Causal Learning Meet Covariates: Empowering Lightweight and Effective Nationwide Air Quality Forecasting

Jiaming Ma[1], **Zhiqing Cui**[3], **Binwu Wang**[1,2,*], **Pengkun Wang**[1,2], **Zhengyang Zhou**[1,2], **Zhe Zhao**[1], **Yang Wang**[1,2,*]
[1]University of Science and Technology of China (USTC), Hefei, China
[2]Suzhou Institute for Advanced Research, USTC, Suzhou, China
[3]Nanjing University of Information Science and Technology
JiamingMa@mail.ustc.edu.cn,202383090003@nuist.edu.cn
{wbw2024,pengkun,zzy0929}@ustc.edu.cn, zz4543@mail.ustc.edu.cn, angyan@ustc.edu.cn

Abstract

Air quality prediction plays a crucial role in the development of smart cities, garnering significant attention from both academia and industry. Current air quality prediction models encounter two major limitations: their high computational complexity limits scalability to nationwide datasets, and they often regard weather covariates as optional auxiliary information. In reality, weather covariates can have a substantial impact on air quality indices (AQI), exhibiting a significant causal association. In this paper, we first present a nationwide air quality dataset to address the lack of open-source, large-scale datasets in this field. Then we propose a causal learning model, CauAir, for air quality prediction that harnesses the powerful representation capabilities of the Transformer to explicitly model the causal association between weather covariates and AQI. To address the high complexity of traditional Transformers, we design CachLormer, which features two key innovations: a simplified architecture with redundant components removed, and a cache-attention mechanism that employs learnable embeddings for perceiving causal association between AQI and weather covariates in a coarse-grained perspective. We use information theory to illustrate the superiority of the proposed model. Finally, experimental results on three datasets with 28 as the baseline demonstrate that our model achieves competitive performance, while maintaining high training efficiency and low memory consumption. The source code is available at CauAir Official Repository.

1 Introduction

Air quality prediction, a fundamental task in smart cities, plays a vital role in pollution control and public health protection [Liang *et al.*, 2023]. In recent years, spatiotemporal graph convolutional networks have become the predominant approach to this task. These models incorporate graph inductive bias by employing GCNs or Transformers as spatial modules to capture correlations among monitoring stations, while using temporal modules such as TCNs or LSTMs to model the evolution of AQI over time. Despite their success in advancing air quality forecasting, these models still face three key challenges:

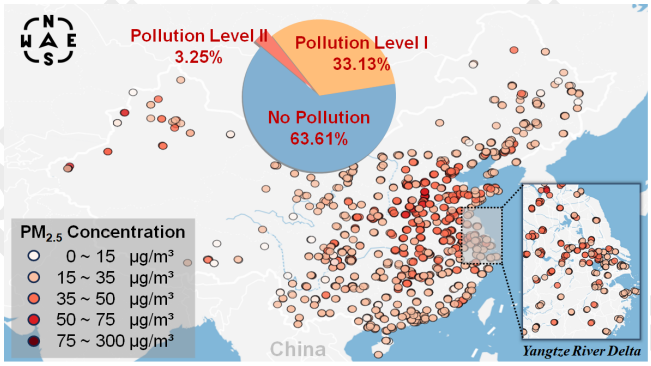

Figure 1: Distribution of nationwide quality monitor stations in China and imbalanced distribution of three pollution levels.

Scalability on the large-scale air quality datasets. While existing models have significantly improved accuracy, their success has been primarily demonstrated on small-scale air quality datasets. A potential reason for this limitation is the lack of large-scale open-source datasets, particularly for nationwide applications, as illustrated in Figure 1. The high computational complexity of these models hampers their practicality [Shao *et al.*, 2022b; Liang *et al.*, 2023]. This challenge is particularly pronounced for Transformer-based models, which currently represent the most advanced architecture. Their time complexity increases quadratically with the number of monitoring stations, complicating their scalability. Furthermore, Transformers integrate attention mechanisms and MLP sub-blocks with skip connections and normalization layers. This intricate architecture leads to high memory requirements [Wang *et al.*, 2024c].

Imbalance pollution level distribution. Researchers may tolerate minor numerical errors in predictions but be strict for

*Yang Wang and Binwu Wang are corresponding authors.

misclassifications of pollution levels, especially for severe pollution level cases. Taking $PM_{2.5}$ as an example, meteorologists classify $PM_{2.5}$ pollution levels into three categories based on current concentration per cubic meter: No Pollution Level ($\leq$ 35 μg/m³), Pollution Level I (35-75 μg/m³), and Pollution Level II ($\geq$ 75 μg/m³). In Figure 2, we reveal an imbalanced air quality pollution level distribution, with Pollution Level II being extremely rare. Deep learning models typically struggle with learning from tail data. These severe level-two pollution events pose serious threats to human health and should be emphasized in air quality prediction.

Causal modeling of AQI and weather conditions. Weather conditions are crucial covariates that significantly influence AQI, i.e., causal association. For example, rainy and typhoon weather typically leads to substantial decreases of $PM_{2.5}$, and severe $PM_{2.5}$ pollution events typically occur under stagnant weather conditions [Zhang *et al.*, 2017]. Weather covariates should serve as valuable resources for improving prediction accuracy. Unfortunately, existing models predominantly concentrate on the spatiotemporal dynamics of AQI, treating weather covariates merely as optional auxiliary information. Most models just employ shallow neural networks to encode weather covariates, subsequently concatenating the resulting high-dimensional embeddings with the final AQI representations before passing the output to the decoder for future predictions. This simplistic approach fails to explicitly capture the causal associations between weather covariates and AQI.

To address the challenges outlined above, we first release LargeAQ, a nationwide air quality dataset collected from 1,341 monitoring stations. This dataset offers comprehensive coverage across 33 major administrative regions in China. In contrast to the commonly used and open-source air quality datasets, detailed in Table 1, LargeAQ is distinguished by its large-scale and long-term coverage. This extensive dataset is anticipated to facilitate significant advancements in air quality prediction.

Furthermore, we propose a causal learning model for nationwide air quality prediction called CauAir, which explicitly models the causal association between AQI and weather covariates to improve the model's predictive performance and its ability to forecast severe pollution levels in the tail distribution. Specifically, CauAir first uses a channel mixing module to separately model spatiotemporal dynamics within AQI and covariates. We propose using a Transformer architecture to model causal association between AQI and covariates. Considering the trade-off between performance and efficiency, we propose a **Cach**e-based **L**ightweight Transf**ormer**, termed CachLormer. This model is a parallelizable adaptation of the vanilla Transformer, achieved by simplifying its complex architecture through the removal of skip connections and normalization layers. CachLormer also employs a cache-attention mechanism with learnable caches to perceive causal association between covariates and AQI. By implementing feature interactions at a coarse granularity, this approach enables robust recognition of causal patterns across heterogeneous data. We also provide a theoretical explanation for the advantages of our model. Our contributions are summarized as follows:

- We introduce LargeAQ, a nationwide large-scale air quality dataset. We anticipate that this pioneering work will create promising avenues for the advancement of air quality prediction techniques.
- We propose CauAir, a lightweight and effective model that explicitly captures causal associations between AQI and weather covariates to achieve notable performance improvements, supported by mutual information theory.
- Experimental results on three datasets with 28 baselines demonstrate that CauAir achieves competitive performance with high efficiency and low memory usage.

Dataset	# Sites	# Months	# Steps	Granularity	Volume	Accessibility
China sites [Yu *et al.*, 2025]	1,200	82	59,720	1 h		
KnowAir [Wang *et al.*, 2020]	184	48	11,688	3 h	76.0M	✓
CCAQ [Chen *et al.*, 2023]	209	28	20,373	1 h	71.7M	✓
LargeAQ (Ours)	**1,341**	**96**	70,128	1 h	1.03 **B**	✓

Table 1: Comparison of commonly used air quality datasets. M: million (10^6). B: billion (10^9).

2 Preliminary

2.1 Task Statement

Let $\mathbf{X}_t \in \mathbb{R}^{N\times c}$ to denote the AQI of N air quality monitoring stations at a given time step t, where c is the number of air pollutants measurements (e.g., $PM_{2.5}$, PM_{10}, or NO_2). We denote $\mathbf{Z}_t \in \mathbb{R}^{N\times f}$ as the weather covariates collected from the vicinity of each station at time step t, where f represents the number of covariates, such as temperature and wind speed. $\mathbf{Z}^p$ and $\mathbf{Z}^f$ refer to the covariates *observed* in the past and those *predicted* for the future, respectively.

Given observed AQI of all stations from the past T time steps $\mathbf{X}_{t-T+1:t} \in \mathbb{R}^{T\times N\times c}$, observed covariates $\mathbf{Z}^p_{t-T+1:t} \in \mathbb{R}^{T\times N\times f}$, and predicted covariates $\mathbf{Z}^f_{t+1:t+L-1} \in \mathbb{R}^{L\times N\times f}$, we aim to learn a function $\mathcal{F}$ that can predict future AQI over next L time steps $\mathbf{Y}_{t+1:t+L-1} \in \mathbb{R}^{L\times N\times c}$: $\left[\mathbf{X}_{t-T+1:t}, \mathbf{Z}^p_{t-T+1:t}, \mathbf{Z}^f_{t+1:t+L-1}\right] \xrightarrow{\mathcal{F}(\cdot)} \mathbf{Y}_{t+1:t+L-1}$.

2.2 Transformer

Transformer has demonstrated its impressive representation capabilities in various fields. The learning process of the vanilla transformer is expressed as follows:

$$\begin{aligned} \mathbf{H}_{\text{hid}} &= \text{Norm}\left(\text{MH-Attn}\left(\mathbf{H}_i\right) + \mathbf{H}_i\right), \\ \mathbf{H}_{\text{o}} &= \text{Norm}\left(\text{FFN}\left(\mathbf{H}_{\text{hid}}\right) + \mathbf{H}_{\text{hid}}\right), \end{aligned} \tag{1}$$

where $\mathbf{H}_i$ and $\mathbf{H}_o$ are the input and output of the transformer, respectively. $\text{FFN}(\cdot)$ uses two-layer MLP with ReLU activation function, and $\text{Norm}(\cdot)$ is the normalization operation. The core of Transformer is the self-attention mechanism, which can be expressed as:

$$\text{Attn}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}, \tag{2}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ represent query, key, and value vectors, respectively, and they come from the mapping transformation of the input. To improve the representation ability, Transformer uses the multi-head self-attention mechanism.

3 Methodology

In this section, we provide the details of the proposed CauAir framework. CauAir initially models the dynamics of the AQI and the covariates independently. Subsequently, we design a lightweight Transformer to explicitly model the causal associations between AQI and the covariates. The overview of CauAir is shown in Figure 2.

3.1 Channel Mixing for Spatiotemporal Learning

We adopt a spatiotemporal learning module to model the spatiotemporal dynamics within both AQI and weather covariates. This module operates as follows: for any input tensor $\mathbf{P} \in \{\mathbf{X}_{t-T+1:t}, \mathbf{Z}^p_{t-T+1:t}, \mathbf{Z}^f_{t+1:t+L-1}\}$, the module first compresses both temporal and feature dimensions to generate node-level representations $\mathbf{H} \in \mathbb{R}^{N \times d_i}$. For example, if $\mathbf{X} \in \mathbb{R}^{T \times N \times c}$ is used as input, d_i of the output would be equal to $T \times c$. The output $\mathbf{H}$ is then fed into a feed-forward network as an encoder with positional embeddings to model spatiotemporal dynamics through mixing feature channels. The forward process of our spatiotemporal learning module is expressed as follows,

$$\begin{aligned} \widetilde{\mathbf{H}} &= \mathrm{FFN}(\mathbf{H}) + \mathbf{E} \in \mathbb{R}^{N \times d_h}, \\ \mathrm{Reshape}\left(\mathbf{P}\right) &\rightarrow \mathbf{H} \in \mathbb{R}^{N \times d_i}, \end{aligned} \tag{3}$$

where $\widetilde{\mathbf{H}}$ denotes the output. $\mathbf{E} \in \mathbb{R}^{N \times d_h}$ is the learnable position embedding to capture high-level features of stations. Here $\mathrm{FFN}\left(\cdot\right)$ includes two-layer MLP with SwiGLU [Shazeer, 2020] as activation function:

$$\begin{aligned} \mathrm{FFN}(\mathbf{H}) &= \left(\mathrm{SiLU}\left(\mathbf{H}W_1\right) \odot \mathbf{H}W_2\right) W_3, \\ \mathrm{SiLU}\left(\mathbf{H}W_1\right) &= \mathbf{H}W_1 \odot \sigma\left(\mathbf{H}W_1\right), \end{aligned} \tag{4}$$

where $W_1, W_2 \in \mathbb{R}^{d_i \times e * d_i}$ and $W_3 \in \mathbb{R}^{e * d_i \times d_h}$ are learnable parameters. $e = 4$ is the expanding coefficient of hidden representation. $\sigma\left(\cdot\right)$ represents the sigmoid function and $\odot$ represents the Hadamard product.

We employ two parameter-independent spatiotemporal learning modules to model the dynamic of the past AQI $\mathbf{X}_{t-T+1:t}$ and historical covariates $\mathbf{Z}^P_{t-T+1:t}$, and the outputs are denoted as $\overline{\mathbf{X}} \in \mathbb{R}^{N \times d_h}$ and $\overline{\mathbf{Z}} \in \mathbb{R}^{N \times d_h}$, respectively.

3.2 Efficient CachLormer for Causal Learning

Weather covariates, such as precipitation and wind speed, play a crucial role in influencing AQI. To explicitly capture these causal associations, we propose employing a Transformer as the backbone of our model, leveraging its strong capability to model dependencies between variables effectively. To mitigate the computational complexity associated with traditional Transformers, we introduce a **Cach**e-based **L**ightweight Transf**ormer** (CachLormer). Its efficiency improvements arise from two key modifications: a simplified architecture and an efficient cache-attention mechanism.

Simplified Architecture. In Equation 1 of a standard Transformer architecture, the structure typically consists of interleaved attention and feedforward (MLP) sub-blocks, designed with the skip connection and normalization layer. In fact, such complex designs are not strictly necessary. Inspired by [He and Hofmann, 2023], we also reformulate the complex structure into a parallelizable version by removing the skip connection and normalization layer, which is shown in Figure 2,

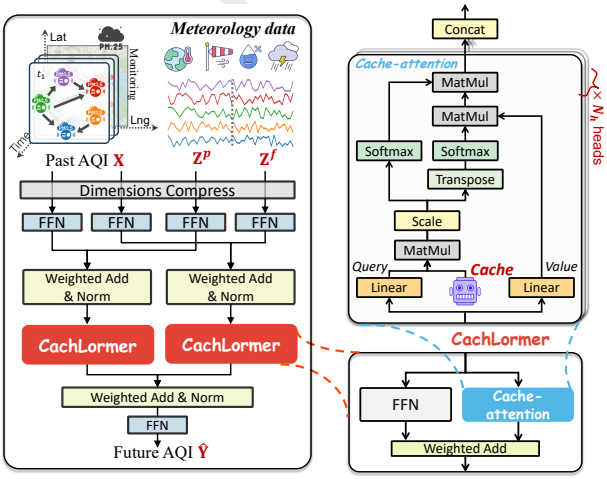

Figure 2: Overview of the proposed CauAir.

$$\begin{aligned} &\mathrm{CachLormer}\left(\mathbf{H}_i\right) \\ &= \delta_0 * \mathrm{MH\text{-}Cachattn}\left(\mathbf{H}_i\right) + \delta_1 * \mathrm{FFN}\left(\mathbf{H}_i\right), \end{aligned} \tag{5}$$

where δ_0 and $\delta_1 \in (0, 1)$ are parameters that control the mixing ratio of two outputs. The feed-forward network here includes a two-layer MLP with SwiGLU activation function. $\mathrm{MH\text{-}Cachattn}\left(\cdot\right)$ represents our design cache-attention, which will be explained in the following section. The elimination of unnecessary components reduces the memory consumption of Transformer.

Cache-attention Mechanism. The traditional self-attention mechanism calculates the attention coefficient by taking the dot product of the key vector $\mathbf{K} \in \mathbb{R}^{N \times d_h}$ and the query vector $\mathbf{Q} \in \mathbb{R}^{N \times d_h}$, resulting in a quadratic complexity of $\mathcal{O}\left(N^2 d_h\right)$. This dense complexity limits the model's scalability on large-scale datasets. To address this issue, we propose cache-attention mechanism, where the underlying motivation for this approach stems from the urban hierarchy theory in the spatiotemporal data analysis field [Guo *et al.*, 2021]: certain fine-grained nodes may exhibit similar spatiotemporal features, allowing us to coarsen these fine-grained nodes into P coarse-regions. Thus, we first randomly initialize a set of caches $\mathbf{E}_p = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_P] \in \mathbb{R}^{P \times d_h}$, which is a learnable parameterized embedding vectors. The hyperparameter P is the number of caches with $P\,(\ll N)$ and d_h is the number of channels. And the embedding vector of k-th ($k \in \{1, \ldots, P\}$) cache $\mathbf{e}_k \in \mathbb{R}^{d_h}$ represents the contextual features of k-th coarse-region. Fine-grained modeling may prevent the model from perceiving robust causal association between two heterogeneous data types, covariates and AQI, while a coarse-grained perspective can alleviate this issue.

Given a query matrix $\mathbf{Q} \in \mathbb{R}^{N \times d_h}$, where $\mathbf{q}_i$ represents the feature embedding of node v_i, we employ a soft assignment mechanism to determine the association between nodes and

caches. This mechanism achieves lossless attention capacity through weighted combinations of attention scores between nodes and cache elements. Specifically, the affinity between node v_i and k-th cache can be computed as:

$$\mathbf{s}_{i,k} = \frac{\mathbf{q}_i \mathbf{e}_k^\top}{\sqrt{d_h}}. \tag{6}$$

Then we obtain two similarities by normalizing the node dimension, and the cache dimension, respectively, with the above calculated assignment weights as follows,

$$\mathbf{s}_{i,k}^{(1)} = \frac{e^{\mathbf{s}_{i,k}}}{\sum_{w=1}^{P} e^{\mathbf{s}_{i,w}}}, \mathbf{s}_{i,k}^{(2)} = \frac{e^{\mathbf{s}_{i,k}}}{\sum_{j=1}^{N} e^{\mathbf{s}_{j,k}}} \in (0,1), \tag{7}$$

where $\mathbf{s}_{:,k}^{(1)}$ records the affinity of each node to k-th cache, and $\mathbf{s}_{i,:}^{(2)}$ represents the affinity of each cache to node v_i. We can indirectly calculate the similarity $a_{i,j}$ between nodes v_i and v_j as follows,

$$a_{i,j} = \sum_{k=1}^{P} \mathbf{s}_{i,k}^{(1)} \mathbf{s}_{j,k}^{(2)} \in (0,1). \tag{8}$$

The similarity calculated by this method inherits normalized properties: for any node v_i, the attention scores from all nodes to v_i sum to 1, i.e.,

$$\sum_{j=1}^{N} a_{i,j} = \sum_{j=1}^{N} \sum_{k=1}^{P} \mathbf{s}_{i,k}^{(1)} \mathbf{s}_{j,k}^{(2)} = \sum_{k=1}^{P} \mathbf{s}_{i,k}^{(1)} \sum_{j=1}^{N} \mathbf{s}_{j,k}^{(2)} = 1. \tag{9}$$

With the cache-attention calculation process from Equation $(6) \sim (8)$, the forward process of the cache-attention can be expressed as follows,

$$\begin{aligned} &\text{Cache-attention}\left(\mathbf{Q}, \mathbf{E}_p, \mathbf{V}\right) \\ &= \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{E}_p^\top}{\sqrt{d_h}}\right) \text{Softmax}\left(\frac{\mathbf{E}_p \mathbf{Q}^\top}{\sqrt{d_h}}\right) \mathbf{V}. \end{aligned} \tag{10}$$

where the query vector $\mathbf{Q} = \mathbf{H}_i W_q$ and the value vector $\mathbf{V} = \mathbf{H}_i W_v$. And $W_q, W_v \in \mathbf{R}^{d_h \times d_h}$ and $\mathbf{E}_p$ are learnable parameters. Our model also uses a multi-head attention mechanism:

$$\begin{aligned} &\text{MH-Cachattn}\left(\mathbf{H}_i\right) = \left[\text{head}_1, \ldots, \text{head}_{N_h}\right] W_o, \\ &\text{head}_w = \text{Cache-attention}\left(\mathbf{H}_i W_q^{(w)}, \mathbf{E}_p^{(w)}, \mathbf{H}_i W_v^{(w)}\right), \end{aligned} \tag{11}$$

where $W_q^{(w)}, W_v^{(w)} \in \mathbb{R}^{d_h \times \frac{d_h}{N_h}}, \mathbf{E}_p^{(w)} \in \mathbb{R}^{P \times \frac{d_h}{N_h}}$ are the w-th head part of W_q, W_v and $\mathbf{E}_p$, respectively. $W_o \in \mathbb{R}^{d_h \times d_h}$ is a learnable parameter to integrate information from N_h heads.

Computational complexity analysis. Compared with the $\mathcal{O}\left(N^2 d_h\right)$ complexity of the self-attention mechanism in Transformer, our cache-attention mechanism is reduced to $\mathcal{O}\left(PN d_h\right)$, where P is be much smaller than N.

Causal Learning between AQI and Covariates

We first deploy a CachLormer, which is denoted as f_0, to model the causal dependencies between past AQI $\overline{\mathbf{X}}$ and past weather covariates $\overline{\mathbf{Z}}$. Specifically, we adopt the addition strategy with $\mathrm{RMSNorm}$ normalization strategy [Zhang and Sennrich, 2019] to deeply fuse AQI and covariate information by addition, compared to independent channel concatenation strategies, the model can analyze deeper causal associations from entangled representations:

$$\tilde{\mathbf{H}}_0' = \text{Norm}\left(\beta_0 \overline{\mathbf{X}} + \gamma_0 \overline{\mathbf{Z}}\right), \tag{12}$$

where β_0 and γ_0 are learnable parameters used to balance the two terms. In order to integrate future weather covariates. Specifically, we first use two parameter-independent spatiotemporal learning modules to capture the dynamics of $\mathbf{X}_{t-T+1:t}$ and $\mathbf{Z}_{t+1:t+L-1}^f$, and the outputs $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Z}}$ are the future AQI representation and the covariate representation, respectively. Then we use a CachLormer module, denoted as f_1, to learn the causal associations between $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Z}}$:

$$\tilde{\mathbf{H}}_1' = \text{Norm}\left(\beta_1 \widetilde{\mathbf{X}} + \gamma_1 \widetilde{\mathbf{Z}}\right). \tag{13}$$

Finally, we add the two components together and feed them into the decoder (e.g., FFN) to predict the future AQI:

$$\hat{\mathbf{Y}} = \text{FFN}\left(\alpha_0 f_0(\tilde{\mathbf{H}}_0') + \alpha_1 f_1(\tilde{\mathbf{H}}_1')\right), \tag{14}$$

where α_0 and α_1 are learnable parameters that weigh the decision strengths of the two components.

3.3 Theoretical Explanation

Theory 1. Equivalence between MSE and mutual information (MI). Minimizing MSE between the predicted values $\hat{\mathbf{Y}}$ and the ground-truth values $\mathbf{Y}$ can be equivalent to maximize the mutual information of $\hat{\mathbf{Y}}$ and $\mathbf{Y}$, $\max I\left(\hat{\mathbf{Y}}, \mathbf{Y}\right)$, in the regression tasks [Jing *et al.*, 2022]:

$$\min \mathbb{E}\left\|\mathbf{Y} - \hat{\mathbf{Y}}\right\|_2^2 \Longleftrightarrow \max I\left(\hat{\mathbf{Y}}; \mathbf{Y}\right). \tag{15}$$

Next, we demonstrate the superiority of our model in modeling causal association between AQI and weather covariates from the mutual information theory.

Theory 2. MI Monotonicity. For any random variables A, B, C, and D, the following inequality of the mutual information $I(\cdot;\cdot)$ holds [Peng *et al.*, 2020]:

$$I(A;B) \geq I(A;C,D) \geq I(A;C). \tag{16}$$

According to Theory 1, our goal is to maximize $I(\hat{\mathbf{Y}};\mathbf{Y})$. Given the fact that weather covariates have significant causal association with air quality, we incorporate these covariates $\mathbf{Z}$ as auxiliary information into the learning process: the learned representations of historical weather covariates and past AQI are input into CachLormer to effectively model the causal association. The learned knowledge is then propagated to future weather covariates to make accurate predictions. Thus, according to Equation 16, we have $I(\hat{\mathbf{Y}};\mathbf{Y}) \geq I\left(\mathbf{Y};\mathbf{X}_{t-T+1:t},\mathbf{Z}\right) \geq I\left(\mathbf{Y};\mathbf{X}_{t-T+1:t}\right)$, indicating that our model has a higher lower bound on the mutual information (MI) between true values $\mathbf{Y}$ and predicted values $\hat{\mathbf{Y}}$. Based on the equivalence between MSE and MI (Theorem 1), we can conclude that weather covariates enhance the upper bound of mutual information, enabling the model to make

more accurate and stable predictions by reducing predictive uncertainty. We further demonstrate the benefits of weather covariates through ablation experiments (as shown in Section 4.5). Compared to approaches that encode weather variables through shallow networks, our model can capture causal association more precisely.

4 Experiments

4.1 Nationwide Air Quality Dataset

Data Description. Due to the lack of open-source large-scale datasets for air quality prediction, we open-source a comprehensive, long-term air quality dataset named LargeAQ. It contains AQI, including $PM_{2.5}$, PM_{10}, NO_2, and so on, from 1,341 monitoring stations across 333 major Chinese cities. We concentrate on $PM_{2.5}$ as the primary pollutant of interest. The dataset spans 8 years from January 1, 2016, to December 31, 2023, with hour granularity. The data are sourced from the China National Environmental Monitoring Center (CNEMC)[1], an online platform that provides real-time AQI for prefecture-level administrative regions across China. We also collect five weather features during this period from the Chinese Weather Website as covariates: temperature, humidity, precipitation, wind shear, and wind speed. A comparison between LargeAQ and several popular air quality datasets is shown in Table 1. In the experiments, we also use two other open-source datasets.

Data Analysis. In weather science, $PM_{2.5}$ pollution is categorized into three levels: No Pollution Level ($\leq$ 35 μg/m³), Pollution Level I (35-75 μg/m³), and Pollution Level II ($\geq$ 75 μg/m³). Figure 3 illustrates the $PM_{2.5}$ distribution across these categories in LargeAQ. We observe a significant imbalance in the distribution of the three pollution levels in the LargeAQ dataset-severe Pollution Level II is rare.

Data Processing. To address missing values in the AQI and weather data, we use the last observed value method. Using the sliding window to process data, the length of both past window and future window are equal to 24. For input features - including AQI, historical covariates, and future covariates - we apply standard normalization to facilitate stable learning. Only the predicted AQI values are denormalized back to their original distribution. For the future weather covariate $\mathbf{Z}^f$, to avoid information leakage caused by using the actual values from our collected weather data, we simulate future weather forecasts by adding synthetic noise. Following common practices in linear regression [Murphy, 2012; Jing *et al.*, 2022], which assume that the predicted values deviate from ground truth by additive Gaussian noise, we generate simulated forecasts by injecting Gaussian noise into the observed weather data. In addition, we also report the performance of models that do not incorporate future weather features, to provide a fair comparison.

4.2 Experimental Setup

Baselines for Comparison

We compare our CauAir with the **28** baselines that belong to the following three categories:

[1]https://www.cnemc.cn/

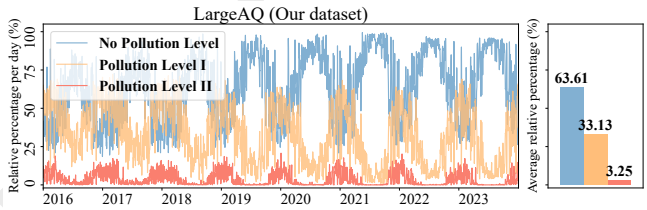

Figure 3: Different $PM_{2.5}$ pollution level relative percentage (%) on each day (left) and average relative percentage (%) in every forecasting window (right) on LargeAQ datasets.

Time-series models: CATS [Lu *et al.*, 2024], CycleNet [Lin *et al.*, 2024a], DLinear [Zeng *et al.*, 2023], DSformer [Yu *et al.*, 2023], SOFTS [Han *et al.*, 2024b], SparseTSF [Lin *et al.*, 2024b], TimeMixer [Wang *et al.*, 2024d], CrossGNN [Huang *et al.*, 2023], Umixer [Ma *et al.*, 2024], and TimeXer [Wang *et al.*, 2024e].

Spatiotemporal forecasting models: AGCRN [Bai *et al.*, 2020], ASTGCN [Guo *et al.*, 2019], BigST [Han *et al.*, 2024a], D^2STGNN [Shao *et al.*, 2022b], GWNet [Wu *et al.*, 2019], STAEformer [Liu *et al.*, 2023], STGCN [Yu *et al.*, 2017], STGODE [Fang *et al.*, 2021], STID [Shao *et al.*, 2022a], STNorm [Deng *et al.*, 2021], STTN [Xu *et al.*, 2020], and RPMixer [Yeh *et al.*, 2024].

Air quality prediction models: AirFormer [Liang *et al.*, 2023], AirPhyNet [Hettige *et al.*, 2024], DeepAir [Yi *et al.*, 2018], GAGNN [Chen *et al.*, 2023], MGSFformer [Yu *et al.*, 2025], and $PM_{2.5}$GNN [Wang *et al.*, 2020].

Implementation Detail

The models are implemented in PyTorch 2.2.0 running on an NVIDIA A100 GPU with 40 GB memory. We adopt Adam optimizer accompanied MultiStepLR learning rate adjustment strategy with a learning rate 0.02 and weight decay 0.004. The cache number P is 32 for LargeAQ, 108 for CCAQ and 10 for KnowAir. The channel dimension of CachLormer is set to 128. We evaluate all models on numerical regression prediction tasks and pollution level category prediction tasks, using three commonly employed metrics: MAE, RMSE, and MAPE for regression prediction and F1 score for pollution level prediction task. All baselines incorporate weather covariates (if available).

4.3 Future $PM_{2.5}$ Forecasting Comparison

We report the performance metrics over 24 horizon forecasts for LargeAQ dataset in Table 2.

Time-series forecasting models typically underperform compared to spatiotemporal learning models, primarily because they focus solely on temporal dependencies while neglecting inter-station correlations. Within the spatiotemporal learning family, Transformer-based models demonstrate superior performance over GCN-based approaches. For instance, D^2STGNN effectively captures $PM_{2.5}$ spatiotemporal dynamics through its decoupling mechanism. Among specific air quality prediction models, DeepAir achieves surprisingly good performance despite its relatively simple architecture. This success can be attributed to its specialized strat-

	Method	Horizon 6			Horizon 12			Horizon 24			Pollution Level Classification		
		MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	No-Pol	Level I	Level II
Time-seris	CATS	9.51	14.41	55.16	10.34	15.55	61.08	11.60	17.25	69.83	0.9070	0.6780	0.3978
	CrossGNN	8.62	13.25	49.78	9.94	14.91	59.35	11.00	16.41	67.45	0.9165	0.7151	0.4165
	CycleNet	8.57	13.33	47.13	9.98	15.05	57.79	11.07	16.59	65.38	0.9161	0.7162	0.4366
	DLinear	8.95	13.43	56.44	10.30	14.97	67.95	11.25	16.27	75.73	0.9108	0.6880	0.4106
	DSformer	8.53	13.14	47.50	9.83	14.77	56.44	10.93	16.28	64.56	0.9190	0.7172	0.4721
	SOFTS	8.40	13.12	48.24	9.81	14.86	57.56	10.96	16.42	65.90	0.9172	0.7158	0.4586
	SparseTSF	9.52	14.23	59.93	10.56	15.75	67.27	12.33	18.11	79.79	0.9042	0.6884	0.4393
	TimeMixer	8.53	13.15	51.62	9.88	14.74	63.79	10.84	16.03	73.12	0.9159	0.7181	0.2765
	TimeXer	7.81	12.02	48.23	8.81	13.20	55.91	9.58	14.08	63.83	0.9217	0.7533	0.4702
	Umixer	8.65	13.19	53.35	9.93	14.75	63.99	10.88	16.01	73.22	0.9153	0.7166	0.3565
Spatioteporal	AGCRN	7.79	11.99	43.16	8.69	13.02	49.58	9.55	14.06	57.94	0.8985	0.5868	0.2820
	ASTGCN	8.33	12.48	47.44	9.34	13.67	53.86	10.28	14.80	64.44	0.9119	0.6829	0.4754
	BigST	8.27	12.48	46.27	9.31	13.55	53.92	10.14	14.58	60.73	0.9123	0.7056	0.4882
	D^{2}STGNN	7.73	11.89	43.05	8.43	12.61	48.18	9.36	14.03	56.13	0.9021	0.6771	0.3432
	RPMixer	8.45	12.83	46.08	9.70	14.32	52.85	10.86	15.82	60.97	0.9216	0.7252	0.5004
	STAEformer	7.52	11.59	46.54	8.50	12.79	51.33	9.42	14.07	57.48	0.9219	0.7590	0.5096
	GWNet	7.75	12.12	44.60	8.68	13.32	49.52	9.39	14.46	60.58	0.9232	0.7576	0.5090
	STGCN	7.77	11.98	43.85	8.66	13.07	51.22	9.54	14.71	58.21	0.8295	0.5802	0.3321
	STGODE	7.82	12.12	44.98	8.74	13.18	51.57	9.72	14.46	60.81	0.8478	0.5221	0.3408
	STID	7.76	12.06	46.17	8.91	13.38	55.48	10.02	14.70	66.84	0.8272	0.6596	0.4750
	STNorm	7.93	12.39	43.47	8.96	13.44	52.43	9.97	14.62	62.91	0.8620	0.6219	0.4275
	STTN	8.11	12.38	46.90	9.12	13.50	53.64	9.97	14.54	62.05	0.9164	0.7150	0.4224
Air Quality	AirFormer	7.95	12.39	46.16	9.04	13.67	53.89	10.10	15.02	64.70	0.8729	0.6833	0.4617
	AirPhyNet	8.77	13.41	54.45	10.05	15.00	63.57	10.98	16.30	70.87	0.9152	0.7114	0.3259
	DeepAir	7.86	12.31	44.23	8.94	13.58	52.40	9.69	14.57	57.69	0.5265	0.4048	0.2955
	GAGNN	10.27	15.39	56.19	11.25	16.40	66.23	11.54	18.81	69.59	0.8963	0.3489	0.1974
	MGSFformer	9.75	14.66	56.27	10.65	15.91	62.24	11.85	17.57	70.13	0.9047	0.6664	0.3874
	PM$_{2.5}$GNN	8.94	13.32	54.01	10.04	14.52	61.54	10.49	15.06	65.33	0.9138	0.6636	0.4659
	Ours without FW	**7.31**	**11.47**	**40.66**	**8.38**	**12.43**	**46.50**	**9.21**	**13.44**	**52.57**	**0.9369**	**0.7824**	**0.5649**
	Ours	**7.23**	**11.29**	**40.53**	**8.11**	**12.26**	**46.13**	**8.74**	**12.96**	**51.49**	**0.9392**	**0.7922**	**0.5820**

Table 2: Performance comparisons on **LargeAQ**. The **best** and **second best** results are bolded by corresponding colors. The third best result is underlined. 'Ours without FW' means we do not use future weather covariates. All experimental results are the average of five independent runs.

egy for encoding weather variable influences, highlighting the crucial role of weather factors in prediction accuracy.

Our CauAir explicitly models the causal associations between covariates and AQI to enhance prediction performance. As a result, our model achieves state-of-the-art performance across all metrics, showing up to 8.27% improvement in $PM_{2.5}$ value prediction compared to existing methods.

4.4 Efficiency Study

Figure 4 illustrates the efficiency comparison between strong baselines and CauAir on the LargeAQ dataset. Each solid circle represents a model, with the model name, training time per epoch, and MAE reported below. The default batch size is 64; for some complex models (marked with $*$ and their batch size), we have to reduce their batch sizes to enable operation on the large-scale dataset.

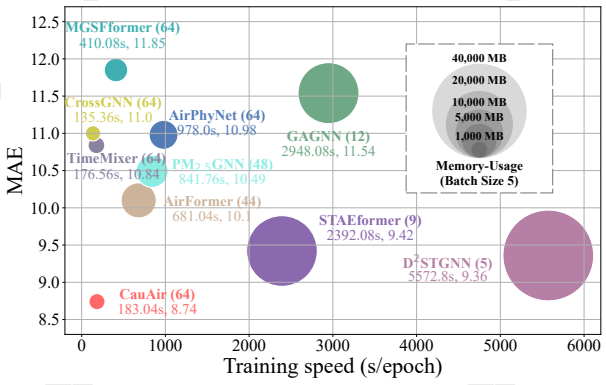

Figure 4: Efficiency study of CauAir on LargeAQ dataset.

We observe that advanced Transformer-based models, such as STAEformer and D^{2}STGNN, suffer from a time complexity that scales quadratically with the number of nodes, due to their reliance on standard self-attention mechanisms. Moreover, their complex architectural designs further increase memory consumption. AirFormer simplifies the Transformer structure, thereby achieving improved efficiency. Our model not only achieves superior predictive performance but also attains the highest computational efficiency, outperforming the SOTA D^{2}STGNN in time and space complexity.

4.5 Ablation study

We conduct ablation experiments on the LargeAQ dataset to verify the effectiveness of each component.

Weather Covariates. We develop the following variants: (1). w/o $\mathbf{Z}^p$: Remove the past weather covariate $\mathbf{Z}^p$ term in input; (2). w/o $\mathbf{Z}^f$: We do not take the future weather covariate $\mathbf{Z}^f$ as input; (3). w/o $\mathbf{Z}^p$ & $\mathbf{Z}^f$: We do not use any weather covariates. The experimental results are shown in Table 3, which shows that both past and future weather covariates are beneficial to improving prediction performance.

Transformer Variant. To evaluate the effectiveness of the proposed CachLormer, we develop the following variants:

Method	Efficiency		Horizon 6			Horizon 12			Horizon 24		
	Time (s)	Memory (MB)	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
w/o $\mathbf{Z}^p$	-	-	7.45	11.54	43.25	8.31	12.45	48.52	8.83	13.02	52.80
w/o$\mathbf{Z}^f$	-	-	7.31	11.47	40.66	8.38	12.43	46.50	9.21	13.44	52.57
w/o $\mathbf{Z}^p$&$\mathbf{Z}^f$	-	-	7.62	11.83	43.75	8.84	13.24	52.79	9.87	14.54	62.22
-Sattn	316	32,538	7.31	11.39	**40.64**	**8.04**	12.48	46.39	8.83	**12.95**	52.82
-VanArc	206	7,756	7.36	11.43	40.79	8.19	12.41	**45.77**	8.81	13.08	**51.70**
+VanTran	321	31,740	**7.08**	**11.04**	41.23	**7.97**	**12.36**	**45.06**	**8.77**	12.98	51.83
Ours	183	7,586	**7.23**	**11.29**	**40.53**	8.11	**12.26**	46.13	**8.74**	**12.96**	**51.49**

Table 3: Ablation study on LargeAQ dataset.

(1). -Sattn: We use standard self-attention mechanism instead of our cache-attention mechanism, using our lightweight architecture; (2). -VanArc: We use standard Transformer architecture in Equation 1 with our cache-attention mechanism; (3). +VanTran: We use standard Transformer architecture to replace CachLormer. We find that CachLormer achieves comparable performance to Transformer while significantly reducing memory consumption and training time (s/epoch).

4.6 Hyperparameter Sensitive Annalysis

We evaluate the sensitivity of the hyperparameter the number of caches P on LargeAQ dataset.As shown in Figure 5, we observe that optimal values of P is equal to 32. A smaller P is insufficient to capture adequate contextual features shared among nodes. When P exceeds these optimal values, performance does not improve significantly, as the model struggles to focus on extracting shared contextual features.

4.7 Performance Analysis on Small-scale Datasets

We evaluate our model against several state-of-the-art approaches using two open-source small-scale datasets: KnowAir and CCAQ. The 24-step average results are presented in Table 4. Our model demonstrates significant performance advantages, with an average improvement of 5.15% and a maximum improvement of 20.55%. These enhancements can be primarily attributed to our explicit modeling of the causal relationships between AQI and weather conditions.

4.8 Related Work

Time series analysis is a fundamental task [Liu *et al.*, 2025a; Liu *et al.*, 2025a; Miao *et al.*, 2024; Huang *et al.*, 2023], which is a specific area of air quality forecasting. Early time series analysis approaches modeled air pollutant emissions and dispersion as dynamic systems using numerical simulations, with notable examples such as the Community Multiscale Air Quality [Byun and Schere, 2006]. However, these methods require extensive theoretical knowledge, carefully selected features and region-specific parameters, making them impractical for real-time air quality monitoring

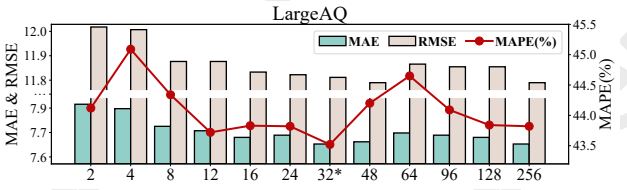

Figure 5: Sensitivity experiments of P on LargeAQ datasets.

Method	CCAQ			KnowAir		
	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
CycleNet	21.30	34.80	32.96	17.28	26.83	56.85
DLinear	21.28	34.62	34.29	17.06	26.20	64.53
SOFTS	21.29	34.69	33.06	17.38	26.84	56.30
TimeMixer	20.84	34.33	32.63	16.78	25.85	65.38
TimeXer	19.64	32.49	31.61	15.70	24.01	58.10
Umixer	21.23	34.58	33.88	17.61	26.87	63.13
AGCRN	19.57	32.65	31.41	16.34	24.81	63.26
BigST	**18.67**	**31.02**	29.37	15.68	24.15	56.52
D^2STGNN	18.82	32.29	**26.30**	15.39	24.31	55.41
GWNet	18.74	31.72	29.11	15.49	23.85	56.73
RPMixer	19.05	32.46	28.91	16.73	25.96	54.07
STAEformer	19.01	31.57	30.34	15.82	24.56	**53.28**
STGCN	19.56	33.34	28.48	15.77	24.25	57.44
STGODE	19.96	33.46	31.10	15.98	25.02	58.00
STID	20.54	34.13	32.86	16.16	24.88	61.41
STTN	19.09	31.83	29.80	15.50	24.08	54.54
AirFormer	20.23	33.16	33.30	16.05	24.70	59.60
AirPhyNet	21.80	35.73	33.57	17.54	26.74	64.58
DeepAir	18.68	31.20	29.43	**14.88**	23.75	55.35
GAGNN	32.91	46.38	41.33	19.40	35.63	71.33
MGSFformer	25.29	40.08	39.44	19.01	29.28	61.75
$PM_{2.5}$GNN	19.75	32.55	28.49	15.10	**22.42**	54.24
Ours	**16.89**	**29.41**	**24.84**	**13.03**	**20.44**	**42.33**

Table 4: Average performance on KnowAir and CCAQ.

systems. Data-driven models emerged with the expansion of urban-scale air quality sensor networks and advances in machine learning algorithms. Recent advances in machine learning, particularly deep learning, have significantly improved prediction accuracy [Liang *et al.*, 2023; Wang *et al.*, 2020]. Researchers have designed cutting-edge models to capture the internal spatiotemporal dynamics of AQI and extract high-dimensional representations. The dominant architecture employs spatiotemporal graph convolutional networks [Miao *et al.*, 2025; Wang *et al.*, 2023; Wang *et al.*, 2024a; Wang *et al.*, 2024b; Ma *et al.*, 2025], utilizing spatiotemporal graphs to represent spatiotemporal data, RNNs or Transformers for temporal dependency modeling, and GNNs or Transformers for spatial dependency extraction [Liu *et al.*, 2024; Shao *et al.*, 2022a; Liu *et al.*, 2025b; Zhang *et al.*, 2025]. These works, however, present some difficulties (e.g., inefficiency) on a nationwide air quality dataset. Moreover, our model differs by explicitly modeling the influence of covariates on air quality, leveraging these additional variables to enhance prediction performance.

5 Conclusion

In this paper, we introduce a nationwide and long-term air quality data from 1,341 monitoring stations. We then propose a lightweight yet effective spatiotemporal causal learning model for air quality prediction. The model employs a novel CachLormer, an efficient Transformer architecture, to explicitly model the causal association between AQI and weather covariates. Compared across three datasets against 28 baselines, our model achieves dominant performance while maintaining high time and memory efficiency.

Contribution Statement

Jiaming Ma and Zhiqing Cui have equal contribution.

Acknowledgements

Thank Xiaolei Wang (homepage: https://quotsoft.net/) for the valuable open-source contribution. And this paper is partially supported by the National Natural Science Foundation of China (No.12227901). The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

References

[Bai *et al.*, 2020] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Proc. of NeurIPS*, 2020.

[Byun and Schere, 2006] Daewon Byun and Kenneth L Schere. Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (cmaq) modeling system. *Applied mechanics reviews*, 2006.

[Chen *et al.*, 2023] Ling Chen, Jiahui Xu, Binqing Wu, and Jianlong Huang. Group-aware graph neural network for nationwide city air quality forecasting. *ACM Transactions on Knowledge Discovery from Data*, 2023.

[Deng *et al.*, 2021] Jinliang Deng, Xiusi Chen, Renhe Jiang, Xuan Song, and Ivor W Tsang. St-norm: Spatial and temporal normalization for multi-variate time series forecasting. In *Proc. of KDD*, 2021.

[Fang *et al.*, 2021] Zheng Fang, Qingqing Long, Guojie Song, and Kunqing Xie. Spatial-temporal graph ode networks for traffic flow forecasting. In *Proc. of KDD*, 2021.

[Guo *et al.*, 2019] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proc. of AAAI*, 2019.

[Guo *et al.*, 2021] Kan Guo, Yongli Hu, Yanfeng Sun, Sean Qian, Junbin Gao, and Baocai Yin. Hierarchical graph convolution network for traffic forecasting. In *Proc. of AAAI*, 2021.

[Han *et al.*, 2024a] Jindong Han, Weijia Zhang, Hao Liu, Tao Tao, Naiqiang Tan, and Hui Xiong. Bigst: Linear complexity spatio-temporal graph neural network for traffic forecasting on large-scale road networks. *Proceedings of the VLDB Endowment*, 2024.

[Han *et al.*, 2024b] Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. Softs: Efficient multivariate time series forecasting with series-core fusion. *arXiv preprint arXiv:2404.14197*, 2024.

[He and Hofmann, 2023] Bobby He and Thomas Hofmann. Simplifying transformer blocks. *arXiv preprint arXiv:2311.01906*, 2023.

[Hettige *et al.*, 2024] Kethmi Hirushini Hettige, Jiahao Ji, Shili Xiang, Cheng Long, Gao Cong, and Jingyuan Wang. Airphynet: Harnessing physics-guided neural networks for air quality prediction. *arXiv preprint arXiv:2402.03784*, 2024.

[Huang *et al.*, 2023] Qihe Huang, Lei Shen, Ruixin Zhang, Shouhong Ding, Binwu Wang, Zhengyang Zhou, and Yang Wang. Crossgnn: Confronting noisy multivariate time series via cross interaction refinement. *In Proc. of NeurIPS*, 36:46885–46902, 2023.

[Jing *et al.*, 2022] Baoyu Jing, Si Zhang, Yada Zhu, Bin Peng, Kaiyu Guan, Andrew Margenot, and Hanghang Tong. Retrieval based time series forecasting. *arXiv preprint arXiv:2209.13525*, 2022.

[Liang *et al.*, 2023] Yuxuan Liang, Yutong Xia, Songyu Ke, Yiwei Wang, Qingsong Wen, Junbo Zhang, Yu Zheng, and Roger Zimmermann. Airformer: Predicting nationwide air quality in china with transformers. In *Proc. of AAAI*, 2023.

[Lin *et al.*, 2024a] Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: enhancing time series forecasting through modeling periodic patterns. *arXiv preprint arXiv:2409.18479*, 2024.

[Lin *et al.*, 2024b] Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. Sparsetsf: Modeling long-term time series forecasting with 1k parameters. *arXiv preprint arXiv:2405.00946*, 2024.

[Liu *et al.*, 2023] Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quanjun Chen, and Xuan Song. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In *Proc. of CIKM*, 2023.

[Liu *et al.*, 2024] Chenxi Liu, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao. Spatial-temporal large language model for traffic prediction. In *25th IEEE International Conference on Mobile Data Management*, pages 31–40, 2024.

[Liu *et al.*, 2025a] Chenxi Liu, Hao Miao, Qianxiong Xu, Shaowen Zhou, Cheng Long, Yan Zhao, Ziyue Li, and Rui Zhao. Efficient multivariate time series forecasting via calibrated language models with privileged knowledge distillation. In *Proc. of ICDE*, 2025.

[Liu *et al.*, 2025b] Chenxi Liu, Shaowen Zhou, Qianxiong Xu, Hao Miao, Cheng Long, Ziyue Li, and Rui Zhao. Towards cross-modality modeling for time series analytics: A survey in the llm era. In *Proc. of IJCAI*, pages 1–9, 2025.

[Lu *et al.*, 2024] Jiecheng Lu, Xu Han, Yan Sun, and Shihao Yang. Cats: Enhancing multivariate time series forecasting by constructing auxiliary time series as exogenous variables. *arXiv preprint arXiv:2403.01673*, 2024.

[Ma *et al.*, 2024] Xiang Ma, Xuemei Li, Lexin Fang, Tianlong Zhao, and Caiming Zhang. U-mixer: An unet-mixer architecture with stationarity correction for time series forecasting. In *Proc. of AAAI*, 2024.

[Ma *et al.*, 2025] Jiaming Ma, Guanjun Wang, Sheng Huang, Kuo Yang, Binwu Wang, Pengkun Wang, and Yang Wang.

Spatiotemporal causal decoupling model for air quality forecasting. *arXiv preprint arXiv:2505.20119*, 2025.

[Miao *et al.*, 2024] Hao Miao, Ziqiao Liu, Yan Zhao, Chenjuan Guo, Bin Yang, Kai Zheng, and Christian S Jensen. Less is more: Efficient time series dataset condensation via two-fold modal matching. *PVLDB*, pages 226–238, 2024.

[Miao *et al.*, 2025] Hao Miao, Yan Zhao, Chenjuan Guo, Bin Yang, Kai Zheng, and Christian S Jensen. Spatio-temporal prediction on streaming data: A unified federated continuous learning framework. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[Murphy, 2012] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[Peng *et al.*, 2020] Zhen Peng, Wenbing Huang, Minnan Luo, Qinghua Zheng, Yu Rong, Tingyang Xu, and Junzhou Huang. Graph representation learning via graphical mutual information maximization. In *Proc. of WWW*, 2020.

[Shao *et al.*, 2022a] Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proc. of CIKM*, 2022.

[Shao *et al.*, 2022b] Zezhi Shao, Zhao Zhang, Wei Wei, Fei Wang, Yongjun Xu, Xin Cao, and Christian S. Jensen. Decoupled dynamic spatial-temporal graph neural network for traffic forecasting. *Proc. VLDB Endow.*, 2022.

[Shazeer, 2020] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[Wang *et al.*, 2020] Shuo Wang, Yanran Li, Jiang Zhang, Qingye Meng, Lingwei Meng, and Fei Gao. Pm2. 5-gnn: A domain knowledge enhanced graph neural network for pm2. 5 forecasting. In *Proceedings of the 28th international conference on advances in geographic information systems*, 2020.

[Wang *et al.*, 2023] Binwu Wang, Yudong Zhang, Xu Wang, Pengkun Wang, Zhengyang Zhou, Lei Bai, and Yang Wang. Pattern expansion and consolidation on evolving graphs for continual traffic prediction. In *Proc. of KDD*, pages 2223–2232, 2023.

[Wang *et al.*, 2024a] Binwu Wang, Jiaming Ma, Pengkun Wang, Xu Wang, Yudong Zhang, Zhengyang Zhou, and Yang Wang. Stone: A spatio-temporal ood learning framework kills both spatial and temporal shifts. In *Proc. of KDD*, pages 2948–2959, 2024.

[Wang *et al.*, 2024b] Binwu Wang, Pengkun Wang, Yudong Zhang, Xu Wang, Zhengyang Zhou, Lei Bai, and Yang Wang. Towards dynamic spatial-temporal graph learning: A decoupled perspective. In *Proc. of AAAI*, pages 9089–9097, 2024.

[Wang *et al.*, 2024c] Binwu Wang, Pengkun Wang, Zhengyang Zhou, Zhe Zhao, Wei Xu, and Yang Wang. Make bricks with a little straw: Large-scale spatio-temporal graph learning with restricted gpu-memory capacity. In *Proc. of IJCAI*, 2024.

[Wang *et al.*, 2024d] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. *arXiv preprint arXiv:2405.14616*, 2024.

[Wang *et al.*, 2024e] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *arXiv preprint arXiv:2402.19072*, 2024.

[Wu *et al.*, 2019] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019.

[Xu *et al.*, 2020] Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv:2001.02908*, 2020.

[Yeh *et al.*, 2024] Chin-Chia Michael Yeh, Yujie Fan, Xin Dai, Uday Singh Saini, Vivian Lai, Prince Osei Aboagye, Junpeng Wang, Huiyuan Chen, Yan Zheng, Zhongfang Zhuang, et al. Rpmixer: Shaking up time series forecasting with random projections for large spatial-temporal data. In *Proc. of KDD*, 2024.

[Yi *et al.*, 2018] Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, and Yu Zheng. Deep distributed fusion network for air quality prediction. In *Proc. of KDD*, 2018.

[Yu *et al.*, 2017] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[Yu *et al.*, 2023] Chengqing Yu, Fei Wang, Zezhi Shao, Lin Wu, and Yongjun Xu. Dsformer: A double sampling transformer for multivariate time series long-term prediction. In *Proc. of CIKM*, 2023.

[Yu *et al.*, 2025] Chengqing Yu, Fei Wang, Yilun Wang, Zezhi Shao, Tao Sun, and Yongjun Xu. Mgsfformer: A multi-granularity spatiotemporal fusion transformer for air quality prediction. *Information Fusion*, 2025.

[Zeng *et al.*, 2023] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proc. of AAAI*, 2023.

[Zhang and Sennrich, 2019] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Proc. of NeurIPS*, 2019.

[Zhang *et al.*, 2017] Henian Zhang, Yuhang Wang, Tae-Won Park, and Yi Deng. Quantifying the relationship between extreme air pollution events and extreme weather events. *Atmospheric Research*, 2017.

[Zhang *et al.*, 2025] Yudong Zhang, Xu Wang, Xuan Yu, Zhaoyang Sun, Kai Wang, and Yang Wang. Drawing informative gradients from sources: A one-stage transfer learning framework for cross-city spatiotemporal forecasting. In *In Proc. of AAAI*, volume 39, pages 1147–1155, 2025.