Expert Systems With Applications 331 (2026) 133099

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Beyond traditional models: Foundation models for accurate particulate matter prediction

Federica Rollo [a,*], Matteo Angelinelli [b], Martina Casari [a], Laura Po [a], Giorgio Pedrazzi [b], Roberta Turra [b]

[a] "Enzo Ferrari" Engineering Department, University of Modena and Reggio Emilia, Via Vivarelli, 10, Modena, (MO), 41121, Italy
[b] CINECA, via Magnanelli, 2, Casalecchio Di Reno, (BO), 40033, Italy

ARTICLE INFO

Keywords:
PM2.5 forecasting
Foundation models
Low-cost sensors
Zero-shot learning
Deep learning

ABSTRACT

Accurate $PM_{2.5}$ forecasting is critical for public health, yet traditional deep learning models trained on location-specific data often lack generalizability across geographic contexts. Time Series Foundation Models (TSFMs) offer zero-shot forecasting capabilities through large-scale pre-training, but their efficacy for air quality prediction using low-cost sensor (LCS) data remains unexplored.

This study presents an empirical benchmarking study of TSFMs for $PM_{2.5}$ prediction using real-world LCS data evaluating Google's TimesFM and IBM's Granite against three deep learning architectures (CNN, LSTM, Transformer) across 34 datasets from 10 cities spanning 6 countries. We evaluate two forecasting strategies: direct prediction of reference station-equivalent data and prediction of LCS data with subsequent calibration.

Our results show that TSFMs consistently outperform traditional deep learning models, particularly for short-term forecasts, while retaining an stable advantage over longer horizons. Zero-shot TSFMs, applied without any task-specific training, perform competitively across all evaluated sites, providing empirical evidence of transferability to unseen sensor deployments. Fine-tuning offers limited additional benefit over the zero-shot configuration. Direct reference station prediction using fine-tuned Granite emerges as the most effective strategy, achieving the lowest errors and highest R^2 values. No clear performance decline is observed in datasets with substantial missing data compared to those with few missing values when linear interpolation is applied to fill gaps. Input window length has limited impact on forecasting accuracy. These findings establish TSFMs as a promising and practically viable direction for air quality monitoring with LCS networks.

1. Introduction

Foundation models (FMs) represent a paradigm shift in artificial intelligence. These large neural networks, pre-trained on massive datasets via self-supervised learning, have demonstrated remarkable generalization capabilities across diverse domains, including urban contexts (Zhang et al., 2024a). In the context of time series analysis, Time Series Foundation Models (TSFMs) such as Google's TimesFM (Das et al., 2024) and IBM's Granite (Ekambaram et al., 2025) have emerged, offering the potential to capture universal temporal patterns and perform forecasting in new domains with little to no fine-tuning, a capability known as zero-shot learning.

Accurate forecasting of $PM_{2.5}$ remains a critical challenge for public health and environmental management, as these ultrafine particles can penetrate deep into pulmonary alveoli and reach the bloodstream, with long-term exposure linked to increased respiratory and cardiovascular mortality (Agency, 2024; Hoek et al., 2013; Rajagopalan et al., 2018). The ability to provide accurate and timely air quality predictions is vital for informing public health policies, enabling real-time alerts, and supporting long-term environmental planning.

Low-cost sensors (LCSs) enable dense $PM_{2.5}$ monitoring networks at a fraction of the cost of traditional reference stations (RSs), but their measurements are sensitive to meteorological conditions, reducing data accuracy (Campo et al., 2023; Casari & Po, 2024; Kelly et al., 2023). Unlike traditional reference stations (RSs), which are equipped with highly accurate but costly instruments and offer limited spatial coverage, LCSs are compact, affordable, and can be widely deployed to form dense monitoring networks (Amegah, 2018; Barkjohn et al., 2021; Campo et al., 2023; Kelly et al., 2023).However, this increased spatial resolution comes at the cost of data accuracy. LCS measurements are

* Corresponding Author.
E-mail addresses: federica.rollo@unimore.it (F. Rollo), m.angelinelli@cineca.it (M. Angelinelli), martina.casari@unimore.it (M. Casari), laura.po@unimore.it (L. Po), g.pedrazzi@cineca.it (G. Pedrazzi), r.turra@cineca.it (R. Turra).

https://doi.org/10.1016/j.eswa.2026.133099
Received 26 March 2026; Received in revised form 20 May 2026; Accepted 31 May 2026
Available online 17 June 2026

notably sensitive to environmental conditions, particularly temperature and relative humidity (RH) (Casari & Po, 2024). Additional meteorological parameters such as wind speed, temperature, and atmospheric pressure further influence both PM concentrations and sensor performance through particle dispersion and chemical transformation processes. This poor data quality creates two fundamental tasks in LCS data processing: **calibration**, mapping raw LCS measurements to RS-equivalent values (Bachechi et al., 2024; Nalakurthi et al., 2024), and **prediction**, forecasting future $PM_{2.5}$ levels to support early warning systems and public health advisories. Accurate 12-h (or longer) forecasts are essential for proactive air quality management, enabling citizens to plan activities and authorities to prepare for pollution peaks.

DL architectures trained from scratch on location-specific data address these tasks effectively, but their performance often degrades in new geographic locations or on different sensor types, as the relationship between LCS readings and true concentrations depends on local meteorological conditions and emission sources (deSouza et al., 2022; Tancev & Pascale, 2020; Zhang et al., 2024b).

In contrast, TSFMs promise to overcome these limitations through their pre-trained, zero-shot capabilities. However, despite their potential, **The efficacy of TSFMs in the specific domain of air quality forecasting using LCS data remains largely unexplored.** Key questions include whether models pre-trained on generic time series can capture $PM_{2.5}$ dynamics, and how their zero-shot and fine-tuned performance compares to DL models trained from scratch. How do their zero-shot and fine-tuned performances compare to traditional DL models trained from scratch on target data? From a practical standpoint, what is the most effective strategy to obtain accurate, calibrated forecasts from LCS networks using these new tools?

To address this gap, this study provides an empirical benchmarking of TSFMs for $PM_{2.5}$ prediction using real-world LCS data, assessing Google's TimesFM and IBM's Granite against three established DL models (LSTM, CNN, Transformer) across 34 datasets from 10 cities and 6 countries, under two prediction strategies: (a) directly predicting RS-equivalent concentrations and (b) predicting raw LCS values with subsequent calibration.

Through this systematic investigation, we aim to provide actionable insights for deploying TSFMs in operational air quality monitoring. Our study is structured around three key objectives:

1. **Benchmarking TSFMs Against Traditional Approaches:** We compare the forecasting accuracy of TSFMs (in zero-shot and fine-tuned modes) against traditional DL models trained from scratch. Zero-shot TSFMs are applied directly to each dataset without any task-specific training, providing empirical evidence of transferability to unseen sensor deployments across diverse geographic contexts. DL models and fine-tuned TSFMs are instead trained and evaluated using chronological train/test splits on each individual sensor dataset. This establishes whether the new paradigm can outperform the old one across both short-term and long-term prediction horizons.
2. **Evaluating Practical Deployment Strategies:** We investigate two practical pathways for obtaining calibrated, RS-equivalent forecasts from LCS data: (i) a direct prediction of RS values using models trained on aligned LCS-RS data, and (ii) an indirect, two-stage "combined strategy" where a TSFM forecast of LCS data is subsequently calibrated by a DL model. This comparison directly informs the design of effective prediction pipelines for settings where an LCS has previously been co-located with an RS.
3. **Analyzing the Role of Missing Data and Temporal Dynamics:** We analyze the impact of missing data on model performance and investigate how input window sizes and forecast horizons affect accuracy, providing practical guidance on data requirements and model configuration.

By addressing these objectives, this study does not propose a new forecasting model or calibration framework. Instead, it contributes a systematic, large-scale empirical comparison of existing TSFMs and traditional DL models under realistic deployment scenarios, offering practical guidance for practitioners in air quality monitoring. Additionally, it offers actionable insights into the optimal configuration of input and prediction windows, providing practical guidance for model deployment in low-cost sensor monitoring networks.

The remainder of this paper is structured as follows. Section 2 reviews related work on air quality forecasting and TSFMs. Section 3 describes the methodology, including problem formulation, data collection, pre-processing, model configurations, and evaluation metrics. Comprehensive results are presented and discussed in Section 4. Finally, Section 5 outlines open challenges and future research directions, while Section 6 concludes the paper.

2. Related work

We provide an overview of existing research relevant to this study. Section 2.1 summarizes the evolution of air quality forecasting methods, from statistical and physics-based models to modern DL architectures. Section 2.2 then introduces the paradigm of TSFMs, discussing their capabilities, limitations, and recent applications in environmental forecasting.

2.1. Air quality forecasting

Air quality prediction has emerged as a critical application domain for DL, driven by the ability of neural architectures to model complex, nonlinear relationships in heterogeneous environmental data (Zhang et al., 2024b). A broad spectrum of DL models has been applied to $PM_{2.5}$ forecasting, encompassing recurrent neural networks (RNNs) and their variants (LSTM, GRU), convolutional neural networks (CNNs), graph neural networks (GNNs), and attention-based architectures including Transformers (Zhou et al., 2024). These approaches address forecasting horizons ranging from short-term (1 h) to long-term (up to 72 h), with prediction errors generally increasing as the forecast window extends.

Recent work emphasizes spatiotemporal dependencies. For example, Song et al. (2022) used attentive multi-task learning with sequence-to-sequence models for simultaneous PM_{10} and $PM_{2.5}$ prediction, while Illescas-Martinez et al. (2025) demonstrated real-time GRU-based forecasting using distributed urban sensor networks. Hybrid approaches combining feature selection with encoder-decoder architectures have also shown promise (Abalo-García et al., 2025; Nguyen et al., 2021). For longer prediction horizons, Zheng et al. (2024) proposed CEEMD-MsI, a multi-stream Informer model that jointly captures temporal and spatial patterns, achieving state-of-the-art performance across diverse environments. A parallel line of research captures spatial correlations via graph-based methods. Su et al. (2023) proposed GCN-TAG, combining distance- and POI-based adjacency matrices with GRU encoders, outperforming LSTM and STGCN baselines over 48-h horizons. Wang et al. (2024) developed GC-SRTCN-L, integrating mix-hop graph convolution with a residual TCN, outperforming ARIMA, LSTM, and graph-based baselines on Beijing data. Zhao et al. (2025) advanced hypergraph modeling with ST-HGAT for higher-order spatial relationships, while Zhu et al. (2026) combined improved graph attention networks with variational mode decomposition and BiLSTM layers. Notwithstanding the strength of multi-station graph approaches, real-world deployment often involves a single sensor instead of a distributed network. Consequently, our study concentrates on *per-sensor forecasting*, acknowledging that graph-based spatial modeling is an orthogonal direction outside the scope of our work.

In parallel to DL approaches, traditional air quality forecasting has relied on physics-based models such as CMAQ (Agency, 2017; Wang et al., 2009) and WRF-Chem (Kumar et al., 2012; Skamarock & Klemp, 2008), which simulate the physical and chemical processes governing pollution transport and formation. While these models offer interpretability and physical consistency, they incur high computational

costs and face challenges in domain adaptation. Statistical methods including ARIMA and Kalman filters are widely adopted as interpretable baselines for air quality forecasting; however, they often struggle to capture the complex spatiotemporal dependencies underlying air pollution dynamics (Arsov et al., 2021; Iskandaryan et al., 2020; Ramadan et al., 2024; Rougier et al., 2023; Saboia, 1977). Recent work has also addressed the critical role of input data quality, particularly from LCSs, with Feng et al. (2025) proposing physics-informed calibration pipelines that combine Köhler theory and Mie scattering corrections with cluster-specific refinement to improve agreement with reference monitors.

Despite these advances, DL models for air quality forecasting are typically trained from scratch on location-specific data, limiting their generalizability across geographic domains and sensor types. This limitation motivates the exploration of TSFMs, which leverage large-scale pre-training to capture universal temporal patterns and enable zero-shot forecasting across diverse environmental conditions.

2.2. Foundation models for time series analysis

Time Series Foundation Models extend FMs to temporal data, enabling cross-domain generalization. However, their application faces challenges including data leakage, lack of training transparency, and limited effectiveness in rare event prediction (Shyalika et al., 2024). Recent work spans two main directions: pre-training FMs specifically for time series, and adapting LLMs via NLP-inspired strategies (Jin et al., 2024). Notable contributions include ChatTime for multimodal integration (Wang et al., 2025a), medical imputation (Bai et al., 2024), finance (Fu et al., 2025), and benchmark evaluations (Aksu et al., 2024). Kim et al. (2024) question the efficacy of self-attention in transformer-based TSFMs, while Shi et al. (2024) show scaling laws translate to time series with Time-MoE, a sparse mixture-of-experts model trained on 300 billion time points.

In the environmental domain, Aurora (Bodnar et al., 2025) outperforms traditional chemical transport models for air quality prediction, Fan et al. (2024) aligns air quality tokens with LLM embeddings for few-shot forecasting, and end-to-end data-driven FMs have shown promise for high-resolution weather forecasting (Allen et al., 2025).

For probabilistic forecasting, Lag-Llama (Rasul et al., 2023) is one of the first open-source FMs explicitly designed for time series. Its decoder-only transformer architecture, pretrained on heterogeneous time series corpora, enables strong zero-shot generalization across domains and delivers full predictive distributions rather than point forecasts, a crucial capability for air quality applications where uncertainty quantification is essential.

Empirical benchmarks reveal mixed results. Mulayim et al. (2024) evaluate several TSFMs on building energy and indoor temperature datasets, finding only marginal gains over statistical baselines such as AutoARIMA on unseen data modalities. Jiang et al. (2026) reach a more optimistic conclusion in the carbon credit pricing domain, showing that TSFMs substantially outperform classical Machine Learning and DL models at long daily horizons.

Recent work aims to uncover *why* TSFMs succeed or fail. Wang et al. (2025b) identify spectral shift (a mismatch between the dominant frequency components in downstream tasks and those represented during pre-training) as a key factor: when downstream series fall outside the pre-trained frequency spectrum, performance degrades systematically. Zhang and Gilpin (Zhang & Gilpin, 2025) show that TSFMs can forecast chaotic systems without being trained on them, suggesting that they learn general statistical patterns rather than system-specific dynamics. For transformer-based TSFMs like TimesFM, the patching mechanism (Das et al., 2024) may help capture local patterns while maintaining computational efficiency, while mixer-based architectures like Granite (Ekambaram et al., 2024) rely on separate temporal and channel-mixing MLPs to model intra-channel dynamics and cross-channel dependencies.

However, none of these studies provides a complete mechanistic explanation of *why* TSFMs outperform DL models when they do. The source of the advantage, whether from pre-training scale, architectural choices, implicit autoregressive bias, or the diversity of training data, remains an open question. Addressing this gap and challenges such as cross-site generalization, data sparsity, rare event prediction, and sensor-specific calibration remains essential for fully leveraging TSFMs in $PM_{2.5}$ forecasting with LCS networks.

Table 1
Datasets specifications. Each dataset is subject to CC BY 4.0.

City *[LCS Type]* (Country)	Sampling Period Months [Date range]
Aosta *[SPS30]* Reggio *[SPS30]* Trento *[SPS30]* (IT) (Casari & Po, 2023)	5 M [2024-02 – 2024-06] 3 M [2024-09 – 2024-12] 3 M [2024-11 – 2025-02]
Badajoz *[OPC-N3]* (ES) (Arroyo et al., 2021)	2 M [2021-03 – 2021-05]
Calgary *[PMS5003]* (CA) (Si, 2019; Si et al., 2020)	5 M [2018-12 – 2019-04]
Bangalore *[PurpleAir PMS5003]* Delhi *[PurpleAir PMS5003]* Hamirpur *[PurpleAir PMS5003]* (IN) (Campmier et al., 2023a,b)	13 M [2019-06 – 2020-07] 18 M [2018-07 – 2020-01] 9 M [2020-03 – 2021-01]
Lima *[IQAir]* Lima *[AirBeam PMS7003]* (PE) (Villanueva et al., 2023)	1 M [2021-11 – 2021-12] 2 M [2021-11 – 2022-01]
Southampton *[SPS30, PMS5003]* (UK) (Bulot, 2022; Bulot et al., 2023)	12 M [2020-07 – 2021-07]

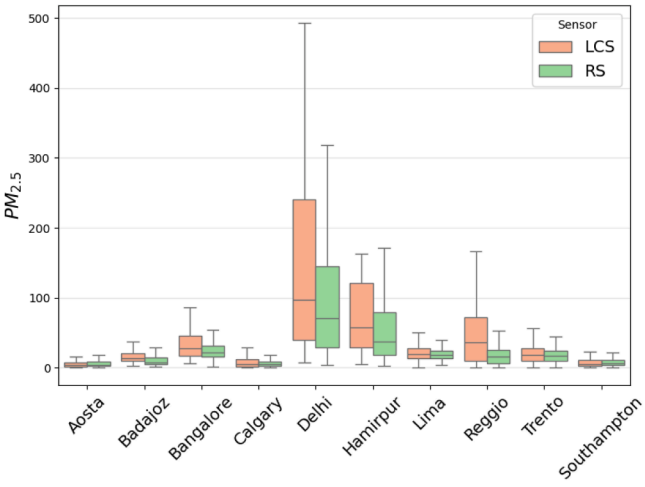


Fig. 1. $PM_{2.5}$ distribution for each location from RSs and LCSs.

3. Methodology

We evaluate $PM_{2.5}$ forecasting models using real-world LCS data from diverse locations. This section describes the problem formulation, datasets, pre-processing, model configurations, and evaluation metrics.

3.1. Problem formulation

Let x_t denote the $PM_{2.5}$ concentration measured by a LCS at time t, and y_t the corresponding RS measurement. Meteorological variables (temperature, humidity, pressure, wind speed) at time t are denoted as $m_t \in \mathbb{R}^M$.

Given a historical window of length I, we aim to obtain RS-equivalent forecasts $\hat{y}_{t+1:t+H}$ for the subsequent H hours. We consider two strategies:

- **Direct prediction:** A prediction model $\mathcal{F}_{\text{direct}}$ maps historical data directly to $\hat{y}$:

$$\hat{y}_{t+1:t+H} = \mathcal{F}_{\text{direct}}(\mathbf{X}_{t-I+1:t}, \mathbf{M}_{t-I+1:t})$$

- **Two-stage combined strategy:** A prediction model $\mathcal{F}_{\text{pred}}$ first predicts LCS values $\hat{x}_{t+1:t+H}$, which are then transformed into $\hat{y}_{t+1:t+H}$ by a separate calibration model $\mathcal{G}_{\text{calib}}$:

$$\hat{y}_{t+1:t+H} = \mathcal{G}_{\text{calib}}(\mathcal{F}_{\text{pred}}(\mathbf{X}_{t-I+1:t}, \mathbf{M}_{t-I+1:t}))$$

The application scope of the two strategies is the same: predicting accurate RS-equivalent measurements when co-located LCS-RS training data are available. In direct prediction, these data are used to train the prediction model. In the two-stage strategy, the same data are used to train the calibration model that maps LCS forecasts to RS-equivalent values. Both strategies therefore require the same type of training data.

3.2. Data collection and pre-processing

Geographic diversification is essential to assess the generalizability of the methodology, as air quality is influenced by meteorological conditions, urban context, and local emission sources. We use publicly available $PM_{2.5}$ data from 10 cities across 6 countries (Table 1), comprising 34 paired LCS-RS time series. As shown in Fig. 1, $PM_{2.5}$ concentrations vary significantly across locations, from severe pollution in Delhi to well-aligned, low concentrations in Aosta and Southampton, providing a robust basis for assessing model performance under diverse environmental conditions.

To ensure data quality and uniformity, each sensor dataset underwent a standardized pre-processing pipeline. First, LCS and RS measurements were temporally aligned to ensure one-to-one correspondence. Selected features included $PM_{2.5}$ concentrations from both sensor types and meteorological variables (pressure, relative humidity, temperature, wind speed); precipitation was excluded due to its rarity outside the Southampton site. Outliers in LCS data were detected using the Interquartile Range (IQR) method and replaced with missing values; corresponding RS and meteorological values at the same timestamps were also marked as missing to preserve alignment. Although this approach is relatively simple, it has been shown to be effective in identifying anomalous observations in environmental and sensor data (AlSalehy & Bailey, 2025; Buchhorn et al., 2024; Dallah & Sulieman, 2022).

Missing data were subsequently filled using linear interpolation, with flagged interpolated values excluded from performance evaluation. This method was adopted to maintain temporal continuity in the time series while avoiding the introduction of artificial variability. Although some gaps may span multiple time steps, linear interpolation provides a simple and robust approximation of the underlying trend and is widely used in environmental time-series pre-processing when the goal is to reconstruct missing observations without imposing strong modeling assumptions (Tawakuli et al., 2025).

Finally, all data were resampled to hourly resolution to ensure temporal continuity. The resulting cleaned and aligned datasets are suitable for downstream tasks within our forecasting framework.

3.3. Models

We employ three DL models for calibration and prediction, along with two TSFMs used for prediction in both zero-shot and fine-tuned settings, resulting in a total of three TSFM configurations.

- **Deep Learning (DL) Models**
  CNNs, while traditionally used in image processing (Cun et al., 1990), have shown strong performance in structured time series analysis. By applying convolutional filters across the temporal axis of the data, CNNs can capture local patterns and short-term dependencies. This makes them suitable for detecting rapid fluctuations in pollutant levels or identifying recurring patterns associated with diurnal or weekly cycles.
  LSTM networks are a type of RNN specifically designed to capture long-range dependencies in sequential data (Hochreiter & Schmidhuber, 1997). Their gating mechanisms allow them to retain relevant information over extended time intervals while mitigating the vanishing gradient problem that hampers RNNs.
  Transformer models, originally developed for natural language processing tasks, have recently demonstrated state-of-the-art results in time series forecasting. Unlike RNNs, Transformers rely on self-attention mechanisms that enable them to model global dependencies in the input sequence without the need for sequential processing (Vaswani et al., 2023). This parallelization enhances computational efficiency and allows the model to focus on relevant temporal features regardless of their position in the sequence.
- **Time Series Foundation Models (TSFMs)**
  Google's TimesFM 2.0[1] (Das et al., 2024) is a large Transformer model adapted for continuous time-series forecasting, with a decoder-only architecture. The raw univariate series is divided into non-overlapping patches of a fixed length. Each patch is then linearly embedded into the dimensional space of the model. An optional small frequency embedding can be added to each patch to encode the sampling rate. This sequence is then passed through stacked decoder layers, each of which consists of residual connections wrapped around an RMS-normalised masked multi-head self-attention layer and a feed-forward layer. During inference, the model produces per-patch representations that are projected via separate linear heads to yield mean and quantile forecasts over the desired horizon. For this study, we utilize the *xreg + timesfm* configuration (hereafter **TimesFM$_{ZS}$**), which integrates both historical $PM_{2.5}$ values and exogenous meteorological covariates within its attention framework, enabling the model to learn interactions between pollutant dynamics and weather conditions.
  IBM's Granite TTM-R2 model[2] (Ekambaram et al., 2025) is an ultra-compact mixer network designed for multivariate forecasting. Architecturally, it begins by projecting each time step's feature vector into a shared latent space of a specified dimension. The model is designed to process a feature vector at each time step. The sequence is then passed through alternating mixer blocks: first, a temporal-mixing MLP processes each channel independently across all time steps and is responsible for capturing intra-channel dynamics; then, a channel-mixing MLP operates at each time step and captures inter-channel relationships. Both sublayers use residual connections and layer normalisation to stabilise training, and a small learnable frequency embedding can be prepended to encode sampling rate information. After the mixer stack, the representation at the final position is fed through a lightweight linear head to produce the desired forecast. With fewer than one million parameters, Granite supports two usage modes: *zero-shot* (**Granite$_{ZS}$**), which leverages pre-trained patterns without additional training; and *fine-tuned* (**Granite$_{tuned}$**), which adjusts model weights on target data to capture domain-specific behaviors. Both configurations natively handle multivariate inputs, making them directly comparable to TimesFM$_{ZS}$ in their ability to integrate exogenous signals.

3.4. Experimental design

The experimental design aims to systematically evaluate the performance of the proposed $PM_{2.5}$ forecasting models across diverse tasks and environmental conditions. We structure our experiments to assess both the models' ability to *calibrate* LCS readings against RS measurements, and their ability to *forecast* future pollutant concentrations under different prediction strategies. After pre-processing, each dataset, corre-

[1] https://huggingface.co/google/timesfm-2.0-500m-pytorch
[2] https://huggingface.co/ibm-granite/granite-timeseries-ttm-r2

sponding to a single LCS-RS sensor pair, is organized chronologically and split into training and test sets, ensuring that temporal dependencies are preserved. DL models and fine-tuned TSFMs are trained and evaluated using chronological train/test splits on each individual sensor dataset. Zero-shot TSFMs are instead applied directly to each dataset without any task-specific training, and their performance across geographically diverse sites provides empirical evidence of transferability to unseen sensor deployments. A formal cross-site evaluation protocol remains outside the scope of this study and is identified as a direction for future work (Section 5).

In the following sections, we detail the experimental scenarios, input window selection, prediction horizons, and the specific configurations of each model tested. This structured setup provides a consistent basis for comparing model performance across locations and time scales.

3.4.1. Calibration task

Calibration aims to estimate the current RS value y_t from historical LCS and meteorological data. We implement this using the three DL architectures trained from scratch on the training dataset, which includes historical aligned LCS-RS time series and meteorological variables, using RS data as target. During test, each model takes an input window $\mathbf{X}_{t-I:t} = [x_{t-I}, \dots, x_t]$ and corresponding meteorological variables $\mathbf{M}_{t-I:t}$ to predict y_t. Input windows of 1, 6, 12, 24, and 48 h are evaluated to assess the impact of historical context on calibration accuracy (Table 2, top section), yielding a total of 15 tested configurations.

3.4.2. Prediction tasks

We implement prediction models as follows:

Scenario X (LCS prediction): Models forecast raw LCS values $\hat{x}_{t+1:t+H}$ using historical LCS data $\mathbf{X}_{t-I+1:t}$ and meteorological variables $\mathbf{M}_{t-I+1:t}$ as input. The training dataset includes LCS values (used as target during training) and meteorological variables. We evaluate DL models (CNN, LSTM, Transformer) trained from scratch on the training dataset, pretrained TSFMs without any additional training or fine-tuning ($\text{TimesFM}_{\text{ZS}}$ and $\text{Granite}_{\text{ZS}}$), and $\text{Granite}_{\text{tuned}}$ fine-tuned on the training dataset.

Scenario Y (Direct RS prediction): Models directly forecast RS values $\hat{y}_{t+1:t+H}$ from the same inputs of Scenario X, i.e., historical LCS data $\mathbf{X}_{t-I+1:t}$ and meteorological variables $\mathbf{M}_{t-I+1:t}$. No historical RS values are included in the input window. The training set includes historical aligned LCS-RS time series and meteorological variables, using RS data as target during training. This implements the direct prediction strategy from Section 3.1. We evaluate DL models and $\text{Granite}_{\text{tuned}}$ that are trained from scratch and fine-tuned on the training dataset, respectively. Zero-shot TSFMs are excluded as they cannot predict variables absent from the input sequence.

Combined Strategy (Two-stage strategy for RS prediction): This implements the two-stage approach introduced in Section 3.1. First, a TSFM from Scenario X generates LCS forecasts $\hat{x}_{t+1:t+H}$. These predictions are then passed to a calibration model (CNN, LSTM, or Transformer) to produce the final RS-equivalent forecasts $\hat{y}_{t+1:t+H}$.

Table 2 provides a complete overview of all experimental configurations. For all prediction tasks, we evaluate prediction horizons $H \in \{1, 12, 24, 48, 72\}$ h. The choice of input window length I depends on the model architecture and its associated constraints. For DL models and $\text{TimesFM}_{\text{ZS}}$, the flexibility of these architectures allows input windows of 12, 24, 48, 72, and 512 h. To ensure consistency across experiments, we constrain the forecasting horizon H such that it does not exceed approximately twice the input window length I, i.e., $H \leq 2I$. For instance, when $I = 12$, we evaluate horizons $H \in \{1, 12, 24\}$. In contrast, Granite models impose fixed input window sizes of 52, 90, 180, and 512 h due to architectural design choices. Consequently, for each prediction horizon H, we select the smallest available input window I that is compatible with the task. For example, $I \geq 52$ is used for $H \in \{1, 12\}$, $I \geq 90$ for $H \in \{1, 12, 24\}$, and so on.

The input window advances one step at a time, producing overlapping predictions for each target time step when $H > 1$. When multiple predictions are available for the same target, we aggregate them by averaging to obtain the final forecast. This approach ensures consistency and robustness, while accounting for varying numbers of valid predictions per timestamp due to series boundaries or excluded interpolated data.

In total, we evaluate 116 Scenario X configurations, 80 Scenario Y configurations, and 1740 Combined Strategy configurations.

3.4.3. Hyperparameter optimization

Hyperparameters for all DL models (LSTM, CNN, Transformer) were systematically optimized using Optuna. Optimization was performed independently for each model, dataset, and task, using a validation set to evaluate configurations based on Mean Squared Error. Early stopping with a patience of 10 epochs was applied to prevent overfitting, and up to 50 trials were conducted per combination. The best-performing configuration (lowest validation loss) was selected, with the tested ranges reported in Table 3. The full list of the best configurations for each city is available in the supplementary materials.

$\text{TimesFM}_{\text{ZS}}$ and $\text{Granite}_{\text{ZS}}$ were employed without performing any hyperparameter tuning, using the default settings recommended by their developers. For $\text{Granite}_{\text{tuned}}$, a batch size of 8 and a learning rate of 0.001 were adopted and kept fixed across all datasets, without any hyperparameter optimization. All remaining hyperparameters were retained from the default configuration. We acknowledge that this creates an asymmetry with the tuned DL models. However, this choice is consistent with the intended out-of-the-box usage of TSFMs. Furthermore, supplementary experiments using the hyperparameter search space reported in Table 3 confirmed that $\text{Granite}_{\text{tuned}}$ exhibits low sensitivity to hyperparameter choices during fine-tuning, with default settings yielding representative performance.

3.5. Evaluation metrics

The evaluation metrics used in the experimental analysis are summarized in Table 4. Each metric compares the calibrated or predicted values with the ground-truth that depends on the task. Interpolated values are excluded from all calculations.

4. Results

In this section, we present the experimental findings of our study. The analysis is structured as follows: Section 4.1 validates the effectiveness of our data pre-processing pipeline, while Section 4.2 provides an exploratory analysis of the environmental variables across cities, highlighting their diversity in terms of both meteorological conditions and pollution levels. Section 4.3 investigates the performance of the calibration models and identifies optimal input window configurations for the DL architectures. Forecasting performance under Scenario X and Scenario Y is addressed in Sections 4.4 and 4.5, respectively, examining the impact of input window size, forecast horizon, missing data, and model architecture. Section 4.6 evaluates the Combined Strategy, and finally, Section 4.7 presents a city-level analysis comparing the best configurations across all approaches.

Experiments were conducted on 34 geographically distributed datasets, each corresponding to the time series from a single sensor. A 75-25 train/test split was employed, and model training and evaluation were performed independently for each dataset to account for local patterns and sensor-specific behaviors. Zero-shot TSFMs ($\text{TimesFM}_{\text{ZS}}$ and $\text{Granite}_{\text{ZS}}$) are applied directly to each dataset without any task-specific training; their performance across datasets from 10 cities and 6 countries therefore provides empirical evidence of transferability to unseen sensor deployments. DL models and $\text{Granite}_{\text{tuned}}$ are instead trained on the first 75% of each sensor's time series and evaluated on the remaining

Table 2
Summary of models and their configurations.

Model	Input Window (I)	Input Features	Output/Target	Prediction Window (H)
Calibration				
CNN / LSTM / Transformer	1/6/12/24/48 h	$\{\mathbf{X}_{t-I:t}, \mathbf{M}_{t-I:t}\}$	y_t	
Prediction Scenario X				
CNN / LSTM / Transformer	12/24/48/72/512 h	$\{\mathbf{X}_{t-I+1:t}, \mathbf{M}_{t-I+1:t}\}$	$\mathbf{X}_{t+1:t+H}$	1/12/24/48/72 h
$\text{TimesFM}_{\text{ZS}}$	12/24/48/72/512 h	$\{\mathbf{X}_{t-I+1:t}, \mathbf{M}_{t-I+1:t}\}$	$\mathbf{X}_{t+1:t+H}$	1/12/24/48/72 h
$\text{Granite}_{\text{ZS}}$	52/90/180/512 h	$\{\mathbf{X}_{t-I+1:t}, \mathbf{M}_{t-I+1:t}\}$	$\mathbf{X}_{t+1:t+H}$	1/12/24/48/72 h
$\text{Granite}_{\text{tuned}}$	52/90/180/512 h	$\{\mathbf{X}_{t-I+1:t}, \mathbf{M}_{t-I+1:t}\}$	$\mathbf{X}_{t+1:t+H}$	1/12/24/48/72 h
Prediction Scenario Y				
CNN / LSTM / Transformer	12/24/48/72/512 h	$\{\mathbf{X}_{t-I+1:t}, \mathbf{M}_{t-I+1:t}\}$	$\mathbf{Y}_{t+1:t+H}$	1/12/24/48/72 h
$\text{Granite}_{\text{tuned}}$	52/90/180/512 h	$\{\mathbf{X}_{t-I+1:t}, \mathbf{M}_{t-I+1:t}\}$	$\mathbf{Y}_{t+1:t+H}$	1/12/24/48/72 h
Combined Strategy (Prediction Scenario X + Calibration)				
$\text{TimesFM}_{\text{ZS}}$ + CNN / LSTM / Transformer	12/24/48/72/512 h	$\{\mathbf{X}_{t-I+1:t}, \mathbf{M}_{t-I+1:t}\}$	$\mathbf{Y}_{t+1:t+H}$	1/12/24/48/72 h
$\text{Granite}_{\text{ZS}}$ + CNN / LSTM / Transformer	52/90/180/512 h	$\{\mathbf{X}_{t-I+1:t}, \mathbf{M}_{t-I+1:t}\}$	$\mathbf{Y}_{t+1:t+H}$	1/12/24/48/72 h
$\text{Granite}_{\text{tuned}}$ + CNN / LSTM / Transformer	52/90/180/512 h	$\{\mathbf{X}_{t-I+1:t}, \mathbf{M}_{t-I+1:t}\}$	$\mathbf{Y}_{t+1:t+H}$	1/12/24/48/72 h

Table 3
Ranges of hyperparameters explored for each DL model type.

Hyperparameter	CNN	LSTM	Transformer	$\text{Granite}_{\text{tuned}}$
Learning rate	$[10^{-5}, 10^{-2}]$ (log)	$[10^{-5}, 10^{-2}]$ (log)	$[10^{-5}, 10^{-2}]$ (log)	$[10^{-5}, 10^{-3}]$ (log)
Dropout rate	[0.0, 0.8]	[0.0, 0.8]	[0.0, 0.8]	[0.0, 0.3]
Batch size	{4, 8, 16, 32}	{4, 8, 16, 32}	{4, 8, 16, 32}	{4, 8, 16}
Channels	[16, 128] (step 16)	—	—	—
Kernel size	[1, 7]	—	—	—
Hidden units	—	[32, 256] (step 32)	—	—
Number of layers	—	[1, 4]	[1, 3]	—
Model dimension	—	—	{32, 64, 128}	—
Attention heads	—	—	{2, 4, 8}	—
Feedforward dim.	—	—	[64, 256] (step 64)	—
Warmup steps	—	—	—	[0, 500]
Learning Rate Scheduler	—	—	—	{*linear*, *cosine*, *cosine with restarts*}

25%, measuring within-site forecasting accuracy. A formal leave-one-location-out or cross-city evaluation is not conducted in this study.

The training set comprises approximately 126,000 observations across all datasets, with an average of 4000 observations per dataset, while the test set contains around 41,000 observations, averaging 1300 measurements per dataset. To ensure unbiased aggregated results, average performance across cities is computed using a **one-sensor-per-city subset (OSC)**, which includes a single representative dataset from each of the 10 cities. This balanced subset ensures that each city contributes equally to the calculation of average metrics, preventing cities with multiple sensors from disproportionately influencing the results. Full results for all 34 individual datasets are provided in the accompanying supplementary materials. In A, we report some detailed visual examples of a representative sensor in Aosta 1.

All experiments were carried out on the Leonardo supercomputer hosted by CINECA. Each node in the Booster partition is equipped with four NVIDIA A100 GPUs, each with 64 GB of VRAM. Training of CNN, LSTM, and Transformer models utilized all four GPUs per node, allowing each model to complete training in approximately 15–25 min, depending on architecture and dataset. Inference with TSFM models in zero-shot mode was performed on a single GPU, requiring less than 5 min per experiment, while fine-tuning of the Granite model on a single GPU took under 15 min for all configurations.

4.1. Data analysis and pre-processing outcomes

The pre-processed datasets are available in the supplementary materials. The effectiveness of the pre-processing pipeline, including outlier removal, resampling, and interpolation, is quantitatively assessed in Fig. 2. This figure compares the alignment of LCS measurements with RS data before and after pre-processing across MAE, RMSE and R^2 metrics, illustrating clear improvements in data quality. The decrease in RMSE following pre-processing indicates that the procedure effectively suppresses anomalous peaks, consistent with the metric's sensitivity to large deviations. Fig. 3 shows the percentage of interpolated data across all datasets, for both training and test sets. Most cities exhibit very low levels of interpolation, indicating high data completeness and/or low presence of outliers. In contrast, Indian cities such as Bangalore, Delhi, and Hamirpur present substantially higher percentages of interpolated data, consistently affecting both training and test sets. This highlights a marked variability in data quality across datasets, with potential implications for model robustness and generalization.

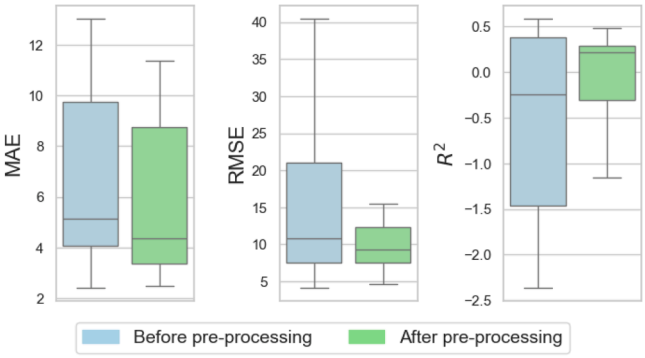


Fig. 2. Performance metrics before and after pre-processing.

Table 5 provides summary statistics of the datasets organized by city. For each city, the number of distinct sensors is reported (corresponding to those detailed in Fig. 3), along with the average number of hourly observations in the final datasets (*After Pre-processing*) and the average

Table 4
Summary of metrics used in the experimental evaluation.

Name and Range	Formula	Definition	Advantages	Limitations
MAE $[0, +\infty)$	$\frac{1}{n}\sum_{i=1}^{n}\lvert z_i - \hat{z}_i\rvert$	The Mean Absolute Error (MAE) quantifies the average magnitude of the absolute differences between predicted values $\hat{z}_i$ and observed values z_i, without considering their direction.	Easily interpretable as it is expressed in the same unit as the target variable.	Scale-dependent; results are influenced by the magnitude of the data, limiting comparability across datasets with different value ranges.
RMSE $[0, +\infty)$	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}$	The Root Mean Square Error (RMSE) measures the square root of the average squared differences between predicted and observed values, giving greater weight to larger errors.	Strongly penalizes large deviations, making it effective for highlighting models that produce significant prediction errors.	Highly sensitive to outliers due to the squaring of errors; scale-dependent and therefore not directly comparable across datasets with different magnitudes.
SMAPE $[0, 1]$	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert z_i - \hat{z}_i\rvert}{(\lvert z_i\rvert + \lvert \hat{z}_i\rvert)/2}$	The Symmetric Mean Absolute Percentage Error (SMAPE) measures the relative difference between predicted and observed values using a symmetric percentage formulation that normalizes the error by the average magnitude of the two values.	Scale-independent, enabling comparison across datasets with different magnitudes; provides an interpretable percentage-based error measure.	Can become unstable when both predicted and observed values approach zero; may still introduce bias in the evaluation of very small values.
R² $(-\infty, 1]$	$1 - \frac{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}$	The coefficient of determination (R^2) represents the proportion of variance in the observed data explained by the predictive model relative to a baseline model using the mean of the observations.	Dimensionless metric that provides an intuitive measure of goodness-of-fit.	Does not directly quantify prediction error; highly dependent on the variance of the observed data; identical absolute errors may lead to very different R^2 values depending on the variability of the ground truth.

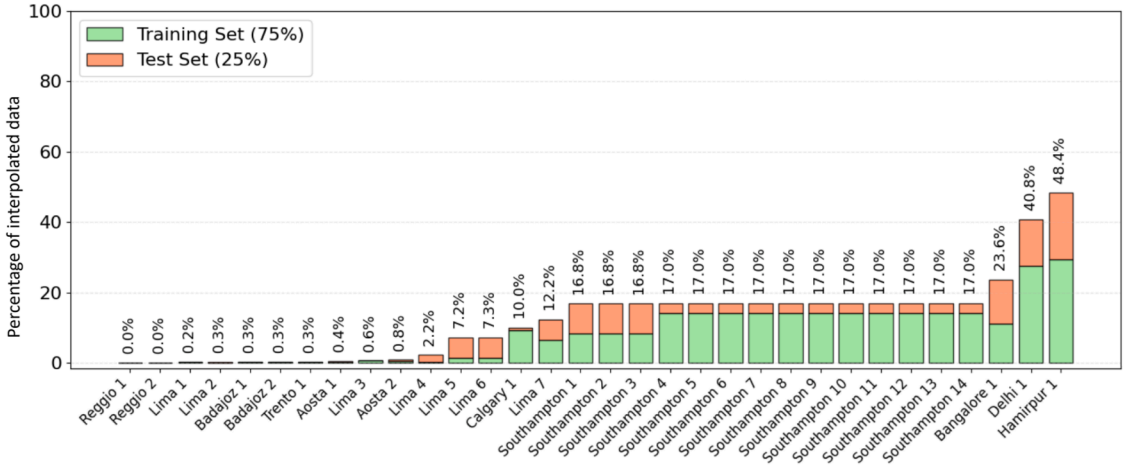


Fig. 3. Interpolated data distribution across datasets.

Table 5
City-level summary of sensors, observations, and interpolated data.

City (Country)	No. of Sensors	Avg No. of Observations	
		After Pre-proc	Interpolated (%)
Aosta (IT)	2	3049	17 (0.56)
Badajoz (ES)	2	1581	5 (0.32)
Bangalore (IN)	1	9745	2304 (23.65)
Calgary (CA)	1	3361	335 (9.97)
Delhi (IN)	1	12,681	5177 (40.82)
Hamirpur (IN)	1	7053	3415 (48.42)
Lima (PE)	7	1184	56 (4.73)
Reggio (IT)	2	2486	0 (0.00)
Trento (IT)	1	2125	7 (0.33)
Southampton (UK)	16	7679	1302 (16.96)

number of interpolated observations (*Interpolated*) with the percentage over the total observations. The last column reflects the extent of linear interpolation applied to fill in missing values, thus indicating the completeness and overall reliability of each dataset. We chose not to further segment the datasets into shorter sequences, as the available data per sensor was already limited. This decision preserves as much temporal continuity as possible for effective model training and evaluation. Interpolated data are used as input to the models during training but are excluded from the calculation of performance metrics.

4.2. Exploratory analysis of environmental variables across cities

Fig. 4 presents the distribution of the main environmental variables in the OSC. The extended analysis conducted across all sensors is available in the supplementary materials and shows that the distributions of sensors within the same city are highly similar. For each location, we report the percentage of observations within predefined value ranges for $PM_{2.5}$, temperature, humidity, pressure, and wind speed. Overall, the figure highlights the strong heterogeneity of environmental conditions across the considered cities, both in terms of pollution levels and meteorological variables. Such variability suggests that models trained and evaluated on these datasets must account for markedly different climatic and environmental regimes.

PM concentration. Substantial variability in $PM_{2.5}$ concentrations is observed across the analyzed cities. Delhi exhibits markedly higher levels, with concentrations reaching values close to 500 $\mu g/m^3$. Hamirpur and Reggio also display relatively elevated concentrations, approaching 200 $\mu g/m^3$. In contrast, the remaining cities generally present $PM_{2.5}$ values below 100 $\mu g/m^3$, indicating considerably lower levels of air pollution. Furthermore, lower variability in $PM_{2.5}$ concentrations is observed in the cities of Aosta, Badajoz, Calgary, Lima, Trento, and Southampton compared with the other cities analyzed.

Temperature. Temperature ranges vary considerably across locations, reflecting their distinct climatic conditions. Calgary is characterized by extremely low temperatures, spanning approximately from −30°C to 15°C, while Aosta and Trento show ranges between about −5°C and 30°C and −5°C to 20°C, respectively. Bangalore exhibits consistently high temperatures, ranging from 20°C up to 45°C, similarly to Delhi and Hamirpur, whose temperatures vary approximately between 10°C and 45°C. Badajoz presents a wide interval from 5°C to 40°C, whereas Lima shows a much narrower and stable range, approximately between 15°C and 25°C. The Southampton dataset spans a broad interval from about −5°C to 35°C.

Humidity. Different humidity distribution patterns can be observed. Aosta, Badajoz, Bangalore, Hamirpur, and Trento display values broadly distributed across the full 0–100% range, indicating high variability throughout the year. Calgary and Delhi show more restricted distributions (approximately 30–80% and 0–65%, respectively). Conversely, Lima, Reggio, the Southampton datasets are characterized by persistently high humidity levels, with values frequently exceeding 60%.

Pressure. Atmospheric pressure values generally lie within the range of approximately 980 to 1040 hPa. In some cities, such as Aosta, the distribution appears relatively uniform across this interval. In others, pressure values tend to concentrate within a narrower band, typically between 1000 and 1040 hPa, indicating more stable atmospheric conditions.

Wind speed. Wind speed shows a strongly right-skewed distribution across all sensors, with most values concentrated at low speeds and only a few high-speed events forming a long tail. This indicates that wind conditions are generally weak, with limited variability and occasional stronger episodes. Wind speed data from Southampton are all around 216 Km/h.

Autocorrelation analysis. Fig. 5 shows the autocorrelation functions (ACFs) of $PM_{2.5}$ concentrations across the LCSs. The vertical dashed blue lines indicate the autocorrelation peaks, highlighting the dominant periodic components in the time series.

In Fig. 5a, corresponding to the datasets from Calgary and Southampton, the ACF exhibits a predominantly monotonic decay with increasing lag, indicating strong short-term persistence and limited pronounced cyclicity. Although weak fluctuations are visible at specific lags (approximately every 24 h), the overall pattern suggests that temporal dependence gradually diminishes without marked periodic oscillations.

In contrast, Fig. 5b, showing the remaining datasets, reveals a pronounced oscillatory pattern on a gradually declining trend. The regularly spaced peaks, roughly every 24 h, indicate a clear daily periodicity in $PM_{2.5}$ levels. The persistence of these peaks over multiple lags points to a stable and recurring daily cycle, likely driven by systematic emission patterns and meteorological conditions. Overall, while all datasets display significant short-term autocorrelation, the strength and periodic structure of temporal dependence vary substantially across cities.

These patterns are consistent with the autocorrelation analysis on RS measurements, confirming that the observed temporal dynamics are robust across sensor types.

4.3. Calibration

We evaluated the performance of three DL models for LCS calibration over five input windows (1h, 6h, 12h, 24h and 48h), as reported in Table 2.

We report in Fig. 6 the performance metrics computed on the OSC. The CNN and LSTM models exhibit quite stable calibration performance across the range of input window sizes and consistently outperform the Transformer. In contrast, the Transformer exhibits increasing errors as the length of the input time window increases. The LSTM achieves the highest median R^2 value with a 48-h input window, while the CNN performs best with a 24-h window.

The outliers observed in the MAE boxplot are primarily associated with the Delhi and Hamirpur datasets. As previously noted, these datasets present higher percentages of missing data and elevated PM concentrations, which have a stronger impact on MAE computation. Notably, the LSTM model demonstrates the lowest variability in MAE values across 12, 24, and 48 h windows, even when considering outliers.

Although **the best-performing DL model and input window vary considerably across cities and datasets** (meaning there is no clear overall winner), by comparing the boxplots and taking into account the outliers, the box width, and the median values, we can conclude that **a 12-h input window represents a reasonable compromise** for the datasets used in these experiments. Compared to shorter and longer windows, the 12-h configuration generally exhibits moderate dispersion, fewer extreme outliers, and competitive median values across both MAE and R^2.

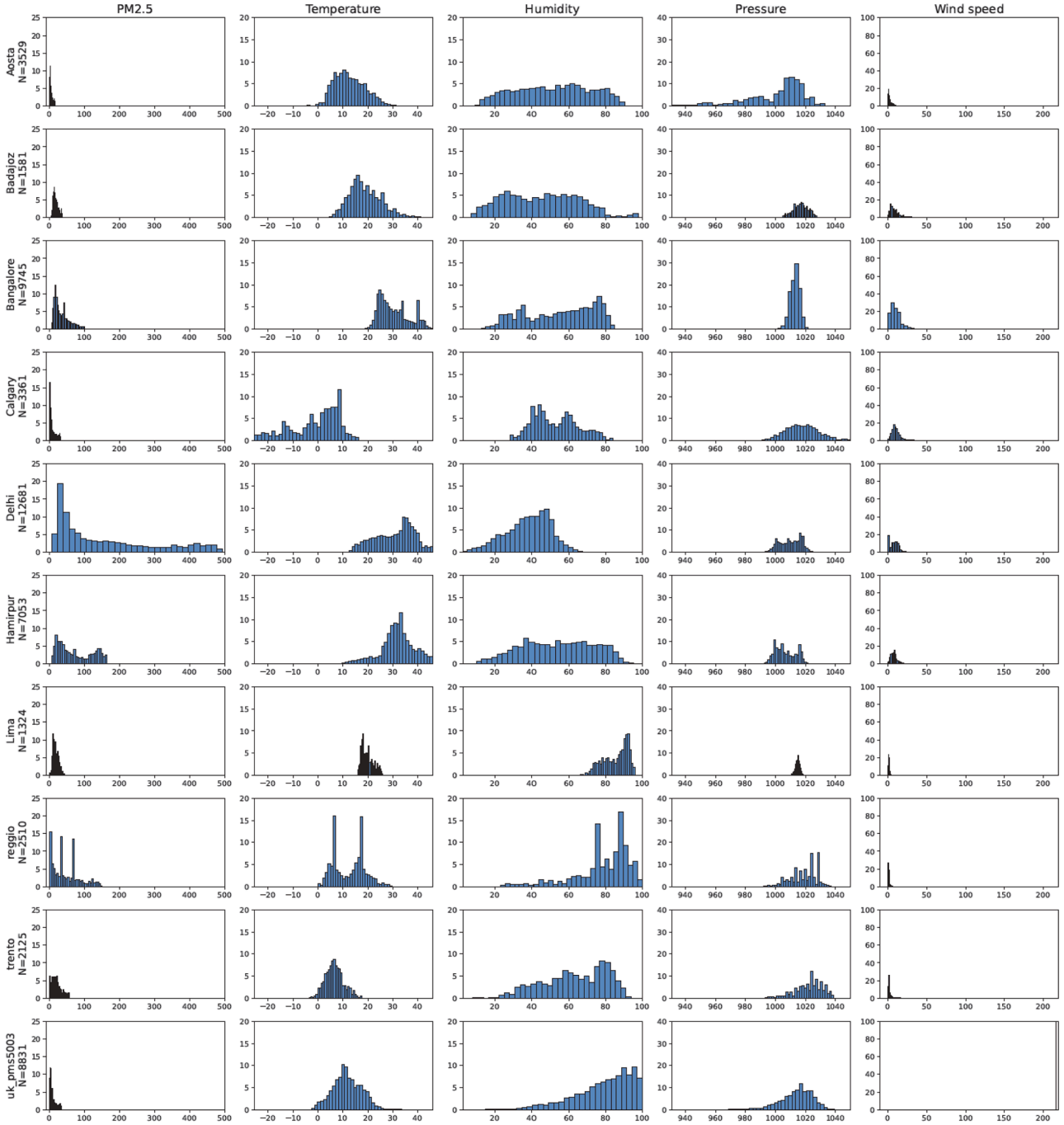


Fig. 4. Data distribution (%) for one-sensor-per-city subset (OSC).

4.4. Prediction scenario X

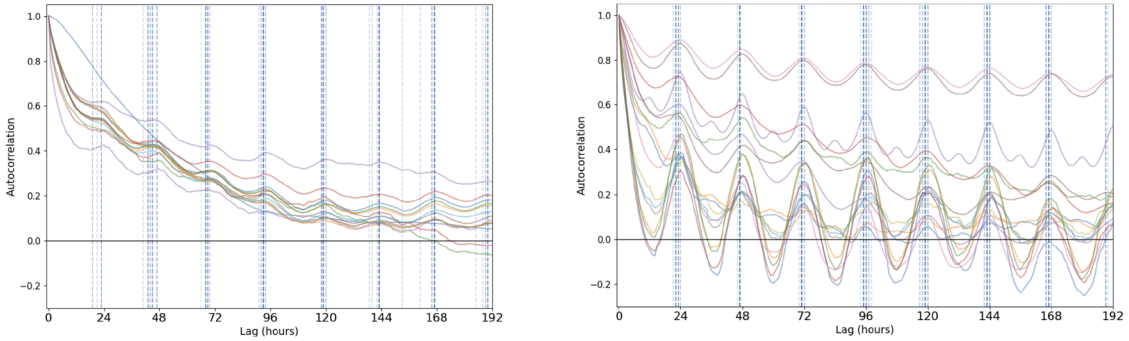
To forecast air quality using data acquired from LCS, we start by considering Scenario X.

First, we conduct a comparative evaluation of DL and TSFM models across multiple input and prediction windows on all datasets. Fig. 7 reports the average MAE across the OSC for each configuration of Scenario X, defined by the combination of input window, prediction window, and prediction model.

When analyzing the MAE values **for a prediction window of 1 h, a clear performance gap emerges between TSFMs and DL models, with TSFMs consistently outperforming DL approaches. As the prediction window increases, the performance gap between DL and TSFM models narrows, although TSFMs still maintain a consistent advantage**. For the longest prediction window (72 h), the DL models achieve performance levels close to those of the TSFMs, suggesting that for longer-term forecasts the relative benefit of the TSFMs is reduced.

Given that the **input window appears to have limited impact on performance**, the smallest input window compatible with the chosen prediction horizon can be considered preferable (e.g., a 24-h input window for a 48-h prediction window in the case of DL models). This suggests that long-term historical context is not necessarily required to obtain accurate forecasts.

With the exception of the $H = 1$ case for DL models, **the performance of all models deteriorates as the prediction horizon in-**



(a) Calgary and Southampton datasets (b) All other datasets

Fig. 5. Autocorrelation in $PM_{2.5}$ data from LCSs across different cities (vertical dashed lines corresponding to the autocorrelation peaks).

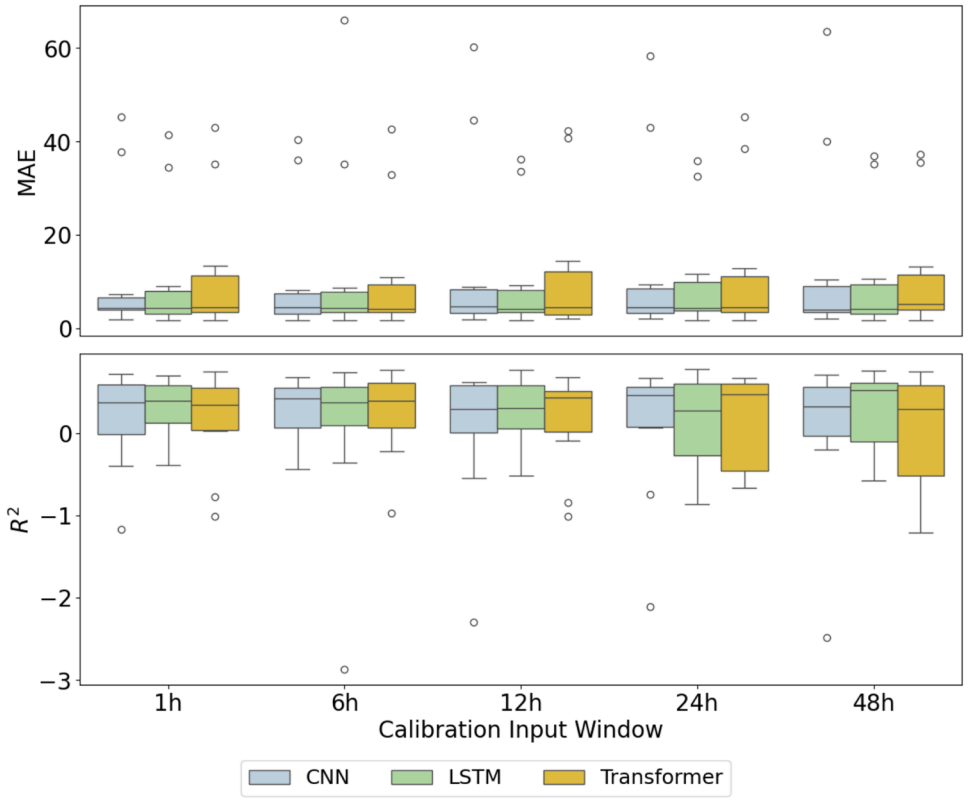


Fig. 6. Calibration performance for different input window sizes on OSC.

creases. For all DL models, the best-performing prediction window is consistently 12 h, whereas for TSFMs it is consistently 1 h. However, such a short forecasting horizon may be of limited practical relevance in real-world applications, where longer prediction intervals are typically required to support operational decision-making.

The analysis of R^2 values further supports these observations. As expected, R^2 decreases with increasing prediction window length. For 1-h forecasts, TSFMs achieve high R^2 values (mean between 0.75 and 0.85), which progressively decline for longer horizons, such as 12 h and beyond.

When comparing Granite_{ZS} and Granite_{tuned}, no substantial differences in predictive performance are observed across the evaluated metrics. This suggests that **fine-tuning the Granite model on the available data does not provide a significant advantage over its zero-shot configuration**. A possible explanation is that the pre-trained model already captures general temporal patterns, limiting the additional benefits that can be obtained through further task-specific training. It is also important to note that performance metrics are computed over the average of multiple overlapping predictions. For each hour in the test set, we obtain 12 different predictions (each corresponding to a different offset within the 12-h prediction window), and the evaluation is based on their average. This may smooth the results, but also introduces additional complexity in model assessment.

In summary, the configurations that offer the best trade-off between predictive performance and required historical data appear to be as follows: for DL models, an input window of 12 h combined with a prediction window of 12 h; for TSFMs, an input window of 52 h for Granite_{ZS} and Granite_{tuned}, and 12 h for Times_{ZS}, paired with a prediction window of 1 h.

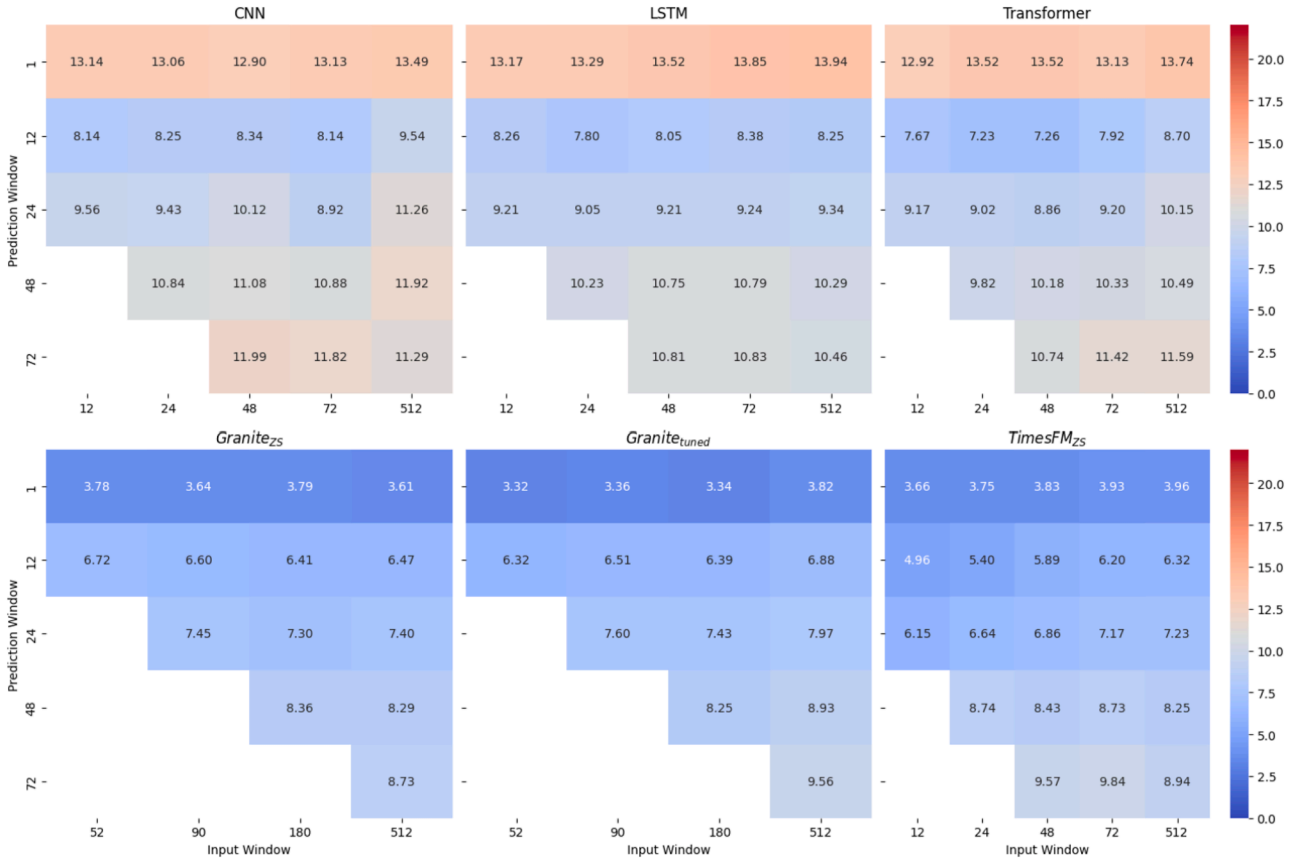


Fig. 7. Scenario X: average MAE of prediction models across the OSC, for different input and forecast window sizes.

4.4.1. *Missing data impact*

Missing observations are a common issue in real-world time series and may affect the reliability of forecasting models. In our setting, missing values arise either from originally unavailable observations or from samples removed during anomaly detection. In both cases, the missing entries are subsequently handled through interpolation before being used for model training and evaluation.

This experiment aims to assess whether the presence of missing data influences forecasting performance. In particular, we analyze whether datasets characterized by a higher proportion of missing values lead to a deterioration in prediction accuracy.

Fig. 8 reports the forecasting results obtained using the 512-h input window, which is the only configuration shared by all models, grouped according to the missing data ranges defined in Fig. 3. The SMAPE values illustrate the impact of missing observations on prediction accuracy. For readability purposes, only Transformer is shown as representative of DL models, as the other DL models exhibit very similar behavior; for the same reason, Granite$_{ZS}$ is excluded from the visualization.

A first relevant observation is that **datasets with a higher proportion of missing data** (Delhi and especially Hamirpur) **do not show a deterioration in forecasting performance** compared with the other datasets. On the contrary, Hamirpur consistently achieves the lowest SMAPE values in almost all configurations. The only exception is the Transformer with a 1-h prediction horizon, which corresponds to the configuration where DL models exhibit the weakest performance, as discussed before.

Another notable aspect concerns the effect of the forecasting horizon. For the Transformer model, the improvement in performance from the 1-h to 12-h prediction horizon is observed across all datasets. The two TSFMs exhibit very similar performance levels and show only marginal differences between the 1-h and 12-h forecasting horizons.

Overall, the results suggest that the proportion of missing data, when handled through interpolation, does not constitute a dominant factor affecting forecasting accuracy in these experiments.

4.4.2. *Comparison with baselines*

To better analyze the behavior of the models, we compared the performance of each model against a simple dummy baseline. This baseline consists of a *lagged window replication model* with rolling aggregation. Given an input window, the model simply replicates the observed values in the window to generate predictions for future timestamps. Because the rolling window advances one step at a time, this strategy produces multiple predictions for the same target timestamp, thus the final forecast is obtained by averaging all the available predictions for that timestamp. This baseline provides a simple reference to assess whether the evaluated models effectively exploit temporal patterns beyond a naive persistence-based strategy.

Fig. 9 reports the values of MAE, SMAPE, and R^2 across different forecast horizons for LSTM, TimesFM$_{ZS}$, and Granite$_{ZS}$. Since the performance was previously observed to vary only marginally with respect to the input window length, for each model and forecast horizon we report the results obtained with the smallest available input window. Overall, **both the dummy baseline and the TSFMs exhibit a gradual degradation in performance as the forecast horizon increases**. This behavior is expected, as longer prediction horizons introduce higher uncertainty and reduce the predictive value of recent observations. In contrast, **the LSTM model shows consistently larger errors** and a noticeably higher variability across all metrics and horizons.

When focusing on $H = 1$, the dummy baseline achieves lower MAE and SMAPE values compared to all models in the figure. This can be explained by the strong short-term autocorrelation present in the $PM_{2.5}$ data, as highlighted in Section 4.2. However, the TSFMs obtain higher

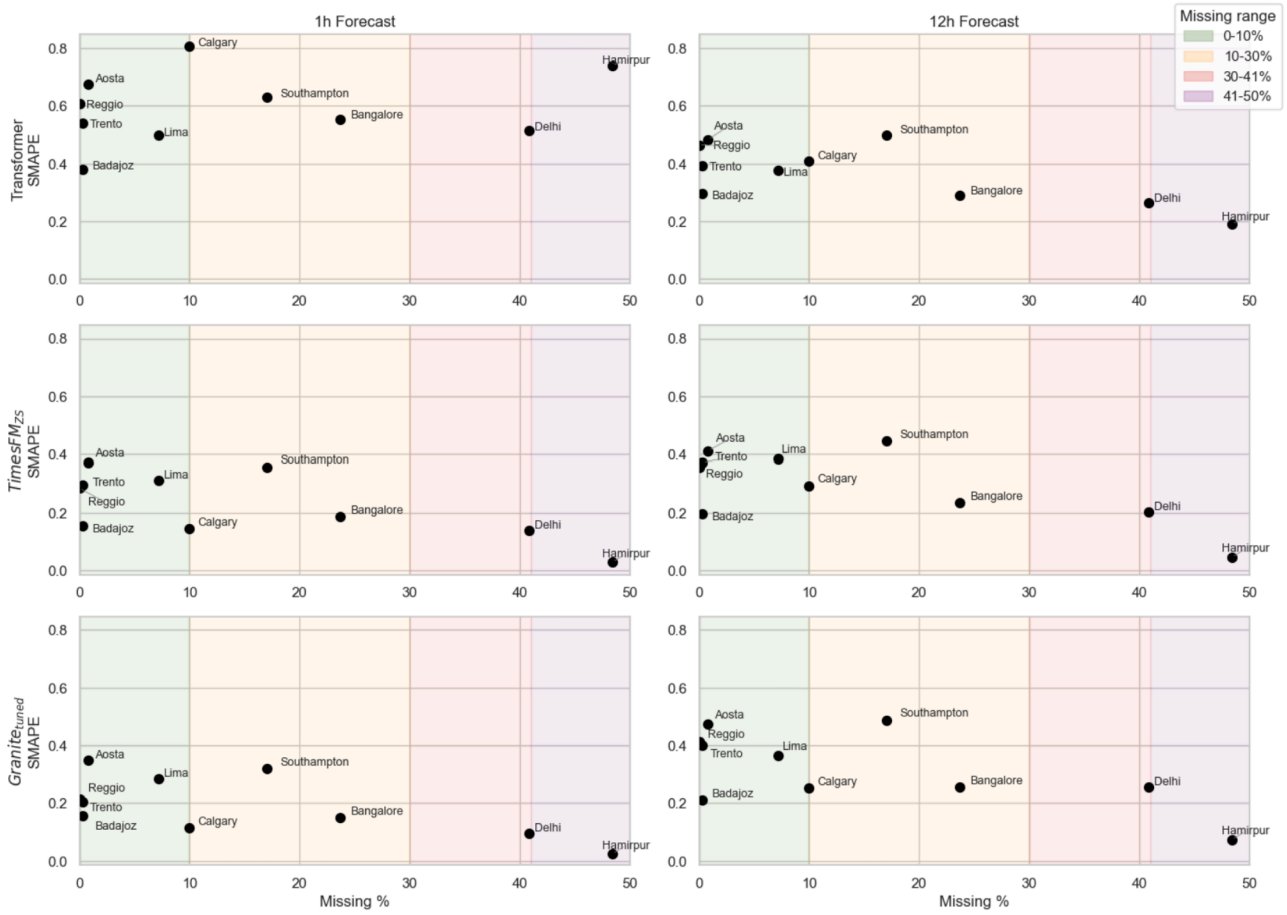


Fig. 8. Scenario X: comparison of Transformer, TimesFM$_{ZS}$ and Granite$_{ZS}$ performance across OSC. Results are shown based on the percentage of missing data of the dataset. Each model is evaluated using its minimal feasible input window.

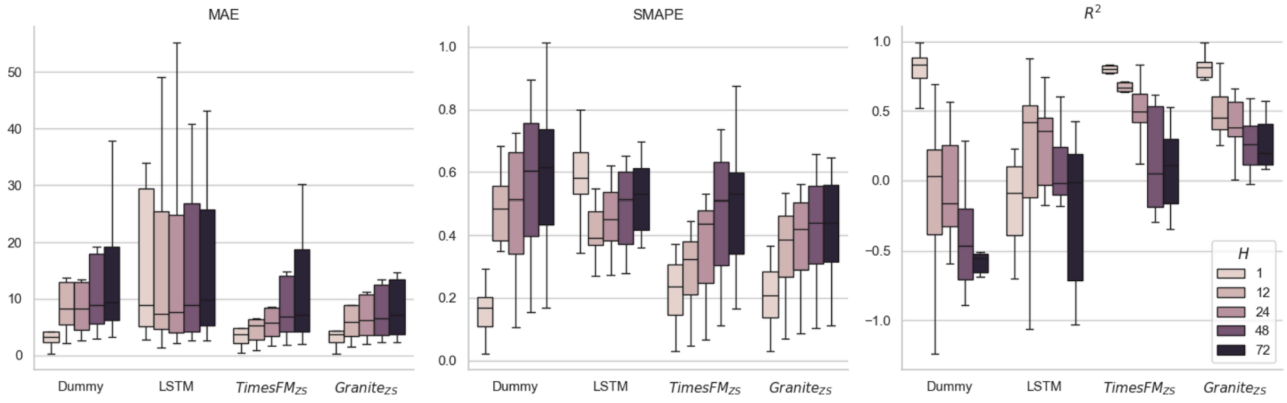


Fig. 9. Scenario X: comparison of LSTM, TimesFM$_{ZS}$, Granite$_{ZS}$ with respect to a dummy baseline model across OSC.

R^2 scores. This indicates that, although the baseline can produce numerically closer predictions in the very short term, the TSFMs better capture the underlying variability of the LCS measurements and explain a larger proportion of its variance.

For longer H, the advantage of the TSFMs becomes more evident. Both TimesFM$_{ZS}$ and Granite$_{ZS}$ consistently outperform the dummy baseline across all three metrics, suggesting that these models are able to exploit temporal patterns beyond simple persistence.

Fig. 10 presents the statistical significance analysis of the performance differences between each prediction model and the dummy baseline across different forecasting horizons. For each dataset and forecasting configuration, we assess whether the difference in MAE between the dummy baseline and the evaluated model is statistically significant by applying a block bootstrap procedure. Specifically, blocks of 24 consecutive observations are randomly sampled to generate bootstrap replicates of the error sequences. For each replicate, the MAE difference between the baseline and the model is recomputed, yielding an empirical distribution of performance differences. From this distribution, we estimate the mean difference and the associated 95% confidence interval. A model is considered significantly better than the baseline when the

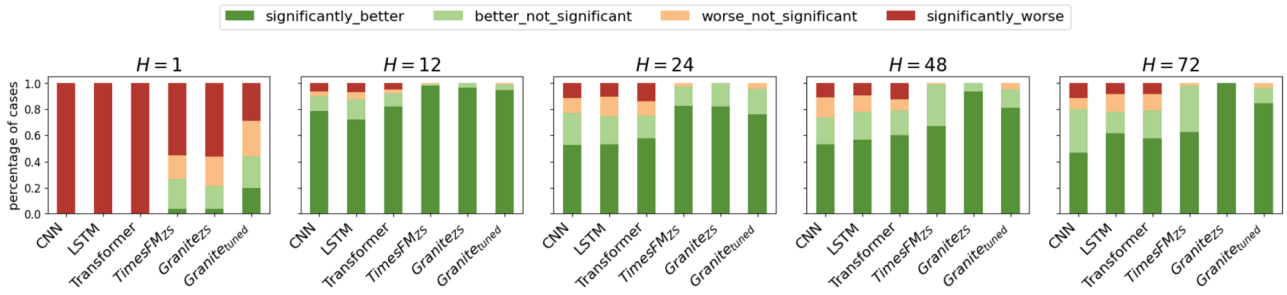


Fig. 10. Scenario X: relative performance of prediction models across horizons. Stacked bars indicate the fraction of evaluation cases where each model outperforms or underperforms the dummy baseline.

Table 6
Scenario X: comparison with ARIMAX across OSC..

	MAE	RMSE	SMAPE	R^2
ARIMAX	23.56	28.11	0.83	-1.81
Best DL model				
Transformer ($I=24$, $H=12$)	7.23	9.44	0.43	0.34
Best TSFM				
Granite$_{tuned}$ ($I=52$, $H=1$)	3.32	5.05	0.25	0.73
TimesFM$_{ZS}$ ($I=12$, $H=12$)	4.96	7.38	0.35	0.56

lower bound of the confidence interval is greater than zero, and significantly worse when the upper bound is less than zero. In all other cases, the performance difference is not considered statistically significant.

Fig. 10 further reports, for each model and forecasting horizon, the proportion of cases across all sensors in which the model performs significantly better than the baseline, better but not significantly, worse but not significantly, or significantly worse. The figure confirms the trends observed in Fig. 9. For $H=1$, DL models consistently perform significantly worse than the baseline. In contrast, TSFMs perform significantly worse in a large number of cases (approximately 25%-55%), although in a smaller fraction of cases (around 5%-20%) they significantly outperform the dummy baseline. As the forecasting horizon increases, the proportion of cases in which the evaluated models significantly outperform the baseline grows. In particular, for TSFMs, there are no longer cases in which the dummy baseline outperforms the model, whereas for DL models there remain instances where the baseline is still significantly better. Among the evaluated approaches, Granite$_{ZS}$ appears to achieve the strongest performance, as it exhibits the highest proportion of cases with statistically significant improvements over the baseline. However, this result should be interpreted with caution, as Granite$_{ZS}$ operates with larger input windows compared to the other models due to architectural constraints.

Table 6 shows the comparison with the statistical model ARIMAX across the OSC subset. ARIMAX was tested using automatic order selection based on AIC, with $p, d, q \in [0, 3]$ optimized independently for each dataset. In the table, when the best DL model or TSFM corresponded to $H=1$, we also included the second-best model with a longer prediction horizon. Overall, both DL models and TSFMs substantially outperform ARIMAX in all evaluation metrics. ARIMAX exhibits the highest errors and a negative R^2 value, indicating poor predictive performance in this setting.

4.4.3. Temporal decay in prediction accuracy

Fig. 11 aims to evaluate the progressive decline in prediction accuracy as the temporal distance between the predicted timestamp and the last timestamp of the input window increases, focusing in particular on a 24-h prediction horizon. We evaluate RMSE and R^2 using two complementary approaches: (1) point-wise comparison, where each predicted value is compared with its corresponding observed value at each prediction step, e.g., first predicted hour, second predicted hour and so on (left panels), and (2) interval-averaged evaluation, where metrics are aggregated over broader ranges (1–8 h, 9–16 h, 17–24 h, and the full window), shown on the right.

In the left panels, **prediction performance clearly deteriorates as the prediction step moves further into the future and stabilizes after 10–15 prediction steps**. DL models exhibit pronounced variability, with a wide spread in RMSE across time steps. In contrast, TSFMs show more compact and stable error profiles. This difference is also reflected in the R^2 metric: DL models frequently yield negative values indicating poor model fit, whereas TSFMs generally maintain positive R^2 scores (an exception is Granite$_{tuned}$, whose R^2 occasionally falls below zero). This behavior is aligned with expectations in autoregressive forecasting, where **short-term predictions benefit from stronger temporal dependencies on recent observations**.

The right panels of Fig. 11 confirm the same trend. DL models consistently underperform TSFMs across all intervals, both in RMSE and R^2. For TSFMs, performance degradation becomes particularly evident beyond the first 8 h, demonstrating the expected temporal decay. These results are consistent with the autocorrelation analysis presented in Section 4.2, which highlights the strong short-term temporal dependence in the data. As the prediction step increases, the predictive information contained in the input window gradually diminishes, leading to a decrease in forecast reliability.

4.4.4. Autoregressive short-horizon prediction biases

Fig. 12 provides a more detailed analysis of model behavior for the $H=1$ and $H=12$, allowing us to assess whether the strong performance of the TSFMs results from a simple replication of the last observation in the input window rather than from the models' ability to learn and exploit temporal dynamics, a known behavior in the literature (Peixeiro, 2022). To analyze this phenomenon, MAE, SMAPE and R^2 were computed between predicted and real values, then the predicted time series was artificially shifted backward by 1 (s1), 2 (s2), and 3 (s3) h.

We first consider the 1-h prediction horizon. LSTM shows relatively stable performance across all metrics and shifts, indicating that its predictions are not strongly tied to a simple replication of the most recent input observation. In contrast, **the TSFMs exhibit lower errors for the s1 shift, suggesting that their predictions tend to resemble the most recent observed value in the input window**. This behavior is also consistent with the strong autoregressive nature of $PM_{2.5}$ data, where consecutive observations are highly correlated (see Section 4.2). Moreover, the metrics obtained without shifting and with the s2 shift are relatively similar, reflecting the limited variability typically observed between adjacent hourly $PM_{2.5}$ values. However, when considering the s3 shift (where predictions are compared with values three hours earlier) the errors increase more noticeably. This pattern indicates that the predictive signal remains closely tied to recent observations, and the mismatch becomes more evident as the temporal shift grows.

This behavior is not observed for the 12-h forecast horizon. In this case, the evaluation metrics remain relatively stable across the original alignment and the shifted versions (s1, s2, and s3). This suggests that, **for**

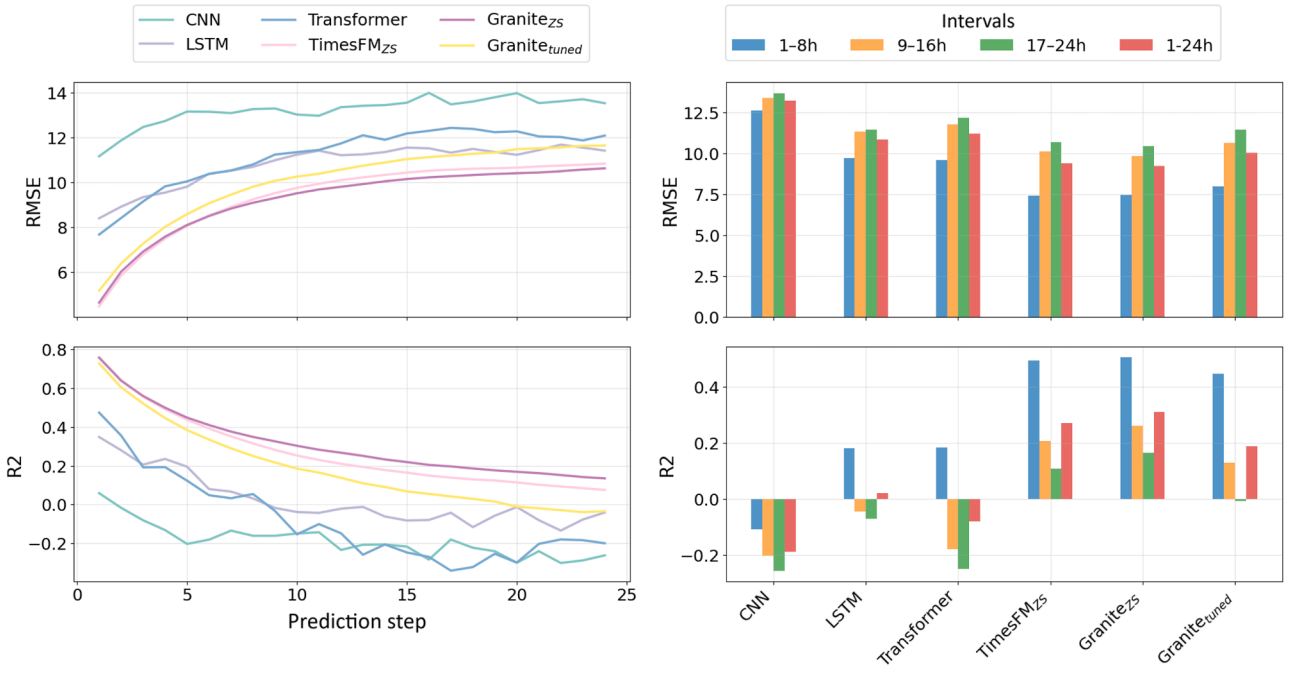
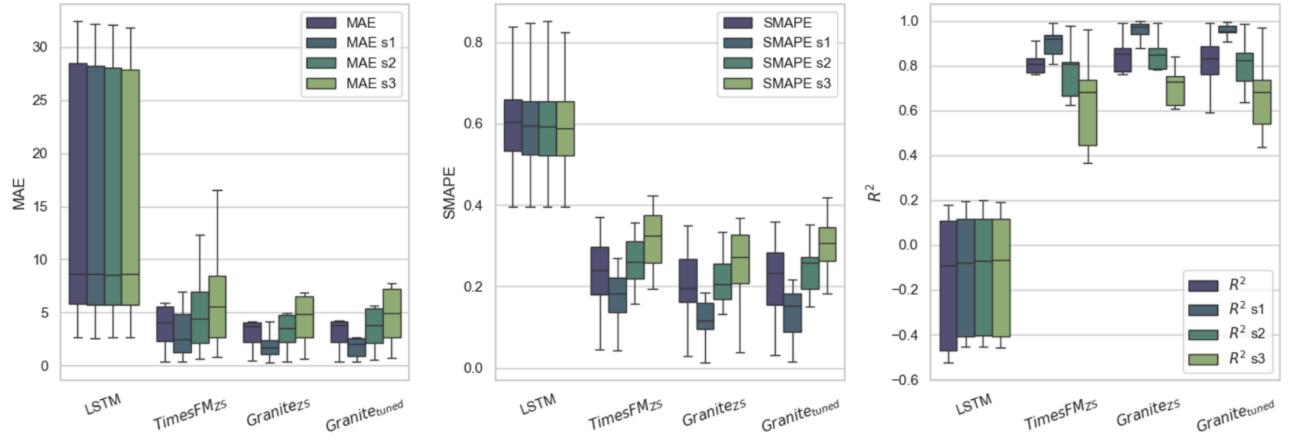
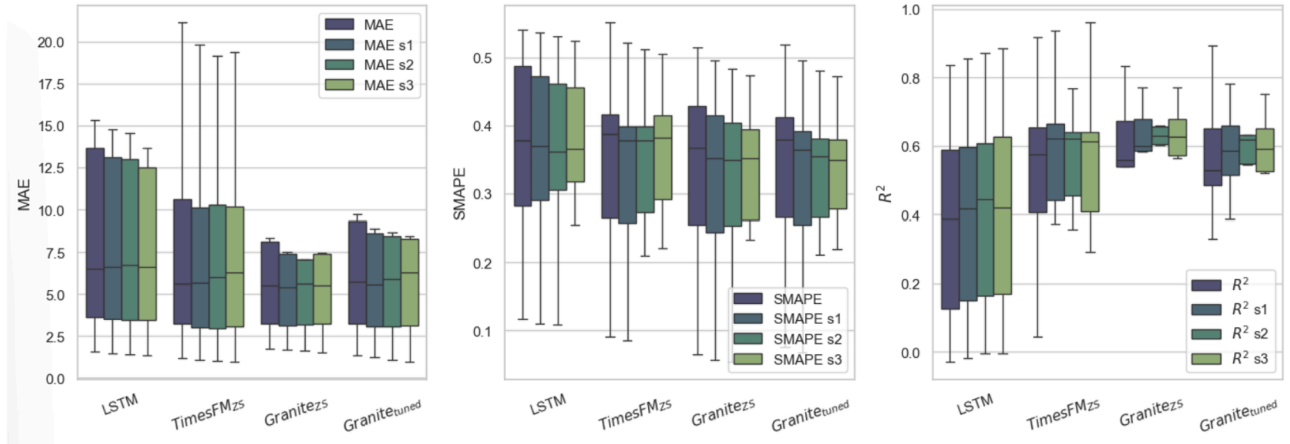


Fig. 11. Temporal decay in prediction accuracy across OSC ($I = 512$, $H = 24$).



(a) $H=1$



(b) $H=12$

Fig. 12. Scenario X: autoregressive short-horizon prediction bias across OSC ($I = 512$).

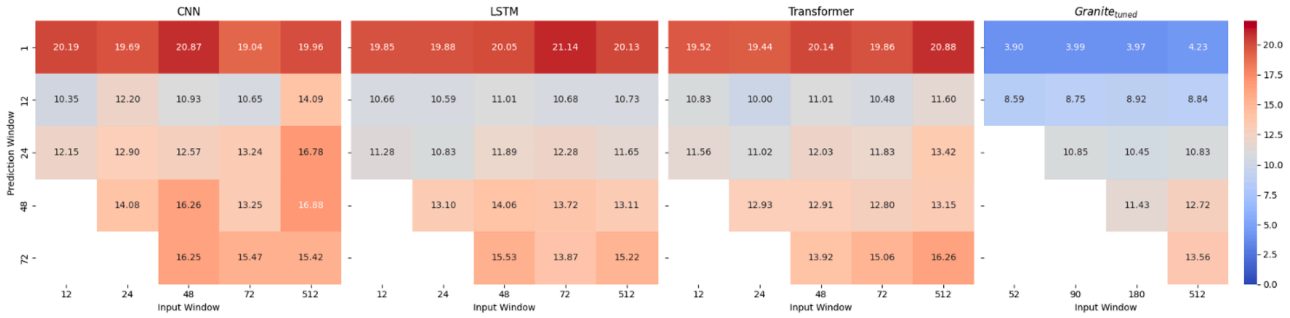


Fig. 13. Scenario Y: average MAE across OSC based on prediction model, input and forecast window size.

longer forecast horizons, the predictions are less dominated by the most recent input observation and instead rely on broader temporal patterns captured within the input window. From an application perspective, this aspect is particularly relevant, as a short-term forecast offer limited practical value in real-world applications.

4.5. Prediction scenario Y

In Scenario Y, we compare the performance of DL models and Granite$_{tuned}$ in predicting RS values, i.e., forecasting y, providing as input historical LCS and meteorological data, using five prediction windows (from a short-term window $H=1$ to a long-term window $H=72$), five input windows for DL models and four input windows for Granite$_{tuned}$.

Fig. 13 shows the average MAE values across different input and prediction windows considering OSC. As can be observed, **the performance trend identified in Scenario X is also confirmed in Scenario Y**: for the DL models, the best prediction window is 12 h, whereas for Granite$_{tuned}$ the best performance is achieved with a prediction window of 1 h. Furthermore, no substantial variation in performance is observed when keeping the prediction window fixed while varying the input window. Finally, **Granite$_{tuned}$ consistently outperforms the DL models across all considered configurations**.

4.5.1. Comparison with scenario X

By comparing the results in Fig. 13 with those in Fig. 7, which use the same color scale, it can be observed that the performance of all models deteriorates in Scenario Y. The only exceptions occur when using Granite$_{tuned}$ with $H=1$ or $H=12$, where performance remains relatively similar across the two scenarios. This suggests that **models face greater difficulty when predicting a variable that is not included in the input window**, as expected. A possible explanation is the limited amount of data available for training or fine-tuning the models.

4.5.2. Missing data impact

Fig. 14 is consistent with the observations from Scenario X, indicating that **under the adopted linear interpolation scheme, no clear deterioration in performance is observed in datasets with a high proportion of missing data**.

Moreover, Granite$_{tuned}$ generally achieves lower and more stable error values compared to Transformer, which exhibits greater variability across different levels of missing data.

4.5.3. Comparison with baselines

As a baseline, we employ a *multivariate linear regression model* that exploits lagged LCS observations and exogenous meteorological variables. For each sensor, the time series is first ordered chronologically and transformed into a supervised learning problem by constructing lagged input vectors. The resulting dataset is split chronologically into training (75%) and test (25%) subsets. We tested all the (I, H) pairs tested in Scenario Y by the other models. For each (I, H) pair, a distinct linear regression model is trained independently for each dataset.

Table 7
Scenario Y: comparison with ARIMAX across OSC.

	MAE	RMSE	SMAPE	R2
ARIMAX	13.37	18.10	0.55	−0.35
Best DL model				
Transformer ($I=24$, $H=12$)	10.00	13.52	0.35	0.44
Best TSFM				
Granite$_{tuned}$ ($I=52$, $H=1$)	3.90	6.10	0.16	0.83
Granite$_{tuned}$ ($I=52$, $H=12$)	8.59	11.85	0.29	0.57

Fig. 15 reports the statistical significance analysis of the performance differences between each prediction model and the baseline using $H=1$ and $H=12$ across different input windows performed using the block bootstrap procedure explained in Section 4.4.2. First we consider DL models. A clear pattern emerges in the short-term forecasting setting ($H=1$) where DL models are predominantly outperformed by the baseline, as indicated by the high proportion of significantly worse cases. Only marginal improvements are observed for larger input windows. For the longer forecasting horizon ($H=12$), the behavior changes substantially. All DL models exhibit a marked increase in the proportion of significantly better outcomes. This trend indicates that **DL models are better able to capture medium-term temporal dependencies**, leading to more consistent improvements over the baseline. For small values of I (e.g., $I=12$), performance is more variable and sometimes inferior to the baseline. **As I increases, the proportion of significantly better results steadily grows**, highlighting the importance of longer historical context for accurate forecasting.

Finally, **Granite$_{tuned}$ consistently achieves the highest proportion of significantly better outcomes** (around 100%), for all input windows and horizons. This suggests a superior ability to exploit temporal dependencies compared to both classical DL architectures and the baseline.

Overall, these results indicate that **forecasting performance is influenced by both the input window length and the prediction horizon**, with longer horizons and larger input windows favoring more complex models over the baseline.

Table 7 reports the comparison with ARIMAX for Scenario Y, confirming the same overall trends observed in Scenario X. ARIMAX performs poorly, yielding higher forecasting errors and a negative R^2.

4.5.4. Autoregressive short-horizon prediction biases

Fig. 16 presents a comparative analysis of model performance for $I=512$ across two prediction horizons, using MAE, SMAPE, and R^2 as evaluation metrics. To investigate the extent to which model predictions resemble recent observations within the input window, we additionally compute the same metrics after shifting the target series backward by 1 (s1), 2 (s2), and 3 (s3) time steps.

For the 1-h forecasting horizon ($H=1$), a clear distinction emerges across models. The DL models exhibit relatively stable performance across all shifts, indicating that **DL models predictions are not strongly driven by a simple replication of the most recent input**

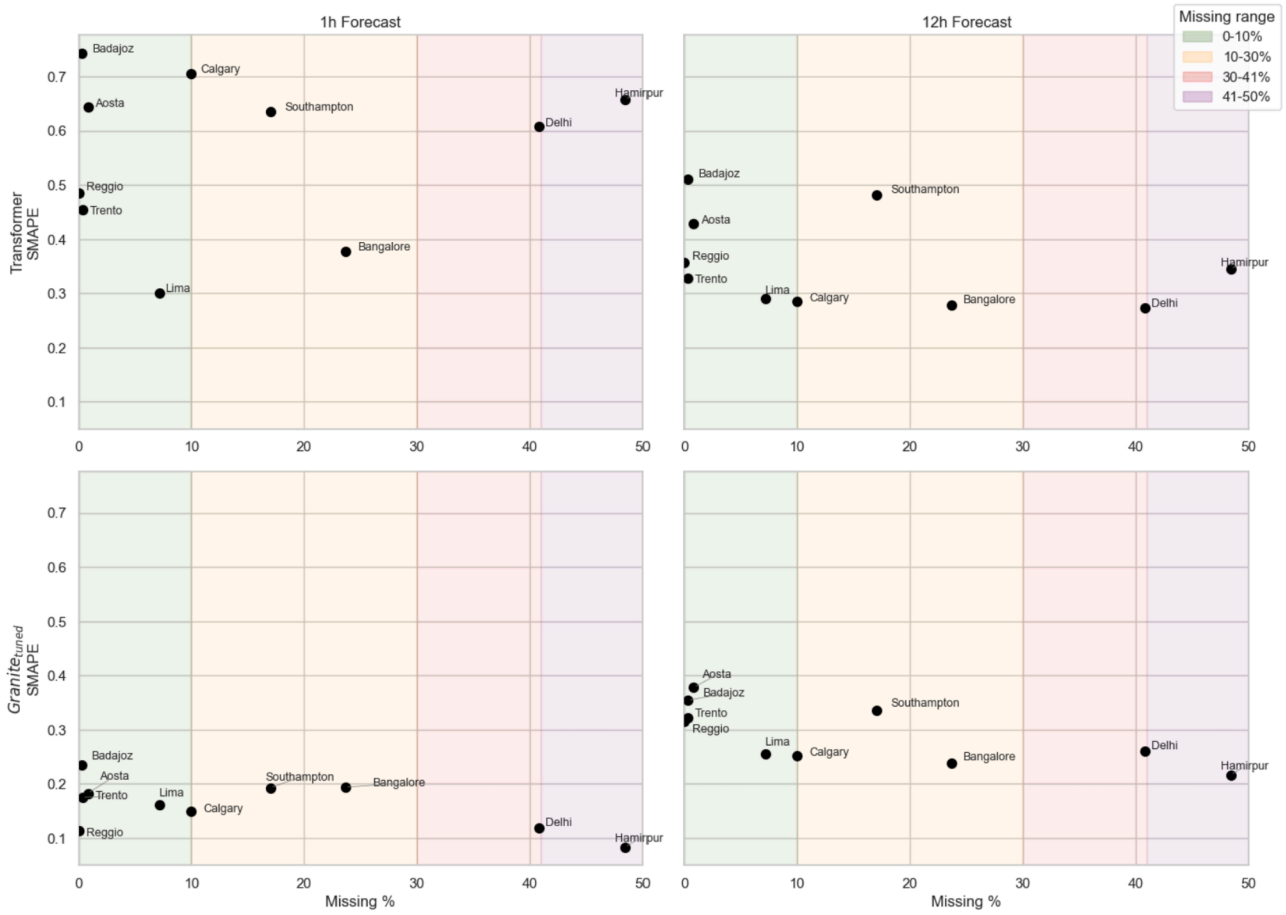


Fig. 14. Scenario Y: comparison of Transformer and Granite$_{\text{tuned}}$ performance across OSC ($I = 512$). Results are shown based on the percentage of missing data of the dataset. Each model is evaluated using its minimal feasible input window.

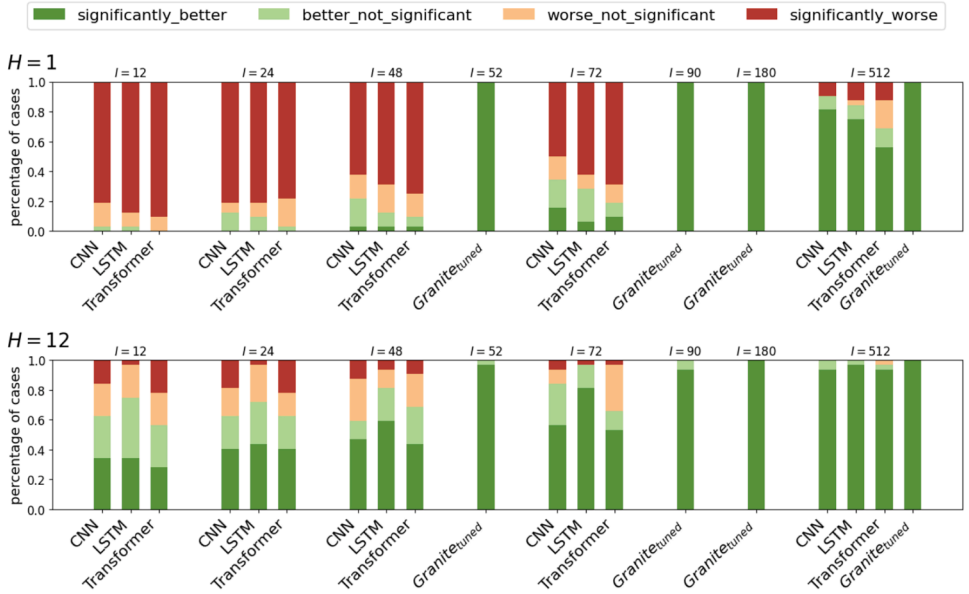
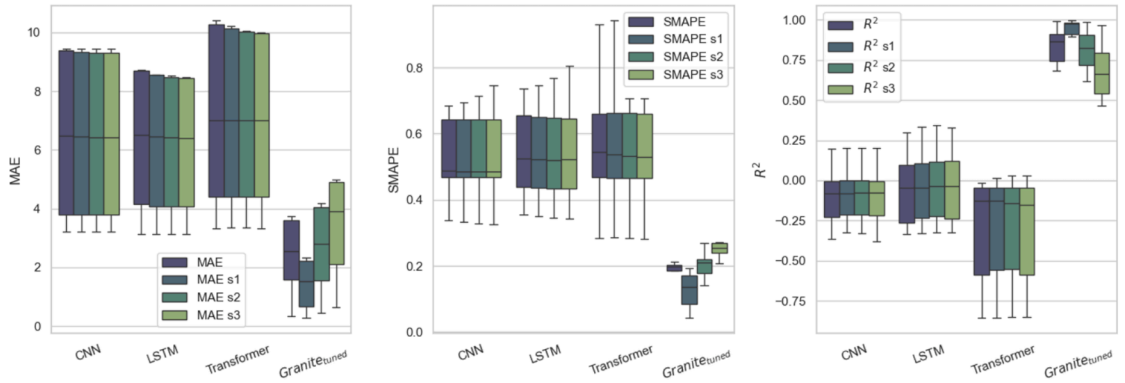
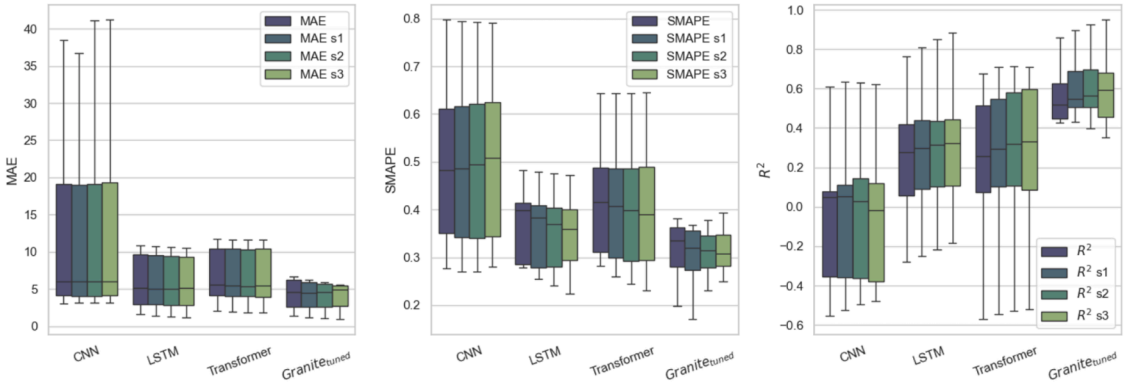


Fig. 15. Scenario Y: relative performance of prediction models across input window sizes considering two forecast horizons. Stacked bars indicate the fraction of evaluation cases where each model outperforms or underperforms the dummy baseline.



(a) 1-hour forecast horizon



(b) 12-hour forecast horizon

Fig. 16. Scenario Y: autoregressive short-horizon prediction bias across OSC ($I = 512$).

values. In contrast, the Granite$_{tuned}$ model shows a marked improvement in the s1 metrics compared to the unshifted case. This behavior suggests that **Granite$_{tuned}$ predictions tend to closely approximate the last observed value in the input window**, revealing a pronounced autoregressive bias. As the temporal shift increases (s2 and s3), the performance of Granite$_{tuned}$ progressively deteriorates, further confirming that the predictive signal is strongly concentrated on the most recent observations. Again, this pattern is consistent with the well-known high autocorrelation of $PM_{2.5}$ time series, where short-term dynamics are dominated by persistence-like behavior.

With $H = 12$, this effect is no longer observed, as all models (including Granite$_{tuned}$) display relatively stable performance across the original and shifted evaluations. This indicates that, **at longer horizons, predictions are less influenced by the most recent input value** and instead rely on broader temporal dependencies captured within the input window.

Overall, these findings highlight that the autoregressive bias is model-dependent and primarily affects the TSFM-based approach in short-term forecasting, while DL models appear less prone to this behavior, as highlighted also for Scenario X in Section 4.4.4.

4.6. Combined prediction strategy

Fig. 17 presents, from left to right, the results of Scenario X (prediction of LCS values), Combined Prediction Strategy (calibration of Scenario X's predictions), and Scenario Y (prediction of RS values). All values are compared to the RS data for evaluation (this is the reason why the first column of this figure reports values different from those presented in Section 4.4, where predictions are compared against LCS measurements). For each model, the figure reports the performance obtained using the best-performing prediction window with the minimal input window, excluding $H = 1$, as such a short forecasting horizon may be of limited practical relevance in real-world applications.

Focusing on the second, third, and fourth columns of Fig. 17, it is possible to compare the impact of the DL models used for calibration when applied to the predictions obtained in Scenario X within the Combined Prediction Strategy. For a given calibration model, **the choice of TSFM used to generate the initial predictions does not appear to significantly affect performance**. In contrast, differences emerge when comparing the calibration architectures. CNN and LSTM generally achieve lower error values (MAE, RMSE, and SMAPE) than the Transformer-based model. However, the Transformer consistently yields positive R^2 values, whereas the CNN and LSTM models occasionally produce negative values. Nevertheless, the median R^2 remains below 0.5 across configurations, suggesting a still limited ability to explain the variance in RS measurements.

The comparison between Granite$_{ZS}$ and Granite$_{tuned}$ shows that **fine-tuning Granite does not lead to performance improvements**, a result that is consistent with what was already observed in Scenarios X and Y.

Furthermore, the results of the Combined Prediction Strategy reported in Fig. 17 indicate that **it is not possible to identify a single combination of TSFM and DL calibration model that consistently outperforms all the others** across the evaluated cities.

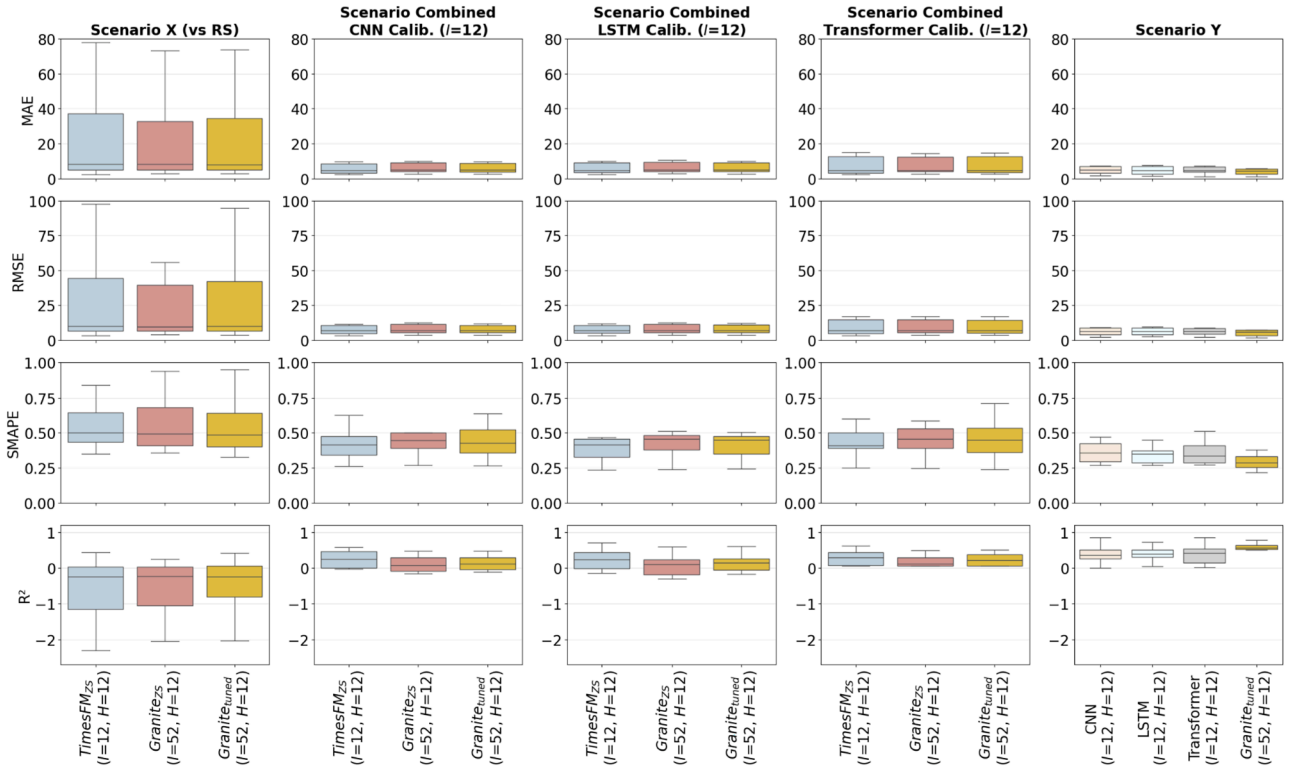


Fig. 17. Comparison of different scenarios across OSC.

4.6.1. Comparison with scenario X

The first column of Fig. 17 (Scenario X) shows the evaluation of the TSFM predictions against the RS measurements without any calibration and can be used as a reference for comparison with the results obtained through the Combined Prediction Strategy. This comparison enables the assessment of both how closely the TSFM predictions align with RS measurements and the ability of the TSFMs to generate RS-equivalent values when combined with the calibration models. Although the predictions refer to different targets (LCS values in Scenario X and RS values in the Combined Prediction Strategy), both represent $PM_{2.5}$ concentrations. Since the ultimate objective is to obtain the most accurate $PM_{2.5}$ measurements from LCSs, this comparison remains meaningful.

The results show that the uncalibrated predictions from Scenario X are characterized by substantially higher MAE, RMSE, and SMAPE values, together with negative R^2 scores. This indicates that **TSFMs alone are not sufficient to accurately estimate $PM_{2.5}$ concentrations from raw LCS data without a dedicated calibration step**, which is consistent with expectations. In the Combined Prediction Strategy, a clear improvement is observed across all evaluation metrics, confirming its effectiveness.

4.6.2. Comparison with scenario Y

Scenario Y (rightmost column of Fig. 17) **shows substantially improved performance compared to both Scenario X and the Combined Prediction Strategy**. All models achieve lower MAE, RMSE, and SMAPE values, together with consistently positive and higher R^2 scores. The distributions are also more compact, suggesting greater stability across cities. Among the evaluated models, as already discussed in Section 4.5, **$Granite_{tuned}$ achieves the best overall performance**, with the lowest errors and the highest R^2 values.

4.7. City-level model comparison

Table 8 reports the best-performing configurations for each city under both the Combined Prediction Strategy and Scenario Y. For cities with multiple sensors, the reported values correspond to the average performance across sensors. The selection of the best configuration was based on the lowest mean MAE. Bold values indicate the best performance for each city, while underlined values denote the second-best results. Predictions with a one-hour horizon ($H = 1$) were excluded from the analysis, as such short-term forecasts offer limited practical value in real-world applications, as discussed in previous sections.

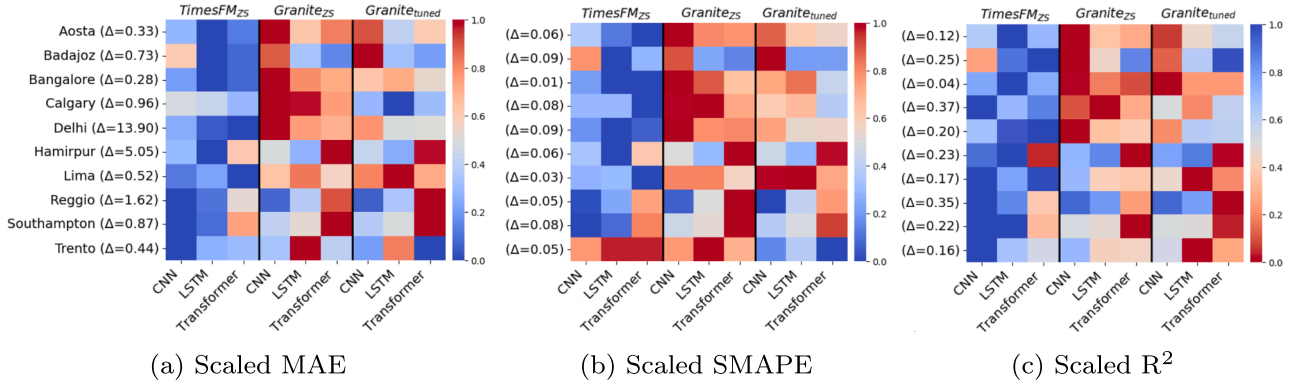
The Raw Data Baseline represents the agreement between RS and pre-processed LCS measurements, prior to the application of any predictive modeling or calibration. As such, it provides a reference point for evaluating the improvements achieved by the Combined Prediction Strategy and Scenario Y. The baseline yields negative R^2 values in most cases, highlighting the need for dedicated strategies to improve the accuracy of LCS measurements.

The Combined Prediction Strategy generally achieves better performance than the baseline across most cities, with the notable exception of Hamirpur, where the R^2 value is negative. Scenario Y further improves upon both the baseline and the Combined Prediction Strategy in nearly all cases. The only exceptions are Hamirpur (just for the MAE value) and Lima, where the Combined Prediction Strategy achieves slightly better results. Considering the data distributions shown in Fig. 4, it is worth noting that Lima is characterized by consistently high humidity levels, which may partially influence the observed low performance due to known humidity-induced biases in PM sensors (Casari & Po, 2024).

In the Combined Prediction Strategy, the optimal model combination varies across cities, with $TimesFM_{ZS}$ + LSTM emerging most frequently. In contrast, under Scenario Y, $Granite_{tuned}$ consistently outperforms the DL models across all cities, confirming the findings reported in Section 4.5. The city-level differences in the best-performing configurations are consistent with the heterogeneity observed in both $PM_{2.5}$ dynamics and meteorological conditions. Cities characterized by long-tailed $PM_{2.5}$ distributions and higher concentration variability, such as Delhi and Hamirpur, tend to exhibit larger prediction errors, confirming the increased difficulty associated with highly heterogeneous pollution dynamics. Interestingly, despite the high variability in $PM_{2.5}$ observed

Table 8
Best configuration comparison in combined prediction strategy and scenario Y.

	Raw Data Baseline			Combined Prediction Strategy				Scenario Y			
City	MAE	RMSE	R^2	Model	MAE	RMSE	R^2	Model	MAE	RMSE	R^2
Aosta (IT)	2.58	4.87	0.48	$TimesFM_{ZS}$ + LSTM	1.92	2.81	0.36	$Granite_{tuned}$	**1.55**	**2.29**	**0.57**
Badajoz (ES)	8.26	11.39	0.11	$TimesFM_{ZS}$ + LSTM	2.82	4.26	0.20	$Granite_{tuned}$	**2.23**	**3.49**	**0.46**
Bangalore (IN)	12.63	16.98	−0.43	$TimesFM_{ZS}$ + LSTM	4.71	6.31	0.55	$Granite_{tuned}$	**4.53**	**6.23**	**0.56**
Calgary (CA)	3.88	5.79	−0.08	$Granite_{tuned}$ + LSTM	2.22	2.86	0.26	$Granite_{tuned}$	**0.96**	**1.74**	**0.92**
Delhi (IN)	69.04	93.23	−0.39	$TimesFM_{ZS}$ + Transf.	35.64	54.01	0.71	$Granite_{tuned}$	**33.51**	**48.16**	**0.77**
Hamirpur (IN)	**23.01**	30.73	0.36	$TimesFM_{ZS}$ + LSTM	49.68	68.52	−0.84	$Granite_{tuned}$	23.39	**30.53**	**0.63**
Lima (PE)	7.15	9.91	−0.34	$TimesFM_{ZS}$ + Transf.	**4.05**	**6.17**	**0.22**	$Granite_{tuned}$	4.19	6.31	0.20
Reggio (IT)	32.59	47.48	−10.20	$TimesFM_{ZS}$ + CNN	8.69	10.78	0.16	$Granite_{tuned}$	**6.09**	**7.85**	**0.55**
Southampton (UK)	4.02	9.44	0.09	$TimesFM_{ZS}$ + CNN	3.22	4.59	0.41	$Granite_{tuned}$	**2.82**	**3.98**	**0.55**
Trento (IT)	6.91	10.57	−0.08	$Granite_{tuned}$ + Transf.	4.90	6.45	0.34	$Granite_{tuned}$	**3.90**	**5.20**	**0.62**



(a) Scaled MAE (b) Scaled SMAPE (c) Scaled R^2

Fig. 18. Performance comparison across all cities in the combined prediction strategy.

in Delhi, the R^2 values for both the Combined Prediction Strategy and Scenario Y remain relatively high, suggesting that the models are able to capture a substantial portion of the underlying variance.

To further investigate the variability of the Combined Prediction Strategy results across model combinations, Fig. 18 presents heatmaps of scaled MAE, SMAPE and R^2 values. Where multiple sensors are available for a given city, we report the average performance across all sensors of that city, as done in Table 8, excluding $H = 1$. These metrics have been row-wise normalized to the $[0, 1]$ interval to highlight model combinations that consistently perform better or worse across all cities. For each city, the difference (Δ) between the best and worst non-scaled MAE, SMAPE and R^2 values is reported in parentheses on the y-axis. This enables the assessment of the practical significance of the observed variability: a small Δ indicates that all model combinations perform similarly, whereas a large Δ indicates substantial performance differences across configurations.

Overall, **model combinations involving $Granite_{ZS}$ tend to exhibit poorer performance** in most cities, with a few exceptions such as Badajoz and Reggio. In contrast, **combinations with $TimesFM_{ZS}$ generally yield the best performance** across the majority of locations, although with some exceptions such as Calgary in terms of MAE and Trento in terms of SMAPE. However, the practical relevance of these exceptions is limited. In Trento, all metrics are associated with very small Δ values, indicating negligible differences among configurations. Similarly, although the best MAE value in Calgary is achieved by $Granite_{tuned}$ + LSTM, the corresponding Δ_{MAE} remains small, while Δ_{R^2} is considerably larger and favors configurations based on $TimesFM_{ZS}$. This pattern suggests a varying degree of model generalizability and potential sensitivity to local data characteristics. Furthermore, even in cities where $TimesFM_{ZS}$ + LSTM does not represent the absolute best configuration, its performance remains extremely close to the top-performing alternative (e.g., Delhi, Reggio, and Southampton). These results support the use of $TimesFM_{ZS}$ + LSTM as a reasonable default choice for the Combined Prediction Strategy.

5. Open challenges and future work

In this section, we outline key challenges encountered in this study and propose directions for future research to advance the use of TSFMs and DL models in PM prediction/calibration.

Incomplete and Sparse Datasets: One limitation is the use of datasets characterized by substantial missing data. Although linear interpolation was employed to fill data gaps, and our experimental analysis does not reveal a clear deterioration in performance as the proportion of missing data increases, the inclusion of interpolated values during training may still introduce bias in other contexts and affect the robustness of the learned models. Future research should investigate more advanced imputation techniques, such as the approach proposed in Hollmann et al. (2025), to handle missing data more effectively. Additionally, alternative training strategies could be explored, such as masking or excluding interpolated observations from the loss computation, thereby preventing them from influencing model optimization.

Leveraging TSFMs for Calibration Tasks: TSFMs have thus far been used predominantly for forecasting tasks. However, their application to sensor calibration, directly correcting LCS measurements, has not been fully explored. Future work could examine the ability of TSFMs to serve as calibration models, either through fine-tuning or conditioning, potentially offering superior accuracy and generalizability compared to standard DL-based approaches.

Linking Model Performance to Data Characteristics: Although this study identifies the best-performing configuration for each city, it does not provide a clear understanding of why certain model configurations outperform others under specific conditions. The observed variability in optimal configurations across cities, especially in Combined Prediction Strategy, suggests that model performance may be influenced by underlying data characteristics, such as the distribution of $PM_{2.5}$ concentrations and the variability of meteorological variables (e.g., temperature, humidity, pressure, and wind speed). Future work should aim to systematically investigate the relationship between dataset properties and

model performance, with the goal of identifying patterns that enable the selection of appropriate configurations based on data characteristics.

Cross-Site and Cross-Sensor Generalization: While the zero-shot TSFMs evaluated in Scenario X are applied without any task-specific training, and their competitive performance across 10 cities and 6 countries provides encouraging evidence of transferability, the present study does not implement a formal cross-site evaluation protocol. Specifically, DL models and fine-tuned TSFMs are trained and tested on chronological splits of the same sensor's time series. This study does not include a dedicated leave-one-sensor-out protocol to quantify cross-site or cross-sensor generalization in a controlled manner. Such experiments would more precisely quantify the extent to which TSFMs and trained models generalize across different sensor types, climatic conditions, and emission contexts, and we identify this as a key direction for future research.

Mechanistic Understanding of TSFM Superiority: While our study provides empirical evidence that TSFMs consistently outperform traditional DL models for $PM_{2.5}$ forecasting, the underlying reasons remain incompletely understood. Future work should systematically investigate potential factors (e.g., pre-training scale, architecture, data diversity), moving beyond empirical benchmarking toward a causal understanding of TSFM capabilities.

Few-shot learning with TSFMs: Recent work has extended TSFMs to few-shot scenarios via in-context fine-tuning (Faw et al., 2025; Xu et al., 2025). However, our results show that full fine-tuning of Granite (using 75% of the data) did not outperform zero-shot, suggesting that few-shot learning would likely provide marginal gains for $PM_{2.5}$ forecasting. Systematic evaluation of few-shot TSFMs in this domain remains an open direction.

6. Conclusion

This study presents an empirical evaluation of TSFMs for $PM_{2.5}$ prediction using LCS data, benchmarking Google's TimesFM and IBM's Granite against CNN, LSTM, and Transformer models across 34 datasets from 10 cities under two forecasting strategies. We systematically benchmark TSFMs against traditional deep learning (DL) models (CNN, LSTM, Transformer) across two forecasting strategies: direct prediction of reference station (RS)-equivalent data and a two-stage combined strategy that predicts LCS measurements and calibrates them to approximate RS data.

Our findings demonstrate that TSFMs consistently outperform traditional DL models across forecast horizons. Zero-shot TSFMs perform competitively across all evaluated sites providing empirical evidence of transferability to unseen sensor deployments. Among the deployment strategies evaluated, Direct RS prediction using fine-tuned Granite achieves the best overall performance.

Further analysis shows that, under linear interpolation, the proportion of missing data does not produce a clear deterioration in performance. TSFMs exhibit an autoregressive bias for 1-h forecasts that dissipates at longer, more practically relevant horizons (e.g., 12 h). Input window length has limited impact on performance, allowing the use of minimal compatible windows. Fine-tuning Granite does not provide a consistent advantage over its zero-shot configuration.

These results position TSFMs as a competitive alternative to conventional DL approaches for air quality forecasting with LCs networks. A formal cross-site evaluation protocol remains an important direction for future work.

CRediT authorship contribution statement

Federica Rollo: Conceptualization, Methodology, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization, Investigation; **Matteo Angelinelli:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft; **Martina Casari:** Conceptualization, Validation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization; **Laura Po:** Conceptualization, Supervision, Project administration, Funding acquisition, Writing – original draft, Writing – review & editing, Investigation; **Giorgio Pedrazzi:** Conceptualization, Supervision; **Roberta Turra:** Project administration.

Data availability

All original $PM_{2.5}$ and meteorological data used in this study are publicly available from the sources cited in Table 1. Detailed references for each dataset are provided in the corresponding row of Table 1.

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.

Acknowledgment

This work was supported by the project "P.A.T.H. - Predictive Active Transport Hub: Predictive routing for active mobility based on air quality and simulated vehicular traffic" within FARD2025, funded by the Engineering Department "Enzo Ferrari" of the University of Modena and Reggio Emilia.

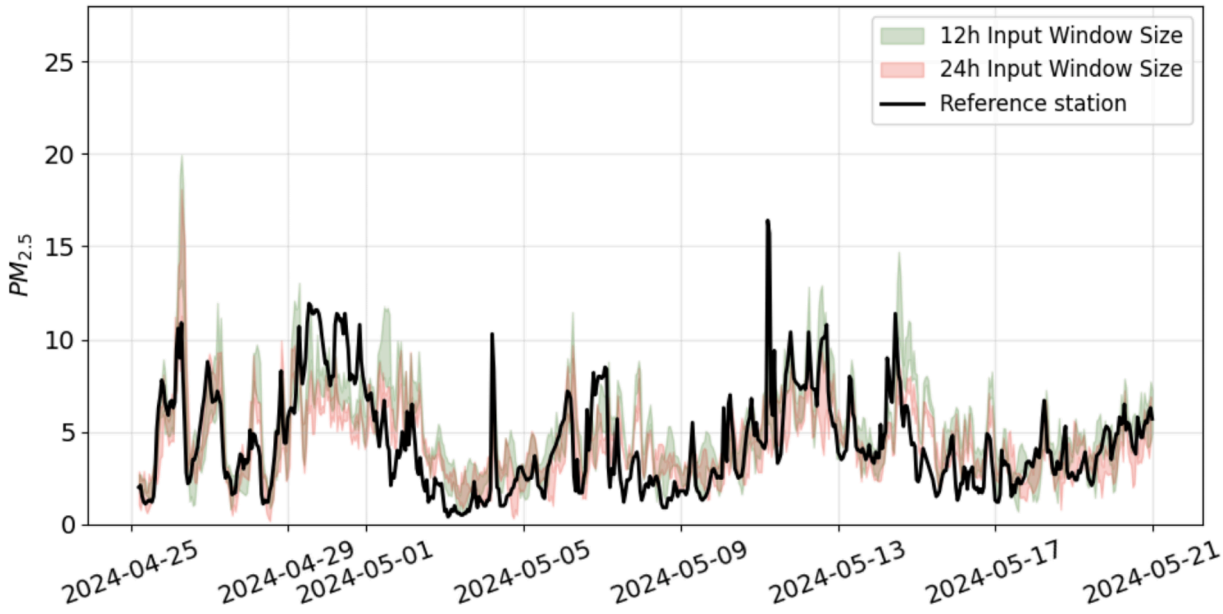
We gratefully acknowledge the Regional Agency for Environmental Protection of the Aosta Valley, Arpae Emilia-Romagna, and the Provincial Agency for Environmental Protection of Trento for providing data from their reference monitoring stations. These stations were co-located with low-cost sensors donated by Wiseair Srl, whose contribution of sensors is also sincerely appreciated.

Appendix A. Sensor-Level Illustrative Examples

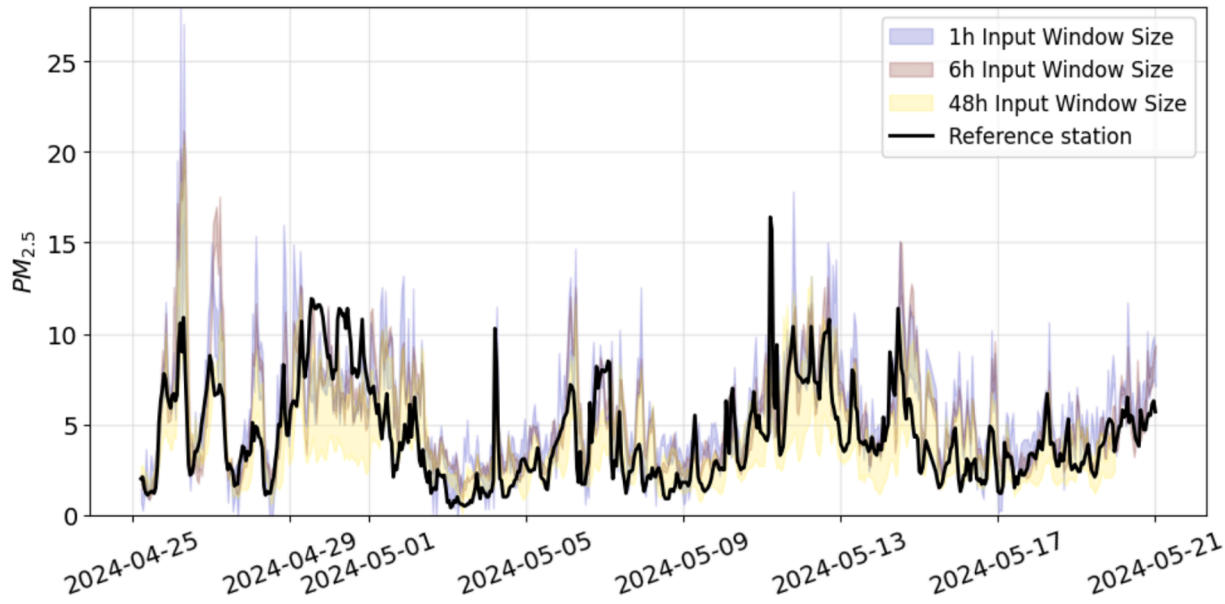
This appendix provides detailed visual examples for a representative sensor (*Aosta 1*) to complement the aggregated results presented in the main manuscript. The figures below illustrate the behavior of different models and strategies across calibration and forecasting tasks.

Calibration Performance Across Input Window Sizes Figs. A.19a and A.19b demonstrate how calibration performance varies with input window length for sensor Aosta 1. The shaded areas represent the range between maximum and minimum values across the three DL models (CNN, LSTM, Transformer). As shown in Fig. A.19, the 1-h and 6-h windows exhibit a systematic tendency to overestimate RS values, particularly during peak events. In contrast, the 12-h and 24-h windows show better agreement with reference measurements, although the 24-h window displays a slight underestimation bias.

Forecasting Performance: Scenario X Fig. A.20 illustrates the predicted time series for sensor *Aosta 1*, comparing the two best prediction windows (i.e., 1 h for TSFMs and 12 h for DL models). DL models with $H = 1$ produce a flattened signal that fails to capture RS variability, whereas with $H = 12$ they approximate the average trend. On the other hand, TSFMs demonstrate robustness across both horizons, though performance declines for the longer 12-h forecast, consistent with the increased difficulty of extended-horizon prediction. Indeed, predicting 12 h ahead is inherently more challenging than forecasting the immediate next hour. The performance gap between DL models and TSFMs narrows in 12-h forecasting, as demonstrated in the aggregated results in the main manuscript.



(a) Input windows: 12 hours (green) and 24 hours (red)



(b) Input windows: 1 hour (blue), 6 hours (brown), and 48 hours (yellow)

Fig. A.19. Calibration performance comparison for different input window lengths (sensor *Aosta 1*). The black line represents reference station (RS) measurements.

Comparison of Prediction Strategies Fig. A.21 compares Scenario X, the Combined Prediction Strategy, and Scenario Y. Scenario Y achieves the closest alignment with RS values, demonstrating superior predictive accuracy. The Combined Prediction Strategy, while improving upon uncalibrated TSFM predictions of Scenario X, exhibits reduced variability and fails to fully recover the true trend, highlighting the challenge of correcting errors inherited from biased initial predictions. Scenario Y avoids this issue and improves overall performance.

Supplementary material

The source code associated with this article can be found at the GitLab repository (https://gitlab.com/martina.casari.93/sensordatacorrection-hygroscopicity/-/tree/main/AIQS4MitH), the pre-processed datasets and the detailed results for all 34 individual sensor datasets are available at https://zenodo.org/records/20607348, and the optimal hyperparameter configurations for each model and dataset are published in the online version of this article, at 10.1016/j.eswa.2026.133099.

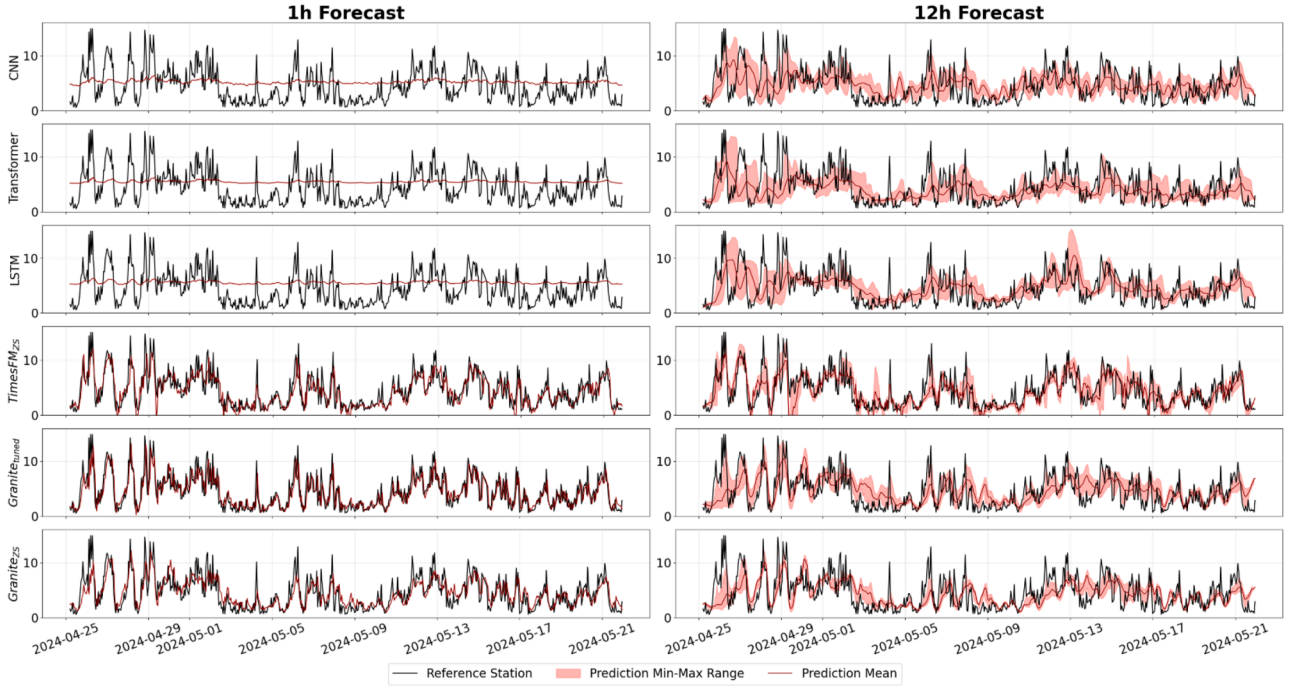


Fig. A.20. Scenario X: forecasting performance for *Aosta 1* sensor. Each model is evaluated using its minimal feasible input window.

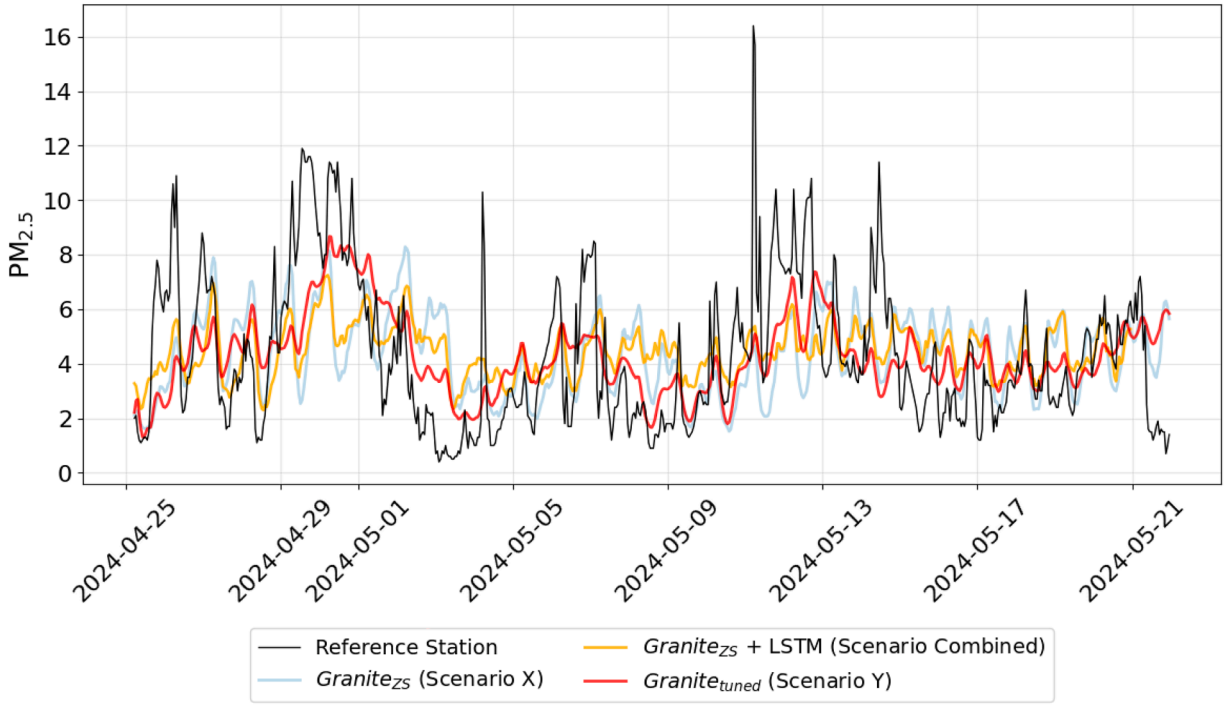


Fig. A.21. Comparison of Scenario X, combined prediction strategy and scenario Y for sensor *Aosta 1* ($I = 512$, $H = 24$).

References

Abalo-García, A., Hernández-García, S., Ramírez, I., & Schiavi, E. (2025). MPD: A meteorological and pollution dataset: A comprehensive study of machine and deep learning methods for air pollution forecasting. *IEEE Access*, *13*, 41282–41299. https://doi.org/10.1109/ACCESS.2025.3547038

Agency, U. S. E. P. (2017). CMAQ: The community multiscale air quality modeling system. www.epa.gov/cmaq.

Agency, U. S. E. P. (2024). Particulate matter (PM) basics. Accessed: 2026-03-26 https://www.epa.gov/pm-pollution/particulate-matter-pm-basics.

Aksu, T., Woo, G., Liu, J., Liu, X., Liu, C., Savarese, S., Xiong, C., & Sahoo, D. (2024). GIFT-eval: A benchmark for general time series forecasting model evaluation, . https://openreview.net/forum?id=Z2cMOOANFX.

Allen, A., Markou, S., Tebbutt, W., Requeima, J., Bruinsma, W. P., Andersson, T. R., Herzog, M., Lane, N. D., Chantry, M., Hosking, J. S., & Turner, R. E. (2025). End-to-end data-driven weather prediction. *Nature*, *641*(8065), 1172–1179. https://doi.org/10.1038/s41586-025-08897-0

AlSalehy, A. S., & Bailey, M. (2025). Improving time series data quality: Identifying outliers and handling missing values in a multilocation gas and weather dataset. *Smart Cities*, *8*(3). https://www.mdpi.com/2624-6511/8/3/82. https://doi.org/10.3390/smartcities8030082

Amegah, A. K. (2018). Proliferation of low-cost sensors. What prospects for air pollution epidemiologic research in sub-saharan africa? *Environmental Pollution*, *241*, 1132 – 1137. https://doi.org/10.1016/j.envpol.2018.06.044

Arroyo, P., Gómez-Suárez, J., Suárez, J. I., & Lozano, J. (2021). Low-cost air quality measurement system based on electrochemical and PM sensors with cloud connection. *Sensors*, *21*(18). https://www.mdpi.com/1424-8220/21/18/6228. https://doi.org/10.3390/s21186228

Arsov, M., Zdravevski, E., Lameski, P., Corizzo, R., Koteli, N., Gramatikov, S., Mitreski, K., & Trajkovik, V. (2021). Multi-horizon air pollution forecasting with deep neural networks. *Sensors*, *21*(4). https://www.mdpi.com/1424-8220/21/4/1235. https://doi.org/10.3390/s21041235

Bachechi, C., Rollo, F., & Po, L. (2024). HypeAIR: A novel framework for real-time low-cost sensor calibration for air quality monitoring in smart cities. *Ecological Informatics*, *81*, 102568. https://www.sciencedirect.com/science/article/pii/S1574954124001109. https://doi.org/https://doi.org/10.1016/j.ecoinf.2024.102568

Bai, Z., Wu, P., Geng, F., Zhang, H., Wang, P., Du, L., Li, Z., Chen, X., Fang, Z., & Wu, Y. (2024). Imputation of doppler cardiograms using a fine-tuned time-series foundation model. In *2024 IEEE International conference on signal, information and data processing (ICSIDP)* (pp. 1–5). https://doi.org/10.1109/ICSIDP62679.2024.10869270

Barkjohn, K. K., Gantt, B., & Clements, A. L. (2021). Development and application of a united states-wide correction for PM2.5 data collected with the purpleair sensor. *Atmospheric Measurement Techniques*, *14*(6), 4617 – 4637. https://doi.org/10.5194/amt-14-4617-2021

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., Gupta, J. K., Thambiratnam, K., Archibald, A. T., Wu, C.-C., Heider, E., Welling, M., Turner, R. E., & Perdikaris, P. (2025). A foundation model for the earth system. *Nature*, *641*(8065), 1180–1187. https://doi.org/10.1038/s41586-025-09005-y

Buchhorn, K., Santos-Fernandez, E., Mengersen, K., & Salomone, R. (2024). Graph neural network-based anomaly detection for river network systems. *F1000Research*, *12*, 991.

Bulot, F. M. J. (2022). Characterisation and calibration of low-cost PM sensors at high temporal resolution to reference grade performances - dataset. https://doi.org/10.5281/zenodo.7198378

Bulot, F. M. J., Ossont, S. J., Morris, A. K. R., Basford, P. J., Easton, N. H. C., Mitchell, H. L., Foster, G. L., Cox, S. J., & Loxham, M. (2023). Characterisation and calibration of low-cost PM sensors at high temporal resolution to reference-grade performance. *Heliyon*, *9*(5). https://doi.org/10.1016/j.heliyon.2023.e15943

Campmier, M., Gingrich, J., Singh, S. et al. (2023a). Seasonally optimized calibrations improve low-cost sensor performance: Long-term field evaluation of purpleair sensors in urban and rural india [dataset]. https://doi.org/10.6078/D1RQ70

Campmier, M. J., Gingrich, J., Singh, S., Baig, N., Gani, S., Upadhya, A., Agrawal, P., Kushwaha, M., Mishra, H. R., Pillarisetti, A., Vakacherla, S., Pathak, R. K., & Apte, J. S. (2023b). Seasonally optimized calibrations improve low-cost sensor performance: Long-term field evaluation of purpleair sensors in urban and rural india. *Atmospheric Measurement Techniques*, *16*(19), 4357–4374. https://amt.copernicus.org/articles/16/4357/2023/. https://doi.org/10.5194/amt-16-4357-2023

Campo, F., Franco, D., de Campos Santos, F., Blanco-Rodríguez, A., Garcia-Ramirez, A. R., Ratão, G., & Hoinaski, L. (2023). Clean - collaborative low-cost environmental and air-quality network. *Environmental Modelling & Software*, *163*, 105664. https://www.sciencedirect.com/science/article/pii/S1364815223000506. https://doi.org/https://doi.org/10.1016/j.envsoft.2023.105664

Casari, M., & Po, L. (2023). AirMLP - SPS30 low-cost sensors and tecora reference station PM 2.5 data. https://doi.org/10.5281/zenodo.10037781

Casari, M., & Po, L. (2024). MitH: A framework for mitigating hygroscopicity in low-cost PM sensors. *Environmental Modelling & Software*, *173*, 105955. https://www.sciencedirect.com/science/article/pii/S1364815224000161. https://doi.org/https://doi.org/10.1016/j.envsoft.2024.105955

Cun, Y. L., Boser, B., Denker, J. S., Howard, R. E., Habbard, W., Jackel, L. D., & Henderson, D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems 2* , pp. 396–404). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Dallah, D., & Sulieman, H. (2022). Outlier detection using the range distribution. In *International conference on recent developments in mathematics* (pp. 687–697). Springer.

Das, A., Kong, W., Rajat, S., & Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. In *Proceedings of machine learning research* (p. 10148 – 10167). (*vol. 235*).

deSouza, P., Kahn, R., Stockman, T., Obermann, W., Crawford, B., Wang, A., Crooks, J., Li, J., & Kinney, P. (2022). Calibrating networks of low-cost air quality sensors. *Atmospheric Measurement Techniques*, *15*(21), 6309–6328. https://doi.org/10.5194/amt-15-6309-2022.

Ekambaram, V., Jati, A., Dayama, P., Mukherjee, S., Nguyen, N. H., Gifford, W. M., Reddy, C., & Kalagnanam, J. (2024). Tiny time mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in neural information processing systems* (pp. 74147–74181). Curran Associates, Inc. (*vol. 37*).

Ekambaram, V., Jati, A., Dayama, P., Mukherjee, S., Nguyen, N. H., Gifford, W. M., Reddy, C., & Kalagnanam, J. (2025). Tiny time mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. In *Proceedings of the 38th international conference on neural information processing systems* NIPS '24. Red Hook, NY, USA: Curran Associates Inc.

Fan, J., Chu, H., Liu, L., & Ma, H. (2024). LLMAir: Adaptive reprogramming large language model for air quality prediction. In *2024 IEEE 30th International conference on parallel and distributed systems (ICPADS)* (pp. 423–430). https://doi.org/10.1109/ICPADS63350.2024.00062

Faw, M., Sen, R., Zhou, Y., & Das, A. (2025). In-context fine-tuning for time-series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025). Proceedings of Machine Learning Research*, *vol. 267*, 16355 – 16374. JMLR.org.

Feng, Z., Zheng, L., & Xue, N. (2025). Physics-informed calibration model for enhanced accuracy in particulate matter monitoring integrating clustering algorithms with field validation. *Expert Systems with Applications*, *277*, 127313. https://www.sciencedirect.com/science/article/pii/S0957417425009352. https://doi.org/https://doi.org/10.1016/j.eswa.2025.127313

Fu, X., Hirano, M., & Imajo, K. (2025). Financial fine-tuning a large time series model. https://doi.org/10.1109/CiFer64978.2025.10975735

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computing*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., & Kaufman, J. D. (2013). Long-term air pollution exposure and cardio-respiratory mortality: A review. *Environmental Health*, *12*(1), 43. https://doi.org/10.1186/1476-069X-12-43

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, *637*(8045), 319–326.

Illescas-Martinez, F., Garcia, L., Garcia-Sanchez, A.-J., Asorey-Cacheda, R., & Garcia-Haro, J. (2025). Air quality forecasting in non-monitored urban areas through machine and deep-learning model. *Expert Systems with Applications*, *284*, 127749. https://www.sciencedirect.com/science/article/pii/S0957417425013715. https://doi.org/https://doi.org/10.1016/j.eswa.2025.127749

Iskandaryan, D., Ramos, F., & Trilles, S. (2020). Air quality prediction in smart cities using machine learning technologies based on sensor data: A review. *Applied Sciences*, *10*(7). https://www.mdpi.com/2076-3417/10/7/2401. https://doi.org/10.3390/app10072401

Jiang, L. L., Cheong, W. S., & Ng, J. S. L. (2026). Multi-step ahead carbon credit price forecasting using time series foundation models. *Applied Soft Computing*, *198*, 115290. https://www.sciencedirect.com/science/article/pii/S1568494626007386. https://doi.org/https://doi.org/10.1016/j.asoc.2026.115290

Jin, M., Zhang, Y., Chen, W., Zhang, K., Liang, Y., Yang, B., Wang, J., Pan, S., & Wen, Q. (2024). Position: What can large language models tell us about time series analysis. In *Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research*, *vol. 235*, 22260 – 22276. JMLR.org.

Kelly, C., Fawkes, J., Habermehl, R., de Ferreyro Monticelli, D., & Zimmerman, N. (2023). Plume dashboard: A free and open-source mobile air quality monitoring dashboard. *Environmental Modelling and Software*, *160*. https://doi.org/10.1016/j.envsoft.2022.105600

Kim, D., Park, J., Lee, J., & Kim, H. (2024). Are self-attentions effective for time series forecasting? In *Advances in neural information processing systems*. (*vol. 37*).

Kumar, R., Naja, M., Pfister, G. G., Barth, M. C., Wiedinmyer, C., & Brasseur, G. P. (2012). Simulations over south asia using the weather research and forecasting model with chemistry (WRF-chem): Chemistry evaluation and initial results. *Geoscientific Model Development*, *5*(3), 619–648.

Mulayim, O. B., Quan, P., Han, L., Ouyang, X., Hong, D., Bergés, M., & Srivastava, M. (2024). Are time series foundation models ready to revolutionize predictive building analytics? In *Proceedings of the 11th ACM international conference on systems for energy-efficient buildings, cities, and transportation* (pp. 169–173).

Nalakurthi, N. V. S. R., Abimbola, I., Ahmed, T., Anton, I., Riaz, K., Ibrahim, Q., Banerjee, A., Tiwari, A., & Gharbia, S. (2024). Challenges and opportunities in calibrating low-cost environmental sensors. *Sensors*, *24*(11). https://www.mdpi.com/1424-8220/24/11/3650. https://doi.org/10.3390/s24113650

Nguyen, M. H., Nguyen, P., Nguyen, K., Le, V. A., Nguyen, T., & Ji, Y. (2021). PM2.5 Prediction using genetic algorithm-based feature selection and encoder-decoder model. *IEEE Access*, *9*, 57338–57350. https://doi.org/10.1109/ACCESS.2021.3072280

Peixeiro, M. (2022). Time series forecasting in python. Simon and Schuster.

Rajagopalan, S., Al-Kindi, S. G., & Brook, R. D. (2018). Air pollution and cardiovascular disease: JACC state-of-the-art review. *Journal of the American College of Cardiology*, *72*(17), 2054 – 2070. https://doi.org/10.1016/j.jacc.2018.07.099

Ramadan, M. S., Abuelgasim, A., & Al Hosani, N. (2024). Advancing air quality forecasting in abu dhabi, UAE using time series models. *Frontiers in Environmental Science*, *vol. 12*

- *2024*. https://doi.org/10.3389/fenvs.2024.1393878
Rasul, K., Ashok, A., Williams, A. R., Ghonia, H., Bhagwatkar, R., Khorasani, A., Darvishi Bayazi, M. J., Adamopoulos, G., Riachi, R., Hassen, N., Biloš, M., Garg, S., Schneider, A., Chapados, N., Drouin, A., Zantedeschi, V., Nevmyvaka, Y., & Rish, I. (2023). Lag-Llama: Towards foundation models for probabilistic time series forecasting, arxiv e-prints. https://doi.org/10.48550/arXiv.2310.08278
Rougier, J., Brady, A., Bamber, J., Chuter, S., Royston, S., Vishwakarma, B. D., Westaway, R., & Ziegler, Y. (2023). The scope of the kalman filter for spatio-temporal applications in environmental science. *Environmetrics*, *34*(1), e2773.
Saboia, J. L. M. (1977). Autoregressive integrated moving average (ARIMA) models for birth forecasting. *Journal of the American Statistical Association*, *72*(358), 264 – 270. https://doi.org/10.1080/01621459.1977.10480989
Shi, X., Wang, S., Nie, Y., Li, D., Ye, Z., Wen, Q., & Jin, M. (2024). Time-MoE: Billion-scale time series foundation models with mixture of experts. arXiv preprint arXiv:2409.16040.
Shyalika, C., Bagga, H. K., Bhatt, A., Prasad, R., Ghazo, A. A., & Sheth, A. (2024). Time series foundational models: Their role in anomaly detection and prediction. arxiv:2412.19286.
Si, M. (2019). Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine learning methods. https://doi.org/10.5281/zenodo.3473833
Si, M., Xiong, Y., Du, S., & Du, K. (2020). Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine-learning methods. *Atmospheric Measurement Techniques*, *13*(4), 1693–1707. https://amt.copernicus.org/articles/13/1693/2020/. https://doi.org/10.5194/amt-13-1693-2020
Skamarock, W. C., & Klemp, J. B. (2008). A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *Journal of Computational Physics*, *227*(7), 3465–3485.
Song, S., Bang, S., Cho, S., Han, H., & Lee, S. (2022). Attentive multi-task prediction of atmospheric particulate matter: Effect of the COVID-19 pandemic. *IEEE Access*, *10*, 10176–10190. https://doi.org/10.1109/ACCESS.2022.3144588
Su, I.-F., Chung, Y.-C., Lee, C., & Huang, P.-M. (2023). Effective PM2.5 concentration forecasting based on multiple spatial–temporal GNN for areas without monitoring stations. *Expert Systems with Applications*, *234*, 121074. https://www.sciencedirect.com/science/article/pii/S0957417423015762. https://doi.org/https://doi.org/10.1016/j.eswa.2023.121074
Tancev, G., & Pascale, C. (2020). The relocation problem of field calibrated low-cost sensor systems in air quality monitoring: A sampling bias. *Sensors*, *20*(21). https://www.mdpi.com/1424-8220/20/21/6198. https://doi.org/10.3390/s20216198
Tawakuli, A., Havers, B., Gulisano, V., Kaiser, D., & Engel, T. (2025). Survey:time-series data preprocessing: A survey and an empirical analysis. *Journal of Engineering Research*, *13*(2), 674–711. https://www.sciencedirect.com/science/article/pii/S2307187724000452. https://doi.org/https://doi.org/10.1016/j.jer.2024.02.018
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. arxiv:1706.03762.
Villanueva, E., Espezua, S., Castelar, G., Diaz, K., & Ingaroca, E. (2023). Smart multi-sensor calibration of low-cost particulate matter monitors. *Sensors*, *23*(7). https://www.mdpi.com/1424-8220/23/7/3776. https://doi.org/10.3390/s23073776
Wang, C., Qi, Q., Wang, J., Sun, H., Zhuang, Z., Wu, J., Zhang, L., & Liao, J. (2025a). ChatTime: A unified multimodal time series foundation model bridging numerical and textual data. In *AAAI conference on artificial intelligence*.
Wang, J., Wu, T., Mao, J., & Chen, H. (2024). A forecasting framework on fusion of spatiotemporal features for multi-station PM2.5. *Expert Systems with Applications*, *238*, 121951. https://www.sciencedirect.com/science/article/pii/S0957417423024533. https://doi.org/https://doi.org/10.1016/j.eswa.2023.121951
Wang, K., Zhang, Y., Jang, C., Phillips, S., & Wang, B. (2009). Modeling intercontinental air pollution transport over the trans-pacific region in 2001 using the community multiscale air quality modeling system. *Journal of Geophysical Research: Atmospheres*, *114*(D4) https://doi.org/10.1029/2008JD010807.
Wang, T., ENNADIR, S., Pertoft, J., Gandler, G. Z., Cao, L., Senane, Z., Katsarou, S., Asadi, S., Karlsson, A., & Smirnov, O. (2025b). Frequency matters: When time series foundation models fail under spectral shift. In *Recent advances in time series foundation models have we reached the 'BERT moment'?* https://openreview.net/forum?id=frTj7liRSi.
Xu, S., Kamarthi, H., Liu, H., & Prakash, B. A. (2025). In-context pre-trained time-series foundation models adapt to unseen tasks. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)* (5386–5390). Association for Computing Machinery. https://doi.org/10.1145/3746252.3760801.
Zhang, W., Han, J., Xu, Z., Ni, H., Liu, H., & Xiong, H. (2024a). Urban foundation models: A survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining* KDD '24 (p. 6633–6643). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3637528.3671453
Zhang, Y., & Gilpin, W. (2025). Zero-shot forecasting of chaotic systems. In *International Conference on Learning Representations (ICLR 2025)*. 93873 – 93899.
Zhang, Z., Zhang, S., Chen, C., & Yuan, J. (2024b). A systematic survey of air quality prediction based on deep learning. *Alexandria Engineering Journal*, *93*, 128–141. https://www.sciencedirect.com/science/article/pii/S1110016824002485. https://doi.org/https://doi.org/10.1016/j.aej.2024.03.031
Zhao, H., Su, Y., Qian, C., & Huang, H. (2025). Forecasting PM2.5 concentrations with spatial-temporal hypergraph attention network: Modeling geospatial correlations and latent associations among stations. *Expert Systems with Applications*, *288*, 128298. https://www.sciencedirect.com/science/article/pii/S0957417425019177. https://doi.org/https://doi.org/10.1016/j.eswa.2025.128298
Zheng, Q., Tian, X., Yu, Z., Jin, B., Jiang, N., Ding, Y., Yang, M., Elhanashi, A., Saponara, S., & Kpalma, K. (2024). Application of complete ensemble empirical mode decomposition based multi-stream informer (CEEMD-msi) in PM2.5 concentration long-term prediction. *Expert Systems with Applications*, *245*, 123008. https://www.sciencedirect.com/science/article/pii/S0957417423035108. https://doi.org/https://doi.org/10.1016/j.eswa.2023.123008
Zhou, S., Wang, W., Zhu, L., Qiao, Q., & Kang, Y. (2024). Deep-learning architecture for PM2.5 concentration prediction: A review. *Environmental Science and Ecotechnology*, *21*, 100400. https://www.sciencedirect.com/science/article/pii/S2666498424000140. https://doi.org/https://doi.org/10.1016/j.ese.2024.100400
Zhu, L., Chen, L., & Chen, H. (2026). Short-term air quality prediction using a multi-scale attention fusion model with 3DIGAT-CBAM-biLSTM based on spatio-temporal correlation. *Expert Systems with Applications*, *298*, 129856. https://www.sciencedirect.com/science/article/pii/S0957417425034712. https://doi.org/https://doi.org/10.1016/j.eswa.2025.129856