E-STGCN: Extreme Spatiotemporal Graph Convolutional Networks for Air Quality Forecasting

Madhurima Panja[1, 0], Tanujit Chakraborty[1, 0, 2 5], Anubhab Biswas[3], Soudeep Deb[4]

[1] Department of Science and Engineering, Sorbonne University Abu Dhabi, UAE.

[2] Sorbonne Center for Artificial Intelligence, Sorbonne University, Paris, France.

[3] University of Applied Sciences and Arts of Southern Switzerland, Switzerland.

[4] Indian Institute of Management Bangalore, India.

Abstract

Modeling and forecasting air quality plays a crucial role in informed air pollution management and protecting public health. The air quality data of a region, collected through various pollution monitoring stations, display nonlinearity, nonstationarity, and highly dynamic nature and detain intense stochastic spatiotemporal correlation. Geometric deep learning models such as Spatiotemporal Graph Convolutional Networks (STGCN) can capture spatial dependence while forecasting temporal time series data for different sensor locations. Another key characteristic often ignored by these models is the presence of extreme observations in the air pollutant levels for severely polluted cities worldwide. Extreme value theory is a commonly used statistical method to predict the expected number of violations of the National Ambient Air Quality Standards for air pollutant concentration levels. This study develops an extreme value theory-based STGCN model (E-STGCN) for air pollution data to incorporate extreme behavior across pollutant concentrations. Along with spatial and temporal components, E-STGCN uses generalized Pareto distribution to investigate the extreme behavior of different air pollutants and incorporate it inside graph convolutional networks. The proposal is then applied to analyze air pollution data ($PM_{2.5}$, PM_{10}, and NO_2) of 37 monitoring stations across Delhi, India. The forecasting performance for different test horizons is evaluated compared to benchmark forecasters (both temporal and spatiotemporal). It was found that E-STGCN has consistent performance across all the seasons in Delhi, India, and the robustness of our results has also been evaluated empirically. Moreover, combined with conformal prediction, E-STGCN can also produce probabilistic prediction intervals.

Keywords: Air quality, graph convolutional networks, extreme value modeling, spatiotemporal forecasting.

1. Introduction

The rapid acceleration of industrialization and urbanization has spurred economic growth globally but has also exacerbated environmental issues, with air pollution ranking among the most pressing concerns (Brunekreef and Holgate, 2002; Shaddick et al., 2020). According to the World Health Organization (WHO), approximately seven million premature deaths per year are linked to air pollution, a figure emphasizing the critical need for effective air quality management policies[1]. The most harmful air pollutants include particulate matters (PM), nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), and carbon monoxide (CO),

---

[0] *Equal Contributions*

[5] *Corresponding author*: *Mail*: tanujit.chakraborty@sorbonne.ae

[1] https://www.who.int/health-topics/air-pollution

which are closely monitored due to their impact on public health, particularly in terms of cardiovascular and respiratory diseases (Lelieveld et al., 2015; Olaniyan et al., 2020). Recognizing this threat, the United Nations has included air quality as a key component of its Sustainable Development Goals (SDGs)[2]. Likewise, the U.S. Environmental Protection Agency and equivalent authorities worldwide have implemented National Ambient Air Quality Standards (NAAQS) to limit pollution levels, which are crucial for safeguarding human health and the environment. Our focus in this article is on India, where the Central Pollution Control Board (CPCB) specifies that the hourly average concentrations of PM with a diameter of 2.5 micrometers or less ($PM_{2.5}$) and PM with a diameter of 10 micrometers or less (PM_{10}) pollutants should not exceed 60 micrograms per cubic meter ($\mu g/m^3$) and 100 $\mu g/m^3$, respectively[3]. However, in practice, air quality levels often surpass these predefined standards. Data from 37 monitoring stations (spatial locations) across Delhi, the capital city of India, collected over five years from 2019 to 2023[4], shows average $PM_{2.5}$ and PM_{10} concentrations exceeding 100 $\mu g/m^3$ and 200 $\mu g/m^3$ respectively, significantly above the recommended limits. The situation deteriorates further with the onset of winter months due to low temperature and Delhi's landlocked geographical location, which hinders the dispersion of pollutants by wind. As a result, Delhi experiences a surge in pollution during winter, increasing the risk of chronic respiratory and cardiovascular diseases, neurological disorders, and a higher burden of mortality (Salvi et al., 2018). Pandey et al. (2021) further highlight that air pollution adversely affects India's economic growth as well. Given these concerns, our study focuses on developing a spatiotemporal forecasting model to improve air quality predictions in urban environments. Such models are essential for informing public behavior and helping authorities implement timely interventions to mitigate health risks.

Research on air quality forecasting can broadly be classified into two categories: physical models and data-driven methods. Traditional physical models rely on fundamental principles of atmospheric science to simulate the emission, transport, and dispersion of pollutants within a target area. A couple of well-known methods in this category are the community multi-scale air quality (Byun and Schere, 2006), and the nested air quality prediction model system (Wang et al., 2014). However, these often require extensive theoretical knowledge, carefully selected features, and region-specific parameters, prohibiting their usage from building a real-time air quality monitoring system. Other methods like Gaussian plume models or the Operational Street Canyon models lack the accuracy needed for real-time forecasting due to their reliance on limited parameters (Vardoulakis et al., 2003; Byun and Schere, 2006). In contrast, data-driven methods, which leverage historical information to capture pollution trends, have shown some promise (Lei et al., 2019). Thus, traditional statistical models such as ARIMA (autoregressive integrated moving average) and dynamic factor models are widely used (see, e.g., Kumar and Jain, 2010), but they are limited in their capacity to handle complex and nonlinear interactions in air quality data. To that end, recent advancements in machine learning, particularly deep learning, have significantly improved forecasting accuracy. For instance, Li et al. (2017) showed that long short-term memory (LSTM) networks can capture complex temporal dependencies and outperform traditional models in air quality forecasting. Du et al. (2019) introduced a novel hybrid deep learning framework by combining bi-directional LSTM and one-dimensional convolutional neural networks. Extant literature on pollution forecasting also include the use of recurrent neural networks (Ong et al., 2016), transformers (Vaswani, 2017), or temporal convolutional network (Samal et al., 2021). However, a critical limitation of many deep learning models is their focus on temporal data alone, often overlooking spatial dependencies between monitoring locations. Since pollutant levels at any given station are influenced by neighboring locations, a spatiotemporal approach is essential to accurately model air pollution dispersion (Zhou et al., 2024). Graph-based modeling strategies have gained significant attention in this regard, with graph neural networks (GNN) and graph convolutional networks (GCN) revolutionizing spatiotemporal forecasting (Scarselli et al., 2008). In the current context, Gao and Li (2021) leveraged GNNs with LSTMs to capture spatiotemporal information in $PM_{2.5}$ level. GCNs are also effective for air quality modeling, particularly due to their ability to perform convolutional operations that propagate information between

nodes in a graph, thus leveraging localized aggregation of features from neighboring nodes (Yu et al., 2018). For a comprehensive discussion in this context, refer to Atluri et al. (2018); Jin et al. (2024).

The application of GCNs for modeling spatial dynamics of air pollution monitoring stations is however limited due to scalability and data sparsity issues. Moreover, existing forecasting architectures often struggle to accurately predict peaks in a time series, which is critical for air pollution forecasting to anticipate exceedances beyond regulatory thresholds and ensure ambient air quality standards are maintained. Due to the catastrophic nature of extreme values in air quality data, it is essential to understand and predict values that exceed the NAAQS threshold for developing effective early warning systems. For that, we turn attention to the extreme value theory (EVT) which provides a statistical framework for analyzing rare events, offering insights into the probability and distribution of extreme pollutant concentrations (Coles et al., 2001). This framework has been successfully applied in various fields, including hydrology, climate studies, and air quality analysis, to predict the likelihood of exceeding established safety thresholds (Horowitz, 1980; Sharma et al., 1999; Ray et al., 2023). In air quality studies, EVT methods such as block maxima (BM) and peaks over threshold (POT) have been employed to model pollutant extremes (Reiss et al., 1997). The probability distribution of these extreme events can be modeled using the generalized extreme value (GEV) distribution for BM and generalized Pareto (GP) distributions for the POT method. These techniques help in estimating the likelihood of extreme occurrences, allowing for the detection of potential rare events. EVT-based studies have been instrumental in forecasting pollution exceedances, thus supporting effective intervention strategies (Kan and Chen, 2004; Sfetsos et al., 2006). The reader is further referred to Martins et al. (2017) for a comprehensive review of EVT tools in air pollution problems.

Interestingly, despite its proven utility, EVT has not been combined with spatiotemporal forecasting methods to build early warning systems for environmental preparedness. Our study aims to bridge this gap by introducing a novel EVT-guided spatiotemporal graph convolutional network (E-STGCN) model, to handle the nonlinear, nonstationary behavior for major air pollutants in Delhi, specifically $PM_{2.5}$, PM_{10}, and NO_2. We examine the extreme behavior of these pollutants across 37 sensor locations using the POT method, modeled by the GP distribution. Integrating these insights into a spatiotemporal GCN (STGCN) enhances its ability to forecast extreme values within time series data. Note that STGCN is a robust class of space-time deep learning models designed for inference on graph structures with temporal dependencies. By leveraging STGCN, we effectively represent the spatial configuration of monitoring stations and address data sparsity. The integration of EVT with STGCN enables more accurate modeling of spatiotemporal embeddings, capturing peaks in pollutant levels with greater precision. Furthermore, our proposed framework is scalable and capable of generating multi-step forecasts for both low and high-frequency spatiotemporal datasets. Unlike traditional GNN models, typically optimized for hourly predictions and shorter horizons (e.g., 12 to 72 hours), E-STGCN handles daily air quality data, providing reliable long-term forecasts at 30-, 60-, and 90-day horizons. Additionally, we apply conformal prediction methods to quantify forecast uncertainties, offering critical probabilistic insights for policy planning.

To establish the efficacy of the proposed algorithm, we rigorously evaluate the model against state-of-the-art approaches. A list of these competing methods and their modeling capabilities are summarized in Table 1. Among time-dependent models, we consider the ubiquitous ARIMA approach (Box et al., 1970) as well as several deep learning techniques including LSTM (Hochreiter and Schmidhuber, 1997), temporal convolutional networks (TCN) (Chen et al., 2020), DeepAR (Salinas et al., 2020), Transformers (Wu et al., 2020), and NBeats (Oreshkin et al., 2019). For the spatiotemporal models, we evaluate the performance of Space-time Autoregressive Moving Average (STARMA) (Pfeifer and Deutrch, 1980), GSTAR (Cliff and Ord, 1975), Fast Gaussian Process (GpGp) (Guinness, 2018), Spatiotemporal Neural Network (STNN) (Saha et al., 2020), STGCN (Yu et al., 2018), and DeepKriging (Nag et al., 2023). In the interest of space, more details of these models are provided in Section S.2 of the supplement.

The remainder of this paper is organized as follows. Section 2 presents the results from the extreme value analysis of air quality data. In Section 3, we introduce the proposed E-STGCN architecture. Section

3

Table 1: Comparison of forecasting frameworks. The columns indicate whether each model can address spatiotemporal correlations, nonlinearity, and stationarity in time series data. Additional columns assess whether the method can produce probabilistic forecasts, scale effectively for large datasets, and handle extreme observations.

Models	Spatiotemporal	Nonlinear	Non-stationarity	Probabilistic Forecasting	Scalability	Extreme Value Handling
ARIMA	✗	✗	✔	✔	✗	✗
LSTM	✗	✔	✔	✗	✔	✗
TCN	✗	✔	✔	✗	✔	✗
DeepAR	✗	✔	✔	✗	✔	✗
Transformers	✗	✔	✔	✗	✔	✗
NBeats	✗	✔	✔	✗	✔	✗
STARMA	✔	✗	✗	✔	✗	✗
GSTAR	✔	✗	✔	✗	✗	✗
GpGp	✔	✗	✗	✔	✔	✗
STNN	✔	✔	✔	✗	✗	✗
STGCN	✔	✔	✔	✗	✔	✗
DeepKriging	✔	✔	✔	✔	✗	✗
Proposed E-STGCN	✔	✔	✔	✔	✔	✔

4 outlines the experimental setup and reports the air quality forecasting results. In Section 5, we discuss the implications of our approach to air quality forecasting. Finally, Section 6 concludes the paper and suggests future research directions.

2. Preliminaries on Extreme Value Theory

Extreme Value Theory (EVT) focuses on analyzing the stochastic behavior of rare or extreme events within a given process. The goal of extreme value analysis is to quantify unusually large or small events and estimate the probability of these extreme occurrences, which differ significantly from the more common observations in the data. EVT deals with the asymptotic distribution of extreme order statistics, especially in the context of large datasets. This theory has been implemented in diverse domains, including earth sciences (Katz et al., 2002), economics and finance (Marimoutou et al., 2009), public health (Thomas et al., 2016), and engineering (Castillo, 2012), among others. As mentioned above, statistical methods for modeling extreme events primarily rely on two approaches, i.e., block maxima and peaks over the threshold. Below, we briefly summarize the EVT methods that were utilized in this study.

2.1. Block Maxima (BM) Approach

The BM method analyzes extreme events in a time series dataset (Gumbel, 1958). Given a sequence of time-dependent observations, this method divides the dataset into equal-sized non-overlapping blocks and considers the maximum value from each block as the extreme values of the time series. The probability distribution of the extremes is modeled using the generalized extreme value (GEV) distribution. To mathematically explain this, let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (iid) random variables with continuous distribution function $F(\cdot)$. Then, as $n \longrightarrow \infty$, the distribution of $M_n = \max_{1 \leqslant i \leqslant n} X_i$ converges to $G(x)$, called the GEV distribution, defined by (following Fisher and Tippett, 1928):

$$G(x) = \begin{cases} \exp\left\{ -\left(1 + \xi_G \left(\frac{x - \mu_G}{\sigma_G}\right)\right)^{-1/\xi_G} \right\} & \text{if } \xi_G \neq 0, \\ \exp\left\{ -\exp\left(-\left(\frac{x - \mu_G}{\sigma_G}\right)\right) \right\} & \text{if } \xi_G = 0. \end{cases}$$

In the above, $\xi_G \in \mathbb{R}$ is the extreme value index and it controls the shape of the distribution, $\mu_G \in \mathbb{R}$ is the location parameter, and $\sigma_G > 0$ is the scale parameter. Depending on the tail behavior of the distribution, which is influenced by ξ_G, the GEV family can be classified into three extreme value distributions: Gumbel ($\xi_G = 0$), Fréchet ($\xi_G > 0$), and Weibull ($\xi_G < 0$). While the Gumbel type distribution is suitable for modeling the extremes of the exponentially decaying-tailed distribution, the Fréchet and the Weibull families are the reference class for the extremes of heavy-tailed and finite-tailed distributions, respectively (Rocco, 2014).

Although the block maxima method has been widely used for extreme value analysis, it has several drawbacks. The partitioning of the dataset in this approach leads to significant information loss, as only the maximum value from each block is retained, potentially missing multiple extreme observations within a block. Also, usually, multiple extreme observations happen within a short time interval which cannot be captured by the block maxima method. POT tries to overcome the disadvantages of the BM approach.

2.2. Peaks Over Threshold (POT) Approach

The POT approach is a key technique in EVT that identifies observations exceeding a pre-selected threshold, known as extreme values (Balkema and De Haan, 1974). By concentrating only on observations above the threshold, the POT approach offers an efficient and accurate mechanism for modeling tail behavior compared to conventional methods that assess the entire distribution. Given a time series dataset $\{e_1, e_2, \ldots, e_l\}$ and a threshold τ^* (any observations that exceed the threshold are called extreme events), the POT approach selects extreme events when $e_i > \tau^*$. The distribution of exceedances over the large threshold τ^* asymptotically follows a Generalized Pareto (GP) distribution. To explain it mathematically, let $\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_n$ be a series of iid random variables with a marginal distribution $Q(\cdot)$. Pickands III (1975) approximated the exceedance distribution for sufficiently large threshold values using a GP distribution, defined by

$$\mathcal{H}(z) = \begin{cases} 1 - \left(1 + \frac{\xi z}{\sigma}\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - \exp\left(-\frac{z}{\sigma}\right) & \text{if } \xi = 0. \end{cases} \tag{1}$$

Here, $\xi \in \mathbb{R}$ is the shape parameter and $\sigma > 0$ is the scale parameter of the GP distribution. The shape parameter ξ plays a key role in determining the qualitative behavior of the GP distribution and influences its domain of attraction. When $\xi = 0$, $\mathcal{H}(z)$ belongs to the Gumbel distribution family, where the probability of extreme observations decreases exponentially, as indicated by its light tails. For $\xi > 0$, $\mathcal{H}(z)$ follows a Fréchet distribution characterized by heavy tails, suggesting more frequent extreme observations. Conversely, when $\xi < 0$, $\mathcal{H}(z)$ corresponds to a Weibull distribution with short tails, implying a lower probability of extreme observations.

The POT approach offers a robust technique for effectively modeling extreme observations with minimum data loss. It is particularly suited for capturing the clustering effect, which is a prominent phenomenon in extreme events. The advantages of this method for modeling extreme air pollution levels are demonstrated in AL-Dhurafi et al. (2018), where the POT approach has been applied to investigate air pollution index exceedances in urban areas of Peninsular Malaysia.

2.3. Methods for Threshold Selection

The choice of threshold plays a key role in identifying the extreme observations in the dataset, thus significantly impacting the effectiveness of the POT approach. If a low threshold is selected, usual observations can be treated as extreme and violate asymptotic assumptions. On the contrary, a high threshold value can

overlook potential extreme observations by treating too few data points as extreme. The threshold selection can be done objectively through a bias-variance trade-off or determined subjectively, with input from domain experts. Among various statistical procedures, the mean excess plot (MEP) is a popular approach for determining the threshold in the POT method (Benktander and Segerdahl, 1960). The mean excess function of the random variable $\mathcal{Z}$ with distribution function $\mathcal{Q}_{\mathcal{Z}(z)}$ and right endpoint z_R is given by

$$\mathrm{ME}\left(\tau^*\right) := E\left(\mathcal{Z} - \tau^* \mid \mathcal{Z} > \tau^*\right) = \int_{\tau^*}^{z_R} \left(\frac{1 - \mathcal{Q}_{\mathcal{Z}}(s)}{1 - \mathcal{Q}_{\mathcal{Z}}(\tau^*)}\right) ds,$$

provided $E\left(\mathcal{Z}\right) < \infty$ (Embrechts et al., 2013). Thus, if we model the statistical properties of exceedance for any arbitrarily chosen random variable $\mathcal{Z}$ among $\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_n$, with GP (σ, ξ) distribution (as in (1)), then the expected value of $\mathcal{Z}$ will be finite if and only if $\xi < 1$ and the mean excess function can be computed as:

$$\mathrm{ME}\left(\tau^*\right) = \frac{\sigma}{1 - \xi} + \frac{\xi}{1 - \xi}\tau^*,$$

where $0 \leqslant \tau^* < \infty$ if $0 \leqslant \xi < 1$ and $0 \leqslant \tau^* \leqslant -\frac{\sigma}{\xi}$ if $\xi < 0$. A natural estimate of the mean excess function, $\widehat{\mathrm{ME}}\left(\tau^*\right)$, is defined by

$$\widehat{\mathrm{ME}}\left(\tau^*\right) = \frac{\sum_{i=1}^{n}\left(z_i - \tau^*\right)\mathrm{I}_{[z_i > \tau^*]}}{\sum_{i=1}^{n}\mathrm{I}_{[z_i > \tau^*]}}; \ \tau^* \geqslant 0.$$

The MEP method considers the set of all points $\{(\tau^*, \widehat{\mathrm{ME}}(\tau^*)) : \tau^* < z_{(n)}\}$ where $z_{(n)}$ is the highest order statistic from the sample. In principle, the MEP will appear linear if the exceedance observations are fitted with GP distribution, which has a finite mean. This plot has been utilized in various fields, including environmental science (Ghosh and Resnick, 2010), finance (Chukwudum et al., 2020), and others.

3. Proposed Methodology

This section introduces the proposed E-STGCN method for spatiotemporal forecasting of air pollution concentration levels in the presence of extreme observations. Specifically, Section 3.1 outlines the mathematical formulation of the spatiotemporal air pollution forecasting problem, while Section 3.2 provides an overview of the E-STGCN architecture, with detailed descriptions of the components within each module of the proposed framework.

3.1. Problem Formulation

In this study, we address the air quality prediction problem as a spatiotemporal forecasting task, where the key challenge is to model temporal patterns from historical data while simultaneously capturing the spatial relationships between multiple air quality monitoring stations. Given a sequence of air pollutant concentrations across N stations at timestamp t, $\mathbf{X}_t = \left\{X_t^1, X_t^2, \ldots, X_t^N\right\} \in \mathbb{R}^N$, our goal is to generate q-step-ahead ($q \geqslant 1$) forecasts for the pollutant levels based on T historical observations. To achieve this, we develop a forecasting model that integrates extreme value theory with STGCN for modeling the spatiotemporal correlations among the N monitoring stations, accounting for the presence of extreme observations in the dataset. Let us use $F_{\mathrm{E\text{-}STGCN}}$ to denote the forecasting function from our algorithm. To understand the objective formally, let $G = \{V, E\}$ denote an undirected graph, where $V \in \mathbb{R}^N$ represents the set of nodes, corresponding to the monitoring stations and the set of edges $E \in \mathbb{R}^{N \times N}$ indicates the spatial correlations between the stations. Mathematically, the air quality forecasting problem can then be expressed as:

$$\widehat{\mathbf{X}}_{T+1}, \widehat{\mathbf{X}}_{T+2}, \ldots, \widehat{\mathbf{X}}_{T+q} = F_{\mathrm{E\text{-}STGCN}}\left(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T; G, W\right)$$

where $\widehat{\mathbf{X}}_{T+i} = \{\widehat{X}_{T+i}^1, \widehat{X}_{T+i}^2, \ldots, \widehat{X}_{T+i}^N\} \in \mathbb{R}^N$ represents the i-step-ahead forecast of air pollution concentrations for the N monitoring stations, computed based on the T historical observations, with G modeling spatial dependencies, and W as the learnable parameters of $F_{\text{E-STGCN}}$.

3.2. E-STGCN Model Overview

The overall architecture of the E-STGCN framework, depicted in Fig. 1, consists of three primary modules: the spatial module, the temporal module, and the EVT module. The spatial module maps the input data onto attributed spatiotemporal graphs and learns the underlying spatial correlations. These learned graph structures and historical air pollutant concentrations are processed through the spatial blocks comprising GCNs and fully connected neural networks, which capture dynamic temporal information and spatial influences. The output from the spatial module is then fed into the temporal module, where the future trajectories of air pollutant concentrations are predicted using recurrent LSTM layers and a fully connected dense layer. The EVT module, a key component of the E-STGCN architecture, fits GP distribution to the historical air pollutant concentrations that exceed permissible thresholds. It then constrains the output of the temporal module using a combination of data-driven loss and POT loss (discussed in Section 3.2.3). Integrating the EVT-based knowledge with the spatiotemporal information learned from the Spatial and temporal modules enables the E-STGCN framework to accurately capture the underlying dynamics of the air pollutant concentrations, particularly in the presence of threshold exceedances. There is an intuitive connection between the proposed E-STGCN and that of Physics-informed machine learning (Karniadakis et al., 2021), which combines (noisy) data with physical models and implements through deep neural networks. However, in our method, instead of a physical model, we use an EVT-based GP distribution-fitted model as the building block for E-STGCN. Our proposed framework is designed to accurately forecast extreme observations in pollution concentration levels by effectively capturing underlying spatiotemporal patterns.

3.2.1. Spatial Module

In the spatial domain, the air pollutant concentrations at different sensor locations influence each other with varying intensities, and most interactions are dynamic. To capture the spatial correlations among the monitoring stations, we employ graph convolution operations. Typically, GCNs allow convolution operations on arbitrary graph structures, enabling the learning of node-order invariant representations. In the E-STGCN framework, we model the historical air pollutant concentrations using GCN by considering the geographical locations of the monitoring stations as nodes, which form the basis of spatial dependencies. Thus, the undirected graph $G = \{V, E\}$, with V nodes and E connecting edges, can be represented using an adjacency matrix A for efficient computer processing. Specifically, the adjacency matrix A is static and is constructed based on the weighted Haversine distance (d_{ij}) between the geographical locations of the ith station (with latitude ϕ^i, longitude λ^i) and the jth station as:

$$d_{ij} = 2R\sin^{-1}\left[\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos\left(\lambda^i\right)\cos\left(\lambda^j\right)\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right],$$

where $\Delta\phi = \phi^i - \phi^j$, $\Delta\lambda = \lambda^i - \lambda^j$, and R represents the earth's radius. The weights of the adjacency matrix, indicating the similarity between the corresponding nodes, are computed using a Gaussian kernel as

$$a_{ij} = \exp\left(-\frac{d_{ij}^2}{\tilde{\sigma}^2}\right), \text{ when } i \neq j \text{ and } \exp\left(-\frac{d_{ij}^2}{\tilde{\sigma}^2}\right) \geqslant \epsilon, \tag{2}$$

where $\tilde{\sigma}^2$ and ϵ are the parameters that control the distribution and sparsity of the adjacency matrix A. Specifically, if the distance between the nodes exceeds $\sqrt{-\tilde{\sigma}^2 \ln \epsilon}$, no edges are considered between the nodes.
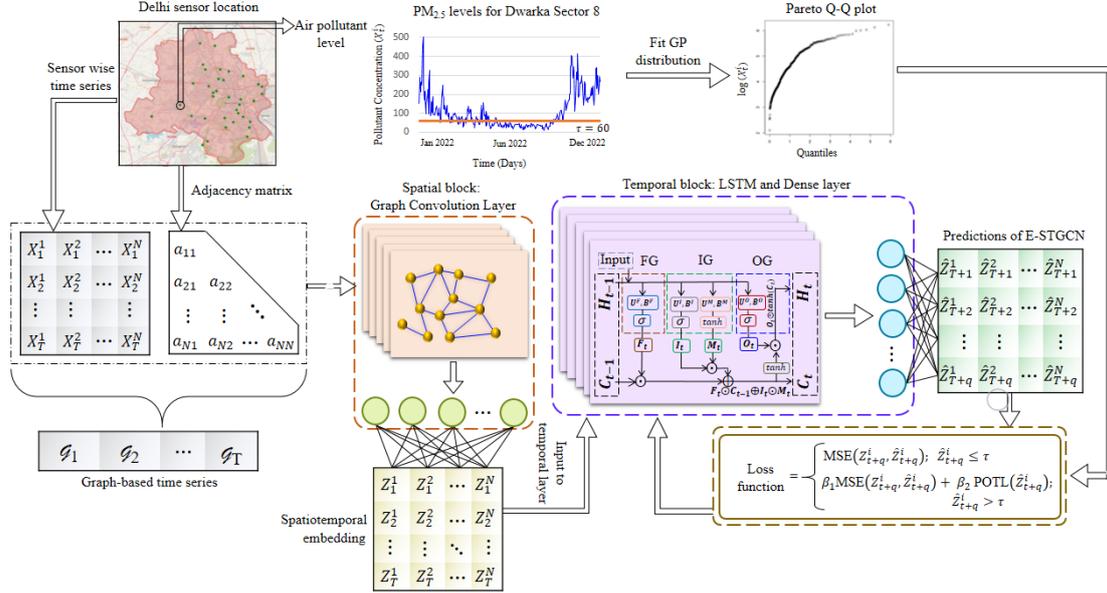
Figure 1: **Extreme Spatiotemporal Graph Convolutional Networks (E-STGCN)**. Daily air pollution concentration levels from different regions of Delhi, along with the corresponding adjacency matrix, are processed through a Graph Convolutional Network (GCN) and a dense layer to generate spatiotemporal embeddings. To account for extreme values, each sensor's time series data is modeled using a Generalized Pareto (GP) distribution. The GCN-embedded data is then passed through an LSTM layer, followed by a dense layer, to produce accurate forecasts. The network is trained using a modified loss function that combines the conventional mean squared error (MSE) loss with a peaks-over-threshold loss (POTL) when predictions exceed a predefined threshold (τ).

With a slight abuse of terminology, let us represent the air pollutant concentrations monitored at N stations over T timestamps as spatiotemporal graphs $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_T\}$, where each graph $\mathcal{G}_t = \{\mathbf{X}_t, A\}$ consists of $\mathbf{X}_t \in \mathbb{R}^N$, representing the pollutant attributes at time t, and $A \in \mathbb{R}^{N \times N}$, providing the structural information for the N stations. To map the non-Euclidean spatiotemporal graphs to spatiotemporal node embeddings, we perform localized convolutions of the node neighborhood using GCN layers. The GCN model mimics the convolutional neural networks (CNN) filters by constructing polynomial filters over neighboring nodes, which can be approximated by Chebyshev's polynomial of order d as:

$$\mathcal{P}_w(L) = \sum_{u=0}^{d} w_u \mathcal{C}_u(\tilde{L}),$$

where $\mathcal{C}_u$ represents the uth degree Chebyshev polynomial of the first kind and $\tilde{L}$ is the normalized graph Laplacian, defined as $\tilde{L} = \frac{2L}{\zeta_{max}} - I_N$. Here, $L = (D - A) \in \mathbb{R}^{N \times N}$ denotes the graph Laplacian, $D \in \mathbb{R}^{N \times N}$ is the diagonal degree matrix (where $D_{ii} = \sum_j A_{ij}$), and ζ_{max} is the largest eigenvalue of L. Thus, the polynomial $\mathcal{P}_w(L) \in \mathbb{R}^{N \times N}$ represents the convolutional filters, with w being the filter weight. Hence, using $\mathcal{P}_w(L)$ on the input data $\mathbf{X}_t$, we can obtain the graph convolutions as:

$$\mathbf{X}_t^{'} = \mathcal{P}_w(L)\mathbf{X}_t. \tag{3}$$

Kipf and Welling (2016) demonstrated that layer-wise linear formulations can be constructed using multiple stacked localized graph convolution operators, employing a 1st-order approximation of L. This allows for the use of a deeper architecture to learn spatial dependencies without the need to explicitly focus

8

on the parameterization of the dth-order polynomial. Consequently, (3) can be simplified as:

$$\mathbf{X}_t^{'} = w_0 \mathbf{X}_t + w_1 \left(\frac{2(D - A)}{\zeta_{max}} - I_N \right) \mathbf{X}_t, \tag{4}$$

where w_0 and w_1 are the weights of the filters shared across all N nodes and I_N is the identity matrix of order N. By applying a stack of K different polynomial filter layers, which corresponds to a sequence of K graph convolution layers with 1st-order approximations, the spatiotemporal node embeddings for $\mathcal{G}_t = \left\{ X_t^1, X_t^2, \ldots, X_t^N, A \right\}$; $t = 1, 2, \ldots, T$, can be computed as:

$$h_t^{i,(0)} = X_t^i$$

$$h_t^{i,(k)} = f_t^{(k)} \left(W_t^{(k)} \frac{\sum_{j \in \mathcal{N}(i)} h_t^{j,(k-1)}}{|\mathcal{N}(i)|} + B_t^{(k)} h_t^{i,(k-1)} \right) ; k = 1, 2, \ldots K$$

$$Z_t^i = \text{Dense} \left(h_t^{i,(K)} \right),$$

where in the kth iteration, the function $f_t^{(k)}$ and filter weights $\{ W_t^{(k)}, B_t^{(k)} \}$ are shared to update the initial embedding using 1-hop localized convolutions, repeated K times based on a neural message passing mechanism (Gilmer et al., 2017). Thus, $h_t^{i,(k)}$ is the embedding of node i at timestamp t during iteration k, computed by taking the mean of its neighboring nodes and its self-embedding from the previous iteration at time t. The final spatiotemporal representation, Z_t^i, from the spatial block, is computed by modeling the GCN output from the Kth layer using a fully connected dense layer. Consequently, the spatiotemporal embedding $\mathbf{Z}_t = \left\{ Z_t^1, Z_t^2, \ldots, Z_t^N \right\} \in \mathbb{R}^N$ generated from $\mathcal{G}_t$ updates $\mathbf{X}_t$ with the encoded information from $(K-1)$-order neighborhood of the central node through K successive filtering operations. It is of the essence here to point out that the 1st-order approximation of the polynomial filter is highly effective and scalable for large-scale graph structures (Yu et al., 2018).

3.2.2. Temporal Module

The temporal module of the E-STGCN framework is designed to model the spatiotemporal embeddings, $\{ \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_T \}$, learned in the spatial block. Due to the complex sequential dependencies within $\{ \mathbf{Z}_t \}$, we employ an LSTM network, a robust variant of RNNs that effectively overcomes the optimization challenges of conventional RNN architectures (Hochreiter and Schmidhuber, 1997). The LSTM layer introduces a specialized memory cell that replaces the standard hidden nodes of RNNs, ensuring greater speed, stability, and accuracy. These memory cells contain self-connected recurrent edges with fixed weights, which allow the stable computation of gradients across many time steps. The LSTM network utilizes a cell state to store the long-term information, while the short-term information is managed through hidden states, similar to the standard RNN structure. The update process for both the cell state and hidden state is controlled by a gating mechanism consisting of three key components: the forget gate, the input gate, and the output gate. These gates enable the network to effectively learn and retain long-term and short-term dependencies within sequential datasets. At each timestamp t, for each of the N nodes, the LSTM receives p lagged values $\underline{z}_t^i = \{ Z_{t-p-1}^i, Z_{t-p-2}^i, \ldots, Z_t^i \} \in \mathbb{R}^p$ along with the previous hidden state vector H_{t-1}^i as input. It then generates the q-steps-ahead projections of Z_t^i along with the new memory M_t^i, updating both the hidden state H_t^i and the cell state C_t^i. The forget gate plays a critical role in determining which information from the previous cell state should be retained. It calculates the forget gate activation vector (F_t^i) as

$$F_t^i = \phi_1 \left(U_{ZF}^i \underline{z}_t^i + U_{HF}^i H_{t-1}^i + B_F^i \right),$$

where $U_{ZF}^i \in \mathbb{R}^{m \times p}, U_{HF}^i \in \mathbb{R}^{m \times m}$, and $B_F^i \in \mathbb{R}^m$ are the weights associated with the forget gate, and m denotes the number of hidden layers. The activation function ϕ_1 is selected as a sigmoidal activation function, which ensures that the element-wise values of $F_t^i \in \mathbb{R}^m$ lies within $[0, 1]$. A value of 1 signifies that

9

the information from the previous cell state C_{t-1}^i is fully retained in C_t^i, while a value of 0 indicates that previous cell state values are completely discarded. Next, to determine how much of the new input vector $\underline{z}_t^i$ should be integrated into the current cell state, the input gate calculates the corresponding activation vector $I_t^i \in \mathbb{R}^m$ as

$$I_t^i = \phi_1 \left(U_{ZI}^i \underline{z}_t^i + U_{HI}^i H_{t-1}^i + B_I^i \right).$$

Thus, based on the current timestamp data $\underline{z}_t^i$, the new memory vector $M_t^i \in \mathbb{R}^m$ is computed as

$$M_t^i = \phi_2 \left(U_{ZM}^i \underline{z}_t^i + U_{HM}^i H_{t-1}^i + B_M^i \right),$$

where $U_{ZI}^i, U_{ZM}^i \in \mathbb{R}^{m \times p}$ and $U_{HI}^i, U_{HM}^i \in \mathbb{R}^{m \times m}$ are the weight matrices, B_I^i and $B_M^i \in \mathbb{R}^m$ are the bias vectors associated with the input gate. The activation function ϕ_2 is chosen as the tanh activation. The activation vectors from the forget gate and the input gate determine the amount of information to be retained from the previous cell state and the current memory state, respectively, for updating the current cell state as

$$C_t^i = F_t^i \odot C_{t-1}^i \oplus I_t^i \odot M_t^i,$$

where $\odot$ denotes the element-wise multiplication. Finally, the current hidden state is calculated in the output gate based on the activation vector $\left(O_t^i \in \mathbb{R}^m \right)$ of the output gate as:

$$H_t^i = O_t^i \odot \phi_2 \left(C_t^i \right), \quad \text{where } O_t^i = \phi_1 \left(U_{ZO}^i \underline{z}_t^i + U_{HO}^i H_{t-1}^i + B_O^i \right),$$

with $U_{ZO}^i \in \mathbb{R}^{m \times p}$, $U_{HO}^i \in \mathbb{R}^{m \times m}$, and $B_O^i \in \mathbb{R}^m$ being the learnable parameters of the output gate. To compute H_t^i and C_t^i, the initial values are set to $H_0^i = C_0^i = 0$. Consequently, the q-steps-ahead forecast of the air pollutant concentrations for the ith node is obtained using a fully connected dense layer as:

$$\left\{ \widehat{Z}_{t+1}^i, \widehat{Z}_{t+2}^i, \dots, \widehat{Z}_{t+q}^i \right\} = \text{Dense} \left(H_t^i \right).$$

The spatiotemporal representations for all the N monitoring stations are modeled similarly, resulting in q-steps-ahead forecasts. The final output generated by the temporal module effectively captures the sequential patterns of the air pollutant series. However, the model struggles to forecast sudden peaks, which are particularly common in Delhi's air pollutant concentrations during winter months. To address this, we design the EVT module within the E-STGCN architecture, enabling the framework to forecast spatiotemporal dependencies in situations of threshold exceedances accurately.

3.2.3. EVT Module

In the field of air pollution control, Roberts (1979) emphasized that rare events often hold more significance than regular observations. Therefore, prior knowledge of these rare occurrences is crucial for accurate modeling and forecasting of air pollutant concentrations. The spatial and temporal modules of the E-STGCN architecture leverage historical pollutant data from various monitoring stations and their geographical locations to predict future trends. However, their inability to differentiate between common and rare events limits their effectiveness in modeling extreme occurrences. To address this issue, the EVT module, a key component of the E-STGCN framework, utilizes extreme value theory to identify the underlying patterns of air pollutant concentrations associated with rare observations. The theoretical insights gained from the EVT enhance data-centric forecasting strategies, enabling more accurate predictions of potential extreme pollutant concentrations. This integration improves air pollution forecasting and provides deeper insights into extreme events.

In the EVT module, we employ the POT approach to analyze the extreme observations and integrate them into the spatiotemporal forecasts of the previous modules. In the POT method (as discussed in Section

2.2), we examine the behavior of exceedances by fitting a GP distribution to the pollutant concentrations that exceed the NAAQS threshold (τ). Following (1), the conditional GP distribution for the ith monitoring station at time t can be mathematically formulated as:

$$P\left[Z_t^i - \tau \leqslant z_t^i \mid Z_t^i \geqslant \tau\right] = \begin{cases} 1 - \left(1 + \frac{\xi^i z_t^i}{\sigma^i}\right)^{-1/\xi^i} & \text{if } \xi^i \neq 0 \\ 1 - \exp\left(\frac{z_t^i}{\sigma^i}\right) & \text{if } \xi^i = 0, \end{cases}$$

where $\xi^i \in \mathbb{R}$ is the shape parameter and $\sigma^i > 0$ is the scale parameter for the GP distribution fitted to the pollutant concentrations of the ith monitoring station. The shape parameter ξ^i is particularly important as it influences the tail behavior of the GP distribution. We compute the log-likelihood of the fitted GP distribution (POT loss), following Farkas et al. (2024), as

$$\text{POTL}\left(z_t^i\right) = -\log\left(\hat{\sigma}^i\right) - \left(1 + \frac{1}{\hat{\xi}^i}\right)\log\left(1 + \frac{\hat{\xi}^i z_t^i}{\hat{\sigma}^i}\right), \tag{5}$$

where $\hat{\sigma}^i, \hat{\xi}^i$ are estimated based on the training data from the ith monitoring station. We then incorporate the log-likelihood function (5) while designing the loss function of the temporal module. This approach enhances the modeling of threshold exceedances by incorporating knowledge from EVT as prior information to the model.

3.2.4. Optimization

The objective function of the E-STGCN framework is formulated as a combination of the data loss, computed by the mean squared error (MSE), and the POT-based loss function depending on the predicted values exceeding a specified threshold τ. Mathematically, this modified loss function can be expressed as:

$$\text{Loss}\left(Z_{t+q}^i, \widehat{Z}_{t+q}^i\right) = \begin{cases} \text{MSE}\left(Z_{t+q}^i, \widehat{Z}_{t+q}^i\right), & \widehat{Z}_{t+q}^i \leqslant \tau \\ \beta_1 \text{MSE}\left(Z_{t+q}^i, \widehat{Z}_{t+q}^i\right) + \beta_2 \text{POTL}\left(\widehat{Z}_{t+q}^i\right), & \widehat{Z}_{t+q}^i > \tau, \end{cases} \tag{6}$$

where β_1 and β_2 are the hyperparameters that regulate the contributions of data loss and the POT-based loss, respectively. Since the loss function is differentiable almost everywhere, we utilize the backpropagation method to train the corresponding weights. If the predictive values do not exceed the threshold τ, then the model is optimized solely based on the data loss, which is the case in the STGCN model. Thus, the proposed E-STGCN framework generalizes the STGCN architecture with the modified loss function that integrates prior distributional knowledge to the predictions in case of threshold exceedances. This enhances the proposal's ability to generate accurate spatiotemporal forecasts in the presence of peaks in the air pollutant concentration levels. A detailed visualization showcasing the working principle of the E-STGCN framework is provided in Fig. 1.

4. Experimental Evaluation

In this study, we assess the efficiency of the proposed E-STGCN framework by comparing its forecasting performance with several temporal and spatiotemporal forecasters. We use daily data on Delhi's air pollutant concentration levels from January 1, 2019, to December 31, 2022, to train the models and generate forecasts for different months of 2023. To demonstrate the generalizability of our proposal, we evaluate its forecasting performance across three forecast horizons, namely short-term, medium-term, and long-term, spanning over 30 days, 60 days, and 90 days, respectively, using a rolling window approach. For the short-term horizon,

forecasts are computed for each of the 12 months of 2023. The forecast window covers two consecutive months in the medium-term horizon, resulting in 6 cases. There are four forecast windows for the long-term horizon, each covering three successive months. Fig. 2 visually represents the training, validation, and test periods used in the forecasting tasks. The following subsections present a brief description of the air pollutant datasets and their global characteristics (Section 4.1), extreme value analysis of air pollutant concentrations (Section 4.2), performance comparison metrics (Section 4.3), implementation of the proposed framework and experimental results (Section 4.4), statistical significance tests of the experimental results (Section 4.5), and uncertainty quantification of the proposal (Section 4.6).

Figure 2: Dataset split for different forecast evaluation window

4.1. Data and Preliminary Analysis

In this study, we focus on forecasting the daily concentration levels of three major air pollutants, namely $PM_{2.5}$, PM_{10}, and NO_2, and analyze their statistical and global features using the data collected from 37 monitoring stations located in Delhi. Pollution concentrations fluctuate significantly throughout the year across various stations, with ranges between $0.08 - 761.95\,\mu g/m^3$ for $PM_{2.5}$, $1.00 - 923.70\,\mu g/m^3$ for PM_{10}, and $0.13 - 428.15\,\mu g/m^3$ for NO_2. The average concentrations are $102.31\,\mu g/m^3$, $203.48\,\mu g/m^3$, and $43.02\,\mu g/m^3$ for $PM_{2.5}$, PM_{10}, and NO_2 respectively. We also compute the five-point summary statistics, standard deviation (sd), coefficient of variation (cv), skewness, and kurtosis for the pollutant concentrations monitored at different stations. Furthermore, we analyze various global time series features, including long-term dependency, stationarity, linearity, and seasonality for the pollutant levels. A brief description of these features is summarized in Section S.1 of the supplement. The results of the descriptive statistics and global features of $PM_{2.5}$, PM_{10}, and NO_2 datasets, as reported in Tables S.1, S.2, S.3 of the supplement, reveal that the air pollutant series from most of the monitoring stations exhibit long-range dependencies, non-stationary behavior, and nonlinear patterns. Additionally, some datasets display weekly and quarterly seasonality.

Next, in Fig. 3, we visualize the spatial distribution of the air pollutant monitoring stations in Delhi. The upper panel of the plot showcases the average pollution levels observed at each station. These plots highlight that monitoring stations in close proximity tend to record similar pollution concentration levels compared to distant ones, underscoring the spatial dependencies in pollutant concentrations. Moreover, we examine the pairwise correlations between pollution levels at different stations. The correlation heatmap, shown in the lower panel of Fig. 3, emphasizes that stations located geographically closer tend to have stronger correlations. The diagonal elements represent the self-correlation of each pollutant series, which naturally equals 1. Interestingly, certain stations (e.g., stations 1 and 5) show significant pairwise correlations with

distant stations, indicating non-local spatial interactions. These observations provide critical insights for modeling both spatial and temporal dependencies during the forecasting process.

Figure 3: Upper panel: Spatial distribution of the monitoring stations in Delhi and average pollution level of (a) $PM_{2.5}$, (b) PM_{10}, and (c) NO_2. The lower panel represents the pairwise correlation between the level of (a) $PM_{2.5}$, (b) PM_{10}, and (c) NO_2 from each station.

4.2. Extreme Value Modeling of Air Quality Data

In this section, we employ the BM method and the POT approach to detect and model the extreme observations in the air pollutant concentrations. Fig. 4 presents the results of extreme value analysis using the BM method for daily pollution concentrations of $PM_{2.5}$, PM_{10}, and NO_2, measured at the Alipur monitoring station from 2019 to 2022. The other stations display similar behavior as well. In our analysis, we consider a block size of 30 days, representing the maximum value in each block with a red circle and the remaining observations with green circles. From the plots, it can be observed that in several blocks, the maximum values are not necessarily extreme. Conversely, in other blocks, multiple extreme values, apart from the maximum, are abandoned by this method. To address these limitations, we employ the POT approach in our study. For determining the optimal threshold in the POT method, we utilize the MEP approach and demonstrate the results for pollution concentrations of $PM_{2.5}$, PM_{10}, and NO_2 monitored at the same station in Fig. 5. The plot highlights the mean excess value for various threshold (τ^*) with a 95% confidence interval. From the MEP, we can observe that the mean excess value becomes linear beyond the green straight line, indicating that the corresponding value represents the threshold. Specifically, the MEP-based thresholds are 583 for $PM_{2.5}$, 658 for PM_{10}, and 116 for NO_2 datasets. However, using these thresholds results in only 0.14% extreme values for $PM_{2.5}$, PM_{10}, and NO_2 dataset, which is insufficient for effective POT analysis. Therefore, in this study, we opt for a subjective method of threshold selection, utilizing the NAAQS established by the CPCB for industrial, residential, rural, and other areas. Domain experts determine these thresholds to protect public health, vegetation, and the environment. Following the NAAQS recommendation, we set the threshold values as 60 $\mu g/m^3$ for $PM_{2.5}$, 100 $\mu g/m^3$ for PM_{10}, and 80 $\mu g/m^3$ for NO_2 pollutants and examine the exceedance of pollution concentration levels over these thresholds. From

13

Tables S.1, S.2, S.3 of the supplement, the average exceedance levels are 61% for $PM_{2.5}$, 77% for PM_{10}, and 10% for NO_2. To verify the iid assumption of the POT approach for these exceedance datasets, we perform the Durbin-Watson (DW) test (Durbin and Watson, 1971), which detects autocorrelation at lag 1 in the residuals from the regression analysis. The DW test p-values (refer to the above-mentioned tables in the supplement) indicate that for most exceedance time series, lag 1 residuals are uncorrelated. However, for certain stations with limited observations above the threshold, the DW test statistic could not be computed. We also demonstrate the fitting of the GP distribution for different air pollutant concentrations with the selected thresholds in Fig. 6.

Figure 4: Block maxima plot for extreme value analysis of (a) $PM_{2.5}$, (b) PM_{10}, and (c) NO_2 pollutant concentration in Alipur, Delhi monitoring station with each month representing a block. Green points indicate the pollution levels and red circles are the maximum values identified for each block.

Figure 5: Mean excess plot for (a) $PM_{2.5}$, (b) PM_{10}, and (c) NO_2 pollutant concentration in Alipur, Delhi monitoring station. The blue solid line indicates the mean excess level, the red dotted lines represent the 95% confidence interval, and the green solid line is the threshold obtained from the mean excess plot.

Figure 6: (a)-(c) Probability density plots of $PM_{2.5}$, PM_{10}, and NO_2 pollutant concentration extremes in Alipur, Delhi monitoring station, respectively. All histograms are fitted with the probability density (blue) of the generalized Pareto distribution.

14

4.3. Forecasting Performance Evaluation Measures

In our experimental evaluation, we employ four key performance indicators, namely Mean Absolute Error (MAE), Mean Absolute Scaled Error (MASE), Root Mean Squared Error (RMSE), and Symmetric Mean Absolute Percent Error (SMAPE), to quantify the performance of different forecasters (Hyndman, 2018). The mathematical formulations of these metrics are as follows:

$$\text{MAE} = \frac{1}{q}\sum_{t=1}^{q}|X_t^i - \widehat{X}_t^i|, \quad \text{MASE} = \frac{\sum_{t=T+1}^{T+q}|\widehat{X}_t^i - X_t^i|}{\frac{q}{T-1}\sum_{t=2}^{T}|X_t^i - X_{t-1}^i|},$$

$$\text{RMSE} = \sqrt{\frac{1}{q}\sum_{t=1}^{q}(X_t^i - \widehat{X}_t^i)^2}, \text{ and } \text{SMAPE} = \frac{1}{q}\sum_{t=1}^{q}\frac{2|\widehat{X}_t^i - X_t^i|}{|\widehat{X}_t^i| + |X_t^i|} \times 100\%,$$

where q denotes the forecast horizon, $\widehat{X}_t^i$ is the forecast of the actual value X_t^i for the ith station at time t, and T is the size of the training sample. By definition, the minimum value of these performance measures suggests the 'best-fitted' model.

4.4. Experimental Setup and Forecasting Accuracy

In this section, we discuss the implementation of the proposed E-STGCN approach for forecasting air pollutant concentrations in Delhi. To train the sequential workflow of our model, we first utilize the 'fgpd' function from the *evmix* package in R. This function computes the maximum likelihood estimates for the scale parameter ($\sigma^i > 0$) and the shape parameter ($\xi^i \in \mathbb{R}$) of the GP distribution, based on the training dataset for the ith station whenever an exceedance over the NAAQS threshold occurs. These estimated parameters provide prior information regarding extreme values in the training data. Subsequently, we implemented the E-STGCN model in Python to generate the spatiotemporal forecasts for the proposed approach. For modeling the spatial dependencies in the dataset, we compute the adjacency matrix (A) based on the weighted Haversine distance, as in (2). This matrix identifies the neighbors for each sensor, organizing their locations into a graphical structure by identifying relevant nodes and edges. Next, we employ CNNs and a dense layer from the *TensorFlow* library to encode the training data's structural and feature-based information. To model the temporal dependencies, the output of the spatial module is passed through an LSTM layer and a dense layer. The weights of the temporal layer are optimized using a custom loss function, which combines the mean squared error loss with a POT-based loss (as in (6)). This modified loss function leverages prior information about the NAAQS exceedances to enhance the accuracy of air pollution forecasts. Once the E-STGCN model and other benchmark forecasters are implemented, we generate out-of-sample forecasts using a rolling window approach for different forecast horizons. Below, we summarize the performance of our proposal and baseline models from temporal and spatiotemporal paradigms based on several key performance indicators. A brief description of the benchmark temporal and spatiotemporal baseline models used in the experimental analysis, along with their implementation details, is outlined in Section S.2 of the supplement.

Tables 2, 3, and 4 present the performance of the proposed model and the baseline architectures in generating short-term forecasts for PM$_{2.5}$, PM$_{10}$, and NO$_2$ levels, respectively. As indicated in Table 2, the proposed E-STGCN model achieves state-of-the-art performance for several months of 2023. In particular, during the onset (November) and end (February) of winter, our proposed framework generates the most accurate forecasts for the PM$_{2.5}$ concentration levels. While NBeats and ARIMA outperformed in December, January, and April, the E-STGCN model improved ARIMA's forecast by 22.4% (based on the MAE metric) in March, and its performance remained competitive with NBeats. During the summer months, from May to September, the proposed framework outperformed the benchmark models, except in July, where LSTM generated better forecasts. In October, the GpGp model recorded the lowest forecast error. For short-term forecasting of PM$_{10}$, the E-STGCN model consistently performed best for the first three months of

2023, as measured by most performance metrics. During April and May, the PM_{10} concentration levels rarely exceeded the NAAQs threshold, leading to similar performance between the E-STGCN and STGCN models, as the framework was trained primarily on the data loss. The performance of the STNN and the NBeats models is better for June and July. However, the E-STGCN model regained its forecasting superiority from August to October. In November and December, the spatiotemporal GpGp and STARMA models showed competitive performance with our method. For 30-day ahead forecasts of NO_2 concentration levels, the performance of the E-STGCN and STGCN models was very similar, as the average exceedance of NO_2 levels over the NAAQS threshold was around 10%, limiting the use of the POT-based loss function. As shown in Table 4, the E-STGCN and STGCN models provided the lowest forecast errors in several months, including January, February, April, May, and August to November. For March, June, and July, the STARMA and GSTAR models performed best, while in December, their performance was competitive with ARIMA.

The 60-day and 90-day-ahead forecasting results, as presented in Tables 5 and 6, demonstrate how the proposed E-STGCN architecture improves upon the baseline models for longer forecast horizons. For both $PM_{2.5}$ and PM_{10} pollutants, our model delivers the most accurate forecasts during the first two 60-day windows, improving forecast accuracy by 9.73% over the best-performing baseline model. In the subsequent two forecast periods (May–June and July–August), the GSTAR model performs best for $PM_{2.5}$ levels, while for PM_{10}, the GpGp and NBeats models provide similar performance to our proposed model. During the September–October period, the E-STGCN framework achieves the lowest forecast error for both pollutants. However, in the final medium-term forecast period of 2023, the ARIMA model surpasses the performance of all other approaches. For the medium-term and long-term forecasting of NO_2 concentration levels, we observe similar patterns to those seen in the short-term forecasts. The proposed E-STGCN and STGCN models generate similar results and outperform the baseline models in most periods, except for July–August (in the medium-term) and the last two long-term forecast windows, where NBeats, DeepAR, and ARIMA perform better. For other forecast windows, the STARMA and GSTAR models offer competitive performance compared to the best-performing frameworks. For the long-term forecasting task of $PM_{2.5}$ and PM_{10}, our model performs best in two out of the four windows, as indicated by all accuracy metrics. In the April–May–June period, NBeats provides the best performance for $PM_{2.5}$, while Transformers deliver comparable results to E-STGCN and STGCN for PM_{10}. During the third forecast window (July–August-September), DeepAR, STNN, and NBeats achieved the best performance for both pollutants.

The experimental results reported in our study align with the *No Free Lunch* theorem, which suggests that any forecasting model performing best on a particular dataset is likely to perform poorly on others (Wolpert and Macready, 1997). Overall, the E-STGCN framework consistently achieved superior forecast performance across most tasks. Among the temporal models, ARIMA and NBeats performed well, while from the spatiotemporal paradigm, most of the baseline architectures, namely STARMA, GSTAR, GpGp, STNN, and STGCN, demonstrated competitive performance. The DeepKriging framework, however, performed poorly in most forecasting tasks due to scalability issues, which hindered its ability to handle medium-sized spatiotemporal datasets. Additionally, the performance of models like LSTM, TCN, DeepAR, and Transformers lagged behind the E-STGCN framework due to their inability to effectively capture the spatial dependencies associated with pollutant concentrations. We also observed that the proposed E-STGCN consistently outperformed or performed similarly to the STGCN model. This advantage is attributed to the training mechanism adopted in E-STGCN, which employs the POT-based loss function. This modified loss function enables the framework to better capture exceedances of pollution concentration levels over the NAAQS threshold, especially during the onset and end of winter months. The performance of the E-STGCN model in the experimental results highlights its strong generalization ability and adaptability, making it capable of providing high forecast accuracy for datasets having extreme observations. Consequently, the E-STGCN framework offers effective and reliable forecasts for air pollutant concentrations across different forecast horizons.

Table 2: Forecasting performance of the proposed E-STGCN model in comparison to the temporal-only and spatiotemporal forecasting techniques for 30 days ahead forecast horizon of $PM_{2.5}$ pollutant (best results are **highlighted**).

Forecast Period	Metric	Temporal-only Model						Spatiotemporal Model						Proposed E-STGCN
		ARIMA	LSTM	TCN	DeepAR	Transformers	NBeats	STARMA	GSTAR	GpGp	STNN	STGCN	DeepKrigging	
JAN	MAE	**54.88**	160.54	108.88	164.82	93.72	61.72	95.87	63.48	86.09	65.90	69.11	179.16	56.07
	MASE	**0.96**	2.85	1.94	2.93	1.63	1.10	1.67	1.10	1.49	1.18	1.20	3.21	0.98
	RMSE	**71.18**	174.93	137.98	178.87	114.86	78.46	112.01	82.27	107.61	84.55	90.21	192.41	75.34
	SMAPE	**31.20**	152.10	79.80	160.70	60.40	36.50	69.40	39.50	53.30	37.50	41.50	187.10	31.60
FEB	MAE	36.26	91.75	74.64	96.10	37.45	38.97	73.25	70.54	37.82	143.11	36.78	142.16	**30.00**
	MASE	1.23	3.11	2.57	3.27	1.27	1.35	2.47	2.40	1.31	5.27	1.26	4.95	**1.01**
	RMSE	45.68	101.86	96.43	105.83	50.23	49.71	88.40	85.35	49.05	168.10	46.61	224.10	**38.81**
	SMAPE	32.00	127.70	74.70	139.70	33.10	34.00	89.10	83.80	33.50	76.60	32.50	167.80	**26.70**
MAR	MAE	33.32	56.96	61.43	60.59	25.88	25.90	36.81	31.18	31.99	74.22	37.43	76.95	**25.83**
	MASE	1.70	2.78	3.03	2.98	1.35	**1.24**	1.79	1.53	1.75	3.93	1.94	3.83	1.27
	RMSE	38.68	61.09	80.25	64.47	**30.17**	31.38	42.52	35.75	36.87	88.74	42.73	80.11	31.84
	SMAPE	38.60	108.50	76.50	120.90	31.90	**30.00**	59.10	44.70	38.20	63.10	42.00	193.20	35.30
APR	MAE	22.49	47.35	47.05	51.54	33.61	**22.44**	50.17	37.83	36.48	46.06	25.31	68.90	25.37
	MASE	1.22	2.52	2.57	2.75	1.94	**1.20**	2.68	2.02	2.16	2.52	1.38	3.75	1.35
	RMSE	**27.92**	53.81	59.37	57.56	38.34	28.55	57.09	45.31	41.67	53.42	30.26	73.88	33.21
	SMAPE	**32.20**	93.50	82.00	108.50	45.10	33.00	107.10	68.90	48.40	71.60	35.90	195.20	38.20
MAY	MAE	35.10	45.77	47.65	49.80	39.24	30.31	42.19	35.32	40.23	42.10	34.01	85.44	**28.95**
	MASE	1.43	1.83	1.93	2.00	1.63	1.22	1.69	1.42	1.72	1.69	1.39	3.52	**1.16**
	RMSE	42.79	56.97	62.95	60.38	45.89	37.60	52.70	45.30	47.76	53.44	40.22	100.97	**37.37**
	SMAPE	48.40	85.80	83.90	100.40	53.70	44.20	80.10	59.30	53.70	76.60	48.70	141.70	**43.80**
JUN	MAE	23.59	22.77	31.88	26.79	53.82	33.72	27.43	24.78	54.13	29.15	22.14	43.22	**22.03**
	MASE	2.20	**1.99**	2.89	2.39	5.22	3.08	2.47	2.24	5.34	2.60	2.03	3.97	2.04
	RMSE	27.23	27.32	40.01	30.91	55.83	40.94	30.66	28.29	57.68	33.82	27.25	46.10	**26.17**
	SMAPE	45.10	58.40	76.80	75.10	78.80	54.00	85.50	73.60	78.00	65.70	45.30	174.60	**42.80**
JUL	MAE	23.00	**12.98**	22.36	16.23	64.44	17.71	20.32	16.44	60.83	17.16	18.85	92.72	15.69
	MASE	3.20	**1.63**	2.84	2.00	8.76	2.47	2.47	1.99	8.39	2.09	2.53	12.38	2.12
	RMSE	26.16	21.20	28.49	19.05	65.33	21.69	23.09	19.42	63.79	20.20	23.40	107.91	18.87
	SMAPE	53.20	41.20	79.40	55.30	99.60	43.00	81.20	60.20	94.60	56.70	46.10	133.90	**41.10**
AUG	MAE	20.00	16.06	25.41	19.67	60.41	18.38	20.52	25.65	58.39	14.87	14.51	35.37	**11.85**
	MASE	3.25	2.30	3.87	2.87	10.15	2.98	3.06	3.84	9.78	2.38	2.38	5.40	**1.84**
	RMSE	22.20	19.01	30.73	22.42	61.48	22.64	24.71	28.93	59.55	18.41	17.73	37.24	**15.00**
	SMAPE	44.00	45.90	88.60	61.10	90.70	38.40	67.40	94.20	89.20	36.60	33.90	160.80	**31.00**
SEP	MAE	22.48	20.76	30.91	24.06	53.97	19.66	22.97	22.30	52.60	28.39	17.33	35.54	**15.06**
	MASE	2.46	2.15	3.26	2.50	6.01	2.12	2.41	2.34	5.84	2.99	1.88	3.74	**1.64**
	RMSE	26.80	25.72	37.13	29.06	57.04	24.10	30.48	29.16	56.34	33.63	21.73	40.25	**18.90**
	SMAPE	49.90	57.10	94.50	70.30	84.10	44.80	67.70	63.90	82.50	91.50	42.30	139.20	**38.60**
OCT	MAE	35.69	75.37	76.73	79.84	35.38	41.20	67.20	65.33	**33.87**	84.67	36.51	68.69	36.37
	MASE	1.93	4.19	4.33	4.45	1.98	2.25	3.72	3.63	**1.89**	4.77	2.04	3.89	1.99
	RMSE	47.38	85.49	90.79	89.45	42.93	52.15	83.63	81.76	**42.76**	94.30	43.08	76.29	44.76
	SMAPE	35.90	111.30	116.80	124.00	35.70	44.90	90.90	86.90	**34.00**	145.70	36.70	62.10	36.70
NOV	MAE	102.46	216.26	177.96	219.77	147.62	119.58	141.29	143.39	136.72	218.78	95.48	239.19	**77.08**
	MASE	1.77	3.74	3.09	3.80	2.52	2.06	2.42	2.46	2.32	3.77	1.65	4.15	**1.33**
	RMSE	119.19	228.49	197.04	231.84	164.44	137.38	157.71	159.42	153.82	230.81	111.04	250.58	**90.35**
	SMAPE	49.30	159.30	116.60	164.40	81.10	61.80	81.10	81.80	72.50	168.90	45.40	198.10	**34.60**
DEC	MAE	74.03	176.33	134.41	181.01	106.47	**68.77**	97.15	106.11	96.54	159.13	104.60	194.85	116.77
	MASE	1.97	4.76	3.63	4.89	2.85	**1.84**	2.59	2.86	2.55	4.37	2.82	5.26	3.15
	RMSE	94.04	186.95	163.73	191.37	122.82	**89.40**	123.92	133.80	116.90	170.62	121.14	205.76	132.24
	SMAPE	39.80	151.90	81.40	159.90	66.10	**35.90**	58.20	68.30	56.80	139.00	64.90	183.90	76.60

Table 3: Forecasting performance of the proposed E-STGCN model in comparison to the temporal-only and spatiotemporal forecasting techniques for 30 days ahead forecast horizon of PM_{10} pollutant (best results are **highlighted**).

Forecast Period	Metric	Temporal-only Model						Spatiotemporal Model						Proposed E-STGCN
		ARIMA	LSTM	TCN	DeepAR	Transformers	NBeats	STARMA	GSTAR	GpGp	STNN	STGCN	DeepKrigging	
JAN	MAE	99.31	276.07	262.61	275.41	165.16	87.24	121.40	82.99	86.54	115.31	89.89	233.97	**82.67**
	MASE	1.23	3.46	3.31	3.45	2.05	1.10	1.49	**1.02**	1.09	1.47	1.12	2.94	1.03
	RMSE	126.85	294.14	307.81	293.53	192.87	112.07	147.81	110.33	109.86	145.59	119.16	273.87	**108.41**
	SMAPE	35.90	175.50	124.70	174.60	70.50	30.80	47.10	29.70	30.30	37.00	32.00	120.60	**29.10**
FEB	MAE	59.32	218.02	212.75	216.10	102.77	68.89	157.65	153.76	106.83	77.69	70.35	201.06	**56.49**
	MASE	1.25	4.72	4.58	4.68	2.14	1.50	3.38	3.32	2.22	1.75	1.48	4.41	**1.21**
	RMSE	77.99	228.33	247.69	226.50	122.29	83.55	175.50	171.40	125.72	101.26	88.41	223.28	**71.36**
	SMAPE	25.40	172.10	116.00	169.00	49.80	29.50	96.70	93.30	52.60	32.20	32.20	140.50	**24.30**
MAR	MAE	51.41	159.77	168.58	154.64	52.33	52.00	54.18	51.87	60.80	252.19	56.17	167.16	**42.54**
	MASE	1.29	3.93	4.20	3.80	1.23	1.31	1.30	1.26	1.61	6.54	1.39	4.12	**1.05**
	RMSE	62.16	166.91	203.62	162.02	63.41	64.42	64.10	61.11	71.00	309.41	70.32	174.16	**51.96**
	SMAPE	29.60	169.10	115.00	158.10	31.30	29.50	33.40	31.60	34.40	78.20	36.20	185.20	**25.30**
APR	MAE	59.82	181.01	171.35	179.36	71.49	67.72	151.14	106.86	61.34	62.35	**53.04**	229.48	53.04
	MASE	1.37	4.18	3.89	4.14	1.61	1.54	3.46	2.45	1.42	1.45	**1.24**	5.38	1.24
	RMSE	75.86	191.41	191.06	189.86	88.30	85.42	165.07	124.18	74.55	77.04	**62.40**	311.57	62.40
	SMAPE	30.90	164.50	130.50	161.40	37.90	36.10	117.10	68.50	31.90	32.70	**27.70**	143.40	27.70
MAY	MAE	**71.01**	176.32	167.71	170.78	79.39	71.75	117.99	88.82	78.77	94.51	77.46	188.01	77.46
	MASE	**1.12**	2.81	2.67	2.72	1.24	1.15	1.86	1.41	1.23	1.53	1.24	3.01	1.24
	RMSE	**91.48**	199.93	205.51	195.07	107.15	92.81	147.03	118.57	105.91	117.57	96.88	211.09	96.88
	SMAPE	**38.00**	167.30	119.70	155.40	42.90	38.60	77.90	49.30	42.50	47.90	41.40	194.50	41.40
JUN	MAE	67.66	117.99	123.54	114.96	47.59	**43.32**	87.20	74.67	69.23	43.44	53.60	130.25	51.65
	MASE	2.00	3.36	3.53	3.27	1.42	1.27	2.46	2.11	2.09	**1.25**	1.58	3.78	1.51
	RMSE	81.88	129.68	148.78	126.93	58.79	56.31	100.34	89.19	82.67	**55.60**	65.15	154.60	65.44
	SMAPE	45.10	150.10	124.00	142.90	36.00	33.40	90.00	70.90	47.00	**33.00**	39.00	137.00	37.90
JUL	MAE	101.08	60.00	72.71	57.24	72.83	**30.07**	38.87	31.80	96.52	60.67	48.40	76.07	48.40
	MASE	5.72	3.21	3.77	3.05	4.37	**1.65**	2.01	1.66	5.83	3.52	2.74	4.11	2.74
	RMSE	108.53	65.60	88.01	63.07	77.47	**36.78**	46.32	38.07	103.92	75.32	55.25	80.74	55.25
	SMAPE	81.10	120.10	117.10	110.20	69.50	**36.90**	62.10	46.70	79.90	56.40	51.70	186.90	51.70
AUG	MAE	60.84	108.73	112.74	105.64	48.29	47.00	71.27	84.91	60.04	87.94	47.44	94.30	**39.70**
	MASE	2.86	4.88	5.02	4.73	2.23	2.03	3.16	3.80	2.88	3.98	2.19	4.25	**1.80**
	RMSE	69.72	116.57	124.58	113.71	56.02	59.81	87.31	97.19	69.23	97.95	55.36	114.38	**47.97**
	SMAPE	42.60	146.70	141.80	138.40	37.20	39.20	71.80	94.80	43.10	104.10	36.00	93.00	**31.00**
SEP	MAE	82.97	92.66	104.00	91.22	52.34	68.16	48.30	47.63	66.26	79.02	44.58	119.67	**35.90**
	MASE	4.60	5.68	6.07	5.59	3.08	3.89	2.95	2.90	3.91	4.93	2.54	7.04	**2.05**
	RMSE	94.06	102.86	122.03	101.53	63.62	81.38	60.46	58.86	78.06	90.40	55.68	166.10	**44.51**
	SMAPE	60.70	130.70	122.30	126.70	46.50	53.40	48.50	46.60	53.70	96.90	41.50	98.90	**35.70**
OCT	MAE	56.62	214.64	202.97	206.32	81.46	94.69	120.44	118.95	67.79	178.89	62.18	218.93	**56.42**
	MASE	**1.54**	6.00	5.57	5.75	2.17	2.61	3.27	3.29	1.80	5.03	1.70	6.12	1.55
	RMSE	76.41	225.67	223.43	217.78	101.38	114.30	148.51	147.03	88.08	192.39	80.64	231.48	**74.86**
	SMAPE	**25.00**	176.90	140.50	162.80	37.90	47.60	66.20	65.60	30.20	134.70	27.80	181.60	25.00
NOV	MAE	160.63	370.52	342.23	365.48	230.28	114.14	164.50	163.94	**101.36**	248.39	105.00	378.61	105.00
	MASE	2.00	4.64	4.27	4.57	2.86	1.43	2.04	2.04	**1.28**	3.08	1.32	4.77	1.32
	RMSE	181.25	387.20	370.87	382.37	255.63	137.63	183.15	182.67	**123.56**	273.40	125.82	406.14	125.82
	SMAPE	49.20	182.20	140.80	176.90	78.10	31.90	51.80	51.80	**27.60**	88.70	29.40	169.70	29.40
DEC	MAE	111.36	320.05	303.66	311.54	174.14	82.85	**131.11**	148.68	112.51	100.10	117.10	326.28	134.80
	MASE	2.02	5.87	5.58	5.71	3.12	1.54	**2.36**	2.71	2.02	1.79	2.14	6.03	2.50
	RMSE	139.68	329.61	342.88	321.36	191.44	100.84	**163.30**	183.15	139.38	118.07	141.74	347.20	157.71
	SMAPE	36.60	184.50	129.80	174.50	66.20	24.70	**45.00**	55.20	36.10	28.90	39.50	168.30	48.20

Table 4: Forecasting performance of the proposed E-STGCN model in comparison to the temporal-only and spatiotemporal forecasting techniques for 30 days ahead forecast horizon of NO_2 pollutant (best results are **highlighted**).

Forecast Period	Metric	Temporal-only Model						Spatiotemporal Model						Proposed
		ARIMA	LSTM	TCN	DeepAR	Transformers	NBeats	STARMA	GSTAR	GpGp	STNN	STGCN	DeepKrigging	E-STGCN
JAN	MAE	13.76	23.48	38.72	26.73	16.49	15.71	17.87	15.76	16.37	28.46	**12.28**	40.02	**12.28**
	MASE	3.33	8.21	15.68	9.85	11.36	4.58	4.99	**2.48**	4.48	8.15	4.74	13.43	4.74
	RMSE	17.01	26.50	41.37	29.63	19.45	19.34	21.80	19.37	19.92	36.38	**15.01**	48.68	**15.01**
	SMAPE	33.70	61.60	160.40	76.20	39.50	41.50	44.20	38.70	39.70	62.50	**28.30**	109.50	**28.30**
FEB	MAE	14.21	25.80	39.27	27.09	16.76	13.64	20.57	18.42	16.95	33.94	**11.92**	33.84	**11.92**
	MASE	**2.32**	4.85	8.16	5.18	4.47	2.65	3.61	2.71	3.58	7.55	2.50	6.94	2.50
	RMSE	17.06	28.35	41.70	29.61	19.36	16.64	23.51	21.46	20.09	47.18	**14.56**	36.77	**14.56**
	SMAPE	34.60	70.80	158.20	76.70	40.20	37.00	56.20	47.40	41.40	67.40	**28.50**	123.20	**28.50**
MAR	MAE	12.39	17.26	31.02	18.48	16.78	12.68	**8.64**	9.94	14.91	21.44	8.96	30.13	8.96
	MASE	4.65	6.43	12.07	6.82	9.83	4.20	3.16	**2.63**	5.53	8.56	3.53	11.56	3.53
	RMSE	14.36	19.22	32.85	20.42	18.54	15.22	**10.60**	11.92	17.07	27.80	11.23	33.18	11.23
	SMAPE	38.90	53.90	157.50	59.50	48.30	39.00	**27.20**	31.40	46.50	58.50	27.40	144.90	27.40
APR	MAE	15.70	22.31	36.47	23.05	19.74	16.72	18.32	18.19	17.41	25.79	**13.63**	37.75	**13.63**
	MASE	3.57	3.15	6.41	3.24	5.02	2.78	2.56	2.94	4.37	6.08	**2.28**	6.95	**2.28**
	RMSE	19.05	25.77	39.24	26.49	22.29	20.62	22.58	22.24	21.18	33.53	**16.22**	40.30	**16.22**
	SMAPE	40.50	60.90	168.20	64.30	49.60	47.70	49.20	50.00	45.20	64.20	**34.40**	196.80	**34.40**
MAY	MAE	14.82	19.30	32.75	19.71	20.06	18.59	14.77	13.94	16.85	27.00	**11.33**	23.67	**11.33**
	MASE	3.37	2.59	5.80	2.59	6.46	6.02	**2.18**	2.62	4.64	7.54	2.20	8.26	2.20
	RMSE	18.05	22.61	35.85	23.03	22.99	22.09	18.47	17.38	20.32	35.38	**14.35**	28.05	**14.35**
	SMAPE	41.70	56.70	155.60	58.60	52.60	55.40	41.30	40.40	47.50	66.00	**32.40**	65.40	**32.40**
JUN	MAE	13.79	12.89	23.74	13.38	24.21	11.55	**7.43**	10.03	13.84	25.04	8.05	26.23	8.05
	MASE	5.61	3.72	9.24	3.77	10.91	5.10	**1.84**	4.18	10.88	14.84	2.71	10.78	2.71
	RMSE	15.80	14.59	25.68	15.09	25.40	13.58	**9.28**	11.79	16.15	33.49	9.88	27.36	9.88
	SMAPE	49.80	48.50	144.50	50.70	70.10	45.60	**29.30**	39.80	51.30	70.60	31.30	197.50	31.30
JUL	MAE	15.21	10.11	19.58	10.19	25.66	8.11	**6.74**	7.92	13.51	23.31	7.56	14.86	7.56
	MASE	6.35	3.40	6.25	3.28	10.64	2.77	**2.15**	2.50	6.09	9.64	2.68	5.08	2.68
	RMSE	16.93	11.74	21.19	11.88	26.97	10.41	**8.48**	9.63	15.33	31.45	9.95	16.94	9.95
	SMAPE	61.60	46.90	153.80	47.50	82.10	47.50	**35.00**	40.10	59.30	74.90	37.20	80.00	37.20
AUG	MAE	17.45	9.59	17.26	9.62	25.21	8.56	7.89	8.40	14.07	19.88	**6.95**	12.12	**6.95**
	MASE	7.14	4.44	6.10	4.22	10.73	3.11	2.44	2.76	7.98	12.51	**2.13**	5.74	**2.13**
	RMSE	19.36	11.41	19.24	11.45	27.10	10.80	9.92	10.40	16.21	28.54	**8.66**	13.93	**8.66**
	SMAPE	71.50	52.90	154.50	53.10	86.80	47.50	40.80	46.60	64.80	75.40	**38.40**	74.90	**38.40**
SEP	MAE	13.27	10.27	20.46	10.32	20.50	9.90	8.50	10.48	12.52	15.40	**6.46**	22.24	**6.46**
	MASE	4.74	3.50	5.65	3.49	7.77	3.24	2.30	2.62	6.35	6.97	**1.87**	6.37	**1.87**
	RMSE	14.91	11.93	22.27	11.98	21.82	11.92	10.62	12.61	14.62	19.91	**8.32**	23.50	**8.32**
	SMAPE	53.50	48.50	148.30	48.80	69.10	49.60	41.00	52.20	53.30	66.50	**32.10**	195.10	**32.10**
OCT	MAE	13.92	19.79	34.59	20.17	18.54	16.77	16.12	16.93	15.79	24.02	**12.09**	32.98	**12.09**
	MASE	2.55	3.05	6.12	3.09	3.81	2.77	2.48	2.61	3.46	4.22	**1.91**	6.01	**1.91**
	RMSE	19.20	24.99	38.91	25.41	23.56	22.57	22.37	23.21	21.30	30.25	**17.30**	37.23	**17.30**
	SMAPE	35.40	55.90	165.90	57.50	46.20	48.50	44.40	47.60	41.90	79.90	**30.30**	136.30	**30.30**
NOV	MAE	19.56	31.39	42.55	31.40	25.12	22.51	18.01	20.86	20.54	34.00	**16.62**	46.43	**16.62**
	MASE	3.11	4.04	5.61	4.03	4.83	3.39	3.02	3.42	4.35	5.18	**2.19**	5.95	**2.19**
	RMSE	24.01	35.90	47.07	35.92	29.31	27.62	22.48	25.60	24.75	42.28	**21.10**	49.98	**21.10**
	SMAPE	43.80	77.40	140.20	77.40	51.40	54.70	42.20	46.30	47.60	84.10	**36.20**	190.20	**36.20**
DEC	MAE	**17.57**	30.04	41.09	30.01	23.40	20.50	18.74	24.55	18.18	31.54	21.44	47.41	21.44
	MASE	2.62	4.24	6.53	4.24	3.30	3.91	**2.56**	3.42	2.99	5.07	3.30	8.38	3.30
	RMSE	**21.30**	33.03	45.11	33.00	26.49	24.97	22.83	28.63	22.00	38.10	24.77	51.95	24.77
	SMAPE	**37.40**	74.20	134.40	74.10	50.80	48.60	40.50	59.90	39.00	76.00	47.90	160.60	47.90

4.5. Statistical Tests for Model Robustness

To validate the robustness of our experimental results, we employ multiple comparison with the best (MCB) test (Koning et al., 2005) and the Diebold-Mariano test (Diebold and Mariano, 2002). The MCB test aims to identify the 'best' forecasting model among all $\mathcal{F}$ architectures based on their performance across $\mathcal{D}$ datasets. For a specific evaluation metric, this non-parametric procedure ranks all models based on their performance across different forecasting tasks and computes the mean rank. The model with the lowest mean rank is considered the 'best' forecasting architecture. Next, the critical distance (CD) for each of the $\mathcal{F}$ models is computed as $\delta_\theta \sqrt{\mathcal{F}(\mathcal{F}+1)/6\mathcal{D}}$, where δ_θ represents the critical value of the Tukey distribution at significance level θ. The CD of the 'best' performing model serves as the reference value against which all other models are compared. We apply the MCB test and visualize the results based on the RMSE metric for $PM_{2.5}$, PM_{10}, and NO_2 in Fig. 7. From these plots, we observe that the proposed E-STGCN architecture achieves the 'best' performance, with a minimum rank of 3.14 for $PM_{2.5}$ and 2.20 for PM_{10} datasets. For NO_2 forecasting, the performance of E-STGCN and STGCN are similar and they jointly obtain the lowest rank of 2.36. Among the baseline models, the STGCN, NBeats, and ARIMA frameworks consistently showcase better performance, having competitive ranks with E-STGCN. Spatiotemporal models such as GSTAR, STARMA, and GpGp outperform the majority of the time-dependent frameworks by effectively capturing space-time correlations. Moreover, the CD values for most of the baseline models lie above the reference value (shaded region), indicating that their performance is significantly worse than the 'best-fitted' E-STGCN model. Additional MCB test results based on the other evaluation metrics are provided in Fig. S.1, in Section S.3 of the supplement. Overall, the MCB test results emphasize that the proposed E-STGCN approach consistently delivers accurate forecasts for various air pollutants, as measured by all the performance indicators.

Next, we employ the Diebold-Mariano (DM) test to assess whether the forecasting performance of the proposed E-STGCN framework significantly differs from that of the baseline models. Specifically, for any

Table 5: Forecasting performance of the proposed E-STGCN model in comparison to the temporal-only and spatiotemporal forecasting techniques for 60 days ahead forecast horizon of different pollutants (best results are **highlighted**).

Pollutant	Forecast Period	Metric	Temporal-only Model						Spatiotemporal Model						Proposed E-STGCN
			ARIMA	LSTM	TCN	DeepAR	Transformers	NBeats	STARMA	GSTAR	GpGp	STNN	STGCN	DeepKrigging	
PM$_{2.5}$	JAN - FEB	MAE	73.48	132.20	116.87	131.90	66.52	59.28	80.56	100.35	58.53	76.26	56.73	84.47	**55.00**
		MASE	1.72	3.11	2.73	3.10	1.54	1.39	1.87	2.36	1.36	1.82	1.32	1.99	**1.28**
		RMSE	86.23	148.60	135.33	148.33	89.53	77.12	96.46	113.61	78.81	94.98	76.44	100.93	**73.00**
		SMAPE	46.50	152.00	126.90	151.30	47.40	45.20	81.50	57.00	39.90	48.80	39.60	77.20	**37.90**
	MAR - APR	MAE	35.03	59.96	49.14	55.51	29.62	35.06	50.95	50.86	37.44	143.20	27.68	68.10	**22.35**
		MASE	1.82	3.02	2.46	2.78	1.58	1.84	2.56	2.55	2.07	7.71	1.39	3.47	**1.13**
		RMSE	40.90	65.06	58.64	60.99	34.89	41.08	57.72	58.23	42.69	154.31	35.06	77.92	**27.69**
		SMAPE	43.70	131.80	93.70	114.10	39.20	42.80	111.50	112.70	46.50	98.80	41.70	144.30	**31.20**
	MAY - JUN	MAE	37.54	40.38	42.47	38.31	46.70	22.53	43.36	**22.15**	49.65	43.12	30.41	37.31	43.53
		MASE	2.14	2.23	2.35	2.11	2.71	1.24	2.41	**1.23**	2.95	2.33	1.70	2.03	2.44
		RMSE	42.70	49.77	51.29	48.09	51.41	32.55	50.27	**30.71**	54.99	52.49	42.70	44.79	56.56
		SMAPE	57.70	97.10	117.20	88.10	66.60	39.10	127.60	**38.70**	68.60	70.40	58.00	99.80	94.50
	JUL - AUG	MAE	30.37	21.33	26.49	18.44	60.75	24.10	29.51	**12.67**	62.11	20.80	29.41	36.68	28.05
		MASE	4.07	2.68	3.38	2.30	8.17	3.25	3.77	**1.60**	8.35	2.60	3.89	4.71	3.70
		RMSE	33.44	24.35	30.50	21.69	61.90	28.32	32.59	**16.06**	64.19	24.31	35.63	38.78	34.11
		SMAPE	60.90	74.00	109.00	59.70	93.40	51.00	132.80	**38.10**	93.60	61.90	56.70	197.50	55.40
	SEP - OCT	MAE	36.12	54.88	63.93	52.05	44.53	40.25	58.62	34.86	44.33	53.65	33.99	57.96	**33.60**
		MASE	2.63	3.95	4.64	3.74	3.32	2.91	4.25	2.54	3.34	3.90	2.50	4.15	**2.46**
		RMSE	51.27	69.38	78.49	66.93	50.97	55.78	76.23	49.13	51.34	68.24	41.05	71.90	**40.34**
		SMAPE	55.00	107.90	148.20	97.40	59.80	64.00	120.70	52.40	59.40	108.80	49.60	123.00	**48.90**
	NOV - DEC	MAE	**90.77**	204.62	195.80	200.48	127.97	132.34	159.92	157.59	120.80	198.17	95.34	224.42	91.09
		MASE	**1.91**	4.30	4.11	4.21	2.67	2.78	3.35	3.31	2.50	4.18	2.00	4.73	1.91
		RMSE	113.64	217.05	210.92	213.17	146.81	149.75	175.57	173.31	139.69	211.16	115.40	244.57	112.98
		SMAPE	**46.50**	169.40	155.70	162.30	74.50	80.80	117.80	113.50	68.20	163.30	49.60	189.00	46.30
PM$_{10}$	JAN - FEB	MAE	94.47	252.19	245.06	245.01	135.51	95.19	120.61	116.70	101.61	102.18	80.35	125.87	**72.53**
		MASE	1.53	3.99	3.87	3.88	2.11	1.53	1.88	1.83	1.58	1.63	1.27	1.97	**1.15**
		RMSE	113.55	269.06	280.32	262.36	163.79	116.76	147.89	143.09	130.47	127.15	106.36	152.58	**94.10**
		SMAPE	35.10	183.00	148.20	171.80	61.20	35.40	61.00	56.50	40.70	40.00	31.40	58.10	**28.20**
	MAR - APR	MAE	61.48	168.41	174.61	164.51	62.63	61.36	103.94	95.72	57.14	165.95	51.94	181.87	**49.59**
		MASE	1.43	3.87	4.02	3.78	1.39	1.42	2.35	2.18	1.32	4.06	1.20	4.19	**1.14**
		RMSE	73.69	178.80	203.57	175.13	78.19	75.55	125.97	117.40	70.86	212.95	63.17	191.54	**60.07**
		SMAPE	33.40	167.10	135.50	159.10	35.80	33.80	76.30	67.70	32.20	60.90	29.20	198.90	**28.10**
	MAY - JUN	MAE	71.99	145.59	145.40	143.09	83.16	67.01	129.13	111.88	67.53	87.48	76.14	77.28	**66.47**
		MASE	1.48	2.95	2.92	2.90	1.65	**1.35**	2.61	2.27	1.40	1.83	1.55	1.53	1.36
		RMSE	89.34	167.01	173.28	164.85	113.32	95.91	150.33	134.00	**87.32**	107.85	100.07	107.10	89.67
		SMAPE	43.60	155.20	133.70	149.30	56.00	42.90	128.80	100.80	41.70	48.30	47.80	50.50	**41.10**
	JUL - AUG	MAE	58.10	87.62	95.24	82.17	59.70	**42.82**	76.47	77.41	70.26	58.97	49.79	72.79	49.79
		MASE	2.90	4.18	4.53	3.90	3.05	**2.04**	3.62	3.69	3.65	2.79	2.46	3.43	2.46
		RMSE	65.40	97.74	109.72	92.87	67.62	**54.78**	91.86	93.15	78.60	72.70	59.03	84.92	59.03
		SMAPE	50.40	142.70	142.60	125.20	52.50	**40.70**	110.10	112.70	58.20	64.00	43.70	105.10	43.70
	SEP - OCT	MAE	64.76	153.86	160.36	148.98	67.84	77.56	119.55	118.63	70.16	132.91	69.95	164.33	**63.70**
		MASE	2.38	5.74	5.96	5.55	2.50	2.88	4.44	4.42	2.62	4.95	2.59	6.14	**2.36**
		RMSE	81.32	174.76	184.58	170.42	87.21	97.58	151.29	150.95	88.90	155.66	84.98	185.21	**81.31**
		SMAPE	40.50	156.20	149.10	145.10	42.50	47.90	93.70	92.00	43.80	119.00	43.20	175.30	**40.00**
	NOV - DEC	MAE	**114.94**	342.22	328.89	338.54	202.89	156.01	204.62	188.63	186.24	175.93	162.09	350.87	162.09
		MASE	**1.71**	5.07	4.85	5.01	2.97	2.30	3.00	2.79	2.71	2.58	2.40	5.20	2.40
		RMSE	**139.63**	356.98	351.94	353.46	226.80	177.83	226.54	210.06	208.70	208.93	187.66	366.66	187.66
		SMAPE	**35.00**	179.80	157.70	175.80	72.60	52.30	81.90	72.00	64.90	61.20	53.80	187.60	53.80
NO$_2$	JAN - FEB	MAE	15.04	29.56	39.64	26.76	17.21	18.62	18.45	18.37	18.52	35.17	**13.50**	42.27	13.50
		MASE	2.85	6.53	9.23	5.79	5.78	4.21	3.85	**2.60**	3.95	7.72	4.69	8.48	4.69
		RMSE	18.17	32.57	42.33	29.93	20.18	22.76	22.59	22.23	21.72	47.45	**16.28**	52.00	16.28
		SMAPE	37.30	91.50	172.00	76.70	41.50	51.50	46.50	47.90	45.90	67.30	**31.90**	108.70	31.90
	MAR - APR	MAE	15.60	21.86	33.86	20.66	17.82	14.20	14.39	15.75	17.59	26.32	**12.04**	27.12	12.04
		MASE	4.38	3.44	6.65	3.25	5.23	3.61	**2.61**	3.03	5.67	6.98	2.74	5.56	2.74
		RMSE	18.92	25.14	36.57	24.02	20.56	18.02	18.53	20.26	21.19	34.52	**15.32**	31.41	15.32
		SMAPE	43.60	68.30	170.60	62.10	48.20	40.80	44.50	47.80	50.30	63.10	**33.80**	103.60	33.80
	MAY - JUN	MAE	16.25	18.31	28.86	16.63	19.45	21.93	15.58	14.83	17.27	22.31	**12.43**	29.55	12.43
		MASE	5.05	3.09	6.64	2.92	7.01	7.33	**2.72**	3.33	6.70	6.69	3.66	6.44	3.66
		RMSE	19.17	21.37	31.48	19.76	21.69	25.46	18.72	17.95	20.48	29.14	**15.34**	32.01	15.34
		SMAPE	50.80	64.00	163.30	55.10	56.20	62.10	54.40	53.40	55.00	65.30	**40.80**	143.80	40.80
	JUL - AUG	MAE	18.64	9.98	18.39	9.90	25.17	**8.69**	9.28	10.07	16.33	24.76	12.05	18.77	12.05
		MASE	7.28	3.48	5.83	3.40	9.63	**2.77**	2.80	2.96	7.62	10.54	4.11	5.96	4.11
		RMSE	20.75	12.36	20.89	12.27	26.64	**11.41**	11.63	12.44	19.01	35.55	14.79	21.13	14.79
		SMAPE	72.20	50.70	154.10	50.30	84.30	**49.00**	52.00	57.70	67.60	78.50	53.30	162.50	53.30
	SEP - OCT	MAE	15.03	15.67	27.75	15.31	19.61	14.92	17.39	19.72	15.21	18.80	**12.42**	19.78	12.42
		MASE	3.82	3.19	5.89	3.11	5.13	3.18	3.42	3.91	4.43	4.66	**2.72**	4.81	2.72
		RMSE	19.63	20.84	32.21	20.47	23.89	20.35	23.63	25.64	19.91	25.21	**16.95**	24.92	16.95
		SMAPE	48.70	55.20	162.60	53.50	57.90	56.50	68.10	84.80	51.10	68.10	**39.90**	84.90	39.90
	NOV - DEC	MAE	20.08	30.52	43.84	30.71	23.70	21.87	21.51	24.54	20.97	41.25	**19.76**	36.24	19.76
		MASE	**2.68**	3.81	5.99	3.83	3.37	2.93	2.82	3.23	3.10	6.36	2.72	5.28	2.72
		RMSE	24.78	35.49	48.37	35.66	28.20	27.05	26.81	29.82	25.37	62.27	**24.71**	40.88	24.71
		SMAPE	43.80	75.00	156.00	75.80	48.50	51.40	49.80	57.90	47.00	77.60	**43.30**	81.60	43.30

Figure 7: MCB plots for (A) PM$_{2.5}$, (B) PM$_{10}$, and (C) NO$_2$ pollutant concentration levels based on RMSE metric. In the figure, for example, 'E-STGCN - 3.14' means that the average rank of the proposed E-STGCN algorithm based on the RMSE error metric is 3.14 for PM$_{2.5}$ dataset; the same explanation applies to other algorithms and datasets. The shaded region depicts the reference value of the test.

baseline architecture $\mathcal{A}$ and the proposed E-STGCN model, we compute the multivariate loss differential series for a given station as:

$$\Lambda^i_{t,\mathcal{A}} = \left| X^i_t - \widehat{X}^i_{t,\mathcal{A}} \right| - \left| X^i_t - \widehat{X}^i_{t,\text{E-STGCN}} \right|,$$

where X^i_t represents the ground truth data for station i at time t with $\widehat{X}^i_{t,\text{E-STGCN}}$ and $\widehat{X}^i_{t,\mathcal{A}}$ being the
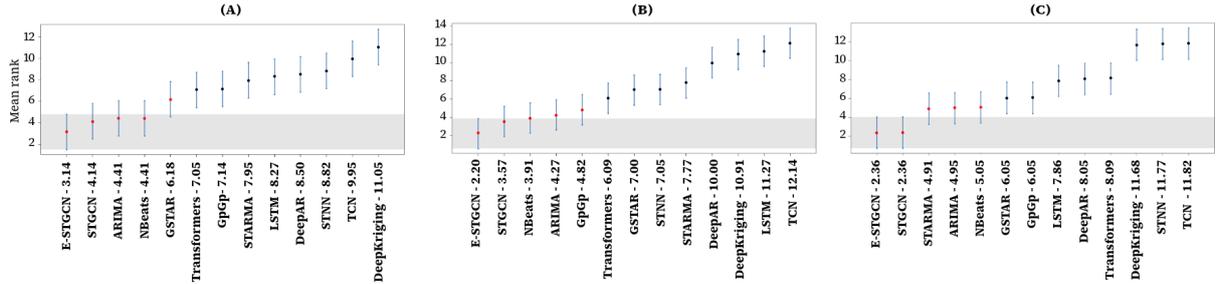
Table 6: Forecasting performance of the proposed E-STGCN model in comparison to the temporal-only and spatiotemporal forecasting techniques for 90 days ahead forecast horizon of different pollutants (best results are **highlighted**).

Pollutant	Forecast Period	Metric	Temporal-only Model						Spatiotemporal Model						Proposed E-STGCN
			ARIMA	LSTM	TCN	DeepAR	Transformers	NBeats	STARMA	GSTAR	GpGp	STNN	STGCN	DeepKrigging	
PM$_{2.5}$	JAN - FEB - MAR	MAE	87.20	108.92	101.25	108.36	52.24	50.23	77.69	80.22	50.66	91.36	47.42	65.84	**46.00**
		MASE	2.47	3.10	2.87	3.09	1.48	1.43	2.20	2.28	1.45	2.66	1.35	1.94	**1.31**
		RMSE	97.95	127.48	118.65	127.00	75.09	66.30	90.16	92.99	68.84	110.19	64.49	76.98	**62.59**
		SMAPE	59.70	143.60	137.90	142.00	41.70	47.90	110.70	110.30	40.20	67.80	37.90	50.20	**36.60**
	APR - MAY - JUN	MAE	44.46	47.61	59.85	43.16	40.92	**25.30**	52.92	47.06	47.07	79.39	40.23	58.46	36.04
		MASE	2.55	2.63	3.36	2.38	2.39	**1.42**	2.95	2.62	2.81	4.72	2.28	3.32	2.05
		RMSE	50.55	55.62	68.24	51.84	45.93	**30.93**	59.96	54.56	52.37	85.98	47.66	70.62	42.94
		SMAPE	59.90	116.20	171.10	96.50	58.00	**40.80**	157.20	133.60	63.50	83.50	58.20	131.70	52.10
	JUL - AUG - SEP	MAE	37.00	20.14	28.17	**19.75**	61.16	41.07	31.89	29.92	61.83	20.28	49.34	43.87	45.66
		MASE	4.64	2.36	3.35	**2.31**	7.69	5.17	3.80	3.56	7.76	2.41	6.18	5.32	5.73
		RMSE	40.83	24.13	32.83	**23.78**	62.78	47.51	35.94	34.54	64.02	24.56	60.88	53.88	57.29
		SMAPE	68.30	64.30	112.40	62.50	93.20	69.80	146.50	132.70	93.10	**61.30**	74.20	135.20	71.50
	OCT - NOV - DEC	MAE	99.14	158.91	176.00	157.95	98.80	136.28	163.98	163.83	96.85	162.02	97.34	177.69	**95.10**
		MASE	2.66	4.29	4.76	4.26	2.65	3.68	4.43	4.43	2.58	4.37	2.62	4.81	**2.56**
		RMSE	126.15	180.61	196.24	179.76	125.13	161.44	190.17	189.36	123.31	183.52	125.07	197.34	**122.44**
		SMAPE	64.30	151.40	187.30	149.20	63.90	110.10	157.20	157.30	61.80	162.00	64.30	199.50	**61.00**
PM$_{10}$	JAN - FEB - MAR	MAE	111.07	219.16	219.77	215.14	109.07	110.28	128.22	118.87	84.76	135.20	68.88	207.63	**68.81**
		MASE	2.04	3.94	3.94	3.86	1.92	2.02	2.28	2.12	1.50	2.51	1.24	3.73	**1.24**
		RMSE	128.20	237.79	251.17	234.10	139.64	129.64	149.21	139.84	113.21	176.45	92.72	235.06	**92.55**
		SMAPE	44.00	174.70	161.10	167.80	52.20	43.80	86.60	73.20	37.20	48.00	29.90	151.50	**29.90**
	APR - MAY - JUN	MAE	69.36	162.31	159.67	155.56	**65.89**	69.75	150.98	123.69	68.75	134.97	71.32	143.21	71.32
		MASE	1.49	3.45	3.37	3.29	**1.39**	1.47	3.20	2.62	1.48	2.97	1.54	3.08	1.54
		RMSE	85.71	180.11	181.27	174.06	88.02	95.17	169.22	144.79	86.74	161.53	**85.70**	163.18	85.70
		SMAPE	40.00	169.50	158.80	154.00	**38.70**	42.40	152.90	112.20	42.50	64.80	40.80	94.30	40.80
	JUL - AUG - SEP	MAE	67.13	89.83	98.76	84.03	57.10	**44.69**	84.31	86.80	70.51	52.56	78.42	71.77	59.51
		MASE	3.48	4.48	4.92	4.18	3.03	**2.21**	4.20	4.35	3.82	2.67	4.04	3.54	3.04
		RMSE	76.53	100.04	111.97	94.82	66.28	**56.63**	98.63	103.53	80.25	65.99	100.54	84.83	76.20
		SMAPE	54.90	144.20	156.50	125.70	50.80	**45.60**	130.20	130.80	57.80	57.30	54.10	104.50	53.30
	OCT - NOV - DEC	MAE	181.14	296.82	300.74	292.10	162.13	246.24	259.03	263.90	155.14	268.84	145.11	284.10	**135.20**
		MASE	3.24	5.32	5.38	5.23	2.86	4.41	4.63	4.73	2.72	4.83	2.60	5.09	**2.41**
		RMSE	212.44	317.30	325.67	312.89	194.85	274.96	293.93	297.78	188.18	291.76	179.30	309.89	**169.99**
		SMAPE	74.10	177.60	175.10	171.40	61.40	122.30	135.20	139.80	57.40	151.00	54.50	164.10	**49.30**
NO$_2$	JAN - FEB - MAR	MAE	15.26	25.05	37.10	24.14	17.52	18.14	19.04	18.62	18.74	30.57	**13.74**	22.91	13.74
		MASE	3.59	6.34	9.93	6.11	6.33	4.74	4.85	**3.07**	4.64	7.52	5.73	7.03	5.73
		RMSE	18.22	28.35	39.76	27.45	20.35	21.85	22.75	22.10	21.95	40.92	**16.56**	26.15	16.56
		SMAPE	40.40	76.40	174.30	71.60	44.80	50.60	54.30	55.00	49.50	69.60	**35.00**	72.90	35.00
	APR - MAY - JUN	MAE	17.56	20.04	31.76	18.86	20.23	**14.51**	19.61	19.57	19.08	23.83	15.58	22.12	15.58
		MASE	4.91	3.17	6.53	3.05	5.85	**2.98**	3.24	3.73	6.69	6.49	3.72	4.02	3.72
		RMSE	20.81	24.36	35.15	23.23	23.11	18.73	23.74	23.30	22.80	31.31	**18.60**	26.23	18.60
		SMAPE	50.20	64.60	173.50	58.50	55.30	47.80	69.50	69.10	56.20	65.80	**44.70**	81.10	44.70
	JUL - AUG - SEP	MAE	19.29	10.10	19.57	**10.00**	25.88	10.09	10.84	11.91	18.41	19.59	15.59	23.39	15.59
		MASE	7.18	3.38	6.06	3.38	9.38	**3.14**	3.16	3.45	8.80	8.51	5.42	8.10	5.42
		RMSE	21.28	**12.74**	22.09	14.01	27.37	13.01	13.71	14.90	20.72	26.72	18.40	32.27	18.40
		SMAPE	72.40	50.50	167.30	**50.00**	83.80	58.90	64.20	71.60	72.30	72.20	60.50	124.50	60.50
	OCT - NOV - DEC	MAE	**19.43**	28.26	42.40	27.22	22.33	25.21	28.42	32.14	20.41	35.98	21.08	43.38	21.08
		MASE	**2.79**	3.86	6.44	3.69	3.54	3.56	4.02	4.63	3.09	5.42	2.97	6.63	2.97
		RMSE	**24.92**	34.10	47.19	33.11	27.58	31.41	35.32	38.73	25.60	41.82	26.91	48.01	26.91
		SMAPE	**43.50**	74.60	179.40	70.00	48.20	68.00	80.00	101.20	47.10	120.90	48.00	196.40	48.00

corresponding forecasts generated by the E-STGCN and model $\mathcal{A}$, respectively. This statistical testing procedure checks whether the expected loss differential is zero using the DM statistic as

$$\text{DM statistic for station } i = \sqrt{q}\frac{\mu_\Lambda^i}{\alpha_\Lambda^i},$$

where q is the forecast horizon, μ_Λ^i and α_Λ^i are respectively the sample mean and standard deviation of the loss differential series $\Lambda_{t,\mathcal{A}}^i$. Using this statistic, we test the null hypothesis $H_0 : \mathbb{E}(\Lambda_{t,\mathcal{A}}^i) \leqslant 0$ against the alternative $H_1 : \mathbb{E}(\Lambda_{t,\mathcal{A}}^i) > 0$, where $\mathbb{E}(\cdot)$ denotes expectation. If the p-value of the test is less than the significance level, we reject H_0 and conclude that the forecasting performance of the E-STGCN framework is superior to that of $\mathcal{A}$ architecture. In our analysis, we conduct the DM test to assess the statistical significance of the performance differences between E-STGCN and the second and third-best-performing baselines, STGCN and NBeats. Figs. 8 and 9 present the test results for forecasting PM$_{2.5}$ and PM$_{10}$ concentrations during the October-November-December period, respectively. These plots evaluate the station-wise forecasting performance of E-STGCN with the benchmarks, where the x-axis represents station indices and the y-axis indicates DM test statistics. A positive DM test statistic value indicates the superiority of the E-STGCN over the baselines, while a negative value suggests that the baselines perform better.

As highlighted in the plots, the E-STGCN method performs similarly to or better than the baselines across most stations, except for PM$_{2.5}$ forecasting of CRRI Mathura Road (station no. 7), where STGCN achieves superior results. Moreover, the significant p-values at 1%, 5%, 10%, and 20% levels are marked using orange, green, blue, and violet-colored stars, respectively. As evident from Fig. 8, E-STGCN significantly outperforms NBeats for PM$_{2.5}$ forecasting in 19 out of 37 stations at 1% significance level. Compared to STGCN, E-STGCN demonstrates significantly different performance for multiple stations at varied levels. For the PM$_{10}$ forecasting, we observe from Fig. 9 that E-STGCN achieves significantly better results than both the STGCN and NBeats for several monitoring stations at 1% and 5% levels of significance. The overall findings of the DM test are consistent with the MAE metrics reported in the experimental evaluations. Hence, this test underscores the statistical significance of our findings. For NO$_2$ concentration levels, the forecasts from the E-STGCN and the STGCN models are very similar due to the absence of many significant extreme observations, resulting in $\Lambda_{t,\text{STGCN}}^i \approx 0$, rendering the DM statistic undefined in this case.
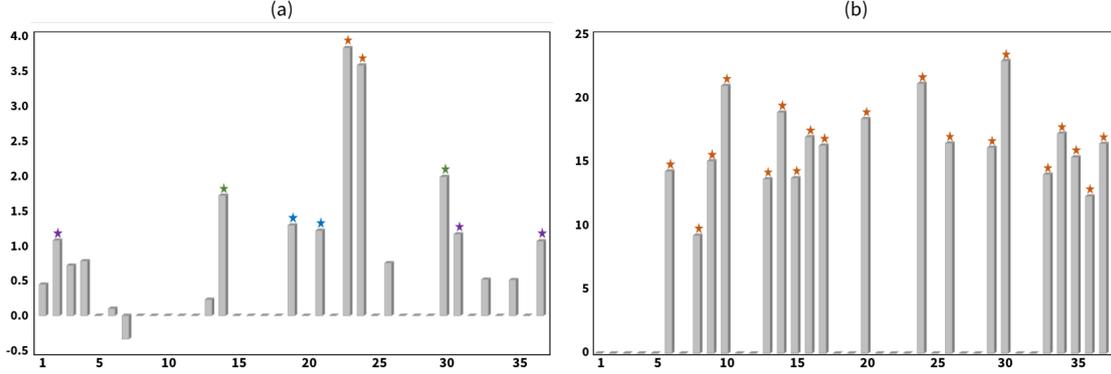
Figure 8: DM test results comparing (a) E-STGCN and STGCN, and (b) E-STGCN and NBeats for forecasting $PM_{2.5}$ pollutant concentrations over the 90-day OCT-NOV-DEC forecast window. The Y-axis represents DM-test statistic values based on the MAE metric, while the X-axis indicates the monitoring station indices. Stars denote significant p-values, with colors representing 1% (orange), 5% (green), 10% (blue), and 20% (violet) significance levels, respectively.
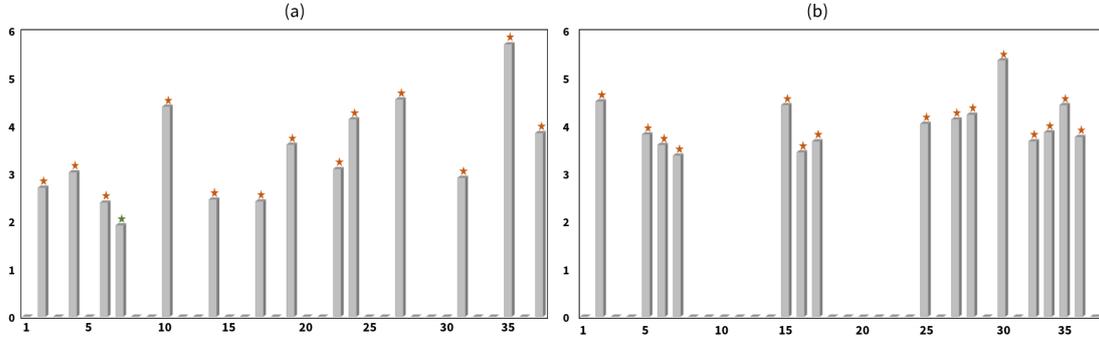

Figure 9: DM test results comparing (a) E-STGCN and STGCN, and (b) E-STGCN and NBeats for forecasting PM_{10} pollutant concentrations over the 90-day OCT-NOV-DEC forecast window. The Y-axis represents DM-test statistic values based on the MAE metric, while the X-axis indicates the monitoring station indices. Stars denote significant p-values, with colors representing 1% (orange) and 5% (green) significance levels, respectively.

4.6. Uncertainty Quantification

In addition to producing the point forecasts of the air pollutant concentrations through the E-STGCN approach, we quantify the uncertainty inherent with these forecasts using the conformal prediction technique (Vovk et al., 2005). This distribution-free approach generates the probabilistic intervals around the point estimates based on a conformal score (γ_t). The computation of γ_t at time t involves modeling p-lagged values of the target series $\mathbf{X}_t$ using both E-STGCN and an uncertainty model $\mathcal{U}$ as follows

$$\gamma_t = \frac{|\mathbf{X}_t - \text{E-STGCN}\left(\mathbf{X}_{t-p}\right)|}{\mathcal{U}\left(\mathbf{X}_{t-p}\right)}.$$

Subsequently, using the sequential nature of $\mathbf{X}_t$ and γ_t, we derive the conformal quantile by applying a weighted aggregation technique with a fixed window $\{\nu_t = \mathbb{1}\left(\chi \geqslant t - \upsilon\right), \ \chi < t\}$ of size υ as

$$\kappa_t = \inf\left\{\omega : \frac{1}{\min\left(\upsilon, \chi - 1\right)} \sum_{\chi=1}^{t-1} \gamma_\chi \nu_\chi \geqslant 1 - \rho\right\},$$

21

where ρ is the significance level. Then, the computation of the conformal prediction interval using the conformal quantiles κ_t can be expressed as,

$$[\text{E-STGCN}\left(\mathbf{X}_{t-p}\right) \pm \kappa_t \, \mathcal{U}\left(\mathbf{X}_{t-p}\right)].$$

In Fig. 10, we present the point and interval estimate of air pollutant concentrations generated by the E-STGCN model along with the results of the two best-performing baselines STGCN and NBeats, as identified by the MCB plots in Section 4.5. The conformal prediction intervals demonstrated in Fig. 10 are calculated with $\rho = 0.20$ for three selected forecasting scenarios. The plot highlights the generalizability of the proposed E-STGCN model in providing valuable insights about air pollutant concentration levels, mainly modeling their threshold exceedance values. These findings are pivotal for environmentalists in designing awareness and mitigation strategies.
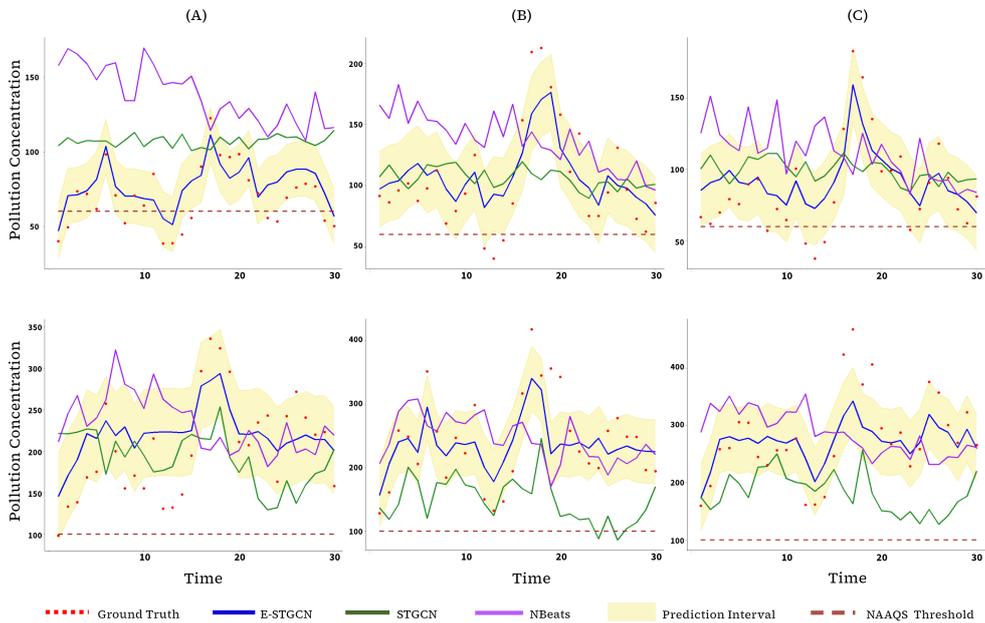

Figure 10: Upper panel presents ground truth (red dots) PM$_{2.5}$ pollutant concentrations monitored at (A) DTU, (B) Dr. Karni Singh Shooting Range, and (C) IGI Airport stations during February 2023 window and corresponding point forecasts of E-STGCN (blue line), STGCN (green line), and NBeats (violet line) framework. The conformal prediction interval (yellow-shaded) of the E-STGCN model quantifies the associated uncertainty. The lower panel highlights similar information about PM$_{10}$ concentrations monitored at the corresponding stations.

5. Policy implications

Rapid urbanization and industrialization have significantly impacted air quality in many developing and underdeveloped countries. In its 2021 Global Air Quality Guidelines (AQGs), WHO recommended critical air pollutants such as PM, NO$_2$, SO$_2$, O$_3$, and CO based on their effects on mortality and human health. Among these, PM and NO$_2$ have gained particular attention from air quality researchers due to their direct links to increased mortality, as evidenced in epidemiological studies (Olaniyan et al., 2020). NO$_2$ is a highly reactive gas primarily emitted from automobile exhaust, power plants, and industrial machinery. In urban areas, NO$_2$ levels are mainly driven by the transportation sector. For instance, the urban regions of North America and Europe often report higher NO$_2$ despite low levels of PM$_{2.5}$ and PM$_{10}$ (Ji et al., 2022). It was found that acute exposure to NO$_2$ can aggravate respiratory diseases, such as asthma and

22

other pulmonary symptoms, although no causal relationship between NO_2 exposure and health mortality was established (Faustini et al., 2014). On the other hand, PM comprises a mix of acids (such as nitrates and sulfates), organic chemicals, metals, soil, dust particles, and allergens. These particles originate from various sources like fossil fuel combustion, industrial emissions, construction activities, wildfires, stubble burning, and household cooking. In Delhi, PM levels frequently exceed the NAAQS, even when NO_2 concentrations remain relatively low (Abirami and Chitra, 2021). It is particularly critical as PM is identified as a causal factor for cardiovascular and respiratory mortality and remains a serious concern for India's capital. Given that the population of Delhi and the national capital region (NCR) of India is particularly vulnerable to PM exposure, which can cause health emergencies, this study aims to forecast PM levels in Delhi by considering extreme behaviors of air pollutants. The proposed E-STGCN framework offers a technological solution for real-time monitoring and forecasting of hazardous air pollutants in Delhi. This approach is particularly valuable when extreme observations and nonlinear patterns characterize the observed spatiotemporal data. The proposed methodology has the potential to advance future research endeavors on enhancing air quality forecasting models and to promote environmental sustainability. Although E-STGCN has been developed specifically for air pollution data in this study, it can also be extended to other applied fields, including epidemiology, seismology, and transportation research, where similar patterns of extreme events and complex dependencies are frequently observed.

6. Conclusion

Accurate air quality forecasting remains a challenging problem due to complex spatiotemporal dependencies in pollutant concentration levels. Pollutants such as $PM_{2.5}$, PM_{10}, and NO_2 often exhibit extreme behaviors while also displaying nonlinear and non-stationary properties. Among these hazardous pollutants, $PM_{2.5}$ and PM_{10} concentrations are consistently high in some of the world's majorly polluted cities, leading to serious health hazards and restricting economic growth. In particular, air pollution levels can intensify with seasonal variations. For instance, in Delhi, the concentration of $PM_{2.5}$ and PM_{10} increases rapidly during winter due to low wind speed, stubble burning, firecracker emissions, and other contributing sources. To address these challenges, public awareness through early warning systems is of paramount importance.

In this study, we propose the E-STGCN model, which aims to provide actionable insights by generating real-time forecasts of air pollutant concentrations. Our approach bridges the gap between existing EVT-based models, which focus on predicting extreme behavior, and data-driven forecasting methods that predict future trajectories without accounting for the tail behavior of the extreme observations. By integrating EVT knowledge with spatiotemporal GCNs, our proposed framework effectively performs spatiotemporal forecasting while tackling extreme observations. Experimental results, conducted on real-world air pollutant data (daily frequencies) of $PM_{2.5}$, PM_{10}, and NO_2 from 37 monitoring stations in Delhi demonstrate that the E-STGCN approach is well-suited for predicting the future dynamics of nonlinear and non-stationary datasets with spatiotemporal dependence and extreme events. Additionally, the model generates appropriate probabilistic bands along with point forecasts, enabling environmental advocates to monitor air pollution trends and design effective control strategies. Further, the forecastability and statistical significance tests conducted in this study verify the effectiveness and robustness of the proposed architecture for air pollution forecasting over various time horizons.

An interesting avenue for future research is to identify various climatic, transportation, and industrial indices that have a causal impact on rising air pollution levels. Future studies could explore how incorporating these causal covariates might enhance the accuracy of the E-STGCN approach. Another potential direction would be to extend the air pollution modeling capabilities of E-STGCN on a global scale, analyzing its impact on pollution-related mortality and morbidity.

References

Abirami, S., Chitra, P., 2021. Regional air quality forecasting using spatiotemporal deep learning. Journal of Cleaner Production 283, 125341.

AL-Dhurafi, N.A., Masseran, N., Zamzuri, Z.H., Razali, A.M., 2018. Modeling unhealthy air pollution index using a peaks-over-threshold method. Environmental Engineering Science 35, 101–110.

Atluri, G., Karpatne, A., Kumar, V., 2018. Spatio-temporal data mining: A survey of problems and methods. ACM Computing Surveys (CSUR) 51, 1–41.

Balkema, A.A., De Haan, L., 1974. Residual life time at great age. The Annals of probability 2, 792–804.

Benktander, G., Segerdahl, C.O., 1960. On the analytical representation of claim distributions with special reference to excess of loss reinsurance, in: Transactions of the h~ ternational Congress of Actuaries, pp. 1–10.

Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 1970. Time series analysis: forecasting and control. John Wiley & Sons.

Brunekreef, B., Holgate, S.T., 2002. Air pollution and health. The lancet 360, 1233–1242.

Byun, D., Schere, K.L., 2006. Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (cmaq) modeling system. Applied mechanics reviews 59, 51–77.

Castillo, E., 2012. Extreme value theory in engineering. Elsevier.

Chen, Y., Kang, Y., Chen, Y., Wang, Z., 2020. Probabilistic forecasting with temporal convolutional neural network. Neurocomputing 399, 491–501.

Chukwudum, Q.C., Mwita, P., Mung'atu, J.K., 2020. Optimal threshold determination based on the mean excess plot. Communications in Statistics-Theory and Methods 49, 5948–5963.

Cliff, A., Ord, J., 1975. Model building and the analysis of spatial pattern in human geography. Journal of the Royal Statistical Society: Series B (Methodological) 37, 297–328.

Coles, S., Bawa, J., Trenner, L., Dorazio, P., 2001. An introduction to statistical modeling of extreme values. volume 208. Springer.

Diebold, F.X., Mariano, R.S., 2002. Comparing predictive accuracy. Journal of Business & economic statistics 20, 134–144.

Du, S., Li, T., Yang, Y., Horng, S.J., 2019. Deep air quality forecasting using hybrid deep learning framework. IEEE Transactions on Knowledge and Data Engineering 33, 2412–2424.

Durbin, J., Watson, G.S., 1971. Testing for serial correlation in least squares regression. iii. Biometrika 58, 1–19.

Embrechts, P., Klüppelberg, C., Mikosch, T., 2013. Modelling extremal events: for insurance and finance. volume 33. Springer Science & Business Media.

Farkas, S., Heranval, A., Lopez, O., Thomas, M., 2024. Generalized pareto regression trees for extreme event analysis. Extremes , 1–41.

Faustini, A., Rapp, R., Forastiere, F., 2014. Nitrogen dioxide and mortality: review and meta-analysis of long-term studies. European Respiratory Journal 44, 744–753.

Fisher, R.A., Tippett, L.H.C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample, in: Mathematical proceedings of the Cambridge philosophical society, Cambridge University Press. pp. 180–190.

Gao, X., Li, W., 2021. A graph-based lstm model for pm2. 5 forecasting. Atmospheric Pollution Research 12, 101150.

Ghosh, S., Resnick, S., 2010. A discussion on mean excess plots. Stochastic Processes and their Applications 120, 1492–1517.

Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E., 2017. Neural message passing for quantum chemistry, in: International conference on machine learning, PMLR. pp. 1263–1272.

Guinness, J., 2018. Permutation and grouping methods for sharpening gaussian process approximations. Technometrics 60, 415–429.

Gumbel, E.J., 1958. Statistics of extremes. Columbia university press.

Herzen, J., et al., 2022. Darts: User-friendly modern machine learning for time series. Journal of Machine Learning Research 23, 1–6.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Horowitz, J., 1980. Extreme values from a nonstationary stochastic process: an application to air quality analysis. Technometrics 22, 469–478.

Hyndman, R., 2018. Forecasting: principles and practice. OTexts.

Ji, J.S., Liu, L., Zhang, J., Kan, H., Zhao, B., Burkart, K.G., Zeng, Y., 2022. No2 and pm2. 5 air pollution co-exposure and temperature effect modification on pre-mature mortality in advanced age: a longitudinal cohort study in china. Environmental Health 21, 97.

Jin, M., Koh, H.Y., Wen, Q., Zambon, D., Alippi, C., Webb, G.I., King, I., Pan, S., 2024. A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Kan, H.D., Chen, B.H., 2004. Statistical distributions of ambient air pollutants in shanghai, china. Biomedical and Environmental Sciences 17, 366–372.

Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nature Reviews Physics 3, 422–440.

Katz, R.W., Parlange, M.B., Naveau, P., 2002. Statistics of extremes in hydrology. Advances in water resources 25, 1287–1304.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 .

Koning, A.J., Franses, P.H., Hibon, M., Stekler, H.O., 2005. The m3 competition: Statistical tests of the results. International journal of forecasting 21, 397–409.

Kumar, U., Jain, V., 2010. Arima forecasting of ambient air pollutants (o3, no, no2 and co). Stochastic Environmental Research

and Risk Assessment 24, 751–760.

Lei, M.T., Monjardino, J., Mendes, L., Gonçalves, D., Ferreira, F., 2019. Macao air quality forecast using statistical methods. Air Quality, Atmosphere & Health 12, 1049–1057.

Lelieveld, J., Evans, J.S., Fnais, M., Giannadaki, D., Pozzer, A., 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature 525, 367–371.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T., 2017. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental pollution 231, 997–1004.

Marimoutou, V., Raggad, B., Trabelsi, A., 2009. Extreme value theory and value at risk: application to oil market. Energy Economics 31, 519–530.

Martins, L.D., Wikuats, C.F.H., Capucim, M.N., de Almeida, D.S., da Costa, S.C., Albuquerque, T., Carvalho, V.S.B., de Freitas, E.D., de Fátima Andrade, M., Martins, J.A., 2017. Extreme value analysis of air pollution data and their comparison between two large urban regions of south america. Weather and Climate Extremes 18, 44–54.

Nag, P., Sun, Y., Reich, B.J., 2023. Spatio-temporal deepkriging for interpolation and probabilistic forecasting. Spatial Statistics 57, 100773.

Olaniyan, T., Jeebhay, M., Röösli, M., Naidoo, R.N., Künzli, N., de Hoogh, K., Saucy, A., Badpa, M., Baatjies, R., Parker, B., et al., 2020. The association between ambient no2 and pm2. 5 with the respiratory health of school children residing in informal settlements: A prospective cohort study. Environmental research 186, 109606.

Ong, B.T., Sugiura, K., Zettsu, K., 2016. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting pm 2.5. Neural Computing and Applications 27, 1553–1566.

Oreshkin, B.N., Carpov, D., Chapados, N., Bengio, Y., 2019. N-beats: Neural basis expansion analysis for interpretable time series forecasting. arXiv preprint arXiv:1905.10437 .

Pandey, A., Brauer, M., Cropper, M.L., Balakrishnan, K., Mathur, P., Dey, S., Turkgulu, B., Kumar, G.A., Khare, M., Beig, G., et al., 2021. Health and economic impact of air pollution in the states of india: the global burden of disease study 2019. The Lancet Planetary Health 5, e25–e38.

Pfeifer, P.E., Deutrch, S.J., 1980. A three-stage iterative procedure for space-time modeling phillip. Technometrics 22, 35–47.

Pickands III, J., 1975. Statistical inference using extreme order statistics. the Annals of Statistics , 119–131.

Ray, A., Chakraborty, T., Radhakrishnan, A., Hens, C., Dana, S.K., Ghosh, D., Murukesh, N., 2023. Pattern change of precipitation extremes in bear island. arXiv preprint arXiv:2312.04502 .

Reiss, R.D., Thomas, M., Reiss, R., 1997. Statistical analysis of extreme values. volume 2. Springer.

Roberts, E., 1979. Review of statistics of extreme values with applications to air quality data: part ii. applications. Journal of the Air Pollution Control Association 29, 733–740.

Rocco, M., 2014. Extreme value theory in finance: A survey. Journal of Economic Surveys 28, 82–108.

Ruchjana, B.N., Borovkova, S.A., Lopuhaa, H., Baskoro, E.T., Suprijanto, D., 2012. Least squares estimation of generalized space time autoregressive (gstar) model and its properties, in: AIP Conference Proceedings-American Institute of Physics, AIP. p. 61.

Saha, A., Singh, K., Ray, M., Rathod, S., 2020. A hybrid spatio-temporal modelling: an application to space-time rainfall forecasting. Theoretical and Applied Climatology 142, 1271–1282.

Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T., 2020. Deepar: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting 36, 1181–1191.

Salvi, S., Kumar, G.A., Dhaliwal, R., Paulson, K., Agrawal, A., Koul, P.A., Mahesh, P., Nair, S., Singh, V., Aggarwal, A.N., et al., 2018. The burden of chronic respiratory diseases and their heterogeneity across the states of india: the global burden of disease study 1990–2016. The Lancet Global Health 6, e1363–e1374.

Samal, K.K.R., Panda, A.K., Babu, K.S., Das, S.K., 2021. Multi-output tcn autoencoder for long-term pollution forecasting for multiple sites. Urban Climate 39, 100943.

Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G., 2008. The graph neural network model. IEEE transactions on neural networks 20, 61–80.

Sfetsos, A., Zoras, S., Bartzis, J.G., Triantafyllou, A.G., 2006. Extreme value modeling of daily pm 10 concentrations in an industrial area. Fresenius Environmental Bulletin 15, 841–845.

Shaddick, G., Thomas, M.L., Mudu, P., Ruggeri, G., Gumy, S., 2020. Half the world's population are exposed to increasing air pollution. NPJ Climate and Atmospheric Science 3, 1–5.

Sharma, P., Khare, M., Chakrabarti, S., 1999. Application of extreme value theory for predicting violations of air quality standards for an urban road intersection. Transportation Research Part D: Transport and Environment 4, 201–216.

Thomas, M., Lemaitre, M., Wilson, M.L., Viboud, C., Yordanov, Y., Wackernagel, H., Carrat, F., 2016. Applications of extreme value theory in public health. PloS one 11, e0159312.

Vardoulakis, S., Fisher, B.E., Pericleous, K., Gonzalez-Flesca, N., 2003. Modelling air quality in street canyons: a review. Atmospheric environment 37, 155–182.

Vaswani, A., 2017. Attention is all you need. Advances in Neural Information Processing Systems .

Vecchia, A.V., 1988. Estimation and model identification for continuous spatial processes. Journal of the Royal Statistical Society Series B: Statistical Methodology 50, 297–312.

Vovk, V., Gammerman, A., Shafer, G., 2005. Algorithmic learning in a random world. volume 29. Springer.

Wang, Z., Li, J., Wang, Z., Yang, W., Tang, X., Ge, B., Yan, P., Zhu, L., Chen, X., Chen, H., et al., 2014. Modeling study of regional severe hazes over mid-eastern china in january 2013 and its implications on pollution prevention and control. Science China Earth Sciences 57, 3–13.

Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE transactions on evolutionary computation 1, 67–82.

Wu, N., Green, B., Ben, X., O'Banion, S., 2020. Deep transformer models for time series forecasting: The influenza prevalence case. arXiv preprint arXiv:2001.08317 .

Yu, B., Yin, H., Zhu, Z., 2018. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, ACM. International Joint Conferences on Artificial Intelligence Organization. pp. 3634–3640. doi:10.24963/ijcai.2018/505.

Zhou, X., Wang, J., Wang, J., Guan, Q., 2024. Predicting air quality using a multi-scale spatiotemporal graph attention network. Information Sciences 680, 121072.

Supplementary material

S.1. Statistical Tests on Air Pollutant Data

Below, we summarize the global features of the dataset used in our analysis and highlight their implementation strategies:

- *Long-term dependency* is a crucial feature in time series processes and has gained significant attention in probabilistic time series modeling. To evaluate long-range dependency, we examine the self-similarity parameter, often referred to as the Hurst exponent, using the *pracma* package in R.

- *Stationarity* is a fundamental property of time series that implies the statistical features, particularly mean and variance, remain constant over time. To assess stationarity, we employ the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test from the *tseries* package in R.

- *Linearity* is another essential characteristic of time series data, playing a critical role in model selection. In our analysis, we apply Teraesvirta's neural network test from the *nonlinearTseries* package in R to determine whether the data-generating process follows a linear trend.

- *Seasonality* refers to recurring patterns in time series that occur at regular intervals. To identify the presence and frequency of seasonal patterns in our dataset, we use Ollech and Webel's test from the *seastests* package in R.

Descriptive statistics and these global features of the three pollutant series ($PM_{2.5}$, PM_{10}, and NO_2) for various monitoring stations in our dataset are presented in Tables S.1, S.2 and S.3.

S.2. Baseline Models

(a) Temporal baseline models:

- **Autoregressive Integrated Moving Average** (ARIMA) is a popular statistical method for time series forecasting (Box et al., 1970). The ARIMA(p, d, q) framework tracks the linear trajectory in a d-order (non-negative integer) differenced stationary time series by combining p historical values of the target series and q prior forecast errors. We utilize the *forecast* package in R statistical software to implement the ARIMA model.

- **Long-short Term Memory** (LSTM) networks, a recurrent neural networks (RNN) architecture, is suitable for modeling long-term dependencies in time series forecasting (Hochreiter and Schmidhuber, 1997). This framework utilizes a gating mechanism with the input gate, forget gate and output gate to regulate the flow of information as short-term and long-term memory.

- **Temporal Convolutional Network** (TCN) combines causal convolutions and dilated convolutions to model the long-term dependencies in a time series dataset (Chen et al., 2020). This architecture has a stable training mechanism due to skip connections in the residual blocks.

- **DeepAR** is a variant of the RNN approach, capable of performing probabilistic forecasting (Salinas et al., 2020). This scalable architecture is suitable for handling complex seasonality in multiple time series.

Table S.1: Descriptive statistics of the PM$_{2.5}$ pollutant levels in different stations. In the table, weekly (W), quarterly (Q), and no (-) seasonality are indicated.

Station	Min	1st quartile	Mean	Median	3rd quartile	Max	Sd	CV	Skewness	Kurtosis	Freq of EVal	Seas	DW test
1 — Alipur	4.81	48.15	127.64	95.71	181.72	758.40	103.66	0.81	1.42	2.36	68.01	-	0.12
2 — Anand Vihar	9.52	52.72	122.39	88.94	162.3	592.28	96.17	0.79	1.52	2.24	69.66	W	0.11
3 — Ashok Vihar	5.95	42.35	109.27	75.6	146.49	601.49	93.26	0.85	1.68	3.13	60.41	-	0.06
4 — Aya Nagar	7.96	37.22	77.86	58.01	98.12	556.03	61.17	0.79	2.33	8.45	47.95	W	0.24
5 — Bawana	7.83	52.92	122.94	94.93	166.03	696.78	93.11	0.76	1.55	3.19	69.38	-	0.05
6 — Burari Crossing	10.97	81.64	105.04	99.79	118.07	565.65	53.53	0.51	2.33	9.28	83.36	-	0.37
7 — CRRI Mathura Road	7.97	42.72	97.18	72.76	125.91	518.18	76.23	0.78	1.78	3.86	59.52	Q	0.24
8 — DTU	9.39	46.63	111.06	83.60	149.51	631.85	86.41	0.78	1.67	3.65	65.27	-	0.84
9 — Dr. Karni Singh Shooting Range	4.39	34.27	89.47	60.34	122.61	571.73	79.41	0.89	1.79	3.89	50.21	Q	0.09
10 — Dwarka Sector 8	8.09	41.38	104	76.03	140.48	600.4	85.35	0.82	1.72	3.88	59.59	-	0.03
11 — IGI Airport	2.68	34.89	82.85	60.57	108.71	506.20	67.44	0.81	1.98	5.51	50.55	-	0.01
12 — Ihbas Dilshad Garden	7.50	43.27	91.83	76.73	119.43	606.41	66.98	0.73	1.70	4.67	62.40	-	0.63
13 — ITO	12.13	53.52	111.17	84.02	142.13	659.29	84.56	0.76	2.04	5.86	69.25	-	0.29
14 — Jahangirpuri	8.77	49.88	128.39	90.71	179.66	658.26	105.42	0.82	1.46	2.03	67.67	-	0.05
15 — Jawaharlal Nehru Stadium	3.79	36.18	95.16	65.22	128.78	513.36	82.48	0.87	1.69	3.17	53.84	-	0.05
16 — Lodhi Road IMD	5.49	38.08	78.41	60.37	99.07	479.72	58.98	0.75	2.05	5.80	50.34	-	0.26
17 — Major Dhyan Chand National Stadium	6.71	38.98	94.15	67.88	126.01	525.64	76.45	0.81	1.63	3.07	55.41	W	0.75
18 — Mandir Marg	4.67	39.87	93.91	73.81	121.22	563.93	74.69	0.8	1.75	4.07	59.45	-	0.18
19 — Mundka	7.38	45.92	120.93	93.04	169.61	698.96	98.22	0.81	1.55	3.12	65.27	-	0.20
20 — Najafgarh	4.50	35.22	83.68	67.57	111.60	574.41	65.59	0.78	1.92	6.09	55.89	Q	0.31
21 — Narela	6.62	47.04	110.17	84.39	149.57	689.10	85.87	0.78	1.62	3.68	65.62	-	0.19
22 — Nehru Nagar	6.06	41.86	116.12	74.08	163.04	554.05	101.59	0.87	1.48	1.96	59.45	W	0.03
23 — North Campus DU	5.58	43.00	95.97	70.35	123.72	570.07	77.70	0.81	1.87	4.43	58.22	-	0.16
24 — NSIT Dwarka	8.33	51.39	100.39	86.92	133.17	406.61	65.08	0.65	1.23	1.72	68.42	-	0.76
25 — Okhla Phase 2	6.18	37.48	99.64	67.85	132.97	547.56	87.48	0.88	1.72	3.22	55.07	W	0.07
26 — Patparganj	4.73	42.34	104.57	74.03	138.85	633.95	87.84	0.84	1.65	3.21	60.41	W	0.25
27 — Punjabi Bagh	6.84	46.34	109.45	79.57	144.22	609.18	88.64	0.81	1.84	4.40	62.67	-	0.04
28 — PUSA DPCC	3.68	37.14	93.82	67.32	130.31	570.61	79.06	0.84	1.64	3.29	54.25	W	0.11
29 — PUSA IMD	7.46	35.75	80.82	60.75	103.87	569.5	64.86	0.80	2.13	7.01	50.89	-	0.06
30 — R K Puram	6.38	41.25	99.97	71.01	134.12	558.86	83.07	0.83	1.61	3.01	58.29	-	0.06
31 — Rohini	6.13	47.64	119.07	85.15	160.09	761.95	99.09	0.83	1.72	3.81	64.79	-	0.89
32 — Shadipur	9.38	37.88	89.67	72.40	119.80	397.45	66.43	0.74	1.40	2.13	58.42	-	0.35
33 — Sirifort	0.08	39.96	96.32	70.71	129.24	573.09	78.07	0.81	1.79	4.38	57.95	W	0.40
34 — Sonia Vihar	6.01	46.37	106.71	77.84	138.55	598.80	86.30	0.81	1.75	3.69	62.60	-	0.34
35 — Sri Aurobindo Marg	5.28	34.88	85.33	60.77	113.41	534.86	72.01	0.84	1.94	5.31	50.68	-	0.17
36 — Vivek Vihar	4.61	45.91	115.35	79.91	156.20	650.88	97.79	0.85	1.57	2.48	62.60	-	0.27
37 — Wazirpur	6.88	48.74	114.83	80.10	150.70	585.79	92.57	0.81	1.69	3.04	65.41	-	0.15

- **Transformers** is a state-of-the-art deep learning architecture that models complex patterns in time series data (Wu et al., 2020). This framework utilizes the multi-head attention mechanism to capture the crucial information in a sequential learning problem.

- **Neural Basis Expansion Analysis for Time Series** (NBeats) is a fully connected neural network architecture designed for time series forecasting (Oreshkin et al., 2019). This model comprises several blocks equipped with a basis expansion mechanism for transforming data into high-dimensional space

28

Table S.2: Descriptive statistics of the PM_{10} pollutant levels in different stations. In the table, weekly (W), quarterly (Q), and no (-) seasonality are indicated.

Station	Min	1st quartile	Mean	Median	3rd quartile	Max	Sd	CV	Skewness	Kurtosis	Freq of EVal	Seas	DW test
1	10.04	98.38	201.54	183.66	284.63	758.40	125.43	0.62	0.76	0.36	74.38	-	0.20
2	16.26	128.95	265.22	249.69	361.79	729.94	153.63	0.58	0.57	-0.38	83.56	-	0.31
3	11.80	106.94	212.58	191	290.86	753.97	130.1	0.61	0.91	0.59	77.47	-	0.16
4	11.39	79.36	152.38	138.25	200.58	665.41	94.31	0.62	1.22	2.20	66.16	W	0.17
5	12.03	127.09	240.84	218.46	335.64	810.85	139.82	0.58	0.73	0.16	82.74	-	0.20
6	20.33	135.13	201.05	180.15	226.15	795.88	99.65	0.50	1.40	2.61	91.99	-	1.00
7	10.22	92.00	205.10	189.72	292.14	721.69	129.26	0.63	0.74	0.22	72.19	-	0.45
8	1.00	114.05	214.90	199.14	295.47	923.70	123.74	0.58	0.78	0.87	78.97	-	0.72
9	8.03	87.93	180.60	166.45	243.90	756.47	110.83	0.61	0.88	0.99	71.37	Q, W	0.52
10	14.63	136.14	253.11	247.16	351.80	807.89	136.80	0.54	0.47	-0.21	85.21	-	0.38
11	12.95	87.57	172.51	148.58	235.35	633.67	105.43	0.61	1.13	1.35	68.22	W	0.01
13	16.00	94.81	170.99	153.03	226.25	691.84	99.54	0.58	1.21	2.17	71.85	-	0.75
14	15.48	125.76	248.47	229.62	339.05	821.78	145.85	0.59	0.72	0.03	81.99	-	0.23
15	10.25	96.43	190.98	175.62	260.25	678.37	115.37	0.60	0.86	0.71	74.04	Q, W	0.46
16	10.06	84.84	161.70	147.87	218.69	611.18	93.82	0.58	0.95	1.05	69.45	-	0.18
17	12.44	98.55	190.49	171.7	261.15	663.12	112.55	0.59	0.82	0.39	74.32	-	0.32
18	17.09	92.99	172.77	159.56	231.77	705.68	98.16	0.57	0.90	0.96	71.58	-	0.46
19	10.67	129.55	256.72	241.74	358.95	790.97	144.14	0.56	0.52	-0.35	84.32	-	0.54
20	8.75	86.20	157.88	147.47	212.83	731.23	93.13	0.59	0.93	1.54	69.18	Q	0.92
21	20.48	128.72	231.14	207.18	314.71	718.18	127.53	0.55	0.75	0.16	85.07	-	0.21
22	10.48	97.90	205.37	181.25	284.89	702.19	129.16	0.63	0.90	0.53	74.11	W	0.29
23	5.17	100.15	194.37	176.32	260.85	735.53	119.71	0.62	1.00	1.10	75.07	-	0.38
25	9.73	108.34	211.05	190.43	282.79	741.37	124.62	0.59	0.86	0.56	78.08	W	0.34
26	8.36	90.98	189.19	170.67	262.95	689.97	118.8	0.63	0.83	0.36	71.85	W	0.23
27	20.60	109.55	206.63	183.98	277.36	768.33	121.4	0.59	0.96	0.79	78.36	W	0.27
28	9.97	104.87	201.78	191.35	277.41	726.86	117.88	0.58	0.66	0.26	76.64	-	0.79
29	13.01	69.98	152.34	130.26	209.25	706.66	99.95	0.66	1.24	2.02	62.12	W	0.22
30	11.52	98.04	193.74	180.34	266.95	699.3	111.82	0.58	0.74	0.46	74.45	W	0.09
31	11.1	116.87	230.09	204.46	318.69	783.4	138.92	0.60	0.82	0.21	80.14	-	0.09
33	10.97	118.90	221.00	213.68	301.66	664.34	121.23	0.55	0.63	0.18	82.05	-	0.50
34	13.00	108.97	213.34	190.00	292.41	720.11	127.22	0.60	0.88	0.50	78.77	-	0.75
35	8.86	72.39	148.27	134.51	202.46	596.06	92.57	0.62	1.00	1.23	63.15	-	0.53
36	12.48	111.68	223.87	198.14	305.24	699.36	132.67	0.59	0.85	0.32	79.18	W	0.33
37	19.3	138.93	246.48	215.07	325.32	793.65	139.44	0.57	0.99	0.68	88.42	-	0.04

and dense layers. The initial layers of the block are used for modeling and predicting past and future observations, while the subsequent layers are designed to remodel the errors and adjust the forecasts.

To implement the above-mentioned deep learning models, we have used the *darts* library in Python (Herzen and et al., 2022).

(b) Spatiotemporal baseline models:

- **Space-time Autoregressive Moving Average** (STARMA) is a modification of the autoregressive moving average framework that incorporates the spatiotemporal auto-correlations (Pfeifer and Deutrch, 1980). This architecture includes both autoregressive and moving average terms that are lagged

Table S.3: Descriptive statistics of the NO$_2$ pollutant levels in different stations. In the table, weekly (W), quarterly (Q), and no (-) seasonality are indicated.

Station	Min	1st quartile	Mean	Median	3rd quartile	Max	Sd	CV	Skewness	Kurtosis	Freq of EVal	Seas	DW test
1	0.70	17.88	34.28	30.93	46.07	120.01	20.60	0.60	0.99	1.00	3.70	-	0.46
2	4.87	44.92	75.56	70.53	100.18	325.77	39.95	0.53	1.06	2.55	41.03	Q	0.25
3	3.54	21.59	40.42	36.01	55.90	183.59	23.53	0.58	0.96	1.57	5.41	-	0.83
4	0.94	13.15	19.48	18.58	23.76	128.24	11.50	0.59	2.29	10.54	0.07	-	NA
5	1.85	11.60	27.25	21.86	36.65	188.92	21.39	0.79	1.98	7.01	2.40	-	0.23
6	1.54	34.17	142.87	86.02	250.16	428.15	130.93	0.92	0.72	-0.90	51.37	Q	1.00
7	0.33	14.00	38.52	20.72	41.88	308.38	43.89	1.14	2.30	5.01	12.67	-	0.29
8	1.06	21.43	40.66	33.89	47.72	276.46	32.42	0.8	2.41	7.97	8.77	-	0.51
9	3.44	28.22	53.46	46.16	72.41	291.46	34.51	0.65	1.64	6.14	20.07	-	0.24
10	6.61	21.27	36.5	31.97	47.19	127.53	20.34	0.56	1.16	1.40	3.70	-	0.46
11	0.64	23.9	44.08	33.03	56.03	417.31	37.91	0.86	3.43	21.22	11.58	-	0.01
12	6.32	26.16	50.41	42.99	67.89	197.44	31.23	0.62	1.17	1.48	16.71	-	0.59
13	8.13	19.87	36.33	28.14	40.49	272.12	29.42	0.81	3.23	14.19	6.99	-	0.23
14	8.63	31.49	60.03	50.28	73.94	237.59	39.34	0.66	1.62	2.63	20.96	-	1.00
15	6.51	40.86	60.95	58.68	78.80	202.93	27.71	0.45	0.55	0.54	23.90	-	0.47
16	0.13	6.49	13.73	9.99	18.13	100.24	11.54	0.84	2.15	7.10	0.21	-	1.00
17	8.97	23.66	42.06	35.99	53.79	159.29	24.17	0.57	1.25	1.68	8.70	-	0.48
18	11.15	39.88	54.12	52.51	69	179.24	22.53	0.42	0.46	0.64	12.26	-	0.67
19	4.38	23.26	37.74	33.95	49.39	117.4	18.32	0.49	0.85	0.47	2.33	-	0.38
20	2.96	11.74	20.73	18.23	27.27	89.38	12.29	0.59	1.36	2.90	0.14	-	NA
21	4.12	26.27	38.89	34.96	48.52	150.4	17.65	0.45	1.21	2.25	3.22	-	0.49
22	8.63	34.86	55.21	48.80	71.95	225.00	27.77	0.50	1.30	3.23	17.53	-	0.93
23	0.96	9.09	25.18	17.37	30.10	205.36	24.75	0.98	2.34	6.74	5.27	-	0.24
24	2.09	19.09	29.37	26.55	37.99	101.78	14.01	0.48	1.20	2.49	0.89	-	0.36
25	7.12	29.30	49.55	45.01	65.23	219.36	25.99	0.52	0.96	1.80	14.59	-	0.08
26	1.78	15.05	31.51	24.23	38.31	140.9	24.21	0.77	1.59	2.44	5.75	-	0.44
27	0.22	33.69	51.52	46.93	63.70	208.14	25.11	0.49	1.58	5.26	11.10	-	0.51
28	6.57	33.97	53.48	54.06	71.61	155.5	24.43	0.46	0.14	-0.63	14.59	-	0.50
29	0.65	12.32	35.21	23.84	47.35	325.08	35.86	1.02	2.57	10.75	9.86	-	0.40
30	0.27	27.91	44.34	43.06	59.15	179.95	22.45	0.51	0.53	1.06	6.71	-	0.70
31	0.30	14.48	25.59	21.53	32.62	146.91	15.69	0.61	1.72	5.22	0.75	-	0.51
32	6.37	24.22	51.60	44.04	73.03	181.41	32.76	0.63	0.95	0.58	19.18	-	0.44
33	0.21	14.10	34.24	25.96	48.55	247.18	27.86	0.81	1.73	5.02	7.05	-	0.20
34	2.87	21.53	35.90	32.32	46.62	111.92	18.41	0.51	0.91	0.77	2.74	Q	0.42
35	2.22	20.39	29.54	28.34	36.58	106.78	12.91	0.44	0.95	2.18	0.27	-	NA
36	0.15	17.76	29.64	25.19	39.45	107.11	16.14	0.54	0.98	0.95	0.75	-	0.67
37	2.05	23.73	41.77	37.17	55.93	139.69	23.67	0.57	0.87	0.60	7.88	-	0.55

in space and time, making it useful for modeling linear trajectories in spatiotemporal systems.

- **Generalized Space-time Autoregressive** (GSTAR) model is a robust spatiotemporal forecasting framework that allows the autoregressive parameters to vary across different locations (Cliff and Ord, 1975; Ruchjana et al., 2012). Unlike the STARMA model, the non-uniform weights of the GSTAR architecture make it more useful for modeling heterogeneous characteristics of sample locations.

- **Fast Gaussian Process** (GpGp) method is a modification of Vecchia's Gaussian process approximation (Vecchia, 1988), designed for analyzing ordered sequences in time series observations (Guinness, 2018). This approach introduces a grouping mechanism for the ordered sequence, which significantly

reduces the computational complexity of traditional Gaussian process forecasting methods.

- **Spatiotemporal Neural Network** (STNN) is a hybrid forecasting approach that combines the classical STARMA model with Support Vector Machine (SVM) and Artificial Neural Networks (ANN) to enhance forecast accuracy (Saha et al., 2020). The STNN operates as an error-remodeling approach, where the original training data is first modeled using the linear STARMA model. Residuals of the STARMA model are then modeled using the SVM architecture, and ANN is subsequently applied to the predicted values of the residuals from SVM to capture the nonlinearities better. The final STNN forecasts are obtained by aggregating the predictions from both the STARMA and ANN frameworks.

- **Spatiotemporal Graph Convolution Network** (STGCN) is a graph-based deep learning framework for performing spatiotemporal forecasting (Yu et al., 2018). This framework comprises two spatiotemporal blocks, each containing a spatial graph convolution layer and two temporal gated convolution layers. The output from the spatiotemporal blocks is modeled with a fully connected dense layer to generate the required forecasts. Since this model combines multiple convolutional layers, it allows for faster training with fewer parameters.

- **Space-Time DeepKriging** (DeepKriging) model is a distribution-free spatiotemporal modeling architecture that is well-suited for handling non-Gaussian and non-stationary processes (Nag et al., 2023). The DeepKriging framework follows a two-step workflow, where radial basis functions and Gaussian kernels capture spatial and temporal trends, respectively. These spatiotemporal basis functions encode the coordinates, enabling effective spatiotemporal interpolation. In the second stage, convolutional LSTM networks are employed to generate forecasts based on the previously learned embeddings, allowing for more accurate spatiotemporal predictions.

To implement the STARMA, GSTAR, GpGp, and STNN models, we utilize the *starma*, *gstar*, *GpGp*, and *TDSTNN* packages in the R statistical software, respectively. We adopt the available implementation provided in the STGCN and DeepKriging model in Yu et al. (2018) and Nag et al. (2023), respectively.

S.3. MCB Plots

The MCB test results for $PM_{2.5}$, PM_{10}, and NO_2 pollutants are summarized in Fig. S.1. The results consider MAE, MASE, and SMAPE evaluation metrics, as discussed in Section 4.3 of the main manuscript. For $PM_{2.5}$, the MCB test results show that the E-STGCN framework achieves the lowest mean rank with values of 2.91 (MAE), 2.48 (SMAPE), and 3.14 (MASE), followed by the STGCN, NBeats, ARIMA, and GSTAR models. The critical distance values of the remaining baseline models lie above the reference value (shaded region), indicating that their performance is significantly worse than the 'best-fitted' E-STGCN model. For PM_{10}, the E-STGCN framework consistently ranks as the 'best' model across all performance indicators, followed by STGCN, NBeats, ARIMA, and GpGp. The performance of the other models is significantly inferior compared to the E-STGCN framework. In the case of forecasting NO_2 levels, the E-STGCN and STGCN frameworks achieved similar rankings, emerging as the 'best' models across all metrics except MASE, where STARMA performed best. Among the remaining models, NBeats, ARIMA, GSTAR, and GpGp performed significantly better than the other approaches.
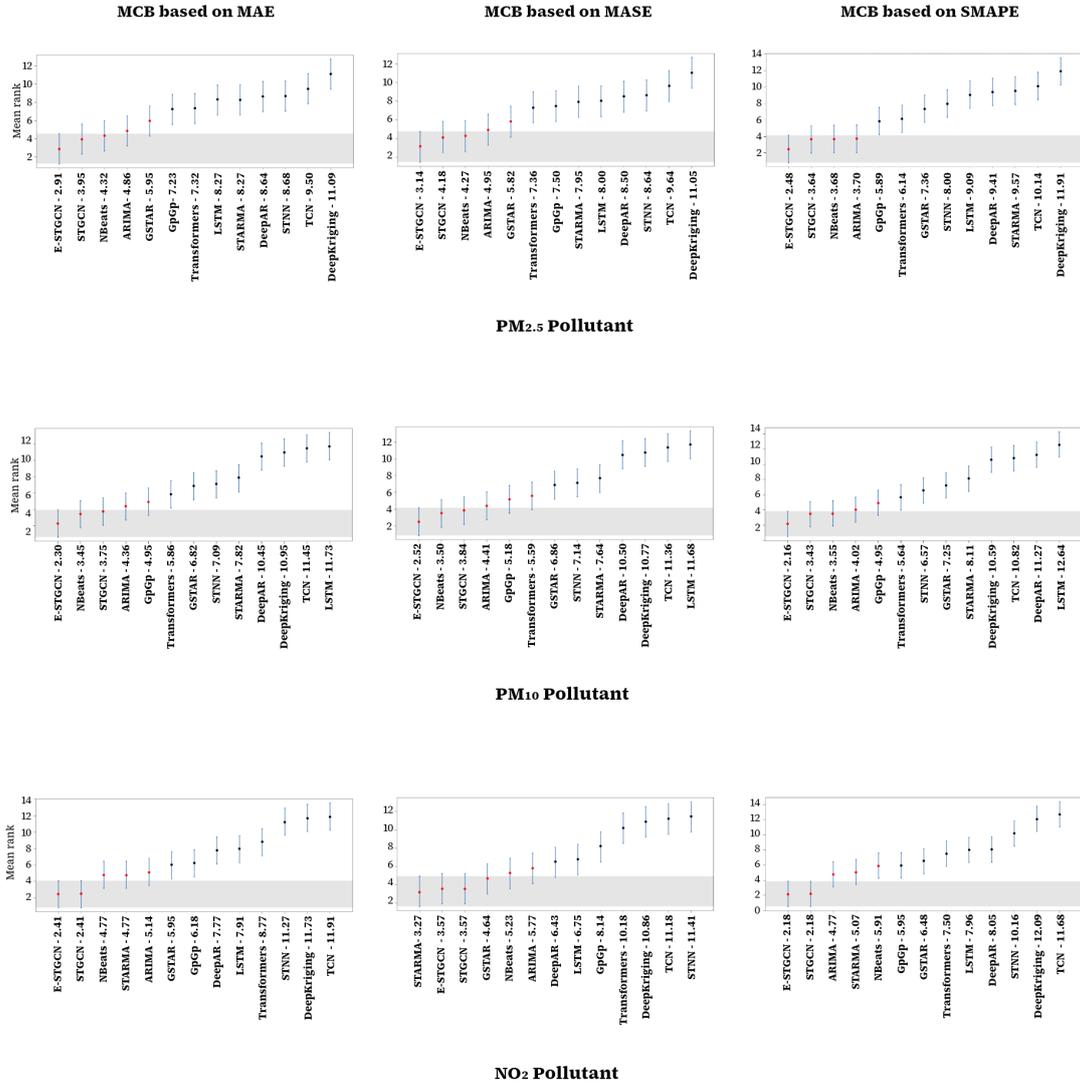
Figure S.1: MCB plot for forecasting pollutant concentrations based on different performance metrics. In the figure, for example, 'E-STGCN - 2.91' means that the average rank of the proposed E-STGCN algorithm for $PM_{2.5}$ forecasting, based on the MAE error metric, is 2.91; the same explanation applies to other algorithms, metrics, and pollutants. The shaded region depicts the reference value of the test.