM) and an artificial neural network (ANN) are examined across regional, grid-based, sector-based and cluster-based spatial configurations to assess their predictive performance.

2. MATERIALS AND METHODS

The study area is the Community of Madrid, in the central Iberian Peninsula (Figure 1). Its 179 municipalities range from the densely populated metropolitan area of the city of Madrid to the mountains of the Central System in the north and north-west.

The study period extends from 1 January 2020 to 31 December 2024, encompassing five complete years for which satellite observations are available at a uniform spatial resolution (see Section 2.1).

2.1 NO_2 data

Two NO_2 datasets were used (Table 1): tropospheric columns from satellite observations and surface concentrations from the regional air quality monitoring network. Satellite-derived tropospheric NO_2 column densities were retrieved from TROPOMI aboard the Sentinel-5P platform. Launched in October 2017 and achieving operational status in June 2018, Sentinel-5P operates in a sun-synchronous polar orbit at 824 km altitude with an equatorial crossing time near 13:30 local solar time. The sensor's orbital configuration yields daily global coverage through 14 orbits per day, and overpasses over the Iberian Peninsula occur between 11:00 and 14:00 Coordinated Universal Time (UTC). The instrument's 2,600 km swath width frequently permits dual coverage of the study region within a single day when consecutive orbital tracks overlap, separated by approximately 100 minutes (Veefkind et al., 2012).

TROPOMI's spatial resolution underwent enhancement in August 2019, improving from 3.5×7.5 km^2 to 3.5×5.5 km^2 at nadir. This study utilizes data exclusively from the post-enhancement period to maintain spatial consistency across the analysis window. The NO_2 retrieval algorithm employs differential optical absorption spectroscopy in the 405-465 nm spectral range, with tropospheric column separation achieved through stratospheric correction procedures (Van Geffen et al., 2019).

Data acquisition was performed through Google Earth Engine (GEE), a cloud-based platform that enables large-scale processing and provides direct access to pre-ingested satellite datasets. While the original NO_2 product distributed by Copernicus corresponds to Level-2 data, GEE supplies a Level-3 version reprojected onto a regular $0.01° \times 0.01°$

latitude–longitude grid, which greatly facilitates spatial querying and integration with ancillary datasets. Within this Level-3 processing chain, embedded quality-assurance (QA) procedures filter out observations with QA values below 0.75 (QA < 0.75), thereby removing pixels affected by cloud contamination, snow or ice surfaces, or retrieval convergence issues. The offline (OFFL) product stream was selected over near-real-time (NRTI) alternatives based on superior retrieval accuracy resulting from extended processing time allowances (GEE, 2025).

Across the 1,827-day study period, at least one valid TROPOMI observation was available over the study area on 1,779 days (97.4%), from a total of 2,574 overpass images; on 664 of these days, two consecutive overpasses provided valid data, which were averaged. Data availability showed a moderate seasonal cycle, decreasing from about 97% of the days in April-September to 92% in October-March (88% in December and January), consistent with the higher frequency of overcast conditions in winter.

Ground-level NO_2 concentration measurements were obtained from the Community of Madrid's air quality monitoring network. These station-based observations were used as reference ground-truth data for model training, validation, and testing. To ensure data continuity and comparability across the analysis period, station selection was restricted to sites maintaining uninterrupted operation from January 2020 through December 2024. This criterion yielded 48 qualifying stations: 24 operated by the Madrid City Council (City of Madrid, 2025) within municipal boundaries and 24 managed by the regional government across the remaining territory (Community of Madrid, 2025). All stations employ chemiluminescence analyzers conforming to European reference method standards, recording NO_2 concentrations at hourly intervals. Measurements are openly available through institutional platforms with typical reporting latency under 24 hours. The spatial distribution of the air quality network is presented in Figure 1, which also shows the mean NO_2 concentration measured at each station over the study period, and in Table S1, which lists the coordinates of each station and its assignment to the spatial units described in Section 2.3. The station mean was computed as the mean of the annual means of the years with at least 90% valid hourly data, which is the minimum data capture required by Directive 2008/50/EC for annual averages (European Commission, 2008). The figure illustrates the gradient from the urban core to the rural and mountain stations of the Central System.

Table 1.- NO_2 data used in the study.

Variable	Unit	Source
Tropospheric NO_2 column density	µmoles/m^2	Sentinel-5P TROPOMI
Surface NO_2 concentration	µg/m^3	Regional air quality stations

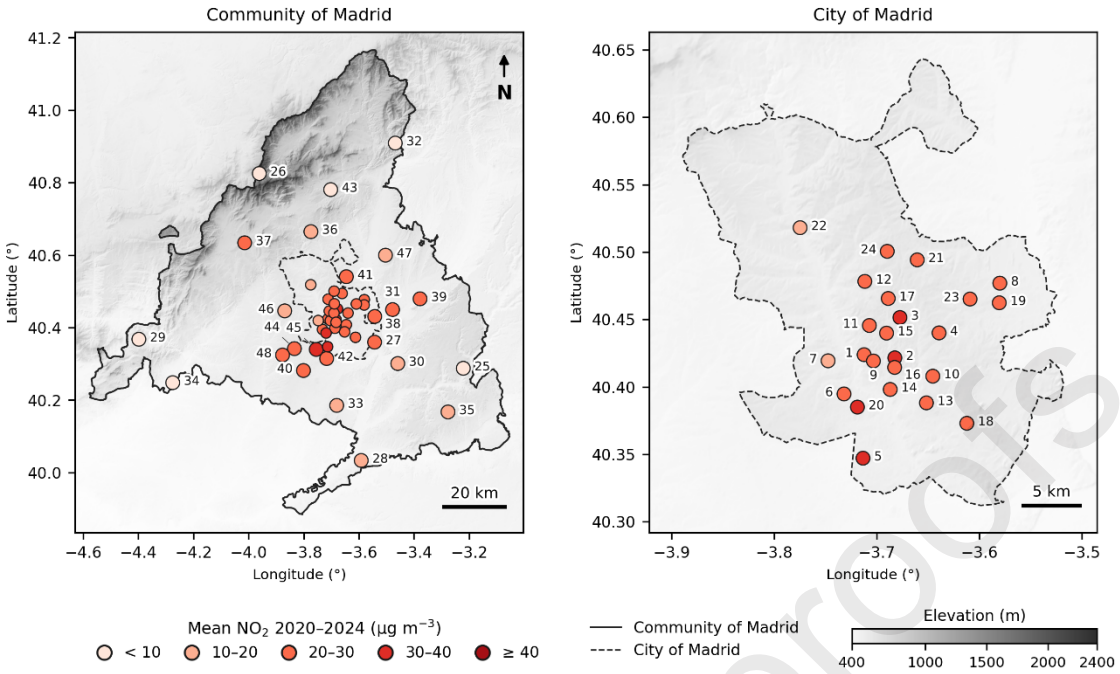


Figure 1.- Air quality monitoring network and mean surface NO_2 concentration at each station (2020–2024) in the Community of Madrid (left) and the City of Madrid (right), over terrain elevation. Station numbers as in Table S1.

2.2 Meteorological data

Meteorological predictors were obtained from two ECMWF reanalyses, openly distributed through the Copernicus Climate Data Store with a latency of about five days. Most variables were taken from ERA5-Land, which provides hourly fields at 0.1° × 0.1° from 1950 to the near-present. ERA5-Land is produced by re-running the ECMWF land-surface model at higher resolution, forced by the atmospheric fields of ERA5. Boundary-layer height (BLH), which is not part of the ERA5-Land variable set, was taken from the hourly single-level fields of ERA5, available at 0.25° × 0.25° (Hersbach et al., 2020).

The following ERA5-Land variables were used directly: zonal (eastward) wind component at 10 m height (10m u-component of wind in ERA5-Land dataset), meridional (northward) wind component at 10 m height (10m v-component of wind), air temperature at 2 m height (2m temperature), surface temperature (skin temperature), surface pressure, surface solar radiation downwards and total precipitation.

BLH was included because it governs the vertical dilution of surface emissions and therefore modulates the relationship between the tropospheric column and the surface concentration (Seibert et al., 2000).

Relative humidity (RH) was not directly available in the ERA5-Land dataset and was therefore derived from 2 m air temperature and 2 m dew point temperature. Specifically, RH was computed using the Magnus formulation (Alduchov and Eskridge, 1996), which provides an accurate approximation of the saturation vapor pressure over water within the temperature range typical of near-surface atmospheric conditions.

First, the saturation vapor pressure $e_s(T)$ and the actual vapor pressure $e(T_d)$ were calculated as:

$$e_s(T) = \exp\left(\frac{17.625 \cdot T}{243.04 + T}\right)$$

$$e(T_d) = \exp\left(\frac{17.625 \cdot T_d}{243.04 + T_d}\right)$$

where T is the air temperature (°C) and Td is the dew point temperature (°C).

Relative humidity was then obtained as:

$$RH = 100 \cdot \frac{e(T_d)}{e_s(T)}$$

Wind direction was derived by calculating the angle from the zonal and meridional wind components averaged over each temporal window (Section 2.3).

$$\theta_{mean} = \arctan(u,v)$$

The cosine and sine of the mean wind direction were then computed, so that close directions on either side of north (e.g., 359° and 1°) are represented by similar values:

$$\cos(\theta_{mean}),\ \sin(\theta_{mean})$$

Journal Pre-proofs

2.3 Data processing

First, all predictor variables were temporally harmonized with the Sentinel-5P overpass schedule. Meteorological inputs were aggregated using two complementary time windows, hereafter referred to as H1 and H2. The first (H1) consisted of averaging each variable over the 11:00-14:00 UTC interval, corresponding to the typical satellite overpass period over the study area. The second (H2) involved computing mean values over the preceding 21-hour interval, spanning from 14:00 of the previous day to 11:00 of the observation day. This latter aggregation is intended to represent the accumulated atmospheric conditions influencing surface-level NO_2 concentrations at the time of satellite acquisition. The same two windows were used in previous studies (Cedeno Jimenez et al., 2023; Cedeno Jimenez and Brovelli, 2023; Morillas et al., 2026).

These predictors were averaged over the overpass window rather than matched to the exact acquisition time because the overpass time varies from day to day (Section 2.1) and both the reference NO_2 analysers and the reanalyses provide hourly values, so that any matching already involves temporal averaging. Over the same region and monitoring network, averaging over the overpass interval increased the TROPOMI-surface correlation from r = 0.78 to r = 0.81 (Morillas et al., 2024b). The H2 window adds antecedent conditions: given NO_2 lifetimes of several hours or longer (Beirle et al., 2011; Shah et al., 2020) and frequent nocturnal inversions in the Madrid basin, midday concentrations also depend on the dispersion conditions of the preceding night and morning.

Temporal variability in NO_2 concentrations was addressed by including time-based features: the month and the day of the week, both represented as cyclical (sine-cosine) variables to capture seasonal and weekly patterns. Surface NO_2 concentrations, averaged over the H1 window, were used only as the target variable and never as predictors, so that the trained models can be applied where no monitoring station exists.

Spatial variability was addressed by grouping the monitoring stations into spatial units according to geometric or environmental criteria. For each unit and day, the target was the mean H1 concentration of its stations, and the TROPOMI column and the meteorological variables were averaged over the pixels and reanalysis grid points within the unit. BLH, whose 0.25° grid is coarser than several units, was taken from the ERA5 grid point nearest to the unit centre when the unit contained none, as it varies smoothly in space. For each configuration, a single model was trained on the pooled unit-day samples, with the unit identifier as a one-hot encoded predictor (Table 2). Four spatial configurations were evaluated (Sections 2.3.1-2.3.4).

The predictor variables considered in the modeling framework are detailed in Table 2.

Table 2.- Predictor variables used for estimating surface-level NO_2 concentrations. All times in UTC.

Variable	Description
Tropospheric NO_2 column density	Tropospheric NO_2 column measured at the time of Sentinel-5P overpass
u H1	Average zonal (eastward) wind component at 10 m height between 11:00 and 14:00
u H2	Average zonal (eastward) wind component at 10 m height between 14:00 of the previous day and 11:00 of the current day
v H1	Average meridional (northward) wind component at 10 m height between 11:00 and 14:00
v H2	Average meridional (northward) wind component at 10 m height between 14:00 of the previous day and 11:00 of the current day
$\cos(\theta_{mean})$ H1	Cosine of average wind direction at 10 m height between 11:00 and 14:00
$\cos(\theta_{mean})$ H2	Cosine of average wind direction at 10 m height between 14:00 of the previous day and 11:00 of the current day
$\sin(\theta_{mean})$ H1	Sine of average wind direction at 10 m height between 11:00 and 14:00
$\sin(\theta_{mean})$ H2	Sine of average wind direction at 10 m height between 14:00 of the previous day and 11:00 of the current day
2m air temperature H1	Average air temperature at 2 m height between 11:00 and 14:00
2m air temperature H2	Average air temperature at 2 m height between 14:00 of the previous day and 11:00 of the current day
Surface temperature H1	Average surface temperature between 11:00 and 14:00
Surface temperature H2	Average surface temperature between 14:00 of the previous day and 11:00 of the current day
Surface pressure H1	Average surface atmospheric pressure between 11:00 and 14:00

Journal Pre-proofs

Variable	Description
Surface pressure H2	Average surface atmospheric pressure between 14:00 of the previous day and 11:00 of the current day
Surface solar radiation H1	Average surface solar radiation between 11:00 and 14:00
Surface solar radiation H2	Average surface solar radiation between 14:00 of the previous day and 11:00 of the current day
Precipitation H1	Average hourly precipitation between 11:00 and 14:00
Precipitation H2	Average hourly precipitation between 14:00 of the previous day and 11:00 of the current day
Relative humidity H1	Average relative humidity between 11:00 and 14:00
Relative humidity H2	Average relative humidity between 14:00 of the previous day and 11:00 of the current day
Boundary-layer height H1	Average ERA5 boundary-layer height between 11:00 and 14:00
Boundary-layer height H2	Average ERA5 boundary-layer height between 14:00 of the previous day and 11:00 of the current day
Month	Month of the year (encoded as sine and cosine)
Day of the week	Day of the week (encoded as sine and cosine)
Spatial unit	Identifier of the grid cell, sector or cluster (one-hot encoded; not used in the regional configuration)

Journal Pre-proofs

2.3.1 Regional aggregation approach

All monitoring stations within the Community of Madrid were considered jointly as a single spatial domain, so that the target was the daily mean of all stations (Figure 1). In this configuration, the study area was assumed to be spatially homogeneous for the purposes of model training.

2.3.2 Grid-based approach

In this approach, the study area was partitioned into a regular grid with a cell size of 10 km. This choice was guided by the 0.1° × 0.1° resolution of ERA5-Land, which provides most of the predictors. Monitoring stations were then assigned to grid cells according to their geographic location.

Cells were retained only if they contained at least one monitoring station and at least one ERA5-Land grid point, which yielded the 24 cells shown in Figure 2. Because ERA5-Land grid points are spaced at 0.1° (about 11 km in latitude and 8.5 km in longitude at this latitude), the cell containing the Puerto de Cotos station, a remote mountain site, included no grid point, and this station was therefore not used in the grid-based configuration (47 stations). BLH was not used as a retention criterion, since only 4 of the 24 cells contain an ERA5 grid point.

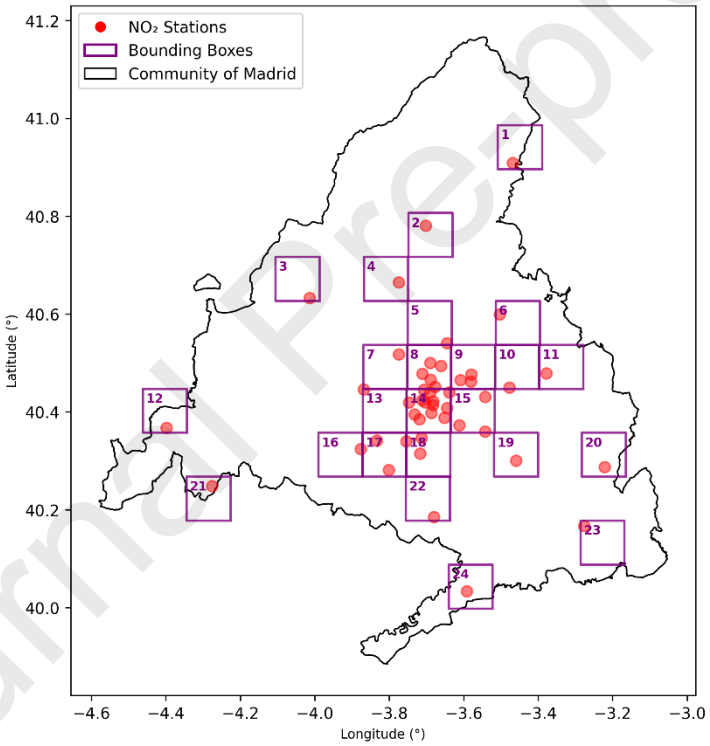


Figure 2.- Spatial distribution of the 10 × 10 km^2 grid cells retained for analysis.

2.3.3 Sector-based approach

In this approach, the Community of Madrid was divided into seven sectors (Figure 3). The aim was to group stations with similar emission environments and meteorological behavior, thereby reducing heterogeneity within each aggregation unit. The sectors were delineated as rectangular bounding boxes along the main gradients of elevation, urbanization and population density, which are shown together with the main road network in Figure S1. They separate the northern mountain ranges (sector 5), the urban core (sector 4), the eastern Henares corridor (sector 6), the south-western metropolitan

ring (sector 3) and the rural west, south and south-east (sectors 1, 2 and 7). Their extent and descriptors are given in Table 3, which shows that the sectors differ systematically in the variables that control the column-surface relationship. The mean surface NO_2 concentration at the overpass time ranges from 3.8 µg/m^3 (sector 1) to 19.1 µg/m^3 (sector 4), and the daily TROPOMI-surface correlation decreases from r = 0.86-0.87 in the metropolitan sectors 3 and 4 to r = 0.60-0.68 in the rural south-east (sector 7) and the mountains (sector 5).

Table 3.- Definition and descriptors of the sectors used in the sector-based approach. Surface NO_2 and TROPOMI column: means over the H1 window, 2020-2024; r: Pearson correlation between daily TROPOMI column and surface NO_2.

Unit	Environment	Latitude [°]	Longitude [°]	Stations	Surface NO_2 [µg/m^3]	TROPOMI column [µmol/m^2]	r
1	West, rural	40.20 / 40.40	-4.50 / -4.25	2	3.8	29	0.77
2	South, semi-rural	40.00 / 40.25	-3.75 / -3.50	2	7.2	44	0.76
3	South-west metropolitan	40.25 / 40.48	-3.95 / -3.79	4	11.3	71	0.86
4	Central urban core	40.30 / 40.57	-3.79 / -3.60	25	19.1	95	0.87
5	North, mountain	40.57 / 40.94	-4.10 / -3.43	6	6.5	37	0.68
6	East, suburban corridor	40.25 / 40.56	-3.60 / -3.33	7	13.3	67	0.78
7	South-east, rural	40.13 / 40.33	-3.33 / -3.15	2	4.9	31	0.60

Journal Pre-proofs

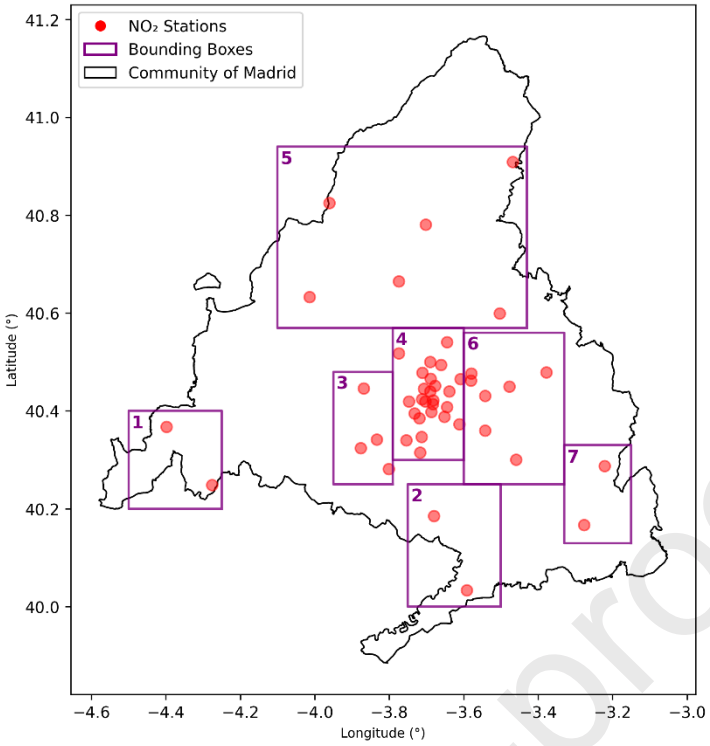


Figure 3.- Sectors used in the sector-based approach.

2.3.4 Simplified coherent clusters approach

Building upon the sector-based configuration, the study area was further simplified into three larger clusters of contiguous bounding boxes representing the dominant environmental and emission regimes (Figure 4). This configuration preserves the main center–periphery contrasts of the region while reducing spatial fragmentation. The clusters are described in Table 4: cluster 1 (central) groups the metropolitan core and the south-western and eastern metropolitan areas (the stations of sectors 3, 4 and 6 except Alcalá de Henares; 35 stations); cluster 2 (northern) contains the mountain and rural stations of sector 5 (6 stations); and cluster 3 (eastern) includes Alcalá de Henares and the rural stations of sector 7 (3 stations). The four stations of sectors 1 and 2, in the west and south of the region, were deliberately excluded, since their mixed topographic and land-use settings did not fit any of the three clusters and their inclusion would have reduced intra-cluster homogeneity.

Table 4.- Definition and descriptors of the clusters used in the simplified coherent clusters approach. Surface NO_2 and TROPOMI column: means over the H1 window, 2020-2024; r: Pearson correlation between daily TROPOMI column and surface NO_2.

Unit	Environment	Latitude [°]	Longitude [°]	Stations	Surface NO_2 [µg/m³]	TROPOMI column [µmol/m²]	r
1	Central, metropolitan	40.20 / 40.58	-4.10 / -3.43	35	17.3	68	0.86

Unit	Environment	Latitude [°]	Longitude [°]	Stations	Surface NO_2 [μg/m³]	TROPOMI column [μmol/m²]	r
2	North, mountain	40.58 / 41.20	-4.10 / -3.43	6	6.6	31	0.59
3	East, peri-urban and rural	40.10 / 40.58	-3.42 / -3.15	3	7.2	39	0.67

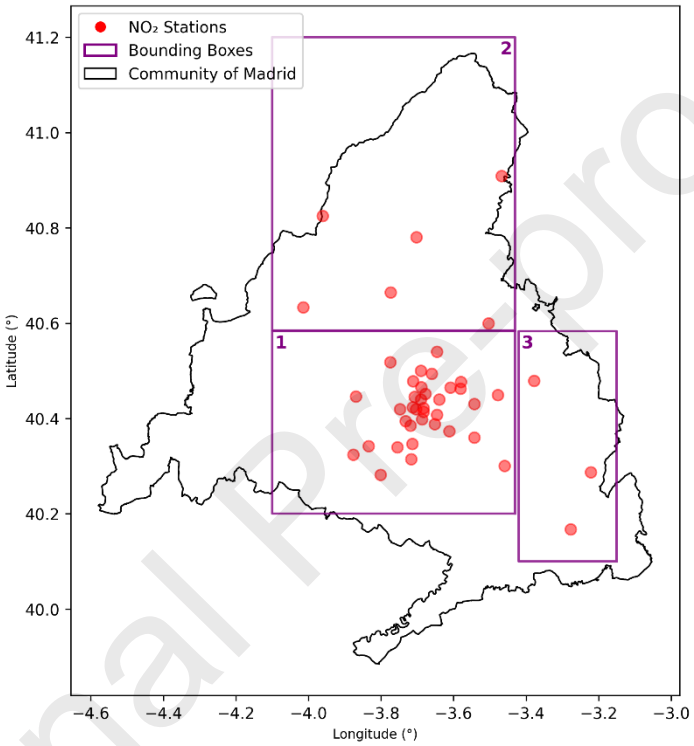


Figure 4.- Spatial clusters used in the simplified coherent clusters approach.

2.4 Machine learning models

Five models were used to estimate surface-level NO_2 concentrations in each spatial configuration: Random Forest (RF), a stacking ensemble of RF and eXtreme Gradient Boosting (XGBoost), a Hybrid XGBoost Optimized model, an SVM and an ANN, described in Sections 2.4.1-2.4.5. Model accuracy was assessed through three complementary metrics: the Root Mean Squared Error (RMSE), reflecting the typical magnitude of estimation errors; the coefficient of determination (R^2), indicating the proportion of observed variance reproduced by the model; and the Mean Absolute Error (MAE), representing the mean absolute discrepancy between estimated and measured values. The mean bias (predicted minus observed) was additionally used to diagnose systematic errors.

Journal Pre-proofs

All models used the same partition by calendar year: 2020-2022 for training, 2023 for validation and 2024 as an independent test year. Contiguous yearly blocks were preferred to a random split because consecutive days share meteorological regimes and pollution episodes, which inflates the apparent skill of randomly partitioned data (Meyer et al., 2018; Roberts et al., 2017). The validation year was used only for model selection; the final models were then refitted on 2020-2023, except the ANN, whose training length had been selected on 2023 (Section 2.4.5). All reported metrics refer to the 2024 test year.

Before training the RF, stacking, XGBoost and SVM models, a common feature-screening step was applied: an auxiliary RF with 100 trees was fitted to the training years, the impurity-based importance of each predictor was computed, and predictors whose importance did not exceed the 25th percentile of the importance distribution were discarded. This retained approximately three quarters of the inputs (e.g., 25 of 34 in the sector-based configuration). Screening predictors by RF importance is an established embedded selection approach for high-dimensional regression (Genuer et al., 2010) and, because it was computed on the training years only, it did not use any information from the validation or test data. The ANN received the complete predictor set, as its weights perform an implicit selection. Principal component analysis (PCA) was not adopted as the main reduction method because it is unsupervised and replaces physically interpretable inputs with linear combinations of them (Jollife and Cadima, 2016); it was nevertheless tested as an alternative (Section 3.4).

Inputs were standardized (z-scores computed with the mean and standard deviation of the training years) for the SVM and the ANN, whose kernel distances and gradient-based optimization depend on the scale of the predictors. No scaling was applied to the tree-based models (RF, stacking and XGBoost), because their splits depend only on the ordering of the values of each variable and are therefore invariant to monotonic transformations (Hastie et al., 2009); this invariance was also verified empirically (Section 3.4).

2.4.1 Random Forest

A RF regressor served as the primary baseline in the modeling framework. This ensemble method operates by training a large collection of decision trees on bootstrap-sampled subsets of the training data, with random feature subsets evaluated at each node split. Combining the outputs of all trees through averaging yields predictions with reduced variance and improved generalization relative to individual decision trees (Breiman, 2001).

The model was configured with 300 decision trees. Its remaining hyperparameters were kept at the default values of scikit-learn (Pedregosa et al., 2011).

2.4.2 Stacking ensemble (RF and XGBoost)

The stacking model combines RF and XGBoost in two levels. First, an RF and an XGBoost regressor (Chen and Guestrin, 2016) were trained on the screened predictors; then, a second XGBoost model (the meta-learner) learned how to combine their predictions. To avoid training the meta-learner on predictions that the base models had made for their own training data, these predictions were generated by 5-fold cross-validation (out-of-fold predictions). The three components used 100 trees each and, for XGBoost, a learning rate of 0.05; these values were fixed a priori.

2.4.3 Hybrid XGBoost Optimized

The Hybrid XGBoost Optimized model applies XGBoost to the screened predictors after tuning its hyperparameters by Bayesian optimization, implemented in the Optuna framework (Akiba et al., 2019) with its tree-structured Parzen estimator sampler (Bergstra et al., 2011). In each of 100 trials, a candidate set of hyperparameters was used to train a model on 2020-2022, and the set yielding the lowest mean squared error on the 2023 validation year was retained. The search covered the number of trees (100-500), the maximum tree depth (3-20), the learning rate (0.01-0.3), the row and column subsampling ratios (0.5-1.0) and the minimum loss reduction required to split a node (0-5). The designation hybrid reflects the integration of RF-based feature screening and Bayesian-optimized XGBoost, as adopted throughout the study.

2.4.4 Support Vector Machine

The SVM was included as a kernel-based alternative. In its ε-support vector regression (ε-SVR) form, it fits a function that ignores deviations smaller than a tolerance ε while keeping the function as smooth as possible, which favours generalization. A radial basis function (RBF) kernel was used to capture nonlinear relationships, with a regularization parameter $C = 10$, $\varepsilon = 0.1$ and the kernel coefficient set to the default 'scale' option of scikit-learn; these values were fixed a priori.

2.4.5 Artificial neural network

The ANN, suited to approximating highly nonlinear relationships, was designed automatically with AutoKeras (Jin et al., 2019), an automated machine learning (AutoML) tool that searches among candidate network architectures. The search was restricted to fully connected (dense) networks of varying depth and width with a single continuous output. Up to 30 candidate architectures were compared by their mean squared error on the 2023 validation year, and the best one was retained. Its training length was then selected on the same year: the network was trained for 20 epochs (complete passes through the training data) and then for up to 20 additional cycles of 10 epochs, and the shortest training whose validation R^2 matched the best value to the second decimal was kept, favouring the simplest of practically equivalent models. Random seeds were fixed and deterministic operations enforced to ensure reproducibility.

3. RESULTS

3.1 Model performance

Table 5 presents the predictive performance obtained for all evaluated machine learning models, including RF, SVM, ANN, and the hybrid ensemble approaches combining RF and XGBoost (stacking model and XGBoost-optimized model), under the different spatial aggregation strategies considered in this study. All values refer to the independent 2024 test year (341 days in the regional configuration and 8029, 2060 and 940 unit-days in the grid-based, sector-based and cluster-based configurations), and Figure 5 displays the same information graphically.

As described in Section 2.3, model training was performed using two alternative temporal intervals: (i) the satellite overpass period alone (H1 only), and (ii) the combination of the overpass period and the extended historical window (H1+H2). Adding the H2 window reduced the RMSE in 15 of the 20 model-clustering combinations. The largest gains were obtained in the regional appoach (0.40-0.58 $\mu g/m^3$ for RF, XGBoost, SVM and ANN, and 0.29 $\mu g/m^3$ for the stacking model), where the predictors are averaged over the whole region and the antecedent conditions add most information. In the cluster-based approach the gains were consistent but smaller (0.04-0.11 $\mu g/m^3$), in the sector-based approach they ranged from 0.02 to 0.08 $\mu g/m^3$ except for the stacking model, whose RMSE increased by 0.20 $\mu g/m^3$, and in the grid-based configuration the effect was negligible (-0.01 to +0.09 $\mu g/m^3$).

In the RF screening step, the TROPOMI column was by far the most important predictor (impurity-based importance of 0.73 in the sector-based approach), followed by the H2 boundary-layer height (0.05); each of the remaining meteorological and temporal variables had an importance below 0.05. This selected feature set is physically consistent with the main mechanisms controlling NO_2 accumulation and transport in urban atmospheres, and agrees with previous interpretability studies conducted for the Madrid region (Morillas et al., 2025).

Journal Pre-proofs

Table 5.- Predictive performance of machine learning models across spatial classification approaches. Values refer to the test year 2024; RMSE and MAE in μg/m³. Epochs: number of ANN training epochs, selected on the 2023 validation year (Section 2.4.5).

Model	Approach	H1+H2 training				H1 training			
		Epochs	RMSE	R^2	MAE	Epochs	RMSE	R^2	MAE
Random Forest	Regional aggregation approach	-	4.43	0.73	3.37	-	4.99	0.65	3.62
	Grid-based approach	-	4.97	0.69	3.31	-	4.88	0.70	3.25
	Sector-based approach	-	3.45	0.82	2.37	-	3.47	0.82	2.38
	Simplified coherent clusters	-	3.54	0.79	2.53	-	3.60	0.78	2.57
Stacking Model (RF+XGBoost)	Regional aggregation approach	-	4.78	0.68	3.55	-	5.07	0.64	3.59
	Grid-based approach	-	4.85	0.70	3.22	-	4.78	0.71	3.15
	Sector-based approach	-	3.57	0.81	2.39	-	3.36	0.83	2.29
	Simplified coherent clusters	-	3.67	0.77	2.52	-	3.71	0.77	2.54

Journal Pre-proofs

Hybrid XGBoost Optimized	Regional aggregation approach	-	4.30	0.74	3.47	-	4.88	0.67	3.58
	Grid-based approach	-	4.59	0.74	3.09	-	4.60	0.73	3.06
	Sector-based approach	-	3.26	0.84	2.28	-	3.35	0.83	2.30
	Simplified coherent clusters	-	3.32	0.81	2.49	-	3.42	0.80	2.50
Support Vector Machine	Regional aggregation approach	-	4.16	0.76	3.12	-	4.55	0.71	3.29
	Grid-based approach	-	4.66	0.73	3.02	-	4.66	0.73	2.94
	Sector-based approach	-	3.16	0.85	2.16	-	3.21	0.85	2.13
	Simplified coherent clusters	-	3.17	0.83	2.29	-	3.27	0.82	2.29
Artificial Neural Network	Regional aggregation approach	80	3.76	0.80	2.75	20	4.29	0.74	3.03
	Grid-based approach	20	4.58	0.74	3.00	20	4.51	0.74	2.88
	Sector-based approach	20	**3.11**	**0.86**	**2.04**	20	3.15	0.85	2.14

Journal Pre-proofs

	Simplified coherent clusters	110	3.39	0.81	2.40	20	3.50	0.79	2.38

Journal Pre-proofs

The sector-based approach achieved the lowest RMSE for all five algorithms and both temporal configurations (Table 5, Figure 5). The best overall result was obtained by the ANN with the sector-based configuration and H1+H2 inputs (RMSE = 3.11 μg/m³, R^2 = 0.86, MAE = 2.04 μg/m³), followed by the SVM (RMSE = 3.16 μg/m³, R^2 = 0.85) and the Hybrid XGBoost Optimized model (RMSE = 3.26 μg/m³, R^2 = 0.84) in the same configuration. The differences among these three models (≤ 0.15 μg/m³) are small compared with those among spatial configurations.

The simplified coherent clusters approach ranked second for every algorithm, with RMSE values 0.02-0.27 μg/m³ higher than those of the sector-based configuration (H1+H2).

By contrast, the grid-based approach produced the highest errors with H1+H2 inputs for all algorithms (RMSE = 4.58-4.97 μg/m³), whereas the regional approach, in which all monitoring stations were treated as a single homogeneous group, yielded intermediate errors (3.76-4.78 μg/m³). These comparisons should be interpreted considering the spatial support of the target: the regional target is the average of all stations and is therefore smoother than the cell averages of the grid-based configuration, which often rely on one or two stations. The advantage of the sector-based approach over the regional one was nevertheless obtained even though its targets are more local, and therefore more variable, than the regional mean.

Overall, the results reveal that the combination of environmentally coherent spatial aggregation and flexible nonlinear learners (ANN, SVM and gradient boosting) provides the most effective framework for estimating ground-level NO_2 concentrations from Sentinel-5P and ERA5-Land data across the study region.

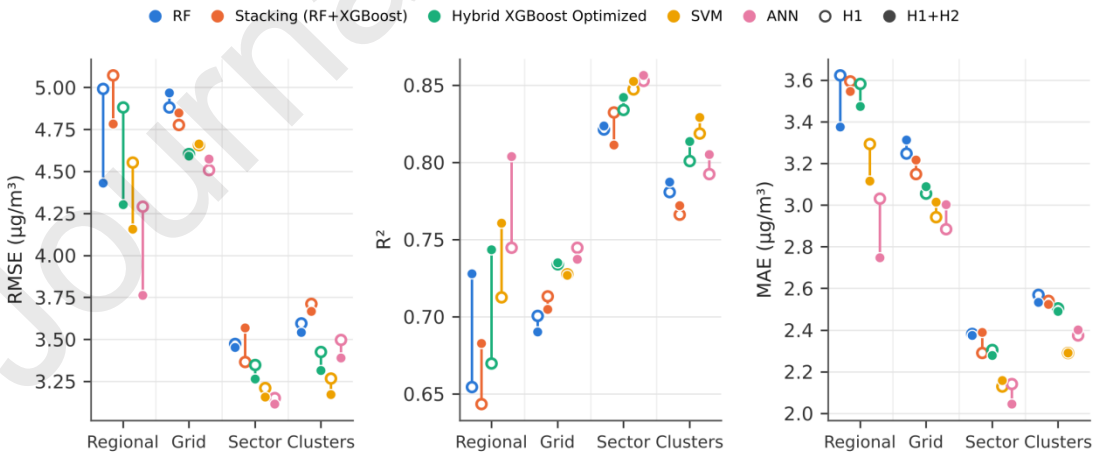


Figure 5.- Test-year (2024) RMSE, R^2 and MAE of the five models for each spatial configuration. Open circles: H1 inputs; filled circles: H1+H2 inputs; the vertical segment shows the change produced by the H2 window.

3.2 Best-performing configuration and performance by sector

The ANN combined with the sector-based approach and H1+H2 inputs achieved the best overall performance (Table 5). To assess the spatial robustness of the results and their dependence on terrain and environment, performance was computed independently for each sector and for all models (Table 6); the characteristics of the sectors are given in Table 3.

Table 6.- Sector-level performance of the five models in the sector-based approach (H1+H2 inputs, test year 2024). HXO: Hybrid XGBoost Optimized; nRMSE: RMSE divided by the mean observed concentration of the sector. The lowest RMSE of each sector is shown in bold; complete metrics for both temporal configurations are given in Table S3. Negative R^2 values indicate errors larger than the variance of the observations.

Sector	Environment	RMSE (µg/m³)					ANN		
		RF	Stacking	HXO	SVM	ANN	R^2	MAE	nRMSE (%)
1	West, rural	2.17	2.32	2.22	2.01	**1.84**	0.57	1.01	59
2	South, semi-rural	2.60	2.56	2.45	**2.38**	2.62	0.74	1.74	40
3	South-west metropolitan	3.87	4.03	**3.62**	3.63	3.89	0.82	2.59	37
4	Central urban core	5.24	5.46	4.88	4.84	**4.37**	0.86	3.23	27
5	North, mountain	2.54	2.33	2.20	**1.83**	2.06	0.39	1.44	38
6	East, suburban corridor	3.96	4.26	**3.76**	3.81	3.84	0.81	2.64	32
7	South-east, rural	2.76	2.82	2.84	2.43	**2.14**	-0.37	1.67	56
All		3.45	3.57	3.26	3.16	**3.11**	0.86	2.04	38

Absolute errors increased with the concentration level for all models, from 1.8-2.8 µg/m³ in the rural sectors 1, 2, 5 and 7 to 3.6-5.5 µg/m³ in the urban core (sector 4) and the metropolitan sectors 3 and 6. Relative errors, however, followed the opposite pattern: the nRMSE of the ANN was lowest in the urban core (27%) and the eastern corridor (32%) and highest in the rural sectors 1 (59%) and 7 (56%). The rural sectors combine low concentrations with a weaker column-surface coupling (r = 0.60-0.77, against 0.78-0.87 in the metropolitan sectors; Table 3), so that the satellite signal carries less information on the day-to-day surface variability; this also explains the low R^2 values in

sectors 5 (0.39) and 7 (-0.37). In sector 7, the day-to-day variability of the target was the smallest of the region (standard deviation of 1.8 µg/m³ in 2024), and all models overestimated its concentrations (ANN bias +1.3 µg/m³, +34%). The squared error therefore exceeded the variance of the observations, which yields a negative R^2, although the predictions still followed the daily variability (r = 0.59). The mountain sector 5, the highest of the region, showed intermediate relative errors (38%), similar to those of the flat southern sector 2 (40%), which indicates that elevation alone does not determine performance and that the intensity of local emissions relative to the regional background is the dominant factor. The ranking of the models was broadly consistent across sectors: the ANN or the SVM provided the lowest RMSE in five of the seven sectors, and the Hybrid XGBoost Optimized model in sectors 3 and 6, by margins of at most 0.05 µg/m³.

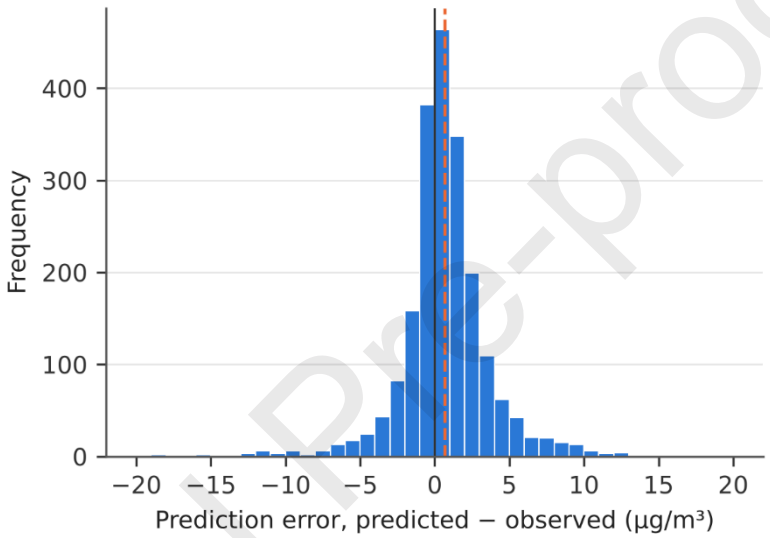


Figure 6.- Frequency distribution of prediction errors (ANN, sector-based approach, H1+H2, test year 2024). Solid line: zero error; dashed line: mean error.

Figure 6 presents the distribution of prediction errors obtained for this configuration. The distribution is centered slightly above zero (mean error +0.67 µg/m³, median +0.59 µg/m³), and 91% of the residuals lie within ±5 µg/m³. This small positive offset is shared by all models in the sector-based configuration (mean bias between +0.67 and +1.43 µg/m³) and is discussed in Section 4.4. There is a very small number of larger residuals in the tails, which reflects sporadic deviations likely associated with atypical meteorological conditions or short-lived emission episodes not fully captured by the predictors.

Journal Pre-proofs

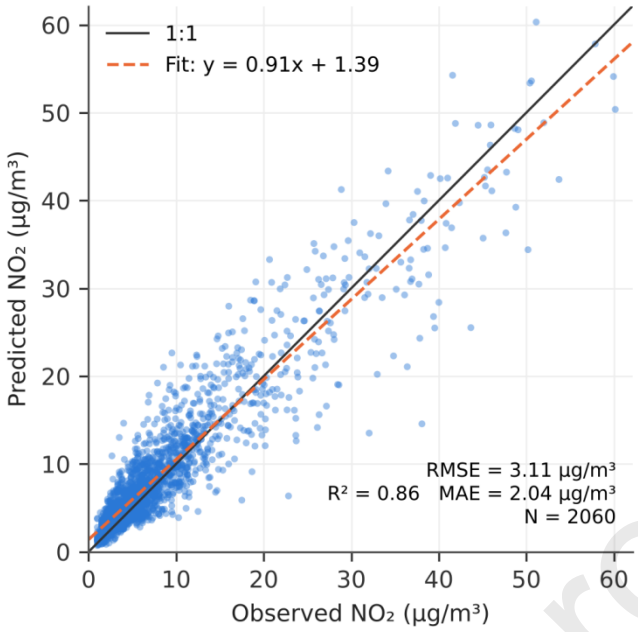


Figure 7.- Scatter plot of observed vs. predicted NO_2 concentrations (ANN, sector-based approach, H1+H2, test year 2024).

Figure 7 shows the scatter plot of observed versus predicted NO_2 concentrations. Predicted values display a strong linear association with the observations (r = 0.93; regression slope of 0.91), with most points close to the 1:1 line at low and moderate concentrations (0-15 µg/m³), which constitute 88% of the test samples. Below 15 µg/m³ the model slightly overestimates (mean error +0.9 µg/m³), whereas above 30 µg/m³ (3% of the samples) it tends to underestimate (mean error -3.5 µg/m³).

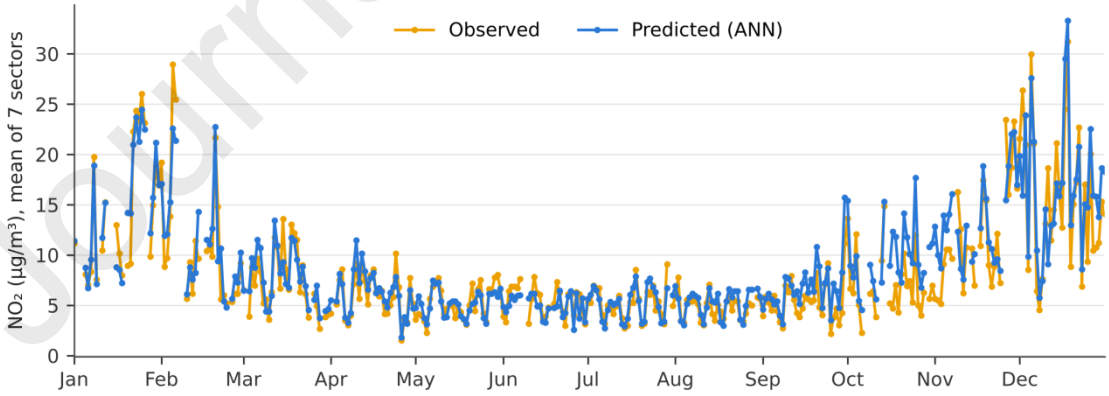


Figure 8.- Observed (orange) vs. predicted (blue) NO_2 concentrations over the year 2024 (ANN, sector-based approach, H1+H2; daily mean of the seven sectors).

The time series in Figure 8 shows the daily mean of the seven sectors over the whole 2024 test year. The model successfully reproduces the main temporal dynamics of

surface-level NO_2, capturing both day-to-day variability and the broader seasonal intensification toward the colder months, including the winter episodes of January-February and December. The main departure is a systematic overestimation in October (mean bias of +2.9 µg/m^3 in the sector-level data) and, to a lesser extent, in September and November (+1.4 and +1.5 µg/m^3), which is the origin of the larger autumn errors described in Section 3.3.

3.3 Seasonal performance

Performance was further disaggregated by season (DJF, MAM, JJA and SON) for all models and spatial configurations (Figure 9; the complete set of metrics for both temporal configurations is given in Table S2). The seasonal pattern was consistent across models and configurations: RMSE was highest in winter (4.3-5.2 µg/m^3 in the sector-based configuration) and autumn (3.8-4.1 µg/m^3), and lowest in spring (1.9-2.4 µg/m^3) and summer (1.6-1.9 µg/m^3). This pattern largely follows the seasonal cycle of the concentrations themselves (mean observed values of 14.1, 6.3, 5.1 and 8.0 µg/m^3 in DJF, MAM, JJA and SON in the sector-based configuration). For the ANN, the relative error was similar in winter, spring and summer (nRMSE of 31%, 30% and 31%), and the explained variance was highest in winter (R^2 = 0.87), when synoptic variability and stagnation episodes produce the widest range of concentrations. Autumn was the most difficult season (nRMSE = 50%, R^2 = 0.71), owing to the positive bias of October noted above (mean bias of +1.85 µg/m^3 in SON).

No single model was the most accurate in every season. In the sector-based configuration, the SVM gave the lowest RMSE in winter and autumn (4.33 and 3.78 µg/m^3, against 4.36 and 3.98 µg/m^3 for the ANN), whereas the ANN was clearly the most accurate in spring and summer (1.89 and 1.59 µg/m^3, against at least 2.20 and 1.88 µg/m^3 for the other models), which explains its lower annual RMSE. In the regional approach, the ANN was markedly better than the other models in winter and autumn (5.02 and 3.95 µg/m^3, against at least 5.90 and 4.55 µg/m^3), whereas in the cluster-based configuration the SVM was the most accurate in both seasons. In summer, when concentrations are low and the boundary layer is deep, the differences among models were small in all configurations.

Journal Pre-proofs

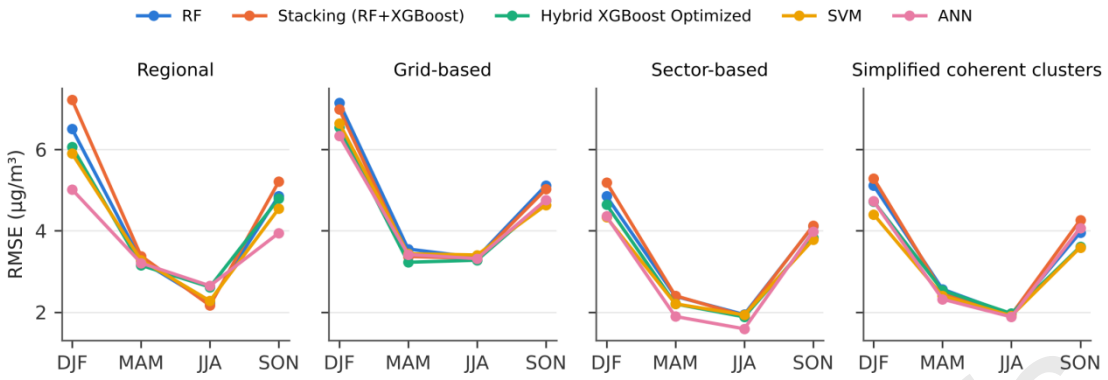


Figure 9.- Seasonal RMSE of the five models for each spatial configuration (H1+H2 inputs, test year 2024). DJF: December-February; MAM: March-May; JJA: June-August; SON: September-November.

3.4 Sensitivity analyses

Additional experiments with the sector-based configuration and H1+H2 inputs tested the methodological choices concerning feature screening, dimensionality reduction, input scaling and data partitioning, each modifying a single element with respect to a reference run (Table S4). Using all predictors instead of the screened subset, replacing the latter by principal components, or standardizing the inputs of the tree-based models changed the RMSE by at most 0.11 µg/m^3; only the ANN clearly benefited from the complete predictor set (+0.17 µg/m^3 when restricted to the screened subset). The partitioning scheme had the largest effect: on the 311 sector-days of 2024 included in the test sets of both partitions, a random 70/15/15 partition lowered the RMSE by 0.18-0.50 µg/m^3 (e.g., 2.84 against 3.34 µg/m^3 for the Hybrid XGBoost Optimized model), which quantifies the optimistic bias introduced when adjacent, strongly autocorrelated days are distributed among the training and test sets.

3.5 Performance during the COVID-19 lockdown

To evaluate the models under an abrupt change in emissions, the best configuration (ANN, sector-based, H1+H2) was retrained excluding the COVID-19 lockdown and de-escalation period in Spain (14 March to 21 June 2020), using the remaining days of 2020-2022 and 2024 for training and 2023 for validation, and was then applied to the excluded period (504 sector-days; Figure S2). The model reproduced the temporal variability of the observations (r = 0.81; RMSE = 2.39 µg/m^3; MAE = 1.77 µg/m^3) but overestimated the mean concentration by 0.95 µg/m^3 (+19%). The bias was concentrated in the strict lockdown (14 March to 3 May; +1.76 µg/m^3, +36%, r = 0.77) and almost disappeared during the de-escalation phases (4 May to 21 June; +0.40 µg/m^3, +8%, r = 0.87), as traffic progressively recovered. Surface NO_2 in Madrid decreased by about 62% during the lockdown (Baldasano, 2020), whereas tropospheric NO_2 columns over western Europe decreased by roughly 20-38% (Bauwens et al., 2020). Because the TROPOMI column is the dominant predictor, the model captured a substantial part of the reduction, but not all of it: the relationship between the column and the surface concentration learned under usual emissions does not fully hold when near-surface traffic emissions

collapse while regional and elevated sources persist. Estimates for periods of abrupt emission change should therefore be interpreted with caution.

4. DISCUSSION

4.1 Contextualizing performance with existing literature

The modeling performance obtained in this work is consistent with results recently reported for satellite-based estimation of ground-level NO_2 concentrations using machine learning approaches. Reported accuracies in the literature vary considerably depending on factors such as the geographical scale of the study, the type and number of predictor variables employed, and the validation methodology adopted (Table 7).

Table 7.- Summary of recent satellite-based surface NO_2 estimation studies.

Study	Region	Meteo inputs	Additional predictors	ML approach	Performance [µg/m³]
Cedeno Jimenez and Brovelli (2023)	Milan	ERA5	-	RF feature selection + Voting ensemble SVR + MLPR	RMSE: 5.53 - 7.24; R^2: 0.76
Shetty et al. (2024)	Europe	ERA5-Land; NCEP-CFSR	VIIRS Nightlights, NDVI	XGBoost	RMSE: 13.07; MAE: 7.77
Chan et al. (2021)	Germany	ERA5	EU-DEM; POLYPHEMUS CTM	ANN	RMSE: 6.32; R^2: 0.64; MAE: 4.76
Rosenda et al. (2025)	Mexico City	ERA5	-	RF	RMSE: 3.81; R^2: 0.972; MAE: 2.91
Yagmur Aydin (2025)	Istanbul	GEOS-CF	NDVI, VIIRS Nightlights, Road Length, Population Density, DEM	CatBoost	RMSE: 16.23, MAE: 11.75

Balamurugan et al. (2023)	Germany	ERA5	DEM, Road density, NDVI, EVI	XGBoost	RMSE: 4.77; R^2: 0.88
Li et al. (2022)	China	NCEP/GFS	NDVI, Latitude, Longitude	Random Forest	RMSE: 7.04; R^2: 0.78,
Sun et al. (2024)	Netherlands, Belgium, Luxembourg, France, and Western Germany	ERA5/ERA5-Land	Road Density, VIIRS Nightlights, Anthropogenic Emissions, Wildfires, Land Cover, DEM	XGBoost Model Ensemble	RMSE: 8.08; R^2: 0.64
Wei et al. (2022)	China	ERA5	Population, VIIRS Nightlights, NDVI, DEM	Deep Forest	RMSE: 4.89; R^2: 0.71

At the European scale, Shetty et al. (2024) combined TROPOMI NO_2 observations with meteorological variables derived from ERA5 and NCEP-CFSR, including planetary boundary layer height and wind-related parameters, within an XGBoost framework, reporting RMSE values of 13.07 μg/m^3 and a MAE of 7.77 μg/m^3 . Similarly, Chan et al. (2021) obtained RMSE values of 6.32 μg/m^3 over Germany using neural-network models driven by meteorological predictors, while Balamurugan et al. (2023) achieved RMSE values of 4.77 μg/m^3 and R^2 values ranging from 0.68 to 0.88 through gradient-boosting approaches incorporating atmospheric precursor emission data. In comparison, the best-performing configuration obtained in the present study (RMSE = 3.11 μg/m^3, R^2 = 0.86, MAE = 2.04 μg/m^3, obtained on an independent test year) falls at the favorable end of this range.

Nonetheless, direct comparison among studies is inherently challenging due to differences in spatial extent, temporal coverage, predictor selection, model architecture, and validation strategies. In particular, the choice of validation strategy can substantially affect the reported accuracy. Studies based on simple random train-test partitions frequently report very high performance metrics (e.g., $R^2 > 0.90$ in Rosenda et al. (2025)), whereas approaches relying on spatial cross-validation or temporally independent testing generally obtain more moderate but more realistic predictive skill (e.g., $R^2 \approx 0.71$ in Wei et al. (2022)). In the present study, model evaluation was conducted on an entire year (2024) that was not used for training or model selection, thereby providing a more robust assessment of the ability of the models to generalize under realistic operational scenarios. For the study region, the sensitivity analysis quantifies the effect of this choice: a random partition lowered the RMSE by 0.18-0.50 μg/m^3 on the same test samples (Table S4).

From a methodological perspective, the proposed framework aligns with a growing body of studies seeking to maximize spatial transferability while minimizing dependence on region-specific ancillary datasets. A particularly relevant reference is the work of Cedeno Jimenez and Brovelli (2023), who estimated daily ground-level NO_2 concentrations in Milan using a combination of TROPOMI observations and ERA5 meteorological variables, demonstrating that relatively simple predictor configurations can still achieve competitive performance in urban environments. In contrast to more data-intensive methodologies requiring detailed emission inventories, traffic-flow datasets, land-use descriptors, or outputs from chemical transport models (Sun et al., 2024; Balamurugan et al., 2023; Yagmur Aydin, 2025), the present approach relies mainly on globally available satellite and reanalysis products. This design choice aims to improve both the reproducibility and the potential transferability of the framework, particularly in regions where high-resolution local ancillary datasets are limited or unavailable.

An important methodological difference with respect to several previous regional studies is the use of ERA5-Land instead of the standard ERA5 reanalysis (Cedeno Jimenez and Brovelli, 2023; Rosenda et al., 2025). ERA5-Land provides the near-surface variables at a finer resolution than ERA5 (0.1° instead of 0.25°), which is relevant in the Community of Madrid, where topography and strong thermal contrasts influence atmospheric mixing and pollutant dispersion.

Two further methodological aspects differentiate the present work from most existing studies. The first is the systematic evaluation of dual temporal aggregation windows, combining meteorological conditions during the satellite overpass period with antecedent atmospheric states over the preceding 21 hours. This approach, explored in a more limited form by Cedeno Jimenez and Brovelli (2023), improved model performance in 15 of the 20 combinations of algorithm and spatial configuration, most clearly in the regional configuration, confirming that carry-over effects from nocturnal boundary-layer stability and early-morning pollutant accumulation provide relevant predictive information at the time of satellite observation. The second contribution concerns the explicit evaluation of multiple spatial stratification strategies, a dimension that remains largely underexplored in the existing literature (Section 4.3).

4.2 Comparative model performance

The three most accurate models of this study (ANN, SVM and Hybrid XGBoost Optimized) achieved similar accuracy in the sector-based configuration (RMSE = 3.11-3.26 $\mu g/m^3$), which indicates that, once the predictors and the spatial units are fixed, the choice among flexible nonlinear learners is less important than the design of the inputs. This is consistent with the absence of a universally superior algorithm in the literature, where gradient boosting (Balamurugan et al., 2023; Shetty et al., 2024), random forests (Li et al., 2022) and neural networks (Chan et al., 2021) have all been reported as the best-performing approach.

A particularly relevant comparison can be established with Morillas et al. (2026), which was conducted over the same study region and employed an equivalent methodological framework, but employing meteorological observations from ground-based weather stations only. In the original study, the ANN combined with the simplified coherent clusters configuration emerged as the best-performing model, achieving an RMSE of 2.44 μg/m³, an R^2 of 0.87, and a MAE of 1.61 μg/m³. These values were obtained with a 70/15/15 partition and are therefore not directly comparable with the year-based results of Table 5. Under a comparable random partition, the most accurate reanalysis-based model of the present study reached an RMSE of 2.84 μg/m^3 on the 2024 samples (Hybrid XGBoost Optimized, sector-based configuration; Section 3.4), i.e., about 0.4 μg/m^3 more than with station-based meteorology. This moderate loss of accuracy is the cost of replacing a dense local meteorological network by globally available reanalysis fields, and it is small in comparison with the gain in transferability.

Under the year-based design, the ANN was again the best-performing model, as in the station-based study, although in a different spatial configuration, whereas under the random partition the Hybrid XGBoost Optimized model ranked first (Table S4). Gradient-boosted trees may benefit particularly from near-duplicate days present in both the training and the test sets, which would explain their advantage under random partitioning; this result illustrates that the validation design can alter not only the absolute metrics but also the ranking of the algorithms.

The SVM model also exhibited strong and consistent performance across spatial configurations in both studies, confirming the robustness of kernel-based approaches for this type of estimation task. RF models, both as standalone predictors and as components of the hybrid framework, provided stable performance across all spatial configurations in both studies while offering the additional advantage of interpretable feature importance rankings, which were used to screen the predictors, although the sensitivity analysis showed that this screening has only a marginal effect on accuracy (Table S4).

4.3 Influence of spatial aggregation in model accuracy

The spatial configurations providing the best predictive performance were not the complete regional aggregation approach nor the regular grid-based strategy aligned with the ERA5-Land resolution. In general, the grid-based configuration failed to improve results relative to regional aggregation and produced the highest errors for all algorithms with H1+H2 inputs, indicating that purely geometric spatial partitioning is insufficient to capture the environmental heterogeneity controlling surface-level NO_2 variability across the Community of Madrid.

This pattern is consistent with our findings in Morillas et al. (2026), where a grid-based configuration defined according to the TROPOMI spatial resolution also underperformed relative to environmentally informed spatial stratifications. The persistence of this behavior across studies using different meteorological input sources suggests that

geometric partitioning alone cannot adequately substitute for environmentally coherent grouping of monitoring stations.

In the present study, the sector-based configuration produced the best overall results, whereas the previous study identified the simplified coherent clusters approach as the optimal strategy. This difference may be related to the coarser resolution of the reanalysis inputs: with station-based meteorology, each cluster was already characterized by locally measured conditions, whereas with ERA5-Land the finer sector subdivision allows each unit to be associated with its own reanalysis conditions, partly compensating for the loss of local meteorological detail. The sector-level analysis (Table 6) further shows that the relative accuracy is governed mainly by the strength of local emissions and of the column-surface coupling rather than by elevation alone.

Collectively, the consistent observation that environmentally informed spatial stratification clearly outperforms both uniform regional aggregation and purely geometric grid-based schemes provides valuable and transferable guidance for the development of satellite-based air quality estimation frameworks in heterogeneous urban environments. These findings indicate that spatial training units should be defined according to shared atmospheric behavior, emission patterns, and topographic conditions rather than arbitrary geometric divisions, regardless of whether meteorological predictors originate from ground-based monitoring networks or global reanalysis products.

4.4 Understanding prediction errors and model limitations

Three limitations emerge from the error analysis. First, concentrations above 30 μg/m^3 are underestimated by about 3.5 μg/m^3 on average (Figure 7), a behaviour common to regression models trained with a squared-error loss and accentuated by the coarse resolution of the predictors, which cannot represent the street-scale conditions that produce the highest peaks. Second, all models showed a positive mean bias in 2024 (+0.67 to +1.43 μg/m^3 in the sector-based configuration). The mean observed concentration at the overpass time decreased from 9.4 μg/m³ in 2023 to 8.2 μg/m^3 in 2024 (-13%), whereas the mean TROPOMI column decreased by only 6%; since the calendar year is not a predictor, the models cannot represent this downward trend in surface concentrations, which is consistent with continued reductions of traffic emissions in the region (MITERD, 2024). Periodic retraining with the most recent years would mitigate this drift in operational applications. Third, as shown by the COVID-19 experiment (Section 3.5), the relationships learned under usual emission conditions only partially hold during periods of abrupt emission change.

A better reproduction of peak concentration events is particularly relevant from a public health and regulatory perspective, as these episodes are precisely those that most frequently trigger exceedances of the legislative thresholds established by European air quality directives. Future research should explore the progressive incorporation of complementary predictors to further refine both overall accuracy and episode-level

performance, while preserving the globally transferable character of the satellite-reanalysis integration approach established in this and the preceding study.

5. CONCLUSIONS

This study validates and extends a methodology for estimating daily surface-level NO_2 concentrations across the Community of Madrid by integrating Sentinel-5P tropospheric column observations with ERA5-Land reanalysis meteorological fields and machine learning approaches. Building upon our previous study (Morillas et al., 2026) that employed the same modeling framework with ground-based meteorological station data, this work consolidates two key methodological findings: the overall benefit of incorporating antecedent atmospheric conditions through an extended dual temporal aggregation window, and the critical role of environmentally coherent spatial stratification of monitoring stations in improving predictive accuracy across a geographically and environmentally heterogeneous region.

The dual temporal aggregation strategy (i.e., combining conditions during the satellite overpass period with those of the preceding 21-hour interval) improved predictive performance in 15 of the 20 combinations of algorithm and spatial configuration, particularly in the regional configuration, confirming that this design choice remains beneficial when station-based meteorology is replaced by globally available reanalysis data. Similarly, the consistent superiority of the sector-based configuration, which yielded the lowest errors for all five algorithms, and of the coherent cluster configuration over purely geometric or region-wide spatial arrangements consolidates the finding that aligning the spatial structure of the training data with underlying environmental gradients is a transferable and effective pathway for improving model accuracy, independent of the meteorological data source employed.

The best-performing configuration (the Artificial Neural Network combined with the sector-based spatial classification) achieved on the independent 2024 test year an RMSE of 3.11 $\mu g/m^3$, an R^2 of 0.86, and a MAE of 2.04 $\mu g/m^3$, relying exclusively on Sentinel-5P observations and reanalysis fields (ERA5-Land and ERA5 boundary-layer height) without incorporating any auxiliary local datasets such as traffic counts, land-use classifications, or emission inventories. The SVM and the Hybrid XGBoost Optimized model achieved comparable accuracy (RMSE of 3.16 and 3.26 $\mu g/m^3$). These results compare favorably with studies that have employed considerably more data-intensive approaches. Relative errors were lowest in the urban core (nRMSE = 27%), where concentrations and population exposure are highest, and the larger absolute errors of winter and autumn reflected the seasonal cycle of concentrations rather than a loss of skill, since relative errors were similar in winter, spring and summer. The models also proved robust under the exceptional emission conditions of the COVID-19 lockdown, which were absent from the training data: they reproduced the day-to-day variability of the observations ($r = 0.81$) with an RMSE of 2.39 $\mu g/m^3$, and the moderate overestimation (+19%), concentrated in the strict lockdown phase, indicates that periodic retraining would further strengthen operational applications during abrupt emission changes. Under a comparable partition, replacing station-based meteorology with

reanalysis fields increased the RMSE by only about 0.4 μg/m³ with respect to the preceding study.

The most significant contribution of this work lies in the demonstration that accurate surface-level NO_2 estimation in a complex urban region can be achieved using exclusively globally available, freely accessible input data. The use of ERA5-Land reanalysis provides spatially continuous and temporally consistent meteorological information at uniform resolution across any geographic region, thereby removing the dependence on local meteorological monitoring infrastructure that could constrain the transferability of the original framework. As a result, the methodology proposed here offers substantially greater scalability and can, in principle, be implemented in any urban area covered by Sentinel-5P observations.

This transferability is of particular relevance in the Spanish regulatory context. The recently adopted European Directive (EU) 2024/2881 (European Commission, 2024) establishes a revised annual NO_2 limit of 20 μg/m^3 by 2030 (half the current threshold of 40 μg/m^3), which the Community of Madrid only achieved for the first time in 2022 (MITERD, 2025). Meeting this more stringent standard will require not only effective emission reduction policies but also robust and spatially representative monitoring systems capable of verifying compliance across entire urban regions. In this context, the framework proposed in this study offers a potentially valuable alternative by enabling reliable NO_2 estimation exclusively from globally available satellite and reanalysis products. Consequently, the methodology could support the expansion of air quality surveillance capabilities in underserved municipalities, contributing to more spatially equitable environmental monitoring and better-informed policy implementation during a critical period for European air quality regulation.

Several challenges must nonetheless be addressed to consolidate the operational potential of this approach. The spatial resolution of ERA5-Land, at approximately 9 km, remains insufficient to fully resolve the fine-scale meteorological gradients that arise within densely built urban environments, where exposure variability can occur at sub-kilometre scales. This constrains the model's capacity to capture sharp, localized concentration peaks driven by street-level wind channelling or spatially concentrated emission sources. Additionally, the reliance on cloud-free satellite observations introduces temporal discontinuities that could compromise the continuity of operational deployments, particularly during winter months when adverse meteorological conditions simultaneously increase NO_2 concentrations and reduce TROPOMI data availability. Finally, the sectors were defined by expert judgement; an objective, data-driven delineation based on globally available descriptors would facilitate the transfer of the framework to other regions.

The development of gap-filling strategies and the upcoming data products of the Sentinel-4 geostationary mission, which will provide hourly observations over Europe represent the most promising avenues for reducing the temporal gaps of the satellite record while preserving the globally transferable and infrastructure-independent character that constitutes the core methodological contribution of this work.

Acknowledgements

This work was supported by:

Project TED2021-130457A-I00, funded by MCIN/AEI/10.13039/501100011033 and European Union “NextGenerationEU”/PRTR".

DIGITWINS4CIUE (project co-funded by the European Health and Digital Executive Agency, HADEA, under the Digital Europe Programme, GA 101084054).

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 2623–2631. https://doi.org/10.1145/3292500.3330701

Alduchov, O.A., Eskridge, R.E., 1996. Improved Magnus Form Approximation of Saturation Vapor Pressure. JApMe 35, 601–609. https://doi.org/10.1175/1520-0450(1996)035

Balamurugan, V., Chen, J., Wenzel, A., Keutsch, F.N., 2023. Spatiotemporal modeling of air pollutant concentrations in Germany using machine learning. Atmos. Chem. Phys 23, 10267–10285. https://doi.org/10.5194/acp-23-10267-2023

Baldasano, J.M., 2020. COVID-19 lockdown effects on air quality by NO2 in the cities of Barcelona and Madrid (Spain). Sci. Total Environ. 741, 140353. https://doi.org/10.1016/J.SCITOTENV.2020.140353

Bauwens, M., Compernolle, S., Stavrakou, T., Müller, J.F., van Gent, J., Eskes, H., Levelt, P.F., van der A, R., Veefkind, J.P., Vlietinck, J., Yu, H., Zehner, C., 2020. Impact of Coronavirus Outbreak on NO2 Pollution Assessed Using TROPOMI and OMI Observations. Geophys. Res. Lett. 47, e2020GL087978. https://doi.org/10.1029/2020GL087978

Beirle, S., Boersma, K.F., Platt, U., Lawrence, M.G., Wagner, T., 2011. Megacity emissions and lifetimes of nitrogen oxides probed from space. Science 333, 1737–1739. https://doi.org/10.1126/SCIENCE.1207824

Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for Hyper-Parameter Optimization. Adv. Neural Inf. Process. Syst. 24.

Borge, R., Artíñano, B., Yagüe, C., Gomez-Moreno, F.J., Saiz-Lopez, A., Sastre, M., Narros, A., García-Nieto, D., Benavent, N., Maqueda, G., Barreiro, M., de Andrés, J.M., Cristóbal, Á., 2018. Application of a short term air quality action plan in Madrid (Spain) under a high-pollution episode - Part I: Diagnostic and analysis from observations. Sci. Total Environ. 635, 1561–1573. https://doi.org/10.1016/J.SCITOTENV.2018.03.149

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324/METRICS

Cao, E.L., 2023. National ground-level NO2 predictions via satellite imagery driven convolutional neural networks. Front. Environ. Sci. 11, 1285471. https://doi.org/10.3389/FENVS.2023.1285471/BIBTEX

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 Sens. 2023, Vol. 15, Page 5400 15, 5400. https://doi.org/10.3390/RS15225400

Cedeno Jimenez, J.R., Pugliese Viloria, A. de J., Brovelli, M.A., 2023. Estimating Daily NO2 Ground Level Concentrations Using Sentinel-5P and Ground Sensor Meteorological Measurements. ISPRS Int. J. Geo-Information 12. https://doi.org/10.3390/IJGI12030107

Cersosimo, A., Serio, C., Masiello, G., 2020. TROPOMI NO2 Tropospheric Column Data: Regridding to 1 km Grid-Resolution and Assessment of their Consistency with In Situ Surface Observations. Remote Sens. 2020, Vol. 12, Page 2212 12, 2212. https://doi.org/10.3390/RS12142212

Chan, K.L., Khorsandi, E., Liu, S., Baier, F., Valks, P., 2021. Estimation of Surface NO2 Concentrations over Germany from TROPOMI Satellite Observations Using a Machine Learning Method. Remote Sens. 2021, Vol. 13, Page 969 13, 969. https://doi.org/10.3390/RS13050969

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 13-17-August-2016, 785–794. https://doi.org/10.1145/2939672.2939785

City of Madrid, 2025. Hourly air quality data since 2001 [WWW Document]. URL https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=f3c0f7d512273410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default (accessed 11.20.25).

Community of Madrid, 2025. Hourly air quality data since 2005 [WWW Document]. URL https://datos.comunidad.madrid/dataset/calidad_aire_datos_historico (accessed 11.20.25).

Das, B.P., Pathan, M.S., Lee, Y.H., Dev, S., 2021. Estimating Ground-level Nitrogen Dioxide Concentration from Satellite Data. Prog. Electromagn. Res. Symp. 2021-November, 1176–1182. https://doi.org/10.1109/PIERS53385.2021.9694752

Deng, F., Chen, Y., Li, L., Wang, C., Cao, L., Peng, C., 2022. Estimation of near-surface NO2based on TROPOMI and provincial control stations' data using machine learning. 2022 5th World Conf. Mech. Eng. Intell. Manuf. WCMEIM 2022 639–644.

https://doi.org/10.1109/WCMEIM56910.2022.10021549

European Commission, 2024. Directive (EU) 2024/2881 of the European Parliament and of the Council of 23 October 2024 on ambient air quality and cleaner air for Europe.

European Commission, 2008. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe.

Garcia Santos, D., Parés, M.E., 2025. Assessing nitrogen dioxide monitoring techniques: a comparative analysis of Sentinel-5 Precursor satellite and ground measurements in Catalonia. Atmos. Meas. Tech. 18, 5569–5590. https://doi.org/10.5194/AMT-18-5569-2025

Genuer, R., Poggi, J.M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recognit. Lett. 31, 2225–2236. https://doi.org/10.1016/J.PATREC.2010.03.014

Goldberg, D.L., Anenberg, S.C., Kerr, G.H., Mohegh, A., Lu, Z., Streets, D.G., 2021. TROPOMI NO2 in the United States: A Detailed Look at the Annual Averages, Weekly Cycles, Effects of Temperature, and Correlation With Surface NO2 Concentrations. Earth's Futur. 9, e2020EF001665. https://doi.org/10.1029/2020EF001665

Google Earth Engine, (GEE), 2025. Sentinel-5P OFFL NO2: Offline Nitrogen Dioxide [WWW Document]. URL https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_NO2?hl=es-419#description (accessed 11.20.25).

Government of Spain, 2026. Air Quality Stations [WWW Document]. URL https://www.miteco.gob.es/es/cartografia-y-sig/ide/descargas/calidad-y-evaluacion-ambiental/estaciones-de-calidad-del-aire.html (accessed 10.11.26).

Government of Spain, 2021. Ley 7/2021, de 20 de mayo, de cambio climático y transición energética.

Griffin, D., Zhao, X., McLinden, C.A., Boersma, F., Bourassa, A., Dammers, E., Degenstein, D., Eskes, H., Fehr, L., Fioletov, V., Hayden, K., Kharol, S.K., Li, S.M., Makar, P., Martin, R. V., Mihele, C., Mittermeier, R.L., Krotkov, N., Sneep, M., Lamsal, L.N., Linden, M. ter, Geffen, J. van, Veefkind, P., Wolde, M., 2019. High-Resolution Mapping of Nitrogen Dioxide With TROPOMI: First Results and Validation Over the Canadian Oil Sands. Geophys. Res. Lett. 46, 1049–1060. https://doi.org/10.1029/2018GL081095

Grzybowski, P.T., Markowicz, K.M., Musiał, J.P., 2023. Estimations of the Ground-Level NO2 Concentrations Based on the Sentinel-5P NO2 Tropospheric Column Number Density Product. Remote Sens. 15. https://doi.org/10.3390/RS15020378

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer Series in Statistics. https://doi.org/10.1007/978-0-387-84858-7

He, S., Dong, H., Zhang, Z., Yuan, Y., 2022. An Ensemble Model-Based Estimation of Nitrogen Dioxide in a Southeastern Coastal Region of China. Remote Sens. 14, 2807. https://doi.org/10.3390/RS14122807/S1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S.,

Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.N., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049. https://doi.org/10.1002/QJ.3803

Jeong, U., Hong, H., 2021. Assessment of Tropospheric Concentrations of NO2 from the TROPOMI/Sentinel-5 Precursor for the Estimation of Long-Term Exposure to Surface NO2 over South Korea. Remote Sens. 2021, Vol. 13, Page 1877 13, 1877. https://doi.org/10.3390/RS13101877

Jin, H., Song, Q., Hu, X., 2019. Auto-keras: An efficient neural architecture search system. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 1946–1956. https://doi.org/10.1145/3292500.3330648/SUPPL_FILE/P1946-JIN.MP4

Jollife, I.T., Cadima, J., 2016. Principal component analysis: A review and recent developments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374. https://doi.org/10.1098/RSTA.2015.0202/115142

Jonson, J.E., Borken-Kleefeld, J., Simpson, D., Nyíri, A., Posch, M., Heyes, C., 2017. Impact of excess NOx emissions from diesel cars on air quality, public health and eutrophication in Europe. Environ. Res. Lett. 12, 094017. https://doi.org/10.1088/1748-9326/AA8850

Khomenko, S., Cirach, M., Pereira-Barboza, E., Mueller, N., Barrera-Gómez, J., Rojas-Rueda, D., de Hoogh, K., Hoek, G., Nieuwenhuijsen, M., 2021. Premature mortality due to air pollution in European cities: a health impact assessment. Lancet Planet. Heal. 5, e121–e134. https://doi.org/10.1016/S2542-5196(20)30272-2

Khomenko, S., Pisoni, E., Thunis, P., Bessagnet, B., Cirach, M., Iungman, T., Barboza, E.P., Khreis, H., Mueller, N., Tonne, C., de Hoogh, K., Hoek, G., Chowdhury, S., Lelieveld, J., Nieuwenhuijsen, M., 2023. Spatial and sector-specific contributions of emissions to ambient air pollution and mortality in European cities: a health impact assessment. Lancet Public Heal. 8, e546–e558. https://doi.org/10.1016/S2468-2667(23)00106-8

Kim, M., Brunner, D., Kuhlmann, G., 2021. Importance of satellite observations for high-resolution mapping of near-surface NO 2 by machine learning. Remote Sens. Environ. 264, 34–4257. https://doi.org/10.1016/j.rse.2021.112573

Kwon, H., Cho, N., 2023. Corrosion Behaviors of Outdoor Bronze Sculptures in an Urban–Industrial Environment: Corrosion Experiment on Artificial Sulfide Patina. Met. 2023, Vol. 13, Page 1101 13, 1101. https://doi.org/10.3390/MET13061101

Li, M., Wu, Y., Bao, Y., Liu, B., Petropoulos, G.P., 2022. Near-Surface NO2 Concentration Estimation by Random Forest Modeling and Sentinel-5P and Ancillary Data. Remote Sens. 2022, Vol. 14, Page 3612 14, 3612. https://doi.org/10.3390/RS14153612

McDuffie, E.E., Smith, S.J., O'Rourke, P., Tibrewal, K., Venkataraman, C., Marais, E.A., Zheng, B., Crippa, M., Brauer, M., Martin, R. V., 2020. A global anthropogenic emission inventory of atmospheric pollutants from sector- And fuel-specific sources (1970-2017): An application of the Community Emissions Data System (CEDS).

Earth Syst. Sci. Data 12, 3413–3442. https://doi.org/10.5194/ESSD-12-3413-2020

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ. Model. Softw. 101, 1–9. https://doi.org/10.1016/J.ENVSOFT.2017.12.001

Ministerio de Transición Ecológica y Reto Demográfico (MITERD), 2025. Evaluación de la Calidad del Aire en España.

MITERD, 2024. Tendencias de la calidad del aire en España 2021-2023.

Morillas, C., Alvarez, S., Pires, J.C.M., Garcia, A.J., Martinez, S., 2024a. Impact of the implementation of Madrid's low emission zone on NO2 concentration using Sentinel-5P/TROPOMI data. Atmos. Environ. 320, 120326. https://doi.org/10.1016/J.ATMOSENV.2024.120326

Morillas, C., Álvarez, S., Pires, J.C.M., García, A.J., Martínez, S., 2025. Linking Satellite and Ground Observations of NO2 in Spanish Cities: Influence of Meteorology and O3. Nitrogen (Switzerland) 6, 32. https://doi.org/10.3390/NITROGEN6020032/S1

Morillas, C., Alvarez, S., Serio, C., Masiello, G., Martinez, S., 2024b. TROPOMI NO2 Sentinel-5P data in the Community of Madrid: A detailed consistency analysis with in situ surface observations. Remote Sens. Appl. Soc. Environ. 33, 101083. https://doi.org/10.1016/J.RSASE.2023.101083

Morillas, C., Giosa, R., Alvarez, S., Serio, C., Masiello, G., Liuzzi, G., Aviles, C., Martinez, S., 2026. Estimating surface-level NO2 concentrations in the Madrid region using Sentinel-5P observations and ground-based meteorological data with machine learning approaches. Atmos. Environ. 373, 121904. https://doi.org/10.1016/J.ATMOSENV.2026.121904

Muñoz Sabater, J., 2019. ERA5-Land hourly data from 1950 to present. https://doi.org/10.24381/cds.e2161bac

Naseer, E., Basit, A., Bhatti, M.K., Siddique, M.A., 2022. Machine Learning for Area-Wide Monitoring of Surface Level Concentration of NO2Using Remote Sensing Data. 2022 Int. Conf. Emerg. Trends Electr. Control. Telecommun. Eng. ETECTE 2022 - Proc. https://doi.org/10.1109/ETECTE55893.2022.10007417

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Rahman, M.M., Shuo, W., Zhao, W., Xu, X., Zhang, W., Arshad, A., 2022. Investigating the Relationship between Air Pollutants and Meteorological Parameters Using Satellite Data over Bangladesh. Remote Sens. 14, 2757. https://doi.org/10.3390/RS14122757/S1

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography (Cop.). 40, 913–929.

https://doi.org/10.1111/ECOG.02881

Rosenda, Y., Herrera, M., Polanco Gaytán, M., Teodoro, R., Santos, A., Saheer, L.B., Mendoza-Cano, O., Macedo-Barragán, R.J., 2025. Estimating Surface NO2 in Mexico City Using Sentinel-5P and Machine Learning. Atmos. 2026, Vol. 17, Page 37 17, 37. https://doi.org/10.3390/ATMOS17010037

Rudke, A.P., Martins, J.A., Hallak, R., Martins, L.D., de Almeida, D.S., Beal, A., Freitas, E.D., Andrade, M.F., Koutrakis, P., Albuquerque, T.T.A., 2023. Evaluating TROPOMI and MODIS performance to capture the dynamic of air pollution in São Paulo state: A case study during the COVID-19 outbreak. Remote Sens. Environ. 289, 113514. https://doi.org/10.1016/J.RSE.2023.113514

Seibert, P., Beyrich, F., Gryning, S.E., Joffre, S., Rasmussen, A., Tercier, P., 2000. Review and intercomparison of operational methods for the determination of the mixing height. Atmos. Environ. 34, 1001–1027. https://doi.org/10.1016/S1352-2310(99)00349-0

Shah, V., J. Jacob, D., Li, K., Silvern, R., Zhai, S., Liu, M., Lin, J., Zhang, Q., 2020. Effect of changing NO*x* lifetime on the seasonality and long-term trends of satellite-observed tropospheric NO2 columns over China. Atmos. Chem. Phys. 20, 1483–1495. https://doi.org/10.5194/ACP-20-1483-2020

Shaw, S., Van Heyst, B., 2022. An Evaluation of Risk Ratios on Physical and Mental Health Correlations due to Increases in Ambient Nitrogen Oxide (NOx) Concentrations. Atmos. 2022, Vol. 13, Page 967 13, 967. https://doi.org/10.3390/ATMOS13060967

Shetty, S., Schneider, P., Stebel, K., David Hamer, P., Kylling, A., Koren Berntsen, T., 2024. Estimating surface NO2 concentrations over Europe using Sentinel-5P TROPOMI observations and Machine Learning. Remote Sens. Environ. 312, 114321. https://doi.org/10.1016/J.RSE.2024.114321

Sun, W., Tack, F., Clarisse, L., Schneider, R., Stavrakou, T., Van Roozendael, M., 2024. Inferring Surface NO2 Over Western Europe: A Machine Learning Approach With Uncertainty Quantification. J. Geophys. Res. Atmos. 129, e2023JD040676. https://doi.org/10.1029/2023JD040676

Van Geffen, J.H.G.M., Eskes, H.J., Boersma, K.F., Maasakkers, J.D., Veefkind, J.P., 2019. TROPOMI ATBD of the total and tropospheric NO2 data products.

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 Sens. Environ. 120, 70–83. https://doi.org/10.1016/J.RSE.2011.09.027

Wei, J., Liu, S., Li, Z., Liu, C., Qin, K., Liu, X., Pinker, R.T., Dickerson, R.R., Lin, J., Boersma, K.F., Sun, L., Li, R., Xue, W., Cui, Y., Zhang, C., Wang, J., 2022. Ground-Level NO2Surveillance from Space Across China for High Resolution Using Interpretable Spatiotemporally Weighted Artificial Intelligence. Environ. Sci. Technol. 56, 9988–9998. https://doi.org/10.1021/ACS.EST.2C03834/ASSET/IMAGES/LARGE/ES2C03834_

0006.JPEG

World Health Organization (WHO), 2021. WHO global air quality guidelines, Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide.

Xing, J., Baek, B.H., Li, S., Wang, C.T., Song, G., Ma, S., Zheng, S., Liu, C., Tong, D., Woo, J.H., Liu, T.Y., Fu, J.S., 2024. A Physically Constrained Deep-Learning Fusion Method for Estimating Surface NO2 Concentration from Satellite and Ground Monitors. Environ. Sci. Technol. 58, 21218–21228. https://doi.org/10.1021/ACS.EST.4C07341

Yagmur Aydin, N., 2025. Machine Learning-Based Ground-Level NO2 Estimation in Istanbul: A Comparative Analysis of Sentinel-5P and GEOS-CF. Appl. Sci. 2025, Vol. 15, Page 10997 15, 10997. https://doi.org/10.3390/APP152010997

Zhang, T., Zheng, B., Huang, R., 2025. Adaptive high-resolution mapping of air pollution with a novel implicit 3D representation approach. npj Clim. Atmos. Sci. 2025 81 8, 180-. https://doi.org/10.1038/s41612-025-01044-6

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proofs