MODELAIR – DELIVERABLE

D4.1 – PERFORMANCE OF NEW TOOLS FOR DIMENSIONALITY REDUCTION AND FLOW PATTERNS IDENTIFICATION IN URBAN FLOWS

This report is part of a project that has received funding from the European Union's Horizon Europe MSCA Doctoral Networks 2021 programme under **Grant Agreement No. 101072559**

Deliverable number: D4.1

Due date: 31st December 2024

Type[1]: R

Dissemination Level[1]: PU

Work Package: WP4

Lead Beneficiary: Universidad Politécnica de Madrid (UPM)

Contributing Beneficiaries: BSC, ARUP, AQC

1

Type	R = Report ADM = Administrative PDE = diss./ex. O = Other DEC = Websites, patents filing, press & media actions, videos, etc.
Dissemination Level	PU = Public CO = Confidential, only for members of the consortium (including the Commission Services) CI = Classified SEN = Sensitive, limited under the conditions of the Grant Agreement

DOCUMENT HISTORY

Deliverable leader: Soledad Le Clainche / Guillermo Barragán / Arindam Sengupta (UPM)

E-mail of lead author: soledad.leclainche@upm.es

Reviewer(s): Alessandro Parente (ULB)

Version	Date	Description
0.1	02/12/24	Draft
1.0	17/12/2024	**Final version**

Abstract

This report describes the activities and results of MODELAIR that focus on the activities related to data analysis and feature extraction described in WP4. The report focuses on describing the tools developed to post-process numerical and experimental databases. Two groups of tools have been developed.

The first group of tools is suitable for the analysis of numerical databases and complex experiments. These tools reduce data dimensionality (from thousand or even millions of grid points- as it is usual in numerical databases) to a few modes representing the main physics of the flow. This information has been exploited to create a reduced-order model combined with deep learning strategies. The model can be used to fill-in multi-parametric databases, enhance data resolution and predict the evolution in time of the flow.

The tools have been developed based on the performance of several fluid dynamics problems modelling laminar and turbulent flows from numerical and experimental databases. The tools have been extended to be applied in databases modelling urban flows. Their performance has been tested in a numerical database modelling a highlight polluted area from Madrid, ES, represented by the district of Tetuan, in the south of the city.

The second group of tools uses the information collected in sensors monitoring air quality spread in cities. These tools aim to find correlation between weather conditions and pollutant concentration in different areas of the city and try to predict the evolution of pollutant concentration as function of external variables. The tools have been tested in databases from Bristol, UK, and Madrid.

Details about air pollution in Bristol connected to the databases analyses are presented in deliverable D2.2.

Details about air pollution in Madrid connected to the databases analyses are presented in deliverable D2.4.

During the next years, the robustness and range of applicability of these tools will be extended and tested in new databases modelling some areas of Brussels, BE. Details about the databases from Brussels are presented in deliverable D2.3.

Also, during the next years, the tools will be tested in complex numerical databases, calculated with a very high level of accuracy, and experimental databases, with the aim at understanding physical principles connected to the presence of air pollution. Details about these numerical and experimental databases are presented in deliverables D3.2, and D3.3.

This report presents a collaborative work between UPM (DC1, DC2), that has developed the tools described in the report, AQC (DC6) that has provided the databases of air quality monitoring sensors from Bristol and provided guidance and support on the analysis, ARUP (DC7) that has guided the collection of databases of air quality monitoring sensors from Madrid and, BSC (DC5) that has provided the numerical databases from Madrid.

Keywords:

Data analysis, post-processing tools, patterns identification, data dimensionality reduction, reduced order models

Acronyms

CFD: Computational fluid dynamics

DMD: dynamic mode decomposition

HOSVD: high order singular value decomposition

LES: Large eddy simulations

ML: machine learning

MSE: Mean squared error

POD: proper orthogonal decomposition

ROM: reduced order models

SVD: singular value decomposition

List of Participants

1	COO	Universidad Politécnica de Madrid	UPM	ES
2	BEN	BARCELONA SUPERCOMPUTING CENTER-CENTRO NACIONAL DE SUPERCOMPUTACION	BSC	ES
3	BEN	UNIVERSITE LIBRE DE BRUXELLES	ULB	BE
4	BEN	KUNGLIGA TEKNISKA HOEGSKOLAN	KTH	SE
5	BEN	OVE ARUP & PARTNERS SA	ARUP	ES
6	BEN	MICROFLOWN TECHNOLOGIES BV	MT	NL
7	AP	BuildWind SPRL	BW	BE
8	AP	BRISTOL CITY COUNCIL	BRIS CC	UK
9	AP	AYUNTAMIENTO DE MADRID	AY MAD	ES
10	AP	UNIVERSITY OF BRISTOL	UoB	UK
11	AP	AIR QUALITY CONSULTANTS LTD	AQC	UK

Table of Contents

1. Introduction

Outdoor air pollution refers to the presence of one or more substances in the atmosphere—such as chemicals, particulate matter, or biological agents—in the form of solid particles, liquid droplets, or gases, at concentrations and duration that can be harmful to human health [1]. Prolonged exposure to high concentrations of air pollutants can lead to significant adverse health effects, including inflammation, oxidative stress, immunosuppression, and mutagenicity in cells throughout the body. These effects can damage vital organs such as the lungs, heart, and brain, potentially culminating in severe outcomes, including mortality [2]. The most common (but not only) air pollutants found in urban areas are ozone ($\boldsymbol{O_3}$), particulate matter (such as $\boldsymbol{PM_{2.5}}$ and $\boldsymbol{PM_{10}}$), sulfur dioxide ($\boldsymbol{SO_2}$), nitrogen dioxide ($\boldsymbol{NO_2}$), carbon monoxide ($\boldsymbol{CO}$), carbon dioxide ($\boldsymbol{CO_2}$) and ammonia ($\boldsymbol{NH_3}$) [3].

The 2024 State of Global Air Report developed by the Health Effects Institute [4] positioned air pollution as the world's second largest risk factor of deaths in 2021, causing approximately 8.1 million deaths that year. This study also reveals that over 90% of the world population live in places where the air pollution levels exceed the World Health Organization air quality standards on fine particulate matter ($\boldsymbol{PM_{2.5}}$) [2], [5]. Recent studies positioned air pollution as the largest environmental health risk in Europe, causing over 300 000 premature deaths in 2021 [6].

Population in cities is constantly increasing due to migration from rural areas which is causing the rise of pollutant emissions into the atmosphere. The United Nations agency through the 11th and 13th sustainable development goals in the 2030 Agenda show the need to address sustainability in urban areas. Urban air pollution refers to air contamination by harmful substances that can threaten both human health and the environment. Understanding air pollution in cities is challenging due to the complex interactions between natural and human-made environmental factors, as well as the spatial and temporal variability of pollutant dispersion and concentration. Understanding how pollutants are emitted and spread throughout an urban area is key to making informed decisions that can improve air quality and control pollution. Despite the efforts, the current available models are unable to provide the required spatio-temporal accuracy to reproduce the pollutant-dispersion patterns encountered in cities[6]. Therefore, precise predictive models for air-quality control are relevant to foreseeing excessive pollutant concentrations episodes and consequently prevent them.

The flow within urban areas is typically turbulent (Re > 4000), presenting significant challenges for analysis due to the wide range of spatio-temporal features inherent in such high-dimensional, nonlinear, and chaotic systems. The similarity of flow characteristics observed across a wide range of fluid flows suggests the existence of dominant processes that underlie different types of complex flows [6]. Although several methods

exist for extracting the features that characterize complex flows, there is growing interest in novel data-driven techniques capable to extract the principal features by performing modal decompositions of the flow [7].

Modal decomposition refers to a group of mathematical methods applied to flow fields that can identify the dominant features and dynamic characteristics of the flow in a low-dimensional coordinate system by separating it into distinct modes. The spatial features of the flow (hereafter referred to as "spatial modes") are typically ranked by their energy content, characteristic growth rates, or frequencies driving the flow motion [6]. This ranking is particularly useful for analyzing complex flows, such as those found in urban areas but also for reduced-order modelling and flow control. Since the spatial modes are ranked based on energy content, it is possible to create a reduced-order model and filter out noise or outliers from the flow field by reconstructing it using only the modes with the highest energy content [6], [8].

Data on complex fluid flow phenomena, such as urban flows, is typically obtained through experimental or numerical methods. On the one hand, experimental methods involve a wide range of setups and rely on the capability of sensors to acquire data. However, sensor data often contains noise due to uncontrolled exposure to external disturbances, which can result in incomplete or unresolved datasets. On the other hand, numerical methods for large-scale domains, such as urban areas, demand substantial computational resources and can require weeks, months, or even years to complete. Thus, reduced order modelling through modal decomposition is key to understanding complex flows while reducing its data dimensionality and computational cost, filtering out noise and non-crucial information [9].

Modal decomposition techniques, such as singular value decomposition (SVD), proper orthogonal decomposition (POD), and dynamic mode decomposition (DMD), are widely used for analyzing various types of flows due to their data-driven nature [7], [8]. These techniques have become powerful tools for flow structure identification and the development of reduced-order models (ROMs). The success of these techniques in fluid mechanics phenomena motivated researchers to develop robust variants such as spectral proper orthogonal decomposition (SPOD) and higher order dynamic mode decomposition (HODMD), which have demonstrated their capabilities to accurately analyze turbulent and multi-scale flows [7]. Combining the acquired physical knowledge of these techniques with new machine-learning strategies it is possible to create fast and efficient tools for identifying the main flow patterns with high accuracy and develop high-fidelity ROMs that reveal new mechanisms related to the flow behavior, thermal effects and pollutant dispersion in urban flows [7].

In this regard, the MODELAIR team at the Technical University of Madrid (UPM) are working in different tools based on modal decomposition and combined with machine learning approaches. Their current work involves the development of tools for the identification of flow patterns connected to air pollution in large-scale urban

flow databases, filling multi-parametric fluid flow databases and the analysis of key-variables affecting air pollution. These tools have been tested using experimental and numerical databases that are strongly related to urban flows, including a Large-Eddy simulation of the Tetuan district of Madrid, Spain provided by the BSC consortium partner.

Also, new tools for data analysis specific to understand pollutant dispersion in cities have been developed to exploit information from air quality monitoring sensors found in cities. These tools have been tested in databases from Bristol, provided by AQC, associated partner, and from Madrid, extracted from open-source repositories and following the guidance of ARUP, consortium partner.

2. Identifying flow patterns connected to air pollution

The identification of flow patterns in urban areas is key for the understanding of the air pollution problem. An adequate understanding of the source of emission location and the pollutant dispersion path over a city will help to prevent and control high pollutant concentration episodes. For this, the UPM team has developed two novel tools capable to deal with large-scale databases from numerical simulations and complex experiments modelling cities:

1) lcSVD-adaption: tool suitable to perform the reconstruction of complete flow fields from sensors, and to reconstruct flow patterns (POD modes).

2) SVD-based hierarchical approach: tool to identify flow patterns where the complete field is divided into different subdomains. The analysis is performed in each domain, and then, the information is grouped. This tool alleviates the computational cost of standard methodologies generally used in the literature to identify flow patterns.

The above-mentioned tools have been applied in a test case of a Large Eddy Numerical Simulation (LES) performed over the Tetuan district in Madrid, a summary of the applied methodology and results is given below.

2.1 Test case: Tetuan district, Madrid-Spain

Madrid, the capital city of Spain, is located centrally on the Iberian Peninsula and houses approximately 3.8 million inhabitants in the city proper and 6.7 million inside the metropolitan area. Madrid, in collaboration with the European Union, is actively pursuing measures to reduce air pollution in urban areas. Implementing stringent emission controls, promoting sustainable transportation, and fostering green initiatives are integral to their joint efforts to create healthier and cleaner city environments. "Plaza de Castilla", is considered Madrid's business center, hosting key national companies, and drawing the attention of real estate developers.

This results in high traffic flow, especially during peak hours and dust generation due to the different construction sites, which naturally leads to higher levels of pollution. Currently, Madrid has 24 air-quality control stations located to acquire real time data of the concentrations of different air pollutants.

The BSC consortium partner has provided a database coming from a LES numerical simulation performed over the district of Tetuan which is located near “Plaza Castilla” providing pressure and velocity data averaged on time. The computational domain consists of 120 million elements in an unstructured mesh over a 3-D reconstruction of the district, as shown in Fig. 1. This database has been used for the development of the above-mentioned tools to define and adapt the new and current modal decomposition methods for macro-data purposes. To make the database suitable for the developed tools, the CFD data has been interpolated into tensor. The spatial dimensions are fixed into a structured grid with 2000 points on each spatial coordinate.

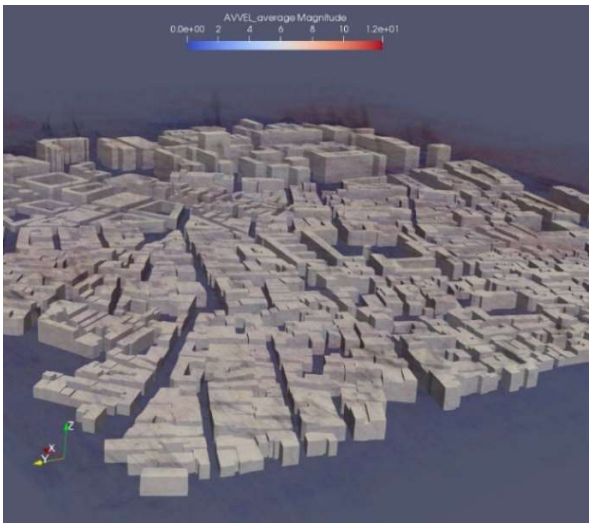


Figure 1: LES numerical simulation - Average Wind Velocity of the Tetuán district, Madrid-Spain.

2.2 Reconstructing flow field using information from sensors

The aim of this tool is to reconstruct an entire urban flow field by the usage of reduced data collected from sparse located sensors using the low-cost singular value decomposition method (lcSVD) and the python library “pysensors” developed for optimal sensor placement as shown in Fig. 2.

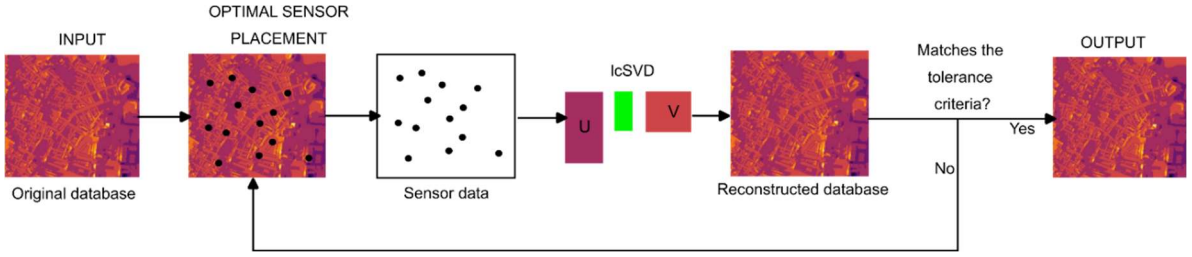


Figure 2: Sketch of the methodology used to reconstruct flow fields using information from sparse sensors.

Low-cost singular value decomposition (lcSVD) [11] is an extension of singular value decomposition (SVD) – a tool generally used for patterns identification that extracts information about the physics of the flow through the proper orthogonal (POD) modes- that addresses the computational challenges associated with analyzing large datasets. By applying SVD to a reduced snapshot matrix—either by reducing spatial or temporal dimensions—it minimizes the size of the data being processed while preserving key flow characteristics. The reduced matrix, referred to as the semi-reduced snapshot matrix, enables efficient reconstruction of the original dataset with significantly lower computational costs. This method is particularly advantageous for analyzing large-scale domains, such as those encountered in urban flows, as it allows for the efficient extraction of dominant flow structures and features without requiring the extensive computational resources typically associated with full SVD.

First, the database in tensor form is reshaped into a matrix suitable for lcSVD, this is achieved by compressing the spatial dimension into one dimension leaving the temporal dimension (in the case of phenomena varying on time). Therefore, the sensor placements are performed using “pysensors”[10], which is a modal decomposition-based tool that using SVD/POD/PCA and Random projection capable to provide the optimal sensor location to reconstruct a database, setting up a fixed number of available sensors. The data collected by these sensors as a reduced snapshot matrix serve an input for lcSVD.

The low-cost SVD (lcSVD) method begins by applying SVD to a reduced snapshot matrix, which factorizes the data into spatial modes, singular values, and temporal coefficients, with the number of retained modes determined by a specified tolerance. The spatial modes are then normalized using QR factorization to correct any non-orthogonality caused by numerical errors. Similarly, the temporal coefficients are re-orthonormalized to ensure they maintain their orthogonality. Afterward, the spatial modes are recovered by combining the reduced data with the corresponding temporal coefficients and singular values. The temporal coefficients are then recovered using the same process, ensuring consistency with the spatial modes. Once both spatial modes and temporal coefficients are reconstructed, the original dataset is approximated by multiplying them with the singular values. Finally, the reconstruction error is quantified to assess the accuracy of the approximation by comparing the original and reconstructed data.

As mentioned above, this methodology has been tested on the database provided by BSC of the flow field around the Tetuan district in Madrid. This dataset contains data regarding the average pressure (AVPRE), average velocity and acceleration (AVVEL and AVVE2). Fig 4 shows the results obtained by applying the lcSVD method for the analysis and reconstruction of the tensor using only 25 sensors over two-dimensional horizontal section of the computational domain. The reconstruction error obtained by the above-mentioned

methodology was 0.94471% and took approximately 0.39388 seconds, showing the robustness, efficiency and high accuracy of the methodology.

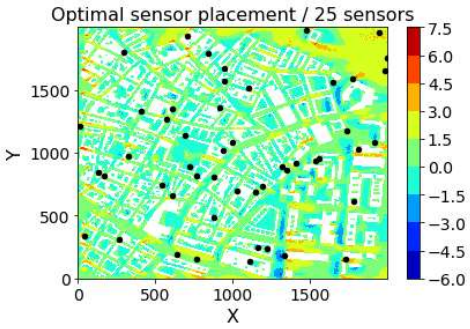


Figure 3: Optimal sensor locations over the urban flows database.

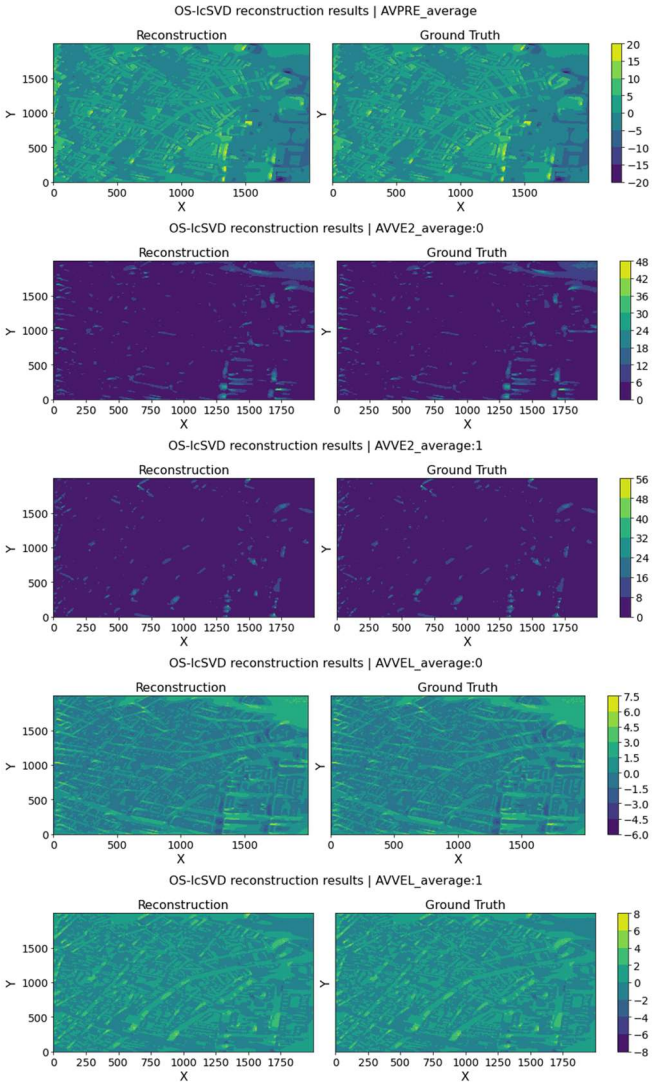


Figure 4: Reconstruction of the different variables related to the urban flow over the district of Tetuan, Madrid using lcSVD and optimal sensor placement.

This approach has been tested on two-dimensional CFD urban flow databases and is being adapted to handle three-dimensional databases obtained through numerical or experimental methods. Furthermore, the methodology is designed to process urban flow databases with temporal variations, making it applicable to time-dependent datasets. It will be applied to both CFD and experimental databases provided by the consortium partners, ensuring its versatility and effectiveness across diverse data sources.

2.3 Hierarchical approach

As highlighted in Section 2.1, urban flow databases obtained through numerical simulations are often prohibitively large, making direct processing computationally expensive. To address this challenge, we propose a methodology that divides the original spatial domain into multiple subdomains, enabling the efficient application of modal decomposition techniques, specifically singular value decomposition (SVD), with parallel computation capabilities.

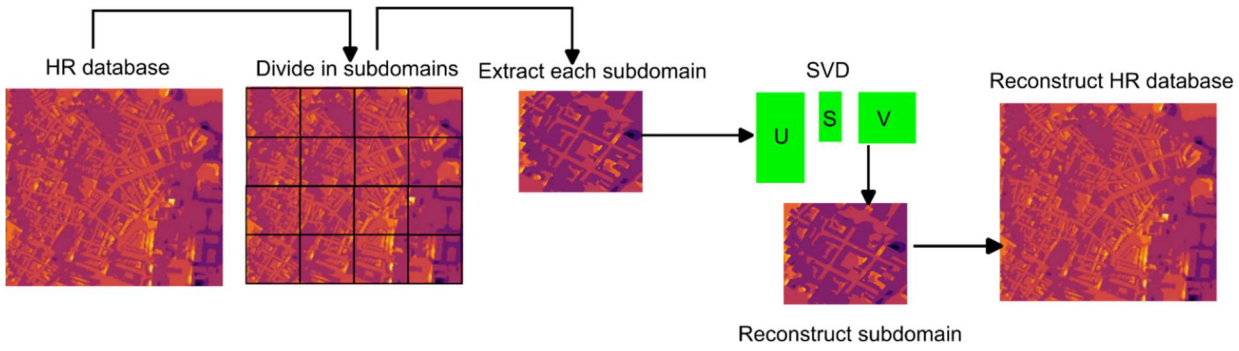


Figure 5: Hierarchical approach combined with singular value decomposition for urban flows database analysis.

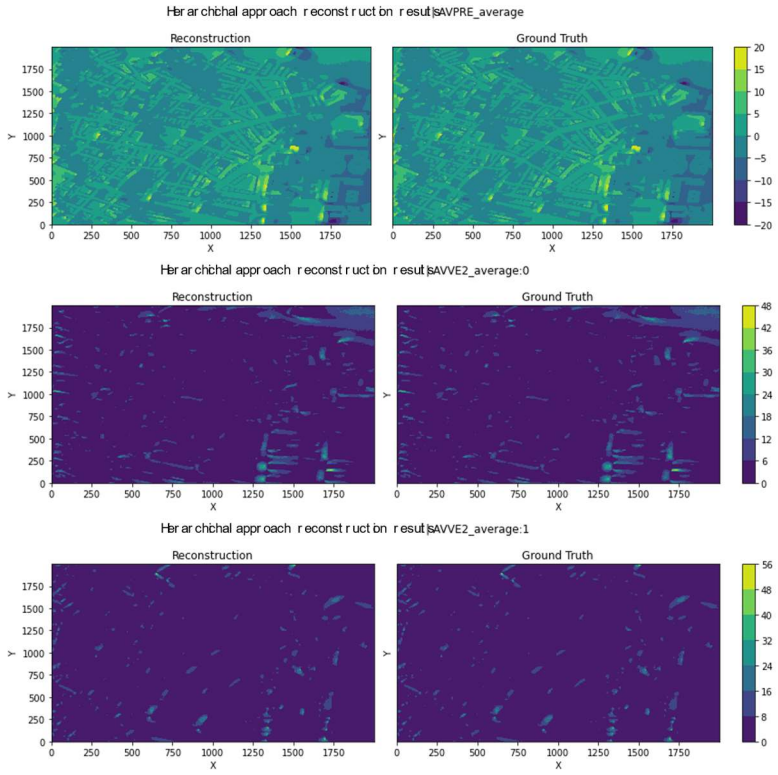
First, the database in tensor form is reshaped into a matrix suitable for SVD, this is achieved by compressing the spatial dimension into one dimension leaving the temporal dimension (in the case of phenomena varying on time). The proposed framework, illustrated in Fig. 5, begins by taking an urban flow database comprised in matrix form as input. The spatial dimensions of the tensor are then divided into *"n"* subdomains to ensure balanced computational loads. Within each subdomain, SVD is applied to extract the dominant flow features, filter noise and outliers, and reduce the dataset's dimensionality. This step facilitates the construction of a reduced-order model (ROM) that captures the essential physics of the fluid flow phenomena.

Singular Value Decomposition (SVD) is a matrix factorization method widely used in fluid dynamics and various other fields due to its ability to reduce the dimensionality of complex data. Developed by Sirovich[12] for fluid dynamics applications, SVD captures the dominant directions of a dataset by decomposing a matrix

into three components: $\boldsymbol{U}$, $\boldsymbol{\Sigma}$ (a diagonal matrix with singular values), and $\boldsymbol{V^T}$. The singular values in $\boldsymbol{\Sigma}$, ordered from largest to smallest, represent the most significant modes of the data, with the larger values corresponding to the dominant dynamics of the system. By retaining only the most significant modes, it is possible to reduce the data's dimensionality without losing critical information, making the method highly efficient for large scale databases as the ones encountered in urban flows.

Finally, the database is reconstructed by aggregating the processed subdomains while maintaining a reconstruction error below a predefined tolerance threshold. This ensures the reliability of the reduced data representation while significantly lowering computational costs, paving the way for scalable analysis of complex urban flow systems.

As in Section 2.3, this approach has been tested on the database provided by BSC of the flow field around the Tetuan district in Madrid and organized involving the average pressure (AVPRE), average velocity and acceleration (AVVEL and AVVE2) variables. The test has been performed by dividing the spatial domain into 16 subdomains with a reconstruction tolerance of 1E-03. Figure 6 shows the reconstruction obtained of the urban flows database using a hierarchical approach with an error of 0.45123%.



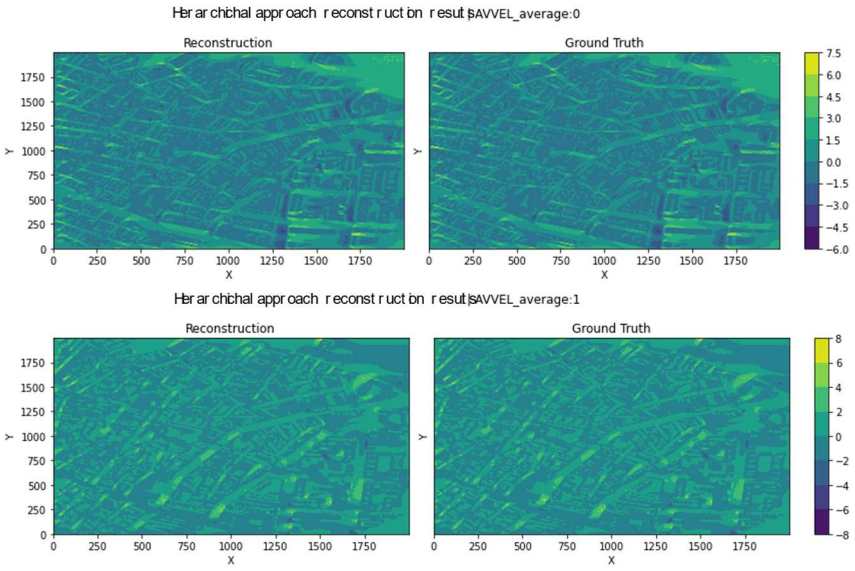


Figure 6: Reconstruction of the different variables related to the urban flow over the district of Tetuan, Madrid using the hierarchical approach.

To increase the performance of the tool and make it suitable for sparse sensor data reconstruction the proposed hierarchical approach will be combined with the methodology described in Section 2.2 by using *pysensors* for optimal sensor placement and lcSVD for the dimensional reduction and reconstruction. Figure 7 shows a scheme of the described new methodology.

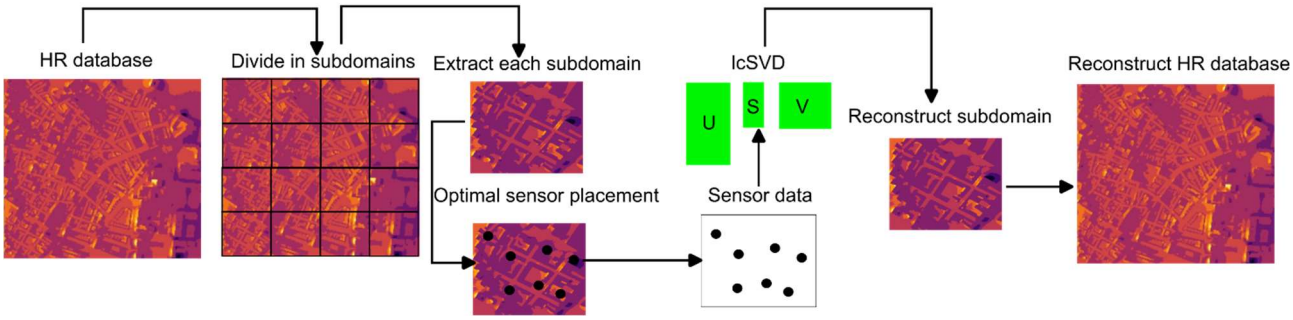


Figure 7: Hierarchical approach combined with optimal sensor placement and lcSVD for reconstructing large databases of urban flows.

3. Filling multi-parametric databases

Performing highly accurate large-scale numerical simulations has high computational cost, requiring huge amounts of computational resources and significant processing time. Simulating complex phenomena, such as urban flows over extensive computational domains like entire cities, is a demanding challenge. The complexity increases when simulating flow behavior under varying conditions, such as changes in Reynolds

number or wind direction, which can be challenging in terms of computational resources and time. To address these challenges, there is a growing need for innovative methodologies capable of generating new data efficiently from reduced computational fluid dynamics (CFD) or experimental databases. The proposed methodology aims to accurately expand multiparametric urban flow databases by generating new data from a limited number of initial datasets, combining modal decomposition, specifically Higher-order singular value decomposition (HOSVD) with machine learning.

The Higher-Order Singular Value Decomposition (HOSVD), also known as orthogonal Tucker decomposition, is a powerful tensor decomposition technique introduced by Tucker in 1966 and later refined by De Lathauwer et al [13]. It is widely used in aerodynamic database compression and modeling due to its ability to extract the most coherent and linearly independent features from high-dimensional data. HOSVD decomposes a tensor into a core tensor and a set of orthonormal mode matrices corresponding to each dimension. These modes, derived through standard SVD applied to tensor fibers, retain the dominant features associated with the largest singular values. This allows for dimensionality reduction, noise filtering, and the creation of compressed yet accurate representations of the original data. With a specified number of modes retained for each spatial dimension it is possible to achieve an optimal balance between accuracy and compression. This physics-based approach has proven effective for capturing the essential flow dynamics in fluid mechanics applications.

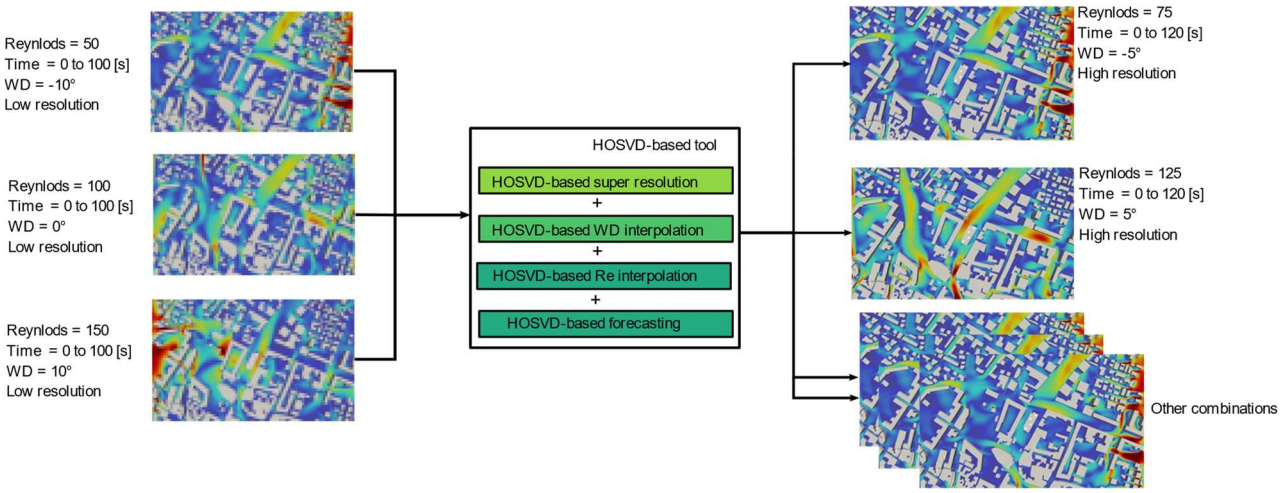


Figure 8: Sketch of the proposed model for filling multi-parametric databases.

Figure 8 illustrates the proposed model for enriching multiparametric databases by combining Higher-Order Singular Value Decomposition (HOSVD) with machine learning techniques. The model begins with a reduced urban flows database in tensor form, obtained either experimentally (sparse sensor data acquisition) or numerically (coarse-mesh CFD simulations), encompassing various flow conditions (e.g., Reynolds number,

wind direction). HOSVD is then applied to decompose the tensor into a set of mode matrices corresponding to each tensor dimension, a core tensor, and a list of singular values. These components serve as inputs for subsequent machine learning applications, such as neural networks, to enhance and expand the database in terms of space and time. These filled up matrices will be combined with the core tensor to reconstruct back the full database. As an output, the model can fill the database with new flow condition data, enhancing the spatial resolution by increasing the number of points or elements, and predicting in time the behavior of the phenomena.

The model has been developed in three different steps, each step focusing on a specific problem: spatial resolution enhancement, temporal forecasting and feature interpolation. In the following paragraphs you will find a summary of the methodologies designed for each step. It is worth mentioning that each step has been developed into an independent publication.

Data resolution enhancement

Data resolution enhancement is also known as super resolution and aims to find the high-resolution flow fields from low-resolution data. The methodology used for data resolution enhancement consists of developing a hybrid approach between modal decomposition and deep learning, combining the HOSVD resulting spatial mode matrices with a fully connected neural network (FNN). In this methodology, the HOSVD is used to extract the underlying physics and main flow patterns extracted while the FNN with an encoder/decoder like architecture is the responsible to extract and learn all the non-linearities of the phenomenon.

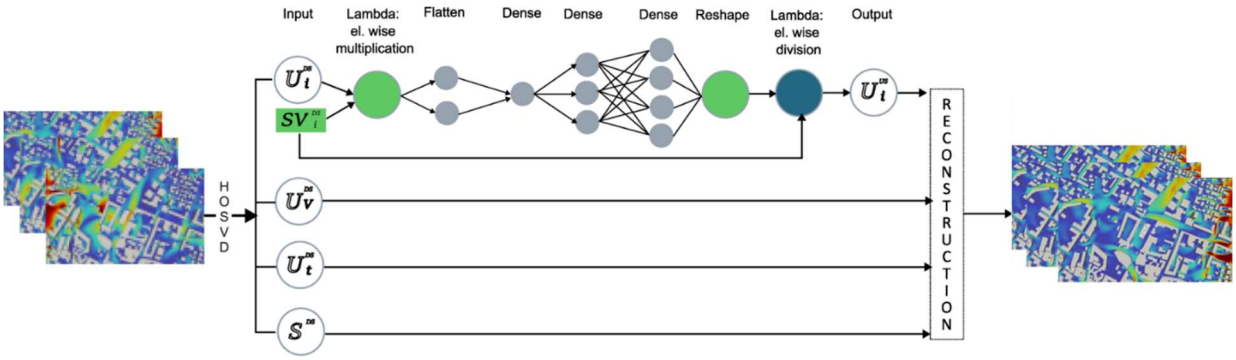


Figure 9: Hybrid modal decomposition and machine learning approach implemented for data resolution enhancement. Where $\boldsymbol{U_i}^{\boldsymbol{DS}}$ are the low-resolution spatial mode matrices, $\boldsymbol{U_V}^{\boldsymbol{DS}}$ the low-resolution feature mode matrix, $\boldsymbol{U_t}^{\boldsymbol{DS}}$ the low-resolution temporal mode matrix and $\boldsymbol{U_i}^{\boldsymbol{US}}$ corresponds to the enhanced-resolution spatial mode matrices.

The low-resolution matrix, with fewer spatial points distributed along each coordinate, limits the number of modes extracted through HOSVD to the total spatial points. For successful spatial resolution enhancement, there must be a balance between the resolution of the input database and the total number of high-energy modes needed to describe the flow. A fully connected neural network with three dense layers is employed, adapting to two distinct scenarios:

- Low-resolution database with limited modes for reconstruction: In this case, truncation of spatial modes for dimensionality reduction is not feasible. The spatial mode matrix is directly fed into a decoder-like neural network. The first two dense layers have a fixed number of neurons, which increase progressively in the forward pass, and employ a "LeakyReLU" activation function to address potential vanishing gradient issues. The final dense layer uses a "linear" activation function and outputs as many neurons as there are spatial points in the high-resolution database.

- Low-resolution database with excess of modes for reconstruction: Here, truncation of spatial modes is possible to reduce dimensionality. The neural network operates as an auto-encoder, with the first dense layer reducing the spatial modes' dimensionality, followed by subsequent layers increasing the dimensionality to match the spatial points of the high-resolution database. The first two layers use a "LeakyReLU" activation function, while the final layer applies a "linear" activation function.

Temporal forecasting

After performing Higher-Order Singular Value Decomposition (HOSVD) on the input tensor, temporal forecasting is employed to predict the future evolution of the data. The temporal matrix extracted from HOSVD serves as the foundation for sequence generation, where sliding windows of data are created for model training. Specifically, fixed-size subsets of temporal data points are used as input (input window), while the subsequent points corresponding to the forecast horizon serve as the target output. The dataset is then partitioned into three subsets: 70% for training the model, 15% for validation and the remaining 15% for evaluating the forecasting performance.

A Long Short-Term Memory (LSTM) network is used for temporal forecasting due to its proven ability to model sequential dependencies and capture long-term temporal patterns. The implemented architecture, while simple and fundamental, is designed to allow for future complexity based on evolving requirements. It includes an input layer for temporal sequence data, a recurrent LSTM layer, and a fully connected dense layer to map the output to the desired forecast horizon. The network is

compiled with a mean squared error (MSE) loss function to minimize prediction error, the Adam optimizer for effective convergence. After training over multiple epochs with validation feedback for dynamic weight adjustments, the model generates predictions for future temporal snapshots using the test set. These predictions are appended to the temporal matrix, effectively updating the dataset. This integration ensures the temporal forecasting model not only captures the dynamic evolution of the temporal data but also complements the dimensionality reduction achieved by HOSVD, enabling precise and efficient predictions.

Feature interpolation

The computational resources required to perform and store large-scale simulations highlight the necessity of developing a methodology capable of generating new data from a limited dataset. The proposed methodology for feature interpolation combines HOSVD with various interpolation techniques, including linear, quadratic, and Kriging interpolation.

Similarly to the data resolution enhancement and temporal forecasting steps taking as input the mode matrix obtained from HOSVD, which corresponds to the features to be interpolated (e.g., Reynolds number, wind direction), along with the labels for all feature values. A singular value decay analysis is first conducted to determine the number of modes needed to capture the flow dynamics. The mode matrix is then truncated to reduce dimensionality, ensuring computational efficiency. The interpolation technique is subsequently chosen based on the distribution of data along each mode, enabling accurate and efficient feature interpolation.

Since each component of the methodology has been developed separately it can be adapted for different purposes. We have developed two different approaches to fill multi-parametric databases that adapt to some of the most common scenarios:

- Data resolution enhancement and forecasting: this approach must be used in cases where the user possesses a low-resolution database for one specific flow condition.
- Feature interpolation and forecasting: this approach must be used in cases where the user possesses a high-resolution database of a flow under various flow conditions (Re, wind direction, etc)
- Data resolution enhancement, feature interpolation and temporal forecasting: this approach must be used in cases where the user possesses a low-resolution database of a flow under various flow conditions (Re, wind direction, etc.)

This methodology has been tested in general fluid dynamics numerical and experimental databases under different flow conditions (laminar, turbulent). In the upcoming years it will be extended and adapted to model realistic urban flows, and will be tested in databases from Madrid, Bristol and Brussels.

4. Variables affecting air pollution

Understanding the dynamics of air pollution requires a multidisciplinary approach that incorporates detailed datasets, computational modelling, and an in-depth analysis of the relationships between key variables

The characteristics and interactions of meteorological factors—such as wind speed, wind direction, temperature, humidity etc. with different pollutants are central to this endeavour. These variables have a considerable influence on the behaviour and distribution of pollutants in urban environments. While tools such as computational fluid dynamics (CFD) models are invaluable for simulating pollution patterns and dynamics, identifying the correlations between key meteorological parameters—such as wind speed, direction, temperature, and humidity—and pollutant concentrations is equally critical.

Studying these co-relations provides essential insights into how different factors collectively impact air quality. For instance, it can help identify conditions that amplify pollutant levels, such as low wind speeds leading to stagnation or high humidity affecting pollutant reactivity. Furthermore, analysing these relationships can reveal underlying trends and dependencies, offering a deeper understanding of urban pollution dynamics that static modelling or simulation alone might overlook.

This section focuses on the development of tools to identify corelations between key variables using datasets obtained for the cities of Bristol, UK, and Madrid, Spain. By uncovering these relationships, this analysis aims to provide a deeper understanding of the interplay between meteorological factors and pollutant concentrations in these urban environments, contributing to data-driven approaches for improving air quality.

4.1 Crucial meteorological variables

Various meteorological variables have unique influences on pollutant concentrations, shaping the dynamics of air pollution. Some of the key meteorological variables include:

- **Wind Speed (U):** Higher wind speeds typically result in lower pollutant concentrations as they enhance the dispersion of pollutants, reducing their accumulation in a specific area. Conversely, low wind speeds can lead to stagnation, increasing pollutant levels.
- **Wind Direction (PHI):** The direction of the wind influences the transport of pollutants, determining which areas experience higher or lower concentrations depending on the source location relative to the monitoring station.

- **Temperature (T):** Temperature affects chemical reactions in the atmosphere. For instance, higher temperatures can increase ozone (O_3) formation, while low temperatures can promote the accumulation of particulates due to reduced vertical mixing.
- **Pressure (P):** High-pressure systems are often associated with stable atmospheric conditions, which can trap pollutants near the ground. Low-pressure systems, on the other hand, can lead to enhanced vertical mixing, dispersing pollutants more effectively.
- **Relative Humidity (RHUM):** High humidity can affect the formation and transformation of certain pollutants, such as particulates, by promoting condensation processes. It also restricts the dispersion of the pollutant in the atmosphere.
- **Cloud Cover (CL):** Increased cloud cover can reduce sunlight availability, potentially limiting photochemical reactions that produce secondary pollutants like ozone.
- **Precipitation:** Rainfall can effectively remove pollutants from the atmosphere through wet deposition, reducing concentrations of particulate matter and soluble gases. However, the scavenging efficiency depends on the intensity and duration of precipitation events.

These meteorological variables, individually and in combination, play a critical role in shaping air quality dynamics of an area. But each variable influences pollutant behaviour in distinct ways, shaping how pollutants are dispersed, transported, or accumulated in the atmosphere.

4.2 Locations and Datasets

Bristol, located in the southwest of England, faces significant challenges related to air pollution, primarily driven by elevated concentrations of nitrogen oxides (NOx). The city regularly experiences pollutant levels that exceed national and European air quality standards, particularly in areas with high traffic volumes [14]. Bristol's unique geography and topography exacerbate its air quality issues. The city is situated along the Avon River and lies near the Severn Estuary, with surrounding hills that can limit air circulation and trap pollutants in the urban basin. The proximity to the sea contributes to fluctuating wind speeds and directions, influencing the dispersion of pollutants. Moreover, weather conditions, such as temperature inversions, can further intensify pollution episodes by preventing vertical mixing of the atmosphere [15].

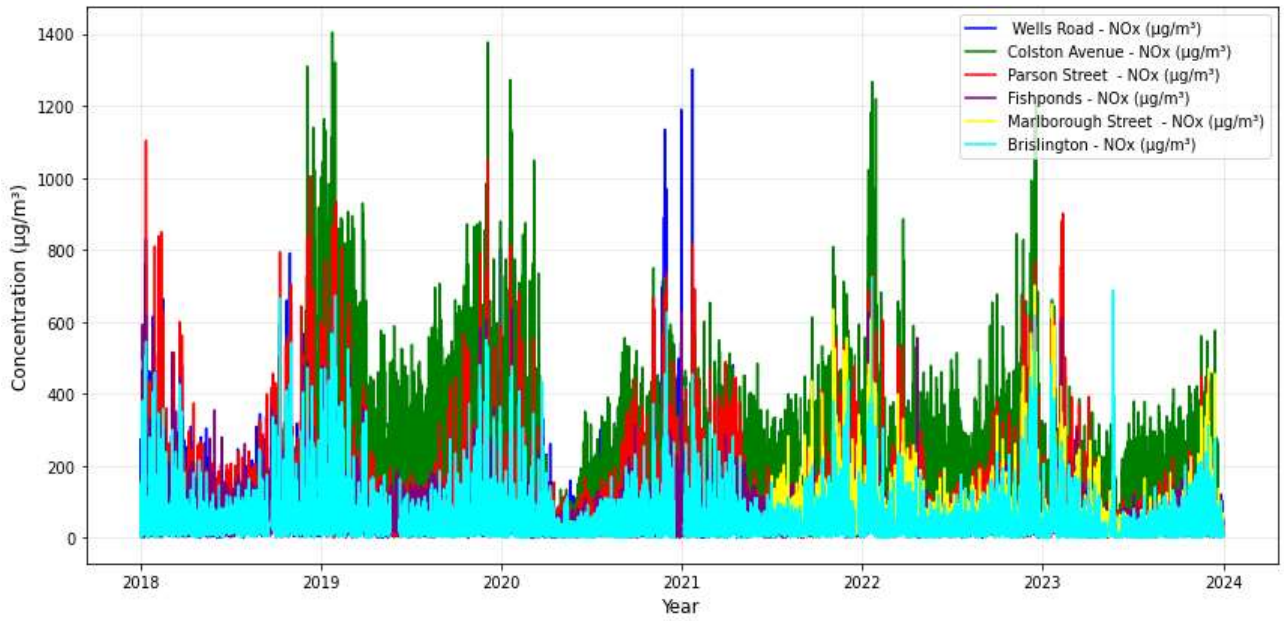


Figure 10: Concentration of NOx at different sites in Bristol City.

The primary sources of pollution in Bristol include vehicular emissions, industrial activities, and residential heating, with road transport accounting for most nitrogen oxide emissions. The city has implemented air quality monitoring networks and sustainable transport initiatives to address these concerns, as documented in its Air Quality Annual Status Reports [16]. More details can be found in deliverable D2.2.

In addition to Bristol, the study also focuses on Madrid, the capital of Spain, which offers the opportunity to explore a completely different topography and set of environmental conditions. Madrid faces notable air quality challenges due to nitrogen dioxide (NO_2) and particulate matter (PM) emissions. The city experiences seasonal fluctuations in pollutant concentrations, influenced by its geographic and climatic conditions. Monitoring data indicates that nitrogen dioxide levels in certain areas of Madrid consistently approach or exceed European Union standards, particularly during peak traffic hours [17].

Madrid's geographical location on a high plateau at approximately 650 meters above sea level contributes to unique air circulation patterns. The surrounding mountain ranges can restrict air movement, leading to the accumulation of pollutants [18]. The main sources of air pollution in Madrid include vehicular emissions, industrial processes, and residential heating systems. Road traffic, particularly diesel vehicles, remains a significant contributor to NO_2 emissions in the city. More detailed information can be found in deliverable D2.4.

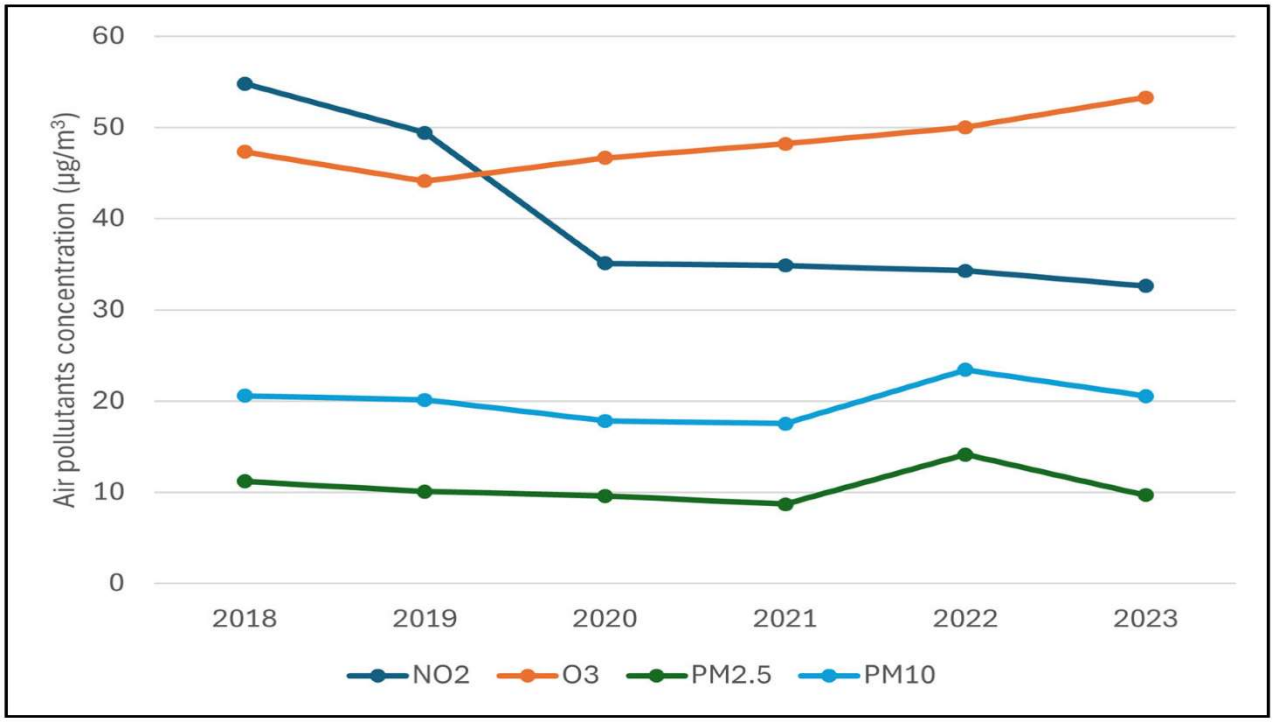


Figure 11: Average annual daily air pollutant concentrations, Madrid (2018–2023) [19].

Dataset

To better understand the dynamics of air pollution, this study utilizes detailed datasets comprising meteorological parameters and pollutant concentrations. These datasets provide a comprehensive basis for examining the relationships between key meteorological variables with pollutants like nitrogen dioxide (NO_2), particulate matter (PM), etc.

The datasets for Bristol were provided by Air Quality Consultants, Bristol, and the Bristol City Council, which include monitoring data for meteorological and pollutant concentrations from various locations across the city. Additionally, publicly available open-source datasets are also available, offering a broader perspective on the environmental and meteorological conditions influencing air quality in Bristol.

The meteorological dataset includes variables structured with hourly values and is organized as follows:

- **STATION ID**: Identifier for the monitoring station
- **YEAR**: Year of observation
- **TDAY**: Day of the year
- **THOUR**: Hour of the day
- **T**: Temperature (°C)
- **U**: Wind speed (m/s)
- **PHI**: Wind direction (angle in degrees)

- **P**: Pressure (kPa)
- **CL**: Cloud cover (oktas)
- **RHUM**: Relative humidity (%)

The figure below shows the locations of the meteorological monitoring sites, and the accompanying table provides details on the time period for which the data is available from each site.

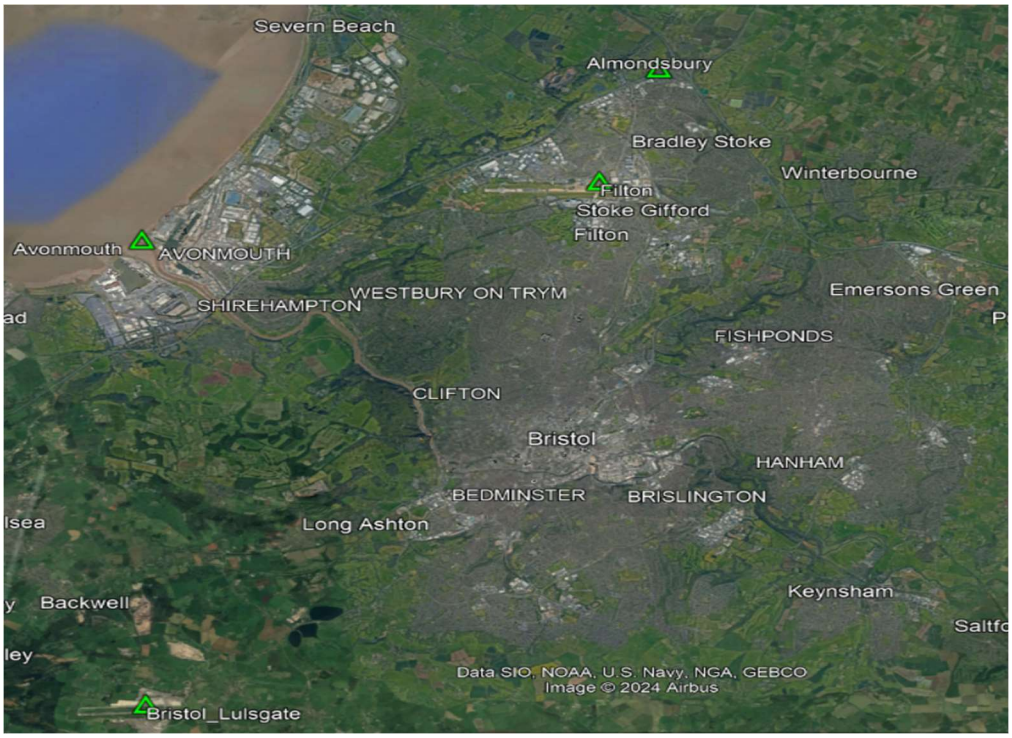


Figure 12: Locations of Meteorological Monitoring Sites for Bristol (Green markers indicate meteorological sites).

SITE LOCATION	TIME PERIOD	SOURCE
1. AVONMOUTH	2015-2020	AQC, Bristol
2. ALMONDSBURY	2019-2023	AQC, Bristol
3. FILTON	2015-2023	AQC, Bristol
4. LULSGATE	2015-2018	AQC, Bristol

Table 1: Site locations, time period and sources for meteorological datasets

Following the meteorological dataset, the analysis also incorporates pollutant concentration data, including key pollutants such as nitrogen oxides (NOx), nitrogen dioxide (NO_2), and nitric oxide (NO). These pollutants were monitored at various locations across Bristol, with data recorded at hourly intervals to capture temporal variations and trends in air quality.

Figure 13 shows a combined representation of the meteorological and pollutant measurement sites. Table 2 provides details on the monitoring sites, including their locations, operational periods, and the source of the dataset used in this study.

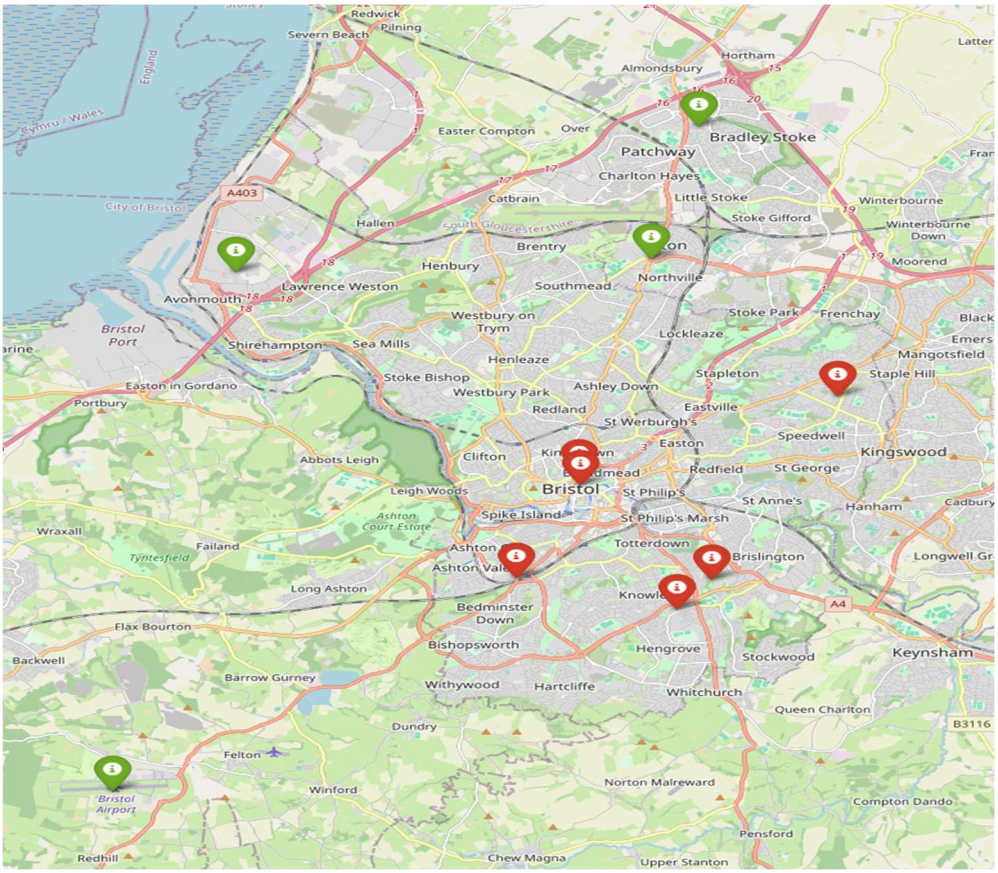


Figure 13: Locations of Meteorological and Pollutant Concentration Monitoring Sites for Bristol (Red markers indicate pollutant concentration sites, and green markers indicate meteorological sites).

SITE LOCATION	TIME PERIOD	SOURCE
1. BRISLINGTON	2018-2023	Bristol City Council
2. COLSTON AVENUE	Dec, 2019-2023	Bristol City Council
3. FISHPONDS	2018-2023	Bristol City Council
4. MARLBOROUGH STREET	Sep, 2021-2023	Bristol City Council
5. PARSON STREET	2018-2023	Bristol City Council
6. WELLS ROAD	2018-2023	Bristol City Council

Table 2: Site locations, time period and sources for pollutant datasets

The dataset from Parson Street also contains data for PM2.5 (Particulate Matter) form Nov,2020 to Dec,2023. Some other additional open sources for the pollutant data include:

1. https://maps.bristol.gov.uk/pinpoint/?service=localinfo&maptype=js&layer=Traffic+survey+locations

2. https://experience.arcgis.com/experience/bcf5a6312bc04ffeb43db67cd57f5439
3. https://opendata.bristol.gov.uk/search?categories=%252Fcategories%252Fenvironment

In the future, traffic data will also be incorporated into the analysis, obtained from the publicly available source - Bristol City Council Pinpoint Map Service. This dataset includes hourly vehicle counts recorded at various traffic survey locations across the city. The data provides valuable insights into traffic patterns and their potential correlation with pollutant concentrations, forming a critical component for further study.
Like Bristol, comprehensive meteorological and pollutant concentration datasets were obtained for the city of Madrid. The datasets include daily measurements of key atmospheric parameters and pollutant levels collected from multiple monitoring stations across the city.

The meteorological dataset for Madrid includes measurements of key meteorological parameters, as presented in section 4.1. However, the pollutant dataset, compared to Bristol, records a wider variety of air pollutants, offering the opportunity to study additional key pollutants. The variables measured include nitrogen oxides (NOx), nitric oxide (NO), nitrogen dioxide (NO_2), particulate matter less than 2.5 µm (PM2.5) and 10 µm (PM10), carbon monoxide (CO), sulphur dioxide (SO_2), ozone (O_3), among others, with concentrations expressed in µg/m³ or mg/m³ depending on the pollutant.

The datasets used in this study, obtained from Madrid's open data portals, provide extensive coverage of meteorological, pollutant, and traffic data. The air quality dataset spans 2001–2024, the meteorological dataset covers 2019–2024, and the traffic dataset includes vehicle counts from 2013–2024. Figure 14 lists all the monitoring stations considered in this study. The Madrid datasets also include co-located meteorological and pollutant measurement sites, offering another avenue for identifying their interrelationships, with their locations marked in Figure 15.

Figure 14: Monitoring stations in Madrid considered for meteorological and pollutant concentration measurements, including co-located stations for integrated analysis.

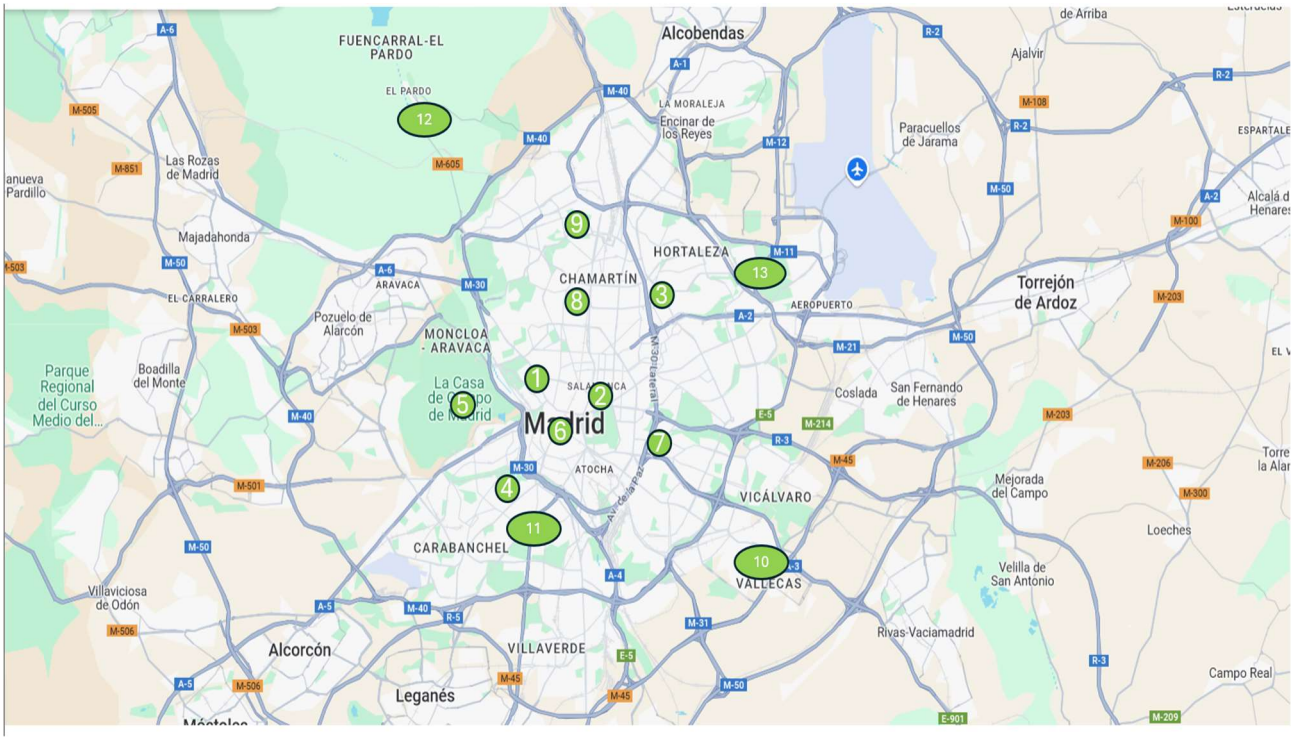


Figure 15: Locations of co-located monitoring stations in Madrid where both meteorological and pollutant concentration measurements were recorded.

This section provides a detailed overview of all the datasets utilized in this study, including meteorological parameters, pollutant concentrations, and traffic data. While these datasets offer valuable insights into the

dynamics of air pollution, several challenges must be addressed to ensure robust analysis. One significant challenge is that the meteorological monitoring sites and pollutant concentration monitoring sites are not co-located for Bristol. This spatial separation complicates direct comparisons between meteorological conditions and pollutant levels. Another issue lies in the pollutant concentration dataset, which contains significant gaps in the recorded data. These missing values, potentially caused by equipment malfunctions or maintenance issues, present challenges for analysis. For Madrid, even though we have co-located monitoring sites, not all of them measure the same variables. Some sites only record temperature as their meteorological component, limiting the scope of comprehensive analysis. To address this, interpolation techniques can be employed to estimate missing variables based on nearby stations or historical data. Addressing these gaps through data repair or imputation is critical to ensure the reliability and validity of the findings derived from this work.

4.3 Methodology

The methodology employed in this study encompasses several steps aimed at systematically analysing the relationships between meteorological variables and pollutant concentrations. The process begins with data augmentation and tensor creation, where the datasets are structured into high-dimensional tensors to enable comprehensive multi-variable analysis. Centring and scaling techniques are applied to standardize the data, ensuring that all variables, regardless of their units or magnitudes, contribute equally to the analysis.

To quantify the interdependencies between variables, a covariance matrix is computed, providing a foundational understanding of the relationships present in the data. Higher-Order Singular Value Decomposition (HOSVD) will then be utilized for dimensionality reduction, allowing the dominant patterns and features within the data to be identified. Finally, hierarchical Higher-Order Dynamic Mode Decomposition (HODMD), a method suitable to identify patterns and correlations between data as function of frequencies, will be implemented to capture complex temporal dynamics and uncover the underlying multi-scale interactions among the meteorological parameters and pollutant concentrations.

This methodology establishes a robust framework for analysing the datasets, providing insights into the intricate relationships influencing air quality in Bristol and Madrid.

Tensor creation

The data used for this analysis was imported from Excel and MET files, encompassing both meteorological and pollutant datasets. Each dataset was structured as a 3-dimensional tensor with the shape: **(stations, variables, timesteps in hours)**.

For the meteorological dataset, the stations included key monitoring sites mentioned in table 1 and figure 15. As an example, for the dataset of Bristol, the variables comprised meteorological parameters such as temperature (T0C), wind speed (U), wind direction (PHI), pressure (P), cloud cover (CL), and relative humidity (RHUM). Similarly, for the pollutant dataset, the stations, as listed in Table 2, include monitoring sites across the city, and provide measured pollutant concentrations for nitrogen oxides (NOx), nitrogen dioxide (NO_2), and nitric oxide (NO).

To facilitate comprehensive analysis, the meteorological and pollutant datasets were concatenated into a single tensor. In this combined tensor:

- The first dimension represents the stations, with meteorological stations followed by pollutant monitoring stations.
- The second dimension corresponds to the variables, with meteorological followed by pollutant concentrations.
- The third dimension is the time axis, capturing hourly timesteps for the observed periods.

This combined tensor serves as the foundation for subsequent analyses, enabling the study of relationships between meteorological conditions and pollutant levels in a unified framework.

Centering and Scaling

Centering and scaling are critical preprocessing steps in data analysis, especially when dealing with variables of differing units and magnitudes, as is the case with meteorological parameters and pollutant concentrations. Centering ensures that each variable has a mean of zero by subtracting its average value, while scaling standardizes the variance by dividing by the standard deviation. These steps help prevent variables with larger magnitudes from disproportionately influencing the results of the analysis, ensuring that all variables contribute equally to identifying patterns and relationships.

In this study, the combined tensor was reshaped into a 2D matrix to facilitate centring and scaling. The daily mean and standard deviation of each variable were computed across all timesteps, and the dataset was subsequently transformed to a standardized scale. This normalization ensures that correlations and covariance analyses are not biased by the relative scale of the variables, enhancing the reliability and interpretability of the results.

The implementation allows the combined tensor to be effectively processed for downstream analysis, such as the computation of the covariance matrix and application of dimensionality reduction techniques. The figure below is just a visual representation of how centring and scaling works.

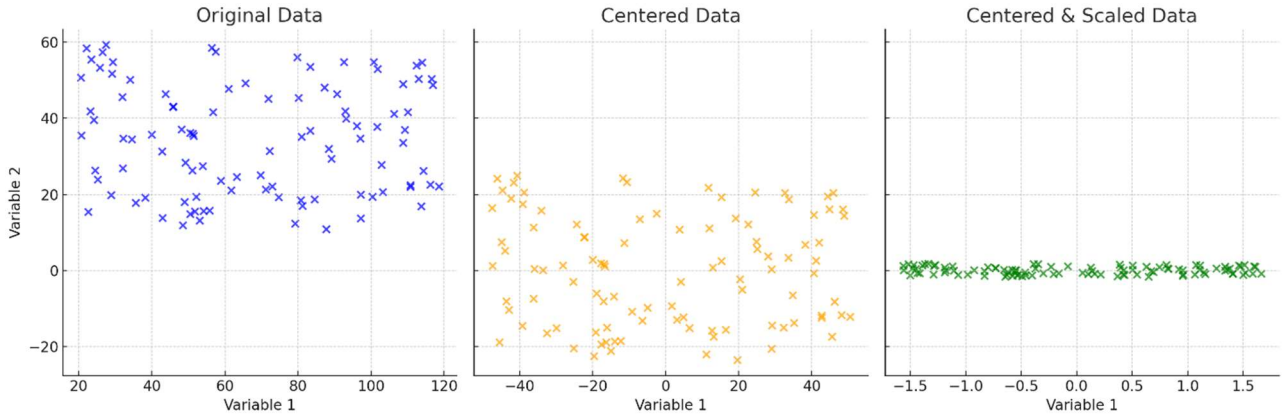


Figure 16: Illustration of Centering and Scaling.

Covariance Matrix

The covariance matrix is a fundamental component of multivariate analysis, offering a detailed representation of the linear relationships between variables. It measures the degree to which two variables change together, with positive values indicating a direct relationship and negative values representing an inverse relationship. For this study, the covariance matrix serves as a critical tool for uncovering the underlying structure within the dataset, which comprises meteorological parameters and pollutant concentrations.

By computing the covariance matrix, it becomes possible to identify key relationships and dependencies across variables, such as how wind speed or temperature correlates with pollutant concentrations like NOx or NO_2. This is particularly important when trying to understand complex, interdependent systems like urban air pollution, where multiple factors interact simultaneously. For example, high covariance between wind speed and pollutant concentrations might indicate that pollutant dispersion is heavily influenced by local wind patterns. Similarly, strong correlations between temperature and NO_2 levels could point to chemical processes influenced by thermal conditions.

In this study, the covariance matrix was computed using the *np.cov* function applied to the centred and scaled dataset. Standardizing the data before this computation ensures that all variables are treated equally, preventing those with larger magnitudes or units from dominating the analysis.

The insights gained from the covariance matrix are instrumental in several ways. First, it helps in identifying dominant variables that might drive pollutant behaviour, providing direction for targeted interventions. Second, the matrix lays the groundwork for dimensionality reduction techniques, such as Higher-Order Singular Value Decomposition (HOSVD), which leverage the covariance structure to extract components and reveal patterns hidden in the data. Lastly, the covariance matrix facilitates the exploration of variable clusters,

allowing for the grouping of variables that exhibit similar behaviour. This is particularly useful for simplifying the analysis of high-dimensional datasets, as it enables a more focused investigation of the most critical factors influencing air quality.

Higher Order Singular Value Decomposition

Higher-Order Singular Value Decomposition (HOSVD) is an extension of the matrix Singular Value Decomposition (SVD) to tensors, which are multi-dimensional arrays. Unlike SVD, which requires data to be flattened into a matrix, HOSVD preserves the inherent multi-dimensional structure of the data. The tensor is decomposed into a core tensor (S) and orthogonal factor matrices (Ux, Uy, and Ut) corresponding to spatial, variable, and temporal dimensions, respectively. This decomposition preserves the multidimensional relationships within the data, enabling efficient dimensionality reduction and analysis of dominant patterns. This makes it particularly suitable for analysing datasets with complex interdependencies, such as those in this study, where meteorological parameters and pollutant concentrations are analysed across multiple stations and time steps.

Further in this study, HOSVD will be applied to the combined tensor, structured as **(stations, variables, timesteps)**. The decomposition involves breaking down the tensor into a core tensor and orthogonal factor matrices, one for each mode. The core tensor captures the essential features of the data, while the factor matrices provide the principal directions along each mode, effectively reducing the dimensionality of the data while retaining its most significant patterns.

HOSVD provides several advantages for this analysis:

1. **Dimensionality Reduction**: It reduces the size and complexity of the dataset while preserving critical relationships.
2. **Pattern Extraction**: The decomposition isolates dominant trends and features, enabling a more focused exploration of interactions between meteorological parameters and pollutants.
3. **Noise Reduction**: By concentrating on the most significant components, HOSVD helps to filter out noise or less relevant variations in the data.

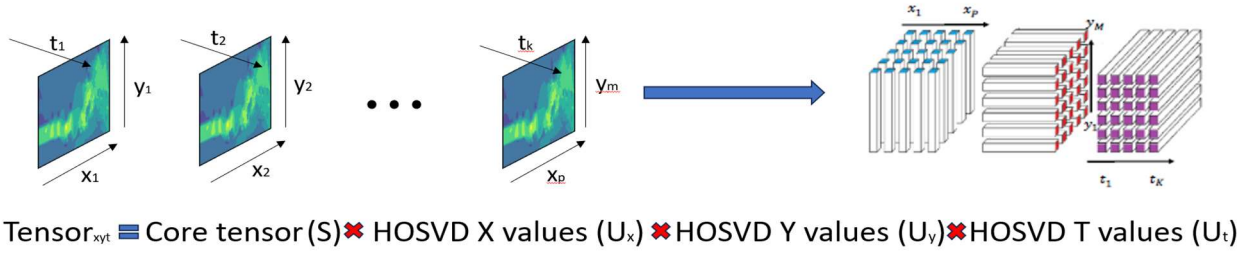


Figure 17: Illustration of Higher-Order Singular Value Decomposition (HOSVD) Process.

This methodology ensures that the complex interdependencies within the dataset are preserved and highlighted, enabling a comprehensive analysis of air pollution dynamics. The results from the HOSVD process serve as the foundation for further analysis, such as the application of multi-hierarchical Higher-Order Dynamic Mode Decomposition (HODMD) to uncover temporal patterns and predict future trends.

4.4 Correlation heatmaps

Each dataset presents its own unique challenges. In the Bristol dataset, while all meteorological and pollutant stations measure the same variables respectively, there are significant chunks of missing data, and the stations are not co-located, complicating direct analysis. On the other hand, Madrid features multiple co-located sites, but not all sites measure the same variables. For example, stations like Plaza España and Calle Farolillo record only temperature as their sole meteorological variable. To address these variations and better understand the relationships between variables, heatmaps of the covariance matrices have been plotted for both Bristol and the Casa de Campo monitoring station of Madrid, providing a visual representation of the interdependencies within the datasets.

The data for the year 2023 from Bristol was used for this analysis as it is the most recent and had fewer missing values compared to other years. Missing data was addressed using simple interpolation techniques to ensure completeness and reliability for further analysis.

Tensor Shape:

- **Combined tensor shape / Hourly tensor shape:** (8, 9, 8760)
- **Daily averaged tensor shape:** (8, 9, 365)

Station Configuration:

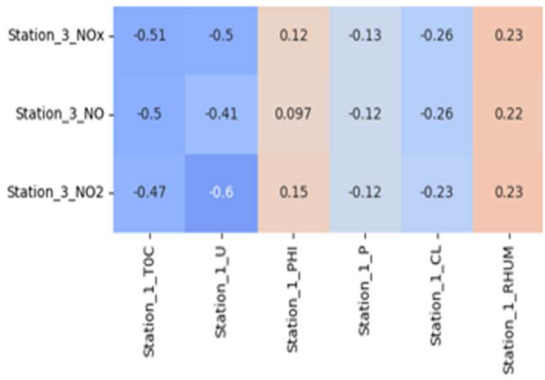
- Meteorological dataset (M)
- Pollutant dataset (P)
    - Station 1 – Almondsbury (M)
    - Station 2 – Filton (M)
    - Station 3 – Brislington (P)
    - Station 4 – Colston Avenue (P)
    - Station 5 – Fishponds (P)
    - Station 6 – Marlborough Street (P)
    - Station 7 – Parson Street (P)
    - Station 8 – Wells Road (P)

Correlation Coefficient

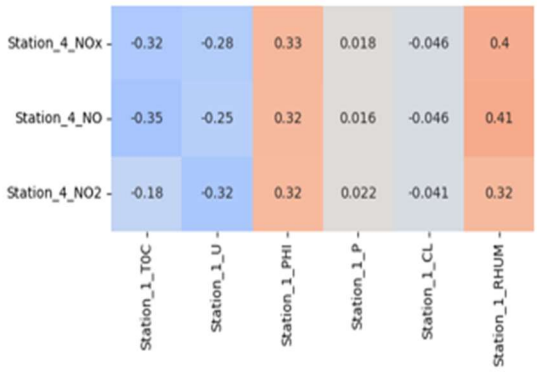
The heatmap displays correlation coefficients ranging from -1 to 1, providing insights into the relationships between variables:

- **Each cell** represents the correlation between two variables.
- **Range and Interpretation:**
  - +1: Perfect positive correlation; as one variable increases, the other increases proportionally.
  - -1: Perfect negative correlation; as one variable increases, the other decreases.
  - 0: No linear relationship; changes in one variable do not systematically affect the other.
- **Positive and Negative Signs:**
  - Positive values (e.g., 0.85) indicate a positive correlation, meaning the variables move in the same direction.
  - Negative values (e.g., -0.4) indicate a negative correlation, meaning the variables move in opposite directions.
- **Magnitude:**
  - **Closer to 1 or -1:** Strong correlation.
  - **Closer to 0:** Weak or no correlation.

The heatmaps below provide a visual representation of the relationships between meteorological variables and pollutant concentrations, helping to identify significant patterns and dependencies within the dataset. The figures below illustrate the relationships between individual meteorological and pollutant stations.



STATION 1 and STATION 3



STATION 1 and STATION 4

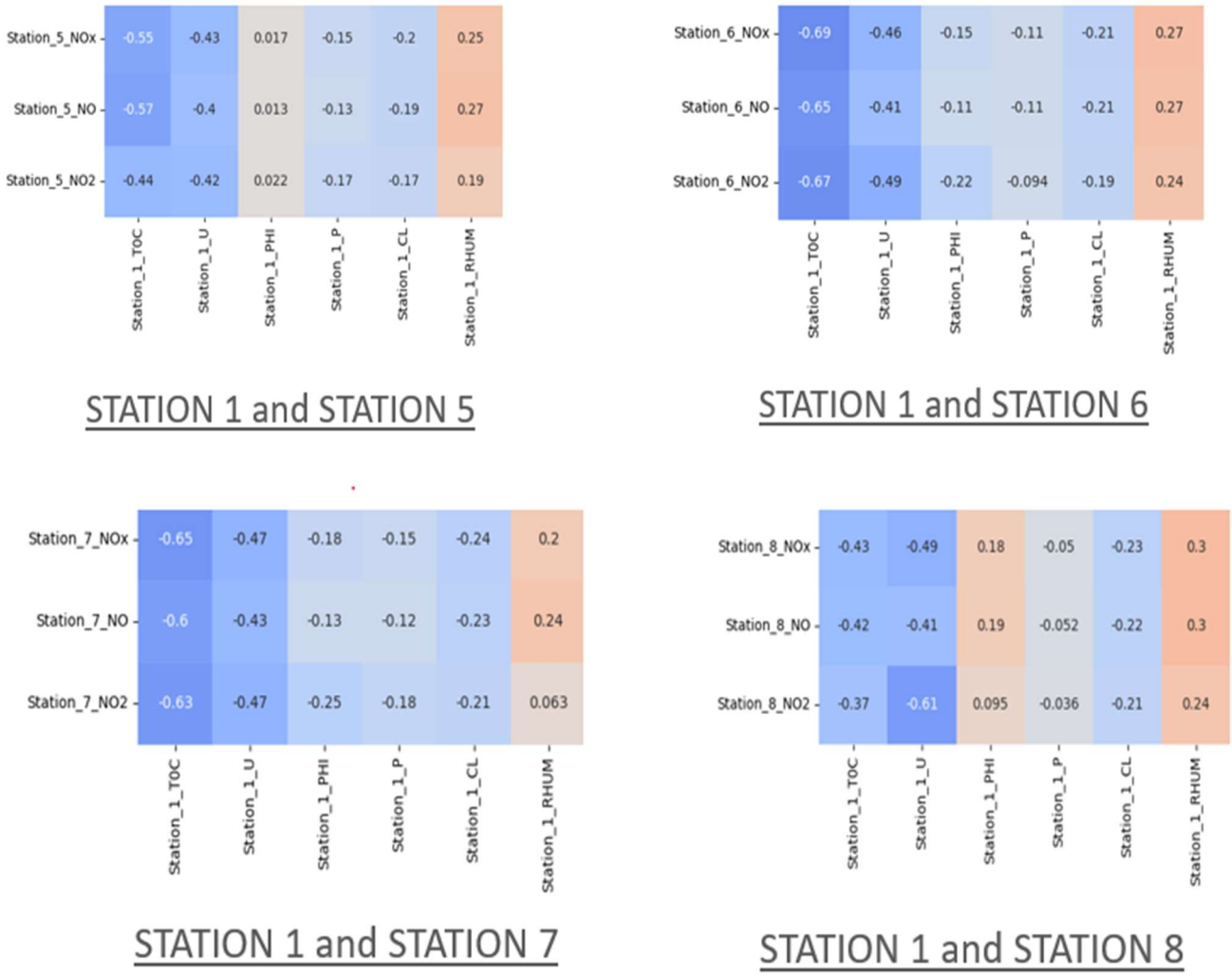
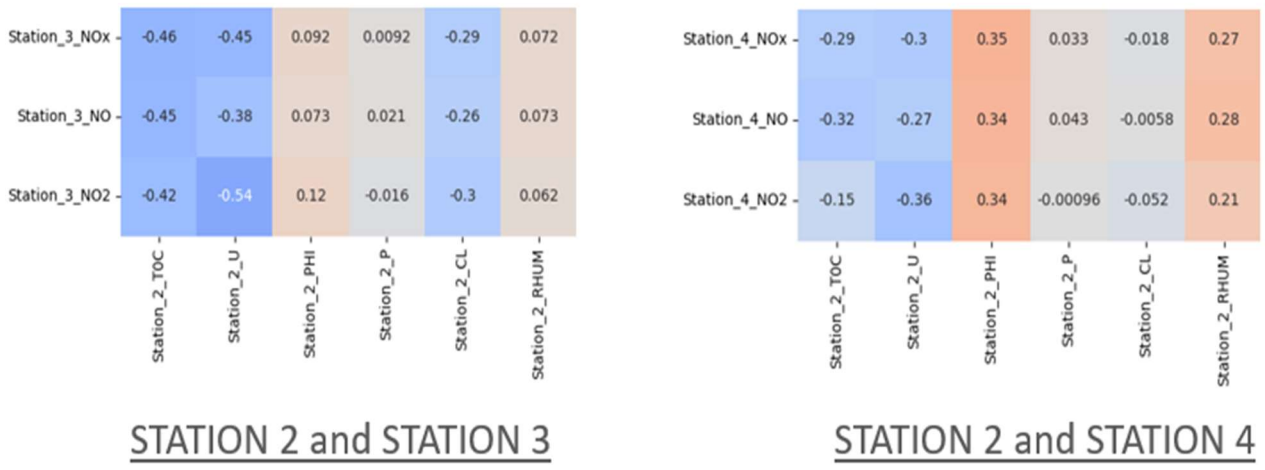


Figure 18: Heatmap displaying the correlation coefficients between meteorological variables at Station 1 (Almondsbury) and pollutant concentrations at Stations 3 to 8 (Brislington, Colston Avenue, Fishponds, Marlborough Street, Parson Street, and Wells Road)



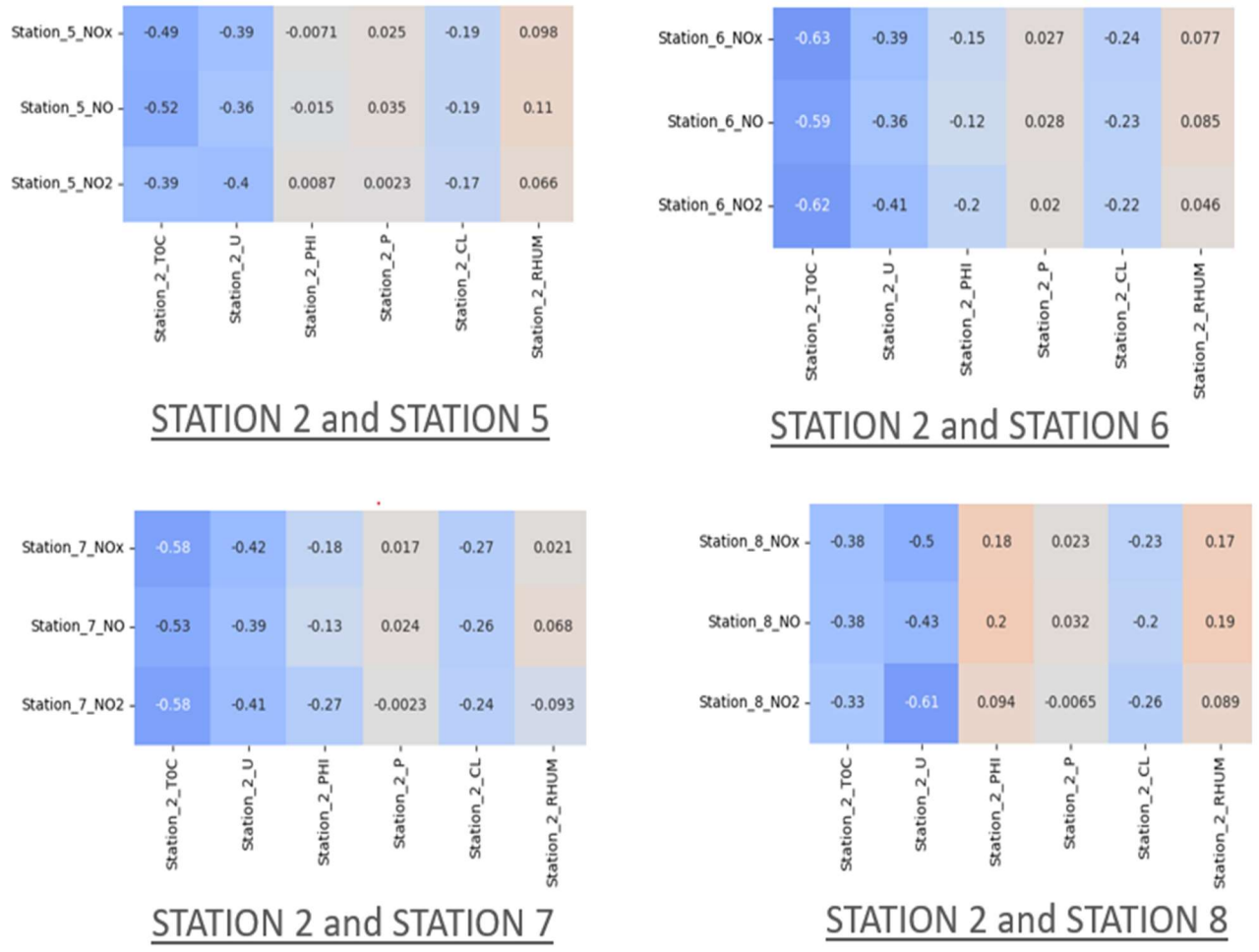


Figure 19: Heatmap displaying the correlation coefficients between meteorological variables at Station 2 (Filton) and pollutant concentrations at Stations 3 to 8 (Brislington, Colston Avenue, Fishponds, Marlborough Street, Parson Street, and Wells Road)

For Madrid, the analysis is conducted on the dataset obtained from the Casa de Campo monitoring site, one of the co-located sites measuring both meteorological and pollutant variables. Casa de Campo stands out as it is a populated urban, residential and business area, while recording a wider range of variables compared to other monitoring stations, making it a valuable site for detailed analysis of correlations. The co-located nature of this site also allows for a more direct comparison between meteorological conditions and pollutant concentrations, providing deeper insights into their relationships.

Tensor Shape:

- **Combined tensor shape / Hourly tensor shape:** (1, 14, 8760)
- **Daily averaged tensor shape:** (1, 14, 365)

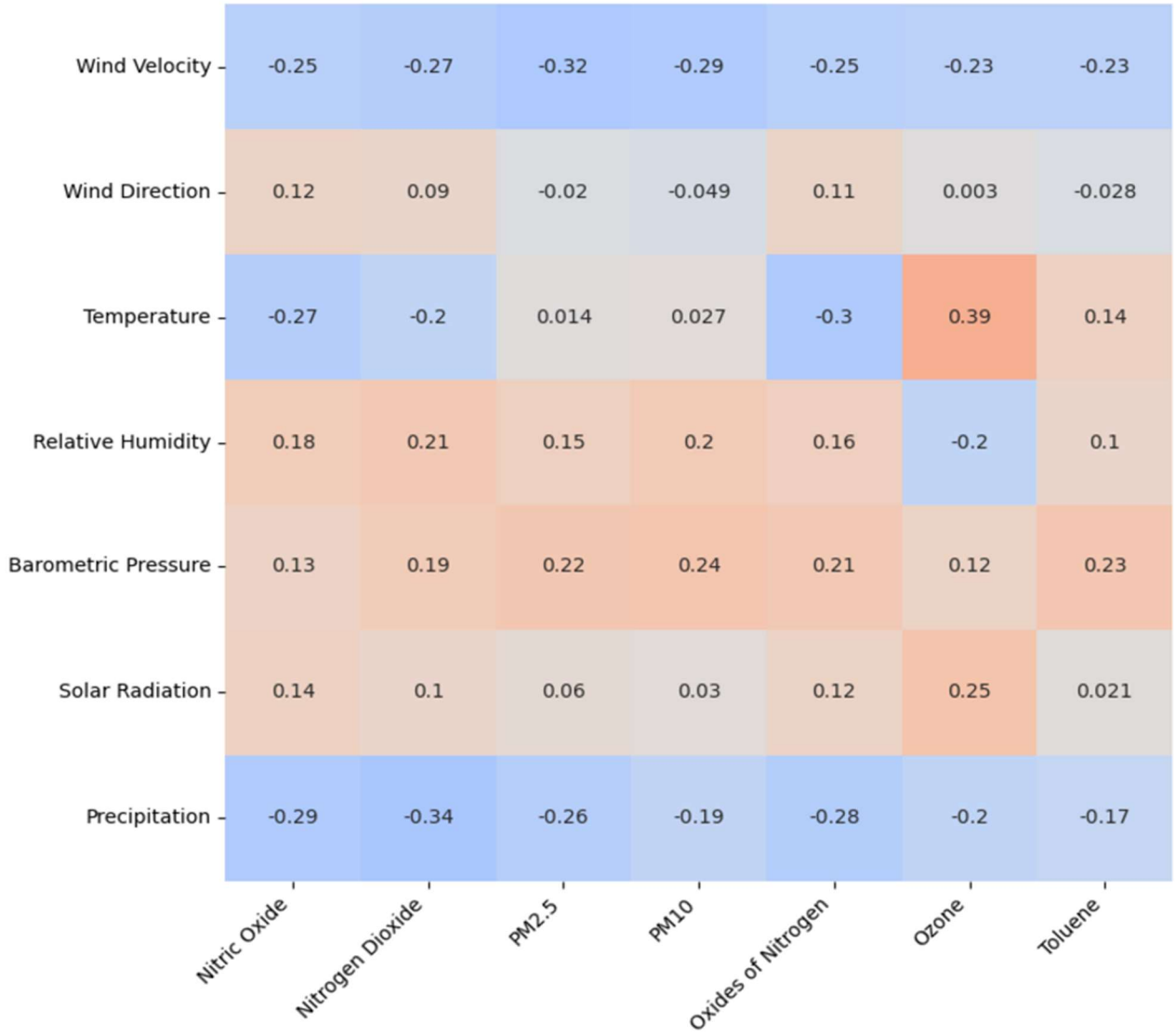


Figure 20: Heatmap displaying the correlation coefficients between meteorological variables and pollutant concentrations at Casa de Campo

4.5 Discussion and future work to exploit information of air quality monitoring sensors in cities

The heatmaps presented in the section above provide a foundational understanding of the correlations between variables in the datasets. While these correlations represent linear relationships, they still offer valuable insights into how meteorological parameters interact with pollutant concentrations. Examining the heatmaps reveals that meteorological variables exhibit considerable relationships with pollutant concentrations and their dispersion. While the strength and nature of these relationships may vary from station to station, it is evident that all meteorological variables play a significant role in shaping pollutant dynamics.

These initial observations allow us to identify significant patterns and trends, which can guide further analysis.

While these observations from the heatmap provide a good starting point, a more detailed analysis is necessary to uncover hidden patterns and dependencies. Higher-Order Singular Value Decomposition (HOSVD) will be applied to the datasets to extract dominant patterns and capture non-linear relationships across multiple dimensions. This technique will help identify the most influential variables and uncover temporal and spatial dynamics.

However, it is essential to address the issue of data repair. The datasets require further preprocessing, including repairing gaps and interpolating missing values, to ensure the completeness and reliability of the analysis. Various interpolation and repair methods will be evaluated to determine the most effective approach. HOSVD will then be implemented to gain a deeper understanding of the interactions between meteorological parameters and pollutant concentrations, laying the groundwork for predictive modelling and air quality management strategies.

5. Conclusions

This report outlines the development and application of tools for data analysis and feature extraction within the MODELAIR project, specifically in WP4. Two groups of tools have been created to address challenges in numerical and experimental database analysis.

The first group focuses on reducing data dimensionality in large numerical datasets, identifying key flow dynamics, and leveraging deep learning to build reduced-order models. These models enhance data resolution, fill multi-parametric databases, and predict flow evolution over time. The tools have been successfully tested on fluid dynamics problems involving laminar and turbulent flows and extended to urban air quality studies. A specific application includes analysing a numerical database for Madrid's Tetuan district, a highly polluted urban area.

The second group of tools processes data from air quality sensors distributed in urban environments to establish correlations between weather conditions and pollutant concentrations. These tools have been tested on sensor databases from Madrid, ES, and Bristol, UK, to predict pollutant evolution based on environmental variables. Future efforts will focus on enhancing these tools' robustness and expanding their applicability to additional datasets, including detailed analyses of Brussels, BE.

This work is the result of the collaboration between the MODELAIR partners from UPM, AQC, ARUP, and BSC contributing expertise and resources, such as sensor (AQC and ARUP) and numerical databases (BSC), to develop, test, and validate the tools described in this report (developed by UPM).

References

1. W. M. Sweileh, S. W. Al-Jabi, S. H. Zyoud, and A. F. Sawalha, “Outdoor air pollution and respiratory health: A bibliometric analysis of publications in peer-reviewed journals (1900 - 2017),” Jun. 01, 2018, BioMed Central Ltd. doi: 10.1186/s40248-018-0128-5.

2. “Air quality, energy and health.” Accessed: Nov. 25, 2024. [Online]. Available: https://www.who.int/teams/environment-climate-change-and-health/air-quality-energy-and-health/health-impacts.

3. H. Gul and B. K. Das, “The Impacts of Air Pollution on Human Health and Well-Being: A Comprehensive Review,” Journal of Environmental Impact and Management Policy, no. 36, pp. 1–11, Oct. 2023, doi: 10.55529/jeimp.36.1.11.

4. Health Effects Institute., “State of Global Air 2024. Special Report.,” Boston, MA, 2024.

5. J. Wise, “Pollution: 90% of world population breathes air that exceeds WHO targets on particulate matter,” BMJ, vol. 380, 2023, doi: 10.1136/bmj.p615.

6. Á. Martínez-Sánchez, E. López, S. Le Clainche, A. Lozano-Durán, A. Srivastava, and R. Vinuesa, “Causality analysis of large-scale structures in the flow around a wall-mounted square cylinder,” J Fluid Mech, vol. 967, Jul. 2023, doi: 10.1017/jfm.2023.423.

7. P. Torres, S. Le Clainche, and R. Vinuesa, “On the experimental, numerical and data-driven methods to study urban flows,” Energies (Basel), vol. 14, no. 5, Mar. 2021, doi: 10.3390/en14051310.

8. K. Taira et al., “Modal analysis of fluid flows: Applications and outlook,” AIAA Journal, vol. 58, no. 3, pp. 998–1022, 2020, doi: 10.2514/1.J058462.

9. Z. Dar, J. Baiges, and R. Codina, “Reduced Order Modeling”.

10. B. de Silva, K. Manohar, E. Clark, B. Brunton, J. Kutz, and S. Brunton, “PySensors: A Python package for sparse sensor placement,” J Open-Source Software, vol. 6, no. 58, p. 2828, Feb. 2021, doi: 10.21105/joss.02828.

11. A. Hetherington and S. Le Clainche, “Low-cost singular value decomposition with optimal sensor placement,” 2023.

12. L. Sirovich and J. D. Rodriguez, “Coherent structures and chaos: A model problem,” Phys Lett A, vol. 120, no. 5, pp. 211–214, Feb. 1987, doi: 10.1016/0375-9601(87)90209-X.

13. L. De Lathauwer, B. De Moor, and J. Vandewalle, “A MULTILINEAR SINGULAR VALUE DECOMPOSITION *,” 2000. [Online]. Available: http://www.siam.org/journals/simax/21-4/30569.html.

14. Bristol City Council. Air Quality Annual Status Report 2023. Available at: https://services.bristol.gov.uk/files/documents/6801-air-quality-annual-status-report-2023.

15. Bristol City Council. Air Quality Annual Status Report 2022. Available at: https://www.bristol.gov.uk/files/documents/5074-air-quality-annual-status-report-2022.

16. Bristol City Council. Air Quality and Pollution. Available at: https://www.bristol.gov.uk/residents/pests-pollution-noise-and-food/air-quality-and-pollution/air-quality.

17. European Environment Agency. Air Quality in Europe - 2022 Report. Available at: https://www.eea.europa.eu/publications/air-quality-in-europe.

18. Story Maps. Madrid: Urban Air Pollution Challenges. Available at: https://storymaps.arcgis.com/stories/c1add8f6180040af8cd27528edd2db6e.

19. Bañuelos-Gimeno, Jorge & Sobrino, Natalia & Arce-Ruiz, Rosa. (2024). Initial Insights into Teleworking’s Effect on Air Quality in Madrid City. 10.20944/preprints202408.1234.v1.