
High-Resolution NO₂, O₃ and PMs Estimation in Puglia: Leveraging AI and Explainability Techniques

[Alessandro Fania](#)[†], [Giovanni Lorusso](#), [Marica De Lucia](#), [Roberto Cilli](#), [Nicola Amoroso](#)^{*}, [Maria Adamo](#), [Mariella Aquilino](#), Loredana Bellantuono, [Antonio Lacalamita](#), Marianna La Rocca, [Tommaso Maggipinto](#), [Angela Morabito](#), Alessandra Nocioni, [Ester Pantaleo](#)[†], Roberto Primerano, [Sabina Tangaro](#), [Roberto Bellotti](#)[‡], [Alfonso Monaco](#)[‡]

Posted Date: 19 December 2025

doi: 10.20944/preprints202512.1736.v1

Keywords: air pollution; remote sensing; machine learning; explainable artificial intelligence; one health



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

High-Resolution NO_2 , O_3 and $\text{PM}_{2.5}$ Estimation in Puglia: Leveraging AI and Explainability Techniques

Alessandro Fania ^{1,2,†}, Giovanni Lorusso ¹, Marica De Lucia ^{3,4}, Roberto Cilli ^{1,2}, Nicola Amoroso ^{2,5,*}, Maria Adamo ⁴, Mariella Aquilino ⁴, Loredana Bellantuono ^{2,6}, Antonio Lacalamita ^{1,2}, Marianna La Rocca ^{1,2}, Tommaso Maggipinto ^{1,2}, Angela Morabito ⁷, Alessandra Nocioni ⁷, Ester Pantaleo ^{1,2,†}, Roberto Primerano ⁷, Sabina Tangaro ^{2,8}, Roberto Bellotti ^{1,2,‡} and Alfonso Monaco ^{1,2,‡}

¹ Dipartimento Interateneo di Fisica M. Merlin, Università degli Studi di Bari Aldo Moro, Via G. Amendola 173, 70125, Bari, Italy

² Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Bari, Via A. Orabona 4, 70125, Bari, Italy

³ Department of Biology, University of Naples Federico II, Via Vicinale Cupa Cintia 21, 80126, Napoli, Italy

⁴ Institute of Atmospheric Pollution Research – National Research Council (IIA-CNR), Via A. Orabona 4, 70125, Bari, Italy

⁵ Dipartimento di Farmacia - Scienze del Farmaco, Università degli Studi di Bari Aldo Moro, Via A. Orabona 4, 70125, Bari, Italy

⁶ Dipartimento di Biomedicina Traslazionale e Neuroscienze (DiBraiN), Università degli Studi di Bari Aldo Moro, Piazza G. Cesare 11, 70124, Bari, Italy

⁷ Apulia Region Environmental Protection Agency (ARPA Puglia), C.so Trieste 27, 70126, Bari, Italy

⁸ Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, Via G. Amendola 165/a, 70126, Bari, Italy

* Correspondence: nicola.amoroso@uniba.it

† These authors contributed equally to this work.

‡ Co-last authors.

Highlights

What are the main findings?

- We produced 300-meter-resolution data-fusion (XGBoost) estimates of NO_2 , O_3 , PM_{10} , and $\text{PM}_{2.5}$ across the Apulia region in Southern Italy from 2019 to 2022.
- Our performance aligns with comparable studies; XAI (SHAP) shows the model reproduces known patterns, with meteorology, land-use/traffic proxies, Sentinel-5P columns, and AOD as principal drivers.

What is the implication of the main finding?

- The high-resolution exposure fields bridge monitoring gaps and enable policy and One-Health-relevant assessments; temporal validation remains robust after monthly/annual aggregation.
- The scalable approach aligns with the 2024 EU Air Quality Directive and is poised to exploit the NASA Multi-Angle Imager for Aerosols (MAIA) for improved particulate estimates.

Abstract

Air pollution remains a major environmental challenge, with severe impacts on human health and ecosystems. Recent advances in satellite technology have transformed air quality monitoring by enabling global, continuous observations of atmospheric pollutants. However, satellite data often lack the precision of ground-based stations. This study aims to develop a machine learning model to predict daily surface concentrations of key air pollutants (NO_2 , O_3 , PM_{10} , $\text{PM}_{2.5}$) at high spatial resolution (300 m) in the Apulia region. Using Regional Environmental Protection Agency (ARPA) station data from 2019 to 2022 and meteorological, geographic, land-use, and temporal variables, we trained an XGBoost model on a 300 m grid. Model performance, assessed by repeated cross-validation, showed an average R^2 of 0.71, with values of 0.77 for NO_2 , 0.78 for O_3 , 0.67 for $\text{PM}_{2.5}$, and 0.64 for PM_{10} . eXplainable AI (XAI) methods confirmed strong alignment with established scientific knowledge, enhancing model reliability and offering insights into pollutant distribution drivers.

Keywords: air pollution; remote sensing; machine learning; explainable artificial intelligence; one health

1. Introduction

Air pollution remains one of the biggest environmental challenges of the 21st century, with detrimental effects on human health and ecosystems. It is estimated to have caused 4.2 million premature deaths worldwide in 2019 [1]. The World Health Organization (WHO) has established air quality guidelines, recommending threshold concentrations for the major pollutants to minimize adverse health impacts, but many urban areas globally continue to exceed these limits, especially in densely populated and industrial regions. The main air pollutants include nitrogen dioxide (NO_2), ozone (O_3), and particulate matter (PM), especially $PM_{2.5}$ and PM_{10} . NO_2 , primarily emitted from traffic and industrial activities, contributes to respiratory stress and acts as a precursor to ozone and secondary particulate matter. Ground-level ozone, beneficial in the stratosphere, becomes harmful at the surface, where it exacerbates asthma and impairs lung function. Long-term exposure to these pollutants (especially NO_2 and O_3) has also been associated with neurodegenerative outcomes such as Alzheimer's, likely via oxidative stress mechanisms [2–4]. Moreover, particulate matter ($PM_{2.5}$ and PM_{10}) poses a significant risk due to its ability to penetrate deep into the respiratory tract. $PM_{2.5}$, with aerodynamic diameter below $2.5 \mu m$, can reach the alveoli and the bloodstream, promoting cardiovascular and pulmonary disease, while PM_{10} can still induce severe respiratory effects, especially in vulnerable populations [5]. Taken together, this evidence frames a concrete scientific requirement: exposure assessment must move beyond sparse point monitoring and provide spatially continuous pollutant fields with both high spatial resolution (hundreds of meters, i.e. neighborhood scale) and daily temporal coverage, so that vulnerable sub-populations can be studied in realistic conditions. Producing such maps in a physically consistent and interpretable way is currently one of the open challenges in atmospheric exposure science.

In recent years, Machine Learning (ML) and Deep Learning (DL) models have emerged as powerful tools to estimate air pollution where monitoring is sparse. Méndez et al. [6] reviewed 155 studies (2011–2021) and reported a dominant use of ML methods such as Support Vector Machines (SVM), Random Forests (RF), and boosted trees, as well as DL architectures (e.g. CNN–LSTM for $PM_{2.5}$ forecasting [7], GRU models for particulate in Beijing [8], and LSTM-based AQI predictors in smaller cities [9]). Typical inputs combine satellite retrievals (column NO_2/O_3 , Aerosol Optical Depth, AOD), reanalysis meteorology (e.g. ERA5), land-use/road/population layers, and sometimes chemical-transport model (CTM) outputs [10,11]. Recent models provide maps at 1–10 km resolution nationally, and in some urban cases down to ~ 1 km [10,11]; most target daily means, with hourly prediction remaining difficult due to stochastic variability and sparse ground truth [12,13]. Despite these advances, the existing literature still faces critical limitations that become especially important when one aims to use the resulting maps for health, regulatory or One Health applications. In fact, several issues recur: domain shift across years or regions, which degrades generalization [12]; sparse and uneven ground networks, which limit spatial cross-validation and leave rural areas under-represented [12]; reliance on static proxies (roads, land use, population) that do not fully capture day-to-day emission variability [13]; vertical representativeness and unit mismatches between satellite columns and true surface concentrations, with gaps due to clouds or retrieval limits [14]; propagation of CTM uncertainties from emission inventories and physics [15,16]; and limited interpretability and uncertainty quantification, which complicates policy and health applications [12,13,17]. In other words, most ML air quality studies either: offer high spatial detail but limited interpretability and uncertain temporal generalization, or ensure robustness at broad scales but lose the fine-scale gradients that drive within-city exposure differences. A key open scientific direction is therefore to build frameworks that integrate heterogeneous predictors (satellite, meteorology, land use, population, temporal cycles), resolve sub-kilometer spatial structure, and remain explainable enough to relate model behavior back

to known atmospheric processes such as photochemistry, dispersion and boundary-layer mixing. A central technical obstacle in that direction is the fusion of satellite products with ground-level concentrations. Ground-based stations provide accurate local concentrations in $\mu\text{g}/\text{m}^3$, but they are sparse and often missing in rural areas. Satellites such as Sentinel-5P provide near-daily, spatially extensive measurements of pollutants like NO_2 and O_3 , but these are reported as tropospheric columns in mol/m^2 , integrating the whole atmospheric column rather than strictly surface air. This introduces spatial smoothing and weakens direct correspondence with near-surface exposure [14]. This mismatch is not only a data availability problem: it is also a physical problem. Satellite tropospheric columns integrate contributions from the free troposphere and are only indirectly sensitive to boundary-layer pollution, whereas health-relevant exposure is driven by near-surface chemistry, emission plumes, and local dispersion. Therefore, a simple linear correlation between column retrievals and surface concentrations is generally weak, particularly for O_3 . In our view, this is exactly where machine learning is scientifically valuable, since non-linear ensemble models can learn interactions between satellite columns, meteorological structure (e.g. wind, mixing height, temperature), and land-use/emission proxies to infer ground-level fields even where no stations are present.

In this work we implement such a framework in a real, operationally relevant setting. We developed a machine learning pipeline to predict daily surface concentrations of NO_2 , O_3 , PM_{10} and $\text{PM}_{2.5}$ across the Apulia (Puglia) region in Southern Italy, at 300 m spatial resolution, using ground measurements from ARPA (Regional Environmental Protection Agency) stations as reference. Our model integrates Sentinel-5P column products; high-resolution meteorological variables from a regionally validated WRF configuration, downscaled and temporally aggregated; fine-scale land-use and anthropogenic indicators (road network density, building density, population); and temporal harmonics. We then apply explainable AI (XAI) via SHAP to interpret which predictors drive each pollutant and to link model behavior to known atmospheric processes. Our approach is not limited to “applying AI in a new geographical area”. Methodologically, it combines Sentinel-5P column products; high-resolution meteorological variables from a regionally validated WRF configuration, downscaled and temporally aggregated; fine-scale land-use and anthropogenic indicators (road network density, building density, population); and explicit explainability via SHAP to interpret which predictors drive each pollutant. This integration enables daily exposure maps at 300 m resolution, which is substantially finer than the 1–10 km scale that is typical for national products, while remaining interpretable in terms of known atmospheric drivers (e.g., NO_2 linked to traffic/industry; O_3 linked to temperature and photochemistry; PM linked to AOD and resuspension). We focused on Apulia because it concentrates densely populated coastal cities and major industrial/port areas (notably Bari and Taranto), including the Acciaierie d'Italia (ex ILVA) steelworks in Taranto, one of Europe's largest steel plants [18–20]. This setting combines intense industry, compact Mediterranean urban form with traffic-dominated NO_2 , strong land–sea breeze systems that modulate recirculation, and complex O_3 photochemistry under high insolation. To the best of our knowledge, there are currently no data-driven models producing daily, 300 m resolution surface maps of air pollution for Apulia. Previous efforts either target specific districts without fine-scale mapping [21], provide national-scale predictions at coarser administrative or grid resolutions using Sentinel-5P and ERA5 [22], or rely on chemical transport models (e.g. WRF–CMAQ, CAMx, WRF–Chem) for scenario analysis and source apportionment [15,16,23], which are computationally expensive and typically run at km-scale resolution. From a scientific perspective, Apulia is not only a case study but also a stress test for the broader field. The region combines intense industrial activity (steel production, port logistics), compact Mediterranean coastal cities with traffic-dominated NO_2 , strong land–sea breeze systems that modulate dispersion and recirculation, and complex O_3 photochemistry under high insolation. A model that can generate daily 300 m maps and explain its own predictions under these conditions provides evidence that high-resolution, explainable air quality estimation is feasible in other coastal industrial regions with limited monitoring. In summary, this study delivers a regional-scale, daily 300 m resolution exposure product for four key pollutants (NO_2 , O_3 , PM_{10} , $\text{PM}_{2.5}$) over Apulia, an assessment of its spatial

and temporal generalization, and an interpretation of pollutant drivers via SHAP. The overarching contribution is to demonstrate an interpretable satellite–meteorology–land-use fusion framework that is transferable to other Mediterranean and coastal industrial regions, rather than a location-specific technical exercise. In this analysis, we excluded sulfur dioxide (SO_2) and carbon monoxide (CO) from the prediction task due to their low ambient concentrations in the region [24,25].

2. Materials and Methods

Figure 1 shows the flowchart of our analysis. We implemented a spatial preprocessing procedure to interpolate missing data, improve spatial resolution, perform spatial and temporal aggregation of input variables, and reproject the data to the chosen reference system. Subsequently, ground-based measurements from ARPA monitoring stations, which serve as ground truth, have been integrated. After creating the final dataset, we developed a machine learning analysis using the XGBoost model in a 5-fold Cross Validation framework to generate predictive maps for the entire area of interest. Finally, we interpreted the model's decisions through an Explainable AI (XAI) analysis using the SHAP algorithm. All the variables used to build the Machine Learning model with the spatio-temporal resolution and the source of the predictors are shown in Table 1. These predictors were selected to jointly capture the main physical and socio-environmental drivers of surface pollution: emission intensity and urban form (population, roads, buildings, land cover), atmospheric transport and mixing/oxidation conditions (wind, stability proxies, humidity, clouds, precipitation, temperature, pressure), regional background loadings from column satellite products (NO_2 , O_3 , AOD), and recurrent temporal cycles (weekly/seasonal). Selection also prioritized variables with consistent regional coverage, minimal missingness, and the finest available resolution, while avoiding perfectly collinear duplicates.

Table 1. List of all variables used to build the AI model. The *Date* and *cell_id* are used only to merge and aggregate data.

| Name | Source / Product | Spatio-temporal resolution |
|----------------------------|-----------------------|---------------------------------------|
| U_wind_component | WRF | 4 km, hourly |
| V_wind_component | WRF | 4 km, hourly |
| Wind_speed | Computed (from WRF) | 4 km, hourly |
| Emissivity | WRF | 4 km, hourly |
| Rain_convective | WRF | 4 km, hourly |
| Rain_non_convective | WRF | 4 km, hourly |
| Vertical_wind | WRF | 4 km, hourly |
| Specific_humidity_2m | WRF | 4 km, hourly |
| Surface_temperature | WRF | 4 km, hourly |
| Cloud_fraction | WRF | 4 km, hourly |
| Surface_pressure | WRF | 4 km, hourly |
| AOD | Computed ¹ | 300 m, daily |
| NO2_sat | S5P (TROPOMI) | $5.5 \times 3.5 \text{ km}^2$, daily |
| O3_sat | S5P (TROPOMI) | $5.5 \times 3.5 \text{ km}^2$, daily |
| Sin_month, Cos_month | Computed | daily |
| Sin_week, Cos_week | Computed | daily |
| Population | WorldPop | 100 m (static) |
| Roads_tot | OSM | 100 m (static) |
| Buildings_tot | OSM | 100 m (static) |
| CORINE classes | CORINE | 100 m (static) |
| Continuous urban fabric | CORINE | 100 m (static) |
| Discontinuous urban fabric | CORINE | 100 m (static) |
| Cell_id | Computed | 300 m (grid) |
| Date acquired | Metadata | daily |

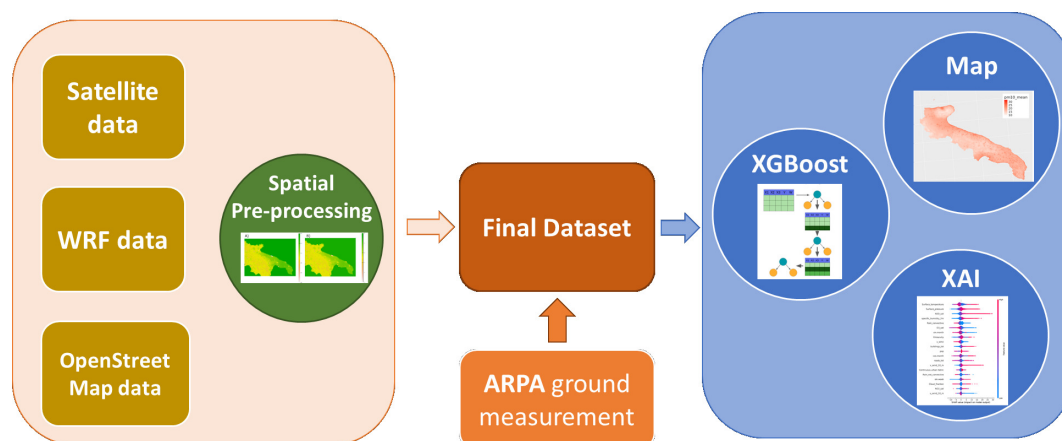


Figure 1. Flowchart of the analysis. After the collection of satellite, WRF and OpenStreetMap data and a spatial preprocessing phase, we implemented a Machine Learning model to predict ground-level daily air pollution over all Apulian region. Afterwards, we applied a XAI feature importance procedure to understand the role of each feature in the prediction.

2.1. Area of Interest

The area of interest is the Apulian region. This region is situated in the South-part of Italy, between approximately 41°E 53'N and 39°E 29'N latitude, and between 15°E 33' and 18°E 31' longitude. The region covers an area of about 19,541 square kilometers, making it one of Italy's larger regions. It is divided into six provinces: Bari, Barletta-Andria-Trani, Brindisi, Foggia, Lecce, and Taranto. From an environmental perspective, Puglia is rich in biodiversity but also faces significant challenges. One of the most critical environmental concerns centers around Taranto, an industrialized city that is home to one of Europe's largest steel plants, ILVA.

2.2. Data

2.2.1. ARPA's Ground Control Units

The ARPA (Regional Environmental Protection Agency) is a public Italian organization dedicated to environmental monitoring [26]. Each Italian region operates its own regional agency, totaling 20 agencies. ARPA is also involved in air quality monitoring, deploying various types of sensors across Italy. These sensors can be fixed or mobile and are situated in specific areas: urban, suburban, rural, industrial, and traffic zones. Utilizing different detection systems (electrochemical sensors, optical particle counters, ultraviolet photometric analyzers), ARPA provides hourly measurements of major atmospheric pollutants. These measurements, reported in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$), offer reliable estimates of human exposure to pollutants. The ARPA network in Puglia is fully compliant with national and European regulations, ensuring the necessary coverage and providing high quality data for air quality assessment. The new European Air Quality Directive, adopted on 23 October 2024 and published in the Official Journal of the European Union on 20 November 2024, does not promote an increase in the number of stations, but rather aims to optimise the existing network and strengthen the role of assessment modelling to achieve a more efficient analysis of air pollution [27]. For this purpose, starting from the ARPA stations in Puglia (52 for NO_2 , 17 for O_3 , 53 for PM_{10} , and 27 for $\text{PM}_{2.5}$) we developed a pipeline, based on machine learning applications, for comprehensive air quality assessment and modeling. In Figure 2 are shown the locations of all control units in Puglia.

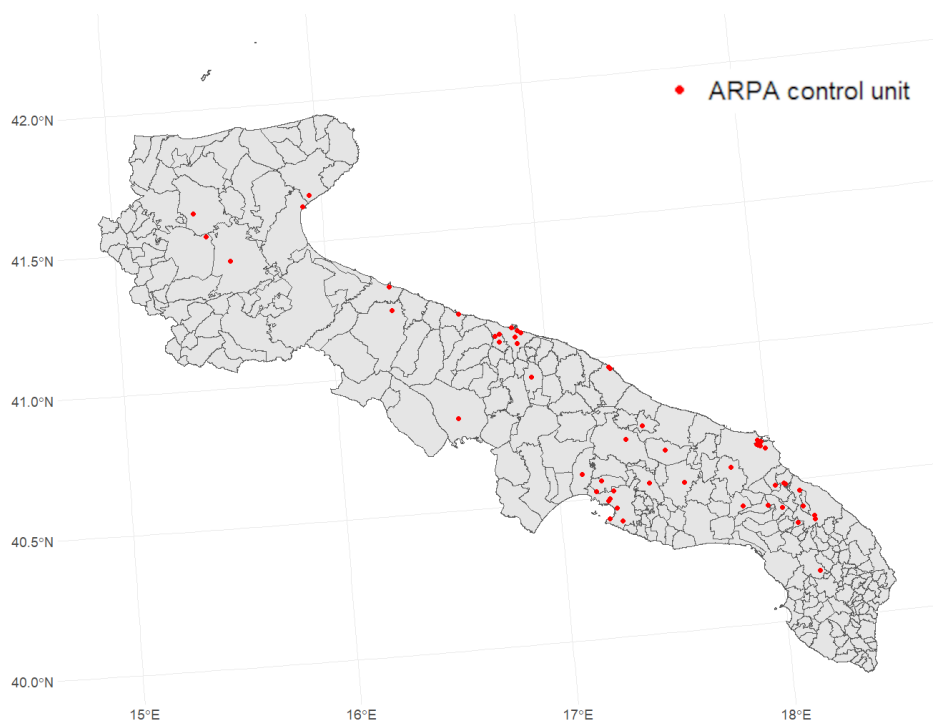


Figure 2. Map of all ARPA control units (red dots) located in the Apulian Region.

2.2.2. Sentinel-5P

Sentinel-5P (Sentinel-5 Precursor) is a satellite mission launched by the European Space Agency (ESA) in October 2017 as part of the Copernicus Earth observation program [28]. Its primary mission is to monitor atmospheric composition and provide critical data on air quality, climate, and the ozone layer. Sentinel-5P is equipped with the TROPOspheric Monitoring Instrument (TROPOMI), an advanced spectrometer that measures a wide range of atmospheric gases, including nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), carbon monoxide (CO), methane (CH_4), formaldehyde ($HCHO$), and aerosols, with a spatial resolution of up to 3.57 km^2 at nadir. The satellite provides near-daily global coverage, with its orbit ensuring that nearly every location on Earth is observed at least once per day.

In our work, we used the tropospheric NO_2 and O_3 variables provided by the Sentinel-5P dataset. The images were downloaded from the Google Earth Engine platform, with a daily temporal resolution. These variables represent the surface density of these pollutants (expressed in mol/m^2), from ground level up to the troposphere. Therefore, these measurements do not account for the effects of urban microclimates and are consequently different from ground-based measurements. They are a good indicator of the presence of atmospheric pollutants, but do not provide data as accurate as the measurements from control units.

Specifically, for O_3 and NO_2 we used the Sentinel-5P Level-3 tropospheric_o3 and tropospheric_no2 bands accessed via Google Earth Engine; given our data-driven approach, no additional vertical correction or profile adjustment was applied beyond projection to the study reference system. Specifically, we found a Pearson correlation coefficient of 0.28 between surface NO_2 and the tropospheric NO_2 column, and 0.10 between surface and tropospheric O_3 . Both dataset can be found at https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_NO2 and https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_O3.

2.2.3. Aerosol Optical Depth

Aerosol Optical Depth (AOD) is a satellite-derived parameter that quantifies the degree to which suspended particles in the atmosphere affect light transmission through absorption or scattering [29]. It serves as an indirect measure of aerosol concentration in the atmospheric column at a given time and location, and is commonly used to estimate particulate matter concentrations ($PM_{2.5}$) [30].

Numerous satellite products provide AOD measurements at varying spatial and temporal resolutions. However, high-resolution AOD retrievals are often subject to artifacts, cloud contamination, and regional biases [31]. To address these limitations and improve the accuracy of daily AOD predictions over Italy, we developed a calibrated model that integrates multiple satellite and reanalysis products as predictors. These include: 300 m Sentinel-3 SYN Aerosol Optical Depth at 550 nm [32], 1 km MODIS MAIAC AOD at 440 nm [33,34], $0.5^\circ \times 0.625^\circ$ MERRA2 AOD reanalysis at 440 nm [35], and $0.75^\circ \times 0.75^\circ$ CAMS AOD reanalysis at 440 nm [36]. We decided to use the AOD measurement at different wavelengths to add variability to the model and avoid highly correlated variables. In fact, these features could cause overfitting and strong fluctuations in model output and interpretability problems [37]. Auxiliary predictors, such as the 1 km MODIS Land Surface Temperature (LST) product, the Shuttle Radar Topography Mission (SRTM) Digital Elevation Model [38], ERA5 weather reanalysis variables (e.g., *T2m*, *wind10m*, etc.) [39], and geographical coordinates (latitude and longitude), were also included in the analysis.

For calibration, we relied on an extensive dataset of ground-based AOD measurements collected from 27 AERONET stations distributed across the Italian peninsula [40]. These measurements, spanning from January 2019 to December 2022, provided approximately 4,300 data points, which served as true labels for calibrating and evaluating satellite-derived AOD products. Our methodology was inspired by Aldabash et al. [41], allowing us to quantitatively assess and refine model performance using this robust observational dataset.

To integrate these observations with satellite and reanalysis data, we applied a stepwise calibration process based on an Ordinary Least Squares (OLS) linear regression model. After extensive testing, the best-fitting solution incorporated MODIS MAIAC, CAMS, and MERRA2 AOD products as predictors, achieving a high level of agreement with AERONET observations ($R^2 = 0.87$) under a leave-one-station-out validation framework. Further details regarding the analysis are provided in the supplementary materials. To ensure spatial consistency and comparability, all variables were resampled and projected onto a common 300-meter grid. AOD processed data used in this study are available under request.

2.2.4. Corine Data

The Corine Land Cover (CLC) dataset, developed by the European Environment Agency (EEA), provides comprehensive land use and land cover for all European countries. Based on satellite imagery, it classifies the landscape into over 40 categories, including urban areas, forests, and water bodies. The dataset has a spatial resolution of 100 meters, which is comparable with the spatial resolution of our model [42].

In our analysis, we used the 5 macro-categories of Corine (1-Artificial surfaces, 2-Agricultural areas, 3-Forest, 4-Wetlands, and 5-Water bodies) to assign land cover classes to our area of interest. Additionally, we calculated the percentage of industrial area occupancy per pixel to identify patterns related to pollutant emissions, particularly NO_2 . Oiamo et al. showed the influence of industrial area in the prediction of different pollutants, included NO_2 [43]. Data can be downloaded from https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_CORINE_V20_100m.

2.2.5. Weather Research and Forecasting Model

Meteorological inputs were produced using the Weather Research and Forecasting (WRF) model, configured over a single domain covering the Apulia region (latitude range 39.61° – 42.41° N; longitude range 14.38° – 18.97° E). The horizontal resolution was 4 km, with outputs stored at an hourly temporal resolution. The dataset spans the period from 31 December 2018, 12:00 UTC to 3 January 2023, 00:00 UTC. Each simulation run covered 181 hours; however, the first 12 hours of each run were discarded to allow for model spin-up and stabilization. Boundary and initial conditions were taken from the ECMWF analysis dataset (0.16° C horizontal resolution, 6-hour temporal resolution). Land cover information was provided by the CORINE Land Cover database.

Table 2 summarizes the WRF domain configuration and runtime settings adopted in this study.

Table 2. WRF model domain and run configuration used in this study.

| Item | Description |
|-------------------------------|---|
| Model version | WRF (operational setup used by ARPA Puglia) |
| Domain extent | Apulia region (39.61°–42.41° N; 14.38°–18.97° E) |
| Horizontal resolution | 4 km |
| Temporal resolution | Hourly output |
| Simulation period | 31 Dec 2018, 12:00 UTC – 3 Jan 2023, 00:00 UTC |
| Run length | 181 h per forecast cycle |
| Spin-up removal | First 12 h of each cycle discarded |
| Lateral/initial forcing | ECMWF analysis (0.16°C; 6-hourly) |
| Land cover dataset | CORINE Land Cover |
| Output variables | 3D atmospheric state (U, V, W, T, water species); surface fields (PSFC, T2, U10, V10, PBLH, radiation fluxes, precipitation); soil variables (soil temperature, soil moisture, runoff) |
| Post-processing for this work | Hourly → daily aggregation; vertical averaging of selected variables (e.g. vertical wind) to reduce redundancy; resampling to 300 m grid via cubic interpolation (Section 2.3.1); projection to the study CRS |

The physical parameterization schemes used in the WRF simulations are reported in Table 3. These choices follow the operational configuration validated for the Apulia region.

Table 3. WRF physical parameterizations adopted in this work. The right column reports the corresponding `namelist.input` options.

| Process | Scheme | WRF option |
|--------------------------------|---------------|-------------------------------------|
| Microphysics | Thompson | <code>mp_physics = 8</code> |
| Longwave radiation | RRTM | <code>ra_lw_physics = 1</code> |
| Shortwave radiation | Dudhia | <code>ra_sw_physics = 1</code> |
| Planetary Boundary Layer (PBL) | YSU | <code>bl_pbl_physics = 1</code> |
| Surface layer | Monin–Obukhov | <code>sf_sfclay_physics = 1</code> |
| Land surface model | Noah LSM | <code>sf_surface_physics = 2</code> |
| Cumulus parameterization | Kain–Fritsch | <code>cu_physics = 1</code> |

Validation. The adopted WRF configuration (including CORINE land cover and the physics schemes in Table 3) has been previously tested and validated over the Apulia region. In particular, Fedele et al. (2015) analyzed spatial biases in WRF simulations with focus on Apulia [44]; Fedele et al. (2015, EMS/ECAM) investigated how different PBL schemes and land cover classifications affect wind speed and temperature biases [45]; Tateo et al. (2017) evaluated ensemble simulations with alternative PBL schemes for wind speed and direction prediction [46]; and Tateo et al. (2018) assessed PBL height prediction using multi-physics ensembles [47]. These studies support the robustness of this setup, which is also used operationally by ARPA Puglia for forecasting and Wind Day alerts.

Finally, all WRF raster outputs were projected to the study reference CRS. The resampled daily fields were then integrated with Sentinel-5P, AOD, land-use (CORINE, OSM), WorldPop population density, and temporal harmonics to build the final machine learning dataset.

2.2.6. OpenStreetMap Data

OpenStreetMap (OSM) is a collaborative, open-source project that provides free and editable geographic data, created and maintained by a global community [48]. It offers detailed mapping information, including roads, buildings, land use, and natural features, making it a valuable resource for applications in geographic information systems (GIS), navigation, urban planning, and environmental monitoring [49].

Specifically, from the OSM website, we downloaded two shapefiles containing both the locations of all buildings and all roads, for the whole Apulia territory. From the first file, we calculated the total number of buildings within each considered grid cell, while from the second file, we computed the total length of roads intersecting each cell. Data can be found here: <https://download.geofabrik.de/europe/italy.html>.

2.2.7. Population Data

To make a model as general as possible, we added information regarding the urban effect. Specifically, we used the dataset provided by WorldPop [50]. This dataset, in GeoTIFF format, contains an estimate of the population in the Italian territory, with a grid resolution of 100 meters. The mapping method employed to build the dataset is Random Forests-based dasymetric redistribution technique developed by Stevens et al. [51]. We extracted values using our reference grid, applying the sum function as the aggregation method at 300 meters. Data are available at https://developers.google.com/earth-engine/datasets/catalog/WorldPop_GP_100m_pop

2.3. Preprocessing

2.3.1. Spatial Resample of WRF Data

The procedure of converting a sampled image from any coordinate structure to other structure is called image resampling. This is a complicated process that involves several geometric transformations of an image and can be summarised in three main steps:

- reconstruction of a continuous intensity surface from a discrete image;
- transformation of the continuous surface into the new coordinate system;
- sampling of the transformed surface to create a new discrete image.

In our analysis, we had to transform the WRF data structure into our prediction grid and increase the resolution of the model to 300 metres. The first step consisted of a spatial interpolation of the original raster image.

Among the three methods described in [52] we chose the Bi-Cubic Interpolation method which provides smoother results, especially compared to the Bi-Linear Interpolation. This method considers a 4x4 grid of surrounding pixels and performs a convolution with the kernel function given in the Equation 1 [53]. This allows for more accurate and natural-looking transitions between pixel values, reducing artifacts like blurring or jagged edges.

$$u(x) = \begin{cases} \frac{3}{2}|x|^3 - \frac{5}{2}|x|^2 + 1 & 0 < |x| < 1 \\ -\frac{1}{2}|x|^3 + \frac{5}{2}|x|^2 - 4|x| + 2 & 1 < |x| < 2 \\ 0 & |x| > 2 \end{cases} \quad (1)$$

The method is particularly used for continuous data, such as satellite imagery or elevation models [54], where maintaining smooth transitions is important, even if it can sometimes introduce slight overshooting where there are sharp changes in the data. The Supplementary Materials include example figures showing surface emissivity, as predicted by the WRF model, before and after applying the resampling method.

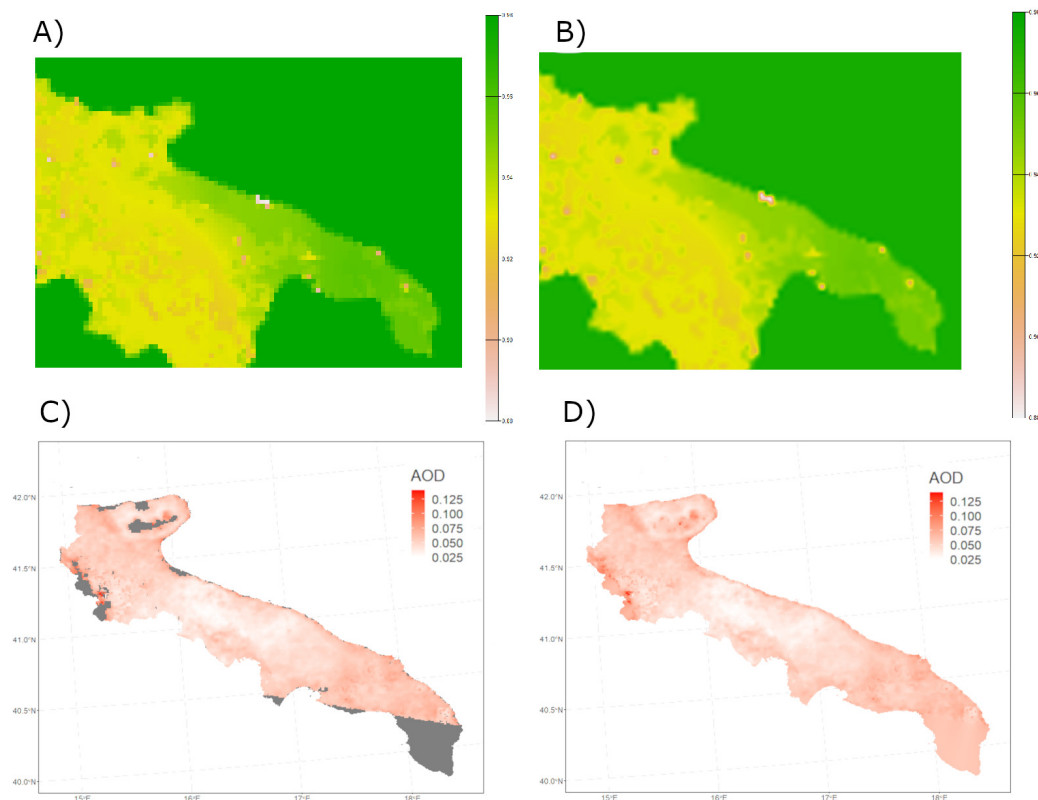


Figure 3. Top left panel (A) Surface Emissivity predicted by the WRF model (4 km resolution), on the top right panel (B) the resampled image to 300 m with bi-cubic interpolation method. Bottom left panel (panel C) AOD values extracted on the grid (the missing values are attributed to cloud covering and orbit of satellite), on the bottom right panel (panel D) the same image in which NA values are replaced using the Inverse Distance Interpolation method.

2.3.2. Spatial Interpolation for Missing Data

As we have described in the previous section, spatial interpolation is a statistical method used to estimate the values of a variable at unsampled locations based on the values observed at known points [55].

In this analysis, we filled the missing values of the input data by using an ensemble approach that combines the predictions obtained with several spatial interpolation methods [56]. This is essential for the Sentinel-5P and AOD data, which contain large gaps due to orbital constraints or cloud cover. The missing values are estimated by averaging the Inverse Distance Weighting (IDW) and the Nearest Neighbours (NN) interpolation methods. An example of this procedure's results is shown in the Supplementary Materials.

2.3.3. Inverse Distance Weighting

Inverse Distance Weighting (IDW) is a widely used spatial interpolation technique that estimates the value of a variable z at an unsampled location x_0 through a weighted average of values at sampled points x_i as shown in Equation 2 [56].

$$z(x_0) = \frac{\sum_{i=1}^n w_i z(x_i)}{\sum_{i=1}^n w_i} \quad (2)$$

The underlying assumption of the IDW method is that points closer to the location of the prediction have a greater influence than points further away. Therefore, the weight assigned to each sample point is inversely proportional to its distance from the prediction location. Furthermore, these distances can

be raised to a power parameter $\beta \geq 0$ that controls the rate of decline of influence with distance, as shown in Equation 3:

$$w_i = d_{Euclidean}^{-\beta}(x_0, x_i) \quad (3)$$

Typical values for β are 1 or 2, corresponding to an inverse or quadratic relationship respectively, in our analysis we have assigned $\beta = 1$.

2.3.4. Nearest Neighbors

The NN interpolation method is another technique in which the value of a variable z at an unsampled location x_0 is determined using the average of the k nearest points to that location, as shown in equation 4.

$$z(x_0) = \frac{\sum_{i=1}^k z(x_i)}{k} \quad (4)$$

This method is based on the assumption that only the points closest to the target location can be considered for the prediction of the variable value. In this case, we choose k to be equal to the rounded square root of the number of sampled cells.

2.4. ML Analysis

Nowadays, ML methods are growing in popularity and are used in many fields to capture the complex data dependencies. In our analysis, we compared the results obtained from two different ML algorithms: linear regression and eXtreme Gradient Boosting (XGBoost).

The XGBoost model is an ensemble method consistently used to study several problems thanks to its scalability in all scenarios[57]. The algorithm is faster than other machine learning techniques and scales to billions of examples in distributed or memory-limited settings [58].

XGBoost is composed of tree models that use a series of cumulative functions to predict the output of a certain task as shown in Equation 5:

$$\hat{y}_i = \Phi(\mathbf{x}_i) = \sum_{k=1}^K f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F} \quad (5)$$

where \mathcal{F} is the space of the Classification and Regression Trees (CART). The parameters of the model are optimized in order to minimize a regularized object which is a sum of the loss function l and a factor expressing the complexity of the model Ω as shown in Equation 6:

$$\mathcal{L}(\Phi) = \sum_{i=1}^N l(\hat{y}_i, y_i) + \sum_{k=1}^K \Omega(f_k) \quad (6)$$

Another important procedure that characterizes XGBoost model is shrinkage. This procedure scales newly added weights by a factor η after each step of tree boosting to reduce the influence of each individual tree and leaves space for future trees [59]. Furthermore, XGBoost implements feature sub-sampling during the training phase, which makes it robust to the over-fitting problem and speeds up the computations.

2.5. Performances Metrics

To assess the performances of our model we used three different metrics:

- the mean absolute percentage error (MAE), defined as

$$MAE = \frac{1}{n} \sum_{i=1}^N |y_i - \hat{y}_i|, \quad (7)$$

- the root mean square error (RMSE), defined as

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (8)$$

- the coefficient of determination between predicted and actual values

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \quad (9)$$

with y_i that represents the actual value, \hat{y}_i the predicted value and \bar{y} la media dei valori reali.

2.6. Explainable Artificial Intelligence: SHAP

Explainable Artificial Intelligence (XAI) refers to a variety of techniques that try to improve transparency and generalization to several AI models [60,61].

In particular, in our study we used the SHAP (SHapley Additive exPlanations) local explanation method to assess the contribution of each feature within the XGBoost model. Using a game theory approach [62,63], SHAP estimates the contribution to the prediction of each predictor, without knowing how the ML model works. An important characteristic of this technique is the ability to provide local explanation, instead of the average explanations given by the XGBoost model. Finally, SHAP provides a link between the values of the predictors, at the specific prediction, and its contribution to the final output. Those correlation are shown in the SHAP importance plot.

Mathematically, given all possible feature subsets F of the total feature set S ($F \subseteq S$) for a feature j the SHAP value is evaluated as the difference between two model outputs, the first obtained including that specific specific feature, the second without. The SHAP value of the j -th feature for the observation x is measured through the addition of the j -th feature to all possible subsets,

$$SHAP_j(x) = \sum_{F \subseteq S - \{j\}} \frac{|F|!(|S| - |F| - 1)!}{|S|!} [f_x(F \cup j) - f_x(F)], \quad (10)$$

where $|F|!$ is the number of feature permutations which precede the j -th feature; $(|S| - |F| - 1)!$ is the number of feature permutations that follow the j -th feature value; $|S|!$ represents the total number of permutations of features; $f_x(F)$ indicates the model prediction f for the sample x , considered a subset F without the j -th feature; $f_x(F \cup j)$ is the output of the same model including the j -th feature [62]. In our analysis we computed the mean SHAP values after a 5-fold CV, repeated 100 times for each considered year.

2.7. Software and Packages

The analyses were conducted using R version 4.4.1 (2024-06-14). The spatial analysis was carried out with the following packages: raster, terra, exactextractr, sf, stars, and gstat. Additionally, the analysis was parallelized using the doPar and foreach packages to accelerate data preprocessing. For the Machine Learning models, the packages XGBoost, Ranger, and caret were employed. The XAI analysis, on the other hand, was performed using Python 3.9.13 with the Spyder version 5 interface and the shap package.

3. Results

After the preprocessing phase, we developed two models for pollutant prediction. Specifically, we used linear regression as a benchmark model and an ensemble learning algorithm, XGBoost. As in our previous work, we found that the Random Forest model achieves similar performance to the XGBoost model but requires more computational time; therefore, we did not include it in the analysis [22].

Table 4 reports the performance metrics evaluated during the training phase.

Table 4. Random cross-validation (5-fold, repeated 10 times). Metrics are mean \pm SD across repeats. MAE and RMSE in $\mu\text{g m}^{-3}$. All p -values are significant at the 1% level.

| | MAE | RMSE | R ² |
|---------------------|--------------------------|--------------------------|-----------------|
| | ($\mu\text{g m}^{-3}$) | ($\mu\text{g m}^{-3}$) | |
| Linear model | | | |
| NO ₂ | 5.60 \pm 0.01 | 7.42 \pm 0.01 | 0.39 \pm 0.01 |
| O ₃ | 11.17 \pm 0.01 | 14.4 \pm 0.01 | 0.50 \pm 0.01 |
| PM _{2.5} | 4.36 \pm 0.01 | 5.94 \pm 0.01 | 0.18 \pm 0.01 |
| PM ₁₀ | 6.91 \pm 0.01 | 9.49 \pm 0.01 | 0.18 \pm 0.01 |
| XGBoost | | | |
| NO ₂ | 3.21 \pm 0.01 | 4.52 \pm 0.01 | 0.77 \pm 0.01 |
| O ₃ | 7.29 \pm 0.01 | 9.48 \pm 0.01 | 0.78 \pm 0.01 |
| PM _{2.5} | 2.66 \pm 0.01 | 3.74 \pm 0.01 | 0.67 \pm 0.01 |
| PM ₁₀ | 4.45 \pm 0.01 | 6.26 \pm 0.01 | 0.64 \pm 0.01 |

Table 4 shows that XGBoost substantially outperforms the linear model for all pollutants in random 5-fold CV repeated 10 times. For NO₂, R² increases from 0.39 (linear) to 0.77 (XGBoost), with a mean absolute error (MAE) reduction of about 2.4 $\mu\text{g}/\text{m}^3$. PM₁₀ and PM_{2.5} also benefit from the non-linear model, with R² rising from 0.18 (linear) to 0.64 and 0.67, respectively. This indicates that particulate matter fields are strongly driven by non-linear interactions among satellite-based AOD, meteorology, and land-use variables, which are not captured by a purely linear formulation. For O₃, the R² reaches 0.78, confirming that—despite the known weak linear relationship between tropospheric column O₃ and surface O₃—the model can still recover useful predictive structure once meteorology and temporal features are included. Overall, random CV demonstrates that tree-based fusion of satellite, meteorology, land-use and temporal harmonics provides accurate daily predictions at 300 m resolution.

Across pollutants, performance is differentiated. NO₂ and the particulate fractions (PM₁₀, PM_{2.5}) exhibit high accuracy and low RMSE, reflecting their strong link to localized emission patterns (traffic, industrial areas) and aerosol load, respectively. O₃ achieves similar apparent R² under random CV, but as we will discuss below, ozone behaves very differently under temporal generalization.

The results reported in Table 4 are obtained through a 5-fold cross-validation (CV) procedure repeated 10 times. After confirming that the XGBoost model performed better than the linear model, we predicted air quality (4 pollutants) for the entire territory of the Apulia region. Figure 4 shows the daily concentration of NO₂, O₃, PM₁₀, and PM_{2.5} predicted by means of our framework and averaged over the considered time interval.

Spatially, the NO₂ map highlights dense urban/industrial corridors, especially around Bari and Taranto, with sharp intra-urban gradients at the 300 m scale. Particulate matter (PM₁₀, PM_{2.5}) shows broader plumes, including higher values in coastal and harbor/industrial areas, consistent with resuspension, port activity, and industrial processes. In contrast, O₃ appears more spatially homogeneous at regional scale, with smoother inland gradients that reflect its secondary, photochemically formed nature and regional transport rather than purely local emissions. These spatial patterns are consistent with known atmospheric behavior of primary vs. secondary pollutants.

To better understand the reasoning behind the model's predictions, we applied the SHAP algorithm, which highlighted the most influential variables and their relationship with the model's output, as shown in Figure 5.

For NO₂, SHAP ranks land-use and anthropogenic predictors (road network density, built-up/industrial fabric from CORINE, population density) and meteorological dispersion variables (wind speed) among the top drivers. High road density and continuous urban fabric push predictions upward, pointing to traffic and industrial combustion as dominant local sources. Wind speed contributes with negative SHAP values at high intensity, consistent with dilution and advection away from sources.

Temperature shows a generally negative contribution at higher values, in line with the known seasonal cycle of NO_2 (winter accumulation under stable conditions vs. enhanced mixing and photochemical removal in summer).

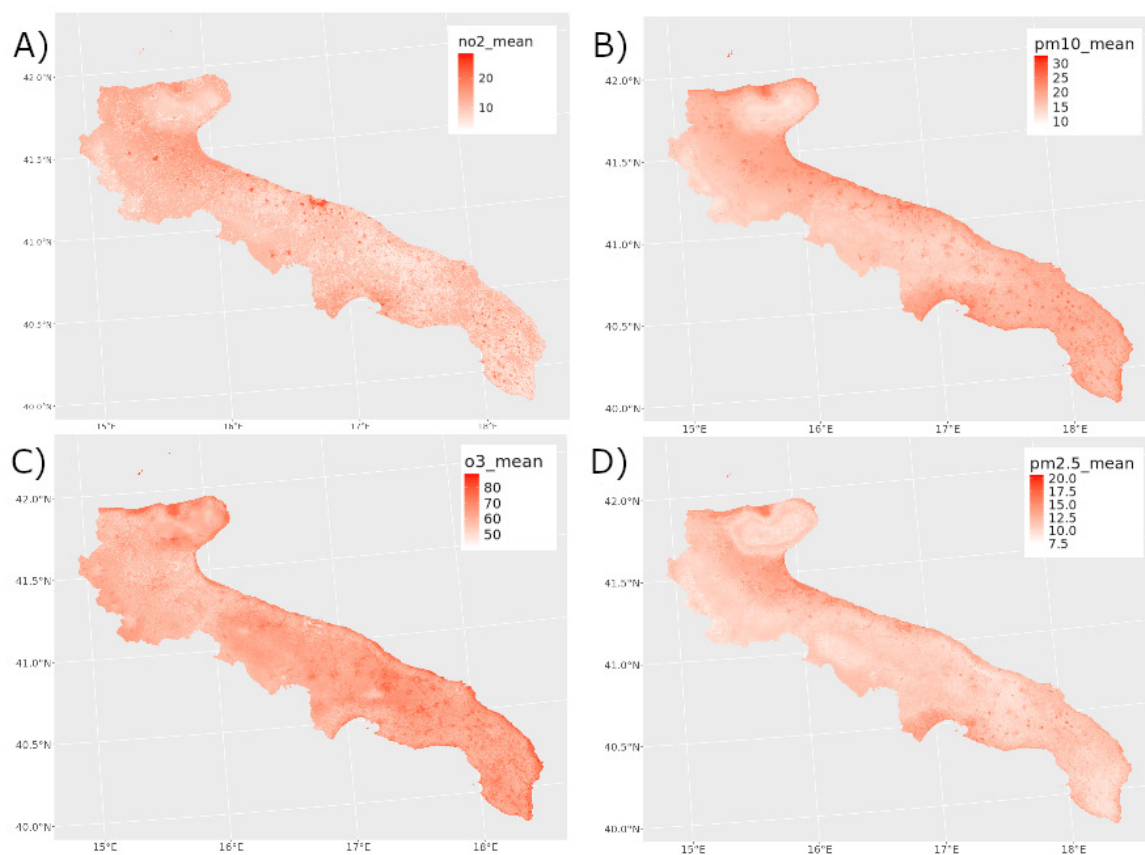


Figure 4. Predicted map of mean NO_2 (top left, panel A), PM_{10} (top right, panel B), O_3 (bottom left, panel C), and $PM_{2.5}$ (bottom right, panel D), using the XGBoost model.

For PM_{10} and $PM_{2.5}$, SHAP attributes a strong positive role to AOD, confirming that the calibrated multi-source AOD field encodes aerosol load relevant to surface PM. Meteorological variables such as wind and precipitation act in the expected direction: higher wind speed and rainfall tend to reduce predicted particulate levels through dispersion and wet scavenging. Land-use features (industrial/urban classes) also contribute positively, but their SHAP magnitude is generally lower than AOD and meteorology, reflecting the more regional nature of particulate burden.

For O_3 , SHAP identifies temperature and emissivity (a proxy linked to surface energy balance and thus photochemical activity) as strong positive drivers, consistent with O_3 being formed photochemically under warm, sunny, stagnant conditions. Wind speed shows non-trivial effects, suggesting both transport and recirculation. Importantly, the Sentinel-5P tropospheric O_3 column appears as a meaningful predictor despite the low raw Pearson correlation (0.10) with ground-level O_3 : SHAP indicates that the model leverages this satellite feature in interaction with meteorology and land cover, rather than as a simple linear proxy. This supports the view that non-linear ML is needed to extract surface-relevant information from column-integrated products.

We acknowledge that the raw Pearson correlation between the tropospheric O_3 column and surface O_3 is low ($R \approx 0.1$), but low marginal linear correlation does not imply that a variable lacks predictive value in non-linear, interaction-based ensembles such as XGBoost. In gradient boosting, trees combine features hierarchically through splits and interactions, so a predictor can become informative conditionally—e.g., above specific temperature thresholds, under particular wind regimes or boundary-layer stability, or in locations with distinct land-use/ NO_x availability—even if its global linear association with the target is weak. This is a well-established property of boosted trees and

modern feature selection theory: univariate correlation is an unreliable proxy for utility in non-linear models, whereas embedded/wrapper approaches and post-hoc attributions are preferred [57,58,64–66]. Consistently, TreeSHAP provides additive, locally accurate attributions for tree ensembles; in our O_3 model, SHAP assigns a non-negligible contribution to the Sentinel-5P column precisely when meteorology and land/emission proxies create favorable photochemical/background conditions, confirming interaction-driven use rather than any one-to-one mapping [62,63].

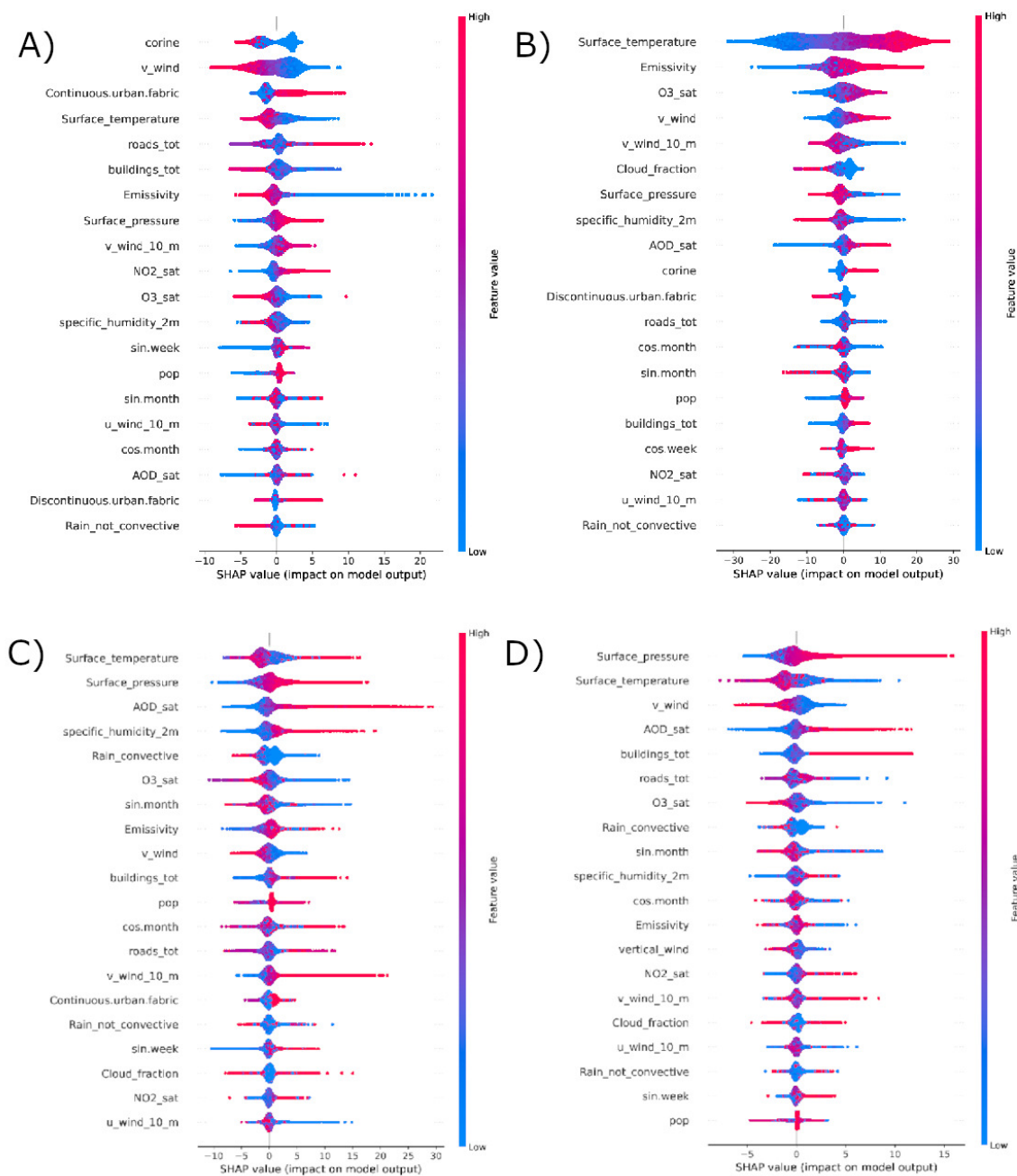


Figure 5. SHAP values distribution for the prediction of NO_2 (top left, panel A), O_3 (top right, panel B), PM_{10} (bottom left, panel C) and $PM_{2.5}$ (bottom right, panel D) concentrations, obtained in cross validation.

Figure 6 shows the trend of the predicted concentrations of NO_2 and O_3 compared to the ground values (our labels), averaged over all ground stations. To obtain this plot we applied a Random CV procedure.

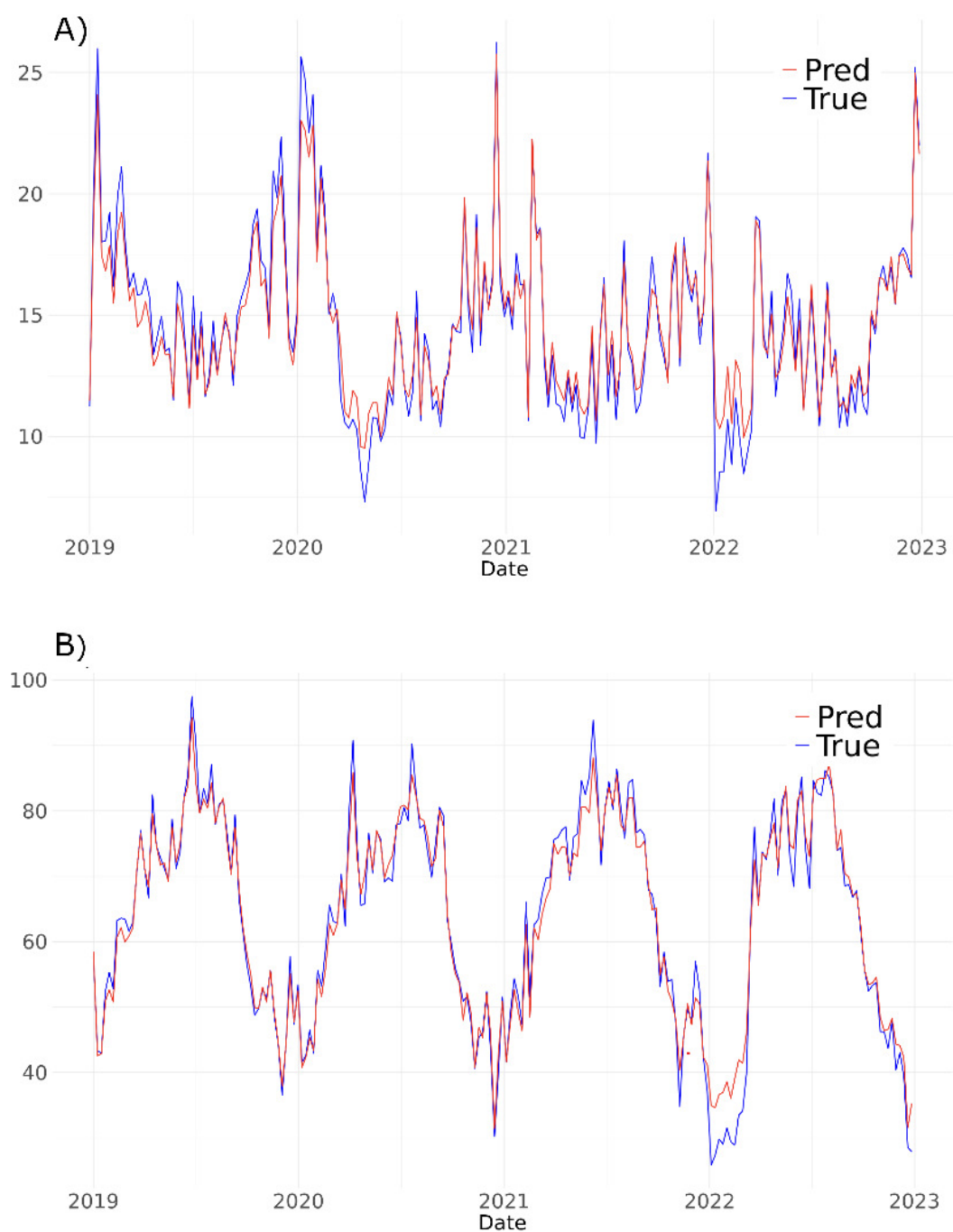


Figure 6. Left panel (panel A) time series of mean true and predicted values of NO_2 concentration ($\mu g/m^3$), on the right panel (panel B) the mean true and predicted values of O_3 concentration.

The temporal evolution in Figure 6 shows that the model reproduces the seasonal cycles of both pollutants. NO_2 peaks in winter and drops in summer, reflecting lower atmospheric mixing heights and more intense combustion/traffic demand in the cold season. O_3 instead peaks in the warm season, consistent with photochemical production under high temperature and radiation. The anti-correlation between the two is evident (high NO_2 / low O_3 in winter and vice versa in summer), and is well captured by the model. The largest discrepancies between prediction and ground truth occur during extreme or transient episodes, which is expected given the limited number of ARPA monitoring

stations and the fact that highly localized spikes (e.g. short-lived plumes, stagnation events) are harder to generalize regionally.

To assess the temporal generalization ability of the model, we adopted a different cross-validation approach. In particular, we applied a LOYO (Leave One Year Out) strategy, splitting the dataset by year.

In this procedure the model was trained for three years (we considered the time interval 2019–2022) and tested for the remaining year. This procedure was repeated to ensure that each year was used at least once for both testing and training. Table 5 shows the performance metrics (MAE, RMSE, R^2) obtained with this approach. Additionally, we summarized the predictions to monthly and yearly averages to observe possible seasonal trends. We decided not to perform spatial cross-validation (which is widely used in the literature), as the small number of measurement stations would lead to a bias in the model.

Table 5. Leave-One-Year-Out (LOYO) cross-validation for the XGBoost model. Metrics are mean \pm SD across LOYO folds. MAE and RMSE in $\mu\text{g m}^{-3}$. Predictions are evaluated at daily scale and after aggregation to monthly and annual means. All p -values are significant at the 1% level.

| | MAE | RMSE | R^2 |
|-------------------------|--------------------------|--------------------------|-----------------|
| | ($\mu\text{g m}^{-3}$) | ($\mu\text{g m}^{-3}$) | |
| Day prediction | | | |
| NO ₂ | 4.60 \pm 0.40 | 7.05 \pm 0.67 | 0.59 \pm 0.05 |
| O ₃ | 10.52 \pm 1.77 | 13.51 \pm 2.30 | 0.53 \pm 0.03 |
| PM _{2.5} | 3.92 \pm 0.18 | 5.45 \pm 0.26 | 0.30 \pm 0.01 |
| PM ₁₀ | 6.28 \pm 0.36 | 8.70 \pm 0.53 | 0.24 \pm 0.06 |
| Month prediction | | | |
| NO ₂ | 3.28 \pm 0.60 | 4.91 \pm 0.99 | 0.76 \pm 0.08 |
| O ₃ | 6.36 \pm 1.99 | 8.43 \pm 2.92 | 0.72 \pm 0.06 |
| PM _{2.5} | 1.89 \pm 0.14 | 2.39 \pm 0.20 | 0.72 \pm 0.05 |
| PM ₁₀ | 3.06 \pm 0.17 | 4.06 \pm 0.24 | 0.51 \pm 0.04 |
| Year prediction | | | |
| NO ₂ | 2.71 \pm 0.78 | 3.88 \pm 1.14 | 0.84 \pm 0.05 |
| O ₃ | 4.17 \pm 1.66 | 4.99 \pm 1.92 | 0.33 \pm 0.15 |
| PM _{2.5} | 1.42 \pm 0.37 | 1.63 \pm 0.41 | 0.79 \pm 0.13 |
| PM ₁₀ | 2.16 \pm 0.49 | 3.35 \pm 0.77 | 0.63 \pm 0.03 |

Table 5 evaluates a much harder task: predicting an unseen year (LOYO). As expected, daily-scale R^2 drops for all pollutants compared to random CV (Table 4), because LOYO tests temporal transferability under different meteorological regimes and emission patterns. NO₂ remains relatively robust ($R^2 = 0.59 \pm 0.05$ daily), confirming that the model captures persistent spatial structure of primary emissions (traffic, industry). PM₁₀ and PM_{2.5} show lower daily R^2 (0.24 and 0.30), which reflects their higher stochastic variability (episodic dust resuspension, Saharan intrusions, industrial events) and the limited number of monitors. Critically, O₃ exhibits $R^2 = 0.53 \pm 0.03$ at daily scale, lower than its random-CV $R^2 = 0.78$. This indicates that ozone is the least temporally transferable pollutant in our framework, consistent with the fact that O₃ depends on interannual meteorology and chemistry (photochemistry, NO titration, boundary-layer recirculation), and that satellite O₃ columns capture large-scale background rather than purely local boundary-layer behavior.

However, when we aggregate predictions to monthly and yearly means, performance markedly improves for all pollutants. For example, NO₂ reaches $R^2 = 0.84 \pm 0.05$ at the annual scale, and PM_{2.5} reaches $R^2 = 0.79 \pm 0.13$. Even O₃, while still more variable, reaches $R^2 = 0.72 \pm 0.06$ at the monthly scale. This suggests that although day-to-day prediction of certain episodes is challenging across

years, the model is robust for exposure assessment at epidemiologically relevant aggregation windows (monthly/annual), which are commonly used in long-term health and policy studies.

Through our model we can produce maps at the municipality level to do a focus on more local realities. Figure 7 shows the average predicted values of NO_2 and PM_{10} in the municipality of Bari and Taranto respectively.

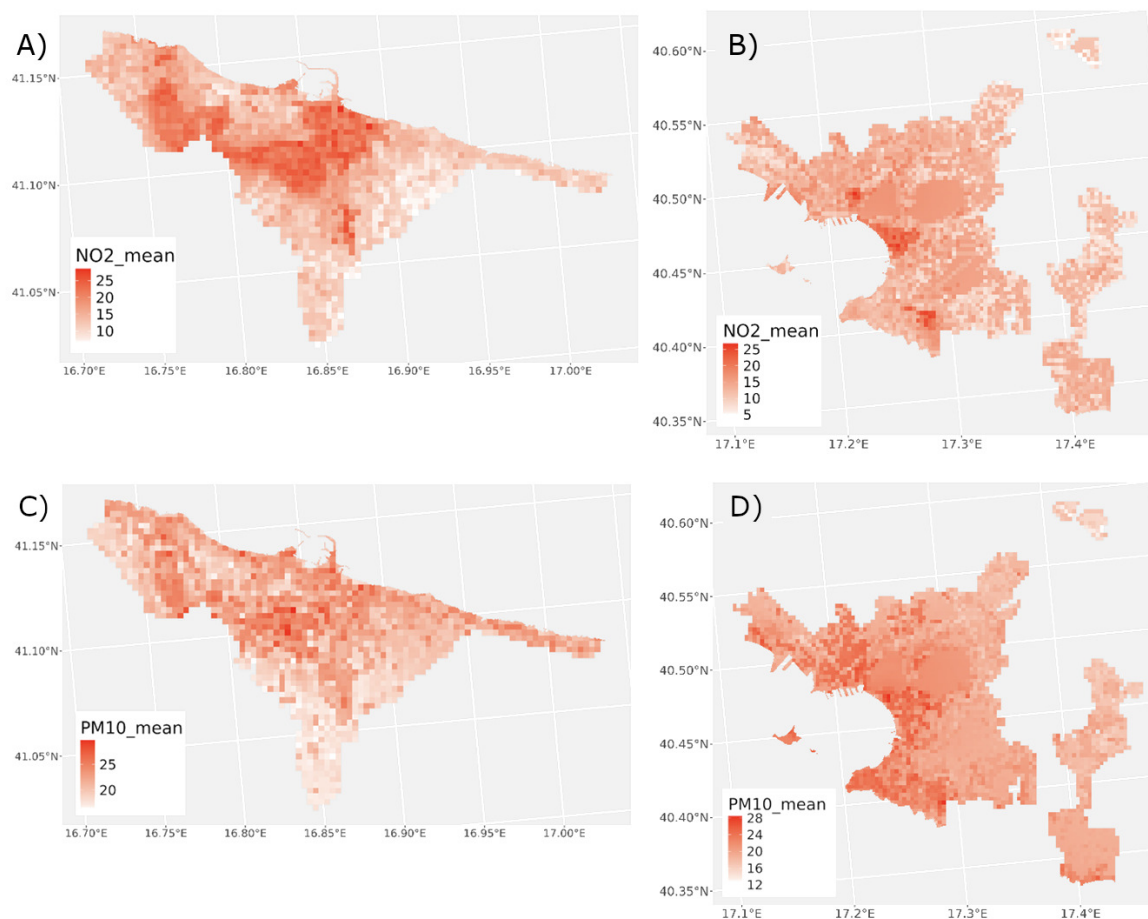


Figure 7. Top left panel (panel A) and bottom left (panel C) map of mean NO_2 and PM_{10} predicted in the area of Bari. On the right, map of mean NO_2 (top, panel B) and PM_{10} (bottom, panel D) in the area of Taranto. All prediction are made with the XGBoost model.

In Bari, NO_2 hotspots align with dense road networks and industrial/port areas, consistent with traffic-related and combustion-related sources. PM_{10} shows elevated levels both in industrial zones and along the coastline. This coastal enhancement is in line with AOD-informed particulate fields and may reflect port activity, industrial processes, and resuspension of marine/harbor aerosols. A similar pattern is observed in Taranto, where NO_2 is highest in the Tamburi district (adjacent to the steel plant area) and along major traffic corridors, while PM_{10} is enhanced over coastal/industrial sectors rather than uniformly over the whole urban core. This confirms that the model can resolve intra-urban patterns that are meaningful for local environmental management.

Figure 8 also highlights a key advantage of the proposed approach. Sentinel-5P tropospheric NO_2 mainly highlights large aggregated plumes (e.g. Taranto), whereas the model's predicted surface NO_2 field captures much finer spatial gradients, including within Bari. When compared with ARPA ground measurements, both Bari and Taranto show elevated NO_2 , but Bari exhibits even higher average concentrations despite the absence of a single dominant point source like the steel plant. This is consistent with population density data (Bari ≈ 2700 inhabitants/ km^2 vs Taranto ≈ 750 inhabitants/ km^2) and intense traffic-related emissions. In other words, the model recovers both

industrial and traffic-driven exposure patterns, which are not equally visible in the satellite column alone.

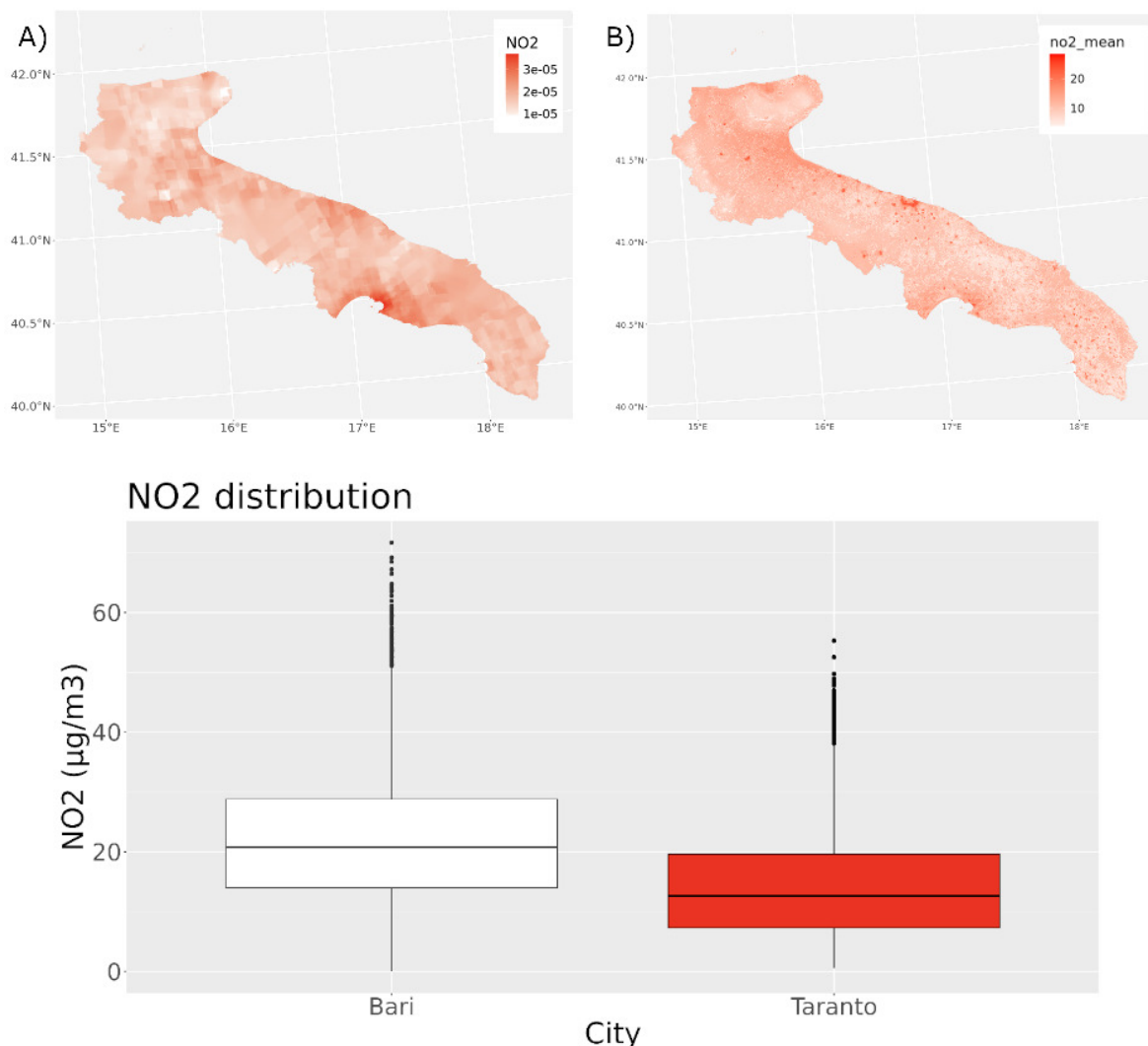


Figure 8. Comparison of NO_2 distribution (left, expressed in mol/m^2) measured by Sentinel 5-P and NO_2 distribution (right, expressed in $\mu\text{g}/\text{m}^3$) predicted by our model. On the bottom, Boxplot of NO_2 concentration (expressed in $\mu\text{g}/\text{m}^3$) from ARPA measurement in the areas of Bari (left) and Taranto (right).

Temporal behavior (Figure 6), spatial distribution at 300 m (Figures 4, 7), and feature attribution via SHAP (Figure 5) together indicate that the model is not only fitting statistical structure, but is reproducing physically and chemically consistent patterns: winter NO_2 buildup under low wind; warm-season O_3 peaks driven by photochemistry; enhanced PM_{10} near coastal/industrial zones tracked by AOD; and stronger spatial persistence of NO_2 hot spots across years compared to the more meteorology-driven O_3 .

4. Discussion

As shown in Table 4, the XGBoost model outperforms the linear model across all pollutants under random CV. This confirms that non-linear, interaction-aware models are required to fuse heterogeneous predictors (satellite columns, meteorology, AOD, land-use, temporal harmonics) and to translate them into surface-level pollutant concentrations at 300 m resolution. The results are consistent with other studies in the literature. Stafoggia et al. achieved an average R^2 of 0.75 for daily predictions of PM_{10} and $\text{PM}_{2.5}$ across Italy at ~ 1 km resolution [67]. Our $\text{PM}_{2.5}$ and PM_{10} R^2 values (0.67 and 0.64 in

Table 4) are slightly lower, which is expected because: we work at a finer spatial scale (300 m instead of ≥ 1 km), increasing spatial variability to be explained; and we rely on a smaller and regionally concentrated monitoring network within Apulia, rather than hundreds of stations nationwide. This resolution effect is well known: moving from kilometer to sub-kilometer grids sharpens intra-urban contrasts (traffic corridors, port areas, industrial districts), which increases variance to be predicted and therefore mechanically lowers apparent R^2 at the pixel level.

The performance of our model in predicting NO_2 and O_3 is also comparable to or better than values reported in the Italian literature. Silibello et al. report $R^2 \approx 0.6$ for NO_2 and $R^2 \approx 0.7$ for O_3 using hybrid modeling over the whole Italian territory at ~ 1 km resolution [68]. In our case, random CV reaches $R^2 = 0.77$ for NO_2 and 0.78 for O_3 (Table 4), again despite the more stringent (300 m) spatial resolution. This is notable because our study region is meteorologically complex (coastal recirculation, alternating land–sea breeze regimes, intense summer photochemistry) and industrially heterogeneous (steel, port logistics, traffic), which typically complicates surface pollutant modeling. The fact that the model maintains high skill in this setting suggests that even with fewer monitoring stations, high-resolution anthropogenic features (road density, port/industrial land-use, continuous urban fabric), and localized meteorology (WRF downscaled and resampled to 300 m) provide enough structure to generalize.

When compared to studies outside Italy, our performance is broadly in line with machine learning air quality models developed in other complex regions. For example, Di et al. reported R^2 values in the 0.7–0.9 range for daily $PM_{2.5}$ across the continental United States at ~ 1 km by combining satellite AOD, chemical transport model (CTM) output, land-use, and meteorology through an ensemble of statistical and machine learning learners [10]. Li et al. obtained R^2 values around 0.7–0.8 for $PM_{2.5}$ in China using deep ensemble approaches and rich station networks [11]. In Beijing and other megacities, deep recurrent models (LSTM, GRU, CNN-LSTM hybrids) used for $PM_{2.5}$ forecasting often report high predictive skill [7,8]; however, these studies typically operate in settings with dozens of stations within a single urban basin and do not always target sub-kilometer mapping of long-term exposure fields. Our framework, by contrast, is designed to deliver exposure fields at 300 m across an entire region that contains multiple distinct urban/industrial systems (Bari, Taranto, coastal logistics, rural inland). In this sense, our approach is closer to regional exposure mapping than to single-city forecasting, while retaining a spatial resolution that is comparable to neighborhood/block scale.

From a public health perspective, this point is non-trivial. Epidemiological and One Health studies increasingly focus on chronic exposure at intra-urban scale (tens to hundreds of meters), because population vulnerability is not uniform within a city: residential areas sit next to major roads; schools and hospitals may lie downwind of ports or industrial stacks; disadvantaged neighborhoods are often co-located with legacy industrial sources and high traffic load. Our 300 m pollutant fields aim exactly at that scale. This is in line with recent air quality exposure work in North America and Northern Europe that emphasizes environmental justice and differential exposure, where within-city gradients are policy-relevant on their own and cannot be inferred from sparse fixed-site monitors alone [10,11].

As expected for data-driven models, LOYO CV exposes reduced daily performance due to year-to-year regime shifts, the limited number and uneven distribution of monitors, and the weaker seasonality and higher stochastic variability of PM. However, aggregating predictions to monthly and annual scales markedly improves R^2 (Table 5), supporting use for exposure assessment and One Health applications. In epidemiology and regulatory assessment, exposure is often averaged over months or years rather than interpreted day by day; our LOYO monthly R^2 values (0.72–0.76 for NO_2 , O_3 , and $PM_{2.5}$) indicate that such aggregated exposure fields are reliable even when an entire year is held out of training. This is important because it shows that the model is not just interpolating noise from individual days, but is actually able to reconstruct the long-term spatial structure of exposure that matters for cardiovascular, respiratory, and neurodegenerative risk analyses [2–4].

The differentiated behavior among pollutants has a clear atmospheric interpretation. NO_2 is a short-lived, primarily emitted pollutant tied to combustion (traffic, port activity, industrial stacks).

Its spatial pattern is persistent in time (urban cores, road networks, industrial/harbor districts), and therefore NO_2 remains relatively robust in LOYO ($R^2 = 0.59$ daily). This agrees with prior work in Italy and across Europe, where NO_2 gradients are dominated by local emission intensity and street-scale dispersion [67,68]. In Apulia we observe the same structure: SHAP shows that NO_2 predictions increase sharply in areas with high road density, dense built-up fabric, and industrial land-use (Figure 5); wind speed carries strong negative SHAP values at high intensity, consistent with pollutant dilution. This supports a mechanistic reading: the model is not only “fitting” data, it is reconstructing physically interpretable source–dispersion relationships.

PM_{10} and $PM_{2.5}$ behave differently. Particulate matter in Apulia reflects both persistent sources (traffic, industry, port logistics, heating) and episodic inputs (resuspension, harbor/ship emissions, Saharan dust intrusions, stagnant meteorology). This mixture of steady and episodic drivers explains why daily LOYO skill is lower ($R^2 = 0.24$ – 0.30) than for NO_2 , and why performance improves so strongly after temporal aggregation (up to $R^2 = 0.79$ for annual $PM_{2.5}$). SHAP indicates that AOD is a dominant positive driver for both PM_{10} and $PM_{2.5}$, while precipitation and stronger winds tend to suppress particulate levels (negative SHAP at high values). This is fully consistent with atmospheric physics: wet scavenging and ventilation remove particles, while dry, stagnant, high-AOD conditions favor accumulation. The spatial patterns we predict (Figures 4 and 7) show PM_{10} and $PM_{2.5}$ enhancements along coastal/industrial belts and port zones, including areas in Bari and Taranto with intense maritime and logistics activity. Similar harbor-related particulate enhancements have been described in other Mediterranean port cities, where ship emissions and bulk handling (e.g. minerals, coal, grain) locally elevate aerosol load [23,69].

O_3 represents the most conceptually challenging pollutant. Unlike NO_2 and primary PM, surface ozone is not directly emitted. It is formed photochemically from nitrogen oxides (NO_x) and volatile organic compounds (VOCs) under sunlight, and then shaped by transport, entrainment from the residual layer, NO titration in traffic corridors, and boundary-layer dynamics. In coastal Mediterranean regions like Apulia, land–sea breeze circulations and recirculation cells can trap and re-inject aged air masses rich in photochemical products. This means that day-to-day O_3 fields depend strongly on meteorology and chemistry, and the surface O_3 level is not simply proportional to the instantaneous tropospheric O_3 column observed by Sentinel-5P. Indeed, previous studies have repeatedly emphasized that satellite tropospheric ozone columns correlate only weakly and inconsistently with surface ozone, because they integrate also the free troposphere and large-scale background [14,70].

Our results are consistent with this view. In random CV, where the model is trained and tested within overlapping seasonal regimes, O_3 achieves $R^2 = 0.78$ (Table 4). Under LOYO, daily O_3 R^2 drops to 0.53 (Table 5), highlighting that interannual variability in solar radiation, stagnation events, and photochemical regimes is difficult to extrapolate. Importantly, SHAP reveals that O_3 predictions are driven by temperature, emissivity (a proxy for surface energy balance and thus photochemical potential), and wind-related variables, with Sentinel-5P O_3 column still contributing positively in interaction with these drivers. That is: the column ozone on its own is not a good linear predictor, but in combination with meteorology and land-use it contains usable signal. This is exactly the scientific motivation for applying non-linear ML instead of simple regression.

From a modeling standpoint, this is also fully consistent with the theory of gradient boosting and modern feature selection: tree ensembles capture interactions and threshold effects without requiring large marginal correlations [57,58,64]. Consequently, correlation filtering is not an adequate criterion to discard features in non-linear pipelines; embedded methods and post-hoc attributions (e.g., SHAP) are more appropriate to evaluate contextual usefulness [62,63,65,66].

The ozone result has two implications. First, it confirms that O_3 is intrinsically harder to generalize temporally than NO_2 or PM, which explains why O_3 -health analyses often rely on multi-year averages or warm-season metrics rather than raw daily values. Second, it suggests clear directions for future improvement: adding explicit predictors of boundary-layer height, stagnation/recirculation indices, and VOC proxies (e.g. land-cover classes indicative of biogenic VOC emissions, or industrial/port VOC

sources) could capture more of the chemistry that drives O_3 , especially in summer. This is aligned with recent calls in atmospheric modeling to hybridize physics and ML, using chemical transport model outputs as structured predictors while still letting data-driven methods learn residual structure [15,16].

SHAP analysis more broadly confirms that the model is physically consistent and therefore interpretable for policy. For NO_2 , strong positive SHAP values are associated with dense built-up fabric, major roads, and industrial land-use; wind speed carries negative SHAP at high intensity, consistent with dispersion. For $PM_{10}/PM_{2.5}$, AOD is the top positive predictor, and precipitation/ventilation act as sinks. For O_3 , temperature and emissivity (radiative forcing at the surface) play a central role, encoding the photochemical engine of ozone formation. This level of interpretability helps ensure that predicted maps can be discussed with air quality agencies and health stakeholders without treating the ML model as an opaque “black box”.

Spatially, the 300 m maps (Figures 4 and 7) and the comparison with ARPA measurements (Figure 8) highlight two important policy-relevant findings. First, NO_2 exposure in Apulia is not confined to a single point source (e.g. the Taranto steel complex), but also reaches high levels in dense urban/traffic environments like Bari, where population density and mobility are high. This means that industrial liability alone cannot explain the observed NO_2 burden: diffuse transport-related emissions in dense coastal cities can produce comparable or higher average exposure, as confirmed by ARPA station data. Second, PM_{10} hotspots align with industrial/port and coastal areas, suggesting that harbors and shoreline logistics contribute to particulate exposure, not only classical industrial stacks. This supports targeted mitigation strategies (harbor electrification, dust suppression, traffic control near port gates) that are already being discussed in several EU port cities.

From a regulatory standpoint, these results are coherent with the 2024 EU Air Quality Directive, which explicitly encourages Member States to combine monitoring networks with modeling and satellite data to generate high-resolution exposure assessments over populated areas, rather than simply multiplying fixed stations. Our framework directly responds to that directive: it fuses remote sensing, meteorology, and land-use to produce daily fields at 300 m, and it remains stable (especially after temporal aggregation) even when an entire year is withheld in training. This is relevant not only for compliance reporting, but also for health impact assessment and One Health planning in regions where industrial activity coexists with densely populated urban waterfronts.

We also underline the One Health dimension. Apulia is a densely inhabited coastal/industrial region where humans, ecosystems, and industrial pressures overlap in a relatively small spatial scale. Chronic exposure to NO_2 , $PM_{2.5}$, and PM_{10} is linked to respiratory and cardiovascular outcomes and, as emerging evidence suggests, to neurodegenerative processes [2–4]. Surface O_3 , especially during warm-season peaks, exacerbates respiratory stress and can impair lung function. By reconstructing pollutant fields at 300 m resolution, including near schools, hospitals, and residential areas adjacent to industrial/port infrastructure, our model can directly support exposure analysis in a One Health framework, i.e. considering human health, environmental quality, and industrial/economic sustainability together.

Limitations remain. The ARPA network in Apulia, while compliant with regulations, is relatively sparse outside the main urban and industrial centers, which can introduce spatial bias and limit the model’s ability to learn fine-scale patterns in rural areas. O_3 remains challenging because satellite tropospheric O_3 columns integrate the entire free troposphere and are only indirectly sensitive to boundary-layer concentrations; this physical mismatch reduces temporal transferability. Interannual regime shifts in meteorology and emissions also reduce LOYO performance, especially at daily resolution. Future work should incorporate additional dynamical predictors (e.g. planetary boundary layer height, stagnation/recirculation metrics), explore hybrid physics–ML approaches that blend WRF/CTM outputs with data-driven learning [15,16], and include explicit uncertainty quantification to flag low-confidence conditions. We also expect that forthcoming missions such as NASA’s MAIA, designed to provide high-resolution particulate information, will further improve $PM_{2.5}/PM_{10}$ estimation and interpretation.

Despite these challenges, our framework already demonstrates that high-resolution (300 m), explainable daily exposure maps for NO_2 , PM_{10} , $PM_{2.5}$, and O_3 are feasible in a coastal industrial Mediterranean region. This is particularly relevant for One Health studies and environmental justice assessments in Bari, Taranto, and similar urban–industrial corridors, where both industrial plumes and traffic emissions matter and where vulnerable populations live close to sources.

5. Conclusions

In this work we developed an explainable machine learning framework to estimate daily surface concentrations of four key air pollutants (NO_2 , O_3 , PM_{10} , $PM_{2.5}$) across the Apulia region at 300 m resolution for the period 2019–2022. The model integrates Sentinel-5P column products, a calibrated multi-source AOD field, high-resolution meteorology from WRF, land-use and anthropogenic indicators (roads, buildings, industrial fabric, population), and temporal harmonics. We assessed predictive skill using both repeated random cross-validation and a Leave-One-Year-Out (LOYO) temporal validation scheme.

Random CV yielded high R^2 values (0.64–0.78 depending on pollutant), confirming that tree-based fusion (XGBoost) captures non-linear interactions between satellite, meteorological, and land-use drivers that linear models cannot. LOYO daily performance dropped, especially for PM and O_3 , reflecting interannual variability in atmospheric conditions and emissions. Crucially, however, after aggregation to monthly and annual means, the R^2 for NO_2 , PM_{10} , $PM_{2.5}$, and partially also O_3 remained high (often > 0.7), indicating that our maps are reliable for long-term exposure assessment at epidemiologically relevant time scales.

O_3 deserves special mention. Although O_3 achieved $R^2 = 0.78$ in random CV, its R^2 in LOYO daily prediction was lower (0.53), highlighting limited temporal transferability. This is consistent with the known physical mismatch between satellite tropospheric ozone columns and near-surface ozone, and with the strong dependence of ozone on photochemistry, boundary-layer dynamics, and NO titration. Nevertheless, SHAP analysis shows that the model uses physically interpretable drivers (temperature, emissivity, wind, column O_3) to reproduce O_3 seasonality and warm-season peaks. This suggests that even ozone, which is traditionally difficult to downscale from satellite alone, can be partially constrained by a non-linear, explainable fusion framework.

From an application perspective, the generated maps resolve intra-urban gradients in Bari and Taranto, distinguishing traffic-related NO_2 exposure from industrial/harbor-related PM_{10} loads. This level of spatial detail is directly relevant for health impact assessment, environmental monitoring, and targeted mitigation strategies in sensitive neighborhoods. The methodology therefore supports the goals of the 2024 EU Air Quality Directive, which encourages improved assessment (as opposed to simply multiplying monitoring stations) and aligns with future satellite missions such as NASA's MAIA, which will provide high-resolution particulate measurements and is expected to further strengthen particulate matter estimation.

In summary, we provide what is, to our knowledge, the first daily, 300 m resolution, explainable AI-based air pollution product for Apulia, integrating satellite, meteorology, land-use and population indicators. The framework is transferable to other Mediterranean and coastal industrial regions with sparse monitoring networks, and sets the stage for One Health and policy-relevant exposure studies focused on both chronic and acute pollution burdens.

Author Contributions: A.F. Conceptualization, methodology, data analysis, software, visualization, writing – original draft; G.L., R.C., A.L., M.L.R., T.M., E.P.: Data curation, validation, investigation, writing – review & editing. N.A., M.A., M.Aq., L.B., M.D.L., A.Mor., A.No., R.P., S.T., A.M., R.B.: Supervision, resources, project administration, writing – review & editing.

Funding: This research received no external funding.

Data Availability Statement: All codes and data used to perform the analysis are available upon request.

Acknowledgments: This paper was funded by the Italian funding within the “Budget MIUR - Dipartimenti di Eccellenza 2023 - 2027” (Law 232, 11 December 2016) - Quantum Sensing and Modelling for One-Health (QuaSiModO), CUP: H97G23000100001. Authors want to thank the Funder: Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16 December 2021 of Italian Ministry of University and Research funded by the European Union - NextGenerationEU. This work was also supported by the Assessment of PM Exposure at intra-urban scale in preparation of MAIA mission (APEMAIA) project, funded by the Italian Space Agency, CALL FOR IDEAS “ATTIVITÀ SCIENTIFICHE A SUPPORTO DELLO SVILUPPO DELLE MISSIONI DI OSSERVAZIONE DELLA TERRA”, Contract n. 2023-39-HH.0.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Organization, W.H. Ambient (outdoor) air pollution. [https://www.who.int/en/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health), 2024.
2. Huang, W.J.; Zhang, X.; Chen, W.W. Role of oxidative stress in Alzheimer’s disease. *Biomedical reports* **2016**, *4*, 519–522.
3. Ionescu-Tucker, A.; Cotman, C.W. Emerging roles of oxidative stress in brain aging and Alzheimer’s disease. *Neurobiology of aging* **2021**, *107*, 86–95.
4. Fania, A.; Monaco, A.; Amoroso, N.; Bellantuono, L.; Cazzolla Gatti, R.; Firza, N.; Lacalamita, A.; Pantaleo, E.; Tangaro, S.; Velichevskaya, A.; et al. Machine learning and XAI approaches highlight the strong connection between O₃ and NO₂ pollutants and Alzheimer’s disease. *Scientific Reports* **2024**, *14*, 5385.
5. Organization, W.H. Health consequences of air pollution on populations. <https://www.who.int/news/item/25-06-2024-what-are-health-consequences-of-air-pollution-on-populations>, 2024.
6. Méndez, M.; Merayo, M.G.; Núñez, M. Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review* **2023**, *56*, 10031–10066.
7. Tsokov, S.; Lazarova, M.; Aleksieva-Petrova, A. A hybrid spatiotemporal deep model based on CNN and LSTM for air pollution prediction. *Sustainability* **2022**, *14*, 5104.
8. Tao, Q.; Liu, F.; Li, Y.; Sidorov, D. Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU. *IEEE access* **2019**, *7*, 76690–76698.
9. Kök, İ.; Şimşek, M.U.; Özdemir, S. A deep learning model for air quality prediction in smart cities. In Proceedings of the 2017 IEEE international conference on big data (big data). IEEE, 2017, pp. 1983–1990.
10. Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M.B.; Choirat, C.; Koutrakis, P.; Lyapustin, A.; et al. An ensemble-based model of PM_{2.5} concentration across the contiguous United States with high spatiotemporal resolution. *Environment International* **2019**, *130*, 104909. <https://doi.org/10.1016/j.envint.2019.104909>.
11. Li, L.; Girguis, M.; Lurmann, F.; Pavlovic, N.; McClure, C.; Franklin, M.; Wu, J.; Oman, L.D.; Breton, C.; Gilliland, F.; et al. Ensemble-based deep learning for estimating PM_{2.5} over California with multisource big data including wildfire smoke. *Environment International* **2020**, *145*, 106143. <https://doi.org/10.1016/j.envint.2020.106143>.
12. Tang, D.; et al. A review of machine learning for modeling air quality: Overlooked but important issues. *Atmospheric Research* **2024**, *296*, 107261. <https://doi.org/10.1016/j.atmosres.2024.107261>.
13. Agbehadji, I.E.; Obagbuwa, I.C. Systematic Review of Machine Learning and Deep Learning Techniques for Spatiotemporal Air Quality Prediction. *Atmosphere* **2024**, *15*, 1352. <https://doi.org/10.3390/atmos15111352>.
14. Petritoli, A.; Bonasoni, P.; Giovanelli, G.; Ravegnani, F.; Kostadinov, I.; Bortoli, D.; Weiss, A.; Schaub, D.; Richter, A.; Fortezza, F. First comparison between ground-based and satellite-borne measurements of tropospheric nitrogen dioxide in the Po basin. *Journal of Geophysical Research: Atmospheres* **2004**, *109*.
15. Baklanov, A.; Zhang, Y. Advances in air quality modeling and forecasting. *Global Transitions* **2020**, *2*, 261–270. <https://doi.org/10.1016/j.glt.2020.11.001>.
16. Gao, Z.; Zhou, X. A review of the CAMx, CMAQ, WRF-Chem and NAQPMS models: Application, evaluation and uncertainty factors. *Environmental Pollution* **2024**, *343*, 123183. <https://doi.org/10.1016/j.envpol.2023.123183>.
17. Considine, E.M.; et al. Evaluation of Model-Based PM_{2.5} Estimates for Exposure Assessment during Wildfire Smoke Episodes in the Western United States. *Environmental Science & Technology* **2023**. Ahead of print / pagination varies by source, <https://doi.org/10.1021/acs.est.2c06288>.

18. Quotidiano, I.F. Ilva, la tormentata storia dell'acciaieria dallo Stato ai Riva fino a Mittal: lavoro, ambiente "svenduto" e un rilancio sempre più difficile. <https://www.ilfattoquotidiano.it/2024/02/19/ilva-storia-acciaieria-stato-riva-mittal-lavoro-ambiente/7451965/>, 2024.
19. Puglia Sviluppo S.p.A.. Mappatura delle aree industriali in Puglia. Technical Report; Available online: <https://old.pugliasviluppo.eu/investiinpuglia/Report%20Mappatura%20Aree%20Industriali.pdf> (accessed on 23 January 2025), 2017.
20. European Parliament. Briefing for Fact Finding Visit to Taranto, Italy. Technical Report; Available online: <https://www.europarl.europa.eu/cmsdata/123280/Background%20Document%20PE571.403EN.pdf> (accessed on 23 January 2025), 2017.
21. Tateo, A.; Campanaro, V.; Amoroso, N.; Bellantuono, L.; Monaco, A.; Pantaleo, E.; Rinaldi, R.; Maggipinto, T. Predicting Air Quality from Measured and Forecast Meteorological Data: A Case Study in Southern Italy. *Atmosphere* **2023**, *14*, 475. <https://doi.org/10.3390/atmos14030475>.
22. Fania, A.; Monaco, A.; Pantaleo, E.; Maggipinto, T.; Bellantuono, L.; Cilli, R.; Lacalamita, A.; La Rocca, M.; Tangaro, S.; Amoroso, N.; et al. Estimation of Daily Ground Level Air Pollution in Italian Municipalities with Machine Learning Models Using Sentinel-5P and ERA5 Data. *Remote Sensing* **2024**, *16*, 1206.
23. Leogrande, S.; Alessandrini, E.R.; Stafoggia, M.; Morabito, A.; Nocioni, A.; Ancona, C.; Bisceglia, L.; Mataloni, F.; Giua, R.; Mincuzzi, A.; et al. Industrial air pollution and mortality in the Taranto area, Southern Italy: A difference-in-differences approach. *Environment International* **2019**, *132*, 105030. <https://doi.org/10.1016/j.envint.2019.105030>.
24. Regione Puglia. Piano Regionale per la Qualità dell'Aria (PRQA)—Allegato 2: Rapporto preliminare di orientamento. Technical Report DGR n. 2436/2019; Available online: https://pugliacon.sit.puglia.it/Documenti/GestioneDocumentale/Allegati/DGR_2436_2019_Allegato_2.pdf (accessed on 23 January 2025), 2019.
25. ISPRA.; ARPA Puglia. Monitoraggio della qualità dell'aria nella Regione Puglia—Report mensile (marzo 2013). Technical Report; Available online: https://www.isprambiente.gov.it/it/garante_aia_ilva/dati-ambientali-rilevanti/dati-rilevati-e-comunicati-dallarpa-puglia/reportmarzo2013.pdf (accessed on 23 January 2025), 2013.
26. Wikipedia. Agenzia regionale per la protezione ambientale. https://it.wikipedia.org/wiki/Agenzia_regionale_per_la_protezione_ambientale, 2024.
27. ITALIANA, G.U.D.R. DIRETTIVA (UE) 2024/2881 DEL PARLAMENTO EUROPEO E DEL CONSIGLIO. <https://www.gazzettaufficiale.it/do/gazzetta/downloadPdf?dataPubblicazioneGazzetta=20250123&numeroGazzetta=6&tipoSerie=S2&tipoSupplemento=GU&numeroSupplemento=0&progressivo=0&numPagina=260&estensione=pdf&edizione=0&home=>, 2024.
28. ESA. Sentinel-5P. <https://esoc.esa.int/content/sentinel-5p>, 2024.
29. NASA. Aerosol Optical Depth. https://earthobservatory.nasa.gov/global-maps/MODAL2_M_AER_OD, 2024.
30. Handschuh, J.; Erbertseder, T.; Baier, F. Systematic evaluation of four satellite AOD datasets for estimating PM_{2.5} using a random forest approach. *Remote Sensing* **2023**, *15*, 2064.
31. Schutgens, N.; Sayer, A.M.; Heckel, A.; Hsu, C.; Jethva, H.; De Leeuw, G.; Leonard, P.J.; Levy, R.C.; Lipponen, A.; Lyapustin, A.; et al. An AeroCom/AeroSat Study: Intercomparison of Satellite AOD Datasets for Aerosol Model Evaluation. *Atmospheric Chemistry and Physics Discussions* **2020**, *2020*, 1–43.
32. Copernicus. Level-2 AOD. <https://sentinels.copernicus.eu/web/sentinel/level-2-aod>, 2024.
33. NASA. Atmosphere Monitoring Service. <https://modis.gsfc.nasa.gov/data/>, 2024.
34. Lyapustin, A.; Wang, Y.; Korokin, S.; Huang, D. MODIS Collection 6 MAIAC algorithm. *Atmospheric Measurement Techniques* **2018**, *11*, 5741–5765. <https://doi.org/10.5194/amt-11-5741-2018>.
35. NASA. Modern-Era Retrospective analysis for Research and Applications, Version 2. <https://gmao.gsfc.nasa.gov/reanalysis/merra-2/>, 2024.
36. Copernicus. Atmosphere Monitoring Service. <https://atmosphere.copernicus.eu/ads-now-contains-20-year-cams-global-reanalysis-eac4-dataset>, 2024.
37. Blalock Jr, H.M. Correlated independent variables: The problem of multicollinearity. *Social Forces* **1963**, *42*, 233–237.
38. USGS. EarthExplorer. <https://earthexplorer.usgs.gov/>, 2024.
39. Copernicus. ECMWF Reanalysis v5 (ERA5). <https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5>, 2024.
40. NASA. AERONET. <https://aeronet.gsfc.nasa.gov/>, 2024.

41. Aldabash, M.; Bektas Balçık, F.; Glantz, P. Validation of MODIS C6. 1 and MERRA-2 AOD using AERONET observations: A comparative study over Turkey. *Atmosphere* **2020**, *11*, 905.
42. Copernicus. CORINE Land Cover. <https://land.copernicus.eu/en/products/corine-land-cover>, 2024.
43. Oiamo, T.H.; Johnson, M.; Tang, K.; Luginaah, I.N. Assessing traffic and industrial contributions to ambient nitrogen dioxide and volatile organic compounds in a low pollution urban environment. *Science of the Total Environment* **2015**, *529*, 149–157.
44. Fedele, F.; Pollice, A.; Guarnieri Calò Carducci, A.; Bellotti, R. Spatial bias analysis for the Weather Research and Forecasting model (WRF) over the Apulia region. In Proceedings of the Proceedings of the GRASPA2015 Conference, Bari, Italy, 2015.
45. Fedele, F.; Tateo, A.; Menegotto, M.; Turnone, A.; Figorito, B.; Guarnieri Calò Carducci, A.; Bellotti, R. Impact of Planetary Boundary Layer parametrization scheme and land cover classification on surface processes: Wind speed and temperature bias spatial distribution analysis over South Italy. In Proceedings of the EMS Annual Meeting & ECAM 2015, Sofia, Bulgaria, 2015.
46. Tateo, A.; Miglietta, M.M.; Fedele, F.; Menegotto, M.; Monaco, A.; Bellotti, R. Ensemble using different Planetary Boundary Layer schemes in WRF model for wind speed and direction prediction over Apulia region. *Advances in Science and Research* **2017**, *14*, 95–102. <https://doi.org/10.5194/asr-14-95-2017>.
47. Tateo, A.; Miglietta, M.M.; Fedele, F.; Menegotto, M.; Ottonelli, S.; Bellotti, R. Multi-physics ensemble using different PBL schemes in WRF model for PBL height prediction over Apulia region. In Proceedings of the Scientific Research Abstracts, 2018, number 153.
48. Wikipedia. OpenStreetMap. <https://en.wikipedia.org/wiki/OpenStreetMap>, 2024.
49. OpenStreetMap. Map features - OpenStreetMap. https://wiki.openstreetmap.org/wiki/Map_features, 2024.
50. WorldPop. WorldPop :: Population counts. <https://hub.worldpop.org/project/categories?id=3>, 2024.
51. Stevens, Forrest R., e.a. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS one* **2015**, *10*, 2.
52. Han, D. Comparison of commonly used image interpolation methods. In Proceedings of the Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013). Atlantis Press, 2013, pp. 1556–1559.
53. Keys, R. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing* **1981**, *29*, 1153–1160.
54. Shaddick, G.; Zidek, J.V.; Schmidt, A.M. *Spatio-Temporal Methods in Environmental Epidemiology with R*; CRC Press, 2023.
55. Lam, N.S.N. Spatial interpolation methods: a review. *The American Cartographer* **1983**, *10*, 129–150.
56. Moraga, P. *Spatial statistics for data science: theory and practice with R*; CRC Press, 2023.
57. Friedman, J.H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* **2001**, pp. 1189–1232.
58. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), New York, NY, USA, 2016; pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
59. Friedman, J.H. Stochastic gradient boosting. *Computational statistics & data analysis* **2002**, *38*, 367–378.
60. Flach, P. Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward. *Proceedings of the AAAI Conference on Artificial Intelligence* **2019**, *33*, 9808–9814.
61. Vollmer, S.; et al.. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* **2020**, *368*, l6927.
62. Lundberg, S.; Lee, S. A unified approach to interpreting model predictions. In Proceedings of the Proceedings of the 31st international conference on neural information processing systems, 2017, pp. 44768–44777.
63. Lundberg, S.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell* **2020**, *2*, 56–67.
64. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, 2009. <https://doi.org/10.1007/978-0-387-84858-7>.
65. Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* **2003**, *3*, 1157–1182.
66. Kuhn, M.; Johnson, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*; Chapman and Hall/CRC: Boca Raton, FL, 2019. <https://doi.org/10.1201/9781315108230>.

67. Stafoggia, M.; Bellander, T.; Bucci, S.; Davoli, M.; De Hoogh, K.; De'Donato, F.; Gariazzo, C.; Lyapustin, A.; Michelozzi, P.; Renzi, M.; et al. Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Environment international* **2019**, *124*, 170–179.
68. Silibello, C.; Carlino, G.; Stafoggia, M.; Gariazzo, C.; Finardi, S.; Pepe, N.; Radice, P.; Forastiere, F.; Viegi, G. Spatial-temporal prediction of ambient nitrogen dioxide and ozone levels over Italy using a Random Forest model for population exposure assessment. *Air Quality, Atmosphere & Health* **2021**, *14*, 817–829.
69. Xie, Q.; Sun, Q. Monitoring the spatial variation of aerosol optical depth and its correlation with land use/land cover in Wuhan, China: a perspective of urban planning. *International Journal of Environmental Research and Public Health* **2021**, *18*, 1132.
70. Fishman, J.; Bowman, K.W.; Burrows, J.P.; Richter, A.; Chance, K.V.; Edwards, D.P.; Martin, R.V.; Morris, G.A.; Pierce, R.B.; Ziemke, J.R.; et al. Remote sensing of tropospheric pollution from space. *Bulletin of the American Meteorological Society* **2008**, *89*, 805–821. <https://doi.org/10.1175/2008BAMS2526.1>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.