IFAC Journal of Systems and Control 31 (2025) 100298

Contents lists available at ScienceDirect

IFAC Journal of Systems and Control

journal homepage: www.elsevier.com/locate/ifacsc

Full length article

Local vs regional neural air pollution forecasting models

Matteo Sangiorgio *, Giorgio Guariso

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Via Ponzio 34/5, I-20133, Milan, Italy

ARTICLE INFO

Article history:
Received 25 November 2024
Received in revised form 24 January 2025
Accepted 10 February 2025
Available online 13 February 2025

Dataset link: www.arpa.piemonte.it, www.arpalombardia.it, www.arpa.veneto.it, www.arpae.it, www.it.windfinder.com

Keywords:
Air pollution forecasting
Nonlinear dynamics
LSTM neural networks
Multi-step prediction
Ozone
Nitrogen dioxide

ABSTRACT

Selecting a suitable dataset to develop a data-based forecasting model is often problematic. This is particularly important in the case of air pollution, where concentration measures are scattered over large areas. On the one hand, the classical approach creates a single-station (local) forecasting model using only the data collected at the same station. This guarantees a training dataset that considers all the site's specific characteristics. On the other hand, these data may be limited and not sufficient to develop a robust predictor. Thus, one may use data from other stations to complement the dataset or develop a unique model considering all the data available within a region/domain. While this approach may be prone to filtering high variations, it may consider information on peculiar episodes that have not occurred in the past to a specific station. This paper discusses the topic of air pollution forecasting using the example of several stations in the Padana Plain, Northern Italy. Local forecasting models are developed using LSTM neural networks for nitrogen dioxide and ozone and hourly data from 2010 to 2023 and then compared with regional models. All these models perform extremely well under various regression-based and classification-based performance indicators, except for a few sites with peculiar characteristics that can be considered at the border of the information domain.

1. Introduction

Air pollution is a global problem that exacerbates respiratory, cardiovascular, and cancer illnesses, reducing the expected life duration of millions of individuals. Forecasting air pollution concentrations is essential to warn sensitive populations, avoid prolonged open-air activities, and plan movements that reduce citizens' exposition (Bai, Li, Wang, Xie, & Li, 2016; Wang, Xu, Zhang, & Dong, 2022; Yang, Zhu, Li, & Li, 2020). Among the most common pollutants worldwide are nitrogen dioxide (NO_2) and ozone (O_3), whose ambient concentrations derive from two quite different processes. NO_2 is a primary pollutant, meaning directly released into the atmosphere and thus its concentration closely depends on the local emissions, mainly produced by thermal engines and some agricultural activities. The meteorological conditions and the emissions close to the measurement site thus determine the evolution of its concentration. On the contrary, ozone is a secondary pollutant. It is not directly emitted by any source and forms in the atmosphere due to a complex series of physical and chemical processes. Its formation requires the presence of precursor gases, in particular, nitrogen oxides (NO_X) and volatile organic compounds (VOC) and is catalyzed by ultraviolet radiation (Agirre-Basurko, Ibarra-Berastegi, & Madariaga, 2006; Chen, Islam, & Biswas, 1998; Lin, Trainer, & Liu, 1988; Singh, Carnevale, Finzi, Pisoni, & Volta, 2011; Zolghadri, Monsion, Henry, Marchionini, & Petrique, 2004). VOCs are mainly emitted by industrial activities and, in particular, by solvents (Finlayson-Pitts & Pitts, 1993; Wang, Chen, Hu, Zhang, & Ying, 2018). The ozone dynamic has several implications: high concentrations are only possible in the daytime and are especially high in summer (when solar radiation is higher); it may appear kilometers away from the precursor's sources, and its presence depends on the balance between the precursor concentrations (Lelieveld & Dentener, 2000). Like all other air pollutants, NO_2 and O_3 concentrations are periodic, with daily and yearly periods. They are due to the periodicity of meteorological conditions and human activities, particularly for NO_2. Both NO_2 and O_3 may have a relevant impact on human health (Huang et al., 2021; Latza, Gerdes, & Baur, 2009; Paoletti, De Marco, & Racalbuto, 2007; Stowell, min Kim, Gao, Fu, Chang, & Liu, 2017; US EPA, 2020; Zhang, Wei, & Fang, 2019). Indeed, in 2021 (World Health Organization, 2021), the World Health Organization (WHO) updated its NO_2 guidelines, decreasing the recommended annual average from 40 to 10 $\mu g/m^3$ and the daily average from 200 down to 25 $\mu g/m^3$. Similarly, the new recommendation for ozone is an average daily maximum 8-h mean in the six consecutive months with the highest running-average concentration (i.e., late spring, summer, and beginning of autumn) below 60 $\mu g/m^3$ compared to the 2006 indication of 100 $\mu g/m^3$ (World Health Organization, 2021).

Simulating the concentration of pollutants, especially ozone, is a complex task that can be performed by one of the several process-based chemical transport models currently available

* Corresponding author.
E-mail address: matteo.sangiorgio@polimi.it (M. Sangiorgio).

https://doi.org/10.1016/j.ifacsc.2025.100298
2468-6018/

(e.g., CMAQ, CALLGRID, TCAM, and CAMx). These models require an extensive dataset, including hourly wind fields over the entire domain, as the concentration at a specific location depends on the emissions and meteorological conditions over a large area. However, the need for detailed data on emissions and meteorological conditions makes these models practically unsuitable for forecasting pollutant concentrations, except for some specific areas and episodes. For this reason, data-driven models have been the preferred approach to air quality forecasting in the last four decades. Initially, they were only linear (ARMAX type and related enhancements) (Bolzern, Finzi, Fronza, Spirito, et al., 1982; Salcedo, Ferraz, Alves, & Martins, 1999), but they became nonlinear with the availability of large observational datasets from punctual stations and the development of machine learning tools (Masood & Ahmad, 2021; Méndez, Merayo, & Núñez, 2023; Pisoni, Farina, Carnevale, & Piroddi, 2009). Specifically, artificial neural networks were unanimously recognized as one of the best tools for addressing prediction tasks in general (Amaranto & Mazzoleni, 2023; Lim & Zohren, 2021; Nielsen, 2019; Tealab, 2018) and air pollution forecasting in particular (Agirre-Basurko et al., 2006; Cabaneros, Calautit, & Hughes, 2019; Ma et al., 2022; Maciąg, Kasabov, Kryszkiewicz, & Bembenik, 2019; Sangiorgio & Guariso, 2023), and are used in this study to examine the performance of a regional model against those of the set of local models, each trained for a specific site. The classical local approach guarantees that the model is identified considering all the site's specific characteristics, i.e., it works within a well-defined information domain. On the other hand, local data may be limited and insufficient to develop a robust predictor. Thus, one may merge data from different stations to create a unique model that considers all the data available in a specific region. While this approach may be prone to filtering high variations, it allows all stations to retain information about critical situations not experienced in the past by each. Recent studies in the domain of streamflow forecasting resulted in regional models performing better in specific locations than models developed with local data (Kratzert, Gauch, Klotz, & Nearing, 2024). The possibility of developing a reliable regional model also allows us to start issuing the forecast on new sites where past data for training a model have yet to be available. This may be interpreted as a special case of transfer learning, i.e., identifying a machine learning model for a data-rich site and using it, without retraining, to another (typically data-scarce) site. This approach has generated an impressive number of papers in the last few years (see, for instance, Fong, Li, Fong, Wong, & Tallon-Ballesteros, 2020; Gupta et al., 2024; Ma et al., 2020, 2022; Yang, Ismail, Li, Zhang, & Fadzli, 2023). However, the source and target domains differ in the classical definition of transfer learning (e.g. Weiss, Khoshgoftaar, & Wang, 2016). This is not precisely the case at hand. We assume that the knowledge domain of all the sites is the same, and the datasets of the various stations are samples (instances) of such a domain. Thus, the paper's purpose is twofold: On one side, we analyze how to efficiently extract the general domain characteristics from the available samples. On the other side, we try to identify the borders of the considered domain. Since such borders are evidently fuzzy, this corresponds to evaluating how the performance of the predictors worsens when moving away from the key domain features. This study explores the differences between local and regional air pollution forecasting models using NO_2 and O_3 measurement sites, all located in Northern Italy's plain, a region with a population of nearly 15 million inhabitants, characterized by frequent air pollution episodes due to the low circulation of air masses (Caserini, Giani, Cacciamani, Ozgen, & Lonati, 2017; Pernigotti, Georgieva, Thunis, & Bessagnet, 2012).

2. Materials and methods

2.1. Air pollution data

The dataset used in the study comprises 26 measurement sites covering an area of about 400 $\times$ 130 km^2 centered around 45° *N* 10° E, representing the flat portion of the Po river basin known as Padana Plain (Fig. 1). Thirteen stations for each pollutant are considered, with an average ratio of missing data of about 4.1% for NO_2 and 4.4% for O_3. They were selected based on data availability and trying to cover all the areas in a relatively uniform way. Data from ten stations serve as the primary testbed (122,712 hourly samples from 2010 to 2023), meaning that they are used for the training (8 years), validation (3 years), and test (3 years) phases of the local models (one for each site) and of the regional model. The other three stations for each pollutant represent the external testbed used to test the performance of the regional model on data never used in the preceding steps (2021–2023 data only, 26,280 samples).

The measured pollutant and the main characteristics of all the stations are listed in Table 1. All stations are at low elevations, and are mainly classified as "urban background". This means they are close to urban centers or located in open spaces within towns, but are less directly impacted by very local emissions than traffic or industrial sites. A few stations are located in rural environments (Colli Euganei, Druento, Masué), and Rimini is 1.6 km from the seashore. The high quality of the considered data is demonstrated by the small fraction of missing values (from 0.5% to 7.7%). In addition, all data were validated by the responsible regional agencies, that also made the necessary adjustments, for instance, to account for substituting some of the sensors. Additionally, a large-scale harmonization between the regional sensor networks was undertaken within the six-year-long PREPAIR Life project (Raffaelli et al., 2020).

Fig. 3 shows the value distribution of the ten primary sites measuring NO_2. The sites are ordered from west to east in order to show the similarities between geographically adjacent sites. The violin plot shows that the most common values are around 20 μg/m^3 and that the concentration rarely reaches 0, except for Colli Euganei and, partly, Rovigo. Monza, close to the most densely populated part of the domain, shows both the highest values and the largest range of variation (the thick black line in the plot shows 25%–75% variation, and the white dot is the median concentration). The similarities between stations also emerge from the analysis of the correlation matrix shown in Fig. 2 (top). Despite referring to quite varied emission and environmental conditions (for instance, the population density differs by up to two orders of magnitude between Milano and Abbadia Cerreto), most concentration variations are common to all stations. The daily and yearly cycles are similar, so peak correlation values reach almost 0.8. The autocorrelation analysis of these time series (Fig. 4) shows that a 24-h period is evident at all sites, even if it is sometimes less pronounced, like in Colli Euganei. Some minor peaks appear with a 12-h periodicity, possibly linked to intraday variations. Similar statistics are reported for ozone. The violin plots of Fig. 5 show that, except for Druento and Biella, the most common values are close to zero. This happens because, as already pointed out, ozone formation is impossible during nighttime. Consistently, the autocorrelation functions for ozone (Fig. 6) exhibit a more substantial daily period that almost goes to zero after twelve hours. Biella is again a partial exception in that this drop in the value is much less evident. The overall average value is slightly less than 50 μg/m^3 and the range of variation is very similar at all sites. Fig. 2 (bottom) shows the correlation of ozone concentration among the ten testbed stations. Their values are higher than NO_2 since all the stations are strongly influenced

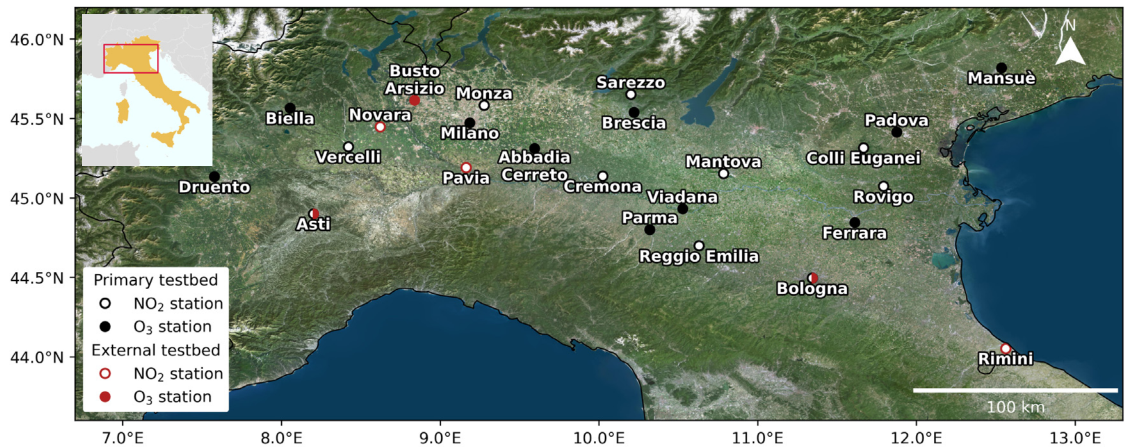

Fig. 1. Position of the NO_2 and O_3 monitoring stations of the primary and external testbeds.

Table 1
Characteristics of the NO_2 and O_3 monitoring stations.

	Site	Testbed	Missing data[%]	Lat [° N]	Lon [° E]	Elevation [m asl]
	Asti	Primary	4.4	44.9	8.2	149
	Bologna	Primary	7.2	44.5	11.3	43
	Colli Euganei	Primary	5.6	45.3	11.6	13
	Cremona	Primary	1.7	45.1	10.0	43
	Mantova	Primary	2.3	45.2	10.8	22
	Monza	Primary	7.7	45.6	9.3	163
NO_2	Novara	External	4.2	45.4	8.6	154
	Pavia	External	1.9	45.2	9.2	77
	Reggio Emilia	Primary	1.4	44.7	10.7	55
	Rimini	External	3.0	44.1	12.6	5
	Rovigo	Primary	4.9	45.0	11.8	4
	Sarezzo	Primary	5.1	45.6	10.2	269
	Vercelli	Primary	3.5	45.3	8.4	134
	Abbadia Cerreto	Primary	6.2	45.3	9.6	64
	Asti	External	6.0	44.9	8.2	149
	Biella	Primary	4.3	45.6	8.1	406
	Bologna	External	0.9	44.5	11.3	43
	Brescia	Primary	6.5	45.5	10.2	70
	Busto Arsizio	External	0.5	45.6	8.8	206
O_3	Druento	Primary	6.4	45.2	7.6	335
	Ferrara	Primary	4.6	44.8	11.6	8
	Mansuè	Primary	5.8	45.8	12.5	11
	Milano	Primary	2.1	45.5	9.2	122
	Padova	Primary	5.9	45.4	11.8	12
	Parma	Primary	4.9	44.8	10.3	60
	Viadana	Primary	2.8	44.9	10.5	37

by the daily cycle and the consequent ozone dynamic. The overall average is as high as 0.83, with some values reaching almost 0.9. The lower values are far from the diagonal, i.e., refer to more distant stations.

2.2. Regional forecasting approach

We propose a framework to assess the predictive performance of a regional model, identified with the data available from different stations, against the local models, one for each site (Fig. 7). First, a predictor is developed for each measurement station (stations **a**, **b**, and **c** in the figure), using local data for training (2010–2017), validation (2018–2020), and test (2021–2023). This represents the classical machine learning application with a clearly defined application domain. Then, a regional model, $f^R(\cdot)$, is developed for each pollutant using all the data of the ten primary-testbed stations for training and validation, testing its performance separately on each measurement site. To further assess the generalization capability of the regional model, its performances are tested on the three additional test stations of the same area (external testbed in Fig. 1), which were not used during the forecasting models identification phases. In terms of performance, this may represent what one could expect by applying a regional model to a new station where past data are unavailable or insufficient to develop a predictor.

The specific task here considered is the forecasting of pollutant concentration for the next 24 h using the 48 past hourly values (only present and lagged concentrations are used as in Gómez-Losada, Asencio-Cortés, Martínez-Álvarez, & Riquelme, 2018). A variety of machine learning models can be employed to perform multi-step forecasting (see, Ayturan, Ayturan, & Altun, 2018; Liao et al., 2020; Zaini, Ean, Ahmed, & Malek, 2022; Zhang et al., 2022 for recent reviews). Here we adopt Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) cells (Hochreiter & Schmidhuber, 1997), that have been widely used for air quality forecasting in recent years (Jin, Zeng, Yan, & Ji, 2021; Seng, Zhang, Zhang, Chen, & Chen, 2021; Song, Huang, & Song, 2019; Zhang et al., 2021). RNNs are designed to process sequential data iteratively, thanks to the cyclic (recurrent) connections between the neurons: At each time step, the current input influences the neurons' internal states, and the prediction is computed. A key feature of RNNs is the parameter sharing across time

NO_2

	Asti	Vercelli	Monza	Cremona	Sarezzo	Reggio Emilia	Mantova	Bologna	Colli Euganei	Rovigo
Asti	-	0.79	0.68	0.71	0.61	0.72	0.69	0.56	0.60	0.62
Vercelli	0.79	-	0.69	0.70	0.64	0.70	0.68	0.60	0.64	0.68
Monza	0.68	0.69	-	0.67	0.60	0.64	0.63	0.56	0.59	0.59
Cremona	0.71	0.70	0.67	-	0.53	0.74	0.76	0.48	0.55	0.65
Sarezzo	0.61	0.64	0.60	0.53	-	0.60	0.56	0.56	0.60	0.51
Reggio Emilia	0.72	0.70	0.64	0.74	0.60	-	0.73	0.57	0.60	0.64
Mantova	0.69	0.68	0.63	0.76	0.56	0.73	-	0.46	0.62	0.71
Bologna	0.56	0.60	0.56	0.48	0.56	0.57	0.46	-	0.53	0.52
Colli Euganei	0.60	0.64	0.59	0.55	0.60	0.60	0.62	0.53	-	0.68
Rovigo	0.62	0.68	0.59	0.65	0.51	0.64	0.71	0.52	0.68	-

O_3

	Druento	Biella	Milano	Abbadia Cerreto	Brescia	Parma	Viadana	Ferrara	Padova	Mansuè
Druento	-	0.78	0.77	0.85	0.76	0.80	0.83	0.81	0.80	0.81
Biella	0.78	-	0.82	0.76	0.82	0.83	0.80	0.77	0.73	0.72
Milano	0.77	0.82	-	0.88	0.87	0.86	0.87	0.82	0.80	0.78
Abbadia Cerreto	0.85	0.76	0.88	-	0.84	0.86	0.90	0.86	0.85	0.84
Brescia	0.76	0.82	0.87	0.84	-	0.87	0.87	0.83	0.80	0.78
Parma	0.80	0.83	0.86	0.86	0.87	-	0.92	0.88	0.84	0.82
Viadana	0.83	0.80	0.87	0.90	0.87	0.92	-	0.89	0.87	0.84
Ferrara	0.81	0.77	0.82	0.86	0.83	0.88	0.89	-	0.86	0.86
Padova	0.80	0.73	0.80	0.85	0.80	0.84	0.87	0.86	-	0.87
Mansuè	0.81	0.72	0.78	0.84	0.78	0.82	0.84	0.86	0.87	-

Fig. 2. Cross correlations for NO_2 (top) and O_3 (bottom) for the primary testbed stations. Cell in position (i, j) reports the Pearson correlation coefficient between data from the station on ith row and on jth column.

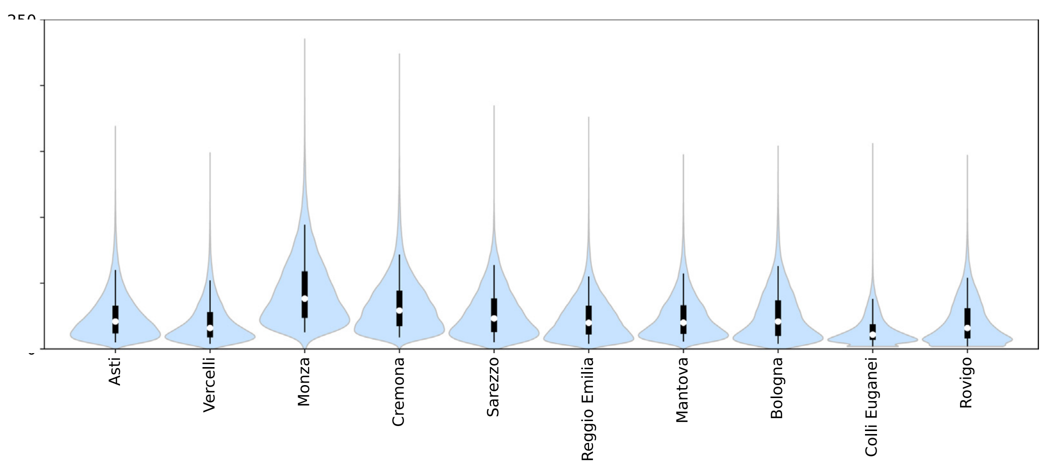

Fig. 3. Distributions of the NO_2 concentration (primary testbed only). The white dot represents the median; the box covers the range between 25 and 75 percentile; the whiskers highlight the interval between 5 and 95 percentile.

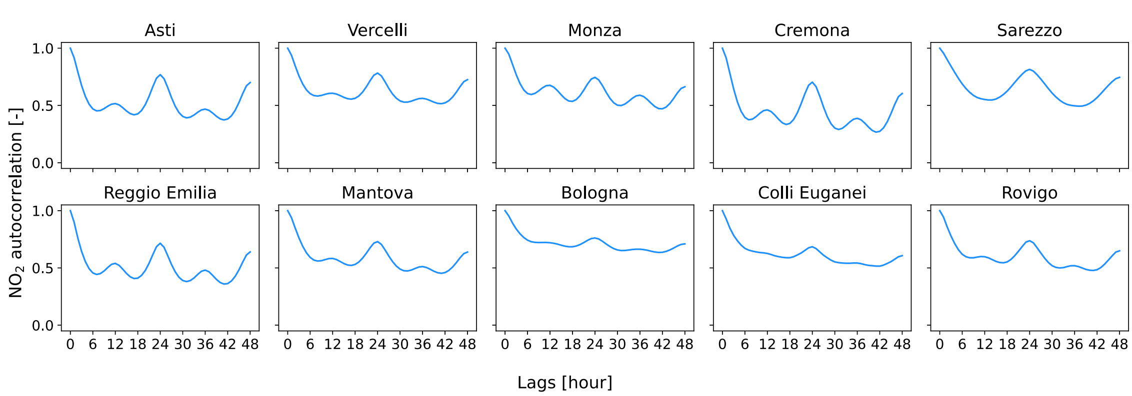

Fig. 4. Hourly autocorrelation of the NO_2 concentration over two days (primary testbed only).

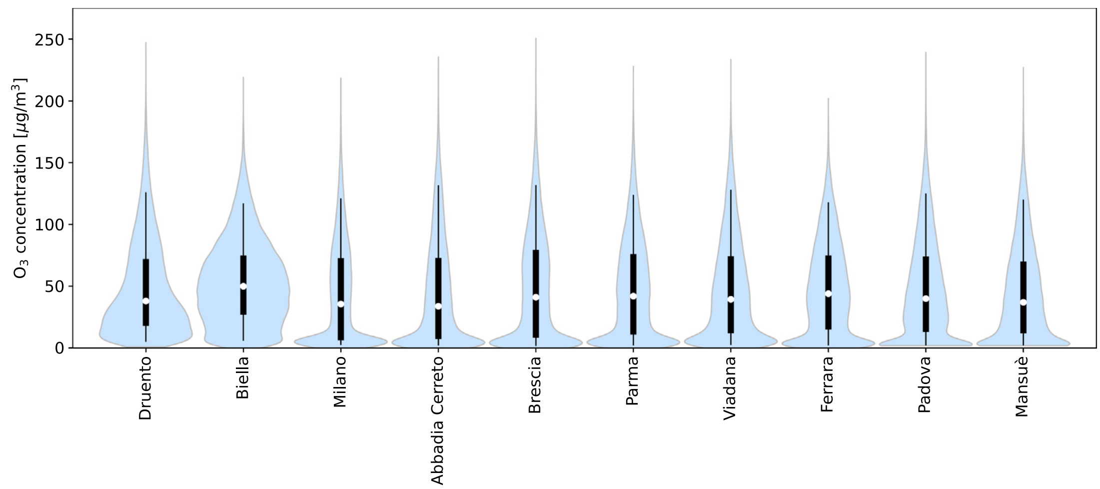

Fig. 5. Distributions of the O_3 concentration (primary testbed only). The white dot represents the median; the box covers the range between 25 and 75 percentile; the whiskers highlight the interval between 5 and 95 percentile.

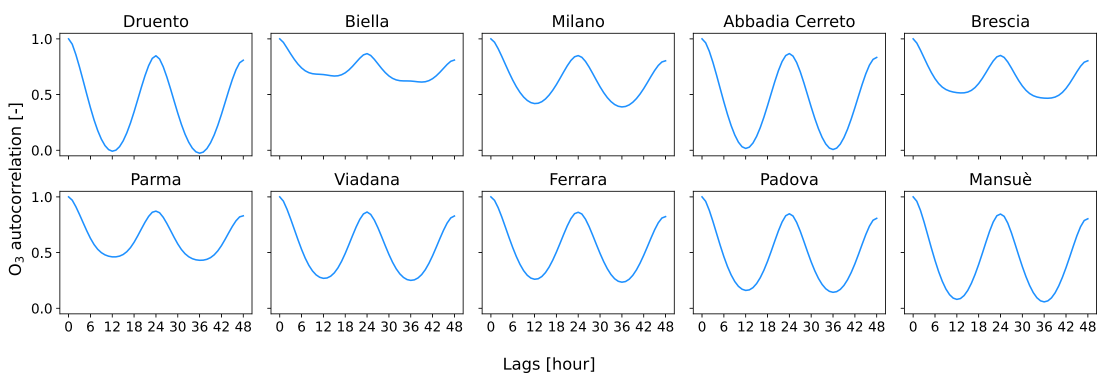

Fig. 6. Hourly autocorrelation of the O_3 concentration over two days (primary testbed only).

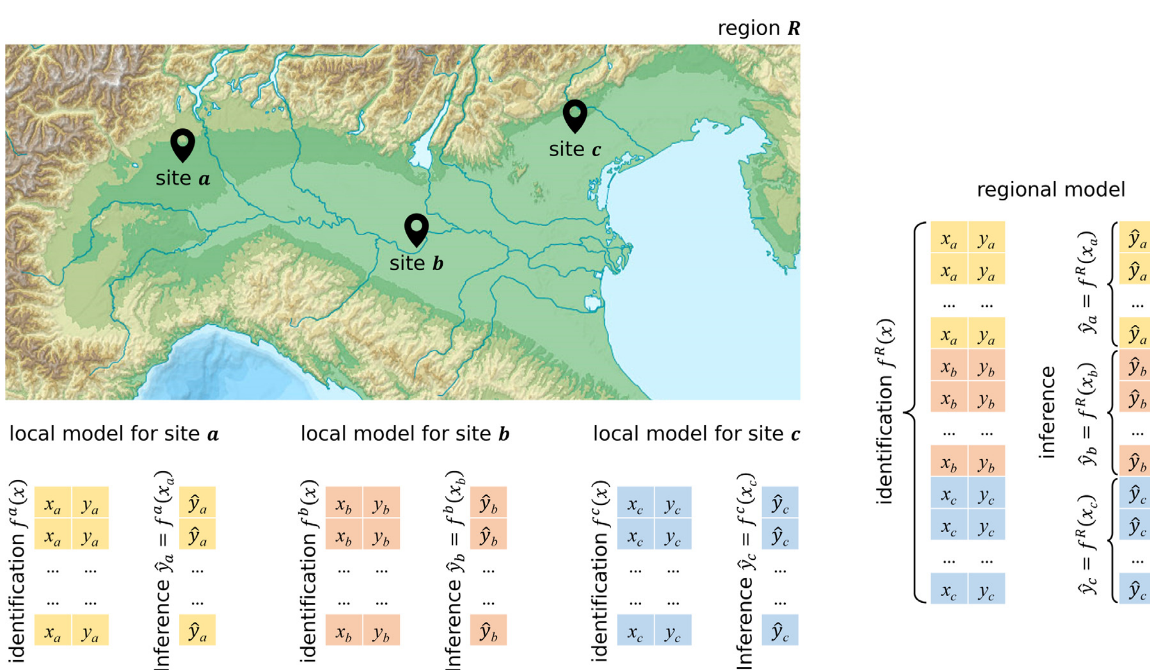

Fig. 7. Identification and inference phases for three local models (sites **a**, **b**, and **c**) and a regional model using the data available from the three sites.

steps, which strongly limits the number of parameters to be trained (it does not depend on the number of steps to be processed) (Hewamalage, Bergmeir, & Bandara, 2021; Sangiorgio & Dercole, 2020). LSTM cells have been preferred over naive recurrent neurons because they have two internal states and three gates that regulate the information flow, making them more robust against vanishing gradients and more efficient at preserving relevant information over long time spans.

Following the approach already used in Sangiorgio and Guariso (2024), each time series is organized into pairs of input/output sequences, $[y(t-48+1),\dots,y(t-1),y(t)]$ and $[y(t+1),y(t+2),\dots,y(t+24)]$. The network can be then trained using a supervised learning approach so that its predictions, $[\hat{y}(t+1),\hat{y}(t+2),\dots,\hat{y}(t+24)]$, approximate as accurately as possible the actual outputs (Masini, Medeiros, & Mendes, 2023; Sangiorgio, Dercole, & Guariso, 2021; Taieb, Bontempi, Atiya, & Sorjamaa, 2012). We initialize the LSTM internal state vector ($s(t-48)$) for each input/output sequence to zero. The model then iteratively processes the 48-h input sequence till the current time step (t) when the prediction starts: $\left[\hat{y}(t+1), s(t)\right] = f\left(y(t), s(t-1)\right)$. Previous predictions ($\hat{y}(t-46),\dots,\hat{y}(t)$) are discarded since they refer to the past. For the following hours, the predicted values are fed $\left[\hat{y}(t+2), s(t+1)\right] = f\left(\hat{y}(t+1), s(t)\right)$.

This means a multi-step prediction is issued at each hour, and thus, the value at a specific instant is predicted 24 times, with increasing precision. This multi-step scheme allows a more complete view of what will happen the following day compared to a traditional setting where the prediction refers to a specific time or the average of the following 24 h. Such a prediction is thus more informative for citizens and decision-makers.

Each LSTM network is trained using Adam (adaptive momentum estimation) optimizer, with gradients computed by the traditional backpropagation through time algorithm. The closed-loop configuration presented in Sangiorgio and Dercole (2020), without teacher forcing, is adopted. The loss function minimized during training is the mean squared difference between predicted and actual concentration values (averaged over the 24-h horizon). The same general neural structure is used for all the models to allow for a straightforward comparison at a reasonable computational cost: two LSTM layers with a fully connected one on top of them. However, the number of neurons in each layer and all the other hyperparameters were optimized for each predictor through an exhaustive search. We tested 5 or 10 neurons in each hidden layer; a learning rate (i.e., the step size of the gradient-based optimization) of 10^{-2} or 10^{-3}; decay of the learning rate of 10^{-3} or 10^{-4}; and batch size (the number of samples used to perform one iteration of the algorithm) of 256 or 512. Two runs with different random seeds were performed for each hyperparameter combination to avoid the effects of a possibly critical initialization.

The neural networks were implemented in Python 3.8 using PyTorch library. The experimental environment Ubuntu 20.04.5 LTS, with a CPU Intel(R) Core(TM) i9-10900K at 3.70 GHz (20 threads) and a GPU NVIDIA Quadro RTX 5000 (16 GB).

3. Results

In the following, we have adopted two traditional metrics

- the R^2-score, which measures the variability of the actual values captured by the model. It is defined as 1 minus the ratio between the Mean Squared Error (MSE, i.e., the squared deviation of model prediction from actual values) and the data variance (Dercole, Sangiorgio, & Schmirander, 2020);
- the Normalized Mean Absolute Error (NMAE), which is the ratio of the Mean Absolute Error (MAE, i.e., the deviation of model prediction from actual values) and the average concentration values. It allows for an easy comparison between models' performances that refer to stations with varied average concentrations.

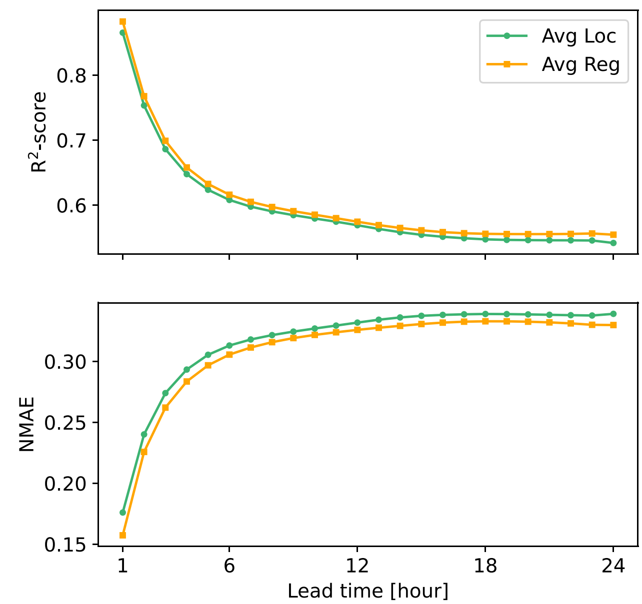

Fig. 8. Overall performance (average of all the primary testbed stations) of the NO_2 local models and the regional model (R^2-score on top; NMAE at bottom). Metrics computed on the testing period (2021–2023).

Fig. 8 compares the average performance of the regional NO_2 model with those of the individual ones in terms of R^2 (upper panel) and normalized mean absolute error (NMAE) in the lower panel.

What immediately emerges is that, on average, the regional model performs slightly better than the local ones on the primary testbed (higher R^2-score and lower NMAE). This means that using the training data of the other stations effectively enlarges the data space, adding helpful information not present in each local dataset. In both cases, the R^2-score, which is close to 0.9 one hour ahead, decreases rapidly on a longer horizon (until about six hours) and then almost stabilizes at a value of around 0.55. In parallel, the NMAE starts with about 15% one hour ahead and, after six hours, reaches a plateau at about 33%. These curves can also be interpreted as the progressive improvements of the forecast of a specific time. Considering, for instance, the value predicted for 8:00 am of day *d*, when the forecast is issued at 8:00 am of the preceding day ($d-1$), the error we can expect is around 33%. The forecast for the same hour issued 18 h later, i.e., at 2:00 am of the day *d*, will have an average error of around 30%. Finally, the value predicted for 8:00 am in the forecast issues at 7:00 am, i.e., one hour in advance, will have an expected error of slightly more than 15%.

A better average performance does not necessarily mean that the regional model outperforms the local ones at all the considered stations. The heat map in Fig. 10 (top) shows that the local model prevails in two cases, Colli Euganei and Vercelli, after step five. These two stations are in areas with very low population density and prevailing agricultural activities and, thus, differ in certain aspects from the rest of the domain. However, if we examine these two cases closely, we can note that the NMAE difference is less than 1%. Thus, even in these cases, the performance of the local and regional approaches is substantially

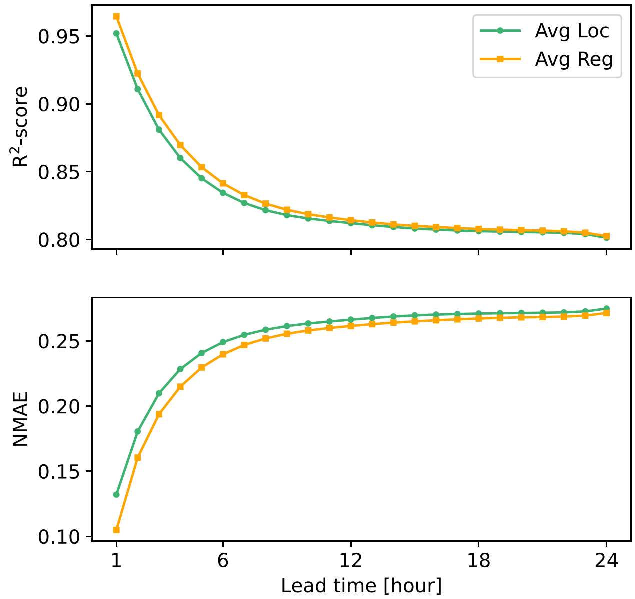

Fig. 9. Overall performance (average of all the primary testbed stations) of the O_3 local models and the regional model (R^2-score on top; NMAE at bottom). Metrics computed on the testing period (2021–2023).

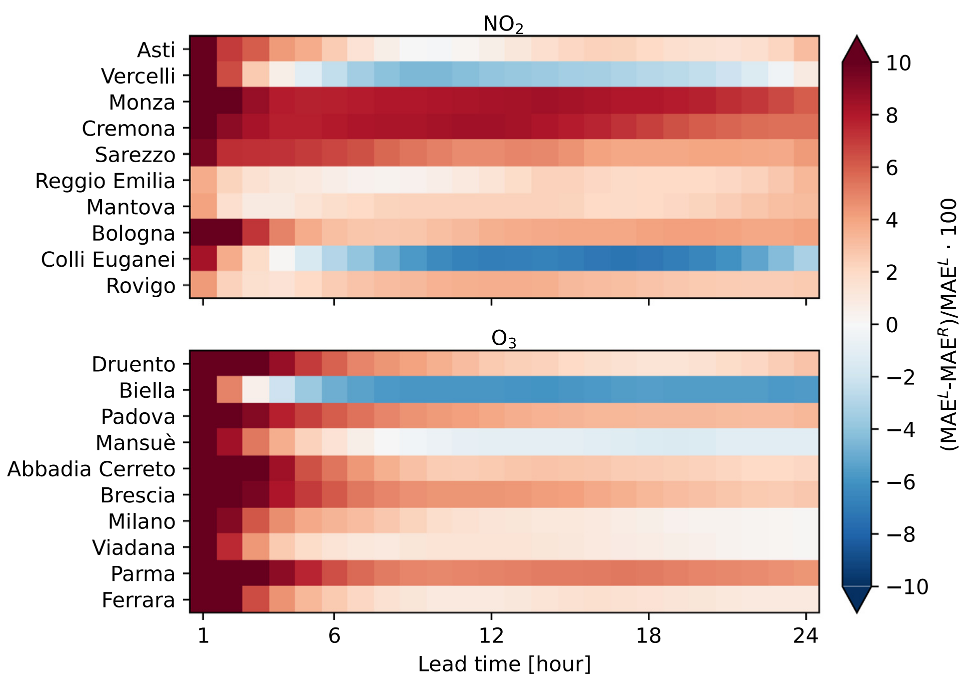

Fig. 10. Comparison between the performance of the regional model used on each station and the local model specific for that station (NO_2 on top; O_3 at bottom). The reported metric is the percentage variation of the MAE obtained with the regional model (MAE^R) with respect to that of the local model (MAE^L). Warm (cold) colors indicate that the regional model outperforms (is outperformed by) the local one. Metrics computed on the testing period (2021–2023).

equivalent. Overall, it is important to note that the regional model always outperforms the local ones in the very short term (1 to 4 h ahead), with a substantial performance improvement of 9%–10% in several cases.

Results for O_3 are quite similar. The average performance of the regional model is again slightly better than that of the local ones (Fig. 9). It can be noticed, however, that R^2 values are higher than those obtained for NO_2 and start from 0.95 to reach the plateau above 0.80. Looking at the results from the NMAE perspective, the values are initially close to 10% and remain below 30% even in the last steps of the forecasting horizon. These results are probably due to the more robust daily component of the O_3 time series that the models can reproduce. Again, Fig. 10 (bottom) shows that the regional model performs better in most sites, especially in the very short term (the improvement over local models from 1 to 4 h ahead is higher than 5% with the only exception of Biella). Biella appears somehow different since the local model performs better for horizons longer than three hours. Indeed, Biella's situation seemed already quite peculiar in terms of the distribution of the concentration values shown in Fig. 5. Biella is at 406 m of elevation, and it is well-known that elevation plays a relevant role in ozone concentrations (Sangiorgio & Guariso, 2023, 2024). Additionally, in Biella, ozone is more persistent, as shown by the analysis of the autocorrelogram (Fig. 6).

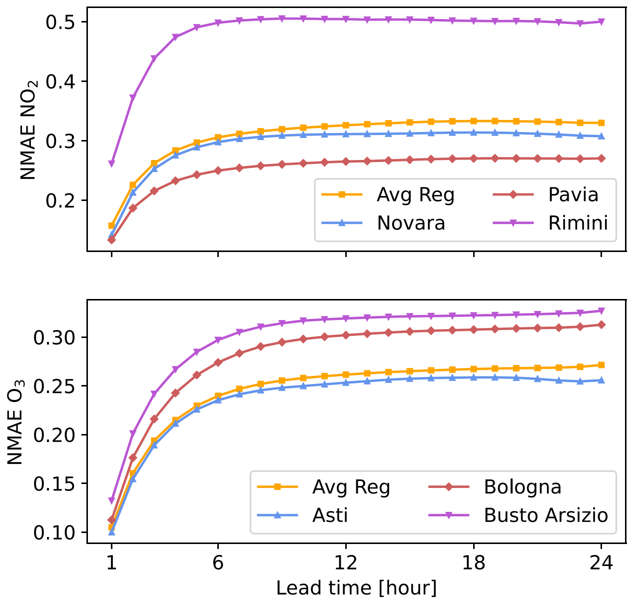

Fig. 11. Performance (NMAE) of the regional model in the prediction of the external testbed stations (NO_2 on top; O_3 at bottom). The average NMAE of the regional model on the primary testbed stations (orange curve) is reported for reference. Metrics computed on the testing period (2021–2023).

Altogether, the average performance of the regional and local forecasting models is very accurate and substantially equivalent, with the regional model slightly prevailing in average terms. Furthermore, O_3, a secondary pollutant that is difficult to simulate, can be forecasted with better results than NO_2. As already illustrated, O_3 concentrations are the output of a complex nonlinear system that may attenuate the possible fast variations due to emission changes, i.e., the input of the chemical-physical process that determines the concentration. NO_2 concentrations, on the contrary, are more directly connected to emission changes and thus are more challenging to forecast with a model only based on past values.

The results of applying the regional models for NO_2 and O_3 to three other sites (external testbed), not included in the original dataset (primary testbed), are shown in Fig. 11 again in terms of R^2 and NMAE. What immediately appears is that regional models perform relatively well and similarly to the original dataset, except for NO_2 in the Rimini station. Its NMAE is about 0.5 five hours ahead, compared to 0.3 of the other sites. Indeed, Rimini has some peculiarities. It lies on the extreme border of the flat portion of Northern Italy, the Padana Plain, considered in the study (see Fig. 1) and is on the shore of the Adriatic Sea. This means its average wind speed, which was not used as input in the current model, is between 10 and 40% higher than the other stations. Consequently, its mean NO_2 concentration is about one-third lower than the domain average.

It is interesting to assess the models' performances also in terms of their ability to predict the exceedance of the threshold, above which environmental agencies are generally bound to take some actions. Interestingly, with this approach, the overall forecasting problem is reduced to a classification one: being above or below the threshold. In practice, however, being 0.1 $\mu g/m^3$ over the threshold is quite different from being 10 or 20 $\mu g/m^3$ over, though being equally a violation of the limit. This is why it is still preferable to frame the overall problem as a regression and check its accuracy in classifying the critical occurrences only afterwards. Again, several metrics can be used to check a model's classification performance. They all derive from the result of the so-called confusion matrix that reports the number CE of correctly predicted exceedances (when the model correctly anticipates an episode that then materializes), the number of false exceedances FE (the model predicts an exceedance that does not happen), the number ME of missed exceedances (the model fails to predict the exceedance), the number CN of correct negative (both the forecast and the actual value are below the threshold), and $N = CE + FE + ME + CN$, the total number of episodes. Counting the number of times the concentration exceeds the limits defined by the WHO, i.e., 25 $\mu g/m^3$ as an average over 24 h for NO_2 and 100 $\mu g/m^3$ as the daily average of the eight most polluted hours for ozone, one can compute the following indicators:

- Precision $= CE/(CE + ME)$, i.e., the number of times the model correctly predicts the exceedance over the total number of exceedances;
- Recall $= CE/(CE+FN)$, i.e., the number of correct exceedance predictions over the number of predicted exceedances;
- F1-Score $= 2CE/(2CE + FN + ME)$, i.e., the harmonic mean of precision and recall;
- Accuracy $= (CE+CN)/N$, i.e., the fraction of correct forecasts among all the episodes considered.

All the indicators are dimensionless fractions with an optimal value of 1.

Table 2 presents the index values for the models under analysis for NO_2 (left) and O_3 (right). The first two columns for each pollutant show the metric averages computed for the local (Avg Loc) and the regional (Avg Reg) models on the primary testbed. These are compared with those of the external testbed. The reported values are extremely high, showing that the models can reliably be adopted to categorize the forecast in terms of exceedance. Again, Rimini represents somehow an exception. The accuracy value is excellent (only 4% lower than the average of the regional model). Still, this metric is known to weigh the missed and the false alarms equally and is not particularly suited when the two classes to be identified are not well-balanced. In our case, NO_2 and O_3 exceedances are between 38 and 40% of all the episodes. So, the two classes can be considered reasonably balanced. Still, the value of the false exceedance is probably much lower than that of the missed alarm, so that they cannot be considered equally important. In this case, one has to refer to the other metrics, particularly recall, whose value is 22% lower than that of the average regional model. This confirms the difficulty of adopting a model developed in the central Po Valley for this station.

4. Discussion

Although forecasting pollution trends is crucial for decision-making, most regulations in force in the European Union and elsewhere quantify air pollution as exceeding concentration thresholds (Deng, Manders, Segers, Heemink, & Lin, 2024). This approach is linked to the concept of population exposure, as it foresees possible damage to human health when the threshold is exceeded. Below the threshold, the impact is assumed to be negligible.

Each performance metric represents a specific viewpoint; thus, a preferred model can be defined only according to a definite performance measure. The "best predictor" does not exist in absolute terms, as selecting a forecasting model is a multi-objective problem. Evidently, a performance measure based on threshold appeals to actual decision-makers for its simplicity and ease of communication. However, a correct choice implies the possibility of evaluating the impact of both false and missed exceedances, requiring a careful analysis of the specific decision setting in which

Table 2
Exceedance analysis for NO_2 (left) and O_3 (right).

	NO_2					O_3				
	Avg Loc	Avg Reg	Novara	Pavia	Rimini	Avg Loc	Avg Reg	Asti	Bologna	Busto Arsizio
Precision	0.810	0.787	0.796	0.854	0.561	0.860	0.857	0.880	0.802	0.876
Recall	0.804	0.819	0.799	0.845	0.635	0.871	0.880	0.812	0.888	0.884
F1-score	0.807	0.803	0.797	0.850	0.596	0.865	0.869	0.845	0.843	0.880
Accuracy	0.892	0.892	0.874	0.889	0.857	0.923	0.926	0.907	0.927	0.928

the model will be applied. From the decision-makers' viewpoint, a forecast slightly below the threshold does not differ significantly from a forecast slightly above. To decide, for instance, about some traffic restrictions, the two forecasts can be considered "almost" equivalent even if, strictly speaking, a value below does not constitute an exceedance while the other does. So, if one adopts a crisp threshold definition, the two cases would be counted on different rows of the confusion matrix. Adopting a fuzzy approach would assign a non-integer value close to but lower than one, for instance, to a 99 forecast (with a 100 threshold). An even lower fractional value, say 0.5, would be assigned to 95 and a value of 0 to a forecast of 90. A similar approach would change the values and the meaning of all the indicators, such as the accuracy or the F1-score. Another issue is related to the trade-off between accuracy and anticipation. In fact, it is intrinsic to forecasting models that their accuracy decreases with the length of the forecasting horizon, so it would be essential to compare the usefulness of more extended anticipation with the possibility of more significant errors. In the future, enriching the model input with other variables, such as meteorology or emission precursors, may increase the precision at longer horizons. Again, a correct choice can be made in each specific application only by weighing the additional precision against the increased complexity of the model and the cost of acquiring the additional data. A further enhancement could be a mechanism, such as periodic retraining, to adapt the model to evolving environmental conditions. This may become necessary if significant changes will become increasingly fast in climate, but also car fleets and other emission sources.

5. Conclusion

This study shows that, within a given application domain, it is possible to develop highly reliable regional models of air quality that can be extended to other sites for both NO_2 and O_3. Since these pollutants represent different air quality issues, it is plausible that this conclusion can be extended to other primary and secondary pollutants, such as particulate matter. Additionally, it has been proved that the models perform excellently as classifiers of exceedances. This ability is particularly important for decision-makers who must adhere to existing regulations, usually defined in terms of thresholds. There is no way to determine if a regional model would be an acceptable predictor for a new station in general terms. An exhaustive search with all the possible combinations of primary and external stations is not only computationally too hard but also inconclusive. There is no guarantee that the model will perform at the new station as it does at the base ones. This means that any conclusion is only valid in a statistical sense. A regional model can, on average, perform similarly or even better than a local one in forecasting air pollutant concentrations one day ahead. However, in some specific cases, a performance worsening is possible due to the differences between a station and the average conditions of those considered in the domain. Collecting other "external" information would be necessary to determine whether the new station belongs to the same knowledge domain as the base ones. The more a station's characteristics (elevation, context, presence of emission source, etc.) differ from those considered, the less reliable a regional model's application is. One has to wait until a sufficiently long dataset becomes available, which allows the training and testing of a local model. However, no administrative or physical boundary of the domain exists: The regional model performances decrease as the station differs from the core of those used to develop such a model. Analogously, in the future, it will be helpful to introduce additional indicators based on a fuzzy definition of thresholds instead of a crisp one. Despite some greater difficulties in their interpretation, this approach has the advantage of being close to quantifying the risk of exceeding the threshold—an instrumental concept to support decision-making.

CRediT authorship contribution statement

Matteo Sangiorgio: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology. **Giorgio Guariso:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization.

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.

Acknowledgments

The authors would like to thank the Regional Environmental Agencies (ARPA) of Piedmont, Lombardy, Veneto (R. Bassan), and Emilia-Romagna (S. Maccaferri) that provided the data arranged for the study.

Data availability

All data used are publicly available from the websites of the Regional Environmental Agencies (ARPA) of Piedmont, Lombardy, Veneto, and Emilia-Romagna (www.arpa.piemonte.it; www.arpalombardia.it; www.arpa.veneto.it; www.arpae.it). Wind data can be found on www.it.windfinder.com.

References

Agirre-Basurko, E., Ibarra-Berastegi, G., & Madariaga, I. (2006). Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area. *Environmental Modelling & Software*, *21*(4), 430–446.

Amaranto, A., & Mazzoleni, M. (2023). B-AMA: A python-coded protocol to enhance the application of data-driven models in hydrology. *Environmental Modelling & Software*, *160*, Article 105609.

Ayturan, Y. A., Ayturan, Z. C., & Altun, H. O. (2018). Air pollution modelling with deep learning: a review. *International Journal of Environmental Pollution and Environmental Modelling*, *1*(3), 58–62.

Bai, Y., Li, Y., Wang, X., Xie, J., & Li, C. (2016). Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. *Atmospheric Pollution Research*, *7*(3), 557–566.

Bolzern, P., Finzi, G., Fronza, G., Spirito, A., et al. (1982). ARMAX stochastic models of air pollution: three case studies. *Mathematical Models for Planning and Controlling Air Quality*, 185.

Cabaneros, S. M., Calautit, J. K., & Hughes, B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. *Environmental Modelling & Software*, *119*, 285–304.
Caserini, S., Giani, P., Cacciamani, C., Ozgen, S., & Lonati, G. (2017). Influence of climate change on the frequency of daytime temperature inversions and stagnation events in the Po Valley: Historical trend and future projections. *Atmospheric Research*, *184*, 15–23.
Chen, J.-L., Islam, S., & Biswas, P. (1998). Nonlinear dynamics of hourly ozone concentrations: nonparametric short term prediction. *Atmospheric Environment*, *32*(11), 1839–1848.
Deng, T., Manders, A., Segers, A., Heemink, A. W., & Lin, H. X. (2024). Ozone exceedance forecasting with enhanced extreme instance augmentation: A case study in Germany. *Environmental Modelling & Software*, Article 106162.
Dercole, F., Sangiorgio, M., & Schmirander, Y. (2020). An empirical assessment of the universality of ANNs to predict oscillatory time series. *IFAC-PapersOnLine*, *53*(2), 1255–1260.
Finlayson-Pitts, B., & Pitts, J., Jr. (1993). Atmospheric chemistry of tropospheric ozone formation: scientific and regulatory implications. *Air & Waste*, *43*(8), 1091–1100.
Fong, I. H., Li, T., Fong, S., Wong, R. K., & Tallon-Ballesteros, A. J. (2020). Predicting concentration levels of air pollutants by transfer learning and recurrent neural network. *Knowledge-Based Systems*, *192*, Article 105622.
Gómez-Losada, Á., Asencio-Cortés, G., Martínez-Álvarez, F., & Riquelme, J. C. (2018). A novel approach to forecast urban surface-level ozone considering heterogeneous locations and limited information. *Environmental Modelling & Software*, *110*, 52–61.
Gupta, S., Park, Y., Bi, J., Gupta, S., Züfle, A., Wildani, A., et al. (2024). Spatial transfer learning for estimating PM 2.5 in data-poor regions. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 385–400). Springer.
Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, *37*(1), 388–427.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.
Huang, S., Li, H., Wang, M., Qian, Y., Steenland, K., Caudle, W. M., et al. (2021). Long-term exposure to nitrogen dioxide and mortality: A systematic review and meta-analysis. *Science of the Total Environment*, *776*, Article 145968.
Jin, N., Zeng, Y., Yan, K., & Ji, Z. (2021). Multivariate air quality forecasting with nested long short term memory neural network. *IEEE Transactions on Industrial Informatics*, *17*(12), 8514–8522.
Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). HESS opinions: Never train a long short-term memory (LSTM) network on a single basin. *Hydrology and Earth System Sciences*, *28*(17), 4187–4201.
Latza, U., Gerdes, S., & Baur, X. (2009). Effects of nitrogen dioxide on human health: systematic review of experimental and epidemiological studies conducted between 2002 and 2006. *International Journal of Hygiene and Environmental Health*, *212*(3), 271–287.
Lelieveld, J., & Dentener, F. J. (2000). What controls tropospheric ozone? *Journal of Geophysical Research: Atmospheres*, *105*(D3), 3531–3551.
Liao, Q., Zhu, M., Wu, L., Pan, X., Tang, X., & Wang, Z. (2020). Deep learning for air quality forecasts: a review. *Current Pollution Reports*, *6*, 399–409.
Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society, Series A*, *379*(2194), Article 20200209.
Lin, X., Trainer, M., & Liu, S. (1988). On the nonlinearity of the tropospheric ozone production. *Journal of Geophysical Research: Atmospheres*, *93*(D12), 15879–15888.
Ma, J., Li, Z., Cheng, J. C., Ding, Y., Lin, C., & Xu, Z. (2020). Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network. *Science of the Total Environment*, *705*, Article 135771.
Ma, W., Yuan, Z., Lau, A. K., Wang, L., Liao, C., & Zhang, Y. (2022). Optimized neural network for daily-scale ozone prediction based on transfer learning. *Science of the Total Environment*, *827*, Article 154279.
Maciąg, P. S., Kasabov, N., Kryszkiewicz, M., & Bembenik, R. (2019). Air pollution prediction with clustering-based ensemble of evolving spiking neural networks and a case study for London area. *Environmental Modelling & Software*, *118*, 262–280.
Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2023). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, *37*(1), 76–111.
Masood, A., & Ahmad, K. (2021). A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance. *Journal of Cleaner Production*, *322*, Article 129072.
Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: A survey. *Artificial Intelligence Review*, *56*(9), 10031–10066.
Nielsen, A. (2019). *Practical time series analysis: Prediction with statistics and machine learning*. O'Reilly Media.
Paoletti, E., De Marco, A., & Racalbuto, S. (2007). Why should we calculate complex indices of ozone exposure? Results from Mediterranean background sites. *Environmental Monitoring and Assessment*, *128*, 19–30.
Pernigotti, D., Georgieva, E., Thunis, P., & Bessagnet, B. (2012). Impact of meteorology on air quality modeling over the Po valley in northern Italy. *Atmospheric Environment*, *51*, 303–310.
Pisoni, E., Farina, M., Carnevale, C., & Piroddi, L. (2009). Forecasting peak air pollution levels using NARX models. *Engineering Applications of Artificial Intelligence*, *22*(4–5), 593–602.
Raffaelli, K., Deserti, M., Stortini, M., Amorati, R., Vasconi, M., & Giovannini, G. (2020). Improving air quality in the Po Valley, Italy: some results by the LIFE-IP-PREPAIR project. *Atmosphere*, *11*(4), 429.
Salcedo, R., Ferraz, M. A., Alves, C., & Martins, F. (1999). Time-series analysis of air pollution data. *Atmospheric Environment*, *33*(15), 2361–2372.
Sangiorgio, M., & Dercole, F. (2020). Robustness of LSTM neural networks for multi-step forecasting of chaotic time series. *Chaos, Solitons & Fractals*, *139*, Article 110045.
Sangiorgio, M., Dercole, F., & Guariso, G. (2021). Forecasting of noisy chaotic systems with deep neural networks. *Chaos, Solitons & Fractals*, *153*, Article 111570.
Sangiorgio, M., & Guariso, G. (2023). Deep neural network adaptation to different environmental contexts: A case study of ozone forecast. *IFAC-PapersOnLine*, *56*(2), 8290–8295.
Sangiorgio, M., & Guariso, G. (2024). Transfer learning in environmental data-driven models: A study of ozone forecast in the Alpine region. *Environmental Modelling & Software*, *177*, Article 106048.
Seng, D., Zhang, Q., Zhang, X., Chen, G., & Chen, X. (2021). Spatiotemporal prediction of air quality based on LSTM neural network. *Alexandria Engineering Journal*, *60*(2).
Singh, V., Carnevale, C., Finzi, G., Pisoni, E., & Volta, M. (2011). A cokriging based approach to reconstruct air pollution maps, processing measurement station concentrations and deterministic model simulations. *Environmental Modelling & Software*, *26*(6), 778–786.
Song, X., Huang, J., & Song, D. (2019). Air quality prediction based on LSTM-Kalman model. In *2019 IEEE 8th joint international information technology and artificial intelligence conference* (pp. 695–699). IEEE.
Stowell, J. D., min Kim, Y., Gao, Y., Fu, J. S., Chang, H. H., & Liu, Y. (2017). The impact of climate change and emissions control on future ozone levels: Implications for human health. *Environment International*, *108*, 41–50.
Taieb, S. B., Bontempi, G., Atiya, A. F., & Sorjamaa, A. (2012). A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Systems with Applications*, *39*(8), 7067–7083.
Tealab, A. (2018). Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, *3*(2), 334–340.
US EPA (2020). Integrated science assessment (ISA) for ozone and related photochemical oxidants (final report, Apr 2020).
Wang, P., Chen, Y., Hu, J., Zhang, H., & Ying, Q. (2018). Attribution of tropospheric ozone to NOx and VOC emissions: considering ozone formation in the transition regime. *Environmental Science and Technology*, *53*(3), 1404–1412.
Wang, J., Xu, W., Zhang, Y., & Dong, J. (2022). A novel air quality prediction and early warning system based on combined model of optimal feature extraction and intelligent optimization. *Chaos, Solitons & Fractals*, *158*, Article 112098.
Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*, 1–40.
World Health Organization (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization.
Yang, J., Ismail, A. W., Li, Y., Zhang, L., & Fadzli, F. E. (2023). Transfer learning-driven hourly PM2. 5 prediction based on a modified hybrid deep learning. *IEEE Access*, *11*, 99614–99627.
Yang, H., Zhu, Z., Li, C., & Li, R. (2020). A novel combined forecasting system for air pollutants concentration based on fuzzy theory and optimization of aggregation weight. *Applied Soft Computing*, *87*, Article 105972.
Zaini, N., Ean, L. W., Ahmed, A. N., & Malek, M. A. (2022). A systematic literature review of deep learning neural network for time series air quality forecasting. *Environmental Science and Pollution Research*, 1–33.
Zhang, L., Liu, P., Zhao, L., Wang, G., Zhang, W., & Liu, J. (2021). Air quality predictions with a semi-supervised bidirectional LSTM neural network. *Atmospheric Pollution Research*, *12*(1), 328–339.
Zhang, B., Rong, Y., Yong, R., Qin, D., Li, M., Zou, G., et al. (2022). Deep learning for air pollutant concentration prediction: A review. *Atmospheric Environment*, *290*, Article 119347.
Zhang, J., Wei, Y., & Fang, Z. (2019). Ozone pollution: a major health hazard worldwide. *Frontiers in Immunology*, *10*, 2518.
Zolghadri, A., Monsion, M., Henry, D., Marchionini, C., & Petrique, O. (2004). Development of an operational model-based warning system for tropospheric ozone concentrations in Bordeaux, France. *Environmental Modelling & Software*, *19*(4), 369–382.