



Research paper

A multivariable sensor-agnostic framework for spatio-temporal air quality forecasting based on Deep Learning

Ignacio-Iker Prado-Rujas^{a,*}, Antonio García-Dopico^{b,c}, Emilio Serrano^a, M. Luisa Córdoba^b, María S. Pérez^a^a Ontology Engineering Group, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Madrid, Spain^b DATSI Computer Science, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Madrid, Spain^c Centro de Investigación en Simulación Computacional, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Madrid, Spain

ARTICLE INFO

Keywords:

Air quality

Sensor data

Spatio-temporal forecasting

Multivariable prediction

Deep Learning

ABSTRACT

Recently, air quality has become a major concern for the protection of the environment and the well-being of people. Air pollution is a key proxy of the quality of life in any city and is crucial in the fight against climate change. Therefore, it is of utmost importance to analyze and forecast the concentration of pollutants. Current efforts in these tasks tend to focus on individual gases and rely on a static view of the measuring sensors.

This work presents an Artificial Intelligence framework that forecasts the concentrations of eleven pollutants in a Region Of Interest for several forecast horizons. The framework is based on Convolutional Long Short-Term Memory networks. Unlike the reviewed state-of-the-art, this framework can work with multiple pollutants simultaneously. Furthermore, it can incorporate exogenous input data by means of independent modules. The framework can also deal with newly installed sensors and recover from missing data. An extensive 10-year hourly dataset was used to train and validate the framework.

The aim of this work is to provide an operational air quality framework for real-world application. This implies not only modeling future air quality but also supporting sensor failures and adaptation to different pollutants. None of the reviewed related works covers all these characteristics to the best of the authors' knowledge. In addition, they are incorporated while scoring error metrics similar to or better than the baselines. Specifically, normalized Root-Mean-Square Error improves 14.1%, 0.8%, and 6.8% compared to the persistence, three-dimensional convolutional, and Long Short-Term Memory models, respectively.

1. Introduction

1.1. Overview and problem description

Today, pollution is one of the main threats to health. Around 4.2 million premature deaths worldwide were caused by diseases induced by outdoor air pollution in 2015 (Landrigan et al., 2018) and in 2019 (World Health Organization, 2023a). Even so, these figures do not take into account household air pollution. Furthermore, around 89% of these deaths occurred in low-income and middle-income countries (World Health Organization, 2022). On top of that, air pollution is particularly harmful to children and the elderly. Due to these facts, air pollution monitoring and modeling have become fundamental for city councils and scientists around the world. On the one hand, **collecting and analyzing air pollutants concentrations** allow to: (1) detect the most polluted areas; (2) identify the causes of such pollution; and (3) evaluate the effect of the adopted policies to reduce it. On

the other hand, **modeling air pollution** is useful to anticipate high-pollution episodes and advise the population accordingly. This advice is particularly important for sensitive areas, such as schools, nursing homes, and hospitals, where there are children, the elderly, and sick people.

Despite the current effort to model air pollution, there is still room for improvement. As Section 1.3 will show, other works present one or more of the following weaknesses: (1) the forecast is limited to just one air pollutant; (2) the historical data used are too short to obtain reliable models; (3) exogenous data are neglected; and (4) the model is tied to a certain number of sensors, not being able to cope with sensor failure.

Problem description: The problem under study in this work is the forecasting of multiple air pollutants in a given Region Of Interest (ROI) and for several forecasting horizons. In short: multivariable spatio-temporal air quality forecasting. Furthermore, the forecast needs to be sensor-agnostic due to the dynamic nature of air pollution (data

* Corresponding author.

E-mail address: iker.prado.ruja@upm.es (I.-I. Prado-Rujas).<https://doi.org/10.1016/j.engappai.2023.107271>

Received 9 May 2023; Received in revised form 4 October 2023; Accepted 5 October 2023

Available online 19 October 2023

0952-1976/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

is harvested with multiple sensors in different locations). This means that the model needs to be able to incorporate new sensors when required and to operate even when some may fail. These characteristics are known as flexibility and robustness (Prado-Rujas et al., 2021b), respectively.

Proposal and contributions: This work develops a Deep Learning (DL)-based framework that combines air pollutant data, weather, and calendar information to predict future air quality. In the validation process, data from 11 air pollutants are collected from 24 real-world sensors that are spread throughout a certain ROI. The characteristics and hazards of these pollutants are described in Section 1.2. The main contribution of this work is the development of a DL-based framework for air quality forecasting with the following characteristics:

- **Spatio-temporal:** Both the input and output of the model comprise multiple instants and locations with fine resolutions.
- **Multivariable:** It can work with multiple air pollutants.
- **Extensive data:** The model is built using an extensive temporal coverage.
- **Exogenous data:** The input of the model can accommodate additional variables (beyond pollutants).
- **Sensor-agnostic:** Adaptability to a variable number of sensors (flexibility). Also: ability to cope with sensor failure (robustness).

The structure of this work is as follows. Section 1 begins by explaining the importance of the topic, describing the problem, and the proposed approach (Section 1.1). The section continues by presenting the characteristics and hazards of the pollutants studied in Section 1.2. Subsequently, Section 1.3 discusses several related works that collect, analyze, and model air pollution data. The framework and its components are described in Section 2. Afterward, Section 3 presents the case study (Section 3.1) and the experiments that were carried out to validate the framework (Section 3.2). The results are discussed later in Section 4, to finish with Section 5 by summarizing conclusions and future research lines.

1.2. Background

According to European Environment Agency et al. (2020), some of the most harmful pollutants for humans are particulate matter (PM), nitrogen dioxide (NO_2), and ground-level ozone (O_3). Particulate matter is usually divided further into PM_{10} and $\text{PM}_{2.5}$, which means particulate matter of 10 or 2.5 micrometers or less in diameter, respectively. Other hazardous gases include sulfur dioxide, nitrogen oxides, carbon monoxide, and hydrocarbons (e.g., benzene and toluene) (Domingo and Rovira, 2020; Kim et al., 2018). The World Health Organization (WHO) has recently published a report on air quality guidelines after a systematic review (World Health Organization, 2021). The European Commission has also established standards for exposure to these pollutants (Directorate-General for Environment, 2008). In this work, the following pollutants are considered:

- **Coarse particulate matter (PM_{10}) and Fine particulate matter ($\text{PM}_{2.5}$):** Generally, particulate matter serves as a proxy indicator of air pollution. Fundamentally, it includes sulfate, nitrates, ammonia, sodium chloride, black carbon, mineral dust, and water. The distinction between PM_{10} and $\text{PM}_{2.5}$ is based on the fact that, while the former cannot penetrate the lungs into the blood system, the latter does. Particulate matter is closely related to mortality and morbidity (World Health Organization, 2022).
- **Ozone (O_3):** This pollutant affects mainly the lungs and can cause asthma (World Health Organization, 2022). Furthermore, it worsens the mortality of chronically ill people. It is formed by the reaction of other pollutants (mainly emitted by cars, such as NO_2) with sunlight, and therefore higher concentrations of O_3 appear during the spring and summer seasons. Thermoelectric plants, industrial activities, and intensive livestock farming are additional sources of O_3 .

- **Nitrogen dioxide (NO_2):** According to World Health Organization (2022), it can cause bronchitis in asthmatic children and reduce lung function and growth. The main source of NO_2 are fossil fuels burned by internal combustion engines.
- **Nitrogen monoxide (NO):** It produces nitric acid when it reacts with water, which contributes to acid rain. In addition to that, it reacts with ozone, resulting in the depletion of the ozone layer. It is also formed in combustion systems.
- **Nitrogen oxides (NO_x):** It is a generic term for denoting nitrogen oxides and includes NO_2 and NO.
- **Sulfur dioxide (SO_2):** World Health Organization (2022) reports that SO_2 can inflame the respiratory system, exacerbating asthma and chronic bronchitis, and can also cause eye irritation. Furthermore, SO_2 forms sulfuric acid when combined with water, which is the main component of acid rain.
- **Carbon monoxide (CO):** The most common source of this pollutant is thermal combustion, both from combustion engines and industry. Prolonged exposure to this toxic gas can cause death.
- **Benzene (BEN):** C_6H_6 (or BEN for simplicity, based on Madrid's Open Data Portal notation, see Madrid City Hall, 2023) is a hydrocarbon that is a key component of crude oil. It can appear due to both natural (e.g., wild fires and volcano eruptions) and human causes (e.g., plastic industry, combustion engines, tobacco, etc.). It is a carcinogen, and no exposure is considered safe.
- **Toluene (TOL):** C_7H_8 is another hydrocarbon that is usually used as a precursor to benzene. Furthermore, it is commonly employed as a solvent for paints and sometimes to improve certain fuels. Even if it is much less toxic than benzene, it could cause unconsciousness and even death.
- **Ethylbenzene (EBE):** C_8H_{10} , yet another hydrocarbon, is obtained from benzene and ethylene and is commonly utilized to produce polystyrene. Moreover, ethylbenzene is produced when gas or coal is burned and can be found in tobacco smoke. As for its effects, it is a precursor to smog and causes eye and throat sensitivity.

These facts evince the importance of air pollution for human health. Therefore, a tool capable of predicting high-risk scenarios would help to take preventive measures. Such a tool should not only predict future values of these gases but also be flexible and robust. In this way, it could be used in a real-world scenario where sensors fail and new ones are installed regularly. Furthermore, the incorporation of new gases into the study would also be relevant. In addition, the integration of meteorological and public holidays information into the tool should improve forecasts. This work combines all these characteristics without losing precision compared to baselines and the literature, as will be shown in Section 3.

1.3. Related work

This section starts with an overview of tools and initiatives to gather, exploit, and visualize air pollution data in Section 1.3.1. Since the current work involved looking for valuable use cases in which to evaluate the developed framework, this research is briefly reflected here. Afterward, Section 1.3.2 discusses some works that use those resources to model air quality.

1.3.1. Air pollution data acquisition and harnessing

Air quality has become a central issue for both health institutions and official agencies (World Health Organization, 2021; European Environment Agency et al., 2020). As a result, many actions are being carried out to boost its research and improvement. The starting point to do so is to build the infrastructure to measure different pollutants at multiple locations over time. Thereafter, these data can be stored in data repositories for its harnessing. In this sense, there are initiatives at the regional (Gobierno de Canarias, 2022; Madrid

City Hall, 2023), national (Ministerio para la Transición Ecológica y el Reto Demográfico, 2022; Environmental Protection Administration in Taiwan, 2022), and global levels (European Environment Agency, 2022; OpenAQ, 2023). Specifically, the nonprofit organization (OpenAQ, 2023) provides an Application Programming Interface (API) to access open data on global air quality of seven pollutants. To contribute to the best possible understanding and to help with data curation, such repositories should provide: temporal coverage, geolocation of the sensors, pollutants covered, measurement units, measurement method, measurement equipment, missing data handling, etc.

Thereupon, these data can be turned into valuable information to understand the dynamics of pollution. World Health Organization (2023b) provides maps at the national level that depict the concentration of $PM_{2.5}$ for several years. Other initiatives focus on a finer spatial granularity and collect sensor-level information. In this sense, the European Environment Agency (EEA) provides web applications for the visualization of contaminants recorded by sensors distributed throughout Europe. As examples, European Environment Agency (2023b) focus on $PM_{2.5}$, while European Environment Agency (2023a) gathers additional pollutants together with an Air Quality Index (AQI).

Since the outbreak of the COVID-19 pandemic, many people's habits have changed, making an impact on pollution trends. For this reason, EEA maintains since then an interactive visualization tool that shows weekly and monthly concentrations of $PM_{2.5}$, PM_{10} , and NO_2 in European cities (European Environment Agency, 2020). Numerous works study the impact of COVID-19 on air pollution, for instance, by analyzing satellite images (Dutheil et al., 2020). In contrast, the work by Berman and Ebisu (2020) uses sensor data from the United States to study this correlation. Moreover, Venter et al. (2020b) combine satellite images and ground sensor data to explore this subject at a global level, providing a visualization tool (Venter et al., 2020a). In addition to this, there are works that study how air pollution favors the spread of coronaviruses. As examples, the work of Fattorini and Regoli (2020) focuses on Italy and that of Travaglio et al. (2021) on England.

1.3.2. Modeling air pollution

As discussed in Section 1.1, the second challenge that can alleviate air pollution is its modeling. According to Mao et al. (2021), air quality forecasting models can be divided into two categories: physical and statistical models. Whereas the former are based on classical, well-studied methods (Wang et al., 2001; Holmes and Morawska, 2006), they are computationally more intensive than the latter (Zlatev and Dimov, 2006). This fact, together with the advances of Artificial Intelligence (AI) (e.g., Lopez-Gomez et al., 2023; Fairfax et al., 2023; Roy and Bhaduri, 2023) make more common to find works falling into the second category, such as the current one. In the remainder of this section, several AI statistical-based air quality forecasting models are discussed.

Deep Neural Networks (DNNs) are widely established as a flexible tool to tackle diverse forecasting tasks, and air pollution is no exception. As a first example, Huang and Kuo (2018) use a combination of several 1-dimensional convolutions and a Long Short-Term Memory (LSTM) network to predict $PM_{2.5}$ on the next hour. They address the problem from a smart city point of view, focusing on Beijing. Furthermore, the model incorporates wind speed and hours of rainfall. Similarly, Wen et al. (2019) predict $PM_{2.5}$ at local and global levels in Beijing and China, respectively. However, in this case they use a combination of Convolutional Neural Networks (CNNs), LSTMs and k-nearest neighbors. Moreover, meteorological data and other physical parameters are incorporated into the model. Qi et al. (2019) aim for the same goal, but define a graph structure and employ regional graph convolutions. Zhou et al. (2019) move onto a finer spatial resolution to predict $PM_{2.5}$, PM_{10} and NO_x in the city of Taipei (Taiwan), up to 4 h ahead. To do so, they develop a Deep Multi-output LSTM neural network, which consists of an Multilayer Perceptron (MLP) with LSTM neurons. Drewil and Al-Bahadili (2022) also concentrate on predicting

$PM_{2.5}$, PM_{10} and NO_x with LSTM methods. Nevertheless, they extend the study for CO and employ a genetic algorithm to tune the model hyperparameters for next day forecasts. Devasekhar and Natarajan (2023) propose a technique to forecast air quality by combining supervised Machine Learning (ML) together with multi-agent systems. However, they do not explore the use of meteorological variables.

Some authors investigate the problem of forecasting in multiple locations. In their work, Mao et al. (2021) develop a framework for predicting air quality in the next 24 h. It consists of a neural network with a temporal sliding LSTM extended model. They train one model per city (12 in total), which includes several stations (65 in total) that collect $PM_{2.5}$. The model considers meteorological data, alongside seasons and months as one-hot encoded vectors. The authors of this work study the effect of long-term prediction horizons (up to 72 h) on model performance. Their findings suggest that the optimal time lag increases with the forecast horizon. Differently, Liu et al. (2022) concatenate a temporal prediction of an AQI with spatial inference for unknown locations. In this case, the AQI is set as the highest value of the available pollutants, that is, O_3 , $PM_{2.5}$, PM_{10} , CO, SO_2 , and NO_2 . They introduce meteorological variables and spatial attributes (e.g., population, road network, elevation, etc.) into the model. In their work, they notice that further forecast horizons imply a worsening in model performance for all seasons. This behavior is studied in Section 3 of this work. De Medrano et al. (2021) study the same city as the current work, i.e., Madrid (Spain). In their work, NO_2 , O_3 , $PM_{2.5}$, and PM_{10} are forecasted for the next 48 h. They develop a chain of models that consider exogenous information, such as weather and anthropogenic features. However, each air pollutant and station are modeled separately, and then the predictions are combined using Bayesian estimation. Some of their results will be compared to the framework proposed in Section 3.

Achieving a finer spatial resolution can be challenging and usually requires using interpolation techniques. Meteosim (2021) develops three platforms for different air pollution-related tasks: (1) simulation of gases dispersion; (2) meteorological forecasting; and (3) air quality and dust forecast for 72 h. In addition, they build tools for threshold exceedance and to detect sources of bad smell. The spatial resolution used is 1 km for meteorological prediction and 250 m for air quality forecast.

All of these works are promising, but lack important generalization capabilities for real-world scenarios. Air pollution forecasting models should: consider multiple pollutants, be trained and validated with extensive and varied data, be able to incorporate exogenous information, and be detached from the number and distribution of sensors. The current work contributes to these lines of research by defining and validating a DL-based spatio-temporal forecasting framework. This framework:

1. can work with multiple air pollutants both as input data and forecasts (up to 11 for the studied case study);
2. is trained with extensive real-world data from 24 sensors gathered over a decade;
3. can naturally integrate exogenous input data to improve the forecasts; and
4. is sensor-agnostic, which allows for:
 - flexibility, in the sense that it can work with a variable set of sensors (in terms of size and geolocations); and
 - robustness, meaning that it can recover from sensor failures and continue producing reliable forecasts.

Table 1 compares the related works studied to the current one based on these characteristics. The methodology to achieve them is developed in Section 2, and the evaluation is carried out in Section 3.

Table 1
Summary of related works regarding statistical air pollution modeling.

	Multiple pollutants	Extensive data	Exogenous data	Sensor-agnostic
Huang and Kuo (2018)	×	~	×	×
Qi et al. (2019)	×	×	×	×
Wen et al. (2019)	×	×	~	×
Zhou et al. (2019)	✓	✓	✓	×
De Medrano et al. (2021)	~	✓	✓	×
Mao et al. (2021)	×	✓	✓	~
Meteosim (2021)	×	?	?	~
Drewil and Al-Bahadili (2022)	✓	~	×	×
Liu et al. (2022)	~	×	~	✓
Devasekhar and Natarajan (2023)	×	×	✓	×
The current work	✓	✓	✓	✓

✓: Covered. ×: Not covered. ~: Partially covered. ?: Unknown.

2. Proposed approach and methodology

This section describes the proposed model, which is built upon a previous work on mobility demand forecasting (Prado-Rujas et al., 2023a). It begins in Section 2.1 by formalizing the components used to build the forecasting framework. Thereupon, Section 2.2 presents the developed air quality forecasting model. It is composed of the spatio-temporal module (Section 2.2.1) and the weather and calendar modules (Section 2.2.2). Finally, Section 2.3 defines how the model is evaluated in both spatial and temporal dimensions.

2.1. Formalization of the approach

Temporal aspects: Consider a set of n_v variables (e.g., pollutant gases) $\mathcal{V}$ that are observed and recorded using measurement instruments over time. The *temporal coverage* $\mathcal{T}_v$ refers to the set of time intervals in which the variable $v \in \mathcal{V}$ is observed. For a given experiment, $\mathcal{T}_v$ can include periods of inactivity or malfunction. The *temporal granularity or resolution* τ_v refers to the time elapsed between consecutive observed data points of v . Throughout a given experiment, τ_v is fixed. The left endpoint of a time interval is used as a tag when referring to it. For instance, $t = 12:00$ refers to the interval [12:00, 13:00) when $\tau_v = 1$ h, but to [12:00, 12:15) when $\tau_v = 15$ min. For simplicity, a common temporal coverage $\mathcal{T}$ and temporal granularity τ are considered for all $v \in \mathcal{V}$ for a given experiment.

Spatial aspects: Each $v \in \mathcal{V}$ can be recorded in several geolocations $\mathcal{G}_v$, which are determined by their latitude and longitude. Therefore, for a given $v \in \mathcal{V}$ the instruments record data points: $X_v = \{x_{g,v}^t \mid t \in \mathcal{T}, g \in \mathcal{G}_v\}$. The *spatial coverage* $\mathcal{A}$ refers to the rectangular ROI that contains all geolocations $\mathcal{G} = \{g \mid g \in \mathcal{G}_v, v \in \mathcal{V}\}$ relevant to a given experiment. A *mesh-grid* $\mathcal{M}_v^t$ is a 2-dimensional array that gathers data points of v based on a regular partition of $\mathcal{A}$. Each $\mathcal{M}_v^t$ corresponds to a certain time interval t . The *shape* of $\mathcal{M}_v^t$ is a pair formed by its number of rows r and columns c . The *spatial granularity or resolution* σ of a mesh-grid is a pair obtained by dividing the base of $\mathcal{A}$ by c and its height by r . It can also be seen as the distance between the centers of consecutive elements of $\mathcal{M}_v^t$ along the longitude and latitude.

Time series forecasting: The number of previous time intervals considered for the prediction is known as *timestep*, and it is denoted by n_x . Similarly, the *number of horizons* (number of future instants predicted) is denoted by n_y throughout this work. The *shift* s is the leap (number of time intervals) between the input and output windows. The parameters n_x, n_y and s must be greater than zero. Furthermore, they are graphically depicted in Fig. 1, together with the mesh-grid shape (r and c).

Problem formalization: As usual, $\mathcal{T}$ is split into three contiguous non-overlapping subsets: $\mathcal{T} = \mathcal{T}_{\text{train}} \cup \mathcal{T}_{\text{val}} \cup \mathcal{T}_{\text{test}}$ (i.e., training, validation, and testing sets). The models are then trained just by inspecting the data observations of $\mathcal{T}_{\text{train}}$ while monitoring the predictions in

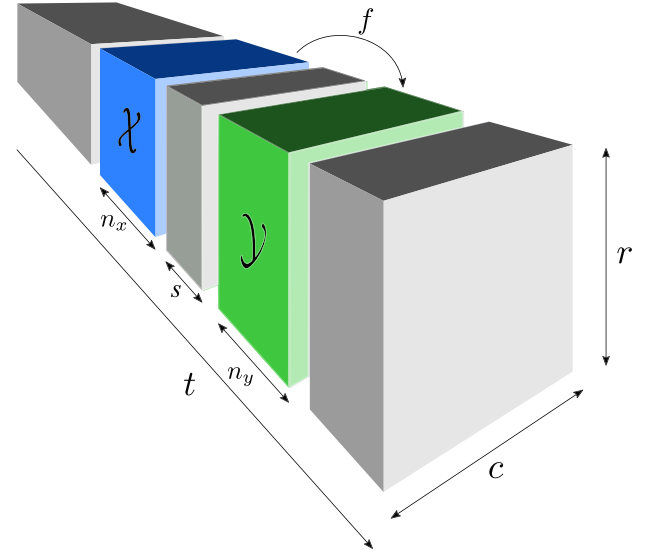


Fig. 1. Spatio-temporal nature of the problem, depicting the employed naming.

$\mathcal{T}_{\text{val}}$. Afterward, their worth is evaluated in $\mathcal{T}_{\text{test}}$. With this set up, the problem under study can be posed as follows:

$$\text{Forecast } \mathcal{Y}^{t^*} = \left\{ y_{g,v}^{t^*} \mid g \in \mathcal{G}_v, v \in \mathcal{V} \right\} \quad \forall t^* \in \mathcal{T}_{\text{test}}$$

$$\text{given } \mathcal{X}^{t^*} = \left\{ x_{g,v}^t \mid t = t^* - s \cdot \tau, g \in \mathcal{G}_v, v \in \mathcal{V} \right\}.$$

In other words: estimate each data observation from the test set (for all available geolocations and variables), given data observations that were recorded s time intervals before. Note that, with this definition, $n_y = 1$ and $n_x = 1$. In general, multiple future instants can be predicted at once ($n_y > 1$). In that case, the set $\mathcal{Y}^{t^*}$ comprises the time intervals $\{t^*, t^* + \tau, \dots, t^* + (n_y - 1) \cdot \tau\}$ for each $t^* \in \mathcal{T}_{\text{test}}$. Similarly, multiple previous time intervals can be considered in the input ($n_x > 1$). Then the set $\mathcal{X}^{t^*}$ comprises the time intervals $\{t^* - (s + n_x - 1) \cdot \tau, \dots, t^* - (s + 1) \cdot \tau, t^* - s \cdot \tau\}$ for each $t^* \in \mathcal{T}_{\text{test}}$. The problem studied is represented graphically in Fig. 1.

2.2. Spatio-temporal air quality forecaster

This section describes the developed model for air quality forecasting based on Section 2.1. First, Section 2.2.1 presents the main component: the spatio-temporal module. Afterward, the weather and calendar modules are explained in Section 2.2.2.

2.2.1. Spatio-temporal module

Mesh-grid construction: As described in Section 2.1, the input to the forecasting model includes observations from the n_x previous time intervals. Afterward, it predicts for n_y future time intervals, for all available geolocations $\mathcal{G}_v$ and for each $v \in \mathcal{V}$. One way to tackle this problem is to train independent models for each pollutant, geolocation, and forecast horizon. However, this approach neglects the correlations between contiguous horizons, geolocations, and pollutants. Arguably, a single model that combines all of these dimensions should achieve better results. In that case, the forecasted set $\mathcal{Y}^{t^*}$ should have $n_y \cdot |\mathcal{G}|$ predictions, for every $t^* \in \mathcal{T}_{\text{test}}$. This shows the need for a model that can work with temporal and spatial information together in a single structure. This gap can be bridged by combining the available time series at different geolocations into mesh-grids for each pollutant. In this way, the set $\mathcal{Y}^{t^*}$ becomes $\{\mathcal{M}_v^{t^*} \mid v \in \mathcal{V}\}$ with $t^* \in \mathcal{T}_{\text{test}}$. Analogously, $\mathcal{X}^{t^*} = \{\mathcal{M}_v^t \mid t = t^* - s \cdot \tau, v \in \mathcal{V}\}$. Therefore, the dataset is now arranged as a 4-dimensional tensor with the following dimensions: time, latitude, longitude and variable (the different pollutants). The

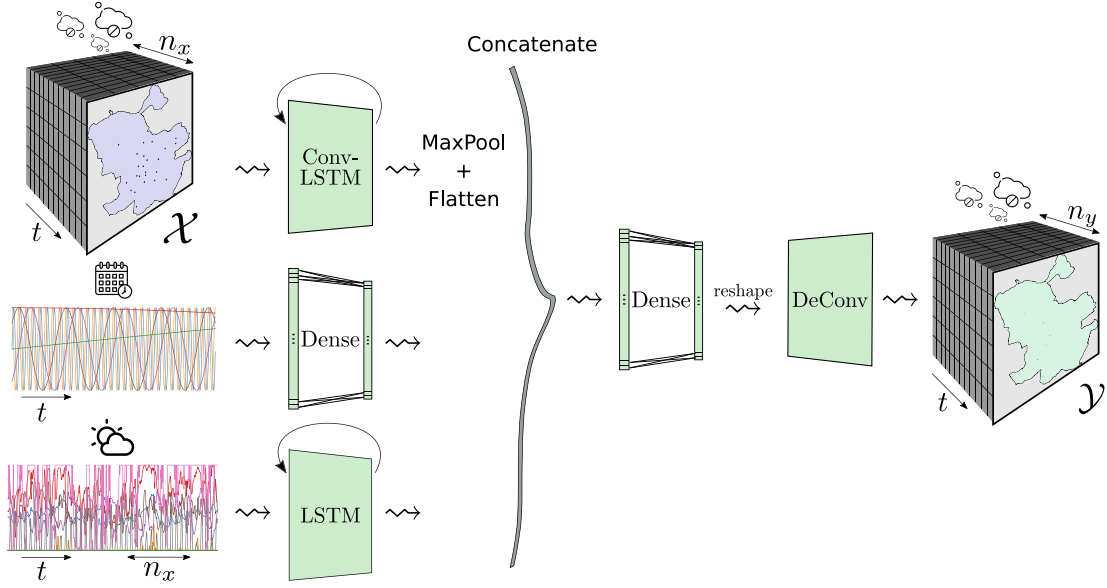


Fig. 2. Schematic of the ST-AQF model.

practicalities of how to achieve this transformation into mesh-grids will be discussed in Section 3.1.1.

Spatio-temporal module: Thanks to the mesh-grid, the pollutant data are fused into a single structure that accounts for both temporal and spatial correlations. Now, it remains to define the model that will work with it, namely the Spatio-Temporal Air Quality Forecaster (ST-AQF). The Convolutional Long Short-Term Memory (ConvLSTM) is a kind of layer obtained by substituting matrix multiplications at each gate of the LSTM with the convolutional operator (Shi et al., 2015). In this way, it combines the LSTM ability to find temporal correspondences with the extraction of 2-dimensional (spatial) features of the convolutional operator. These DNNs have successfully been applied to related problems, such as precipitation forecasting from radar images (Agrawal et al., 2019), solar irradiance forecasting with satellite images (Gallo et al., 2022) or ground sensors (Prado-Rujas et al., 2021a), and road networks identification using satellite images (Kamdem De Teyou, 2020). The deconvolution is another kind of layer that establishes the same connectivity patterns as the convolutional operator, but in the opposite direction. Thus, it is also known as transposed convolution (see Zeiler et al., 2010 for details and Dumoulin and Visin, 2018 for a visual description). The steps that the spatio-temporal module performs are depicted in Fig. 2 (see the top module) and are as follows:

1. The air pollutant mesh-grid is fed to the ConvLSTM layer, which produces a 3-dimensional tensor.
2. This tensor is then shrunk by a max-pooling layer, flattened, and passed by a fully-connected dense layer.
3. The resulting vector is reshaped into a 4-dimensional tensor to match the output dimensions: time, latitude, longitude, and pollutant.
4. The next step is a deconvolution. In this way, the prediction $\mathcal{Y}^{t^*}$ is obtained as $n_y \cdot |\mathcal{V}|$ mesh-grids of shape (r, c) .
5. The final step is to undertake the opposite conversion from the one that yielded the mesh-grids. In so doing, the $n_y \cdot |\mathcal{G}|$ predictions are obtained for every $t^* \in \mathcal{T}_{\text{test}}$.

This version of the ST-AQF model only includes the spatio-temporal module. Two additional components are described in the next section, namely, the weather and calendar modules.

2.2.2. Weather and calendar modules

The input of the ST-AQF model described in Section 2.2.1 can be enriched with exogenous data. In other words, the model can be extended with multiple input datasets $\{\mathcal{X}_1, \dots, \mathcal{X}_p\}$ beyond $\mathcal{X}^{t^*}$ to improve its predictions. These exogenous datasets cannot overlap with $\mathcal{Y}^{t^*}$, that is: $\mathcal{X}_i \cap \mathcal{Y}^{t^*} = \emptyset, \forall i \in \{1, \dots, p\}$. In this manner, the ST-AQF has as many modules as input datasets. Then, it performs feature extraction on each of them, and this information is combined. Based on the type of exogenous data, each module will perform different operations:

- **Weather module:** It considers weather information from the same time window as $\mathcal{X}^{t^*}$, i.e., n_x instants. Since these data are a time series, it is fed to a recurrent LSTM layer. The information extracted is then concatenated with that from other modules (see the bottom module in Fig. 2).
- **Calendar module:** It considers temporal information, such as time of the day and holidays. In this case, these values are extracted just for the interval t^* . Therefore, the input vector is fed to a fully-connected dense layer. Afterward, the output is concatenated in the same way as the weather module (see the middle module in Fig. 2).

Thus, four different variants of the forecasting model are considered based on the modules that comprise it:

- **ST-AQF:** It only includes the spatio-temporal module described in Section 2.2.1. It has 807147 trainable parameters.
- **ST-AQF^W:** It considers both the spatio-temporal and the weather modules. It has 1003243 trainable parameters.
- **ST-AQF^C:** It considers both the spatio-temporal and the calendar modules. It has 922987 trainable parameters.
- **ST-AQF^{*}:** It includes all three modules: spatio-temporal, weather, and temporal. It has 1119083 trainable parameters.

In this manner, the information of exogenous datasets can be incorporated in a natural way by means of independent modules. The ST-AQF^{*} model that includes the three modules can be depicted in Fig. 2. Furthermore, plots of the four ST-AQF variants can be found in Prado-Rujas et al. (2023c), including input and output shapes for each layer.

2.3. Evaluation methodology

The standard error metrics Root-Mean-Square Error (RMSE), normalized RMSE (nRMSE), and Mean Absolute Error (MAE) are used to evaluate the trained models. However, as described in Section 2.1, the aim is to forecast y^{t^*} when $t^* \in \mathcal{T}_{\text{test}}$. Therefore, the evaluation metrics are considered as a function of the horizon, geolocation, and the evaluated variable. To this end, the metrics were adapted following the notation from Section 2.1 as follows:

$$\text{RMSE}(h, g, v) = \sqrt{\frac{1}{|\mathcal{T}_{\text{test}}|} \sum_{t^* \in \mathcal{T}_{\text{test}}} (y_{g,v}^{t^*+h} - p_{g,v}^{t^*+h})^2}, \quad (1)$$

$$\text{nRMSE}(h, g, v) = \frac{\text{RMSE}(h, g, v)}{\frac{1}{|\mathcal{T}_{\text{test}}|} \sum_{t^* \in \mathcal{T}_{\text{test}}} y_{g,v}^{t^*}}, \quad (2)$$

$$\text{MAE}(h, g, v) = \frac{1}{|\mathcal{T}_{\text{test}}|} \sum_{t^* \in \mathcal{T}_{\text{test}}} |y_{g,v}^{t^*+h} - p_{g,v}^{t^*+h}|, \quad (3)$$

where $h \in \{0, \dots, (n_y - 1) \cdot \tau\}$ represents the forecast horizon, $g \in \mathcal{G}_v$ is a geolocation for the variable $v \in \mathcal{V}$, $y_{g,v}^{t^*+h} \in \mathcal{Y}^{t^*}$ is a data observation, and $p_{g,v}^{t^*+h}$ represents the predicted value for the same time interval, geolocation and variable as $y_{g,v}^{t^*+h}$.

Thus, error metrics are obtained for every forecast horizon h , geolocation g , and variable v . To summarize these values, the median across geolocations is calculated for each horizon and variable. Afterward, the average over horizons is also computed, yielding one error value per variable for the studied model. Throughout this work, this is how all error metrics are computed for every model and variable. The tables and figures in Section 3 depict the RMSE and nRMSE. The MAE is also included in Prado-Rujas et al. (2023c) for completeness.

3. Experiments and results

This section presents the experiments that were carried out to validate the methodology described in Section 2. It starts by describing the case study and the data used in Section 3.1. Then, the section continues with the different experiments in Section 3.2.

3.1. Case study: Madrid

Section 3.1.1 describes the air quality data and how it was processed. Subsequently, Section 3.1.2 outlines the temporal information used. Finally, Section 3.1.3 presents the meteorological variables considered in the experiments.

3.1.1. Air pollutants data

The air quality data used in this work were collected in the city of Madrid (Spain) and obtained from the open data portal of the [Madrid City Hall \(2023\)](#). Its temporal coverage goes from the 1st of January 2010 to the 31st of December 2019, being $\tau = 1$ h. There are a total of 24 air quality stations, whose relative geolocations are depicted in Fig. 3. Each collects several of the 11 pollutants studied, which were described in Section 1.2. The number of active monitoring stations varies over time and by pollutants. For instance, SO₂ and PM_{2.5} are covered by at least four sensors for more than 99% of the observations. Other pollutants are generally covered by all 24 sensors, such as NO, NO₂, and NO_x.

All air quality data were standardized and normalized. These two feature scaling techniques are shown in Eqs. (4) and (5):

$$\text{stand}(x) = \frac{x - \mu(X_v)}{\sigma(X_v)}, \quad (4)$$

$$\text{norm}(x) = \frac{x - \min(X_v)}{\max(X_v) - \min(X_v)}, \quad (5)$$

where $x \in X_v$ and $v \in \mathcal{V}$, X_v are data observations of the variable v , $\mu(X_v)$ is the mean value of X_v , $\sigma(X_v)$ is the standard deviation of X_v ,

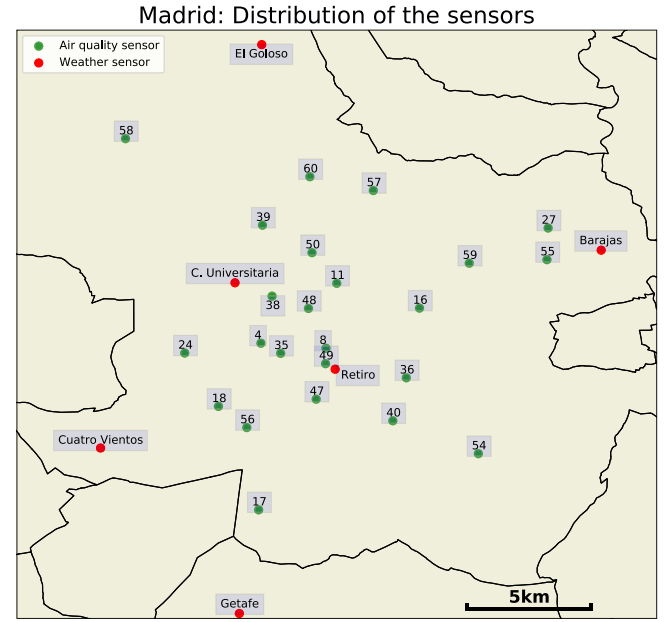


Fig. 3. Madrid's metropolitan area, scattered with the geolocations of the 24 air pollution sensors (in green) and 6 weather stations (in red).

$\min(X_v)$ is the minimum value of X_v , $\max(X_v)$ is the maximum value of X_v .

To build the mesh-grids, nearest and linear interpolation techniques were applied to the available data at every time interval. Section 3.2.1 analyzes which combination of scaling and interpolation methods achieves better results. The interpolation transforms several time series $\{x_{g,v}^t \mid t \in \mathcal{T}\}$ from different geolocations g into a time series of mesh-grids $\{\mathcal{M}_v^t \mid t \in \mathcal{T}\}$, for each variable $v \in \mathcal{V}$. The shape of the mesh-grid was fixed at (35, 30). In this way, the spatial granularity σ is close to (500, 500) meters. Therefore, the final dataset that gathers the ten years for the eleven variables has shape (87648, 35, 30, 11), which corresponds to $|\mathcal{T}| \times r \times c \times |\mathcal{V}|$ total values. This is the dataset fed to the spatio-temporal module described in Section 2.2.1. Hourly boxplots and statistics of the pollutant data can be found in Appendix B. Furthermore, the processed data are available on Zenodo (Prado-Rujas et al., 2023b) for reproducibility purposes.

3.1.2. Calendar data

Exploration of the data (see Section 3.1.1) highlights the periodic behavior of pollution variables. Several useful parameters can be extracted for each prediction from the ongoing time interval $t = t^* - s \cdot \tau$, as the following:

- **time of the day** normalized to $[-1, 1]$ by the sine and the cosine,
- **time of the week** normalized to $[-1, 1]$ by the sine and the cosine,
- **time of the year** normalized to $[-1, 1]$ by the sine and the cosine,
- **weekday or weekend** in $\{0, 1\}$, and
- **holiday** in $\{0, 1\}$ (obtained from [Comunidad de Madrid, 2023](#)).

This information is extracted for each t , and therefore they are not treated as time series (see Section 2.2.2). These data are fed to the calendar module described in Section 2.2.2.

3.1.3. Meteorological data

Weather conditions influence the air quality observed at a given location, as discussed by Abirami and Chitra (2021), among others. Data from six weather stations were obtained from the Agencia Estatal de Meteorología (AEMET) (see Fig. 3). Specifically, the variables used are as follows:

Table 2

Summary of the GPUs utilized for the experiments in this work.

GPU model	Number of GPUs	Compute capability	Memory
NVIDIA GeForce GTX TITAN X	4	5.2	12 GB
NVIDIA GeForce GTX TITAN Black	2	3.5	6 GB
NVIDIA GeForce GTX TITAN	1	3.5	6 GB
NVIDIA GeForce RTX 2070	1	7.5	8 GB
NVIDIA GeForce RTX 2080	1	7.5	12 GB

- **Temperature:** Measured every hour in tenths of degrees Celsius. The average number of missing observations across the sensors is 2.2%.
- **Precipitation:** Measured every hour in tenths of millimeters. The average number of missing observations across the sensors is 3.4%.
- **Wind:** Measured every hour in tenths of degrees (direction) and kilometers per hour (speed). Angles are generally not an appropriate input for DNNs (e.g., 0° and 359° should be close to each other). Therefore, direction and speed were converted to xy -coordinates (i.e., a wind vector) in kilometers per hour. The average number of missing observations across sensors is 9.1%.
- **Humidity:** Measured as a percentage at 00:00, 07:00, 13:00, and 18:00 (UTC). The remaining hours are forward filled from the latest available ones.

Missing data were filled on the basis of observations from the station with the least gaps (i.e., Barajas). If these values were also missing, the next most reliable station was checked (i.e., Retiro). Furthermore, all weather variables are normalized as per Eq. (5). These data are treated as time series, as discussed in Section 2.2.2. Therefore, a window of n_x consecutive time intervals is fed to the weather module. This corresponds to the same time intervals as for $\mathcal{X}^*$. Additional statistics on weather variables can be found in Appendix B.

3.2. Experiments

This section starts with the experimental setup and hyperparameter analysis and tuning in Section 3.2.1. Afterward, Section 3.2.2 compares the error scored by the ST-AQF model with several baselines, and presents an analysis of the parameter n_x . Thereafter, Sections 3.2.3 and 3.2.4 detail how to incorporate multiple pollutants and exogenous data into the ST-AQF model, respectively. Finally, experiments on the robustness and flexibility of the model are presented in Section 3.2.5. Since the ST-AQF model considers 11 variables, the number of plots in each experiment has been limited (keeping NO_2 in all as a proxy for air pollution). In any case, the graphs for the remaining pollutants can be found in Prado-Rujas et al. (2023c).

3.2.1. Experimental setup and hyperparameter tuning

The models presented were developed using the Keras library (Chollet et al., 2015) with the Tensorflow backend (Abadi et al., 2015). Additionally, the NumPy, Pandas, and PyTables Python libraries were used for data manipulation and storage. The experiments were carried out using three Ubuntu machines, each with 12 cores, 24 logical processors, and three graphic cards (see Table 2). The data used for training cover a decade between 2010 and 2019, as explained in Section 3.1.1. Specifically, the years 2010 to 2016 were used for training (~70%), 2017 for validation (~10%), and 2018 and 2019 for testing (~20%). The errors for all experiments are calculated on the test set. Early stopping was employed with patience of 10 epochs on the validation loss.

Fine-tuning the optimizer, loss function, and learning rate: In this experiment, the models were trained without using weather nor calendar data for 100 epochs, only historical air quality. Besides, $n_x = 4$, $s = 1$, $n_y = 4$, $n_v = 11$, and the scaling and interpolation methods were fixed to standardized and nearest-neighbors, respectively. The 96 cases studied were derived from:

- 8 optimizers: Adam, Nadam, Adamax, Adadelata, Adagrad, SGD, RMSprop, and Ftrl.
- 3 loss functions: Mean Squared Error (MSE), MAE and Mean Squared Logarithmic Error (MSLE).
- 4 learning rates: 10^{-3} , 10^{-4} , 10^{-5} , $5 \cdot 10^{-6}$.

As for the optimizers, the ones that stood out the most were Adam, Nadam, Adamax, and RMSprop, with Adamax the most consistent. Furthermore, setting the learning rate at 10^{-3} derived the best results. Experiments indicate that the best loss function is specific to pollutants, obtaining better results both for MAE and MSE. All in all, fixing the optimizer to Adamax and the learning rate to 10^{-3} rather shows a balance between these two. Specifically, the former obtains an improvement of just 2% on average (over variables and horizons). Given these facts and to facilitate the reproducibility of the experiments, the loss was chosen to be MSE. Furthermore, the same experiment was implemented for $n_v = 1$, fixing the learning rate to 10^{-3} and varying the optimizers and loss function in the same way. In this case, each model works with just one variable, and therefore the number of trained models is 11 times higher. These 264 additional models provided conclusions very similar to those obtained in the case of $n_v = 11$. Therefore, the Adamax algorithm with a learning rate of 10^{-3} was used to optimize the MSE loss function in the training phase for all trained models. The full table of results including different metrics (RMSE, nRMSE, MAE) can be found in Prado-Rujas et al. (2023c) for completeness.

Fine-tuning the scaling and interpolation methods: The scaling methods studied were raw (without conversion), standardized (see Eq. (4)) and normalized (see Eq. (5)). Regarding the interpolation methods, nearest-neighbors and linear interpolation were considered on the ROI. These six datasets are available on Zenodo (Prado-Rujas et al., 2023b). As in the previous analysis, the models were trained without using weather nor calendar data for 100 epochs. Besides, $n_x = 4$, $s = 1$, $n_y = 4$, and the optimizer, loss function, and learning rate were fixed to Adamax, MSE, and 10^{-3} , as previously discussed. In the case of $n_v = 1$, 66 models were trained since each gas is modeled separately. The results show that normalization is the worst scaling method of the three studied. Differently, standardizing is the best and most consistent. It scores the smallest nRMSE most of the time (or follows raw scaling very closely in other cases). Taking into account the standardized scaling, nearest-neighbors interpolation behaves better in almost every case. However, looking closer, it can be seen that this improvement is small. Specifically, it is 1.8% better on average over variables and horizons, having minimum and maximum values of -2.2% and 8.4% , respectively. Therefore, the best combination of scaling and interpolation methods for $n_v = 1$ is standardization with nearest-neighbors interpolation. The nRMSE scored by the ST-AQF model with these parameters are shown in the upper part of Table 3. Regarding $n_v = 11$, 6 models were trained, since each of them considers all 11 variables at once. The results were similar for the scaling method, but linear interpolation behaves better in this case. However, again, this improvement is small (2.3% on average, and with minimum and maximum values of -1.1% and 7%). The nRMSE scored by the ST-AQF model with these parameters are shown in the lower part of Table 3. Nearest-neighbors interpolation features some benefits when compared to linear interpolation, because of its nature. These advantages include faster calculation times and fewer memory requirements. Therefore, and considering the small improvement, nearest-neighbors is used over linear interpolation for the remaining of this work. As mentioned above, the complete table of results can be found in Prado-Rujas et al. (2023c).

Summary of the fine-tuning results:

- Adamax algorithm with a learning rate of 10^{-3} is used to optimize the MSE loss function, both for $n_v = 1$ and $n_v = 11$.
- The scaling method that yields the best results is standardization for both n_v values.

Table 3Comparison between $n_v = 1$ and $n_v = 11$, taking for each case the best combination of scaling and interpolation methods.

#	Scaling	Interp.	n_v	Pollutant	Median nRMSE over $\mathcal{C}_v$				Average
					1 h	2 h	3 h	4 h	
1	Standardized	Nearest-neighbors	1	PM _{2.5}	0.454	0.525	0.557	0.582	0.53
2				PM ₁₀	0.422	0.511	0.563	0.6	0.524
3				O ₃	0.234	0.279	0.357	0.43	0.325
4				NO	1.157	1.329	1.579	1.804	1.467
5				NO ₂	0.338	0.433	0.544	0.633	0.487
6				NO _x	0.617	0.725	0.873	1.007	0.806
7				SO ₂	0.217	0.252	0.29	0.312	0.268
8				CO	0.35	0.413	0.488	0.552	0.451
9				TOL	0.922	1.127	1.245	1.329	1.156
10				BEN	0.639	0.777	0.892	0.969	0.819
11				EBE	1.533	1.643	1.717	1.786	1.67
12	Standardized	Linear	11	PM _{2.5}	0.646	0.638	0.642	0.66	0.647
13				PM ₁₀	0.719	0.721	0.731	0.743	0.728
14				O ₃	0.527	0.537	0.555	0.57	0.547
15				NO	1.543	1.574	1.712	1.863	1.673
16				NO ₂	0.478	0.529	0.601	0.661	0.567
17				NO _x	0.745	0.824	0.942	1.043	0.888
18				SO ₂	0.489	0.489	0.5	0.513	0.498
19				CO	0.474	0.496	0.537	0.586	0.523
20				TOL	1.312	1.29	1.306	1.357	1.316
21				BEN	1.148	1.139	1.163	1.182	1.158
22				EBE	2.151	2.15	2.15	2.151	2.15

The training time is almost 1 h for all models with $n_v = 1$ and 2 h 20 min for the one model with $n_v = 11$ (for 100 epochs in all cases). The nRMSE per forecast horizon is obtained as the median across the available geolocations for each variable, and the average of these median values is displayed in the last column of the table.

- The nearest-neighbors interpolation method behaves better for $n_v = 1$, whereas linear interpolation is more suited for $n_v = 11$. All in all, nearest-neighbors is chosen over linear interpolation due to its benefits.
- Table 3 shows that the nRMSE for $n_v = 1$ is better than the one for $n_v = 11$, for all variables and horizons. This will be further explored in Section 3.2.3.

3.2.2. Modeling separate pollutants

In this section, the results obtained by the ST-AQF are compared with the baseline models and the literature for a single pollutant (i.e., $n_v = 1$). Additionally, an experiment is carried out to understand the impact of the size of the input window n_x .

Comparison between the ST-AQF and baseline models: As in Section 3.2.1, $n_x = 4$, $s = 1$, $n_y = 4$ are fixed. The ST-AQF model was compared with several baseline models for each pollutant:

- **Persistence:** Simple model that assumes that the forecasted variable remains unchanged. Therefore, its prediction is a copy of the observation for $t = t^* - s \cdot r$. It is a tough contestant for short forecasting horizons.
- **Conv3D:** It is based on the ST-AQF base model, where the ConvLSTM layer is replaced by a 3D-convolution layer.
- **LSTM:** Initially, it flattens the mesh-grid to obtain a vector on every time interval. Then, it is passed by an LSTM layer with 16 neurons, two fully-connected layers, and the output is reshaped back to the mesh-grid form.

Table 4 shows the nRMSE scored for every model and pollutant, highlighting in bold the best result per pollutant and horizon. The median nRMSE across the available geolocations $\mathcal{C}_v$ of each variable v is calculated for every forecast horizon. The last column displays the average of those nRMSE values across horizons. In addition, the training time is displayed for each case (which considers early stopping, as discussed in Section 3.2.1). Table 4 reveals that the nRMSE degrades as the forecast horizon moves forward over time. This can be caused by the accumulation of errors in the prediction, which implies that 4 h ahead forecasts are more unstable than those of 1 h. To complement Table 4, the Almost Stochastic Order (ASO) test (Del Barrio et al., 2018; Dror et al., 2019) was carried out as implemented by Ulmer

et al. (2022). Table 5 presents its results for the daily RMSE scored on the test set (over 700 days for each horizon). The ST-AQF model with $n_v = 1$ is compared with the three baseline models for every pollutant with a confidence level $\alpha = 0.05$. Table 5 shows that the ST-AQF is stochastically dominant over the persistence and LSTM models ($\epsilon_{\min} < 0.5$) in practically every case. The differences are less clear-cut for the Conv3D, where $\epsilon_{\min} < 0.5$ for three pollutants and $\epsilon_{\min} \geq 0.5$ for the other eight. This can be expected since the only difference between the ST-AQF and the Conv3D baseline is the replacement of the ConvLSTM layer by a 3D-convolution. In any case, Table 4 shows that their error metrics are within the same range.

Comparison between the ST-AQF and the literature: Direct comparison of multivariable spatio-temporal models is usually tricky. Differences in data sources, forecast horizons, timestep, and evaluation metrics make it a complex task. In addition, many do not disclose the code or trained models. Aznarte (2017) and García and Aznarte (2020) forecast 1-hour ahead NO₂ measurements in two locations of Madrid, as in this work. Specifically, the best RMSE reported by Aznarte (2017) for station 4 (see Fig. 3) is 8.1% better than the one scored by the ST-AQF model (for the same station and horizon). As for García and Aznarte (2020), the ST-AQF obtains 33.3% better RMSE for station 8 (see Fig. 3). The more recent work of De Medrano et al. (2021) models four pollutants in the city of Madrid (see Section 1.3). However, they do not share the model for intellectual property reasons. Furthermore, they forecast for the next 48 h, and hence a straightforward comparison is not possible. However, it serves as a proxy to show that error metrics are akin. Specifically, the RMSE scored by the ST-AQF is 8.1% better for PM₁₀, and 5.6%, 18.3%, and 7.3% worse for PM_{2.5}, NO₂ and O₃, respectively. The remaining seven air pollutants are not considered in this work.

Comparison between different n_x values: Fig. 4 compares the nRMSE of the same ST-AQF model with increasing values of n_x , for two of the pollutants (NO₂ on the left and O₃ on the right). The impact can be seen on every horizon, obtained as the median of the nRMSE over the available geolocations. The n_x parameter takes values between 1 and 24 h. The results for the other pollutants are similar, obtaining smaller nRMSE for higher values of n_x . Thus, the remaining experiments are carried out with $n_x = 4$ as a compromise between model performance and computational load.

Table 4Comparison between the ST-AQF model and the baseline models (persistence, Conv3D, and LSTM) for $n_v = 1$.

#	Pollutant	Model	Training time	Median nRMSE over G_v				Average
				1 h	2 h	3 h	4 h	
1	PM _{2.5}	ST-AQF	17	0.462	0.521	0.56	0.593	0.534
2		Persistence	–	0.459	0.579	0.632	0.683	0.588
3		Conv3D	31	0.466	0.531	0.562	0.583	0.536
4		LSTM	6	0.484	0.536	0.564	0.602	0.547
5	PM ₁₀	ST-AQF	56	0.435	0.516	0.563	0.601	0.529
6		Persistence	–	0.422	0.553	0.633	0.693	0.575
7		Conv3D	31	0.437	0.52	0.561	0.603	0.53
8		LSTM	8	0.466	0.536	0.585	0.626	0.553
9	O ₃	ST-AQF	56	0.225	0.272	0.358	0.433	0.322
10		Persistence	–	0.221	0.368	0.481	0.568	0.409
11		Conv3D	31	0.245	0.285	0.36	0.435	0.331
12		LSTM	17	0.218	0.299	0.374	0.433	0.331
13	NO	ST-AQF	56	1.166	1.319	1.59	1.819	1.474
14		Persistence	–	1.176	1.766	2.156	2.397	1.874
15		Conv3D	31	1.162	1.337	1.611	1.823	1.483
16		LSTM	19	1.172	1.433	1.621	1.745	1.493
17	NO ₂	ST-AQF	54	0.34	0.423	0.539	0.632	0.484
18		Persistence	–	0.362	0.58	0.73	0.83	0.625
19		Conv3D	31	0.334	0.436	0.549	0.641	0.49
20		LSTM	20	0.339	0.466	0.562	0.626	0.498
21	NO _x	ST-AQF	56	0.612	0.714	0.875	1.009	0.803
22		Persistence	–	0.635	0.972	1.202	1.345	1.038
23		Conv3D	31	0.598	0.729	0.885	1.017	0.807
24		LSTM	20	0.602	0.77	0.89	0.971	0.808
25	SO ₂	ST-AQF	52	0.205	0.249	0.279	0.304	0.259
26		Persistence	–	0.16	0.231	0.278	0.311	0.245
27		Conv3D	10	0.241	0.264	0.292	0.315	0.278
28		LSTM	17	0.329	0.352	0.373	0.397	0.363
29	CO	ST-AQF	54	0.354	0.424	0.5	0.56	0.459
30		Persistence	–	0.329	0.499	0.612	0.685	0.531
31		Conv3D	31	0.34	0.4	0.48	0.54	0.44
32		LSTM	19	0.377	0.465	0.558	0.623	0.506
33	TOL	ST-AQF	56	0.927	1.125	1.247	1.329	1.157
34		Persistence	–	0.96	1.314	1.523	1.642	1.36
35		Conv3D	31	0.971	1.146	1.266	1.341	1.181
36		LSTM	20	1.051	1.183	1.298	1.359	1.223
37	BEN	ST-AQF	55	0.672	0.812	0.919	0.989	0.848
38		Persistence	–	0.63	0.897	1.068	1.172	0.942
39		Conv3D	31	0.66	0.795	0.906	0.981	0.835
40		LSTM	13	0.754	0.883	0.977	1.03	0.911
41	EBE	ST-AQF	54	1.504	1.626	1.692	1.754	1.644
42		Persistence	–	1.687	1.986	2.076	2.172	1.98
43		Conv3D	26	1.528	1.629	1.684	1.75	1.648
44		LSTM	19	1.709	1.817	1.869	1.915	1.828

Training time is displayed in minutes (accounting early stopping). The nRMSE per forecast horizon is obtained as the median across the available geolocations for each variable, and the average of these median values is displayed in the last column of the table.

Table 5ASO test comparing the ST-AQF model to the baseline models (persistence, Conv3D, and LSTM) for $n_v = 1$.

Pollutant	ST-AQF vs. persistence	ST-AQF vs. Conv3D	ST-AQF vs. LSTM
PM _{2.5}	0.011	1.0	0.160
PM ₁₀	0.048	0.898	0.030
O ₃	0.006	0.603	0.068
NO	0.004	0.684	0.096
NO ₂	0.001	0.755	0.275
NO _x	0.002	0.562	0.310
SO ₂	1.0	0.006	0.0
CO	0.016	0.733	0.001
TOL	0.086	0.343	0.017
BEN	0.230	1.0	0.000
EBE	0.982	0.204	0.002

Bold format is applied when $\epsilon_{\min} < 0.5$, which expresses an upper bound to the amount of violation of stochastic order (see Ulmer et al., 2022). The experiment is available at Prado-Rujas et al. (2023c).

3.2.3. Multivariable modeling

Section 3.2.2 presented the results obtained by the ST-AQF, together with a comparison of different input sizes (i.e., different n_x). In all cases, the models were trained for a single input and target variable, that is, $n_v = 1$. In this section, an experiment is carried out to study how the ST-AQF model behaves when progressively incorporating variables. This characteristic is known as extensibility in this context (Prado-Rujas et al., 2021b). To this end, two variables are chosen and several ST-AQF models are trained, while progressively including more variables. Initially, two models were trained for NO₂ and PM₁₀ (i.e., $n_v = 1$), and they were combined for the case $n_v = 2$. Afterward, SO₂ and CO were incorporated ($n_v = 4$), then NO and PM_{2.5} ($n_v = 6$), followed by NO_x and O₃ ($n_v = 8$), and finally the remaining ones ($n_v = 11$). The aim is to study whether the model benefits from having multiple variables at once and if there are differences between distinct values of n_v .

Evaluation for several n_v : Fig. 5 presents the results of this experiment for the pollutants studied: NO₂ (Fig. 5(a)) and PM₁₀ (Fig. 5(b)). The graphs represent the nRMSE scored by every model on each forecast horizon. For both pollutants, the model with $n_v = 1$ scores

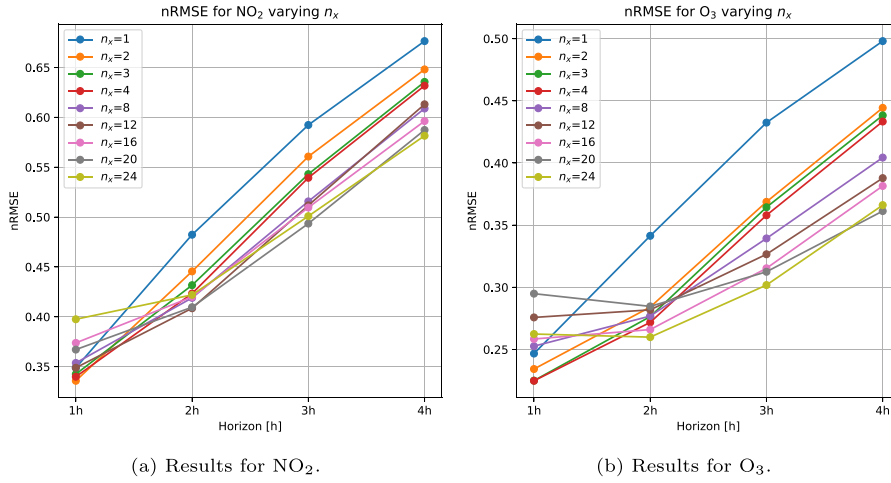


Fig. 4. Comparison of the nRMSE scored by the ST-AQF for two different variables, with $n_x \in \{1, 2, 3, 4, 8, 12, 16, 20, 24\}$, i.e., input windows between 1 and 24 h. The same experiment for the remaining pollutants can be found in Prado-Rujas et al. (2023c).

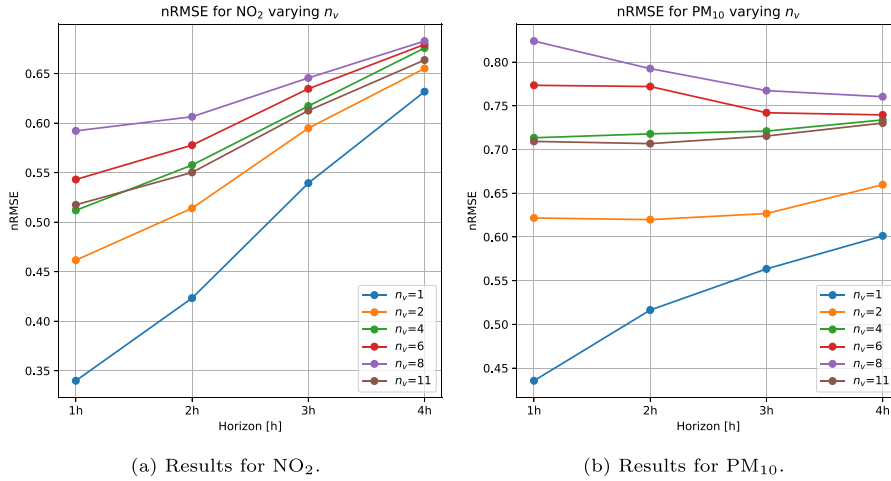


Fig. 5. Comparison of the nRMSE scored by the ST-AQF for two different variables, with $n_v \in \{1, 2, 4, 6, 8, 11\}$, i.e., varying the number of variables considered in the input and output.

the lowest error, which increases for every consecutive value of n_v , on all horizons. However, it can be seen that the case $n_v = 11$ stands between $n_v = 2$ and $n_v = 4$, being very close to the latter. Admittedly, this experiment merely serves as a proxy of the full-length one (all combinations of incorporating variables one at a time). Considering a single pollutant, the total number of models to explore would be the factorial of 10, more than 3.5 million cases. Furthermore, this figure increases to almost 40 million when considering the 11 pollutants. Fig. 5 presents a small subset of this experiment. However, it indicates that a careful combination of multiple variables could improve the forecasting results of separate models. Nevertheless, sensitivity analysis should be conducted to understand whether some pollutants lower the prediction error of others.

Fine-grained evaluation: Additionally, the ST-AQF with $n_v = 1$ and $n_v = 11$ are compared at the sensor and horizon levels in Fig. 6. As before, Fig. 6(a) refers to NO₂ and Fig. 6(b) to PM₁₀. The graphs show the improvement for the case $n_v = 1$ when compared to $n_v = 11$. This improvement is calculated as a percentage by means of the skill, see Eq. (6):

$$\text{Skill} = \left(1 - \frac{\text{error}_{\text{proposed}}}{\text{error}_{\text{reference}}} \right) \cdot 100 \quad (6)$$

The skill is calculated taking as reference the case $n_v = 11$ and $n_v = 1$ as proposed, with the nRMSE as error metric. The figure shows that the skill is greater than 30% or even 40% for the 1-hour ahead forecasts and that it reduces for further horizons. Of the 148 cases plotted in Fig. 6, only once (sensor 58 for 4-hours ahead from Fig. 6(a)) the skill is negative with a value of -1.3% . However, there are meaningful differences between sensors. These variations can be caused by the specific location or weather particularities (i.e., being close to a park, being more or less exposed to wind, etc.).

Further horizons: Table 3 from Section 3.2.1 presented the nRMSE of ST-AQF for $n_v \in \{1, 11\}$, taking in each case the best combination of scaling and interpolation methods. A different point of view is to compare the nRMSE of the two n_v values for each pollutant and horizon. The table highlights that a model with $n_v = 1$ outperforms the results obtained by a model with $n_v = 11$ for every pollutant. This is expected since the model architecture is unchanged, but the number of variables to predict is 11 times higher. However, Table 3 also shows that the improvement is monotonically decreasing over the forecast horizons for all variables. Therefore, this leads to the assumption that models with $n_v = 1$ and $n_v = 11$ should score comparable nRMSE figures for horizons further than 4 h.

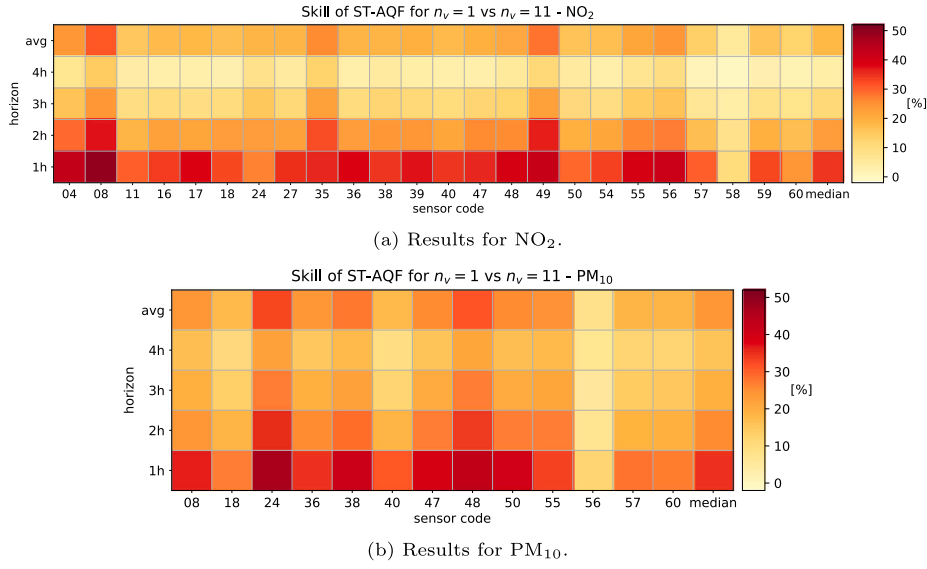


Fig. 6. Comparison of the ST-AQF with $n_v = 1$ against $n_v = 11$, expressed as a percentage of improvement of the nRMSE. The results are expressed at sensor and horizon levels, including the median per horizon across sensors and the average per sensor across horizons. The same experiment for the remaining pollutants can be found in [Prado-Rujas et al. \(2023c\)](#).

3.2.4. Ablation study

Section 2.2.2 described four variations of the ST-AQF model based on exogenous input data. This section presents experiments to explore the differences between these models, namely: ST-AQF, ST-AQFⁱ, ST-AQF^w, and ST-AQF* (see Section 2.2.2 for reference). Fig. 7 shows a comparison between the four cases for four pollutants: NO₂, O₃, PM_{2.5} and SO₂. In all cases, the parameters are $n_x = 4$, $n_y = 4$, $s = 1$. Besides, n_v is set to one to isolate the effects of incorporating exogenous data.

Regarding NO₂ (see Fig. 7(a)), the ST-AQF* and ST-AQF^w obtain a smaller error than its peers for forecasts of 2, 3 and 4 h. Specifically, it achieves a skill of more than 6.5% for the third horizon, compared to ST-AQF (see Eq. (6)). It seems that the temporal module does not provide much useful information for the ST-AQF or the ST-AQF^w. However, the weather module is improving the results of ST-AQF and ST-AQFⁱ on most horizons. The case of O₃ (Fig. 7(b)) is similar, but it has the particularity that ST-AQF* reaches a skill of almost 15% compared to the ST-AQF on the third horizon.

The results for PM_{2.5} (Fig. 7(c)) are similar to the previous cases, but also improve in the 1-hour ahead forecast. The skills of the ST-AQF^w and ST-AQF* models are close to or above 5% for the three further horizons, compared to ST-AQF. Finally, the SO₂ presents a more stable behavior. In this case, the temporal module is providing some information to the ST-AQF* model, as Fig. 7(d) shows. All in all, it seems that the choice of model should be based on the variable under study.

3.2.5. A variable number of sensors

Once the air quality sensors are operational, they will start streaming a continuous flow of measurements to a central server. Unavoidably, some of those measurements will be invalid or empty due to hardware breakdowns or external factors. In this vein, *robustness* is defined as the ability to recover from sensor failure and continue producing reliable predictions. In addition, some of these sensors may become obsolete and be removed from the network, and some additional ones may be installed. In this scenario, *flexibility* is defined as the ability of the forecasting framework to accept a different number of inputs and outputs without changing its structure. These concepts are based on previous work (Prado-Rujas et al., 2021a,b, 2023a). This section seeks to study the robustness and flexibility properties of the ST-AQF model.

The robustness feature was evaluated by simulating sensor failure and calculating how the error of the test set degrades when an increasing number of sensors fail. The experiment was carried out for individual pollutants ($n_v = 1$) to isolate the effects of missing information. Besides, the usual parameters were fixed: $n_x = 4$, $n_y = 4$, $s = 1$. The experiment was designed as follows:

1. Choose a variable v that is observed by a maximum of n_s sensors and load the ST-AQF model trained on v .
2. Fix the number of faulty sensors: $f \in \{1, \dots, \lceil n_s/2 \rceil\}$.
3. Choose f random sensors from the available ones n_s .
4. Calculate the RMSE from the entire testing period excluding those f sensors:
 - (a) In each time interval $t^* \in \mathcal{T}_{\text{test}}$, load the $n_s - f$ available sensors.
 - (b) Build the mesh-grid with data from those sensors.
 - (c) Feed it to the ST-AQF, obtaining the output mesh-grid.
 - (d) Undo this mesh-grid to obtain predictions for the n_s sensors.
 - (e) Calculate the squared error of t^* : $\text{SE}^{t^*} = (y_{g,v}^{t^*+h} - p_{g,v}^{t^*+h})^2$ (see Section 2.3 for the notation).
 - (f) Once all time intervals have been completed, obtain the RMSE per horizon and sensor (see Eq. (1)):

$$\sqrt{\frac{1}{|\mathcal{T}_{\text{test}}|} \sum_{t^* \in \mathcal{T}_{\text{test}}} \text{SE}^{t^*}}.$$
5. Go to step 3 until the r repetitions are completed.
6. Go to step 2 until the $\lceil n_s/2 \rceil$ cases are completed, then finish.

The result of the experiment is an array of shape $(\lceil n_s/2 \rceil, r, n_y, n_s)$. To compare it to the original RMSE, the skill (see Eq. (6)) is applied on every horizon and sensor, for each simulation.

Fig. 8 shows the results of the experiment for NO₂ and PM_{2.5}. The maximum number of sensors n_s is 24 and 7, respectively, and the number of repetitions r was fixed at 20. To present the results, the skill was averaged across repetitions, and then the median was calculated across the sensors. Fig. 8(a) shows the results for NO₂. It can be seen that RMSE degrades, on average, less than 15% on the four horizons when 25% (or less) of the sensors are not working. When that number increases to half of them (12 sensors), the worsening stays around

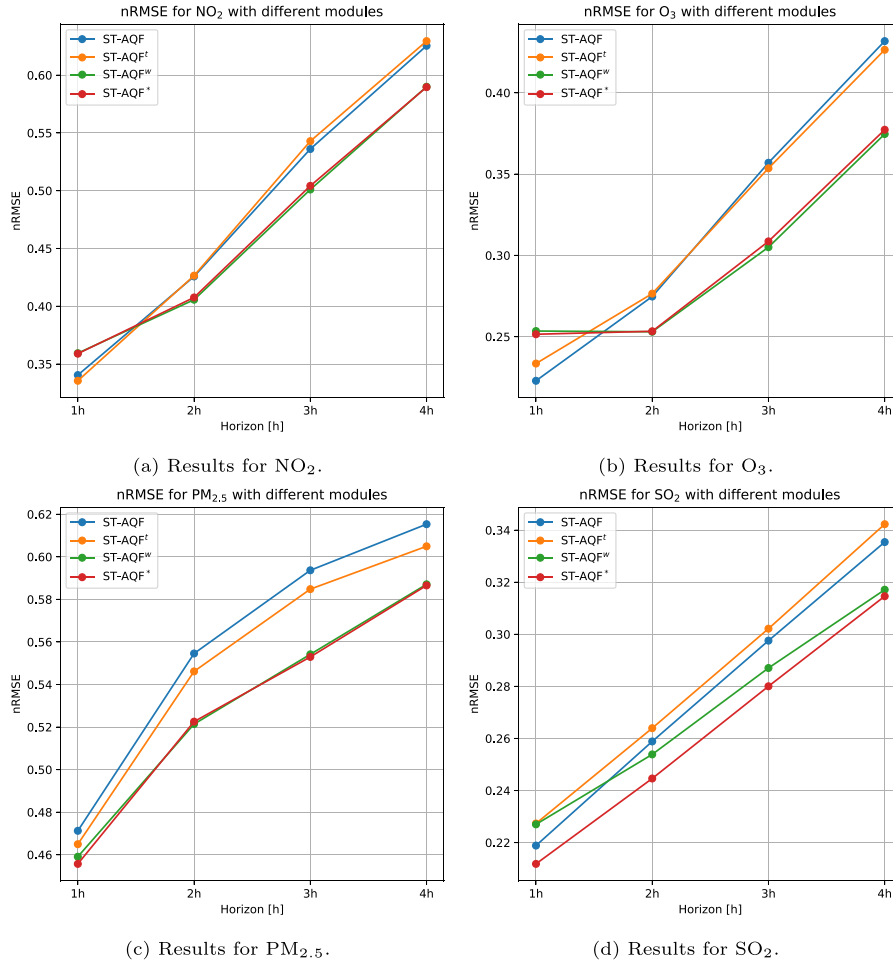


Fig. 7. Comparison of the ST-AQF with different modules: just the air quality module (ST-AQF), including the temporal module (ST-AQF^t), the weather module (ST-AQF^w), or both (ST-AQF^{*}). The same experiment for the remaining pollutants can be found in [Prado-Rujas et al. \(2023c\)](#).

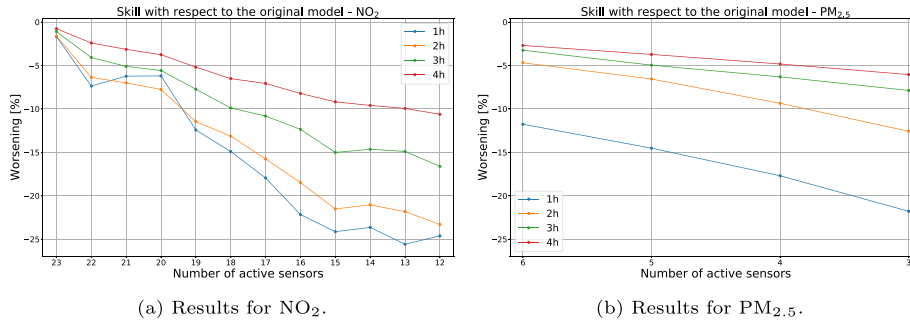


Fig. 8. Robustness test: Evaluation of the skill of the ST-AQF when an increasing number of sensor fail to record valid data. The baseline for the comparison considers that all sensors are active (24 in the case of NO₂ and 7 in the case of PM_{2.5}). The same experiment for the remaining pollutants can be found in [Prado-Rujas et al. \(2023c\)](#).

or below 25%. Similarly, [Fig. 8\(b\)](#) presents the results for PM_{2.5}. The RMSE degradation also stays below 25% when more than half of the sensors do not send information in this case. Having a smaller number of sensors, the experiment for PM_{2.5} presents higher stability than that of NO₂. In both cases the RMSE degradation is almost ordered by the horizon: further horizons are less affected by missing data than the closest ones. Interestingly, the 1-hour ahead forecast degrades slightly less than the 2-hour one for 3 and 4 missing sensors in the case of NO₂.

This could be caused by the random nature of the experiment, and a higher number of repetitions could smooth it out.

The flexibility is inherent to the ST-AQF model thanks to the integration of the mesh-grid. The time series at sensor level are incorporated into the fixed-size mesh-grid (see [Section 3.1.1](#)). In this way, the model is able to work with a variable number of sensors at both the input and the output. Furthermore, the robustness experiment shows this in practice. In addition, the models in [Section 3.2.2](#) also employed a variable number of sensors for each pollutant when $n_v = 1$, without

changing the structure of the framework. It should be noted that the maximum number of sensors n_s is restricted by their distribution and the shape (r, c) of the mesh-grid. If there is a need to increase n_s beyond that limit, then the spatial granularity σ should be set to a finer value.

4. Discussion

The proposed model for spatio-temporal air quality forecasting was trained for the prediction of up to eleven different pollutants recorded in the city of Madrid (Spain). Furthermore, the model was evaluated from different perspectives to study the beneficial properties it can provide, beyond scoring small error metrics.

Initially, several alternatives were studied in Section 3.2.1 in terms of mesh-grid design, data scaling, and hyperparameter tuning. This allowed many alternatives to be left out and to detect which ones were the most consistent between different pollutants. Thereupon, the Adamax algorithm with a learning rate of 10^{-3} was chosen to optimize the MSE loss function. Additionally, the scaling and interpolation methods were fixed to standardize and nearest-neighbors, respectively.

The nRMSE scored by the ST-AQF was compared to the baselines in Section 3.2.2. These baseline models were the persistence, Convolutional 3D, and LSTM models. The ST-AQF obtained better results than its peers on most pollutants and horizons, as Table 4 showed. Subsequently, Section 3.2.3 studied the ability of the ST-AQF model to incorporate pollutants, i.e. $n_v > 1$. Even if the best results were obtained for $n_v = 1$, the difference decreases as the forecast horizon moves away (see Fig. 6). Therefore, for horizons beyond four hours, the results for $n_v = 1$ and $n_v = 11$ could be comparable.

An ablation study compared the four different variants of the ST-AQF model in Section 3.2.4. The aim of this was to discern which modules improved the forecasts. On the basis of the results, the weather module seems to be more important for air quality forecasting than the calendar one. Furthermore, the model that incorporates all modules obtained the smaller nRMSE for most of the pollutants and horizons studied. These results were shown in Fig. 7, see the ST-AQF*.

In the scenario studied, robustness and flexibility are beneficial characteristics for the operational aspect of such a framework. Section 3.2.5 studied how the error evolves when an increasing number of sensors fail to read air quality data. Up to half of the available sensors were excluded from the input for two gases. This resulted in a degradation of only 25% or less both for NO_2 and $\text{PM}_{2.5}$, as Fig. 8 showed.

Even if the system is flexible and robust, some limitations should be taken into account. The mesh-grids necessary to cover the entire ROI are obtained by interpolation from known data, i.e., real measured data. These estimates of air pollution are reasonable, although they have not been acquired with instruments. The only solution to improve this limitation is to incorporate more sensors into the network. Despite the model's ability to withstand the malfunction of multiple sensors, if too many of them fail, the system as a whole is not feasible. In particular, at least three operational measurements are needed to continue forecasting future air quality. The only solution to avoid this pitfall is to replace the faulty instruments, in order to keep a sufficient number of sensors working.

5. Conclusions and future work

Air quality is of utmost importance for human health, since it directly affects our respiratory system and general well-being. Given this fact, governments, organizations, and individuals are encouraged to:

- take measures to reduce pollution levels,
- promote cleaner energy sources,
- enhance public transportation, and

- implement environmental policies.

Gathering, analyzing, and forecasting the concentration of pollutant gases enable to build these effective policies that can improve air quality. This work presents a framework for forecasting air quality on multiple geolocations and horizons while integrating data sources of different nature. This framework is called Spatio-Temporal Air Quality Forecaster (ST-AQF).

ST-AQF has been trained and tested on an extensive 10-year dataset of the city of Madrid (Spain). This case study showed that ST-AQF can work with multiple pollutants at once with fine spatio-temporal resolutions (~ 500 m and 1 h, respectively). Besides, this Deep Learning (DL)-based model is able to integrate exogenous information (such as weather and temporal features) naturally, improving the results obtained. Furthermore, the experiments indicate that the model presented is robust and flexible. In other words, it is capable of adapting to sensor failure and changes in the sensor network, and even to a variable number of pollutants. All these characteristics facilitate the use of the framework in very different and dynamic scenarios. For example, in situations where new sensors are installed, others fail, and different pollutants are meaningful in different contexts. These are distinctive aspects of the ST-AQF compared to other similar frameworks.

Regarding future work, further sensitivity analysis should be performed to discern whether some pollutants improve the forecasting results of others (see Section 3.2.3). This analysis could be complemented with a parameter-tuning on the ST-AQF model for different values of n_v . Additionally, a study could be conducted to investigate whether the use of mesh-grids in the weather module improves the forecasting results. Uncertainty analysis would also be of interest to improve the feasibility of the proposed framework. Alternative architectures and additional imputation methods (such as DL-based methods, see Ni and Cao, 2022) could also be considered to improve the results. Such approaches include graph-based neural networks (Ge et al., 2021; Jiang et al., 2022) and transformers (Chen et al., 2022; Jamil and Roy, 2023). Furthermore, additional desired characteristics could be investigated, such as spatial or temporal scalability and fine-tuning of a pre-trained model for a different ROI. In addition, the framework could incorporate a traffic forecasting model to benefit from richer information, such as the one from Méndez et al. (2023). On another note, the framework presented could also be integrated into a larger project, such as the Environmental Insights Explorer from Google (2023). The purpose of this project is to provide cities and researchers with data and modeling capabilities to facilitate emission reduction.

Having reliable, adaptive, robust, and flexible air quality forecasting models is highly valuable. However, action must be taken on the basis of their results and historical data analysis to tackle air pollution and the otherwise inevitable climatic change. As Landrigan et al. (2018) discuss, such actions include the deployment of monitoring systems, the identification of pollution sources, the enforcement of pollution reduction, and the development and smooth transition towards clean energy and transportation alternatives.

CRedit authorship contribution statement

Ignacio-Iker Prado-Rujas: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Antonio García-Dopico:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Emilio Serrano:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **M. Luisa Córdoba:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **María S. Pérez:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition, Project administration.

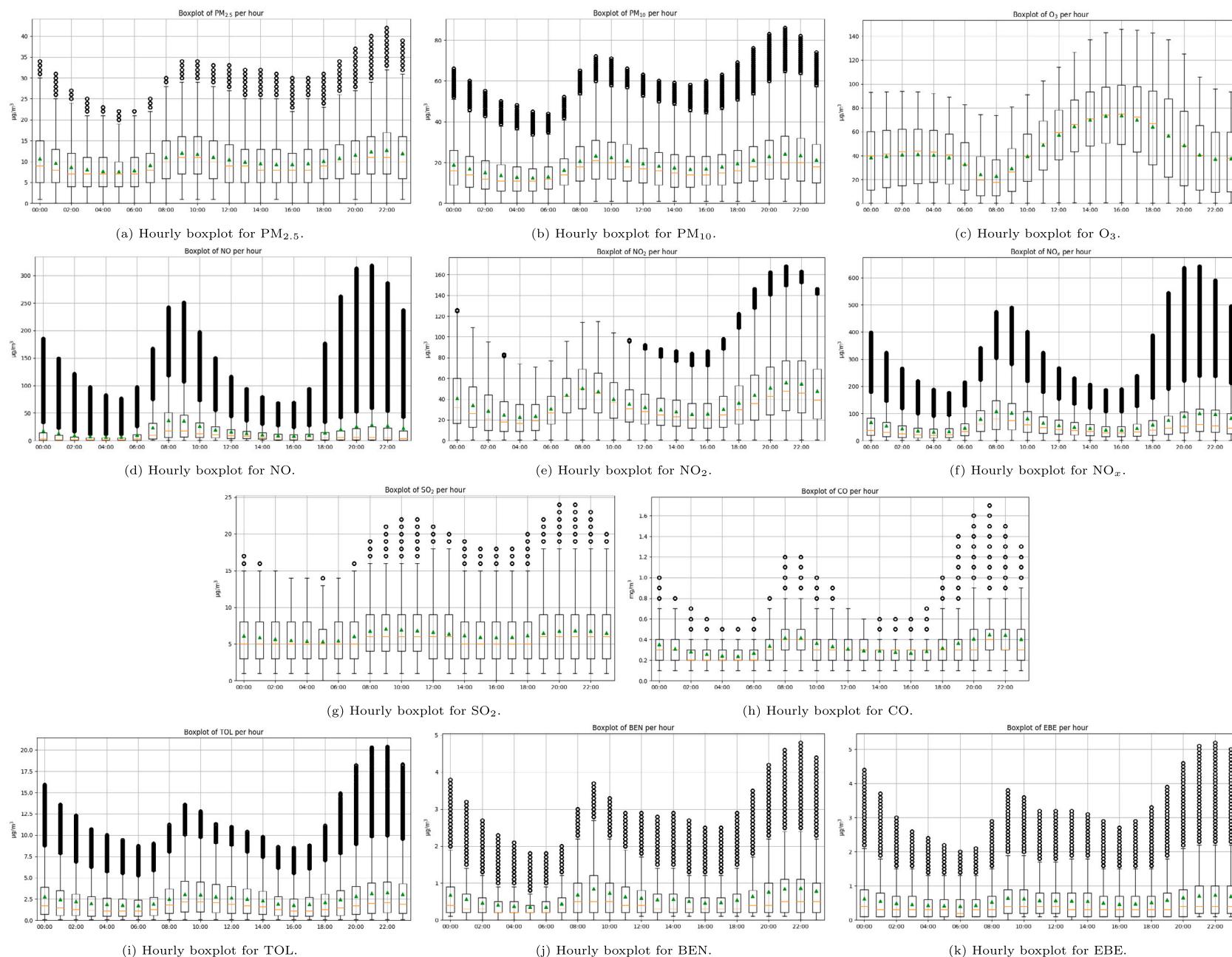


Fig. B.9. Hourly boxplots of each pollutant (2010–2019). The orange lines and green triangles show the median and mean values, respectively. The 2% highest values were removed to facilitate visualization. Whiskers are calculated from the first and third quartiles (Q1 and Q3). Defining the interquartile range as $IQR = Q3 - Q1$, the upper whisker extends to the last datum less than $Q3 + 1.5 \cdot IQR$. Similarly, for the lower whisker: $Q1 - 1.5 \cdot IQR$.

Table B.6
Summary of the pollutant data employed (2010–2019).

#	Pollutant	Units	Number of observations	Active sensors per hour			Values recorded by the sensors						
				min	median	max	mean	std	min	25%	50%	75%	max
1	PM _{2.5}	µg/m ³	497512	1	6	7	10.1	6.41	0.0	5.0	9.0	14.0	32.0
2	PM ₁₀	µg/m ³	1012481	1	12	13	18.62	12.95	0.0	9.0	16.0	25.0	66.0
3	O ₃	µg/m ³	1187574	10	14	14	46.83	31.19	0.6	18.96	46.27	69.65	123.7
4	NO	µg/m ³	2037902	17	24	24	17.12	28.41	0.0	2.0	6.0	18.0	178.0
5	NO ₂	µg/m ³	2037535	17	24	24	36.71	25.73	1.0	16.0	31.0	52.0	119.0
6	NO _x	µg/m ³	2037818	17	24	24	63.23	65.01	2.0	20.0	40.0	81.0	380.0
7	SO ₂	µg/m ³	847376	6	10	10	6.25	3.83	0.0	3.0	5.0	8.0	19.0
8	CO	mg/m ³	846442	6	10	10	0.33	0.17	0.1	0.2	0.3	0.4	1.0
9	TOL	µg/m ³	493950	2	6	6	2.41	2.39	0.1	0.7	1.5	3.5	12.5
10	BEN	µg/m ³	495486	2	6	6	0.59	0.61	0.0	0.2	0.4	0.7	3.2
11	EBE	µg/m ³	494176	2	6	6	0.55	0.58	0.0	0.1	0.3	0.8	3.4

The abbreviations are as follows. Min: minimum, max: maximum, std: standard deviation, 25%: 25th percentile, 50%: 50th percentile (median), 75%: 75th percentile.

Table B.7
Summary of the weather data employed (2010–2019).

#	Variable	Units	Number of observations	Active sensors per hour			Values recorded by the sensors						
				min	median	max	mean	std	min	25%	50%	75%	max
1	Temperature	1/10 °C	514606	3	6	6	151.92	88.97	−90.0	82.0	140.0	218.0	419.0
2	Precipitation	1/10 mm	508128	4	6	6	0.43	3.61	0.0	0.0	0.0	0.0	430.0
3	Wind-x	km/h	478661	3	5	6	−0.1	7.24	−59.0	−3.5	0.0	3.8	58.0
4	Wind-y	km/h	478661	3	5	6	−1.44	8.13	−59.2	−4.8	0.0	3.5	74.8
5	Humidity	%	85818	4	6	6	56.84	23.09	5.0	38.0	55.0	76.0	100.0

The abbreviations are as follows. Min: minimum, max: maximum, std: standard deviation, 25%: 25th percentile, 50%: 50th percentile (median), 75%: 75th percentile. For the minimum, median and maximum number of humidity sensors, only the available hours in the day are considered (00:00, 07:00, 13:00, 18:00), see Section 3.1.3.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work has been supported by the Autonomous Region of Madrid (Spain) through the program CABAHLA-CM (GA No. P2018/TCS-4423). This work is also funded in part by the Euratom research and training program 2019–2020 ENTENTE (GA No. 900018). The authors would also like to thank the Agencia Estatal de Meteorología (AEMET) for kindly providing the meteorological data, and Javier Campoy for his support and ideas during the development of this work.

Appendix A. Complete results, datasets and code

The complete set of graphs, the code to produce the datasets, and the code to reproduce the experiments are available on GitHub (Prado-Rujas et al., 2023c). Additionally, the air quality datasets are available on Zenodo (Prado-Rujas et al., 2023b).

Appendix B. Pollutants and weather variables

This section presents an overview of the pollutants (Table B.6), the weather variables used (Table B.7), and the hourly boxplots of the pollutant data (Fig. B.9) for the studied period 2010–2019. Outliers (the 2% highest measurements) were removed from the pollutant analysis (Table B.6 and Fig. B.9) to obtain a clearer description of the data.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from <https://tensorflow.org>. (Accessed 22 July 2023).

Abirami, S., Chitra, P., 2021. Regional air quality forecasting using spatiotemporal deep learning. *J. Clean. Prod.* 283, 125341. <http://dx.doi.org/10.1016/j.jclepro.2020.125341>.

Agrawal, S., Barrington, L., Bromberg, C., Burge, J., Gazen, C., Hickey, J., 2019. Machine Learning for Precipitation Nowcasting from Radar Images. <http://dx.doi.org/10.48550/arXiv.1912.12132>, arXiv preprint.

Aznarte, J.L., 2017. Probabilistic forecasting for extreme NO2 pollution episodes. *Environ. Pollut.* 229, 321–328. <http://dx.doi.org/10.1016/j.envpol.2017.05.079>.

Berman, J.D., Ebisu, K., 2020. Changes in U.S. air pollution during the COVID-19 pandemic. *Sci. Total Environ.* 739, 139864. <http://dx.doi.org/10.1016/j.scitotenv.2020.139864>.

Chen, Y., Chen, X., Xu, A., Sun, Q., Peng, X., 2022. A hybrid CNN-Transformer model for ozone concentration prediction. *Air Quality, Atmosphere & Health* 15 (9), 1533–1546. <http://dx.doi.org/10.1007/s11869-022-01197-w>.

Chollet, F., et al., 2015. Keras. Software available from <https://keras.io>. (Accessed 22 July 2023).

Comunidad de Madrid, 2023. Regional holidays - Open data portal of the region of madrid. Type: dataset, https://datos.comunidad.madrid/catalogo/dataset/festivos_regionales_locales_historico. (Accessed 22 July 2023).

De Medrano, R., de Buen Remiro, V., Aznarte, J.L., 2021. SOCAIRE: Forecasting and monitoring urban air quality in Madrid. *Environ. Model. Softw.* 143, 105084. <http://dx.doi.org/10.1016/j.envsoft.2021.105084>.

Del Barrio, E., Cuesta-Albertos, J.A., Matrán, C., 2018. An optimal transportation approach for assessing almost stochastic order. In: *The Mathematics of the Uncertain*. Springer, pp. 33–44. http://dx.doi.org/10.1007/978-3-319-73848-2_3.

Devasekhar, V., Natarajan, P., 2023. Prediction of air quality and pollution using statistical methods and machine learning techniques. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* 14 (4), <http://dx.doi.org/10.14569/IJACSA.2023.01404103>.

Directorate-General for Environment, 2008. Air quality standards. https://environment.ec.europa.eu/topics/air/air-quality/eu-air-quality-standards_en. (Accessed 22 July 2023).

Domingo, J.L., Rovira, J., 2020. Effects of air pollutants on the transmission and severity of respiratory viral infections. *Environ. Res.* 187, 109650. <http://dx.doi.org/10.1016/j.envres.2020.109650>.

- Drewil, G.I., Al-Bahadili, R.J., 2022. Air pollution prediction using LSTM deep learning and metaheuristics algorithms. *Measurement: Sensors* 24, 100546. <http://dx.doi.org/10.1016/j.measen.2022.100546>.
- Dror, R., Shlomov, S., Reichart, R., 2019. Deep dominance - How to properly compare deep neural models. In: Korhonen, A., Traum, D.R., Màrquez, L. (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Association for Computational Linguistics, pp. 2773–2785. <http://dx.doi.org/10.18653/v1/p19-1266>.
- Dumoulin, V., Visin, F., 2018. A guide to convolution arithmetic for deep learning. <http://dx.doi.org/10.48550/arXiv.1603.07285>, arXiv preprint.
- Dutheil, F., Baker, J.S., Navel, V., 2020. COVID-19 as a factor influencing air pollution? *Environ. Pollut.* 263, 114466. <http://dx.doi.org/10.1016/j.envpol.2020.114466>.
- Environmental Protection Administration in Taiwan, 2022. Taiwan air quality monitoring network. Type: dataset, https://airtw.epa.gov.tw/CHT/Query/His_Data.aspx. (Accessed 22 July 2023).
- European Environment Agency, 2020. Air quality and COVID-19. https://www.eea.europa.eu/ds_resolveuid/6RV1E9C8JD. (Accessed 22 July 2023).
- European Environment Agency, 2022. Air quality statistics. Type: dataset, https://www.eea.europa.eu/ds_resolveuid/DAS-20-en. (Accessed 22 July 2023).
- European Environment Agency, 2023a. European air quality index. <https://airindex.eea.europa.eu/Map/AQI/Viewer/>. (Accessed 22 July 2023).
- European Environment Agency, 2023b. European city air quality viewer. https://www.eea.europa.eu/ds_resolveuid/DAS-271-en. (Accessed 22 July 2023).
- European Environment Agency, González Ortiz, A., Guerreiro, C., Soares, J., 2020. Air Quality in Europe: 2020 Report. Publications Office of the European Union, LU, <http://dx.doi.org/10.2800/786656>.
- Fairfax, E., Zhu, E., Clinton, N., Maiman, S., Shaikh, A., Macfarlane, W.W., Wheaton, J.M., Ackerstein, D., Corwin, E., 2023. EEAGER: A neural network model for finding beaver complexes in satellite and aerial imagery. *J. Geophys. Res. Biogeosci.* 128 (6), e2022JG007196. <http://dx.doi.org/10.1029/2022JG007196>.
- Fattorini, D., Regoli, F., 2020. Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy. *Environ. Pollut.* 264, 114732. <http://dx.doi.org/10.1016/j.envpol.2020.114732>.
- Gallo, R., Castangia, M., Macii, A., Macii, E., Patti, E., Aliberti, A., 2022. Solar radiation forecasting with deep learning techniques integrating geostationary satellite images. *Eng. Appl. Artif. Intell.* 116, 105493. <http://dx.doi.org/10.1016/j.engappai.2022.105493>.
- García, M.V., Aznarte, J.L., 2020. Shapley additive explanations for NO₂ forecasting. *Ecol. Inform.* 56, 101039. <http://dx.doi.org/10.1016/j.ecoinf.2019.101039>.
- Ge, L., Wu, K., Zeng, Y., Chang, F., Wang, Y., Li, S., 2021. Multi-scale spatiotemporal graph convolution network for air quality prediction. *Appl. Intell.* 51, 3491–3505. <http://dx.doi.org/10.1007/s10489-020-02054-y>.
- Gobierno de Canarias, 2022. Calidad del aire - Portal de datos abiertos. Type: dataset <https://datos.canarias.es/catalogos/general/dataset/calidad-del-aire-de-canarias>. (Accessed 22 July 2023).
- Google, 2023. Environmental insights explorer. <https://insights.sustainability.google>. (Accessed 22 July 2023).
- Holmes, N.S., Morawska, L., 2006. A review of dispersion modelling and its application to the dispersion of particles: An overview of different dispersion models available. *Atmos. Environ.* 40 (30), 5902–5928. <http://dx.doi.org/10.1016/j.atmosenv.2006.06.003>.
- Huang, C.J., Kuo, P.H., 2018. A deep CNN-LSTM model for particulate matter (PM_{2.5}) forecasting in smart cities. *Sensors* 18 (7), 2220. <http://dx.doi.org/10.3390/s18072220>, Number: 7, Publisher: Multidisciplinary Digital Publishing Institute.
- Jamil, S., Roy, A.M., 2023. An efficient and robust Phonocardiography (PCG)-based Valvular Heart Diseases (VHD) detection framework using Vision Transformer (ViT). *Comput. Biol. Med.* 158, 106734. <http://dx.doi.org/10.1016/j.combiomed.2023.106734>.
- Jiang, B., Chen, S., Wang, B., Luo, B., 2022. MGLNN: Semi-supervised learning via multiple graph cooperative learning neural networks. *Neural Netw.* 153, 204–214. <http://dx.doi.org/10.1016/j.neunet.2022.05.024>.
- Kamdem De Teyou, G., 2020. ConvLSTM for spatio-temporal feature extraction in time-series images. In: *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*. URL <https://www.climatechange.ai/papers/neurips2020/12>. (Accessed 22 July 2023).
- Kim, D., Chen, Z., Zhou, L.F., Huang, S.-X., 2018. Air pollutants and early origins of respiratory diseases. *Chronic Dis. Transl. Med.* <http://dx.doi.org/10.1016/j.cdtm.2018.03.003>.
- Landrigan, P.J., Fuller, R., Acosta, N.J.R., Adeyi, O., Arnold, R., Basu, N.N., Baldé, A.B., Bertollini, R., Bose-O'Reilly, S., Boufford, J.I., Breyse, P.N., Chiles, T., Mahidol, C., Coll-Seck, A.M., Cropper, M.L., Fobil, J., Fuster, V., Greenstone, M., Haines, A., Hanrahan, D., Hunter, D., Khare, M., Krupnick, A., Lanphear, B., Lohani, B., Martin, K., Mathiasen, K.V., McTeer, M.A., Murray, C.J.L., Ndahimananjara, J.D., Perera, F., Potočnik, J., Preker, A.S., Ramesh, J., Rockström, J., Salinas, C., Samson, L.D., Sandilya, K., Sly, P.D., Smith, K.R., Steiner, A., Stewart, R.B., Suk, W.A., Schayck, O.C.P.v., Yadama, G.N., Yumkella, K., Zhong, M., 2018. The Lancet Commission on pollution and health. *Lancet* 391 (10119), 462–512. [http://dx.doi.org/10.1016/S0140-6736\(17\)32345-0](http://dx.doi.org/10.1016/S0140-6736(17)32345-0), Publisher: Elsevier.
- Liu, C.C., Lin, T.C., Yuan, K.Y., Chiueh, P.T., 2022. Spatio-temporal prediction and factor identification of urban air quality using support vector machine. *Urban Clim.* 41, 101055. <http://dx.doi.org/10.1016/j.uclim.2021.101055>.
- Lopez-Gomez, I., McGovern, A., Agrawal, S., Hickey, J., 2023. Global extreme heat forecasting using neural weather models. *Artif. Intell. Earth Syst.* 2, e220035. <http://dx.doi.org/10.1175/AIES-D-22-0035.1>.
- Madrid City Hall, 2023. Calidad del aire. Datos horarios años 2001 a 2021 - Portal de datos abiertos. Type: dataset <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbc4b2e4b284f1a5a0/?vgnextoid=f3c0f7d512273410VgnVCM2000000c205a0aRCRD>. (Accessed 22 July 2023).
- Mao, W., Wang, W., Jiao, L., Zhao, S., Liu, A., 2021. Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. *Sustainable Cities Soc.* 65, 102567. <http://dx.doi.org/10.1016/j.scs.2020.102567>.
- Méndez, M., Merayo, M.G., Núñez, M., 2023. Long-term traffic flow forecasting using a hybrid CNN-BiLSTM model. *Eng. Appl. Artif. Intell.* 121, 106041. <http://dx.doi.org/10.1016/j.engappai.2023.106041>.
- Meteosim, 2021. Sistema de pronóstico y de diagnóstico sobre la calidad del aire en andalucía. <https://meteosim.com/meteosim-desarrolla-de-un-sistema-de-pronostico-y-de-diagnostico-sobre-la-calidad-del-aire-en-andalucia/>. (Accessed 22 July 2023).
- Ministerio para la Transición Ecológica y el Reto Demográfico, 2022. Histórico de la calidad del aire en España. Type: dataset https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/evaluacion-datos/historico/calidad_aire.aspx. (Accessed 22 July 2023).
- Ni, Q., Cao, X., 2022. MBGAN: An improved generative adversarial network with multi-head self-attention and bidirectional RNN for time series imputation. *Eng. Appl. Artif. Intell.* 115, 105232. <http://dx.doi.org/10.1016/j.engappai.2022.105232>.
- OpenAQ, 2023. Air quality open data. Type: dataset. Available from: <https://openaq.org> (accessed 22/07/2023).
- Prado-Rujas, I.I., García-Dopico, A., Serrano, E., Pérez, M.S., 2021a. A flexible and robust deep learning-based system for solar irradiance forecasting. *IEEE Access* 9, 12348–12361. <http://dx.doi.org/10.1109/ACCESS.2021.3051839>.
- Prado-Rujas, I.I., Serrano, E., García-Dopico, A., Córdoba, M.L., Pérez, M.S., 2021b. Predicción espacio-temporal: más allá del error/precisión. In: *Congreso Español de Informática 2021. CEDI'21, Málaga, Spain*, <http://dx.doi.org/10.5281/zenodo.5530048>.
- Prado-Rujas, I.I., Serrano, E., García-Dopico, A., Córdoba, M.L., Pérez, M.S., 2023a. Combining heterogeneous data sources for spatio-temporal mobility demand forecasting. *Inf. Fusion* 91, 1–12. <http://dx.doi.org/10.1016/j.inffus.2022.09.028>.
- Prado-Rujas, I.I., Serrano, E., García-Dopico, A., Córdoba, M.L., Pérez, M.S., 2023b. Datasets for a deep learning system for spatio-temporal air quality forecasting. <http://dx.doi.org/10.5281/zenodo.7764528>.
- Prado-Rujas, I.I., Serrano, E., García-Dopico, A., Córdoba, M.L., Pérez, M.S., 2023c. Results, graphs and code for a deep learning system for spatio-temporal air quality forecasting. <https://github.com/iipr/air-quality>. (Accessed 22 July 2023).
- Qi, Y., Li, Q., Karimian, H., Liu, D., 2019. A hybrid model for spatiotemporal forecasting of PM_{2.5} based on graph convolutional neural network and long short-term memory. *Sci. Total Environ.* 664, 1–10. <http://dx.doi.org/10.1016/j.scitotenv.2019.01.333>.
- Roy, A.M., Bhaduri, J., 2023. DenseSPH-YOLOv5: An automated damage detection model based on DenseNet and Swin-Transformer prediction head-enabled YOLOv5 with attention mechanism. *Adv. Eng. Inform.* 56, 102007. <http://dx.doi.org/10.1016/j.aei.2023.102007>.
- Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.k., Woo, W.c., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. NIPS '15*, MIT Press, Cambridge, MA, USA, pp. 802–810, URL <https://dl.acm.org/doi/abs/10.5555/2969239.2969329>. (Accessed 22 July 2023).
- Travaglio, M., Yu, Y., Popovic, R., Selley, L., Leal, N.S., Martins, L.M., 2021. Links between air pollution and COVID-19 in England. *Environ. Pollut.* 268, 115859. <http://dx.doi.org/10.1016/j.envpol.2020.115859>.
- Ulmer, D., Hardmeier, C., Frellsen, J., 2022. deep-significance - Easy and meaningful statistical significance testing in the age of neural networks. arXiv preprint [arXiv: 2204.06815](https://arxiv.org/abs/2204.06815).
- Venter, Z.S., Aunan, K., Chowdhury, S., Lelieveld, J., 2020a. Air pollution changes during COVID-19 lockdowns. <https://nina.earthengine.app/view/lockdown-pollution>. (Accessed 22 July 2023).
- Venter, Z.S., Aunan, K., Chowdhury, S., Lelieveld, J., 2020b. COVID-19 lockdowns cause global air pollution declines. *Proc. Natl. Acad. Sci.* 117 (32), 18984–18990. <http://dx.doi.org/10.1073/pnas.2006853117>, Publisher: National Academy of Sciences, Section: Physical Sciences.
- Wang, Z., Maeda, T., Hayashi, M., Hsiao, L.F., Liu, K.Y., 2001. A nested air quality prediction modeling system for urban and regional scales: Application for high-ozone episode in Taiwan. *Water Air Soil Pollut.* 130 (1), 391–396. <http://dx.doi.org/10.1023/A:1013833217916>.
- Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., Chi, T., 2019. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Sci. Total Environ.* 654, 1091–1099. <http://dx.doi.org/10.1016/j.scitotenv.2018.11.086>.

- World Health Organization, 2021. WHO Global Air Quality Guidelines: Particulate Matter (PM_{2.5} and PM₁₀), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization, URL <https://apps.who.int/iris/handle/10665/345329>. (Accessed 22 July 2023).
- World Health Organization, 2022. Ambient (outdoor) air pollution. [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health). (Accessed 22 July 2023).
- World Health Organization, 2023a. Ambient air pollution attributable deaths. <https://www.who.int/data/gho/data/indicators/indicator-details/GHO/ambient-air-pollution-attributable-deaths>. (Accessed 22 July 2023).
- World Health Organization, 2023b. Annual mean PM_{2.5} concentration in urban areas. [https://www.who.int/data/gho/data/indicators/indicator-details/GHO/concentrations-of-fine-particulate-matter-\(pm2-5\)](https://www.who.int/data/gho/data/indicators/indicator-details/GHO/concentrations-of-fine-particulate-matter-(pm2-5)). (Accessed 22 July 2023).
- Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R., 2010. Deconvolutional networks. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2528–2535. <http://dx.doi.org/10.1109/CVPR.2010.5539957>.
- Zhou, Y., Chang, F.J., Chang, L.C., Kao, I.F., Wang, Y.S., 2019. Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. *J. Clean. Prod.* 209, 134–145. <http://dx.doi.org/10.1016/j.jclepro.2018.10.243>.
- Zlatev, Z., Dimov, I., 2006. *Computational and Numerical Challenges in Environmental Modelling*. Elsevier, ISBN: 978-0-444-52209-2.