POLITECNICO DI TORINO

MASTER's Degree in COMPUTER ENGINEERING

MASTER's Degree Thesis

Multimodal Transformer Architecture for Urban Air Quality Forecasting: An Ablation Study

Supervisors

Prof. Andrea BOTTINO

Luca BARCO

Lorenzo INNOCENTI

Giacomo BLANCO

Candidate

Christian D'ALLEVA

MARCH 2026

Multimodal Transformer Architecture for Urban Air Quality Forecasting: An Ablation Study

Christian D'Alleva

Abstract

Multimodal learning has become increasingly relevant in environmental modeling, where heterogeneous sources such as satellite observations and meteorological measurements are jointly used to estimate urban air pollutant concentrations. Despite encouraging predictive performance, it remains unclear which modalities and temporal components truly contribute useful information, and which instead introduce redundancy or noise. Understanding these relationships is essential for designing efficient and interpretable models rather than progressively more complex architectures. This thesis proposes a multimodal framework for urban air quality forecasting over Milan. The approach integrates multi-sensor satellite imagery (Sentinel-1, Sentinel-2, Sentinel-3, and Sentinel-5), land-cover data, and a digital elevation model together with meteorological observations. These data are fused within a self-supervised pretraining strategy based on an adaptation of the VideoMAE architecture, enabling the extraction of representative spatio-temporal features describing air quality dynamics. The learned latent representation is subsequently combined with weather forecasts in a downstream prediction task aimed at estimating pollutant concentrations up to seven days ahead. The final output is a map of Milan representing predicted concentrations of PM_{10}, $PM_{2.5}$, O_3, NO_2, and SO_2. An ablation study is conducted to evaluate the contribution of each modality and to analyse cross-modal interactions. Model behaviour is assessed through variations in predictive performance and comparative modality importance.

Table of Contents

List of Figures

List of Tables

Acronyms

AQ	Air Quality.
ARIMA	Autoregressive Integrated Moving Average.
ANN	Artificial Neural Network.
LSTM	Long Short-Term Memory.
WHO	World Health Organization.
SO2	Sulfur Dioxide.
NO2	Nitrogen Dioxide.
O3	Ozone.
PM	Particulate Matter.
IARC	International Agency for Research on Cancer.
CTMs	Chemical Transport Models.
SVMs	Support Vector Machines.
SGD	Stochastic Gradient Descent.
CNN	Convolutional Neural Networks.
CAMS	Copernicus Atmosphere Monitoring Service.
ViTs	Vision Transformers.

Chapter 1

Introduction

Air pollution represents one of the most pressing environmental and public health challenges worldwide. According to World Meteorological Organization [1], air pollution is often linked to health problems and economic losses. Beyond its impact on human health, poor air quality also affects ecosystems, reduces visibility through haze formation, damages infrastructure, and threatens food and water security. Certain atmospheric pollutants, such as tropospheric ozone, act as climate forcers, while aerosols contribute to atmospheric cooling, further highlighting the complex interactions between air quality and climate systems.

Accurate air quality forecasting plays a crucial role in mitigating these impacts by predicting future pollutant concentrations and enabling early warning systems for vulnerable populations. Recent forecasting systems typically rely on a combination of meteorological information, emission inventories, and atmospheric chemistry models to estimate future pollution levels. Ground monitoring stations provide reliable measurements of pollutant concentrations, while satellite observations offer large-scale spatial coverage of atmospheric variables.

Atmospheric processes involve highly non-linear interactions between meteorological conditions, emission sources, chemical transformations, and transport phenomena. Traditional statistical approaches, such as regression models and ARIMA methods [2], often struggle to capture these complex dynamics. In recent years, machine learning approaches, including Artificial Neural Networks and Long Short-Term Memory networks [3], have demonstrated improved predictive capabilities by modeling non-linear relationships in environmental data.

More recently, the Transformer architecture has emerged as a powerful modeling paradigm capable of capturing long-range dependencies in complex data. Originally developed for natural language processing, Transformers have also demonstrated strong performance in computer vision tasks such as image classification, object detection, semantic segmentation, and video understanding. Their self-attention mechanism enables the modeling of global spatial and temporal relationships while reducing the need for strong inductive biases, making them particularly suitable for multimodal environmental data.

In the context of air quality prediction, the integration of multiple heterogeneous

data sources has become increasingly important. Data fusion approaches aim to combine complementary information from satellites, ground monitoring networks, and environmental datasets to generate improved estimates and forecasts of pollutant concentrations. Unlike traditional data assimilation techniques that update the state of a numerical model, data fusion methods integrate diverse data modalities to produce a new predictive representation. Satellite-derived measurements, such as aerosol optical depth, are particularly valuable for capturing variability at scales that cannot be fully resolved by ground stations alone.

Building on these developments, this thesis investigates a multimodal Transformer-based framework for air quality forecasting over the metropolitan area of Milan. The proposed approach leverages heterogeneous environmental data sources, including satellite imagery, meteorological variables, land cover information, and digital elevation models, to predict pollutant concentrations measured by ground monitoring stations. By learning joint representations across heterogeneous environmental modalities, the proposed model focuses on predicting future pollutant concentration trends at the level of individual ground monitoring stations.

In addition to the forecasting framework, this work includes a systematic ablation study designed to evaluate the contribution of each data source to the predictive performance. By progressively analyzing different modality combinations, the study provides a clearer understanding of the relative importance of satellite imagery, weather conditions, land cover, and elevation information in the air quality forecasting task. This analysis allows us to quantify the value that each environmental data source brings to station-level pollutant prediction and to assess how multimodal fusion improves forecasting accuracy in complex urban environments.

Nevertheless, several challenges remain. Air pollution dynamics are influenced by unpredictable factors such as sudden emission changes, extreme meteorological events, and non-linear chemical reactions in the atmosphere. These phenomena make accurate forecasting particularly difficult, especially for extreme pollution episodes.

Chapter 2

Air Pollution in Urban Environments: Sources, Impacts, and Socio-Economic Drivers

Urban air quality is a complex and multifaceted issue, shaped by diverse socio-economic conditions that vary across regions of the world and even within individual cities. Urban areas, by their very nature, concentrate populations, materials, and activities, which makes them both hotspots of pollutant emissions and major receptors of their impacts. Air pollution operates across multiple spatial and temporal scales: from immediate, localized effects on human health and material degradation; to regional phenomena such as acidification and forest decline unfolding over decades; and ultimately to global processes capable of altering environmental conditions for both humans and ecosystems over centuries. In this context, cities function primarily as significant sources of air pollution.

Historically, outdoor air pollution was predominantly an urban phenomenon, and both literature and historical records indicate that its impacts were substantial. [4] These impacts may, however, have been underestimated, as earlier populations were generally less critical of their living conditions and lacked the scientific tools necessary to assess long-term health effects, such as those associated with carcinogenic substances. Moreover, many historical accounts focused on aesthetic concerns, such as odor and surface soiling, which, while undesirable, are not necessarily harmful to health. It should also be noted that until the Second World War, public attitudes toward pollution were often ambivalent, as industrial emissions were frequently perceived as symbols of economic growth and prosperity. Following the Second World War, global demographic and urban dynamics changed rapidly. The world population increased from approximately 2.5 billion to 5.9 billion by the mid-1990s, with 1.2 billion residing in more developed countries and 4.7 billion in less developed regions, including 1.2 billion in China alone. During the same period, global urbanization, defined as the proportion of people living in settlements with more than 2,000 inhabitants, rose from less than 30% to approximately 44%. Urbanization levels reached an average of 73% in more developed countries and 36%

in less developed countries. This rapid urban growth, particularly in developing regions, has been driven by both agricultural mechanization and the expansion of industrial, commercial, and public service sectors, prompting large-scale rural-to-urban migration. In Asia, Latin America, and Africa, urbanization has frequently been accompanied by the expansion of slums and informal settlements. In some cases, environmental conditions have been further exacerbated by the relocation of polluting industries from industrialized countries, where environmental regulations are stricter and labor costs higher. Consequently, regions experiencing high birth rates and significant immigration often face severe environmental challenges arising from unplanned urban expansion and the emergence of megacities.

2.1 Major Air Pollutants

According to World Health Organization (WHO)[5], air pollution is contamination of the indoor or outdoor environment by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere. Household combustion devices, motor vehicles, industrial facilities and forest fires are common sources of air pollution. Pollutants of major public health concern include particulate matter, ozone, nitrogen dioxide and sulfur dioxide.

Sulfur Dioxide (SO_2) Sulfur dioxide is a colorless gas that is readily soluble in water and is primarily generated by the combustion of sulfur-containing fossil fuels used for domestic heating, industrial processes, and electricity generation. Emissions of SO_2 can be effectively reduced through the use of low-sulfur fuels, such as natural gas or low-sulfur oil, as alternatives to coal. In large-scale industrial and power generation facilities, particularly in industrialized countries, flue gas desulfurization technologies are widely implemented and have proven to be effective in controlling SO_2 emissions. Sulfur dioxide exposure is closely linked to increased hospital admissions and emergency room visits for asthma and other respiratory conditions, especially among vulnerable populations such as children, the elderly, and individuals with pre-existing respiratory diseases.

Nitrogen Oxides (NO_x) Nitrogen oxides are formed through the oxidation of atmospheric nitrogen during high-temperature combustion processes. Nitrogen dioxide (NO_2), a major component of NO_x, is a reddish-brown, water-soluble gas and a strong oxidizing agent. Ambient NO_2 concentrations primarily originate from fuel combustion associated with transportation, residential heating, industrial activities, and power generation. In domestic environments, sources include fuel-burning appliances such as furnaces, fireplaces, gas stoves, and ovens. Most NO_x emissions, particularly from motor vehicles, are initially released as nitric oxide (NO), a relatively non-toxic compound that is subsequently oxidized in the atmosphere to form the secondary pollutant NO_2. Emission reductions can be achieved through combustion optimization techniques, such as low-NO_x burners in power plants and

lean-burn engines in vehicles, as well as through the use of catalytic converters in exhaust systems. Exposure to nitrogen dioxide (NO_2) has been shown to irritate the airways and exacerbate existing respiratory diseases, particularly asthma. In addition to its direct effects, NO_2 plays a critical role as a precursor in the formation of ground-level ozone, a pollutant strongly associated with respiratory morbidity.

Ozone (O_3) Ozone and associated photochemical oxidants represent secondary atmospheric pollutants, as they are synthesized through chemical pathways rather than being emitted directly from primary sources. This synthesis involves a complex series of atmospheric reactions driven by the absorption of solar radiation by nitrogen dioxide (NO_2) molecules. Such energy transfers initiate the dissociation of NO_2, leading to the formation of ground-level ozone through catalytic cycles.[6]

The presence of O_3 in the troposphere poses a significant threat to biological systems, specifically through the induction of inflammatory lung injury and the exacerbation of chronic respiratory conditions such as asthma and Chronic Obstructive Pulmonary Disease (COPD). Beyond respiratory distress, clinical evidence correlates prolonged exposure with oncogenesis and systemic impairments affecting neurological, metabolic, and reproductive health.

Particulate Matter (PM) Particulate matter consists of a complex mixture of inhalable solid particles and liquid droplets suspended in air, including sulfates, nitrates, ammonium, sodium chloride, black carbon, mineral dust, and water. PM is commonly classified based on aerodynamic diameter, with PM_{10} (particles smaller than 10 µm) and $PM_{2.5}$(particles smaller than 2.5 µm) being the most relevant for regulatory purposes and human health. Coarse particles (2.5–10 µm) are primarily derived from natural and mechanical processes, such as wind-blown dust, sea spray, pollen, agricultural activities, road dust, and mining operations. Fine particles ($PM_{2.5}$) originate from both primary sources, including fuel combustion in vehicles, industries, and power plants, and secondary sources formed through atmospheric chemical reactions involving gaseous precursors. In indoor environments, the largest sources of particulate matter are typically the combustion of solid or liquid fuels in open hearths or inefficient, poorly ventilated stoves and heaters. Household activities such as cooking, space heating, lighting, preparation of animal fodder, water heating, and beverage brewing can significantly contribute to indoor PM levels. In outdoor environments, PM sources are location-specific but generally include traffic and transportation, industrial activities, power generation, construction, waste burning, and biomass fires. The health effects of particulate matter, particularly particles with aerodynamic diameters below 10 µm (PM_{10}) and 2.5 µm ($PM_{2.5}$), are among the most extensively documented. Fine particles are capable of penetrating deep into the respiratory system and entering the bloodstream, where they contribute to cardiovascular diseases (including ischemic heart disease), cerebrovascular events (such as stroke), and respiratory illnesses. Both short- and long-term exposure to PM are associated with increased morbidity and mortality from cardiovascular and

respiratory causes. Long-term exposure has also been linked to adverse perinatal outcomes and lung cancer. In 2013, PM was classified as a carcinogen by the World Health Organization's International Agency for Research on Cancer (IARC), and it is now widely used as a key indicator for assessing the health impacts of air pollution. By contrast, coarse particles with diameters greater than 10 µm have not been consistently associated with significant health effects.

2.1.1 Broader Impacts, Urban Planning and Mitigation Strategies

Beyond human health, urban air pollution affects materials, vegetation, including urban agriculture, and visibility. The extent of material damage depends not only on pollutant concentrations but also on environmental factors such as temperature, humidity, and interactions among pollutants. Nevertheless, impacts on human health and well-being remain the primary driver of air pollution control strategies, even though urban health outcomes are influenced by multiple social, economic, and environmental determinants. [4] Outdoor air pollution is recognized as a major global public health challenge. In 2019 alone, it was responsible for approximately 4.14 million non-accidental premature deaths worldwide, affecting both urban and rural populations. In 2016, 54% of the global population resided in urban areas, where $PM_{2.5}$, NO_2, and ground-level ozone are considered the most harmful pollutants for human health. Projections indicate that by 2050, approximately 70% of the world's population will live in urban areas, potentially increasing premature deaths attributable to outdoor air pollution to 6.6 million annually. [7] The impacts of urban air pollution can be mitigated through effective urban planning and integrated land-use strategies. Earlier approaches, which emphasized strict separation between industrial and residential areas, are now considered outdated, as they often result in increased commuting, traffic congestion, and overall emissions. Measures such as driving restrictions, economic incentives (e.g., congestion pricing and green taxes), parking limitations, and pedestrian zones have achieved partial success, though they have sometimes faced opposition from commercial sectors. Moreover, such restrictions may unintentionally promote the expansion of commercial centers outside city cores, leading to increased total traffic volumes. Contemporary urban planning emphasizes integrated land use aimed at minimizing transportation demand and reducing overall emissions. The inclusion of green spaces and urban parks plays a key role in enhancing environmental quality, particularly in residential areas. While restructuring options are limited in existing cities, infrastructure solutions such as ring roads that divert traffic away from city centers remain a viable mitigation strategy. [4]

Chapter 3

Related Works

3.1 Overview of Air Quality Prediction Challenges

Air pollution constitutes a major environmental and public health challenge, contributing to the intensification of the climate crisis and to broader environmental degradation. Exposure to contaminated air significantly increases the risk of respiratory and cardiovascular diseases, making the accurate prediction of pollutant concentrations essential for effective mitigation strategies and informed environmental management. Air pollutants are generally defined as a mixture of particulate matter and gaseous compounds, including SO_2, $PM_{2.5}$, PM_{10}, NO_2, and O_3. Forecasting their concentration is inherently complex due to the presence of dynamic dependencies in both spatial and temporal dimensions. From a spatial perspective, pollutant concentrations at a given monitoring site are influenced by surrounding locations and by meteorological factors such as temperature, humidity, and wind direction. Pollutants can disperse across neighboring areas, and the magnitude and direction of this influence vary over time. This phenomenon, referred to as dynamic spatial correlation, implies that air-quality prediction cannot be treated as an isolated single-site task; instead, models must capture nonlinear interactions among geographically distributed monitoring stations. From a temporal perspective, pollutant levels evolve continuously and exhibit dynamic temporal correlation. Historical observations contribute unequally to future concentrations, meaning that past time steps exert varying levels of influence on prediction time steps. Consequently, predictive models must dynamically adjust the weight assigned to historical data and account for real-time temporal dependencies.[8]

3.2 Evolution of Air Quality Forecasting Methodologies: Non Deep-Learning Approaches

Prior to the integration of deep learning architectures and multimodal fusion strategies, air pollutant concentration prediction primarily utilized two paradigms: deterministic physical models and classical statistical models. This classification follows the taxonomy established by Zhang et al. [8], which serves as the foundational framework

for analyzing the field's progression.

Deterministic (Physics-based) Models Deterministic approaches leverage established principles of atmospheric physics and chemistry to simulate the explicit dynamics of pollutants. These simulations account for discrete processes including emission, transport, transformation, and removal within the atmosphere. To execute these simulations, researchers employ numerical formulations, commonly implemented as Chemical Transport Models (CTMs), that solve complex differential equations governing meteorological and chemical interactions. Standard frameworks in this category include the Weather Research and Forecasting model with Chemistry (WRF-Chem)[9], the Community Multiscale Air Quality (CMAQ) modeling system, and the Comprehensive Air Quality Model with Extensions (CAMx) [10],[11],[12].

These systems function by integrating high-dimensional data, specifically meteorological variables, emission inventories, and specific chemical reaction mechanisms. While such integration ensures high interpretability and physical consistency, the reliance on rigid mathematical formulations introduces significant constraints. The most critical limitation involves the inability of these models to represent the highly nonlinear interactions occurring between heterogeneous environmental factors. This lack of flexibility often causes predictive performance to degrade in urban environments, where pollutant formation is driven by complex, multi-source interactions that exceed the resolution of standard physical equations.

Statistical Methods In contrast to the explicit modeling of physical laws, classical statistical methods treat pollutant concentration as a stochastic process governed by structured temporal dependencies. These data-driven approaches aim to identify relationships within historical measurements without requiring a prior understanding of atmospheric chemistry. A prominent example of this paradigm is the Autoregressive Integrated Moving Average (ARIMA) model, which synthesizes autoregressive and moving-average components to forecast short-term fluctuations [2].

The efficacy of ARIMA and its derivatives depends on the availability of substantial historical datasets and the application of specific empirical rules. Despite their utility in stable conditions, these methods generally assume linear or near-linear relationships within the underlying data. Consequently, they fail to capture the complex, non-linear correlations inherent in atmospheric pollutants, particularly during rapid environmental shifts or extreme weather events. This fundamental inability to model deep nonlinearity necessitates a transition toward the more robust feature extraction capabilities found in contemporary deep learning frameworks.

Machine Learning Methods To address the inherent linearity constraints of classical statistical models, researchers shifted toward machine learning architectures capable of mapping nonlinear relationships without the need for explicit atmospheric chemistry modeling. This paradigm shift replaced rigid physical equations with data-driven algorithms, most notably Artificial Neural Networks (ANNs) and Support

Vector Machines (SVMs). Liu et al. [13] provide a foundational comparison between SVM and Random Forest approaches, illustrating the superior flexibility of these methods in high-dimensional air quality forecasting.

The adoption of ANNs in air pollution forecasting has increased significantly, primarily because these models offer a viable alternative to deterministic systems that are sensitive to parameter quality and high computational costs. Unlike physics-based approaches that require exhaustive databases of input parameters, many of which are often unavailable, ANNs operate effectively without an in-depth understanding of the underlying environmental dynamics. This operational independence stems from the use of nonlinear transfer functions, such as the sigmoid function, which allows the network to approximate the complex interactions between predictors like wind speed or traffic density and the resulting target pollutant concentrations. Despite these advantages, the nature of ANNs introduces a degree of uncertainty regarding model interpretability. Furthermore, traditional machine learning models generally learn shallow feature representations, which limits their ability to extract deep spatio-temporal dependencies from large-scale, heterogeneous datasets. This structural limitation prevents shallow models from fully capturing the multi-scale variations inherent in urban air quality.[3]

Following the growing adoption of data-driven techniques, recent research has explored the integration of heterogeneous environmental data sources to enhance air-quality prediction performance. In this context, the study Urban Air Pollution Forecasting: A Machine Learning Approach Leveraging Satellite Observations and Meteorological Forecasts [14] represents a significant advancement by combining satellite remote sensing with traditional meteorological information for urban pollution forecasting on the Milan metropolitan area. The researchers combine multiple data sources , Sentinel-5P satellite observations, meteorological data, topographical information, temporal variables, and weather forecasts, to predict short-term concentrations of five major air pollutants. A key contribution of the work is the development of monitoring-station-independent models, allowing pollution prediction even in regions lacking ground measurement infrastructure. The authors trained separate prediction models for each pollutant, using historical environmental and weather data within a defined time window. Three machine-learning techniques were compared: Linear Regression, Gradient Boosting Regression and Stochastic Gradient Descent (SGD) Regression. Each model estimates pollutant concentrations for the following day based on past satellite, meteorological, topological, temporal, and forecast features. Experiments carried out in the Milan metropolitan area show that the models effectively predict next-day pollutant concentrations, achieving an average prediction error of approximately 30%.

Motivation Toward Deep and Multimodal Learning While traditional machine learning offers improvements over deterministic and classical statistical approaches, these methods remain insufficient for modeling the high-dimensional nature of contemporary environmental data. Air pollution formation is influenced by a

simultaneous convergence of meteorology, emission inventories, geography, mobility patterns, and remote sensing observations. These diverse inputs produce strong nonlinear and cross-modal correlations that exceed the processing capacity of shallow architectures. Conventional approaches typically operate on single-source or low-dimensional inputs, which leads to a failure in modeling three critical dimensions: long-range temporal dependencies, spatial interactions among distributed monitoring stations, and cross-modal relationships between meteorological and urban data. The inability to synthesize these disparate data streams effectively necessitates the adoption of deep and multimodal learning frameworks. Such frameworks are designed to learn hierarchical representations and jointly model spatial, temporal, and heterogeneous environmental information at a scale previously unattainable by shallow machine learning methods.

3.3 Deep Learning Approaches

Deep learning architectures have significantly advanced the field of air pollutant concentration forecasting by enabling the extraction of hierarchical, nonlinear, and high-dimensional representations from heterogeneous environmental data. Unlike shallow artificial neural networks, these models utilize multiple stacked nonlinear layers to progressively abstract complex feature interactions. This hierarchical depth facilitates the modeling of intricate atmospheric dynamics, which typically involve strong spatial dependencies, long-range temporal correlations, and cross-variable nonlinear interactions. Based on their structural capacity, deep learning approaches for air quality (AQ) prediction are categorized into three primary modeling paradigms.

1. **Spatial modeling architectures**: such as Convolutional Neural Networks (CNN) and Graph Convolutional Networks (GCN), focus on terrestrial and topological distributions.

2. **Temporal modeling architectures**: including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), capture sequential dependencies and accumulation patterns.

3. **Spatio-Temporal Deep Learning (STDL)**: frameworks integrate these modules to jointly model the simultaneous transport and transformation of pollutants.

3.3.1 Modeling Spatial Correlations

Convolutional Neural Networks (CNN) based approaches Convolutional Neural Networks are deep feedforward architectures composed of stacked layers designed to extract multiscale, shift-invariant features. These networks utilize convolutional filters and nonlinear activation functions, such as the Rectified Linear Unit (ReLU), to learn localized spatial dependencies. To enhance translation invariance and reduce computational dimensionality, practitioners frequently employ pooling

operations, such as max or average pooling. In the context of AQ prediction, CNNs are utilized to identify spatial correlations among monitoring regions. Researchers often transform pollutant and meteorological measurements into one-dimensional or two-dimensional tensors to facilitate the learning of latent spatial patterns. However, because monitoring sites are frequently distributed irregularly across a landscape, the naive reshaping of data into regular grids can distort original geographic relationships. To mitigate this distortion, certain studies standardize monitoring locations into spatial grids where each cell represents a discrete region. These structured inputs allow the CNN to extract meaningful neighborhood-level features centered on a target prediction site. Further advancements, inspired by computer vision, treat satellite imagery or pollution-related visual data as primary inputs for CNN-based models. These works employ deep convolutional stacks to learn spatial features from images, often incorporating dropout layers to prevent overfitting. Despite their efficacy in Euclidean domains, CNNs inherently assume grid-like structures, which restricts their utility when modeling the irregular spatial topologies common in environmental monitoring networks.

A significant application of this technology is found in the study by [15], which proposes an ensemble model to predict concentrations from visual data. This framework utilizes three modified CNN architectures, VGG-16, Inception-v3, and ResNet50, as base learners, with a feed-forward neural network acting as a meta-learner. Through transfer learning from the ImageNet dataset, these CNNs are adapted for regression tasks by replacing final classification layers with a single linear output layer optimized via Mean Squared Error (MSE) loss. The meta-learner then nonlinearly combines the outputs of these three models to produce a refined final estimate.

Graph Convolutional Networks (GCN) based approaches To address the Euclidean limitations associated with CNNs, researchers have increasingly adopted Graph Convolutional Networks (GCN) for AQ prediction. Unlike traditional convolutional layers, GCNs operate on non-Euclidean graph structures, which are inherently suited for modeling irregularly distributed monitoring stations. In these applications, monitoring sites are represented as graph nodes, while the edges encode spatial relationships such as geographic proximity or similarity in pollutant dynamics. Spectral-based GCNs define convolution operations within the graph Fourier domain by utilizing the graph Laplacian matrix. By propagating information along these defined edges, the model effectively captures spatial interactions, such as the physical transport of pollutants between neighboring regions. Consequently, GCNs preserve the intrinsic topology of monitoring networks more effectively than grid-based approaches, enabling adaptive and precise spatial feature extraction.

3.3.2 Modeling Temporal Correlations

Air pollutant concentrations evolve over time through a combination of accumulation, chemical reactions, and meteorological dynamics. These time-varying processes

necessitate the use of temporal modeling as a fundamental component of air-quality prediction systems. Deep learning approaches capture such dependencies by utilizing convolutional, recurrent, and attention-based sequence architectures.

Convolutional Temporal Models Although originally designed for spatial feature extraction, Convolutional Neural Networks (CNNs) can also process sequential data through one-dimensional convolutions. Fully convolutional sequence models demonstrate that temporal dependencies can be learned via stacked convolutional filters operating across discrete time windows. In air-quality forecasting, deep temporal CNNs extract patterns from historical pollutant measurements to predict short-term concentrations of pollutants. A primary advantage of convolutional sequence modeling lies in parallel computation, which allows for more efficient GPU utilization than is possible with recurrent architectures.

A more effective variant within this category is the Temporal Convolutional Network (TCN) [16]. This architecture employs dilated causal convolutions to enlarge the receptive field without increasing computational overhead. By adjusting dilation factors, TCNs capture long-range temporal dependencies while maintaining stable gradients during backpropagation. Unlike recurrent networks, convolutional models operate on fixed receptive fields, making them faster to train and easier to optimize. However, the lack of explicit memory states in these models can limit their adaptability when temporal patterns vary dynamically over long horizons.

Recurrent Neural Networks Recurrent architectures explicitly model sequential dependencies by propagating hidden states across successive time steps. The classical Recurrent Neural Network (RNN) introduces feedback connections that enable information persistence, allowing current predictions to depend on a sequence of previous observations.

To address the vanishing and exploding gradient problems inherent in standard RNNs, Long Short-Term Memory (LSTM) networks incorporate gated memory cells. These cells control the flow of information through input, output, and forget gates, enabling the model to learn long-term relationships within pollutant time series. A simplified alternative is the Gated Recurrent Unit (GRU), which merges these operations into update and reset gates. GRUs require fewer parameters and typically achieve faster convergence while maintaining performance comparable to LSTMs in many air-quality prediction scenarios.

Despite becoming standard tools for modeling nonlinear temporal dependencies in environmental data, the sequential nature of recurrent models limits parallelization. Some researchers attempt to overcome these hurdles by utilizing encoder-decoder structures. Specifically, the study by [17] introduces the EDSModel, which aims to improve prediction accuracy by capturing deep temporal patterns in both pollutant and meteorological data.

The EDSModel architecture follows a dual-component design. The encoder utilizes a Read-first LSTM (RLSTM), which enhances the traditional LSTM by interrelating

the gating mechanisms to filter redundant input information. This modification improves long-term feature extraction and captures the complex relationships between pollutants and their meteorological drivers. Following the encoding phase, the decoder employs a multi-layer LSTM to generate future pollutant concentration forecasts based on the compressed contextual features.

3.4 Spatio-Temporal Deep Learning for Air Quality Forecasting: Advances and Challenges

Spatio-Temporal Deep Learning (STDL) has emerged as a promising framework for air quality forecasting by jointly modeling spatial transport processes and temporal pollutant dynamics. By integrating spatial encoders (e.g., CNN/GCN) with temporal sequence models (e.g., LSTM/GRU/TCN), these hybrid architectures capture long-term dependencies, nonlinear correlations, irregular monitoring networks, and non-stationary pollutant behavior. Such capabilities are particularly valuable for large-scale, multimodal forecasting scenarios that combine meteorological, satellite, emission, and sensor data. However, key challenges remain. Many models struggle to simultaneously model both long- and short-term dependencies, exhibit limited transferability across heterogeneous urban environments, lack interpretability for decision-making, demand substantial computational resources, and face difficulties in effectively integrating heterogeneous data sources. Further research is therefore required to improve generalization, efficiency, explainability, and multimodal data fusion in STDL-based AQI prediction systems.

Addressing these constraints, Guo et al.[18] developed an end-to-end framework designed for citywide, high-resolution forecasting that moves beyond traditional site-specific paradigms. Their approach accounts for the spatial heterogeneity driven by urban morphology by utilizing dense monitoring data from 417 micro-stations in Lanzhou, China. This micro-station data enables the generation of fine-grained $PM_{2.5}$ forecasts at a spatial resolution of 500 m $\times$ 500 m and a temporal frequency of one hour. The architecture, termed Air-PredNet, utilizes a spatio-temporal transformation module to convert discrete station observations into structured spatial maps. These maps serve as inputs for a ConvLSTM-based prediction module, which treats sequences of pollutant distributions as high-dimensional tensors. By processing these sequences, the model captures the underlying fluid dynamics and transport processes required for multi-step ahead forecasting across the entire urban grid.

Alternative approaches to capturing temporal patterns involve the synthesis of attention mechanisms and recurrent units. Liu et al. [19]introduced a hybrid framework integrating a Transformer encoder with a Bidirectional Long Short-Term Memory (BiLSTM) network to refine AQI forecasting accuracy. Their study utilized a comprehensive dataset from Shijiazhuang, Beijing, and Tianjin, spanning from November 2013 to February 2025. This dataset encompasses the six primary pollutants: $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3. The Transformer–BiLSTM

architecture exploits the global dependency modeling of the self-attention mechanism to identify long-range temporal associations. These global features are subsequently processed by the BiLSTM layer, which extracts localized, bidirectional temporal patterns to enhance feature representation. To ensure numerical stability during this process, the raw time-series underwent Butterworth filtering for noise reduction and normalization to the range $[-1, 1]$. The final AQI values were generated via a fully connected layer, with the optimization process employing AdamW and a learning rate scheduler to minimize Mean Squared Error (MSE) loss.

3.5 VideoMAE Self-Supervised Pre-Training Architecture

In this thesis, the following architecture is adopted as the pre-training framework for the air quality forecasting task. The model is trained in a self-supervised manner to learn latent representations from satellite image time series through a reconstruction-based objective. The paper proposes Video Masked Autoencoders (VideoMAE), a self-supervised framework designed to improve the data efficiency of video transformer pre-training. Prior work has shown that vision transformers typically require large-scale labeled datasets or pre-training on extensive image corpora, which limits their applicability to video understanding tasks due to the relatively small size of available video datasets. To address this limitation, the authors extend masked autoencoding, previously successful in natural language processing and image representation learning, to the video domain. VideoMAE formulates self-supervised video pre-training as a reconstruction task in which a substantial portion of spatiotemporal video tokens is masked and subsequently reconstructed. Recognizing the strong temporal redundancy inherent in videos, the method introduces an extremely high masking ratio (90-95%) together with a tube masking strategy that removes contiguous regions across time. This design mitigates information leakage between adjacent frames and prevents models from exploiting low-level temporal correlations, thereby encouraging the learning of higher-level semantic representations.

Architecture The proposed approach enables effective training of vanilla Vision Transformer (ViT) architectures directly on video datasets without reliance on external image data or supervised initialization. The VideoMAE framework introduces several architectural and methodological adaptations to address challenges specific to masked video modeling, particularly temporal redundancy and information leakage across frames. The overall pipeline performs self-supervised pre-training by reconstructing masked spatiotemporal regions using an asymmetric encoder–decoder architecture built upon a vanilla Vision Transformer (ViT) backbone with joint space–time attention. To improve efficiency, the method first applies temporal downsampling, where video frames are paired together in temporal strategy to reduce redundancy among consecutive frames. This process compresses the temporal dimension while preserving essential motion information, enabling more computationally

efficient training. The input video is subsequently tokenized through joint space–time cube embedding, which divides videos into 3D spatiotemporal cubes, called tublets, treated as tokens. A central contribution of VideoMAE lies in its tube masking strategy with extremely high masking ratios (90–95%) . Given the lower information density of videos compared to images, aggressive masking increases reconstruction difficulty and encourages the model to learn high-level semantic representations rather than relying on local spatiotemporal correlations. The proposed temporal tube masking further enforces identical masking patterns across frames, ensuring that masked regions remain unavailable throughout the temporal axis. This design prevents the model from recovering missing content through neighboring frames and reduces information leakage, particularly for regions exhibiting limited motion. The framework employs a vanilla ViT backbone with joint space–time self-attention to capture global spatiotemporal dependencies among the remaining visible tokens. Although joint attention mechanisms typically incur high computational cost, the extremely high masking ratio substantially reduces encoder input tokens, thereby improving computational efficiency during pre-training. Empirical results demonstrate that VideoMAE achieves competitive or state-of-the-art performance across multiple benchmark datasets while requiring significantly less training data. VideoMAE architecture is presented in Figure3.1

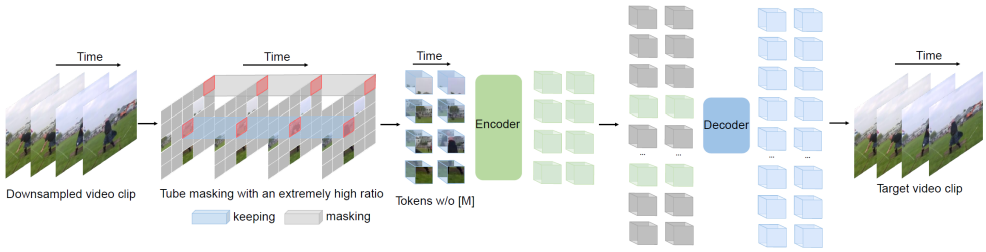

Figure 3.1: VideoMAE [20] architecture

Chapter 4

Dataset

This chapter introduces the dataset sources employed in the present study, which are directly related to the context of air quality forecasting. The dataset integrates multiple complementary data sources, including Sentinel-1, Sentinel-2, Sentinel-3, Sentinel-5P, Open-Meteo, land cover information, and a digital elevation model. The temporal range considered spans from 2018 to 2024, a period during which all these sources provide continuous data coverage, although some of them are available over longer historical intervals. All datasets were preprocessed as necessary to ensure temporal and spatial consistency across the observation frames, thereby enabling a coherent multimodal analysis. Detailed preprocessing methodologies and considerations are discussed in the following chapters. The integration of these diverse data sources forms the foundation of an ablation study designed to evaluate the relative contribution of each source within a multimodal air quality forecasting framework. This approach allows for a systematic assessment of the most informative datasets in the urban air quality context and establishes a baseline for further investigation into previously unexplored aspects of urban air quality dynamics.

In the subsequent sections, the characteristics of each data source are described in detail, including the specific parameters and channels analyzed, providing an overview of the information leveraged for this study.

4.1 Ground Stations

The ground-truth dataset used in this study consists of air pollutant measurements collected from ground monitoring stations and obtained from the open data portal of the Municipality of Milan [21]. The dataset provides daily pollutant concentration measurements expressed in $\mu g/m^3$ and spans multiple years; however, the analysis period was restricted to May 2018–December 2024 to ensure temporal alignment with the availability of satellite observations.

The study focuses on five atmospheric pollutants employed in the computation of the Air Quality Index (AQI) according to the European Environmental Agency guidelines:

- Particulate matter (PM_{10} and $PM_{2.5}$)

- Nitrogen dioxide (NO_2)
- Sulfur dioxide (SO_2)
- Ozone (O_3)

Measurements are collected from five monitoring stations distributed across the Milan metropolitan area. Not all stations record every pollutant, and the dataset contains occasional missing observations. These ground-based measurements serve as reference data for training and validating deep learning models aimed at air quality prediction.

As described by the Municipality of Milan, pollutant concentration limits and health-related standards follow regulatory thresholds defined by law, including daily and annual limit values for PM_{10} and NO_2, as well as exposure thresholds for ozone based on the maximum daily 8-hour moving average. The dataset therefore provides both historical pollutant time series and standardized environmental indicators relevant for assessing human health impacts and supporting predictive modeling of air quality conditions. Figure 4.1 illustrates the spatial distribution of the monitoring stations, along with the specific pollutants measured at each location.

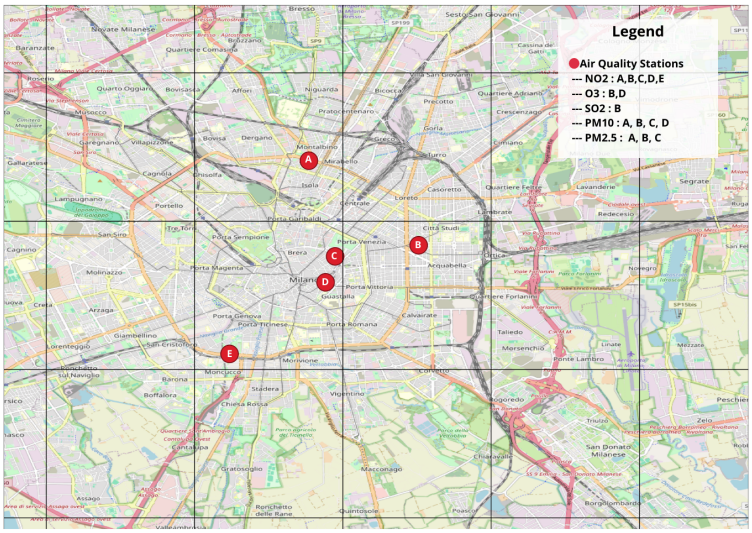

Figure 4.1: Spatial distribution of monitoring stations in Milan and the pollutants measured at each location.

4.2 Satellite Sources

Sentinel 1 The Sentinel-1 mission [22] represents the European radar observation component of the Copernicus Programme, a joint initiative of the European Commission (EC) and the European Space Agency (ESA) aimed at providing operational environmental monitoring and security services through satellite and in-situ observations. Sentinel-1 delivers continuous Earth observation data to support applications

related to land monitoring, ocean surveillance, climate studies, and environmental assessment.

The mission consists of a constellation of two sun-synchronous polar-orbiting satellites, Sentinel-1A and Sentinel-1B, operating in the same orbital plane with a 180° phase difference. Equipped with a C-band Synthetic Aperture Radar (C-SAR) instrument, Sentinel-1 provides all-weather, day-and-night imaging capabilities independent of cloud cover and illumination conditions. Each satellite operates with a 12-day repeat cycle, which is reduced to six days when both satellites are operational, enabling frequent and systematic observations suitable for long time-series analyses.

Sentinel-1 acquires data through multiple imaging modes that balance spatial resolution and swath coverage, achieving resolutions up to approximately 5 m and swath widths reaching 400 km. The mission follows a pre-programmed acquisition strategy designed to ensure global coverage, conflict-free operations, and the generation of consistent long-term datasets. These characteristics make Sentinel-1 particularly suitable for monitoring surface dynamics such as soil moisture, vegetation structure, land deformation, cryosphere evolution, and marine conditions.

A key feature of Sentinel-1 is its dual-polarisation radar capability, which enables the transmission and reception of microwave signals with different polarisation states. In dual-polarisation acquisitions, the system commonly operates in VV (vertical transmit–vertical receive) and VH (vertical transmit–horizontal receive) configurations. The VV channel primarily captures surface scattering mechanisms and is sensitive to surface roughness and moisture conditions, making it particularly informative for monitoring bare soil, urban areas, and water surfaces. In contrast, the VH channel is more responsive to volume scattering generated by structurally complex targets such as vegetation canopies, where multiple scattering interactions cause depolarisation of the radar signal.

Table 4.1 presents a summary of the Sentinel-1 data channels considered in this study. For each selected band, the historical mean and historical standard deviation are reported in order to characterize the statistical distribution of the observed measurements. The validity of the observations is determined according to predefined band ranges, which identify the interval of physically consistent values for each channel. Measurements falling outside these ranges are classified as non-valid data and are subsequently replaced with the corresponding historical mean value. Additional details regarding the preprocessing methodology are provided in the following chapters.

Table 4.1: S1 Source Channels

Name	Channel	Band Range	Hist Mean	Hist Std	Measure Unit
S1	VV	[0.0, 0.5]	0.0670945	0.1094125	Unitless
S1	VH	[0.0, 0.5]	0.0318255	0.0588595	Unitless

Sentinel 2 Sentinel-2 [23] is a high-resolution, wide-swath multispectral Earth observation mission developed within the European Copernicus Programme . The mission is designed to provide systematic and frequent global observations of the Earth's surface, supporting environmental monitoring and operational services such as land management, agriculture, forestry, disaster response, and risk assessment.

The Sentinel-2 constellation consists of twin satellites operating in the same sun-synchronous orbit with a 180° phase difference, enabling a global revisit time of approximately five days at the Equator. Each satellite is equipped with the Multi-Spectral Instrument (MSI), which acquires imagery across thirteen spectral bands at spatial resolutions of 10 m, 20 m, and 60 m, with a swath width of 290 km. This configuration ensures high spatial detail combined with large-area coverage, making Sentinel-2 particularly suitable for long-term multispectral time-series analysis.

Operating at an altitude of approximately 786 km with a 10-day orbital cycle and a mean local solar time of 10:30 a.m. at the descending node, the mission employs a sun-synchronous orbit to maintain consistent illumination conditions across acquisitions. This orbital design minimizes variations caused by shadows and solar angle changes, thereby ensuring temporal consistency essential for monitoring surface dynamics over time.

The Sentinel-2 mission continues the legacy of earlier multispectral missions such as SPOT and Landsat while ensuring long-term data continuity through successive satellite deployments.

Sentinel-2 multispectral imagery provides observations across thirteen spectral bands spanning the visible, near-infrared (NIR), and shortwave infrared (SWIR) regions of the electromagnetic spectrum. These bands [24] enable detailed characterization of land surface properties, atmospheric conditions, and vegetation dynamics.

- **B01 - Aerosol** Used for aerosol detection and atmospheric correction.
- **B02 - Blue** Supports soil and vegetation discrimination, forest mapping, and identification of anthropogenic features; improves visibility in shadowed areas and shallow water analysis.
- **B03 - Green** Enhances contrast between clear and turbid water and supports vegetation and surface feature detection.
- **B04 - Red** Useful for vegetation type identification, soil characterization, and urban area mapping due to strong interaction with vegetation reflectance properties.
- **B05, B06, B07 - Red Edge Bands** Primarily employed for vegetation classification and monitoring of plant physiological conditions.
- **B08 - Near Infrared** Sensitive to vegetation biomass and canopy structure; widely used for vegetation analysis and shoreline mapping.
- **B8A - Narrow NIR** Supports vegetation classification and condition assessment.

- **B09 - Water Vapour** Used for atmospheric water vapour detection and correction.
- **B10 - Cirrus** Designed for cirrus cloud detection.
- **B11 - SWIR 1** Enables estimation of soil and vegetation moisture content and improves discrimination between vegetation types, snow, and clouds.
- **B12 - SWIR 2** Provides complementary information for moisture monitoring and land surface characterization, particularly for distinguishing snow, clouds, and vegetation conditions.

In this study, only Sentinel-2 bands B03, B04, B08, and B11 are employed for model training, due to their relevance for air quality forecasting applications and used to compute the NDVI (Normalized Difference Vegetation Index). These bands provide complementary information related to vegetation dynamics, surface characteristics, and moisture content, which are closely associated with atmospheric and environmental conditions. Future work may explore the inclusion of additional spectral bands to further enhance model performance where belived beneficial.

Table 4.2 provides a comprehensive overview of the Sentinel-2 spectral channels considered in this study. For each band, the historical mean and historical standard deviation are reported. The specified band ranges define the interval within which observations are regarded as valid. Values falling outside these predefined ranges are treated as non-valid measurements and are subsequently replaced with the corresponding historical mean value. This preprocessing approach was adopted to ensure data consistency and stability throughout the dataset. Further details regarding the adopted preprocessing methodology are discussed in the subsequent chapters.

Table 4.2: Sentinel 2 Source Channels

Channel	Band Range	Hist Mean	Hist Std	Measure Unit
B03	[6.49, 559.01]	416.3233	116.7453	Unitless
B04	[3.31, 484.13]	344.5523	98.6301	Unitless
B08	[0.95, 307.80]	166.1215	93.7028	Unitless
B11	[0.40, 69.78]	47.3316	17.7123	Unitless

Sentinel 3 Sentinel-3 [25]is part of the European Copernicus Programme and is designed to provide accurate and reliable measurements of sea surface topography, sea and land surface temperature, and ocean and land colour. The mission supports a wide range of applications including ocean forecasting, environmental and climate monitoring, vegetation analysis, inland water observation, fire monitoring, and cryosphere assessment. Sentinel-3 ensures continuity with previous Earth observation missions such as ERS and ENVISAT while introducing improved sensor performance and spatial coverage.

The Sentinel-3 constellation consists of two satellites, Sentinel-3A and Sentinel-3B, launched in 2016 and 2018, respectively. The mission carries an advanced payload composed of three primary instruments.

- The ***Sea and Land Surface Temperature Radiometer*(SLSTR)** provides high-accuracy measurements of sea and land surface temperatures through multi-spectral observations in visible, shortwave infrared, and thermal infrared bands, enabling enhanced atmospheric correction and global temperature monitoring.
- The ***Ocean and Land Colour Instrument*** **(OLCI)** acquires multispectral imagery across 21 spectral bands at 300 m spatial resolution, supporting observations of ocean colour, vegetation dynamics, and land surface properties.
- The ***Surface Topography Mission*** **(STM)** includes a Synthetic Aperture Radar Altimeter (SRAL), a Microwave Radiometer (MWR), and a Precise Orbit Determination (POD) system, enabling accurate measurements of sea level, inland water height, and cryospheric parameters.

Sentinel-3 operates in a near-polar, sun-synchronous orbit with an inclination of approximately 98.65° and a descending node crossing time of 10:00 a.m. mean local solar time. This orbital configuration ensures consistent illumination conditions and global coverage, particularly over high-latitude regions.

The Sea and Land Surface Temperature Radiometer (SLSTR) [26] onboard Sentinel-3 is a dual-view, multichannel radiometer designed primarily for accurate retrieval of land and sea surface temperatures. The instrument measures top-of-atmosphere (TOA) radiation across nine spectral bands spanning the visible and near-infrared (VNIR), shortwave infrared (SWIR), and thermal infrared (TIR) regions. In addition to temperature estimation, shorter wavelength channels support atmospheric correction by enabling the detection and characterization of clouds and aerosols, thereby improving geophysical parameter retrieval.

SLSTR data products are generated separately for each viewing geometry and are provided at two spatial resolutions: 500 m for solar reflectance channels (S1-S6) and 1 km for thermal infrared channels (S7-S9 and auxiliary bands). For the purposes of this study, only the SLSTR thermal infrared bands S7, S8, and S9 are utilized, because of they are related to air quality and atmospheric temperature monitoring applications.

Table 4.3 presents a summary of the Sentinel-3 spectral bands employed in this study, together with their corresponding attributes, including historical mean values, historical standard deviations, and measurement units. It should be noted that value ranges for the individual bands are not reported, as a masking procedure was applied to exclude non-valid observations. The presence of this mask filter ensures that only reliable data are retained, thereby eliminating the necessity of defining explicit band ranges.

Sentinel 5p The Copernicus Sentinel-5 Precursor (Sentinel-5P) [27] mission is the first Copernicus initiative dedicated to atmospheric monitoring. It is the result

Table 4.3: Sentinel 3 Source Channels

Channel	Band Range	Hist Mean	Hist Std	Measure Unit
S7	–	287.6223	43.9959	K
S8	–	271.7466	44.0112	K
S9	–	269.7650	43.5205	K

of a collaborative effort among the European Space Agency (ESA), the European Commission, the Netherlands Space Office, industry stakeholders, data users, and the scientific community. The mission employs a single satellite equipped with the TROPOspheric Monitoring Instrument (TROPOMI), co-funded by ESA and the Netherlands, to perform high-resolution measurements of atmospheric constituents. These data support monitoring of air quality, stratospheric ozone, ultraviolet radiation, and climate.

Sentinel-5P was launched on 13 October 2017 from the Plesetsk Cosmodrome in Russia. It operates in a near-polar, sun-synchronous orbit with an ascending node equatorial crossing at 13:30 local solar time. This configuration allows coordination with NASA's Suomi-NPP satellite to utilize co-located cloud mask data from the VIIRS instrument. The orbit has an inclination of approximately 98.7°, a reference altitude of 824 km, and a 16-day repeat cycle with 14 orbits per day. TROPOMI addresses three primary environmental themes: air quality, stratospheric ozone, and climate change monitoring and forecasting. Near the Earth's surface, aerosols, ozone, and reactive gases such as nitrogen dioxide directly affect human health, ecosystem integrity, and the built environment.

As part of the Copernicus Space Component, Sentinel-5P complements Sentinel-4 and Sentinel-5 within a constellation serving Copernicus Atmosphere Monitoring Service (CAMS). Together, these missions provide critical information on atmospheric variables such as ozone, surface UV, air quality, and climate-relevant trace gases. Sentinel-5p [28] deliver key data products including NO_2, SO_2, HCHO, CHOCHO, aerosols (AER_AI_340_380, and AER_AI_354_388), CO, CH_4, and stratospheric O_3 quality parameters, providing daily global coverage for climate, air quality, and ozone/UV applications.

Table 4.4 summarizes the relevant information concerning this data source. Owing to occasional data unavailability caused by satellite measurement errors, primarily resulting from severe cloud coverage, invalid or missing observations were replaced with the historical mean value of the corresponding spectral channel. This imputation strategy was adopted to preserve temporal continuity within the dataset and to avoid disruptions in the sequential frame structure required for subsequent analyses. A detailed discussion of this preprocessing procedure is provided in the following chapters.

Table 4.4: Sentinel 5p Source Channels

Channel	Band Range	Hist Mean	Hist Std	Measure Unit
CO	[0.0, 0.1]	0.0333234	0.0335877	mol/m^2
HCHO	[0.0, 0.001]	0.0001753	0.0002062	mol/m^2
NO2	[0.0, 0.0003]	0.0001058	0.0001213	mol/m^2
O3	[0.0, 0.36]	0.1464774	0.1476350	mol/m^2
SO2	[0.0, 0.01]	0.0008097	0.0012906	mol/m^2
AER_AI_340_380	[-1.0, 5.0]	-0.1960849	0.6126001	mol/m^2
AER_AI_354_388	[-1.0, 5.0]	-0.2090579	0.5606826	mol/m^2

4.3 Weather Source

Open-Meteo [29] is a weather forecasting service that leverages open-data numerical weather prediction outputs provided by national meteorological agencies. These institutions distribute numerical weather prediction models that are freely accessible; however, effective utilization of such datasets typically requires specialized expertise in binary data formats, spatial grid systems, map projections, and the theoretical foundations of atmospheric modeling.

In contrast to many proprietary weather APIs, Open-Meteo ensures full transparency by providing open access to its source code and explicitly documenting all underlying data sources, while appropriately acknowledging the contributions of national weather services. The API is available free of charge for non-commercial applications. Despite its cost-free availability, the platform maintains a high level of forecast accuracy through the integration of multiple regional and global meteorological models updated at high temporal frequency, thereby enabling precise location-specific predictions worldwide. The service provides hourly forecasts, medium-range predictions extending up to 16 days, and access to approximately 80 years of historical weather observations. Its low-latency architecture and Cross-Origin Resource Sharing (CORS) support make Open-Meteo particularly suitable for data-driven platforms, including agricultural decision-support systems that require rapid access to environmental information.

The present study concentrates on historical meteorological observations related to the city of Milan. The analysis incorporates the variables summarized in Table 4.5, which reports the historical mean values, historical standard deviations, and corresponding measurement units for each considered parameter. The absence of predefined value ranges for the selected variables is due to the high reliability and precision of the observational datasets, both in terms of measurement accuracy and data completeness. Consequently, no preliminary data cleaning or range-based filtering procedures were deemed necessary prior to the analytical process. Furthermore, due to the lack of a reliable and freely accessible meteorological forecasting provider offering continuous coverage for the city of Milan across the entire temporal range considered in this study, Open-Meteo data were adopted as the forecasting reference for the period between 2018 and 2022. For the remaining years included in the

analysis, Open-Meteo provides direct historical coverage, thereby ensuring dataset consistency and temporal continuity throughout the study period.

Table 4.5: Open-Meteo Source Channels

Channel	Band Range	Hist Mean	Hist Std	Measure Unit
temperature_2m_mean	–	14.2958	8.0316	°C
relative_humidity_2m_mean	–	73.8182	14.2194	%
dew_point_2m_mean	–	8.9964	7.1274	°C
precipitation_sum	–	3.4904	7.7766	mm
surface_pressure_mean	–	999.8788	7.4880	hPa
wind_speed_10m_mean	–	7.6151	3.3577	km/h
cloud_cover_mean	–	54.0178	30.3061	%
shortwave_radiation_integral	–	13.8941	7.8626	W/m^2

4.4 Topological Sources

In this study, topological data are employed as static sources to provide information on the terrain surrounding monitoring stations, allowing the analysis of the relationship between terrain characteristics and pollutant concentrations. Two types of topological data are utilized.

First, the Digital Elevation Model (DEM) provides information on terrain morphology, offering a constant observation throughout the reference period. For each monitoring station, altitude is extracted at three scales: precise point altitude, altitude within a 100-meter radius, and altitude within a 1-kilometer radius.

Second, land cover and land use data are obtained from the Copernicus Urban Atlas 2018 [30], which provides high-resolution, inter-comparable maps for urban areas. The original 27 land cover classes are aggregated into 10 broader categories, including urban, roads, railways, ports, airports, extraction areas, non-utilized land, green areas, open spaces, and water bodies. The Urban Atlas is a joint initiative of the European Commission Directorate-General for Regional and Urban Policy with support from the European Space Agency and the European Environment Agency.

Chapter 5

Methodology

The difficulty of training video transformers from scratch is amplified by the limited size of available video datasets when compared with image datasets. Data scarcity constrains the ability of transformer models to learn robust spatio-temporal representations without external supervision. The challenge is particularly pronounced in the context of remote sensing applications for air quality analysis. In this domain, large-scale pre-trained architectures tailored to satellite observations are not available, and datasets that integrate multiple sources of environmental information remain limited. Satellite observations, meteorological measurements, and geospatial descriptors are typically collected by different systems and often lack unified datasets suitable for large-scale supervised training. The Video Masked Autoencoder (VideoMAE) [20] architecture addresses these limitations through a self-supervised video pre-training strategy.

The present work adapts the VideoMAE architecture, originally designed for natural video, to the analysis of satellite image time series. In this formulation, sequences of satellite observations are interpreted as videos in which each frame captures a snapshot of atmospheric and surface conditions at a specific time. The architecture is employed as the backbone of a self-supervised pre-training stage that integrates multiple heterogeneous data sources. Satellite imagery is combined with meteorological variables, land cover information, and digital elevation models through a fusion strategy incorporated into the reconstruction objective.

Two considerations motivate the application of the VideoMAE framework to satellite-based air quality analysis.

- First, satellite image time series can be naturally interpreted as short video sequences, where each acquisition corresponds to a frame that captures the state of the atmosphere at a specific time.
- Second, the limited number of satellite acquisitions available for a given region and time period constrains the use of conventional transformer training strategies, which typically require very large datasets. The self-supervised formulation of VideoMAE alleviates this limitation.

The encoder learned representations are subsequently used in a downstream

prediction task, where they are combined with weather forecasting data to estimate future air quality conditions. The predictive model generates a sequence of seven station level predictions corresponding to the following seven days. Each value represents the predicted ground-level concentrations of five atmospheric pollutants over the study area. Supervision is provided by measurements collected from ground monitoring stations, which record local pollutant concentrations and enable the model to align its spatial predictions with observed air quality data.

5.1 Data Alignment

When multiple data sources are combined within a single learning framework, the alignment of the corresponding modalities becomes a critical requirement for effective model training. Multimodal alignment refers to the process of establishing consistent semantic relationships between heterogeneous data representations. In these settings, alignment is typically achieved by mapping different modalities into a shared representation space in which their similarity can be measured while accounting for potential ambiguities and long-range dependencies.

The present work integrates several data sources characterized by heterogeneous spatial resolutions and temporal sampling rates. These differences require a preprocessing stage in which the data are transformed into a common reference grid. All modalities are therefore resampled to match the spatial resolution of the Sentinel-5P images after the preprocessing step, resulting in images of size 336×384 pixels. Temporal alignment is performed by enforcing a uniform temporal granularity of one observation per day. After this harmonization step, the dataset can be organized into samples defined by a past observation window and a future forecasting horizon.

Once the data have been aligned spatially and temporally, the processing pipeline follows two distinct paths depending on the learning stage. The first path corresponds to the pre-training phase, where the objective is to learn a representation of the environmental dynamics within the observed region. In this phase, only the past scope is considered. Satellite imagery, meteorological records, land cover information, and the digital elevation model are integrated and provided to the self-supervised pre-training procedure.

The second path corresponds to the downstream prediction task. In this stage, the previously described data sources representing the past context are combined with additional information describing the future conditions. This future context is provided through weather forecast records, which supply estimates of the expected meteorological evolution.

5.2 Data Preprocessing

The integration of heterogeneous data sources is fundamental to the predictive performance of the proposed air quality forecasting model. These sources are categorized into four distinct modalities: satellite imagery, meteorological data, ground-station

recordings, and static topographic data, including Land Cover and Digital Elevation Models (DEM). Each modality requires specific transformations to ensure structural compatibility and numerical stability. While every sources maintain a daily temporal granularity, the feature sets for each task differ; the pretraining phase leverages the full suite of spatial and meteorological data, whereas the forecasting task incorporates these alongside future meteorological projections and ground-truth station observations.

Satellite data are represented as a spatial grid covering the Milan metropolitan area at a resolution of 50×50 meters.To fit the model requirements all the sources are padded to be aligned with the standard. These data frequently exhibit issues such as missing values or noise due to sensor malfunction and cloud interference. To mitigate the impact of these anomalies, the Copernicus project's validity metadata is used to construct a binary masking map for each spectral channel. This mask identifies unreliable pixels, allowing the model to distinguish between genuine signals and imputed values during the reconstruction phase. Any identified out-of-range values are replaced with channel-specific historical means, a necessary step as the model architecture requires complete input tensors without null entries.

This dataset is subdivided into historical observations, which represent measured ground truth, and historical forecasts, which represent the predicted atmospheric states available at that time. These features are provided in a tabular format with a single daily value for the study area. Similarly, topographic data from Land Cover and DEM sources serve as static spatial descriptors. These provide context regarding the urban and geographical characteristics surrounding monitoring sites and are integrated into the model without cleaning.

The target variable for the forecasting task is derived from ground-station recordings. These are mapped onto a grid of the same resolution as the satellite data, though the resulting tensors are inherently sparse; only pixels corresponding to the physical coordinates of a station contain valid pollutant concentrations. It is important to note that the availability of these labels varies, as not every station monitors every pollutant on a daily basis. To ensure these disparate scales and distributions are suitable for the task, a rigorous normalization protocol is enforced.

Standardization is achieved via Z-score normalization, where the historical mean and standard deviation are calculated exclusively from valid, non-masked data points for each channel. A specialized transformation is reserved for ground pollutant concentrations due to their characteristic distribution, which often features rare but significant peaks. These values undergo a logarithmic scaling prior to Z-score normalization. This dual-stage transformation squashes the dynamic range of the target variable, preventing high-concentration outliers from dominating the loss function and degrading the overall predictive accuracy of the model. Detailed information about historical means o standard deviations is provided in the dataset chapter.

5.3 Pre-Training

The pre-training stage aims to learn a representation of the atmospheric conditions over the study area using historical satellite observations. For each training sample, a sequence of sixteen consecutive days of satellite imagery is considered as input. Representation learning is formulated as a reconstruction task within an asymmetric encoder-decoder architecture derived from the ImageMAE [31] framework originally proposed for images.

The input time series is first partitioned into non-overlapping spatial patches of size p_1^2, producing a regular spatial grid that is consistent across all frames. After spatial partitioning, consecutive frames are grouped into temporal segments composed of n frames. The spatial grid is then extended along the temporal dimension, in this way is possible to extract 3D cubes with dimensions $p_1^2 \times n$. These cubes are referred to as tublets, and the parameter n is defined as the tublet size. This procedure produces a sequence of joint spatio-temporal embeddings that represent localized atmospheric dynamics. The resulting representation reduces the effective spatial and temporal resolution of the input, which makes the application of the attention mechanism computationally feasible.

Temporal downsampling is introduced through the choice of the tublet size. Following the configuration adopted in the VideoMAE framework, the present work groups frames using a tublet size of two. This design captures short-term temporal variations while reducing the sensitivity of the model to missing information introduced during the preprocessing stage. In satellite time series, missing observations may occur due to factors such as cloud coverage or acquisition gaps. The aggregation of consecutive frames partially mitigates the influence of these missing values by distributing the temporal information across each tublet.

The masking strategy plays a central role in the learning process because the model operates on video-like sequences and naive masking scheme applied independently to each frame could allow the model to infer masked information by exploiting correlations across neighboring frames. To avoid this form of information leakage, the masking procedure follows the temporal tube masking strategy proposed in VideoMAE. Under this scheme, if a spatial patch is masked in one frame, the corresponding patches at the same spatial location in all other frames are also masked. The resulting masks form tubes that extend across the entire temporal dimension, which forces the model to infer missing content from the surrounding spatial and temporal context rather than from direct temporal alignment. The reconstruction task is further increased in difficulty by adopting a masking ratio close to ninety percent, as recommended in the original VideoMAE formulation. Its important to underline that the pretraining phase needs to handle masking at two levels: a pixel level drived by the specific validity range and the patch level masking drived from the temporal tube masking. The masking strategy takes into account these constraints producing a masking patter that preserve the desired signal rateo.

The encoder follows the standard Vision Transformer backbone and applies a

joint space-time attention mechanism that allows each token to interact with every other token within the sequence through multi-head self-attention. This interaction enables the model to capture high-level spatio-temporal dependencies within the remaining visible regions of the input.

It is also important to note that spatial and temporal structures are handled at two distinct stages of the representation pipeline. The first stage corresponds to the construction of tublets, where spatial patches are aggregated across short temporal intervals. The second stage occurs within the transformer encoder, where the joint space-time attention mechanism models relationships between tublets across the entire sequence.

Multimodal fusion Multimodal learning integrates heterogeneous sources of information in order to exploit complementary signals that are not available within a single modality. Each modality captures a specific aspect of the physical processes that influence atmospheric composition and their integration allows the model to combine complementary environmental descriptors and reduces the limitations associated with relying on a single data source.

In the proposed framework, multimodal fusion is performed directly at the input level before the encoder stage. After spatial and temporal alignment, the different modalities are merged through a channel-wise concatenation operation. This operation produces a unified tensor in which all modalities share the same spatial grid and temporal structure. Each frame of the resulting sequence therefore contains the combined representation of satellite observations, meteorological variables, land cover information, and elevation data.

The encoder processes this fused representation as a single spatio-temporal signal. To maintain an order in this 3D context, a sinusoidal positional encoding distinguish between the tokens both spatially and temporally. By operating on concatenated inputs and drived by the positional encoding, the model is allowed to learn joint representations that capture interactions between the sources.

5.4 Downstream Task: Multimodal Spatiotemporal Forecasting of Air Pollution

The downstream task addressed in this work consists of forecasting atmospheric pollutant concentrations over Milan Metropolitan Area for the next K days. The prediction is performed using a multimodal transformer architecture that integrates information from past satellite observations and numerical weather forecasts. The model follows an encoder–decoder paradigm and is designed to learn complex spatiotemporal dependencies governing pollutant dispersion and accumulation. The custom architecture is presented in Figure5.1

Problem Formulation Let $X \in \mathbb{R}^{C \times T_{past} \times H \times W}$ denote a sequence of satellite observations over a spatial grid, C the number of channels, T_{past} the number of past

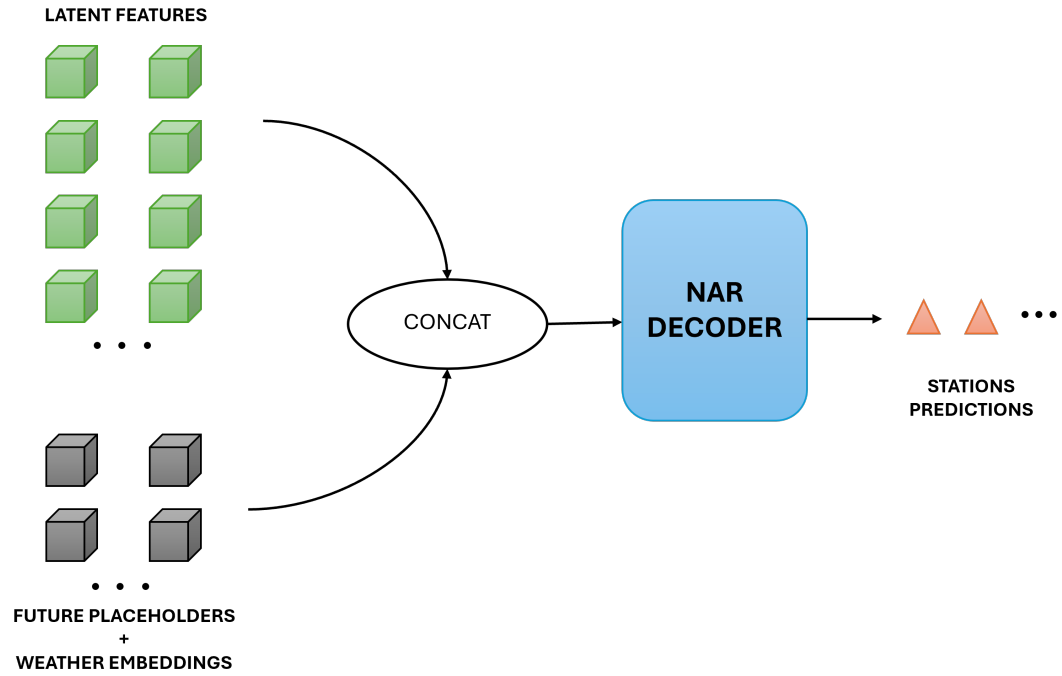

Figure 5.1: Downstream custom architecture

time steps, and $H \times W$ the spatial resolution. The objective is to predict future pollutant concentration maps over a prediction horizon of $T_{future} = K$ days.

Additionally, let $W_f \in \mathbb{R}^{T_{future} \times F}$ denote the meteorological forecast variables available for the future time horizon, where F is the number of weather features.

The learning objective can therefore be expressed as:

$$f(X_{t-T_{past}:t}, W_{t:t+T_{future}}) \rightarrow Y_{t+1:t+T_{future}} \tag{5.1}$$

where Y represents the future pollutant concentration maps over the spatial grid.

The pretrained encoder processes the all tokens extracted from past satellite imagery and produces latent embeddings representing the observed spatiotemporal dynamics:

$$Z = \text{Encoder}(X) \tag{5.2}$$

To retain spatial and temporal structure, the model uses factorized positional embeddings consisting of a spatial component and a temporal component. For a token located at spatial patch p and time index t, the positional encoding is defined as

$$PE(p,t) = PE_{space}(p) + PE_{time}(t) \tag{5.3}$$

This formulation allows the model to separately capture spatial layout and temporal ordering while maintaining a compact representation. Despite the single positional encoding of the Pretaining task, here a dimension-specific positional encoding is necessary to order the tokens of both the past and the future context.

Thanks to this approach, each token can be precisely identified in a spatio-temporal scope.

Future Token Construction and Weather Conditioning Future pollutant states are the prediction target, therefore, the model introduces a set of learnable mask tokens representing future spatiotemporal locations that must be predicted by the decoder.

For each future time step $t \in [1, T_{future}]$ and spatial patch p, a token is initialized using a shared learnable mask embedding. Spatial and temporal positional encodings are then added to these tokens to provide structural context.

Meteorological forecast information is incorporated through a dedicated weather embedding layer represented by a lightweight multi-layer perceptron:

$$E_w = \text{WeatherEmbedding}(W_f) \tag{5.4}$$

The resulting weather representation is broadcast across spatial locations and added to the corresponding future tokens:

$$T_{future}(t, p) = T_{mask} + PE_{space}(p) + PE_{time}(t) + E_w(t) \tag{5.5}$$

This mechanism conditions future predictions on expected atmospheric dynamics such as wind speed, temperature, or humidity.

Multimodal Fusion and Decoder Prediction The final input sequence to the decoder is constructed by concatenating the encoded past observations with the future tokens:

$$Z = [Z_{vis}, T_{future}] \tag{5.6}$$

A transformer decoder processes this joint representation and predicts pollutant concentration values for all future tokens simultaneously.

The decoder operates in a non-autoregressive manner, meaning that all future time steps are predicted in parallel. To preserve temporal causality, an attention mask is applied so that tokens corresponding to earlier prediction days cannot attend to tokens representing later days. This constraint ensures that the prediction for day t only depends on past observations and previous forecast steps, preventing information leakage from future predictions.

Training Objective Ground-truth pollutant measurements are obtained from monitoring stations, which provide accurate observations but are spatially sparse. Since the transformer model predicts dense pollutant concentration maps over the entire spatial grid, a direct supervision using only station pixels would result in extremely sparse training signals. To address this limitation, a label densification strategy based on Gaussian spatial propagation is employed.

The measurements recorded at station locations are spatially propagated to nearby pixels using a Gaussian kernel. This procedure generates smooth local concentration fields while preserving the original station values. This formulation produces two outputs:

- a **densified concentration map** Y_{dense} obtained through Gaussian interpolation
- a **support map** which represents the amount of nearby station support for each pixel.

Pixels located close to monitoring stations receive higher support values, while pixels far from stations retain low confidence or remain undefined. This map therefore provides a natural weighting mechanism reflecting the reliability of the interpolated supervision.

To account for varying reliability across spatial locations, the model is trained using a confidence-weighted mean squared error loss.

The final training objective is defined as

$$\mathcal{L} = \frac{\sum_i C_i(\hat{y}_i - y_i)^2}{\sum_i C_i} \tag{5.7}$$

where $\hat{y}_i$ is the predicted concentration, y_i is the densified ground truth, and C_i is the confidence value associated with pixel i.

Prediction Output The decoder outputs a sequence of tokens corresponding to the predicted pollutant concentration values for each spatial patch and future time step. These predictions are then reshaped back into spatial maps to produce the final pollution forecasts over the study region.

This framework enables the model to jointly reason over satellite-derived environmental observations and meteorological forecasts, capturing the evolution of atmospheric pollutants.

Chapter 6

Results

This section presents the empirical evaluation of the proposed multimodal transformer for air quality forecasting. The analysis focuses on an ablation study designed to quantify the contribution of each data modality and to examine the interactions between modalities within the model. Model behaviour is evaluated through variations in predictive performance and by comparing the relative importance of the available inputs.

The proposed architecture predicts ground-level concentrations of five atmospheric pollutants by integrating heterogeneous sources of information, including satellite imagery, meteorological tabular data, land cover maps, and digital elevation model data. This study adopts a comparative framework in which multiple input modality configurations are evaluated, while Sentinel-5P observations and meteorological measurements are consistently retained as the core input components. The model produces that estimate pollutant concentration across the study area. These predictions are evaluated against ground-truth observations derived from monitoring stations.

The experimental dataset spans the period from 2018 to 2024. Data collected in 2022 are used for validation, while 2023 is reserved for testing. All remaining years are allocated to the training split. This temporal partitioning prevents information leakage across time and allows the evaluation to reflect the model's ability to generalize to unseen temporal conditions.

The results indicate that the model effectively captures large-scale dependencies and temporal trends. In particular, the architecture demonstrates the ability to reproduce their temporal evolution within a forecasting horizon of seven days.

6.1 Experimental Setup

Pretraining configuration The pretraining stage processes a frames time series. Each frame is partitioned into non-overlapping patches of size 24×24, resembling the VideoMAE frame to patches proportion, which are aggregated into temporal tokens spanning 2 consecutive frames. Each training sample contains a temporal context of 16 frames.

The encoder embedding dimension is set to 768 with 12 attention heads and a feed-forward expansion ratio of 4, following the VideoMAE architectural settings. The decoder embedding dimension is 512 with 8 attention heads.

Training is performed with a learning rate of 10^{-4}. The temporal forecasting horizon considered in the dataset preparation is 7 days, while the temporal context length used during pretraining is 16 frames. The complete list of architectural and training parameters used during pretraining is reported in TableA.1.

Downstream configuration In the downstream configuration, the encoder representation is processed by a non-autoregressive (NAR) decoder composed of 2 layers, which is responsible for producing pollutant forecasts. The decoder operates on the latent features and generates predictions corresponding to the monitoring stations.

The model is trained with a learning rate of 10^{-4}. The output consists of predicted pollutant concentration trends over the monitoring stations for a forecasting horizon of 7 days, using the preceding 16 temporal observations as context. A detailed description of the downstream model configuration is provided in TableA.1.

Evaluation metrics Model performance is evaluated using error metrics computed with respect to the station-based supervision signal augmented through the Gaussian blob representation. A confidence map is used to weight prediction errors according to spatial proximity to monitoring stations. Pixels located closer to a station are assigned higher confidence values, while distant pixels receive lower weights. This weighting scheme reflects the spatial reliability of the ground-truth information.

The primary evaluation metric used throughout the ablation analysis is the Mean Absolute Error (MAE), computed as the average absolute deviation between predicted and reference values across the entire test set. MAE is widely used in air quality forecasting because it provides a direct and interpretable measure of prediction error. Additional metrics are reported to provide a more detailed characterization of model performance. These include pollutant-specific MAE values, day-wise MAE across the forecasting horizon, and the corresponding Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). Together, these measures provide a structured view of the model behaviour across pollutants, and forecast lead times.

6.2 Quantitative Analysis

Table6.1 reports the Mean Absolute Error (MAE) obtained by the downstream models for each pollutant and averaged across all pollutants. The comparison summarizes the results of the ablation study, where different combinations of input modalities are progressively integrated into the forecasting architecture. The baseline configuration, denoted as S5_OM, relies exclusively on Sentinel-5P atmospheric observations and meteorological variables. This configuration achieves the lowest overall MAE among the evaluated models, with an average error of 9.47. The result indicates that the

atmospheric measurements combined with weather information provide a strong predictive signal for short-term pollutant forecasting.

The introduction of additional modalities produces heterogeneous effects across pollutants. When Sentinel-1 radar observations are incorporated, the overall MAE increases slightly to 9.67. Despite this marginal degradation in global performance, a reduction in the O_3 error is observed, suggesting that radar-derived surface information may provide complementary cues for ozone estimation. The integration of Sentinel-2 infrared imagery leads to a larger increase in the aggregated MAE, primarily due to higher errors in O_3, PM_{10}, and $PM_{2.5}$ predictions. A similar trend appears when Sentinel-3 observations are introduced , where the overall MAE rises to 9.91 despite a slight reduction in SO_2 error.

Static geographical descriptors, represented by land cover and digital elevation model data , produce moderate changes in performance. The overall MAE increases to 9.77, indicating that the inclusion of static spatial context does not systematically improve predictive accuracy within the evaluated configuration. However, these variables appear to stabilize the predictions for some pollutants, particularly PM_{10} and $PM_{2.5}$, whose errors remain comparable to those of the baseline.

Models that combine multiple satellite modalities show a further increase in prediction error. The most complex configuration, which integrates all available modalities, records the highest aggregated MAE of 10.75.

The inclusion of additional data sources does not systematically degrade model performance. In particular, the configuration combining the baseline inputs with Sentinel-1 observations and static geographical descriptors, namely land cover and digital elevation model data, achieves the lowest error for O_3 and SO_2 while maintaining one of the best overall MAE scores across all evaluated configurations.

These results suggest that increasing the number of input modalities does not necessarily lead to improved predictive performance. The baseline configuration already captures the dominant atmospheric patterns governing pollutant concentration, while additional modalities may introduce redundant or weakly correlated signals that complicate the learning process. The ablation analysis therefore highlights the central role of Sentinel-5P atmospheric measurements and meteorological variables in the proposed forecasting framework, while the contribution of auxiliary modalities appears to be pollutant-dependent and less consistent across configurations. Similar results are captured by the RMSE analysis which are summarized by Table6.5. Each table's row represent a model that uses the checkmarked sources, where S5 stands for Sentinal 5p. Following this pattern all the satellite sources are compared along with the Land cover and Digital Elevation Model which are abbreviated with 'L&D'.

The daily MAE values reported in Table6.2 reveal a consistent temporal pattern across most model configurations. Prediction errors are generally higher for the first forecasting day and decrease during the following days, typically reaching their minimum between days three and five. This behaviour suggests that the model benefits from the temporal structure imposed by the forecasting decoder, which allows intermediate predictions to stabilize as the temporal context is processed. After the

Table 6.1: MAE performance metrics per pollutant.

S5	OM	S1	S2	S3	L&D	NO2	O3	PM10	PM25	SO2	All
✓	✓					**18.05**	17.57	**9.01**	**6.69**	2.24	**9.47**
✓	✓	✓				19.36	16.14	9.21	6.83	2.24	9.67
✓	✓		✓			18.50	20.47	9.60	7.30	2.13	10.11
✓	✓			✓		19.55	17.10	9.48	7.11	1.98	9.91
✓	✓				✓	18.92	18.77	9.25	6.80	1.98	9.77
✓	✓	✓	✓			18.55	15.93	9.99	7.62	2.37	10.23
✓	✓	✓		✓		19.67	17.57	9.46	7.21	2.07	9.98
✓	✓		✓	✓		18.97	18.59	9.76	7.14	2.15	10.10
✓	✓	✓			✓	18.49	**15.60**	9.29	6.97	**1.97**	9.65
✓	✓	✓	✓	✓		19.73	19.88	9.72	7.06	2.12	10.20
✓	✓		✓		✓	18.47	17.93	9.47	7.09	2.18	9.89
✓	✓			✓	✓	18.96	16.29	9.40	6.99	2.27	9.78
✓	✓	✓	✓		✓	18.36	20.40	9.58	7.09	2.39	10.00
✓	✓	✓		✓	✓	19.99	19.19	9.02	6.77	2.08	9.67
✓	✓		✓	✓	✓	18.74	17.75	9.82	7.22	2.08	10.08
✓	✓	✓	✓	✓	✓	19.58	21.78	10.40	7.65	2.21	10.75

mid-horizon period, the error tends to remain stable or increase slightly, indicating that the predictive uncertainty grows again toward the end of the forecasting window. The same trend is observed by Table6.4 and Table6.6

Table6.3 reports the MAPE values computed per pollutant and averaged across the dataset. These results highlight substantial variability in predictive accuracy depending on the pollutant type. The largest relative errors are observed for SO_2, which consistently shows the highest MAPE values across all model configurations. This behaviour is likely related to the lower concentration levels and higher temporal variability typically associated with SO_2, which amplify the percentage-based error metric. The observed MAPE values are consistent with those reported in others studies on air quality forecasting, indicating that the predictive performance achieved by the proposed model falls within the typical error range documented in the literature.

6.3 Qualitative Analysis

A qualitative evaluation was conducted to complement the quantitative metrics by examining the temporal behaviour of the predicted pollutant concentrations. For each pollutant channel (SO_2, NO_2, O_3, PM_{10}, and $PM_{2.5}$), the predicted values were compared with the corresponding ground-truth time series derived from monitoring stations. The comparison is illustrated through five plots, each reporting the evolution of the predicted signal alongside the reference measurements.

The visual inspection of these plots indicates that the model is able to capture the overall temporal dynamics of pollutant concentrations. In particular, the predictions follow the general trend observed in the ground truth, reproducing both the rising and decreasing phases of the signal across the forecasting horizon. The predicted curves also reflect the main shape variations present in the observed data, including

Table 6.2: Daily MAE performance.

S5	OM	S1	S2	S3	L&D	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
✓	✓					11.37	11.08	11.08	11.08	**11.05**	11.12	11.13
✓	✓	✓				11.63	11.11	10.97	10.86	**10.80**	10.81	10.89
✓	✓		✓			12.86	11.97	11.75	**11.69**	11.75	11.93	12.04
✓	✓			✓		12.10	11.36	11.05	11.03	**11.01**	11.02	11.10
✓	✓				✓	12.24	11.81	11.40	11.46	11.23	**11.19**	11.20
✓	✓	✓	✓			12.47	11.59	11.13	10.85	10.82	**10.74**	10.77
✓	✓	✓		✓		12.65	11.85	11.40	11.18	**10.99**	11.01	11.10
✓	✓		✓	✓		12.62	11.80	11.49	11.43	11.22	**11.20**	11.24
✓	✓	✓			✓	11.79	10.83	10.50	**10.41**	10.42	10.54	10.55
✓	✓	✓	✓	✓		13.65	12.43	11.92	11.62	11.44	11.23	**11.22**
✓	✓		✓		✓	12.55	11.66	11.20	11.14	10.96	**10.95**	11.14
✓	✓			✓	✓	12.07	11.24	10.90	10.81	10.80	**10.74**	10.90
✓	✓	✓	✓		✓	12.78	12.13	11.70	11.66	11.55	11.56	**11.51**
✓	✓	✓		✓	✓	13.14	12.16	11.55	11.39	11.01	11.00	**10.89**
✓	✓		✓	✓	✓	12.22	11.58	11.18	11.16	**11.11**	11.23	11.27
✓	✓	✓	✓	✓	✓	13.53	12.93	12.47	12.37	12.15	12.03	**11.97**

several local fluctuations and peak events. This behaviour suggests that the model effectively exploits the temporal context encoded in the input modalities and learns a representation that reflects the large-scale dynamics of atmospheric pollution.

Despite this capability, the amplitude of the predicted peaks is typically lower than the corresponding ground-truth values. The model therefore tends to produce more conservative estimates, attenuating extreme variations in concentration levels. This behaviour is consistent across the five pollutant channels and reflects a smoothing effect introduced by the forecasting architecture and by the supervision mechanism based on Gaussian blobs.

A different behaviour emerges when analysing the spatial structure of the predicted maps. Although the model produces dense outputs covering the entire study area, the spatial variability of the predictions remains limited. Instead of reconstructing distinct spatial patterns that reflect the local distribution of pollutant sources, the model tends to reproduce a repeated patch-level structure across the output grid. In practice, the predicted maps exhibit a recurrent spatial pattern whose values vary in magnitude over time following the predicted temporal trend.

As a consequence, the model is able to encode the temporal evolution of pollutant concentrations but struggles to reconstruct meaningful spatial variability across patches. The spatial information therefore remains weakly expressed in the generated maps, which limits the ability of the architecture to produce spatially informative forecasts. These observations highlight a gap between the temporal predictive capabilities of the model and its ability to generalize spatial representations within the patch-based decoding framework. The analysis over the the test set trends computed as a mean of the station's recorded concentration per-pollulant is presented in those Figures 6.1, 6.2, 6.3, 6.4, 6.5.

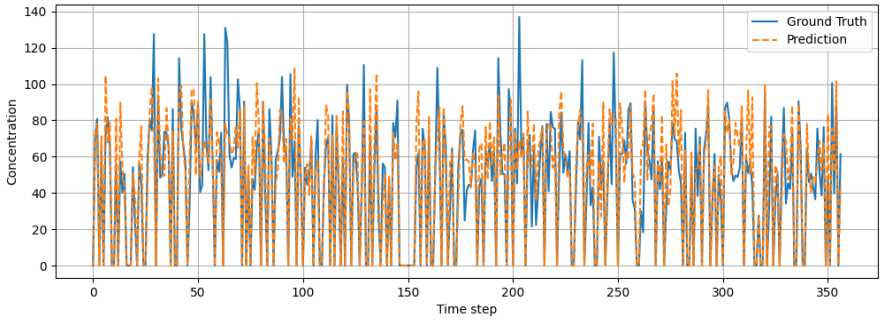

Figure 6.1: NO_2 concentration's trend over the test set

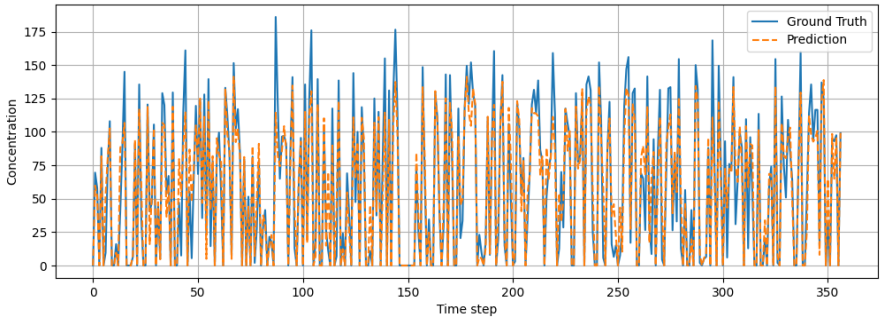

Figure 6.2: O_3 concentration's trend over the test set

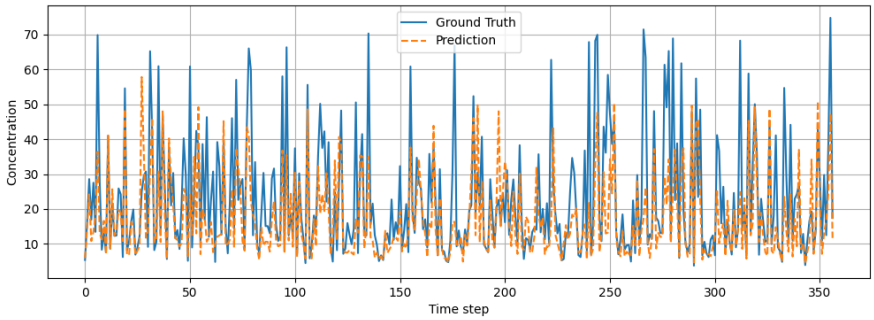

Figure 6.3: PM_{10} concentration's trend over the test set

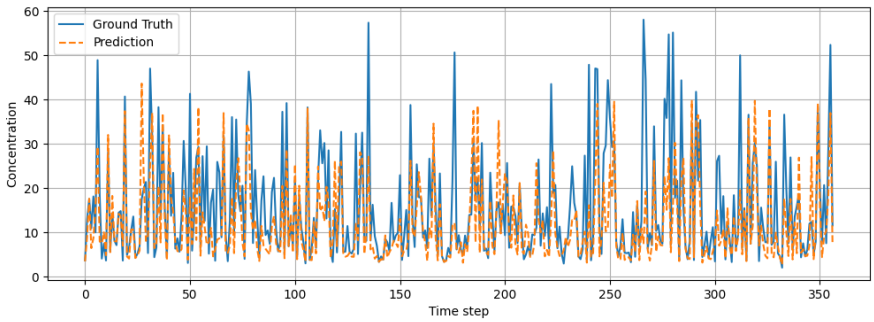

Figure 6.4: $PM_{2.5}$ concentration's trend over the test set

Table 6.3: MAPE performance metrics per pollutant.

S5	OM	S1	S2	S3	L&D	NO2	O3	PM10	PM25	SO2	All
✓	✓					33.66	54.28	41.95	51.58	79.40	52.17
✓	✓	✓				33.75	48.47	42.47	50.36	78.57	50.73
✓	✓		✓			32.24	61.45	45.13	56.37	73.96	53.83
✓	✓			✓		33.89	**40.77**	42.25	49.00	67.90	**46.76**
✓	✓				✓	34.86	48.89	42.31	49.47	66.19	48.35
✓	✓	✓	✓			31.56	46.10	44.93	53.60	85.29	52.30
✓	✓	✓		✓		33.55	45.35	43.42	51.71	71.10	49.03
✓	✓		✓	✓		31.34	51.52	42.02	**47.60**	75.60	49.61
✓	✓	✓			✓	32.67	44.32	45.44	56.21	**64.46**	48.62
✓	✓	✓	✓	✓		33.23	51.61	43.66	50.01	72.60	50.22
✓	✓		✓		✓	33.23	43.64	42.22	52.59	76.18	49.57
✓	✓			✓	✓	33.74	48.08	42.19	50.32	82.26	51.32
✓	✓	✓	✓		✓	33.07	47.26	42.40	50.42	86.59	51.95
✓	✓	✓		✓	✓	32.47	46.06	**39.75**	47.63	70.92	47.37
✓	✓		✓	✓	✓	**30.89**	57.04	42.01	47.89	70.93	49.75
✓	✓	✓	✓	✓	✓	33.25	51.80	43.49	50.83	73.63	50.60

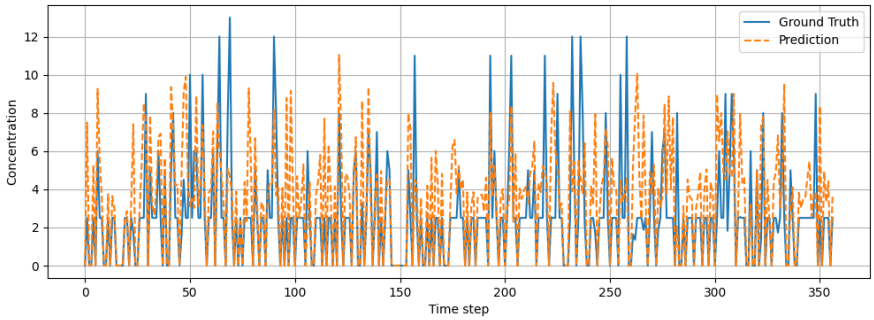

Figure 6.5: SO_2 concentration's trend over the test set

Table 6.4: Daily average MAPE performance per configuration.

S5	OM	S1	S2	S3	L&D	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
✓	✓					50.06	50.83	51.28	52.06	53.11	53.80	54.07
✓	✓	✓				47.38	49.37	50.14	50.89	51.84	52.52	52.93
✓	✓		✓			51.55	51.64	52.65	53.12	54.97	55.86	57.00
✓	✓			✓		**45.50**	**45.61**	**45.60**	**45.97**	**46.96**	48.31	49.38
✓	✓				✓	48.36	48.29	47.71	47.86	48.06	48.74	49.40
✓	✓	✓	✓			51.50	52.17	52.06	52.05	52.70	52.77	52.82
✓	✓	✓		✓		48.42	48.58	48.37	48.47	48.92	49.71	50.70
✓	✓		✓	✓		48.13	48.16	48.63	49.21	50.22	51.10	51.84
✓	✓	✓			✓	47.83	47.72	48.11	48.01	48.78	49.74	50.14
✓	✓	✓	✓	✓		51.40	50.83	49.92	49.93	49.51	50.16	49.79
✓	✓		✓		✓	48.28	48.25	48.58	48.84	49.89	50.93	52.23
✓	✓			✓	✓	49.47	49.33	50.30	50.54	52.55	52.72	54.34
✓	✓	✓	✓		✓	51.95	51.49	50.93	51.50	52.36	52.84	52.57
✓	✓	✓		✓	✓	46.11	47.13	47.58	47.60	47.44	**47.79**	**47.93**
✓	✓		✓	✓	✓	47.01	48.36	48.78	49.39	50.70	51.68	52.32
✓	✓	✓	✓	✓	✓	50.60	50.30	49.79	50.69	50.47	51.21	51.12

Table 6.5: RMSE performance metrics per pollutant.

S5	OM	S1	S2	S3	L&D	NO2	O3	PM10	PM25	SO2	All
✓	✓					**25.07**	24.08	**14.61**	**10.62**	3.33	17.58
✓	✓	✓				26.35	20.76	14.89	10.79	3.47	17.18
✓	✓		✓			25.73	28.47	15.44	11.39	3.47	19.25
✓	✓			✓		26.37	21.86	15.47	11.36	3.11	17.62
✓	✓				✓	26.53	24.48	14.97	10.77	3.17	18.18
✓	✓	✓	✓			25.87	20.72	15.89	11.94	3.62	17.36
✓	✓	✓		✓		26.69	22.72	15.39	11.45	3.29	17.93
✓	✓		✓	✓		26.23	23.45	15.89	11.42	3.41	18.07
✓	✓	✓			✓	25.39	**20.14**	14.85	10.79	**3.06**	**16.71**
✓	✓	✓	✓	✓		26.67	25.54	15.57	11.19	3.42	18.67
✓	✓		✓		✓	25.93	23.07	15.33	11.15	3.53	17.76
✓	✓			✓	✓	25.89	21.12	15.21	11.09	3.61	17.23
✓	✓	✓	✓		✓	25.52	25.77	15.46	11.31	3.85	18.42
✓	✓	✓		✓	✓	26.90	24.03	14.94	10.88	3.41	18.19
✓	✓		✓	✓	✓	25.67	23.77	16.06	11.65	3.26	18.04
✓	✓	✓	✓	✓	✓	26.34	26.86	16.79	12.18	3.50	19.27

Table 6.6: Daily average RMSE performance per configuration.

S5	OM	S1	S2	S3	L&D	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
✓	✓					17.97	17.48	17.52	17.48	**17.45**	17.55	17.63
✓	✓	✓				18.34	17.37	17.07	16.97	16.80	**16.77**	16.89
✓	✓		✓			20.04	18.87	18.96	**18.75**	19.06	19.30	19.73
✓	✓			✓		19.03	17.87	17.31	17.27	**17.14**	17.21	17.39
✓	✓				✓	19.54	18.76	18.15	18.06	17.63	**17.52**	17.54
✓	✓	✓	✓			19.32	17.96	17.25	16.88	16.77	**16.54**	16.61
✓	✓	✓		✓		19.82	18.44	17.90	17.46	17.24	**17.20**	17.30
✓	✓		✓	✓		19.55	18.49	17.95	17.76	**17.51**	17.52	17.60
✓	✓	✓			✓	18.31	16.96	16.37	16.24	**16.23**	16.34	16.44
✓	✓	✓	✓	✓		21.33	19.33	18.62	18.08	17.94	**17.55**	17.57
✓	✓		✓		✓	19.77	18.22	17.47	17.32	**17.01**	17.04	17.31
✓	✓			✓	✓	18.86	17.54	16.97	16.77	**16.72**	16.73	16.91
✓	✓	✓	✓		✓	19.92	18.96	18.30	18.03	17.94	17.87	**17.86**
✓	✓	✓		✓	✓	20.79	19.03	18.02	17.74	17.21	17.17	**17.06**
✓	✓		✓	✓	✓	19.18	18.41	17.79	17.75	17.71	**17.70**	17.71
✓	✓	✓	✓	✓	✓	20.84	19.94	19.27	19.08	18.74	18.50	**18.42**

Chapter 7

Conclusion

This thesis developed and evaluated a multimodal Transformer-based framework for air quality forecasting in the metropolitan area of Milan. By integrating heterogeneous data sources, including satellite imagery, meteorological variables, land cover information, and elevation data, the proposed approach demonstrates how multimodal data fusion can capture the complex temporal dynamics that characterize urban air pollution without requiring pollulant-specific models, just a single holistic approach.

The experimental evaluation shows that the proposed model achieves robust performance across the considered forecasting tasks, with metrics consistent with the difficulty and variability inherent to air quality prediction. In particular, qualitative analysis indicates that the model effectively captures pollutant dynamics at the ground-station level. The predicted concentration curves closely follow the observed trends, suggesting that the architecture is highly responsive to temporal patterns in the input data and capable of modeling pollutant fluctuations.

The systematic ablation study highlight that the composition of the multimodal input plays a critical role. Among the considered sources, Sentinel-5P satellite observations and meteorological variables emerged as the most influential contributors to predictive performance, indicating their central importance for modeling urban air quality dynamics.

Despite the strong temporal predictive capability, the analysis also reveals a limitation in terms of spatial granularity. While the model successfully internalizes the overall atmospheric context of the metropolitan area, it exhibits difficulty in reconstructing fine-grained local variations in pollutant concentration. This suggests that although the multimodal fusion strategy captures the general atmospheric state, the current architecture lacks either the inductive biases or the spatial feature resolution required to distinguish subtle micro-environmental differences within the urban landscape. This limitation in representing spatial heterogeneity highlights an important direction for future research.

Appendix A

Appendix A

Parameter	Downstream	Pretraining
Image Size	[336, 384]	[336, 384]
Patch Size	24	24
Encoder Embedding Dim	768	768
Encoder Depth	12	12
Encoder Heads	12	12
Decoder Embedding Dim	512	512
Decoder Depth	2	8
Decoder Heads	8	8
MLP Ratio	4.0	4.0
Tubelet Size	2	2
Mask Ratio	0	0.9
Drop rate	0	0
Attn. drop rate	0	0
Drop path rate	0	0
Number of Frames	16	16
Learning Rate	1×10^{-4}	1×10^{-4}
Forecast Days	7	–

Table A.1: Comparison of downstream and pretraining model configurations.

Bibliography

[1] World Meteorological Organization. *Global Air Quality Forecasting and Information System (GAFIS)*. WMO Community Knowledge Hub. Accessed: March 1, 2026. 2020. URL: https://community.wmo.int/site/knowledge-hub/programmes-and-initiatives/global-atmosphere-watch-programme-gaw/global-air-quality-forecasting-and-information-system-gafis (cit. on p. 1).

[2] Lanyi Zhang, Jane Lin, Rongzu Qiu, Xisheng Hu, Huihui Zhang, Qingyao Chen, Huamei Tan, Danting Lin, and Jiankai Wang. "Trend analysis and forecast of PM2.5 in Fuzhou, China using the ARIMA model". In: *Ecological Indicators* 95 (2018), pp. 702–710. ISSN: 1470-160X. DOI: https://doi.org/10.1016/j.ecolind.2018.08.032. URL: https://www.sciencedirect.com/science/article/pii/S1470160X18306319 (cit. on pp. 1, 8).

[3] Sheen Mclean Cabaneros, John Kaiser Calautit, and Ben Richard Hughes. "A review of artificial neural network models for ambient air pollution prediction". In: *Environmental Modelling & Software* 119 (2019), pp. 285–304. ISSN: 1364-8152. DOI: https://doi.org/10.1016/j.envsoft.2019.06.014. URL: https://www.sciencedirect.com/science/article/pii/S1364815218306352 (cit. on pp. 1, 9).

[4] Jes Fenger. "Urban air quality". In: *Atmospheric Environment* 33.29 (1999), pp. 4877–4900. ISSN: 1352-2310. DOI: https://doi.org/10.1016/S1352-2310(99)00290-3. URL: https://www.sciencedirect.com/science/article/pii/S1352231099002903 (cit. on pp. 3, 6).

[5] World Health Organization. *Air Pollution*. https://www.who.int/health-topics/air-pollution. Accessed: 2026-03-01. 2026 (cit. on p. 4).

[6] Pavlos Vongelis, Nikolaos G Koulouris, Petros Bakakos, and Nikoletta Rovina. "Air pollution and effects of tropospheric Ozone (O3) on public health". In: *International journal of environmental research and public health* 22.5 (2025), p. 709 (cit. on p. 5).

[7] Pierre Sicard, Evgenios Agathokleous, Susan C. Anenberg, Alessandra De Marco, Elena Paoletti, and Vicent Calatayud. "Trends in urban air pollution over the last two decades: A global perspective". In: *Science of The Total Environment* 858 (2023), p. 160064. ISSN: 0048-9697. DOI: https://doi.org/

10.1016/j.scitotenv.2022.160064. URL: https://www.sciencedirect.com/science/article/pii/S0048969722071649 (cit. on p. 6).

[8] Bo Zhang, Yi Rong, Ruihan Yong, Dongming Qin, Maozhen Li, Guojian Zou, and Jianguo Pan. "Deep learning for air pollutant concentration prediction: A review". In: *Atmospheric Environment* 290 (2022), p. 119347. ISSN: 1352-2310. DOI: https://doi.org/10.1016/j.atmosenv.2022.119347. URL: https://www.sciencedirect.com/science/article/pii/S1352231022004125 (cit. on p. 7).

[9] Pengfei Wang, Peng Wang, Kaiyu Chen, Jun Du, and Hongliang Zhang. "Ground-level ozone simulation using ensemble WRF/Chem predictions over the Southeast United States". In: *Chemosphere* 287 (2022), p. 132428. ISSN: 0045-6535. DOI: https://doi.org/10.1016/j.chemosphere.2021.132428. URL: https://www.sciencedirect.com/science/article/pii/S0045653521029003 (cit. on p. 8).

[10] Stephen F. Mueller and Jonathan W. Mallard. "Contributions of Natural Emissions to Ozone and PM2.5 as Simulated by the Community Multiscale Air Quality (CMAQ) Model". In: *Environmental Science & Technology* 45.11 (2011). PMID: 21545154, pp. 4817–4823. DOI: 10.1021/es103645m (cit. on p. 8).

[11] Youn-Seo Koo, Dae-Ryun Choi, Hi-Yong Kwon, Young-Kee Jang, and Jin-Seok Han. "Improvement of PM10 prediction in East Asia using inverse modeling". In: *Atmospheric Environment* 106 (2015), pp. 318–328. ISSN: 1352-2310. DOI: https://doi.org/10.1016/j.atmosenv.2015.02.004. URL: https://www.sciencedirect.com/science/article/pii/S1352231015001211 (cit. on p. 8).

[12] Tin Thongthammachart, Shin Araki, Hikari Shimadera, Shinnosuke Eto, Tomohito Matsuo, and Akira Kondo. "An integrated model combining random forests and WRF/CMAQ model for high accuracy spatiotemporal PM2.5 predictions in the Kansai region of Japan". In: *Atmospheric Environment* 262 (2021), p. 118620. ISSN: 1352-2310. DOI: https://doi.org/10.1016/j.atmosenv.2021.118620. URL: https://www.sciencedirect.com/science/article/pii/S1352231021004428 (cit. on p. 8).

[13] Huixiang Liu, Qing Li, Dongbing Yu, and Yu Gu. "Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms". In: *Applied Sciences* 9.19 (2019). ISSN: 2076-3417. URL: https://www.mdpi.com/2076-3417/9/19/4069 (cit. on p. 9).

[14] Giacomo Blanco, Luca Barco, Lorenzo Innocenti, and Claudio Rossi. "Urban air pollution forecasting: a machine learning approach leveraging satellite observations and meteorological forecasts". In: *2024 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv)*. IEEE. 2024, pp. 421–426 (cit. on p. 9).

[15] Nabin Rijal, Ravi Teja Gutta, Tingting Cao, Jerry Lin, Qirong Bo, and Jing Zhang. "Ensemble of Deep Neural Networks for Estimating Particulate Matter from Images". In: *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*. 2018, pp. 733–738. DOI: 10.1109/ICIVC.2018.8492790 (cit. on p. 11).

[16] K. Krishna Rani Samal, Korra Sathya Babu, and Santos Kumar Das. "Multi-directional temporal convolutional artificial neural network for PM2.5 forecasting with missing values: A deep learning approach". In: *Urban Climate* 36 (2021), p. 100800. ISSN: 2212-0955. DOI: https://doi.org/10.1016/j.uclim.2021.100800. URL: https://www.sciencedirect.com/science/article/pii/S2212095521000304 (cit. on p. 12).

[17] Bo Zhang, Guojian Zou, Dongming Qin, Yunjie Lu, Yupeng Jin, and Hui Wang. "A novel Encoder-Decoder model based on read-first LSTM for air pollutant prediction". In: *Science of The Total Environment* 765 (2021), p. 144507. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2020.144507. URL: https://www.sciencedirect.com/science/article/pii/S0048969720380384 (cit. on p. 12).

[18] Rong Guo, Qiang Zhang, Xin Yu, Ying Qi, and Bu Zhao. "A deep spatio-temporal learning network for continuous citywide air quality forecast based on dense monitoring data". In: *Journal of Cleaner Production* 414 (2023), p. 137568 (cit. on p. 13).

[19] Xinni Liu, Kai Su, Shubin Wang, and Kamarul Hawari Ghazali. "Intelligent prediction of air quality index based on the transformer-BiLSTM model". In: *Scientific Reports* 15.1 (2025), p. 41838 (cit. on p. 13).

[20] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. *VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training*. 2022. arXiv: 2203.12602 [cs.CV]. URL: https://arxiv.org/abs/2203.12602 (cit. on pp. 15, 25).

[21] Comune di Milano. *Portale del Dato - Open Data Comune di Milano*. Accessed: 2026-03-01 (cit. on p. 16).

[22] Copernicus Earth Observation program. *Sentinel 1 overview*. https://sentiwiki.copernicus.eu/web/s1-mission. Accessed: 2026-03-01 (cit. on p. 17).

[23] Copernicus Earth Observation program. *Sentinel 2 overview*. https://sentiwiki.copernicus.eu/web/s2-mission. Accessed: 2026-03-01 (cit. on p. 19).

[24] Copernicus Earth Observation program. *Sentinel 2 bands*. https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/bands/. Accessed: 2026-03-01 (cit. on p. 19).

[25] Copernicus Earth Observation program. *Sentinel 3 overview*. https://sentiwiki.copernicus.eu/web/s3-mission. Accessed: 2026-03-01 (cit. on p. 20).

[26] Copernicus Earth Observation program. *Sentinel 3 bands*. https://user.eumetsat.int/resources/user-guides/sentinel-3-slstr-level-1-data-guide. Accessed: 2026-03-01 (cit. on p. 21).

[27] Copernicus Earth Observation program. *Sentinel 5 overview*. https://sentiwiki.copernicus.eu/web/s5p-mission. Accessed: 2026-03-01 (cit. on p. 21).

[28] Copernicus Earth Observation program. *Sentinel 5 bands*. https://sentiwiki.copernicus.eu/web/s5p-applications. Accessed: 2026-03-01 (cit. on p. 22).

[29] Open-Meteo. *OpenMeteo*. https://open-meteo.com/en/docs. Accessed: 2026-03-01 (cit. on p. 23).

[30] Copernicus Earth Observation program. *Land Cover*. https://land.copernicus.eu/en/products/urban-atlas. Accessed: 2026-03-01 (cit. on p. 24).

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. *Masked Autoencoders Are Scalable Vision Learners*. 2021. arXiv: 2111.06377 [cs.CV]. URL: https://arxiv.org/abs/2111.06377 (cit. on p. 28).

Dedications

Dedications for everyone