applied sciences

MDPI

Article

Citizen Science and STEM Education with R: Reproducible Learning from Open Urban Air Quality Data

Jesús Cáceres-Tello [1], José Javier Galán-Hernández [2], María Belén Morales Cevallo[3], Eloy López-Meneses[4] , *

[1] Department of Computer Science and Engineering, Faculty of Computer Science, Complutense University of Madrid, Spain; jescacer@ucm.es
[2] Department of Computer Science, University of Alcalá, Spain; jose.galan@uah.es
[3] Faculty of Marketing and Communication, Universidad Ecotec, Samborondón, Ecuador; elopmen@upo.es
[4] Department of Education and Social Psychology, Pablo de Olavide University, Sevilla, Spain; elopmen@upo.es
* Correspondence: jescacer@ucm.es

Featured Application

The workflow proposed in this study can be readily applied to environmental education, air-quality management, and citizen-science initiatives. It offers a fully reproducible framework that integrates open urban data with R-based analysis and forecasting, enabling educators, students, and local administrations to examine real atmospheric patterns, assess pollution dynamics, and design data-driven sustainability actions. The same approach can be adapted to other cities or environmental domains where open data and civic participation intersect.

Abstract

Open urban environmental data offer a unique opportunity to connect scientific research, education, and citizen participation. This study presents a reproducible workflow developed in the Quarto–R environment to analyse and model air-quality dynamics in Madrid between 2020 and 2024. The workflow integrates data acquisition, validation, harmonisation, exploratory analysis, and forecasting using the Prophet model. The analysis focuses on nitrogen dioxide (NO_2) and ozone (O_3) as representative pollutants of traffic emissions and photochemical processes. Results show a marked decline in NO_2 concentrations across traffic stations and a parallel rise in O_3 levels in suburban areas, reflecting the combined effects of emission control and regional transport.

Beyond its scientific contribution, the Quarto–R workflow functions as a pedagogical tool that embeds transparency, traceability, and active learning throughout the analytical process. By enabling students and researchers to reproduce every step—from raw data to interpreted results—it strengthens data literacy and fosters a deeper understanding of urban sustainability. The framework exemplifies how open data and reproducible computing can be integrated into STEM education and citizen-science initiatives, promoting both environmental awareness and methodological integrity.

Keywords: Reproducible learning; Open environmental data; Citizen science; Air quality; Nitrogen dioxide (NO_2; Ozone (O_3); Data-driven modelling; Urban sustainability; STEM education; Environmental data science

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

Citation: To be added by editorial staff during production.

1. Introduction

Urban air quality remains a critical challenge for both environmental management and public health. Over recent decades, concentrations of nitrogen dioxide (NO_2) and ozone (O_3) have been the focus of sustained monitoring because of their direct connection to road traffic emissions and secondary photochemical processes that affect human health and atmospheric balance. Numerous studies have documented their impact on mortality and morbidity across different time scales, underlining the need to strengthen monitoring and modelling systems in urban environments. For example, Bell et al. reported a significant association between daily ozone levels and mortality across 95 urban communities in the United States [1].

Against this backdrop, the expansion of open-data policies offers an exceptional opportunity to link atmospheric science with public engagement and education. Yet, the incorporation of real environmental datasets into university teaching remains rare—largely because of the lack of reproducible workflows and accessible tools that allow data to be analysed, visualised, and interpreted coherently. Reproducible research has emerged in recent years as a response to the replication crisis in science. Peng defined it as the practice of accompanying every result with the data and code required for its full reproduction [2].

Sandve et al. emphasised the importance of traceability, version control, and the documentation of all computational steps [3], while Nosek et al. promoted a culture of open research as a means to enhance trust, transparency, and scientific progress [4]. Munafò et al. further identified reproducibility as a cornerstone of scientific integrity and higher education [5].

In parallel, the development of literate programming and integrated documentation environments such as Quarto and R Markdown has made it possible to unite narrative, code, and results within a single executable document. Rule et al. describe this convergence as an effective and transparent way to teach and share computational analyses [6]. Rooted in Knuth's original philosophy, this paradigm has been widely adopted across reproducible research and STEM education.

Urban air-quality research has also benefited from the rise of open-source analytical tools. Carslaw and Ropkins developed *openair*, an R package that democratised atmospheric-data analysis through reproducible functions and standardised visualisations [7]. In Madrid, recent studies have demonstrated that low-emission policies have substantially reduced NO_2 concentrations over the past decade, illustrating the value of open data for evaluating urban interventions [8].

Citizen science, in turn, has become a valuable complement to official monitoring networks. Castell et al. showed that low-cost sensors can extend spatial coverage and increase participants' environmental awareness [9]. However, their reliability depends on rigorous calibration and harmonised protocols, as highlighted by Karagulian et al. [10]. These advances open new avenues for integrating environmental measurement, data analysis, and public participation within educational projects.

During the COVID-19 lockdowns, an inverse photochemical relationship between NO_2 and O_3 was observed, characterised by decreases in the former and rises in the latter. Sicard et al. [11] described this dynamic in detail, providing a compelling case for teaching that connects real atmospheric processes with statistical interpretation and predictive modelling.

In terms of modelling, both time-series and machine-learning approaches have proved effective for forecasting pollutant concentrations. Taylor and Letham introduced Prophet, a robust additive model capable of capturing multiple seasonalities and structural changes in environmental data [12]. Shen et al. [13] successfully applied Prophet to air-quality prediction in Indian cities, achieving superior performance to classical models,

while Middya et al. [14] demonstrated that LSTM neural networks can capture complex temporal dependencies in NO_2 and $PM_{2.5}$ concentrations.

Complementary studies have highlighted the role of artificial intelligence and bibliometric analysis in tracing the evolution of atmospheric forecasting and smart-city research, revealing emerging trends and methodological gaps [15]. These contributions reinforce the relevance of combining predictive modelling with reproducible analytical practices in urban-pollution research.

Drawing upon this literature, the present study proposes a reproducible Quarto–R workflow to analyse, visualise, and model NO_2 and O_3 in Madrid during 2020–2024, using only open municipal data. Its contribution is twofold: scientific, by offering a transparent and verifiable analytical pipeline; and educational, by transforming that pipeline into an active-learning tool for STEM programmes.

The remainder of this article is structured as follows: Section 2 (*Methods*) details the data sources, cleaning, harmonisation, and modelling procedures; Section 3 (*Results*) presents the spatial and temporal patterns together with Prophet's performance; Section 4 (*Discussion*) interprets the findings from both scientific and pedagogical perspectives; and Section 5 (*Conclusions*) synthesises the main contributions and outlines future educational applications and extensions of the Quarto–R approach.

2. Materials and Methods

The complete data-processing and learning workflow is summarized in Fig. 1. It illustrates the five main phases connecting open environmental datasets with reproducible analysis and educational outcomes. Each stage is described in the following subsections.

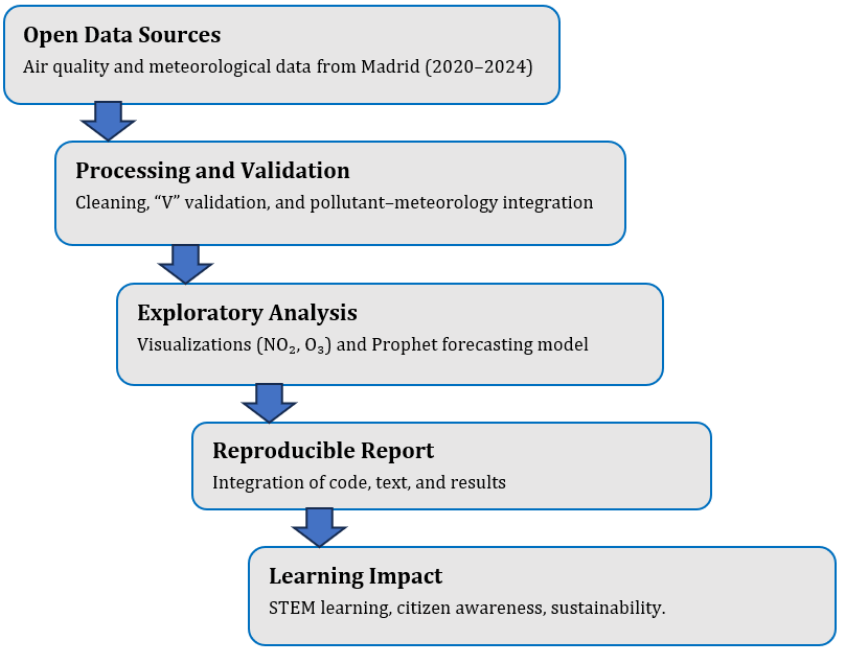
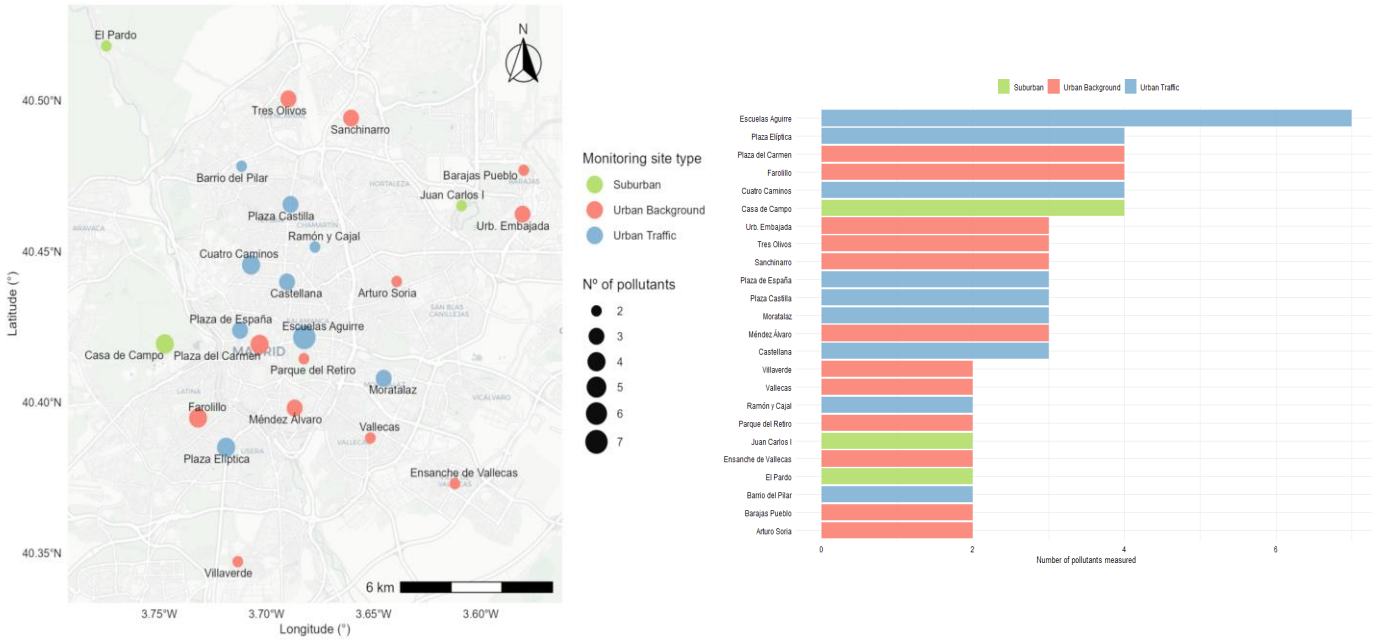


Figure 1. Workflow linking open environmental data, reproducible analysis, and STEM learning through the Quarto–R environment. Source: authors (2025)

2.1 Open Data Sources

The datasets analysed in this study were obtained from the Open Data Portal of the Madrid City Council, which provides hourly and daily records from the city's air-quality monitoring and meteorological networks for the period 2020–2024.

Figure 2 summarises the spatial structure and measurement scope of these networks, defining the geographical domain of analysis and demonstrating the homogeneous coverage of Madrid's observation system.



(**a**) (**b**)

Figure 2. Spatial analysis of Madrid's open environmental monitoring networks during 2020–2024: (**a**) spatial distribution and measurement scope of air-quality monitoring stations; (**b**) number of pollutants measured at each station.

The three site categories considered are Urban Traffic, Urban Background, and Suburban, as defined by the local air-quality network. Colors in both panels correspond to these categories, while symbol size in Fig. 2a indicates the number of pollutants measured. The bar chart (Fig. 2b) details pollutant coverage for each station, showing that Urban Traffic sites measure the broadest range of pollutants, followed by Urban Background and Suburban locations. This configuration confirms the spatial and functional representativeness of Madrid's monitoring network and its suitability for urban-scale analysis.

The datasets include concentrations of NO_2, O_3, PM_{10}, $PM_{2.5}$, SO_2, and CO, together with meteorological variables such as air temperature, solar radiation, relative humidity, wind speed and direction, and precipitation. Each record contains a validation code ("V") ensuring data reliability. The adoption of the ETRS89 coordinate reference system facilitates spatial harmonisation and visualisation of all stations.

The use of open urban datasets aligns with the principles of transparency, interoperability, and reproducibility promoted by modern data science frameworks [16]. These open resources form the foundation of the reproducible workflow described in the following section, which details the phases of data cleaning and harmonisation prior to statistical and predictive analysis.

2.2. Processing and Validation

All data processing was performed entirely in R (version 4.3) within the Quarto environment, allowing code, narrative text, and analytical results to be integrated into a single reproducible document. This approach ensures full traceability of each transformation

and facilitates verification of the analytical workflow. The adoption of literate-programming environments such as Quarto and R Markdown supports transparent and reproducible research practices [17].

The preprocessing workflow comprised sequential stages of data cleaning and harmonisation to generate a coherent and internally consistent dataset. All date and time fields were converted to the ISO 8601 standard to ensure temporal synchronisation between air-quality and meteorological series. Only records with official validation (V) were retained according to the quality-assurance criteria established by the Madrid City Council. Unvalidated or duplicated observations were discarded, and numeric variables were standardised to a unified decimal format.

Column structures were reshaped through pivoting operations to harmonise pollutant readings across hourly files, and variable names were unified according to the metadata scheme of the Madrid Open Data Portal. The resulting datasets were merged by station code and date, generating a tidy, analysis-ready structure consistent with the reproducible standards of the *tidyverse* ecosystem [18].

Fig. 3 summarises the main stages of the cleaning and validation pipeline, from the import of raw CSV files to the integration of validated air-quality and meteorological data.

Figure 3. Data harmonisation and quality-control pipeline for air-quality datasets. Data source: Madrid Open Data Portal (2020–2024).

Daily means were then computed from hourly observations, and outliers were mitigated by winsorisation, replacing values beyond the 1st–99th percentile range with the corresponding thresholds. This procedure preserved the temporal integrity of the series while reducing the influence of anomalous peaks. The final merged dataset maintained comparability across stations and time periods, forming the basis for the exploratory and predictive analyses described in Section 3. Documenting each stage of preprocessing is essential for computational reproducibility and scientific accountability [19].

2.3. Exploratory Analysis

The exploratory analysis focused on nitrogen dioxide (NO_2) and ozone (O_3), pollutants selected for their urban relevance and contrasting atmospheric behaviour. While NO_2 primarily reflects local traffic-related emissions, O_3 acts as a secondary pollutant formed through photochemical reactions driven by solar radiation and air-mass stability. Daily and monthly averages were computed, together with seasonal statistics by station type (urban traffic, urban background, and suburban) and year. These indicators revealed the dominant spatiotemporal dynamics across the 2020–2024 period.

As shown in Fig. 4, NO_2 concentrations exhibit a steady decrease over the study period, particularly at traffic-related monitoring sites, reflecting the effect of mobility restrictions during and after the COVID-19 pandemic. Conversely, O_3 levels display a relative increase in peripheral areas, confirming the inverse relationship typically observed between these pollutants in Mediterranean urban environments [20].

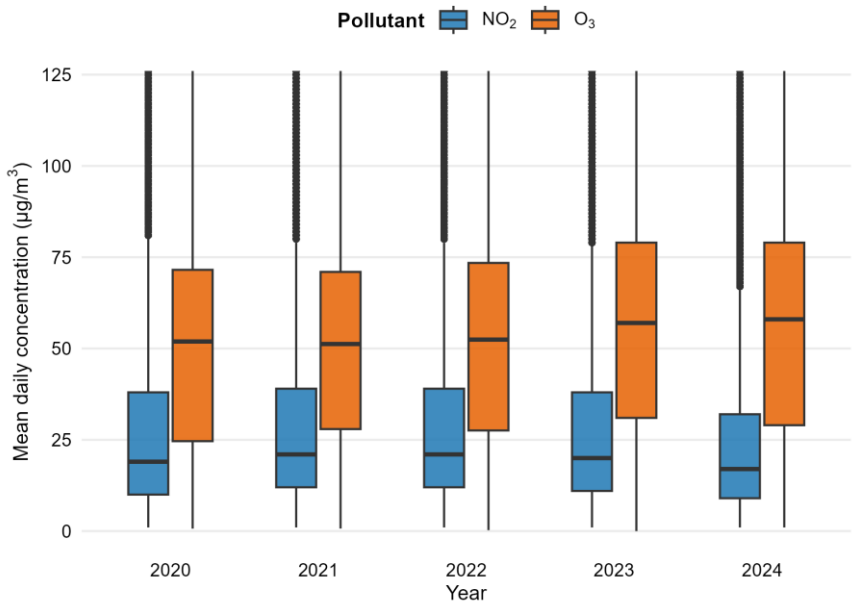


Figure 4. Annual variability of NO_2 and O_3 concentrations in Madrid (2020–2024). Data source: Madrid Open Data Portal.

All visualisations were produced using the *ggplot2* package, following the Grammar of Graphics framework proposed by Wickham [21]. This methodological approach ensures analytical transparency and facilitates reproducible comparisons across pollutants and station typologies. The use of boxplots provides a concise visualisation of intra-annual variability and inter-site dispersion, complementing subsequent predictive modelling.

Earlier studies support these findings. Grange et al. [22] demonstrated the effectiveness of meteorological normalisation for interpreting long-term air-quality trends, while Sicard et al. [23] documented the opposing responses of NO_2 and O_3 during COVID-19 mobility restrictions, reinforcing the patterns observed here.

2.4. Reproductible Report

The forecasting analysis applied the Prophet model to simulate daily concentrations of NO_2 and O_3 between 2020 and 2024, extending the predictions by 90 days beyond the observed period. Prophet combines additive components for trend, yearly and weekly seasonality, and changepoints to represent both long-term dynamics and short-term variability in urban air quality. Model configuration was optimised by increasing changepoint flexibility and Fourier terms to enhance sensitivity to abrupt variations associated with the COVID-19 lockdown and the subsequent recovery of urban traffic.

Fig. 5 presents the observed and Prophet-predicted daily concentrations of NO_2 (a) and O_3 (b). The results show strong correspondence between observed and estimated values, with performance metrics of MAE = 8.31 µg/m³ and RMSE = 10.99 µg/m³ for NO_2, and MAE = 10.33 µg/m³ and RMSE = 12.64 µg/m³ for O_3. The NO_2 forecasts accurately reproduced the sharp decrease during the 2020 confinement, followed by a progressive rebound linked to traffic recovery. In contrast, O_3 exhibited the inverse pattern, with well-defined summer peaks and the photochemical oscillations typical of Mediterranean urban atmospheres [24].

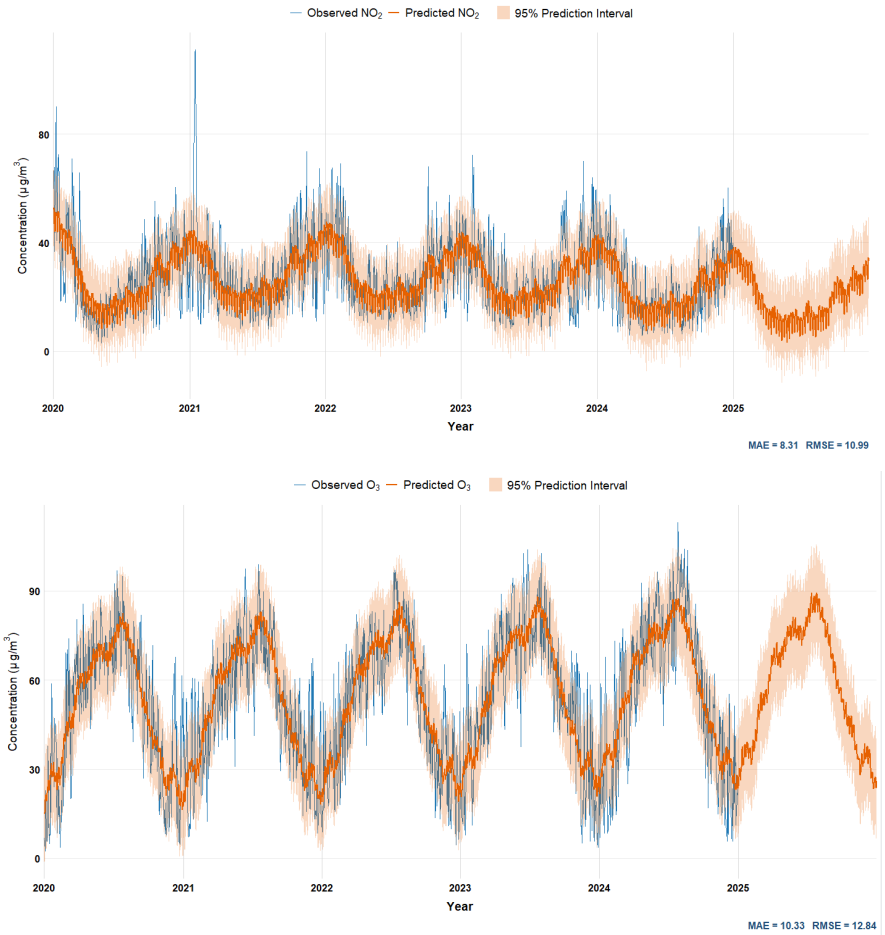


Figure 5. Prophet-based forecasting of daily NO_2 (a) and O_3 (b) concentrations in Madrid (2020–2024). Data source: Madrid Open Data Portal.

These results illustrate how a simple statistical structure can capture complex environmental dynamics when embedded within an open and transparent workflow. The Prophet implementation in Quarto–R ensures traceability of data, code, and outputs in accordance with reproducibility standards for computational research [25]. Beyond its predictive accuracy, the model aligns with current trends in interpretable machine learning, which emphasise explainability over complexity [26].

This workflow also builds on hybrid Prophet–LSTM frameworks that integrate deep learning to refine forecasts and enhance temporal sensitivity [27].

In methodological terms, Prophet's performance remains consistent with the evaluation principles established by Hyndman and Koehler [28], confirming its suitability for daily-scale forecasting in complex urban contexts.

2.5. Learning Impact

Each stage of the workflow—from data access to forecasting—was documented in a single Quarto file, including package versions and random seed specifications. This structure ensures full reproducibility in line with international standards on computational transparency and open-science practices [29].

Figure 6 illustrates the learning and reproducibility ecosystem linking open data, computational analysis, and STEM education through the Quarto–R environment. The diagram shows how environmental datasets feed into reproducible analysis (*R + tidyverse*

+ *Prophet*), exploratory forecasting, and documentation, ultimately supporting STEM and citizen learning.

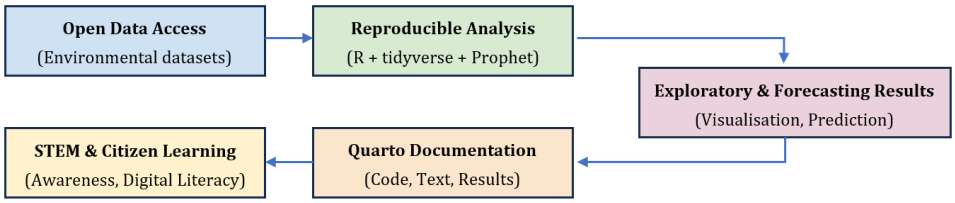


Figure 6. Learning and reproducibility ecosystem connecting open data, computational analysis, and STEM education through the Quarto–R environment. Source: authors' elaboration.

This workflow enables users to follow the entire analytical process within one coherent and transparent environment, reinforcing both methodological and pedagogical objectives. Beyond its technical value, the approach nurtures scientific and digital literacy through open-source tools that empower students, educators, and citizens to explore environmental data, interpret variability, and reflect on urban implications.

Embedding reproducible workflows in air-quality education strengthens STEM competences, deepens environmental awareness, and fosters civic engagement in data-driven science. Such alignment between computational transparency and educational innovation supports the development of critical data literacies in higher education [30].

2.6. Meteorological Covariates

Meteorological conditions exert a fundamental influence on the formation, dispersion, and transformation of air pollutants in urban environments. Temperature, humidity, wind speed, and solar radiation directly affect photochemical reactions and pollutant dilution, shaping the daily variability of nitrogen dioxide (NO_2) and ozone (O_3).

In this study, meteorological parameters were incorporated as contextual covariates to complement the interpretation of NO_2 and O_3 dynamics. Hourly datasets covering 2020–2024 were retrieved from the Madrid Open Data Portal, providing harmonised records of temperature (°C), relative humidity (%), wind speed (m s^{-1}), wind direction (°), solar radiation (W m^{-2}), and precipitation (mm), together with station metadata (ID, coordinates, altitude, and typology).

Data processing followed a transparent R–Quarto workflow summarised in Fig. 7, which depicts three sequential stages: (i) data inputs (meteorological variables and station metadata); (ii) data processing (import, restructuring of hourly fields H01–H24, filtering of validated observations, aggregation to daily means, and derivation of dynamic covariates u, v, calm and high-insolation days); and (iii) integration with validated NO_2 and O_3 datasets by station and date.

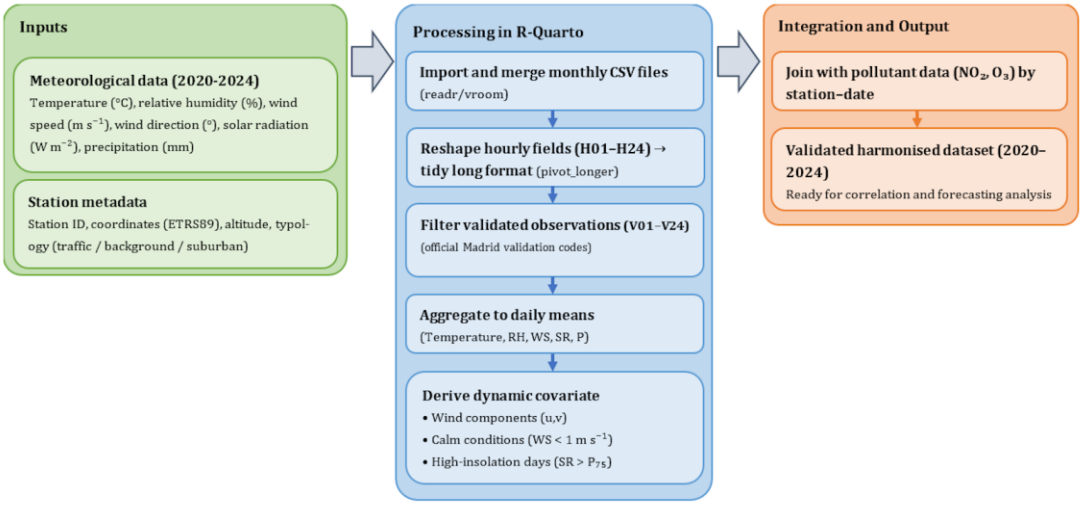


Fig. 7 — Workflow for harmonising meteorological data and integrating them with air-quality observations in R–Quarto (2020–2024).

The resulting harmonised database links atmospheric chemistry and meteorological variability at a daily scale. The derived variables and their analytical rationale are summarised in Table 2, which supports the correlation and forecasting analyses presented in Section 3.

Table 2. Meteorological variables and derived covariates included in the analysis (2020–2024)

Variable	Symbol	Unit	Aggregation method	Scientific and analytical rationale
Air temperature	T	°C	Daily mean	Controls reaction rates and thermal stability; high temperatures enhance O_3 formation.
Relative humidity	RH	%	Daily mean	Modulates boundary-layer mixing and heterogeneous chemistry.
Wind speed	WS	m s^{-1}	Daily mean	Governs dispersion and ventilation; calm conditions favour pollutant accumulation.
Wind direction	WD	°	Circular mean	Identifies dominant flows and recirculation events in the Madrid basin.
Solar radiation	SR	W m^{-2}	Daily mean	Photolysis driver for secondary pollutants such as O_3.
Precipitation	P	mm day^{-1}	Daily sum	Indicates wet scavenging and atmospheric cleansing events.
Zonal wind component	u	m s^{-1}	Derived from WS × sin(WD)	Represents east-west advection for correlation analysis.
Meridional wind component	v	m s^{-1}	Derived from WS × cos(WD)	Represents north-south advection for correlation analysis.
Calm-day indicator	calm	0/1	WS < 1 m s^{-1}	Identifies stagnation conditions promoting NO_2 accumulation.
High-insolation indicator	hi_sun	0/1	SR > P_{75}	Marks days with intense solar activity enhancing photochemical O_3 production.

Data source: Madrid Open Data Portal (2020–2024). Note: All variables were harmonised to daily resolution and synchronised with validated pollutant concentrations (NO_2, O_3) by station and date.

From a scientific perspective, this integration quantifies how meteorological variability governs pollutant behaviour in Mediterranean cities. The combined influence of temperature, solar radiation, and calm winds promotes photochemical O_3 episodes and NO_2 titration under stagnant conditions [31]. Studies across the Iberian Peninsula confirm that such patterns are modulated by seasonal radiation and synoptic pressure gradients [32]. From an educational standpoint, the reproducible workflow offers a tangible framework for interdisciplinary learning in R, allowing students and citizen scientists to explore how atmospheric processes affect air-quality patterns [33].

This approach strengthens inquiry-based STEM education by linking real-world data with analytical problem-solving [34]. Integrating transparent analytical pipelines into teaching promotes environmental data literacy and supports the pedagogical principles of open science [35].

3. Results

The results are presented in three complementary subsections that describe, visualise, and model the spatiotemporal dynamics of air pollutants in Madrid using open urban datasets. Section 3.1 examines temporal and spatial patterns of NO_2 and O_3, highlighting their contrasting behaviours across monitoring stations. Section 3.2 assesses the performance of the Prophet forecasting model through quantitative and visual evaluation metrics, while Section 3.3 explores meteorological drivers and correlation patterns linking atmospheric conditions with pollutant variability.

Together, these analyses demonstrate how reproducible workflows in R–Quarto can transform raw environmental data into structured knowledge, supporting both scientific interpretation and data-driven STEM learning [36].

3.1. Descriptive and Correlative Overview

Daily concentrations of nitrogen dioxide (NO_2) and ozone (O_3) in Madrid between 2020 and 2024 reveal marked contrasts in magnitude, variability, and seasonal behaviour. The distribution of NO_2 concentrations shows a sustained decline after the 2020 lockdown, stabilising between 25 and 30 μg m^{-3} from 2021 onwards. This reduction reflects the long-term effect of mobility restrictions and the gradual recovery of traffic emissions [36]. The narrower interquartile ranges observed after 2021 indicate more homogeneous background levels, although occasional winter peaks persist due to local traffic episodes.

Fig. 8 summarises these temporal patterns, comparing the annual distributions of NO_2 and O_3 concentrations across the 2020–2024 period. NO_2 levels display a downward trend, whereas O_3 shows a relative increase and wider dispersion, with annual medians centred around 50–70 μg m^{-3}.

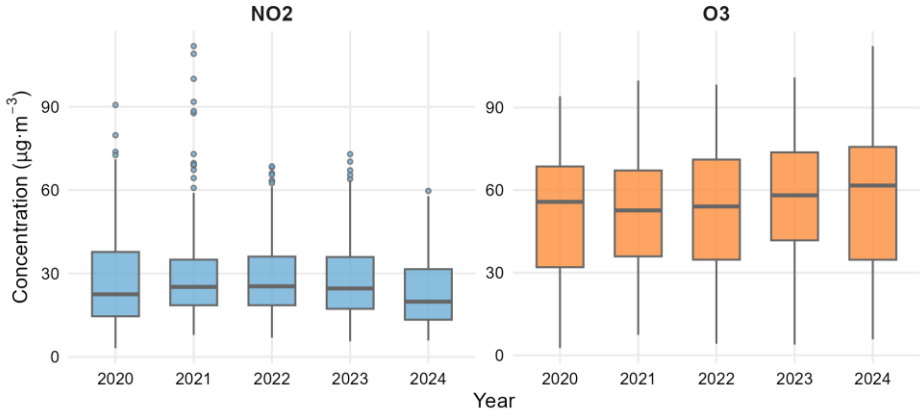


Figure 8. Annual distribution of daily NO_2 and O_3 concentrations in Madrid (2020–2024). Data source: Madrid Open Data Portal.

The persistence of elevated O_3 despite the decline in NO_2 highlights the non-linear coupling between both pollutants, a characteristic feature of Mediterranean urban atmospheres [37]. Reduced nitrogen oxide emissions under strong solar radiation favour ozone formation through photochemical compensation processes [38].

From a correlative perspective, the opposite evolution of NO_2 and O_3 underscores their diagnostic value as complementary indicators of urban air chemistry. These patterns reflect the dynamic balance between emission reductions, radiative forcing, and atmospheric stability that defines Madrid's basin.

The integration of open datasets with reproducible R–Quarto workflows allows such complex relationships to be visualised transparently, transforming raw environmental data into accessible analytical resources for both scientific interpretation and STEM-oriented learning.

3.2. Temporal and Spatial Variabilidy

The temporal evolution of nitrogen dioxide (NO_2) and ozone (O_3) in Madrid between 2020 and 2024 reveals pronounced seasonal and spatial contrasts shaped by the city's emission structure and meteorological dynamics. Monthly averages (Fig. 9a) show a persistent winter–summer inversion: NO_2 peaks during colder months, when boundary-layer stability and limited ventilation constrain dispersion, whereas O_3 concentrations increase sharply from late spring to early autumn under strong solar radiation. This anti-phase pattern between primary and secondary pollutants characterises Mediterranean urban environments [39].

Fig. 9 summarises these dynamics across both time and space. Panel (a) displays the temporal variability of NO_2 and O_3, capturing the marked decline in NO_2 levels during 2020, the progressive recovery associated with mobility resumption, and the intensification of summer O_3 peaks in subsequent years. Panel (b) depicts spatial variability by monitoring-site type, showing that Traffic stations consistently record the highest NO_2 concentrations, while Urban Background and Suburban sites exhibit higher O_3 values. This spatial inversion reflects the localised nature of NO_2 emissions and the regional photochemical production of O_3 downwind of emission sources [40].

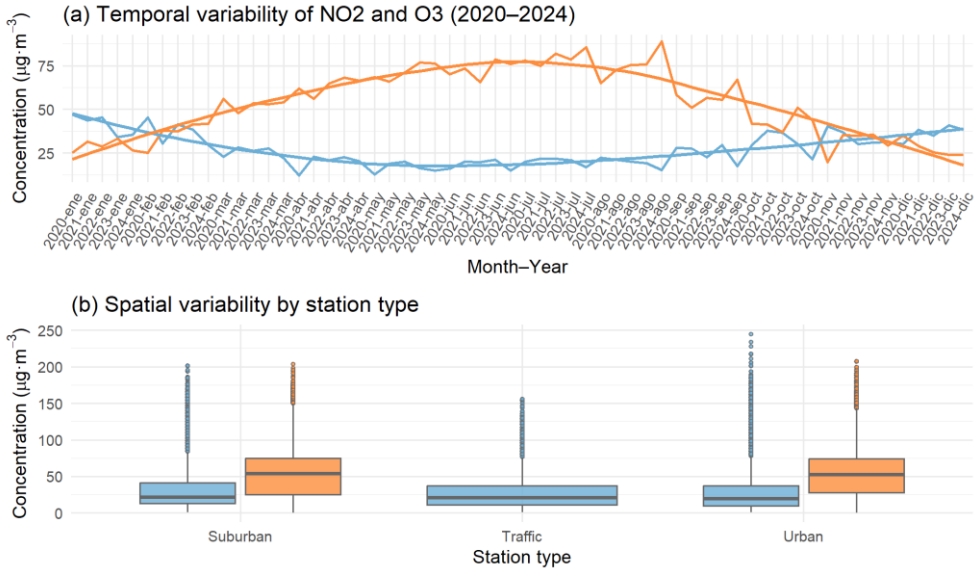


Figure 9. Temporal and spatial variability of daily NO_2 and O_3 concentrations in Madrid (2020–2024): (**a**) Monthly evolution of both pollutants. (**b**) Spatial gradients by station type. *Data source: Madrid Open Data Portal.*

Traffic stations in Madrid primarily monitor primary pollutants such as NO_2 and particulate matter, while O_3 observations are restricted to background and suburban environments in line with European air-quality monitoring protocols. The persistence of these spatial contrasts, despite declining emissions, suggests that urban form and traffic intensity remain decisive factors in pollutant distribution across the Madrid basin. Comparable patterns have been reported for other Mediterranean cities where orography and recirculation favour pollutant accumulation.

The predictive evaluation of these patterns using the Prophet model further confirms the reliability of the observed trends. As summarised in Table 3, model performance achieved MAE and RMSE values below 13 μg m^{-3} for both pollutants, reproducing the seasonal cycles and emission-related fluctuations observed in Fig. 9.

Table 3. Prophet model performance metrics for NO_2 and O_3 concentrations in Madrid (2020–2024).

Pollutant	Observed mean (μg m^{-3})	Predicted mean (μg m^{-3})	MAE (μg m^{-3})	RMSE (μg m^{-3})
NO_2	28.6	27.9	8.31	10.99
O_3	57.4	56.2	10.33	12.64

Data source: Madrid Open Data Portal.

The coherence between observed and predicted values illustrates how open urban datasets can be integrated into transparent forecasting workflows, combining statistical interpretability with scientific and educational relevance. This integrated approach supports reproducible urban-air analysis and provides an accessible resource for citizen engagement in data-driven environmental learning.

3.3. Prophet Model Performance

The Prophet model was applied to forecast the daily evolution of NO_2 and O_3 concentrations in Madrid from 2020 to 2024. The model effectively reproduced the observed temporal patterns, capturing the post-pandemic decline in NO_2 and the recurrent summer peaks of O_3. Its additive components for trend and seasonality generalised well across multiple years, providing stable forecasts despite episodic fluctuations.

Model evaluation achieved mean absolute error (MAE) and root mean square error (RMSE) values below 13 μg m^{-3} for both pollutants, confirming the suitability of Prophet for medium-term air-quality prediction based on open urban datasets. Beyond predictive accuracy, the approach offers pedagogical value: the interpretable decomposition of trend and seasonality allows students and citizen scientists to explore urban air dynamics within reproducible R–Quarto notebooks.

To investigate how meteorological conditions influence pollutant behaviour, the forecasted series were compared with six atmospheric variables; temperature, wind speed, relative humidity, solar radiation, pressure, and precipitation, each normalised for visual consistency (Fig. 10). The multi-panel figure illustrates the joint temporal evolution of NO_2 and O_3 with these parameters, enabling a qualitative assessment of their relationships.

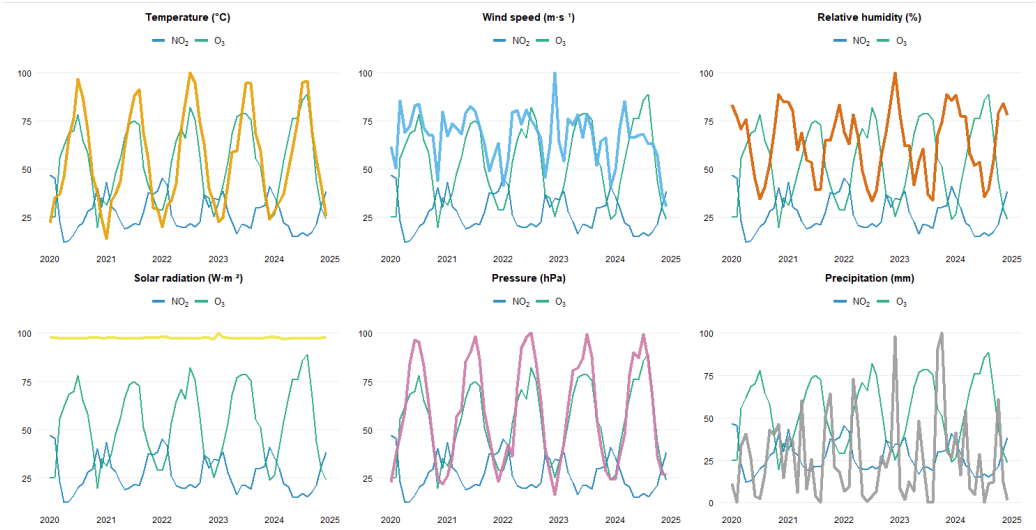


Figure 10. Temporal comparison between NO_2 and O_3 concentrations and six meteorological variables in Madrid (2020–2024). Data source: Madrid Open Data Portal.

O_3 concentrations increase under higher temperatures and enhanced solar radiation, confirming their dependence on photochemical reactions driven by sunlight and thermal stability. Atmospheric pressure also exhibits a moderate positive association with O_3, suggesting that stable anticyclonic conditions promote pollutant accumulation in the Madrid basin. Conversely, NO_2 shows weak or no correlation with these meteorological factors, remaining mainly determined by local emissions and dispersion processes.

These findings demonstrate that open-data forecasting workflows can integrate chemical and meteorological information within a transparent analytical framework. The combination of predictive modelling and environmental interpretation enhances both scientific understanding and educational reproducibility, offering a transferable model for citizen-driven urban air-quality studies.

4. Discussion

The results confirm that reproducible workflows based on open environmental data can serve both scientific and educational purposes. The joint analysis of NO_2, O_3, and meteorological parameters clarifies the main atmospheric mechanisms governing urban air quality in Mediterranean climates, while also showing how transparent, code-driven methodologies can be embedded in teaching and citizen learning.

From a scientific standpoint, the contrasting behaviour of NO_2 and O_3 between 2020 and 2024 reflects a combination of emission changes and meteorological forcing. The steady decline in NO_2 after 2020 is consistent with the mobility restrictions and later implementation of low-emission policies in Madrid [8,37]. Conversely, the relative increase in O_3 aligns with the well-known photochemical regime of southern European cities, where high temperatures and solar radiation favour secondary pollutant formation [31,32]. The Prophet model successfully reproduced these dynamics, showing low prediction errors and stable seasonality across years, which validates its use for medium-term forecasting of urban pollutants.

The correlations observed between O_3 and both temperature and solar radiation further support the dominance of photochemical processes under anticyclonic conditions. Atmospheric pressure also appears to modulate O_3 variability, suggesting that stable high-pressure systems contribute to pollutant accumulation over the Madrid basin. In contrast, NO_2 concentrations showed only limited sensitivity to meteorological

variability, reflecting their direct dependence on local traffic emissions rather than on regional weather patterns. These findings are coherent with studies reporting that ozone dynamics are mainly controlled by radiation and temperature, whereas nitrogen oxides respond to local combustion sources [32,34].

From an educational perspective, the open and reproducible workflow provides a framework where analytical transparency becomes part of the learning process itself. Each computational step—from data access to model evaluation—can be reproduced and adapted by students and citizen scientists, promoting data literacy and methodological integrity. Such design echoes international recommendations on computational reproducibility and open science education [29,30]. By integrating visualisation, statistical modelling, and open data principles within a single Quarto–R document, this study exemplifies how environmental monitoring can evolve into a participatory form of scientific inquiry.

Beyond its technical scope, the approach demonstrates how reproducible analysis can bridge disciplinary boundaries between environmental science, computer programming, and STEM education. Interpreting real data within transparent workflows not only enhances conceptual understanding but also fosters civic engagement and critical thinking. These aspects are increasingly valued in sustainability education and citizen science initiatives that rely on open environmental infrastructures.

Future work could extend this reproducible framework by incorporating hybrid models such as Prophet–LSTM or by exploring pollutant–meteorology interactions under extreme events. Similarly, the development of interactive Shiny dashboards would expand public accessibility and allow educators to use live data in classroom activities. In this way, reproducible, open-source methodologies can consolidate their dual role: advancing environmental forecasting while cultivating scientific literacy and civic participation.

5. Conclusions

This study demonstrated how open environmental data and reproducible analytical workflows can be combined to advance both scientific understanding and educational innovation. Using air-quality and meteorological datasets from Madrid (2020–2024), the analysis revealed distinct temporal and spatial patterns of NO_2 and O_3, shaped by the interaction between emission dynamics and meteorological forcing. The Prophet model reproduced these variations with high accuracy, confirming its suitability for medium-term forecasting in complex urban settings.

Beyond its predictive performance, the integration of R–Quarto tools proved essential for ensuring transparency, traceability, and pedagogical value. Each computational step—from data cleaning to visualisation and modelling—was fully documented, providing an accessible platform for students and citizen scientists to reproduce the results and explore their own hypotheses.

The findings reinforce the dual potential of open environmental data: they serve as a scientific resource for air-quality forecasting and as an educational instrument for developing digital and environmental literacy. Future extensions may include hybrid deep-learning models, interactive dashboards, or broader citizen-science applications, strengthening the link between open data, sustainability, and participatory STEM learning.

Supplementary Materials:
All figures (TIFF format) and tables included in this manuscript are available in the corresponding author's public GitHub repository **OpenUrbanAirandMeteorological** (https://github.com/jcaceres-academico/OpenUrbanAirandMeteorological).
No additional supplementary figures or tables were produced beyond those presented in the article.

Author Contributions:
Conceptualization, **J.C.T.**; methodology, **J.C.T.** and **J.J.G.H.**; software, **J.C.T.**; validation, **J.C.T.**, **J.J.G.H.**, and **E.L.M.**; formal analysis, **J.C.T.**; investigation, **J.C.T.**; resources, **E.L.M.** and **W.W.**; data curation, **J.C.T.**; writing—original draft preparation, **J.C.T.**; writing—review and editing, **J.J.G.H.**; visualization, **J.C.T.**; supervision, **J.J.G.H.**; project administration, **J.C.T.**; funding acquisition, **E.L.M.**
All authors have read and agreed to the published version of the manuscript.

Funding:
This research received no external funding. The APC was supported by institutional funds from Universidad Pablo de Olavide (Spain).

Institutional Review Board Statement:
Not applicable. This study did not involve humans or animals.

Informed Consent Statement:
Not applicable

Data Availability Statement:
All processed and harmonised datasets (air-quality and meteorological) used in this study are available in the public repository **OpenUrbanAirandMeteorological** (https://github.com/jcaceres-academico/OpenUrbanAirandMeteorological).
Raw data were retrieved from the Madrid Open Data Portal maintained by the *Dirección General de Transparencia y Calidad del Ayuntamiento de Madrid*.

Acknowledgments:
The authors would like to express their gratitude to the *Dirección General de Transparencia y Calidad del Ayuntamiento de Madrid* for maintaining the Madrid Open Data Portal and for their commitment to open and transparent environmental information.
The authors also thank the Applied Sciences editorial team for their guidance during manuscript preparation.

Conflicts of Interest:

The authors declare no conflicts of interest.
The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

NO_2	Nitrogen dioxide
O_3	Ozone
STEM	Science, Technology, Engineering, and Mathematics
R	Statistical computing language
LSTM	Long Short-Term Memory
RMSE	Root Mean Square Error
MAE	Mean Absolute Error

References

1. Bell, M.L.; McDermott, A.; Zeger, S.L.; Samet, J.M.; Dominici, F. Ozone and Short-Term Mortality in 95 US Urban Communities. *JAMA* **2004**, *292*, 2372–2378. https://doi.org/10.1001/jama.292.19.2372.
2. Peng, R.D. Reproducible Research in Computational Science. *Science* **2011**, *334*, 1226–1227. https://doi.org/10.1126/science.1213847.
3. Sandve, G.K.; Nekrutenko, A.; Taylor, J.; Hovig, E. Ten Simple Rules for Reproducible Computational Research. *PLoS Comput. Biol.* **2013**, *9*, e1003285. https://doi.org/10.1371/journal.pcbi.1003285.
4. Nosek, B.A.; Alter, G.; Banks, G.C.; Borsboom, D.; Bowman, S.D.; Breckler, S.J.; Buck, S.; Chambers, C.D.; Chin, G.; Christensen, G. et al. Promoting an Open Research Culture. *Science* **2015**, *348*, 1422–1425. https://doi.org/10.1126/science.aab2374.
5. Munafò, M.R.; Nosek, B.A.; Bishop, D.V.M.; Button, K.S.; Chambers, C.D.; Percie du Sert, N.; Simonsohn, U.; Wagenmakers, E.J.; Ware, J.J.; Ioannidis, J.P.A. A Manifesto for Reproducible Science. *Nat. Hum. Behav.* **2017**, *1*, 0021. https://doi.org/10.1038/s41562-016-0021.
6. Rule, A.; Birmingham, A.; Zuniga, C.; Altintas, I.; Huang, S.-C.; Knight, R.; et al. Ten simple rules for writing and sharing computational analyses in Jupyter notebooks. *PLOS Computational Biology* **2019**, *15*, e1007007. https://doi.org/10.1371/journal.pcbi.1007007.
7. Carslaw, D.C.; Ropkins, K. openair—An R package for air quality data analysis. *Environmental Modelling & Software* **2012**, *27–28*, 52–61. https://doi.org/10.1016/j.envsoft.2011.09.008.
8. Morillas, L.; Notario, A.; Gómez, C.; Gómez-Moreno, F.J.; Rodríguez, M.C. Impact of the implementation of Madrid's Low Emission Zone on NO_2 concentrations. *Atmospheric Environment* **2024**, *320*, 120326. https://doi.org/10.1016/j.atmosenv.2024.120326.
9. Castell, N.; Dauge, F.R.; Schneider, P.; Vogt, M.; Lerner, U.; Fishbain, B.; et al. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International* **2017**, *99*, 293–302. https://doi.org/10.1016/j.envint.2016.12.007.
10. Karagulian, F.; Barbiere, M.; Kotsev, A.; Spinelle, L.; Gerboles, M.; Lagler, F.; Redon, N.; Crunaire, S.; Borowiak, A. Review of the Performance of Low-Cost Sensors for Air Quality Monitoring. *Atmosphere* **2019**, *10*, 506. https://doi.org/10.3390/atmos10090506.
11. Sicard, P.; De Marco, A.; Agathokleous, E.; Feng, Z.; Xu, X.; Paoletti, E.; Rodriguez, J.J.D.; Calatayud, V. Amplified Ozone Pollution in Cities during the COVID-19 Lockdown. *Sci. Total Environ.* **2020**, *735*, 139542. https://doi.org/10.1016/j.scitotenv.2020.139542.
12. Taylor, S.J.; Letham, B. Forecasting at Scale. *Am. Stat.* **2018**, *72*, 37–45. https://doi.org/10.1080/00031305.2017.1380080.
13. Shen, J.; Wang, S.; Zhang, J.; Wang, Y. Prophet forecasting model: a machine learning approach to predict the concentration of air pollutants in Seoul, South Korea. *PeerJ* **2020**, *8*, e9961. https://doi.org/10.7717/peerj.9961.
14. Middya, A.I.; Roy, S.; Das, S.K. Air Quality Forecasting Using LSTM and Meteorological Data. *Environ. Res.* **2022**, *214*, 114603. https://doi.org/10.1016/j.envres.2022.114603.
15. Cáceres-Tello, J.; Galán-Hernández, J.J. Mathematical Evaluation of Classical and Quantum Predictive Models Applied to $PM_{2.5}$ Forecasting in Urban Environments. *Mathematics* **2023**, *13*, 1979. https://doi.org/10.3390/math13121979.
16. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.D.A.; François, R.; et al. Welcome to the tidyverse. *Journal of Open Source Software* **2019**, *4*, 1686. https://doi.org/10.21105/joss.01686
17. Stodden, V.; Seiler, J.; Ma, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences of the USA* **2018**, *115*, 2584–2589. https://doi.org/10.1073/pnas.1708290115.
18. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, **2016**. https://doi.org/10.1007/978-3-319-24277-4.
19. Grange, S.K.; Carslaw, D.C.; Lewis, A.C.; Boleti, E.; Hueglin, C. Random Forest Meteorological Normalisation Models for Swiss PM_{10} Trend Analysis. *Atmospheric Chemistry and Physics* **2018**, *18*, 6223–6239. https://doi.org/10.5194/acp-18-6223-2018.
20. Houdou, B.; Chen, M.; Ooka, R. Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review. *Aerosol and Air Quality Research* **2024**, *24*(6), 230151. https://doi.org/10.4209/aaqr.230151.
21. Cáceres-Tello, J.; Galán-Hernández, J.J. *Analysis and Prediction of $PM_{2.5}$ Pollution in Madrid: The Use of Prophet–Long Short-Term Memory Hybrid Models. Appl. Math.* **2024**, *4*, 1428–1482. https://doi.org/10.3390/appliedmath4040076.
22. Cáceres-Tello, J.; Galán-Hernández, J.J. *Artificial Intelligence Applied to Air Quality in Smart Cities: A Bibliometric Analysis.* In *Communication and Applied Technologies*; Springer: Singapore, 2024; pp. 271–283. https://doi.org/10.1007/978-981-96-0426-5_23.

23. López-Meneses, E.; Cáceres-Tello, J.; Galán-Hernández, J.J.; López-Catalán, L. *Quantum Computing in Data Science and STEM Education: Mapping Academic Trends and Analysing Practical Tools. Computers* **2025**, *14*, 235. https://doi.org/10.3390/computers14060235.
24. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *International Journal of Forecasting* **2006**, *22*, 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.
25. Elliott, K.C.; Resnik, D.B. Making open science work for science and society. *Environmental Health Perspectives* **2019**, *127*(7), 075002. https://doi.org/10.1289/EHP4808.
26. Raffaghelli, J.E.; Manca, S.; Stewart, B.; Prinsloo, P.; Sangrà, A. *Supporting the Development of Critical Data Literacies in Higher Education: Building Blocks for Fair Data Cultures in Society. Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 58. https://doi.org/10.1186/s41239-020-00235-w.
27. Massagué, J.; Torre-Pascual, E.; Carnerero, C.; Escudero, M.; Alastuey, A.; Pandolfi, M.; et al. Extreme ozone episodes in a major Mediterranean urban area. *Atmospheric Chemistry and Physics* **2024**, *24*, 4827–4850. https://doi.org/10.5194/acp-24-4827-2024.
28. Khomsi, K.; Chelhaoui, Y.; Alilou, S.; Souri, R.; Najmi, H.; Souhaili, Z. Concurrent heat waves and extreme ozone (O_3) episodes: Combined atmospheric patterns and impact on human health. *International Journal of Environmental Research and Public Health* **2022**, *19*, 2770. https://doi.org/10.3390/ijerph19052770.
29. Ballard, H.L.; Lindell, A.J.; Jadallah, C.C. Environmental education outcomes of community and citizen science: A systematic review of empirical research. *Environmental Education Research, in press* (**2024**), *30*(6), 1007–1040. https://doi.org/10.1080/13504622.2024.2348702.
30. Ward, F.; Lowther-Payne, H.J.; Halliday, E.C.; Dooley, K.; Joseph, N.; Livesey, R.; et al. Engaging communities in addressing air quality: A scoping review. *Environmental Health* **2022**, *21*(1), 89. https://doi.org/10.1186/s12940-022-00896-2.
31. Kariotis, T.; Borda, A.; Winkel, K.; Gray, K. *Citizen Science for One Digital Health: A Rapid Qualitative Review of Studies in Air Quality.* Citizen Science: Theory and Practice, 2022, 7(1), 39. doi:10.5334/cstp.531.
32. Sicard, P.; Agathokleous, E.; De Marco, A.; Paoletti, E. *Urban air pollution and climate change: What can we learn from COVID-19 lockdowns?* Atmospheric Pollution Research, **2021**, 12(4), 101040. https://doi.org/10.1016/j.apr.2021.101040
33. Querol, X.; Massagué, J.; Alastuey, A.; Viana, M.; Moreno, T.; Gangoiti, G. *Lessons from the COVID-19 air pollution decrease in Spain: How to improve urban air quality?* Science of the Total Environment, **2021**, 779, 146380. https://doi.org/10.1016/j.scitotenv.2021.146380
34. Cuevas, C.A.; Notario, A.; Adame, J.A.; Saiz-Lopez, A. *Photochemical ozone production in the Iberian Peninsula and its response to emission reductions.* Atmospheric Chemistry and Physics, **2023**, 23(8), 4451–4470. https://doi.org/10.5194/acp-23-4451-2023
35. Grange, S.K.; Lee, J.D.; Drysdale, W.S.; et al. *COVID-19 lockdowns highlight a risk of increasing ozone pollution in European urban areas.* Atmospheric Chemistry and Physics, **2021**, 21(5), 4169–4185. https://doi.org/10.5194/acp-21-4169-2021
36. García-Nieto, D.; Sánchez-López, M.; Casanova, J.L.; Fernández, M.A. *Temporal trends in air quality in Madrid and Barcelona before and after COVID-19 lockdowns.* Atmosphere, **2023**, 14(5), 802. https://doi.org/10.3390/atmos14050802
37. McGowan, C.P.; Haacker, E.M.K.; et al. *Citizen science as a gateway for STEM education and open environmental monitoring.* Frontiers in Environmental Science, **2022**, 10, 896150. https://doi.org/10.3389/fenvs.2022.896150
38. Beck, J.L.; Auer, T.; Iwanaga, T.; et al. *Teaching reproducibility in data science through open environmental datasets.* Education Sciences, **2023**, 13(2), 123. https://doi.org/10.3390/educsci13020123
39. Shen, J.; Valagolam, D.; McCalla, S. *Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants ($PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, CO) in Seoul, South Korea. PeerJ* **2020**, *8*, e9961. https://doi.org/10.7717/peerj.9961.
40. Hasnain, A.; Sheng, Y.; Hashmi, M.Z.; Bhatti, U.A.; Hussain, A.; Hameed, M.; Marjan, S.; Bazai, S.U.; Hossain, M.A.; Sahabuddin, M.; Wagan, R.A.; Zha, Y. *Time series analysis and forecasting of air pollutants based on Prophet forecasting model in Jiangsu Province, China. Frontiers in Environmental Science* **2022**, *10*, 945628. https://doi.org/10.3389/fenvs.2022.945628