

Interpretable long-horizon air pollution forecasting using a transformer-based framework

Zhiguo Zhang ^a, Xiaoliang Ma ^{a,b,*}, Magnuz Engardt ^c, Daniel Schlesinger ^d,
Christer Johansson ^e

^a Department of Civil and Architectural Engineering, School of Architecture and Built Environment, KTH Royal Institute of Technology, Stockholm, Sweden

^b Centre for Digital Futures, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden

^c Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

^d SLB-analys, Environment and Health Administration, City of Stockholm, Stockholm, Sweden

^e Department of Environmental Science, Stockholm University, Stockholm, Sweden

ARTICLE INFO

Keywords:

Multi-horizon forecasting

Explainable AI

Meteorological forecast integration

Air quality and health risk

ABSTRACT

Accurate and interpretable air pollution forecasting is essential for public health and proactive intervention. While deep learning models have achieved strong performance in long-horizon forecasting, they often require aligned historical sequences and provide limited interpretability. To address these limitations, we propose a novel Transformer-based framework that integrates historical observations with auxiliary forecasts to enhance predictive performance while providing built-in interpretability. The framework introduces two core components: a Structured Temporal Embedding Module (STEM) for preserving the semantic integrity of heterogeneous inputs, and an eXplainable Target-Oriented Cross-Attention (X2-Attention) mechanism that aligns future predictions with known inputs to yield fine-grained attribution for each forecasted timestep across both temporal and feature dimensions. Evaluated on diverse urban and suburban datasets in Stockholm, our model simultaneously delivers reliable interpretability and superior predictive accuracy. Compared to five baselines (TimeXer, Crossformer, iTransformer, Transformer and LSTM), it exhibits significantly lower Mean Squared Error (MSE) for NO_x and competitive performance for PM₁₀ across horizons up to 720 hours. Ablation studies reveal that incorporating auxiliary forecasts reduces MSE by 29% for NO_x and 15% for PM₁₀. Moreover, the X2-Attention mechanism generates plausible feature attribution validated through perturbation analysis. Its attributions are consistent with GradientSHAP while avoiding prohibitive computational overhead, thereby facilitating efficient air quality forecasting.

1. Introduction

Air pollution is a pressing global environmental challenge, imposing substantial detrimental effects on public health and economic development (WHO, 2024). Consequently, accurate and interpretable multi-horizon forecasting of air pollutant concentrations is essential for developing effective environmental policies and enabling proactive public health interventions. Accurate forecasts allow individuals to implement timely mitigation measures (Laudan et al., 2025), while interpretability is indispensable for building public trust and ensuring the credibility of decision-support systems (Tang et al., 2024).

Air pollutant forecasting is primarily approached through two paradigms: physics-based models and data-driven models. Operational

systems often rely on numerical models grounded in physical principles (CAMS, 2017; Ottosen et al., 2015). Although these models provide physical insights, they suffer from high computational costs and error accumulation over extended horizons (Brasseur & Jacob, 2017; Gidhagen et al., 2005). In contrast, data-driven approaches have demonstrated superior performance in capturing non-linear relationships between pollutant concentrations and key drivers such as meteorology and traffic flow (Li et al., 2025; Méndez et al., 2023; Qiu et al., 2023). However, maintaining accuracy over longer horizons remains a significant challenge. Many models rely exclusively on historical observations (Gerges et al., 2024; Hu et al., 2025; Yu et al., 2025). This reliance ignores forward-looking information, such as anticipated atmospheric conditions and scheduled emission dynamics, which directly influence long-term pollutant dynamics.

* Corresponding author.

E-mail addresses: zhiguo@kth.se (Z. Zhang), liang@kth.se (X. Ma), magnuz.engardt@smhi.se (M. Engardt), daniel.schlesinger@slb.nu (D. Schlesinger), christer.johansson@aces.su.se (C. Johansson).

<https://doi.org/10.1016/j.eswa.2026.131719>

Received 4 October 2025; Received in revised form 11 February 2026; Accepted 14 February 2026

Available online 18 February 2026

0957-4174/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

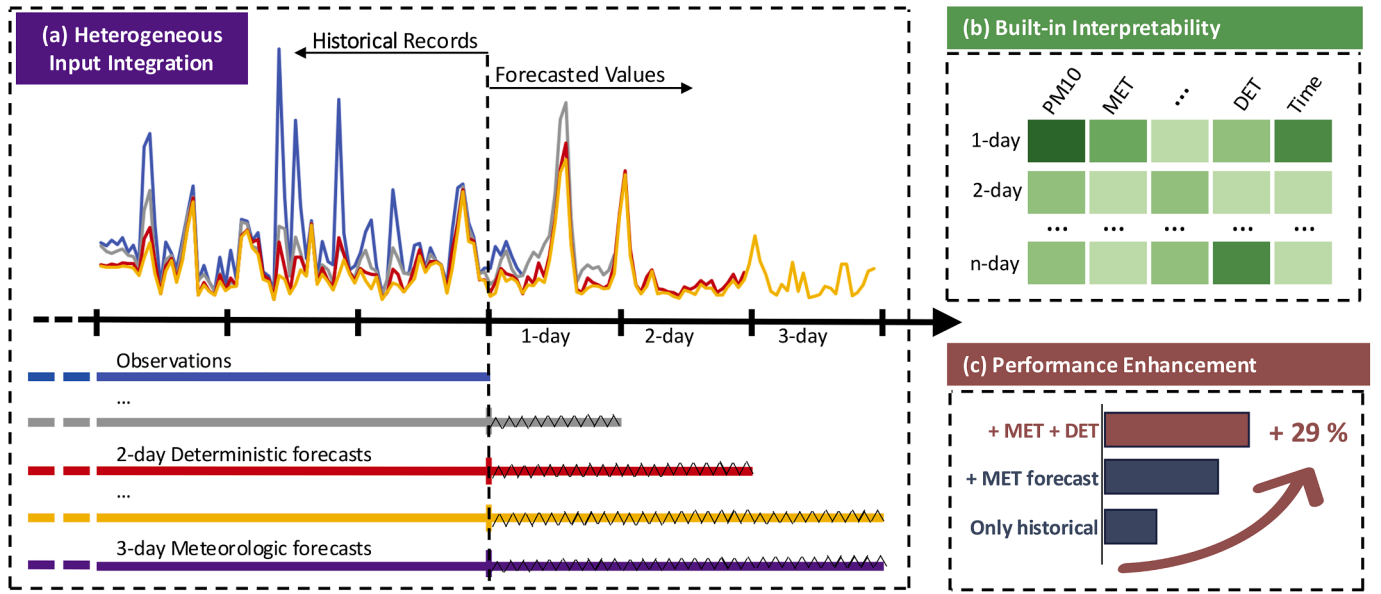


Fig. 1. Interpretable long-horizon forecasting with heterogeneous inputs. (a) Integration of historical records (solid lines) and multi-source forecasts (jagged segments) with varying temporal coverage. (b) Fine-grained interpretability, where the heatmap illustrates the contribution of each feature segment (columns) to specific future horizons (rows). (c) The proposed model achieves an average 29% MSE reduction for NO_x within 1–3 day horizon by effectively leveraging forward-looking variables.

Integrating these auxiliary forecasts introduces a dual challenge of heterogeneous data alignment and complex dependency modeling. As illustrated in Fig. 1(a), forward-looking data from different forecasting systems often vary in variable types, horizons, and temporal resolutions. Standard alignment techniques frequently distort intrinsic temporal patterns (Bertrand et al., 2023; Liu et al., 2024). Furthermore, pollutant dynamics are driven by the asynchronous coupling of historical accumulation and forecasted external drivers (Di et al., 2019). This complexity requires models that can jointly represent temporal dynamics and cross-feature interactions. Classic neural architectures, including Recurrent Neural Networks (RNNs) or standard Transformers, generally prioritize either temporal trends or feature correlations, which can limit their effectiveness in long-horizon air quality prediction (Munkhdalai et al., 2022; Wang et al., 2023; Zhang et al., 2023a).

Beyond predictive accuracy, interpretability remains a persistent bottleneck due to structural limitations. In Transformer-based architectures, attention mechanisms operate on latent representations that have undergone multiple non-linear projections and normalizations (Liu et al., 2023; Zhou et al., 2021). This semantic ambiguity means that resulting attention weights reflect abstract correlations between hidden states rather than traceable contributions of physical variables, thereby obscuring the model's decision-making logic. While interpretable models like the Temporal Fusion Transformer (TFT) introduce variable selection, they typically generate feature weights for the entire forecast sequence (Lim et al., 2021). Such horizon-static explanations lack the resolution to reveal how these attributions dynamically evolve for each individual future timestep. Neglecting this attribution drift is particularly problematic in long-term forecasting, where the influence of historical observations naturally diminishes over time while the relevance of forward-looking variables gradually takes precedence.

To address these limitations, we propose a novel Transformer-based framework that unifies predictive performance with fine-grained interpretability. The framework consists of two key components: a Structured Temporal Embedding Module (STEM) and an eXplainable Target-Oriented Cross-Attention (X2-Attention) mechanism.

Unlike existing architectures, our design prioritizes semantic preservation to support joint interpretability across temporal and feature dimensions. Specifically, STEM module encodes heterogeneous inputs, including pollutant measurements, meteorological forecasts, and deter-

ministic model outputs, into a structured representation that retains the identity of each feature. Building on this, the X2-Attention mechanism utilizes learnable queries that represent future forecast targets. By operating directly on physically meaningful segments of each feature, the model establishes an explicit link between future predictions and specific inputs, as conceptually shown in Fig. 1(b). This design preserves the long-sequence modeling capabilities and provides joint temporal-feature attribution, which simultaneously identifies key driving features and their critical time windows for each individual forecasted timestep.

Our primary contributions are summarized as follows:

- **Structured Encoding:** We develop the STEM module to effectively align heterogeneous observations and auxiliary forecasts, significantly enhancing forecasting performance, as shown in Fig. 1(c).
- **Explainable Mechanism:** We propose the X2-Attention mechanism to provide fine-grained importance attributions for each forecasted timestep, revealing the dynamic shift of predictive drivers across the forecast horizon.
- **Validated Interpretability:** We validated the empirical fidelity of X2-Attention through perturbation analysis and consistency evaluations with GradientSHAP, providing plausible explanations without the prohibitive computational overhead of post-hoc methods.

Extensive experiments on urban and suburban datasets in Stockholm validate predictive accuracy of the proposed framework against five competitive baselines for prediction horizons up to 720 hours, including TimeXer (Wang et al., 2024), Crossformer (Zhang & Yan, 2023), iTransformer (Liu et al., 2023), Transformer (Zhou et al., 2021), and LSTM-Seq2Seq (Zhou et al., 2019). Overall, our model delivers built-in, fine-grained interpretability alongside superior predictive accuracy, highlighting the potential of forecast-informed, interpretable architectures for long-horizon air pollutant predictions.

2. Related work

This section reviews three areas relevant to our contribution: the integration of auxiliary forecasted variables, multi-horizon forecasting modeling, and interpretability in data-driven models.

2.1. Air pollution forecasting with forecasted variables

Air quality dynamics are governed by complex interactions between pollutant emissions and atmospheric physics. Traditionally, physics-based models have been the dominant choice for operational forecasting. At regional scales, chemical transport models such as WRF-Chem and CMAQ simulate pollutant formation and transport using meteorological forecasts and emission inventories in large-scale (Appel et al., 2017; CAMS, 2017; Tuccella et al., 2012). At the micro-scale, the Operational Street Pollution Model leverages localized information, such as street geometry, to provide high-resolution forecasts (Ottosen et al., 2015). While these models offer physical consistency, they are computationally intensive and sensitive to input uncertainties, often leading to error accumulation over long horizons (Gidhagen et al., 2005).

In contrast, many data-driven methods have integrated forecasted variables, such as meteorology and remote sensing, to enhance predictive performance and extend lead times (Li et al., 2025; Tang et al., 2024; Zhang et al., 2023b). Meteorological forecasts have been used to predict NO_x up to 72 hours, $\text{PM}_{2.5}$ up to 120 hours, and O_3 up to 216 h (Liu et al., 2024; Zhang et al., 2023a, 2024). Deterministic model outputs have also been integrated for bias correction (Qiu et al., 2023; Sayeed et al., 2021). However, these efforts typically rely on simple concatenation or interpolation to align multi-source data (Bertrand et al., 2023; Liu et al., 2024), which may obscure intrinsic temporal patterns and fail to maintain the semantic distinction between different data sources. Inspired by patch-based embedding in computer vision and time-series modeling (Dosovitskiy et al., 2020; Nie et al., 2022), this study proposes a structured embedding strategy that accommodates variable-length inputs while preserving their semantic and temporal integrity.

2.2. Multi-Horizon air pollution forecasting

Multi-horizon forecasting methods generate predictions for all future time steps simultaneously, which provides continuous trajectories and mitigates the error accumulation common in single-step models (Bi et al., 2022; Lim et al., 2021). Early multi-horizon forecasting primarily utilized RNNs and their variants (Li et al., 2023; Zeng et al., 2022). However, their sequential computation and limited receptive field constrain scalability and long-range dependency modeling (Bai et al., 2018).

Transformer-based models have become prominent due to self-attention and parallel computation. Representative variants, such as Informer and Autoformer, introduced sparse attention and series decomposition to handle long-horizon forecasting (Wu et al., 2021; Zhou et al., 2021). While effective for capturing temporal trends, these models often construct tokens by merging all variables at each time step, which can obscure cross-feature interactions and reduce flexibility for heterogeneous inputs (Feng et al., 2023; Tian et al., 2024; Yu et al., 2023). In response, feature-centric models such as iTransformer apply attention across the variable dimension (Liu et al., 2023), and PatchTST utilizes sub-series patching to enhance local semantic extraction (Nie et al., 2022). While these methods emphasize cross-variable interactions, they may do so at the expense of representing the intricate temporal dynamics within each feature. Hybrid designs such as Crossformer and TimeXer attempt to bridge temporal and feature dimensions (Wang et al., 2024; Zhang & Yan, 2023; Zhang et al., 2025). Their effectiveness has primarily been validated on datasets with strong periodicity like electricity consumption (Chen, Céspedes). Air quality forecasting involves highly irregular and volatile dynamics driven by the asynchronous coupling of historical accumulation and sudden meteorological or emission changes (Di et al., 2019). However, the capacity of these models to handle such complexities remains largely unproven.

2.3. Interpretability in air pollution forecasting

Interpretability is a central concern in environmental applications where model transparency affects trust and adoption (Houdou et al.,

2024; Rudin, 2019). Existing interpretability approaches are categorized as model-intrinsic or model-agnostic (Molnar, 2020). Intrinsic interpretability integrates explanatory structures directly into the model. Some hybrid pipelines combine deep learning with linear or correlation-based components (Munkhdalai et al., 2022; Yang et al., 2024), while tree-based ensembles provide global feature importance rankings (Breiman, 2001; Hou et al., 2022). Although effective for initial variable screening, these methods typically produce temporally invariant rankings. By aggregating contributions across the entire input window, they fail to distinguish between immediate precursors and distant historical trends, thereby smoothing over critical, time-sensitive physical signals.

In Transformer-based forecasting, attention weights are often treated as a source of transparency, yet these explanations remain ambiguous due to structural limitations. In standard Encoder-Decoder frameworks, the original physical identity of variables is often attenuated as raw inputs pass through deep non-linear encoding layers (Wu et al., 2021; Zhou et al., 2021). The resulting attention maps reflect abstract correlations between latent hidden states rather than traceable contributions of physical variables, creating a semantic gap.

The Temporal Fusion Transformer (TFT) introduced an influential design by combining variable selection networks (VSN) with self-attention (Lim et al., 2021). Although VSN provides input-side attribution to identify relevant features at each historical moment, these features are subsequently compressed into a fused representation before decoding. This fusion obscures the specific contributions of individual features during the prediction phase. Furthermore, TFT's explanations are largely horizon-static, pertaining to the entire forecast sequence as a single entity rather than to each future timestep. This neglects the inherent attribution drift in atmospheric dynamics, where governing drivers evolve as the forecast lead-time increases.

Post-hoc model-agnostic techniques, such as Shapley Additive Explanations (SHAP) and Permutation Importance, are frequently applied as external evaluators. Permutation method assumes feature independence, a premise that fails to hold under the strong multicollinearity of atmospheric variables (Casalicchio et al., 2019). SHAP offers a more theoretically grounded feature attribution based on Cooperative Game Theory (Gu et al., 2021; Liu et al., 2024; Lundberg, 2017). However, applying SHAP to transformer-based models presents significant practical barriers. The method's reliance on extensive sampling makes it computationally prohibitive. Additionally, the reliability of post-hoc explanations remains a concern, as they can be manipulated to conceal model biases, potentially leading to misleading interpretations and flawed conclusions (Rudin, 2019; Slack et al., 2020).

3. Methodology

3.1. Problem definition

Multi-horizon air pollution forecasting aims to predict the future concentrations of a target pollutant over a prediction horizon ℓ . Let $\mathbf{x}_{t+1:t+\ell} = [x_{t+1}, x_{t+2}, \dots, x_{t+\ell}] \in \mathbb{R}^\ell$ denote the sequence of future target values, where x_k is the pollutant concentration at time k . We learn a mapping function $\mathcal{F}$ that predicts

$$\hat{\mathbf{x}}_{t+1:t+\ell} = \mathcal{F}\left(\tilde{\mathbf{x}}_{t-n+1:t}, \tilde{\mathbf{z}}_{t-n+1:t+m_i}\right), \quad (1)$$

where $\tilde{\mathbf{x}}_{t-n+1:t}$ denotes the historical pollutant observations in a look-back window of length n , and $\tilde{\mathbf{z}}$ is the set of C auxiliary variables available at training and inference,

$$\tilde{\mathbf{z}}_{t-n+1:t+m_i} = \left\{ \mathbf{z}_{t-n+1:t+m_i}^i \in \mathbb{R}^{T_i} \mid i = 1, \dots, C, T_i = n + m_i \right\}. \quad (2)$$

Each feature vector $\mathbf{z}^i$ concatenates n historical values with m_i known forecasted values. The forecast length m_i may differ across features. For instance, meteorological forecasts usually covers a shorter period than the prediction horizon, i.e., $m_{\text{met}} < \ell$, whereas temporal encodings, such

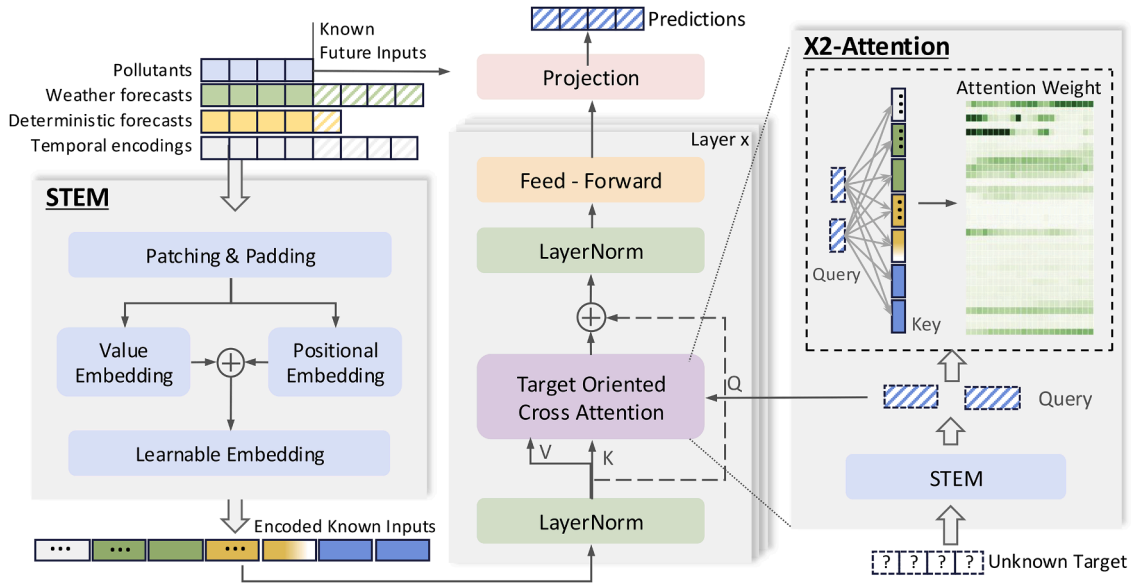


Fig. 2. Overview of the proposed forecasting framework. Solid boxes are historical inputs, and slanted boxes are known forecasted variables. Question marks ('?') denote unknown prediction targets. STEM encodes heterogeneous inputs into a unified representation. Future targets are posed as learnable queries that attend to the encoded inputs through X2-Attention. The attention weights provide target-oriented interpretability.

as hour-of-day or day-of-week, extend over the entire horizon, yielding $m_{\text{temporal}} = \ell$.

3.2. Framework overview

The proposed framework integrates two core components: a Structured Temporal Embedding Module (STEM) for encoding heterogeneous inputs and an eXplainable Target-Oriented Cross-Attention (X2-Attention) mechanism for target-oriented explanations, as illustrated in Fig. 2. Inputs include historical pollutant measurements, forecasted meteorological data, deterministic air quality model outputs, and temporal encodings. These inputs are embedded into a unified patch-based representation through the STEM module, which preserves local temporal patterns and semantic consistency across features. Unknown forecasting targets are represented as learnable queries via the STEM module. These target-oriented queries then interact with the encoded input representation through cross-attention mechanism. This operation explicitly computes the relevance of each input patch to each future prediction patch, thereby capturing complex temporal and feature dependencies. The resulting attention weights serve as a direct, built-in attributions of feature importance for each forecast timestep. Moreover, the attended representations are passed through position-wise feed-forward layers to produce patch-level predictions, which are then reassembled into a continuous sequence to form the final multi-horizon forecast.

3.3. Structured temporal embedding module (STEM)

STEM encodes all known inputs into a structured representation, accommodating their different temporal lengths, illustrated in Fig. 3. Specifically, for each input feature vector $\mathbf{z}^i \in \mathbb{R}^{T_i}$ with $T_i = n + m_i$, we partition the sequence into $N_i = \lceil T_i / L_{\text{pat}} \rceil$ non-overlapping segments of length L_{pat} . The operator $\lceil \cdot \rceil$ is the ceiling. If T_i is not divisible by L_{pat} , the final segment is completed with zero-padding. These segments, or patches, are then stacked as rows to form the patch matrix $\mathbf{Z}_{\text{pat}}^i \in \mathbb{R}^{N_i \times L_{\text{pat}}}$ for the feature i , which is then projected into a shared d -dimensional latent space using a linear transformation as follows:

$$\mathbf{E}_v^i = \mathbf{Z}_{\text{pat}}^i \cdot \mathbf{W}_v + \mathbf{b}_v, \quad \mathbf{W}_v \in \mathbb{R}^{L_{\text{pat}} \times d}, \quad \mathbf{b}_v \in \mathbb{R}^d. \quad (3)$$

where $\mathbf{E}_v^i \in \mathbb{R}^{N_i \times d}$ is the produced value embedding. The operation reduces the temporal resolution from the timestep-level to patch-level while preserving local temporal patterns within each segment.

A fixed sinusoidal positional encoding (Vaswani et al., 2017) is applied to preserve the internal temporal order within each variable and facilitate alignment across heterogeneous inputs. For the i -th feature vector with N_i patches, the corresponding positional encoding matrix $\mathbf{E}^{\text{fix_pos}} \in \mathbb{R}^{N_i \times d}$ is computed element-wise by:

$$\mathbf{E}^{\text{fix_pos}}(j, 2k) = \sin\left(\frac{j}{10000^{2k/d}}\right), \quad (4)$$

$$\mathbf{E}^{\text{fix_pos}}(j, 2k+1) = \cos\left(\frac{j}{10000^{2k/d}}\right), \quad (5)$$

where $j \in \{0, \dots, N_i - 1\}$ is the patch index and $k \in \{0, \dots, d/2 - 1\}$ is the embedding dimension index.

The initial embedding for the c -th feature is obtained by summing its value embedding and the corresponding positional encoding:

$$\mathbf{E}_{\text{ini}}^i = \mathbf{E}_v^i + \mathbf{E}^{\text{fix_pos}}. \quad (6)$$

Subsequently, concatenating all variables along the patch dimension across features yields a matrix:

$$\mathbf{E}_{\text{raw}} = \text{Concat}(\mathbf{E}_{\text{ini}}^1, \dots, \mathbf{E}_{\text{ini}}^i, \dots, \mathbf{E}_{\text{ini}}^C), \quad (7)$$

where $\mathbf{E}_{\text{raw}} \in \mathbb{R}^{M \times d}$ and $M = \sum_i N_i$.

A learnable positional embedding $\mathbf{E}^{\text{learn_pos}} \in \mathbb{R}^{M \times d}$ is further introduced to preserve the relative order after concatenation, which can be jointly optimized with other model parameters. The final encoded inputs are produced by applying Layer Normalization:

$$\mathbf{E}_{\text{inp}} = \text{LayerNorm}(\mathbf{E}_{\text{raw}} + \mathbf{E}^{\text{learn_pos}}). \quad (8)$$

The resulting structured embedding $\mathbf{E}_{\text{inp}} \in \mathbb{R}^{M \times d}$ preserves the semantic identity of heterogeneous inputs and provides a compact representation for the downstream attention mechanism.

3.4. Explainable target-Oriented cross-Attention (X2-Attention)

Self-attention applies the same sequence as queries, keys, and values, as illustrated in Fig. 4. While effective for prediction, this symmetric design captures only internal relationships among known inputs. It produces a coarse measure of feature importance aggregated over the

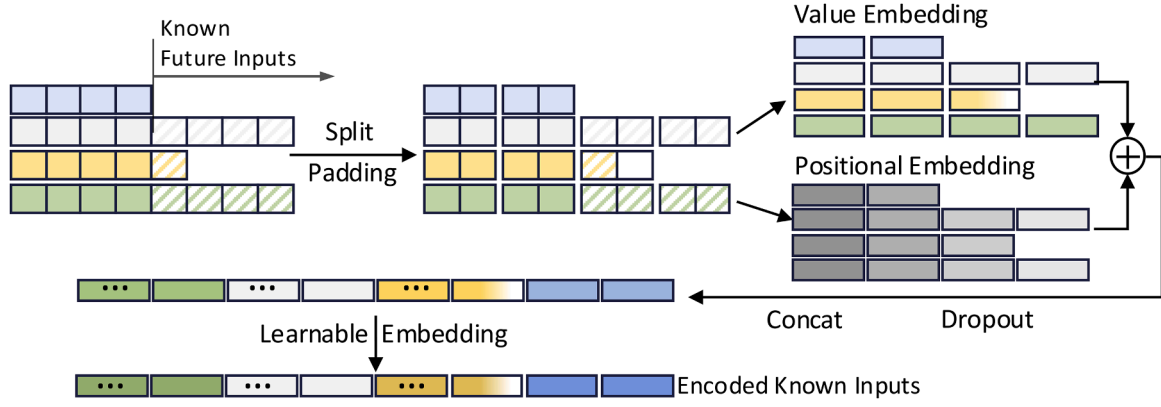


Fig. 3. Structured Temporal Embedding Module (STEM). Solid-filled boxes denote historical inputs and slanted boxes represent known future inputs.

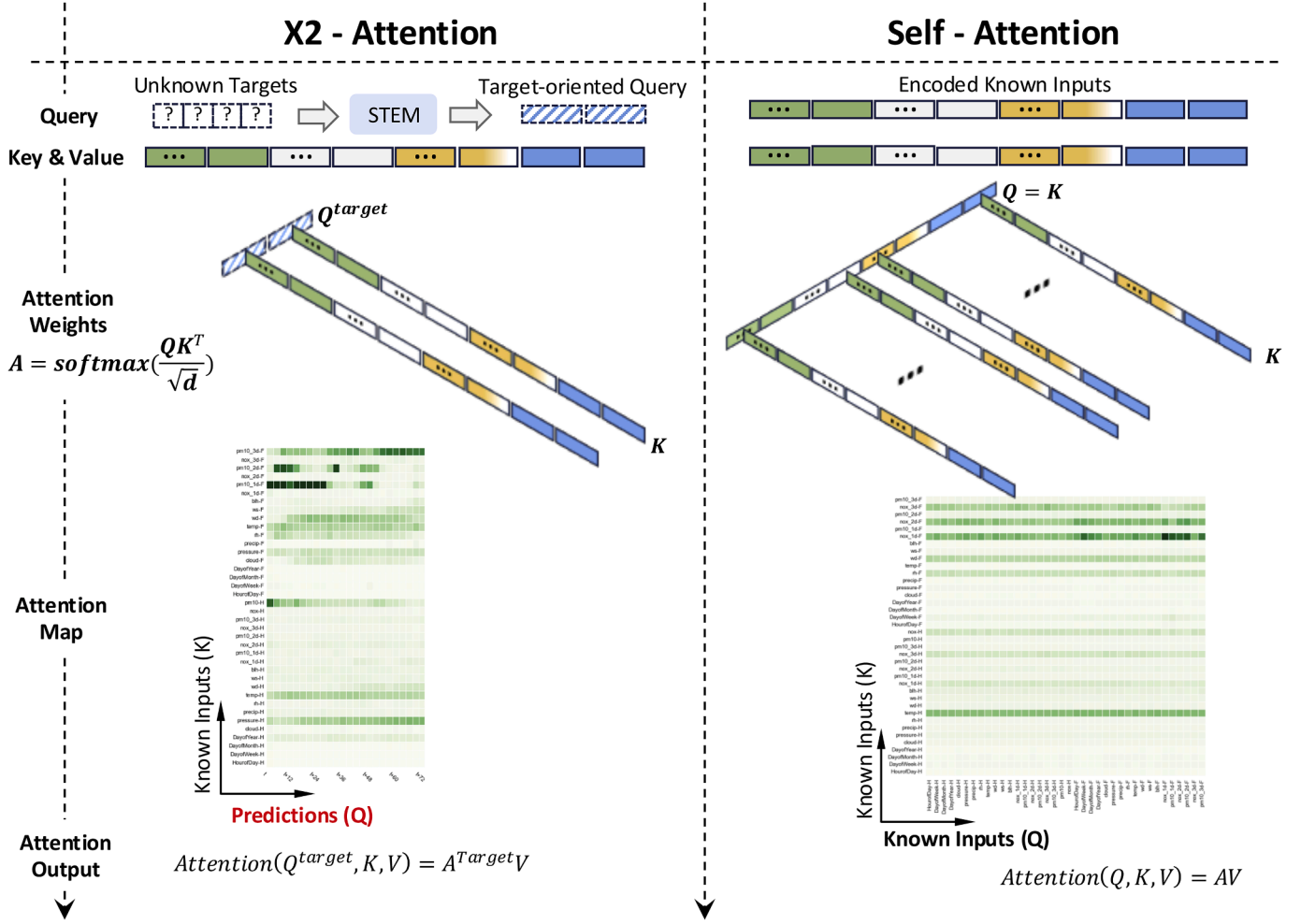


Fig. 4. A comparison between X2-Attention and standard self-attention, including the construction of queries, keys, and values, as well as the computation of attention weights, attention maps, and outputs. Unlike self-attention, which captures dependencies among known inputs, X2-Attention explicitly connects each prediction target to relevant inputs.

entire forecasting range, rather than offering timestep-wise attributions for individual prediction patch.

The proposed X2-Attention mechanism explicitly aligns known inputs with future targets through learnable, target-specific queries. Specifically, we construct a proxy sequence $\tilde{\mathbf{X}}_{t-\ell+1:t} \in \mathbb{R}^{\ell \times s}$ from the most recent ℓ time steps of target pollutants. This proxy captures the autoregressive nature of pollutant dynamics and provides a temporal context. The proxy sequence is then processed by the STEM module to

generate target-oriented query embeddings:

$$\mathbf{E}^Q = \text{STEM}(\tilde{\mathbf{X}}_{t-\ell+1:t}),$$

where $\mathbf{E}^Q \in \mathbb{R}^{N_Q \times d}$ serve as target-driven anchors to guide the attention mechanism. $N_Q = s \times \lceil \ell / L_{\text{patch}} \rceil$ denotes the number of patches and each query patch represents a specific time interval within the forecast horizon.

Cross-attention is then performed between the learned embeddings $\mathbf{E}^Q$ and the encoded input sequence $\mathbf{E}_{\text{inp}}$. These embeddings are projected into the query, key, and value spaces:

$$\mathbf{Q}^{\text{target}} = \mathbf{E}^Q \cdot \mathbf{W}_q, \quad (9)$$

$$\mathbf{K} = \mathbf{E}_{\text{inp}} \cdot \mathbf{W}_k, \quad (10)$$

$$\mathbf{V} = \mathbf{E}_{\text{inp}} \cdot \mathbf{W}_v, \quad (11)$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d}$ are corresponding learnable projection matrices. The scaled dot-product attention is computed by:

$$\text{Attention}(\mathbf{Q}^{\text{target}}, \mathbf{K}, \mathbf{V}) = \mathbf{A} \cdot \mathbf{V} = \text{Softmax}\left(\frac{\mathbf{Q}^{\text{target}} \mathbf{K}^T}{\sqrt{d}}\right) \cdot \mathbf{V}. \quad (12)$$

where $\mathbf{A} \in \mathbb{R}^{N_Q \times N}$ are attention weights that quantify the contribution of each known input patch to every future prediction patch.

3.5. Multi-Head attention and output projection

To enhance representational capacity, the model adopts Multi-Head Attention (MHA). For each of the H attention heads, the query, key, and value embeddings are projected into a subspace of dimension $d_h = d/H$. The attention for a single head is computed as:

$$\text{head}_h = \text{Attention}\left(\mathbf{E}^Q \cdot \mathbf{W}_q^h, \mathbf{E}_{\text{inp}} \cdot \mathbf{W}_k^h, \mathbf{E}_{\text{inp}} \cdot \mathbf{W}_v^h\right). \quad (13)$$

Outputs from all H heads are concatenated and then projected back to the full latent dimension:

$$\text{MHA}(\mathbf{E}^Q, \mathbf{E}_{\text{inp}}) = \text{Concat}_h(\text{head}_1, \dots, \text{head}_H) \cdot \mathbf{W}_o. \quad (14)$$

where $\mathbf{W}_q^h, \mathbf{W}_k^h, \mathbf{W}_v^h \in \mathbb{R}^{d \times d_h}$ are projection matrices for the h -th head and $\mathbf{W}_o \in \mathbb{R}^{d \times d}$ is a output projection matrix that restores the embedding to its original dimension.

After the attention operation, a position-wise Feed-Forward Network with Gaussian Error Linear Unit (GeLU) activation function is applied to increase model depth and non-linearity (Hendrycks & Gimpel, 2016):

$$\mathbf{Z}_{\text{processed}} = \text{GELU}(\mathbf{Z}_{\text{MHA}} \cdot \mathbf{W}_1 + \mathbf{b}_1) \cdot \mathbf{W}_2 + \mathbf{b}_2, \quad (15)$$

where $\mathbf{Z}_{\text{MHA}} \in \mathbb{R}^{N_Q \times d}$ denotes the output of the attention module; $\mathbf{W}_1 \in \mathbb{R}^{d \times d_{\text{ff}}}$, $\mathbf{W}_2 \in \mathbb{R}^{d_{\text{ff}} \times d}$, and $d_{\text{ff}} > d$.

Its output is further mapped $\mathbf{Z}_{\text{processed}} \in \mathbb{R}^{N_Q \times d}$ to a predicted patch:

$$\hat{\mathbf{x}}_{\text{patched}} = \mathbf{Z}_{\text{processed}} \cdot \mathbf{W}_{\text{out}} + \mathbf{b}_{\text{out}}, \quad (16)$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{d \times L_{\text{patch}}}$. The resulting patches are reshaped and concatenated to reconstruct the continuous multi-horizon forecast $\hat{\mathbf{x}}_{t+1:t+\ell}$.

3.6. Built-in interpretability

The X2-Attention mechanism provides intrinsic interpretability without additional post-hoc computations. Since the query embeddings are constructed to represent future prediction patches, the attention weights $\mathbf{A} \in \mathbb{R}^{N_Q \times M}$ directly quantify the contribution of each input patch to each forecast patch, where N_Q is the number of forecast patches and M is the total number of input patches. By controlling the patch size, the model trades off computational efficiency with granularity of interpretations. Smaller patch sizes yield more fine-grained, nearly timestep-level attributions at a higher computational cost.

To extract more practical insights from raw attention weights, we aggregate them at the feature level. Each element a_{jk} of $\mathbf{A}$ represents the contribution of the k -th input patch to the j -th forecast patch. Let $\mathcal{P}^i$ denote the set of input patch indices corresponding to feature i , where $i = 1, \dots, C$. The total attention weight assigned to feature i by forecast patch j is defined as:

$$a_{ji}^{\text{feat}} = \sum_{k \in \mathcal{P}^i} a_{jk}. \quad (17)$$

Table 1

Overview of input variables categorized by data sources. Variables with the suffix -H represent historical records, whereas those with -F describe forecasted values.

Data Source	Variables
Pollutant Measurements	NO _x -H, PM ₁₀ -H
Deterministic Forecasts	nox_1d-H / F, nox_2d-H / F, nox_3d-H / F, pm10_1d-H / F, pm10_2d-H / F, pm10_3d-H / F
Meteorological Forecasts	wind_direction-H / F, wind_speed-H / F, pressure-H / F, temperature-H / F, precipitation-H / F, cloud-H / F, relative_humidity-H / F, boundary_layer_height-H / F
Temporal Encodings	HourofDay-H / F, DayofWeek-H / F, DayofMonth-H / F, DayofYear-H / F

This aggregation yields a feature-level attribution matrix $\mathbf{A}^{\text{feat}} \in \mathbb{R}^{N_Q \times C}$, where each element a_{ji}^{feat} quantifies the importance of input feature i to forecast patch j for a given instance.

For broader comparisons, we compute global feature importance rankings by averaging the feature-level attributions across all forecast patches and all instances:

$$\bar{\alpha} = \mathbb{E}_{j, \text{instance}} \left[\mathbf{a}_j^{\text{feat}} \right], \quad (18)$$

where $\bar{\alpha} \in \mathbb{R}^C$ summarizes the average contribution of each input feature over the entire prediction horizon and dataset, yielding a global ranking of feature importance. This aligns with common interpretability practices, such as the MDI used in tree-based models (Scornet, 2023) and permutation-based importance scores (Hooker et al., 2021), which also produce feature-level rankings aggregated over all samples.

4. Experimental setup

4.1. Datasets and pre-processing

The city of Stockholm operates a comprehensive air quality management system, which relies on a suite of deterministic air quality models at regional, urban, and street scales and provides short-term forecasts for the entire metropolitan area (Engardt et al., 2021). Our study leverages this operational infrastructure to enhance multi-day forecasting capabilities.

We utilize data from three representative street-canyon monitoring stations in central Stockholm: Hornsgatan (HO), Folkungagatan (FO), and Sveavägen (SV). These sites provide a rigorous testbed under dense urban conditions. The dataset covers the period from August 2020 to October 2023 with hourly resolution. To enhance long-term forecasting, we incorporate auxiliary variables from several external forecasting systems. Meteorological forecasts are sourced from the Swedish Meteorological and Hydrological Institute (SMHI, 2025) and the Norwegian Meteorological Institute (MET Norway, 2025), providing hourly predictions for key variables such as temperature, wind speed, humidity, and boundary layer height, with lead times up to 72 hours. Additionally, we integrate deterministic forecasts for NO_x and PM₁₀ from the Stockholm Air Quality Forecast System (Engardt et al., 2021). This system aggregates predictions from the Copernicus Atmosphere Monitoring Service (CAMS) (CAMS, 2017), the Airviro Gaussian dispersion system (Kilbo Edlund et al., 2024), and the Operational Street Pollution Model (OSPM) (Ottosen et al., 2015). These deterministic forecasts are available for 1-day, 2-day, and 3-day horizons.

To evaluate model generalization across heterogeneous environments, we incorporated data from Hallunda (HA), a suburban station adjacent to a six-lane motorway arterial. Unlike the inner-city street canyons (HO, FO, SV), HA represents an open-terrain micro-environment characterized by distinct emission sources and dispersion dynamics. Specifically, pollutant dynamics at HA are driven by high-speed vehicular exhaust and road dust resuspension under rapid ventilation conditions, which contrast sharply with the low-speed traffic and

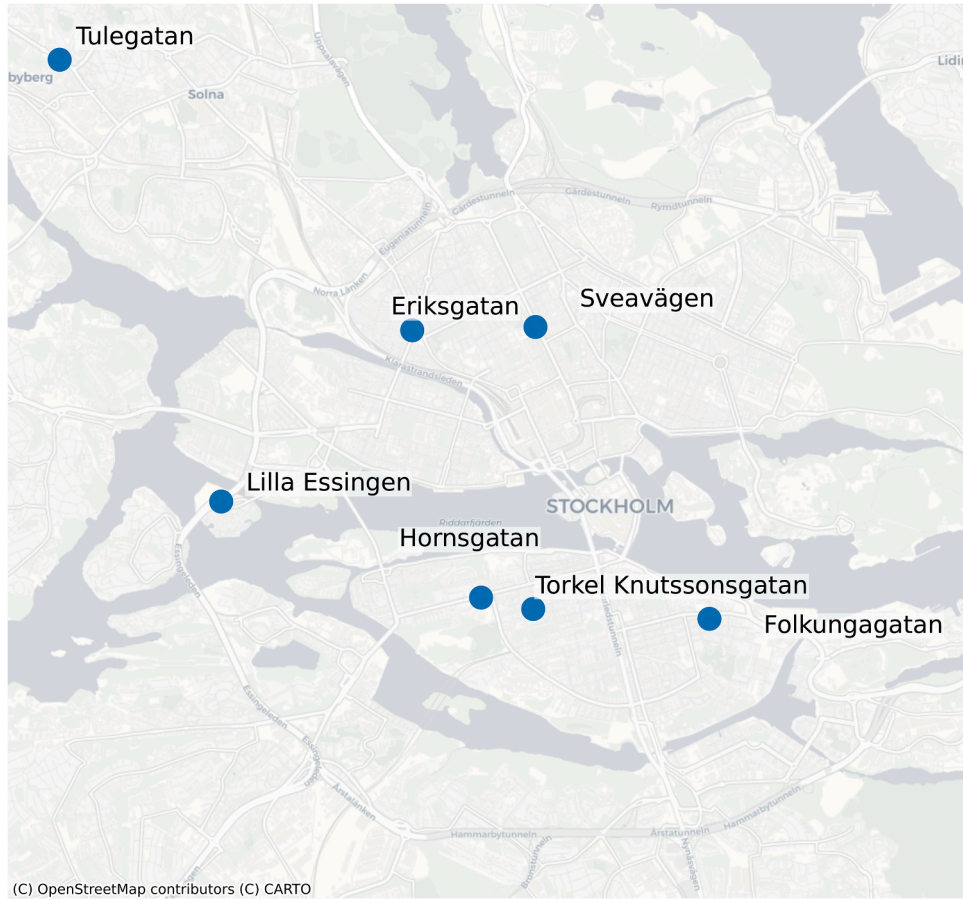


Fig. 5. The geographical distribution of air quality monitoring stations across Stockholm, Sweden.

trapping effects typical of street canyon stations. Detailed locations of both the urban street canyon and suburban highway sites are provided in Figs. 5 and D.17.

A critical feature of our dataset is the dual availability of all auxiliary variables as both historical records and future predictions. This allows the model to utilize both retrospective patterns and prospective conditions. In contrast, pollutant concentrations are available solely as historical observations. Temporal encodings are included to capture cyclical patterns. We denote historical input variables with the suffix -H and forecasted variables with -F. A complete list of input features is summarized in Table 1.

Data pre-processing includes missing value imputation using historical average interpolation to preserve periodicity (Zhang et al., 2024). Physically implausible values, such as negative pollutant concentrations, are treated as missing values prior to imputation. Outliers are retained to reflect real-world extreme events. All features are normalized to the range [0, 1] to ensure numerical stability. The data is split chronologically into 70 % for training, 20 % for validation, and 10 % for testing.

4.2. Baseline models

To evaluate the effectiveness of the proposed framework, we compare it against five baselines representing the state-of-the-art paradigms for long-horizon forecasting. These baseline models include recurrent neural networks (LSTM-Seq2Seq), attention-based models that focus on either temporal- or feature-centric dependencies (Transformer and iTransformer), and hybrid designs that jointly model both dimensions (Crossformer and TimeXer).

The LSTM-Seq2Seq (LSTM) model is a standard recurrent sequence-to-sequence architecture (Zhou et al., 2019). An encoder-only vari-

ant of standard Transformer is included as a fundamental attention-based model that primarily captures temporal dependencies (Zhou et al., 2021). iTransformer extends Transformer by treating each variable as a token, enabling it to better capture cross-variable dependencies (Liu et al., 2023). Crossformer is selected as a hybrid modeling that applies hierarchical dual attention across both temporal and variable dimensions (Zhang & Yan, 2023). Finally, TimeXer also is a hybrid strategy, developed for forecasting with auxiliary variables (Wang et al., 2024). It separates target and auxiliary inputs using self-attention for intrinsic target dynamics and cross-attention for incorporating external forecasts.

All baselines adopted public implementations (Wang et al., 2024) and were evaluated under identical experimental conditions. Notably, these baseline models are designed to handle temporally aligned historical input. To ensure a fair comparison, we applied a unified pre-processing strategy. Forecast variables with shorter lengths than prediction window were temporally aligned using zero-padding, and all auxiliary features were concatenated with its historical records to form an input sequence. This strategy allowed each baseline model to utilize forecasted variables without changing its original input design.

For interpretability evaluation, we use GradientSHAP (Lundberg, 2017), a widely used and theoretically grounded post-hoc explanation technique, as a reference. To the best of our knowledge, this is the only existing method capable of providing timestep-wise interpretability for multi-horizon forecasting. GradientSHAP estimates feature attributions by integrating gradients along the path between a baseline and the actual input, incorporating noise for robustness. Formally, the attribution for feature i is computed as

$$\text{GradSHAP}_i = (x_i - x'_i) \mathbb{E}_\epsilon \left[\int_0^1 \frac{\partial F(x' + \alpha(x - x') + \epsilon)}{\partial x_i} d\alpha \right], \quad (19)$$

where x is the input sample, x' is the baseline input, α is the interpolation coefficient, and ϵ denotes a noise term sampled from a predefined distribution.

4.3. Implementations

All computational experiments were conducted on a single NVIDIA A40 GPU. Models were optimized using the Adam optimizer (Kingma & Ba, 2014) ($\beta_1 = 0.9$, $\beta_2 = 0.999$) to minimize Mean Squared Error (MSE). An early stopping strategy was employed based on validation loss, with a patience threshold of three epochs.

All models were trained and tested under the same settings across nine forecast horizons covering 1 to 30 days, where each horizon $\ell \in \{24, 48, 72, 96, 120, 144, 168, 336, 720\}$ represents the length of the predicted sequence from $t + 1$ to $t + \ell$. For each horizon, the look-back window was dynamically set to $n = \ell + 24$, following existing finding that longer input sequences benefit long-horizon forecasting (Wang et al., 2024). A fixed patch length of 12 hours was used to capture diurnal patterns, and all models employed a single encoder and decoder layer to maintain architectural comparability.

Hyperparameter tuning was conducted using Optuna (Akiba et al., 2019), following identical procedures for all models. The search space was defined based on the official implementations and recommended settings of the respective baselines. Key hyperparameters included the embedding dimension $d \in \{2^5, 2^6, \dots, 2^9\}$ and the number of attention heads $n_{\text{heads}} \in \{4, 8, 16\}$. Additional implementation details are provided in Appendix A.

4.4. Evaluation metrics

Model performance is evaluated using four standard regression metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2), defined as follows:

$$\text{MSE} = \frac{1}{\ell} \sum_{k=1}^{\ell} (x_{t+k} - \hat{x}_{t+k})^2, \quad \text{MAE} = \frac{1}{\ell} \sum_{k=1}^{\ell} |x_{t+k} - \hat{x}_{t+k}|, \quad (20)$$

$$\text{MAPE} = \frac{100\%}{\ell} \sum_{k=1}^{\ell} \left| \frac{x_{t+k} - \hat{x}_{t+k}}{x_{t+k}} \right|, \quad R^2 = 1 - \frac{\sum_{k=1}^{\ell} (x_{t+k} - \hat{x}_{t+k})^2}{\sum_{k=1}^{\ell} (x_{t+k} - \bar{x})^2} \quad (21)$$

where x_{t+k} and $\hat{x}_{t+k}$ denote the ground truth and predicted values at future time step $t + k$, respectively, and $\bar{x}$ is the mean of the observed values.

To evaluate the accuracy of health warnings for peak concentrations, we employ the Critical Success Index (CSI) and False Alarm Ratio (FAR) (Wilks, 2011). As established metrics in atmospheric science, they quantify the model's success in capturing these peak values and its tendency to issue false warnings, respectively:

$$\text{CSI} = \frac{H}{H + M + FA}, \quad \text{FAR} = \frac{FA}{H + FA} \quad (22)$$

where H , M , and FA represent Hits, Misses, and False Alarms, respectively. Detailed definitions are provided in Appendix B.

Computational efficiency and deployment feasibility are evaluated using four metrics: (1) Parameters: total trainable weights; (2) FLOPs: floating-point operations per forward pass; (3) Inference Latency: average time required for a single inference step; and (4) Peak GPU Memory: maximum GPU memory footprint during inference. These metrics collectively quantify the trade-off between performance and computational cost.

5. Results and discussion

This section evaluates the proposed framework's forecasting performance, computational efficiency, and interpretability. Predictive accuracy is compared against five baselines, with assessments extending to

peak concentration detection and generalization tests. Further analyses quantify the impact of auxiliary variables via ablation studies and assess resource consumption. Finally, X2-Attention mechanism is validated by comparing its consistency and fidelity with the GradientSHAP benchmark.

5.1. Multi-horizon forecasting performance

5.1.1. Forecast visualization

Fig. 6 displays representative NO_x forecasts from the three top-performing models for 24-hour and 336-hour horizons. At the 24-hour horizon, all models captured the general trends but missed high-frequency fluctuations. At the 336-hour horizon, Crossformer and LSTM models consistently overestimated NO_x concentrations across multiple diurnal cycles at Folkungagatan and Hornsgatan. Conversely, the proposed model aligned closely with the observed concentrations. At Sveavägen, three models effectively captured periodic traffic-associated patterns. Additional forecast visualizations for all models regarding NO_x and PM_{10} are provided in Figs. C.14 and C.15.

5.1.2. Overall model performance

We evaluated forecasting performance for both NO_x and PM_{10} across nine horizons ranging from 24 to 720 hours. Table 2 summarize forecasting performance for NO_x and PM_{10} averaged across all street canyon sites. Detailed two-sided paired t-tests and corresponding effect sizes (Cohen's d) are provided in Tables C.11 and C.12.

The proposed model consistently outperformed all baseline models across nearly all horizons. For NO_x , it achieved significant improvements from 48 to 336 hours. Compared to Crossformer, which delivers the second-best performance, our model reduces the average MSE by 8.1%. For PM_{10} , the model obtains the lowest MAE for seven of the nine horizons. While the standard Transformer remains competitive for PM_{10} at specific horizons, our model maintained superior stability when averaged over all prediction lengths ($p < .01$ for both MAE and MSE). Notably, given the limited sample size ($n = 5$ independent runs), Cohen's d is employed to quantify the magnitude of improvement where p -values may lack sufficient statistical power. For instance, in 24-hour NO_x forecasting, our model achieves a substantial effect size ($d = 1.15$) relative to Crossformer, despite a marginal p -value (0.06).

Among the baseline models, the optimal architecture is highly pollutant-specific, revealing a trade-off between capturing temporal dependencies and modeling cross-variable correlations. Crossformer delivered the strongest baseline performance for NO_x , likely due to its joint modeling of temporal and cross-feature dependencies. This approach effectively captures the dual dynamics of NO_x driven by emission patterns and meteorological variables. In contrast, the Transformer proved more effective for PM_{10} . Since PM_{10} behavior is largely governed by autocorrelative structures and transient meteorological drivers, the pointwise attention mechanism of the Transformer proved more robust than complex cross-variate architectures.

Models prioritizing feature interactions over temporal structure, such as iTransformer and TimeXer, exhibited significant performance degradation. The iTransformer's inversion of attention compresses sequential signals essential for tracking emissions, while TimeXer's failure suggests that exogenous alignment cannot compensate for a weak endogenous temporal backbone. Preserving the integrity of the temporal structure is therefore essential for pollutant dynamics.

The LSTM model remained competitive up to 120-hour but significantly deteriorated at longer horizons. For NO_x at the 720-hour horizon, LSTM yields an MSE of 484 compared to 378 for our model. This degradation reflects the limited memory capacity of recurrent architectures (Bai et al., 2018; Qi et al., 2019). In our experimental setup, LSTM maintained stability up to 120 hours likely due to the integration of auxiliary meteorological and deterministic forecasts, which provided explicit forward-looking information. However, Transformer-based models generally surpass recurrent architectures beyond this point.

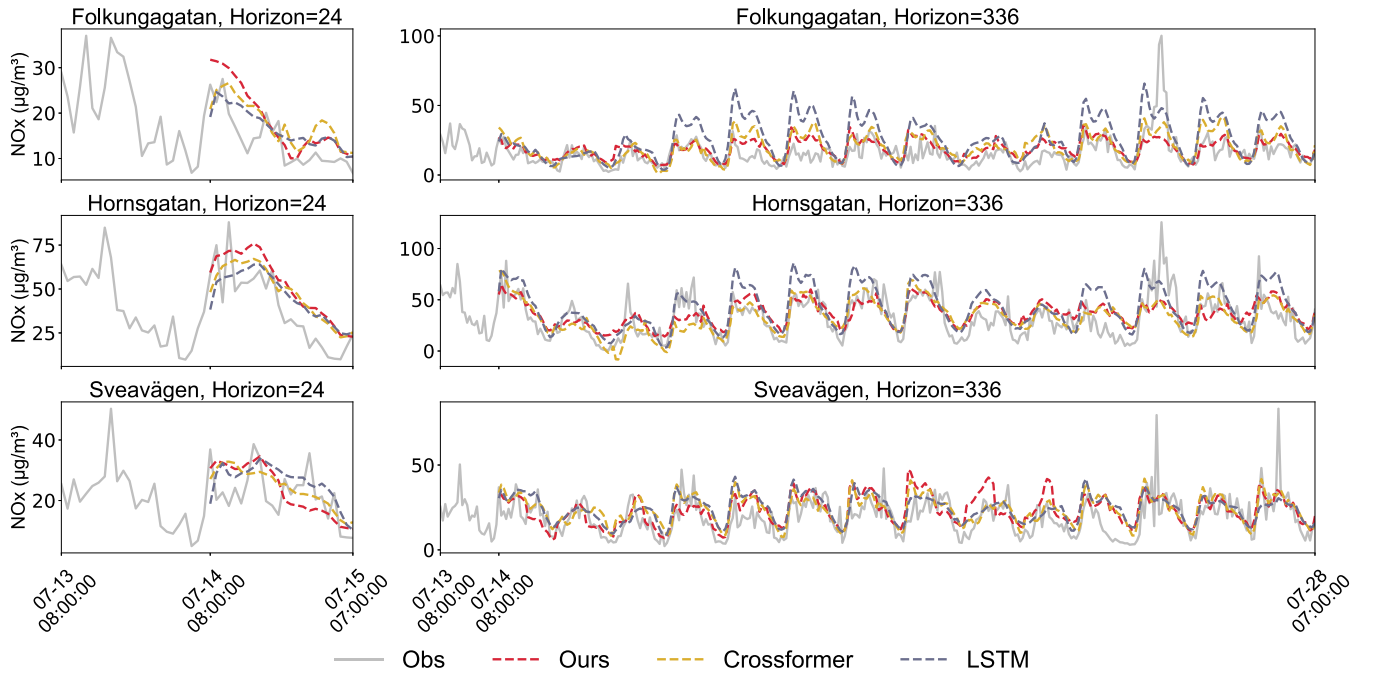


Fig. 6. Representative forecasts of NO_x concentrations at street canyon stations over 24-hour and 336-hour (14-day) horizons, starting from 08:00 on July 14, 2023. The gray solid line indicates the observed NO_x concentrations (Obs).

Table 2

Multi-horizon forecasting performance for NO_x and PM_{10} , averaged across all monitoring stations over five independent runs. For each forecast horizon, the best-performing model is highlighted in bold, and the second-best is underlined. The 'AVG' reports the mean performance across all horizons, while "1st Count" indicates the number of first-place ranks a model achieved across all forecast horizons. The units for MAE and MSE are $\mu\text{g}/\text{m}^3$ and $(\mu\text{g}/\text{m}^3)^2$, respectively.

Models	Metric	Ours		Crossformer		TimeXer		iTransformer		Transformer		LSTM	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
NO _x	24	10.68	265.19	10.85	268.43	12.59	365.51	12.18	354.29	10.91	275.73	<u>10.75</u>	265.50
	48	11.18*	278.47	12.24	316.82	13.36	403.26	12.77	377.15	12.05	313.67	<u>11.22</u>	<u>279.95</u>
	72	11.24*	283.88*	12.59	339.12	13.54	398.17	13.05	382.29	12.12	321.07	<u>11.51</u>	<u>292.75</u>
	96	11.63**	298.11**	12.42	330.32	13.42	384.85	12.98	378.44	13.10	361.51	<u>12.15</u>	<u>324.56</u>
	120	12.10**	318.24*	13.24	366.65	13.29	375.97	13.25	392.66	13.76	381.86	<u>12.77</u>	<u>351.39</u>
	144	12.61	331.65*	<u>12.88</u>	<u>350.25</u>	13.35	377.16	13.49	406.04	13.85	387.23	13.03	359.68
	168	12.71*	340.22*	<u>13.21</u>	<u>357.77</u>	13.36	378.12	13.73	412.22	13.54	380.10	13.21	363.49
	336	13.01**	358.22**	<u>13.89</u>	<u>391.43</u>	13.49	393.14	14.19	427.79	13.95	396.90	14.08	403.20
	720	13.53	378.39	<u>13.75</u>	<u>383.43</u>	14.50	446.96	15.21	470.27	14.91	439.04	15.45	484.52
	AVG	12.08**	316.93**	12.78	<u>344.91</u>	13.43	391.46	13.43	400.13	13.13	361.90	<u>12.68</u>	347.23
PM ₁₀	1 st Count	9	9	0	0	0	0	0	0	0	0	0	0
	24	4.02**	44.06**	4.34	47.90	4.48	51.19	4.51	52.59	4.57	51.20	<u>4.15</u>	<u>46.94</u>
	48	4.38**	51.11**	4.60	<u>52.32</u>	4.87	58.18	5.00	62.81	4.73	54.80	<u>4.44</u>	<u>52.70</u>
	72	<u>4.57</u>	53.53	5.30	64.25	5.04	61.68	5.19	68.07	4.58	53.58	4.52	<u>53.80</u>
	96	4.53**	53.04**	4.89	57.86	5.04	61.24	5.11	62.67	4.92	<u>57.35</u>	<u>4.73</u>	58.01
	120	4.75**	56.89	5.17	61.20	5.08	62.30	5.17	63.56	5.17	59.91	<u>4.92</u>	<u>59.67</u>
	144	4.84*	<u>58.70</u>	5.61	70.26	5.16	63.86	5.29	66.15	<u>5.01</u>	58.36	5.01	60.65
	168	4.87	<u>58.54</u>	5.36	64.74	5.13	63.69	5.26	65.86	<u>4.94</u>	56.89	4.96	58.72
	336	4.94	<u>53.74</u>	5.27	60.74	5.04	57.50	5.37	61.52	<u>4.99</u>	53.58	5.09	55.48
	720	5.22	<u>50.34</u>	5.98	70.10	<u>5.12</u>	53.36	5.43	56.57	5.07	48.34	5.74	57.52
AVG		4.68**	53.33**	5.17	61.04	5.00	59.22	5.15	62.20	4.89	<u>54.89</u>	<u>4.84</u>	55.94
	1 st Count	7	5	0	0	0	0	0	0	1	4	1	0

*/**: Ours outperforms the second-best model at $p < .05/0.01$ using two-sided paired t-test.

Second-best models: Crossformer for NO_x and Transformer for PM_{10} .

To standardize the comparison across pollutants, we evaluated relative error using R^2 and MAPE, as shown in Fig. 7 and Table C.4. For NO_x , which is dominated by periodic traffic emissions (Karner et al., 2010), our model maintained R^2 values above 0.2 even at 720 hours. In contrast, PM_{10} exhibited lower temporal persistence due to transient factors like road surface moisture and dust (re-)suspension (Denby et al., 2013). Consequently, while the model achieved an R^2 of approximately 0.4 at 24 hours, performance declines rapidly, approaching zero by 336 hours and yielding negative R^2 values by 720 hours. MAPE followed similar trends.

Location-specific analyses provided in Tables C.5–C.10 further confirms these findings. For NO_x , the proposed framework reduced MSE by 3.6%, 8.8%, and 6.6% compared to the second-best model at Folkungagatan, Hornsgatan, and Sveavägen, respectively. Similar trends were observed for PM_{10} , although the magnitude of improvement varied by location due to local environmental conditions.

5.1.3. Accuracy in peak concentrations

Forecasting peak pollutant concentrations is critical for issuing public health warnings. We defined peak concentrations as values ex-

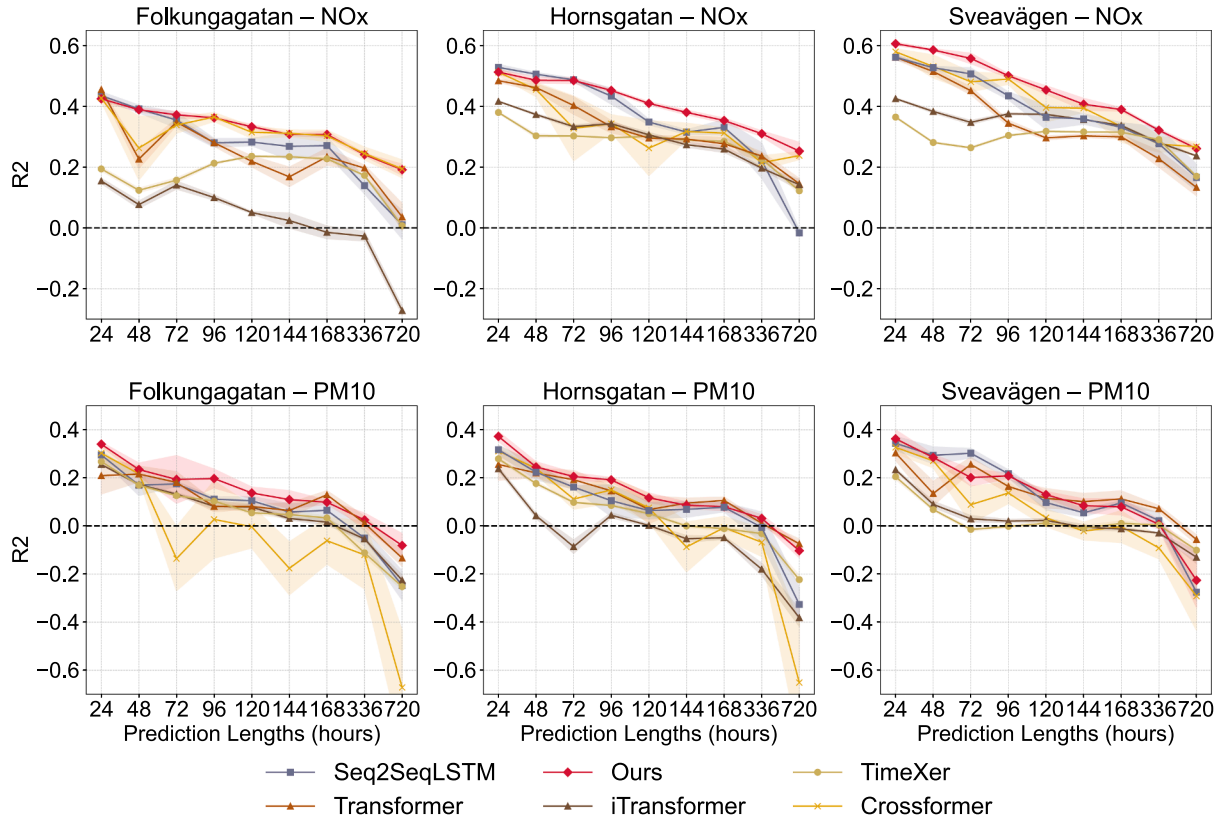


Fig. 7. R^2 for each pollutant at each station across nine forecasting horizons. Results and their 95% confidence intervals are computed on the test set over five independent experiments. The x-axis uses equal spacing across horizons. Negative R^2 values indicate performance worse than the mean predictor and lack practical utility.

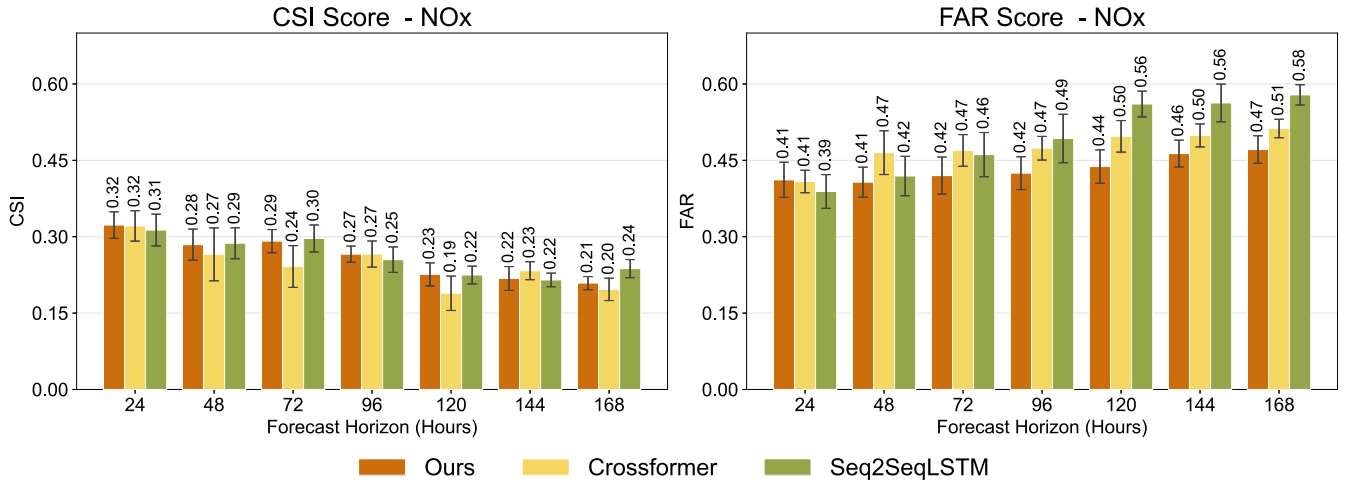


Fig. 8. Detection accuracy for peak NO_x concentrations measured by CSI and FAR scores. Results show the average across all stations and five independent runs. Corresponding mean values are labeled on each bar.

ceeding the 90th percentile of the test set. Given the actionable lead time for interventions and predictive reliability, we evaluate accuracy within a 168-hour horizon (7 days) using CSI and FAR. Higher CSI values indicate superior event detection, while lower FAR values denote fewer false alarms. Fig. 8 illustrates the evolution of peak prediction accuracy across forecasting horizons for NO_x , comparing our model against the second-best baselines, Crossformer and LSTM.

The proposed framework demonstrates a superior balance between detection sensitivity and precision. For NO_x , CSI results indicate that

our model remains competitive with the LSTM and outperforms the Crossformer model. Concurrently, Our model consistently achieves the lowest false alarm rates across all horizons and maintains values between 0.40 and 0.47. The baseline models exhibit a significant increase in false alarms as the forecast horizon extends, particularly LSTM. Minimizing false positives is essential for operational air pollution warning systems to prevent erroneous interventions and public warning fatigue. For PM_{10} , the Transformer baseline yielded the highest event detection scores, whereas our model maintained the lowest false warning rates, as detailed in Fig. C.16.

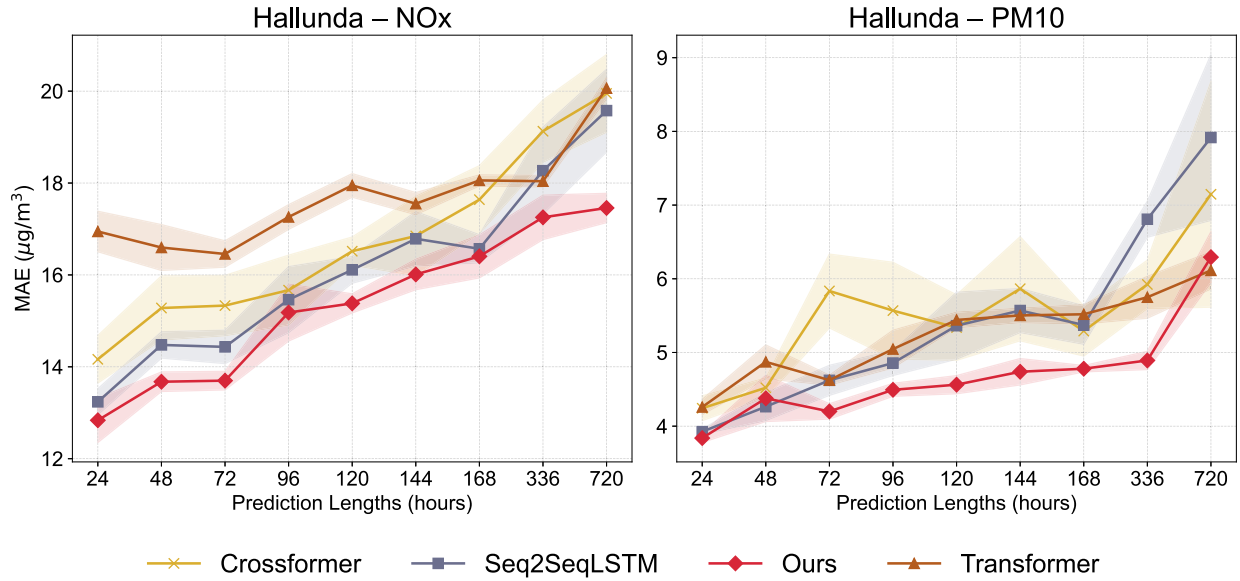


Fig. 9. MAE for each pollutant at Hallunda across nine forecasting horizons. Results and their 95% confidence intervals are computed on the test set over five independent experiments. The x -axis uses equal spacing across horizons.

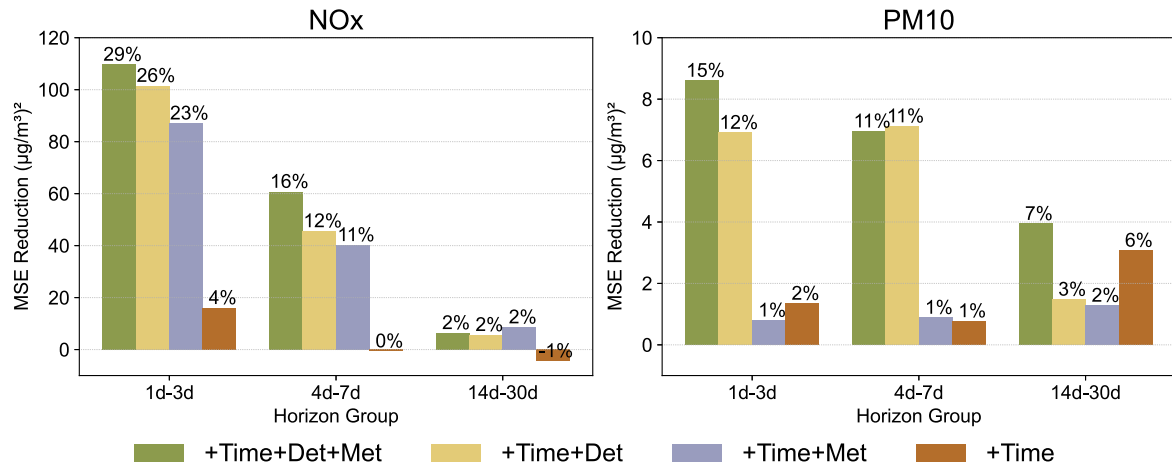


Fig. 10. Ablation study on the contribution of forecasted auxiliary variables for NO_x (left) and PM_{10} (right). The bars illustrate the absolute and percentage MSE reduction relative to the historical-only (OH) baseline. Results are averaged across all stations and grouped by forecast horizons: 1–3 days, 4–7 days, and 14–30 days.

5.1.4. Generalization to suburban station

We extended our evaluation under distributional shifts at the Hallunda (HA) suburban station. The results of Kolmogorov-Smirnov tests confirm a significant distributional shift between the suburban and urban datasets ($p < .001$), while diurnal analysis reveals divergent patterns between highway commuting and city traffic, as detailed in Fig. D.18 and D.19. These discrepancies make this station a robust out-of-distribution testbed.

Fig. 9 summarizes the MAE comparison against the leading baselines identified in the urban experiments, while Tables D.13, D.14, and D.15 provide a comprehensive comparison of MSE, R^2 , and MAPE metrics, along with significance tests. Despite the distinct emission and diffusion mechanisms of the open highway terrain, our proposed framework demonstrates remarkable robustness. For NO_x , our model achieved the lowest MAE across all forecasting horizons and maintained a positive R^2 even at the challenging 720-hour mark where other models degrade significantly. For PM_{10} , our model achieves superior performance in nearly all horizons with significantly lower MSE and MAPE, although

the Transformer still remained competitive at specific intervals. The architecture effectively captured the regular yet high-speed traffic patterns of suburban highways without overfitting to the stochastic turbulence typical of urban street canyons.

5.2. Diagnostics and efficiency analysis

5.2.1. Contribution of forecasted variables

We conducted an ablation study to quantify the contribution of forecasted auxiliary variables. Each input type was incrementally added to a baseline model (OH), which relies exclusively on historical records. The configurations include future temporal encodings (+Time), meteorological forecasts (+Time + Met), and deterministic model outputs (+Time + Det), culminating in a full combination (+Time + Det + Met).

Fig. 10 illustrates that forecasted variables substantially improve performance, particularly within 1–3 day horizon. Introducing future temporal encodings (+Time) alone leads to modest MSE reductions. The addition of meteorological forecasts (+Time + Met) further reduces MSE

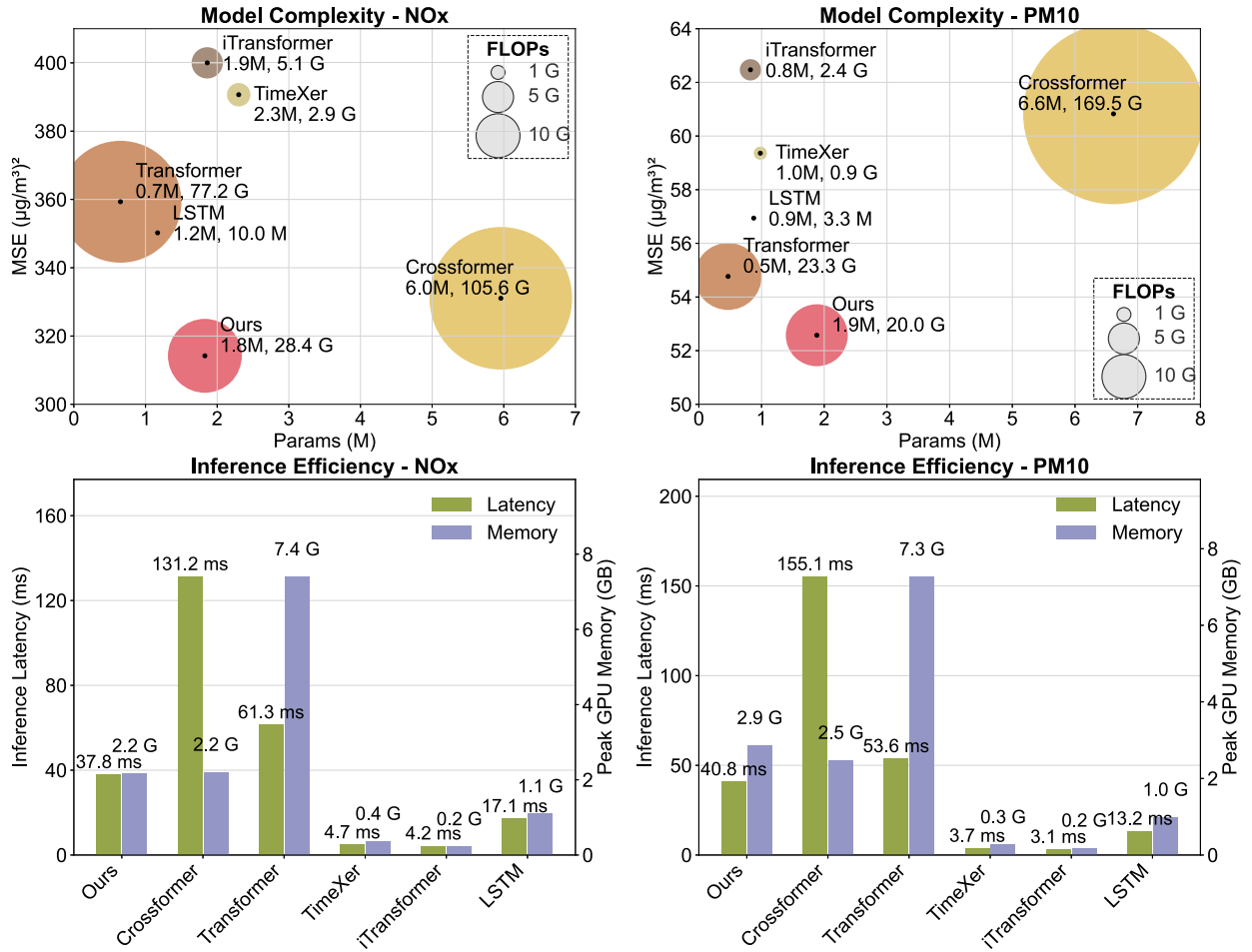


Fig. 11. Comparison of forecasting performance and computational efficiency for NO_x (left) and PM_{10} (right). Top: Model complexity analysis visualizing the trade-off between prediction error and parameter count (G: Giga, M: Mega), where bubble size indicates FLOPs. Bottom: Inference efficiency measured by latency and peak GPU memory usage.

for NO_x by 23%, though contributing marginally for PM_{10} . In contrast, deterministic model outputs (+Time+Det) yield notable improvements for both pollutants, with 26% for NO_x and 12% for PM_{10} . Physics-based deterministic forecasts are particularly valuable within 3 day forecasting. The full configuration (+Time+Det+Met) achieved the best performance by reducing MSE by 29% for NO_x and 15% for PM_{10} . This highlights the complementary contribution among different forecast sources.

As forecast horizon extends to 4–7 days, a period that surpasses the 72-hour availability of the auxiliary forecasts, the direct benefit diminishes. Nevertheless, this limited-horizon future information continued to enhance performance. The full model maintained a 16% and 11% MSE reduction for NO_x and PM_{10} , respectively. The model leverages autocorrelation structures, allowing early-window information to propagate to later timesteps. At 14–30 day horizons, performance converged toward the historical-only baseline.

5.2.2. Computational efficiency

Beyond predictive accuracy, computational efficiency is a critical factor for practical deployment. Fig. 11 details model complexity and inference efficiency. The results are averaged across all stations and forecast horizons.

The proposed model exhibits a favorable balance between model complexity and predictive performance. For NO_x , it employs a compact architecture with only 1.8 M parameters and 28.4 GFLOPs. This

is significantly more efficient than Crossformer (6.0 M parameters, 105.6 GFLOPs) and standard Transformer (77.2 GFLOPs). For PM_{10} , our model achieved lower computational cost than the Transformer (20.0 GFLOPs vs. 23.3 GFLOPs) while delivering higher predictive accuracy.

Inference latency and memory footprint further validate deployment feasibility. For NO_x , the standard Transformer encounters a memory bottleneck with a peak usage of 7.4 GB, whereas our model requires only 2.2 GB. The average inference latency of our model is 37.8 ms, significantly faster than second-best models (Crossformer, 131.2 ms). While lightweight baselines like TimeXer achieve lower latency and reduced memory usage, they typically compromise long-horizon forecasting accuracy. Similar efficiency trends were observed for PM_{10} . This efficiency of our model is primarily attributed to patch-based embedding and cross-attention mechanism, which facilitate superior accuracy alongside manageable computational costs.

5.3. Evaluation of interpretability

The reliability of the interpretability provided by the X2-Attention mechanism is validated through a comparative evaluation against GradientSHAP benchmark. This assessment focuses on the consistency of feature attribution and the faithfulness of the generated explanations via perturbation analysis. To facilitate detailed temporal insights, 72-

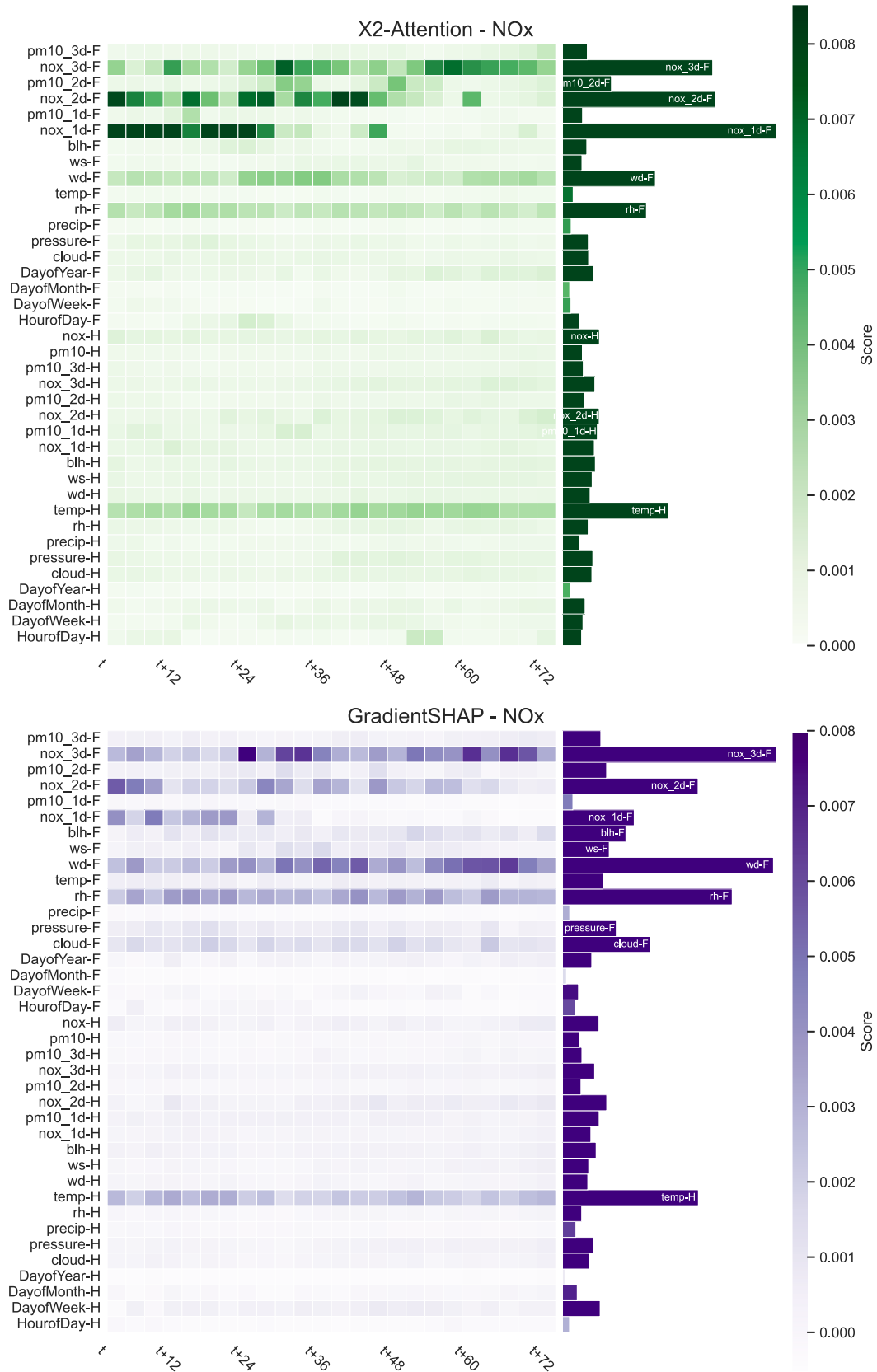


Fig. 12. Comparison of feature attribution for NO_x 72-hour forecasting at Sveavägen. The upper and lower panels correspond to the proposed X2-Attention mechanism and the GradientSHAP method, respectively. Left: Temporal heatmaps of feature importance. Right: Aggregated global feature rankings.

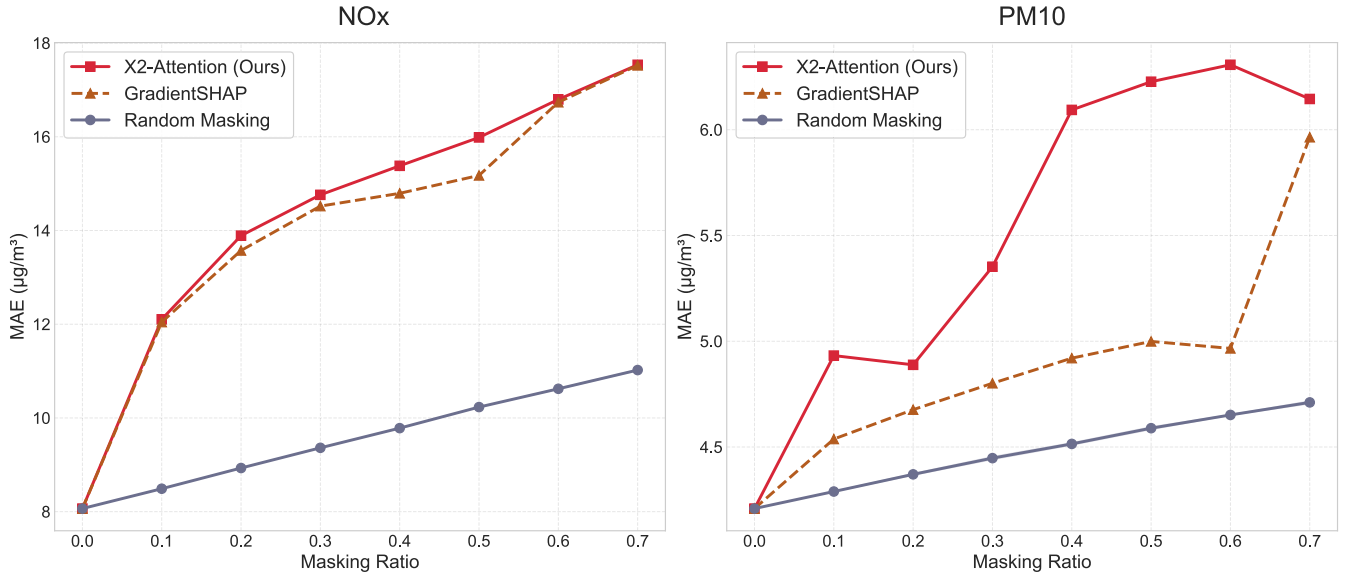


Fig. 13. Empirical fidelity evaluation via perturbation analysis at Sveavägen.

Table 3

Correlation between X2-Attention and GradientSHAP attribution scores.

Pollutant	Pearson Correlation (p-value)	Kendall's Tau (p-value)
NO _x	0.61 ($p < .001$)	0.60 ($p < .001$)
PM ₁₀	0.62 ($p < .001$)	0.78 ($p < .001$)

hour forecasts are partitioned into 3-hour segments, enabling attribution analysis at a fine-grained resolution .

5.3.1. Consistency with gradientshap

The X2-Attention mechanism provides intrinsic interpretability by generating attribution scores directly during inference. As defined in Eq. 17, it produces an attribution matrix that quantifies the relevance of each input segment to specific forecast patches. In contrast, GradientSHAP (Lundberg, 2017) is a post-hoc method that estimates feature contributions by integrating gradients across multiple randomized input perturbations. We averaged the attribution scores across all test instances to derive a representative importance profile, as shown in Fig. 12. The heatmaps visualize the temporal evolution of feature importance, where darker shades indicate higher influence, while the bar charts rank the global features ranking aggregated over the forecast window.

Both methods consistently capture similar temporal patterns, as evidenced by the comparative heatmaps. For NO_x, dominant contributors include deterministic forecasts for the upcoming days (no_1d-F, nox_2d-F, nox_3d-F), forecasted wind direction (wd-F), and historical temperature (temp-H). These results align with the established urban pollution dynamics, where deterministic forecasts encapsulate emission patterns and meteorological factors govern pollutant dispersion (Zhang et al., 2024). For PM₁₀, as detailed in Fig. E.20, both methods show a clear decay in the relevance of past measurements as the forecast horizon increases. The global feature rankings further underscore this consistency. Six of the top ten NO_x features identified by X2-Attention appear in the GradientSHAP ranking, and the overlap is even stronger for PM₁₀, where nine of the top ten features identified by X2-Attention match those of GradientSHAP.

We further quantified this agreement using two metrics: Pearson correlation (Sedgwick, 2012) on flattened attribution scores to assess the

overall similarity in their importance distributions, and Kendall's Tau (Arndt et al., 1999) to evaluate the consistency of global feature rankings.

As summarized in Table 3, both pollutants exhibit strong, statistically significant correlations. For NO_x, the Pearson correlation and Kendall's Tau reached 0.61 and 0.60, respectively. This agreement is even more pronounced for PM₁₀, with coefficients of 0.62 and 0.78. These findings confirm that X2-Attention generates explanations highly consistent with established GradientSHAP benchmarks, suggesting its plausibility as an attribution tool.

5.3.2. Fidelity evaluation via perturbation analysis

To quantify the empirical fidelity of the generated explanations, we conducted a faithfulness evaluation using the Most Relevant First (MoRF) strategy (Hooker et al., 2019; Samek et al., 2016). This involves iteratively masking the top $k\%$ of the most critical input segments identified by each interpretability method and replacing them with the global mean of the training set. A steeper increase in prediction error indicates that the identified temporal segments of individual features were more essential to the model's predictive accuracy.

Fig. 13 compares the degradation curves of X2-Attention against GradientSHAP and a random masking baseline across masking ratios from 0.0 to 0.7. Removing features identified by X2-Attention leads to the most significant performance drop for both NO_x and PM₁₀. Notably, the X2-Attention curve consistently remains above the GradientSHAP curve, indicating superior precision in localizing influential input patterns upon which the model relies. This contrast is particularly pronounced in the PM₁₀ analysis, where X2-Attention induces a sharp rise in error even at low masking ratios. Both methods significantly outperform the random baseline, confirming that the identified high-attention regions are meaningful drivers of prediction rather than noise.

Beyond consistency and fidelity, a pivotal advantage of the proposed X2-Attention is its computational efficiency. Unlike post-hoc methods like GradientSHAP that rely on intensive sampling and repetitive gradient computations, X2-Attention achieves built-in interpretability. While generating equivalent feature attributions using GradientSHAP required approximately 40 hours on a NVIDIA A40 GPU, X2-Attention produces these scores without any additional computational cost. This efficiency makes our framework well-suited for practical applications with constrained resources.

6. Conclusion and future work

This study proposes a novel Transformer-based framework designed for interpretable multi-horizon air pollution forecasting. The architecture integrates the Structured Temporal Embedding Module (STEM) and the Target-Oriented Cross-Attention (X2-Attention) mechanism. This design effectively harmonizes heterogeneous inputs, including historical observations and multi-source forecasts, to enhance long-horizon predictive performance. A distinctive advantage of this framework lies in delivering advanced predictive accuracy alongside plausible, fine-grained interpretability for each forecasted timestep across both temporal and feature dimensions.

Extensive evaluations across diverse urban and suburban environments in Stockholm demonstrate superior performance of the proposed model. Compared to five competitive baselines, the framework achieves significant reductions in averaged prediction error across nine horizons (24–720h). For NO_x , the model maintains meaningful predictive skill with an R^2 value above 0.2 even at the 30-day horizon, while PM_{10} reliability is sustained for up to 14 days. Furthermore, ablation studies underscore the value of incorporating auxiliary forecasts for improving predictive accuracy. The inclusion of meteorological predictions and deterministic model outputs yields average MSE reductions of 29% for NO_x and 15% for PM_{10} within 72-hour window. Although the predictive benefit diminishes beyond the temporal coverage of these forecasted inputs, their positive influence persists for up to 7-day.

The pivotal contribution of this study lies in its built-in, fine-grained interpretability across the forecast window. Unlike computationally intensive post-hoc methods, the X2-Attention mechanism provides feature attributions without additional overhead. These attributions were rigorously validated through perturbation analysis and exhibited strong consistency with the GradientSHAP benchmark. Notably, while the identified attention weights provide a plausible proxy for the model's decision logic, they primarily reflect empirical associations and do not inherently establish the underlying physical causality of pollutant dispersion. Future work will incorporate counterfactual interventions and other causal inference frameworks to further disentangle the complex physical drivers of pollutant dispersion.

Furthermore, we aim to enhance the framework's generalization by validating it across diverse climatic zones and incorporating high-resolution spatial proxies, such as traffic dynamics and satellite imagery. Translating these technical insights into intuitive visualization tools will be a priority, ultimately bridging the gap between complex deep learning architectures and evidence-based environmental governance.

Author contributions

X.M. conceived the projects and provided overall supervision of the study. Z.Z. carried out the primary work, including data processing, formal analysis, modeling, and visualization, and wrote the manuscript. M.E., D.S., C.J., and X.M. participated in ongoing discussions and made significant revisions. All authors contributed to the final manuscript by reading and approving the submitted version.

Data and Code availability

The data and code used in this paper can be obtained upon request from the corresponding author (liang@kth.se).

Competing Interests

The authors declare no competing interests.

Data availability

Data will be made available on request.

Declaration of competing interest

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

Acknowledgments

This research was supported by the Centre for Digital Futures through the iHorse+ 2024 and iHorse 2021 projects, as well as by Richterska Stiftelsen (grant no. 2024-00992). The authors acknowledge Chalmers e-Commons at Chalmers University of Technology for providing computational resources. Computational support was also provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council under grant agreement no. 2022-06725.

Appendix A. Additional Training Details

During preliminary trials, a default batch size of 512 was used to improve computational efficiency. However, at the longest forecast horizons ($\ell = 336$ and 720 hours), memory constraints caused out-of-memory (OOM) errors for Transformer and Crossformer models. To address this, the batch size was reduced to 128 and 256, respectively, for those specific combinations.

To maintain stable training dynamics under varying batch sizes, we adopted the Linear Scaling Rule (Goyal et al., 2017), which suggests that the learning rate should scale proportionally with the batch size:

$$\text{learning_rate} \propto \text{batch_size}$$

We defined a base learning rate $lr_{base} = 1 \times 10^{-4}$ associated with a reference batch size $bs_{base} = 64$. For each trial with batch size bs_{new} , the target learning rate was computed as:

$$lr_{target} = lr_{base} \times \frac{bs_{new}}{bs_{base}}$$

Optuna then sampled the actual learning rate within the range $[lr_{target}, 10 \times lr_{target}]$ on a logarithmic scale.

Appendix B. Metrics for Peak Concentration Assessment

We evaluate the performance of peak warnings using a binary classification framework. Time points are categorized as events or non-events based on whether the pollutant concentration exceeds a specific threshold. This categorization yields four outcomes (Wilks, 2011):

1. **Hit (H)**: Both forecast and observation exceed the threshold.
2. **Miss (M)**: The observation exceeds the threshold, but the model predicts a value below it, signifying a failure to issue a necessary warning.
3. **False Alarm (FA)**: The model predicts a peak event that does not occur in observations.
4. **Correct Rejection (CN)**: Both forecast and observation remain below the threshold.

While this framework parallels the standard confusion matrix, specific metrics are chosen to address the imbalanced nature of environmental extremes where calm periods (CN) predominate. We exclude the standard Accuracy metric ($\frac{H+CN}{H+M+FA+CN}$) because it is heavily dominated by the large number of CN cases. In contrast, the CSI metric ($\frac{H}{H+M+FA}$) excludes CN, providing a more robust assessment of the model's ability to capture rare peak events.

Appendix C. Supplementary Results for Urban Stations

C.1. Forecast visualizations

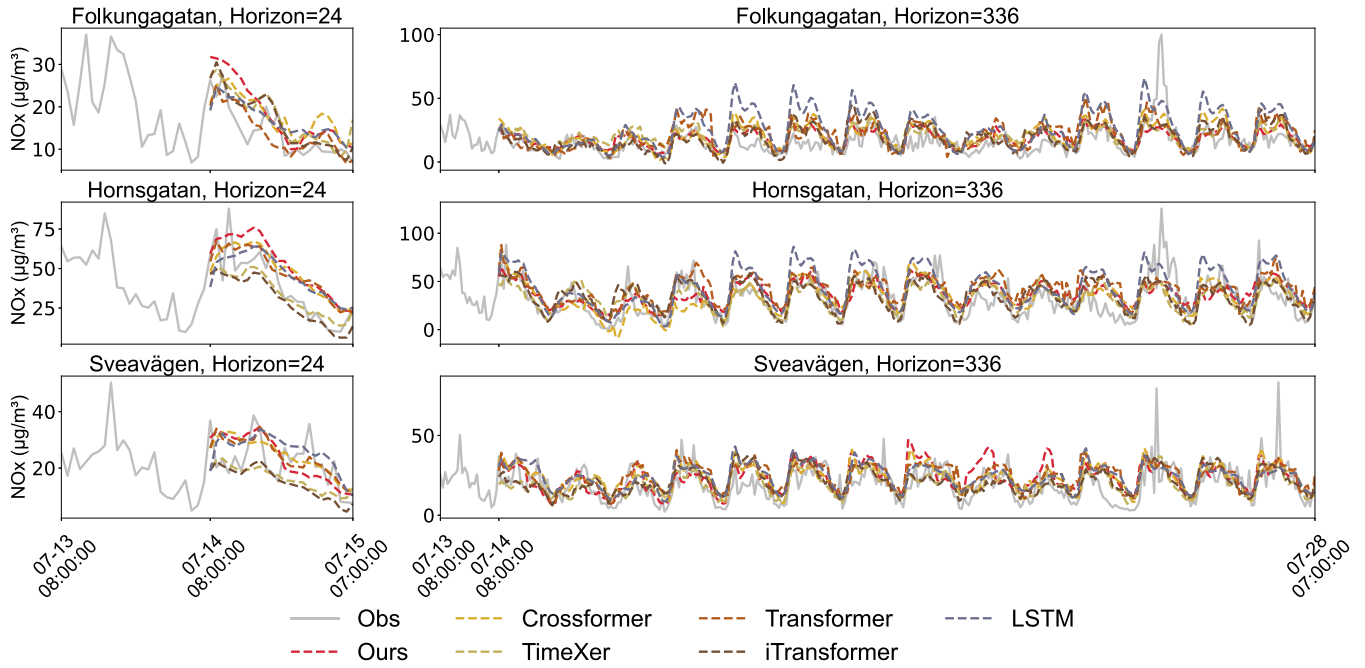


Fig. C.14. Representative forecasts of NO_x concentrations at three stations for all models, starting from 08:00 on July 14, 2023, over 24-hour and 336-hour (14-day) horizons. The gray solid line indicates the measured NO_x concentrations.

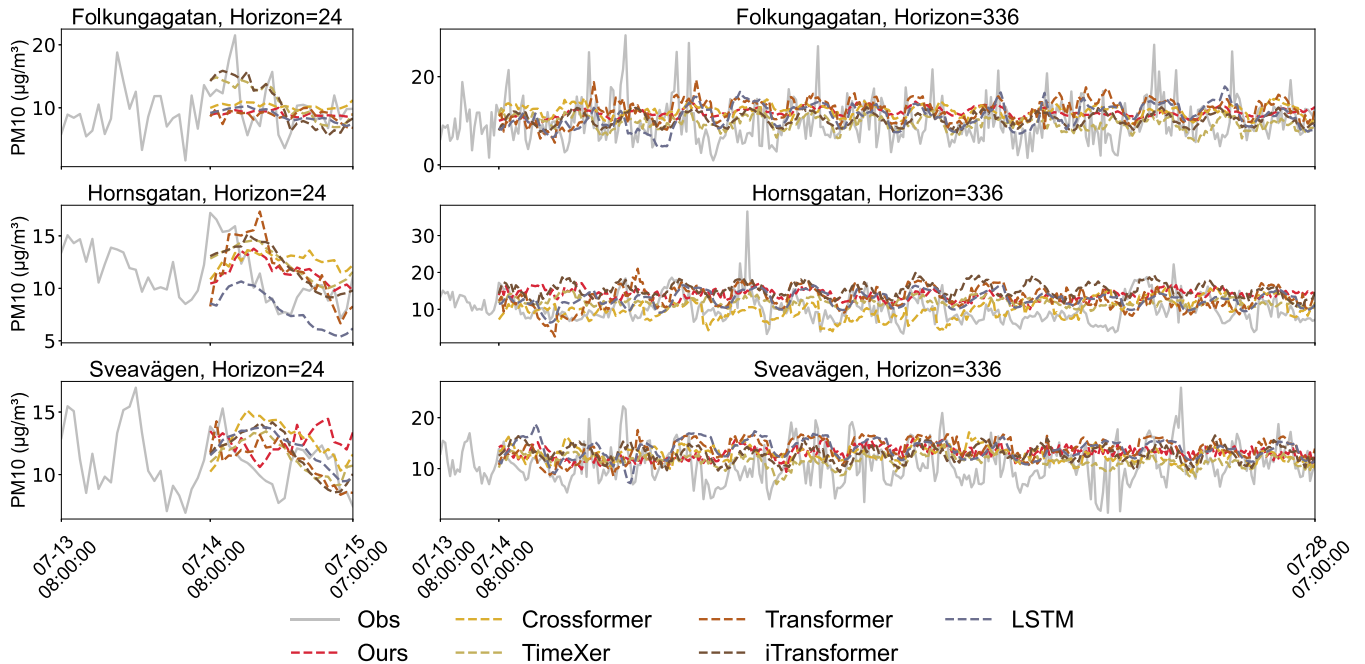


Fig. C.15. Representative forecasts of PM_{10} concentrations at three stations for all models, starting from 08:00 on July 14, 2023, over 24-hour and 336-hour (14-day) horizons. The gray solid line indicates the measured NO_x concentrations.

C.2. Detailed performance metrics

Table C.4

Multi-horizon forecasting performance for NO_x and PM₁₀, averaged across all monitoring stations over five independent runs. For each forecast horizon, the best-performing model is highlighted in bold, and the second-best is underlined. The ‘AVG’ reports the mean performance across all horizons, while “1_{st} Count” indicates the number of first-place ranks a model achieved across all forecast horizons.

Models	Metric	Ours		Crossformer		TimeXer		iTransformer		Transformer		LSTM	
		R ²	MAPE	R ²	MAPE	R ²	MAPE	R ²	MAPE	R ²	MAPE	R ²	MAPE
NO _x	24	0.52	<u>0.94</u>	<u>0.51</u>	0.96	0.31	0.98	0.33	0.93	0.50	1.06	<u>0.51</u>	0.98
	48	0.49	<u>1.03*</u>	0.41	1.14	0.24	1.04	0.28	<u>1.01</u>	0.41	1.09	<u>0.47</u>	0.97
	72	0.47*	<u>1.07</u>	0.39	1.20	0.24	<u>1.07</u>	0.27	1.09	0.40	1.17	<u>0.44</u>	1.04
	96	0.44**	1.13	<u>0.40</u>	1.15	0.27	<u>1.06</u>	0.27	1.13	0.32	1.29	0.38	1.05
	120	0.40	<u>1.15*</u>	<u>0.33</u>	1.21	0.28	1.08	0.24	1.19	0.27	1.32	<u>0.33</u>	1.17
	144	0.37	<u>1.18</u>	<u>0.34</u>	1.18	0.28	1.10	0.21	1.24	0.25	1.28	<u>0.32</u>	1.21
	168	0.35	1.18	<u>0.32</u>	<u>1.17</u>	0.28	1.14	0.20	1.32	0.27	1.32	0.30	1.18
	336	0.29**	1.30	<u>0.24</u>	1.33	0.23	<u>1.31</u>	0.15	1.44	0.22	1.59	0.21	1.38
	720	0.24	<u>1.07</u>	<u>0.23</u>	1.03	0.10	1.09	0.03	1.18	0.11	1.25	0.05	1.20
	AVG	0.40**	<u>1.12**</u>	<u>0.35</u>	1.15	0.25	1.10	0.22	1.17	0.31	1.26	0.34	1.13
	1 st Count	9	1	0	1	0	3	0	1	0	0	0	3
PM ₁₀	24	0.36**	0.41**	0.31	0.45	0.25	0.48	0.24	0.48	0.26	0.48	<u>0.32</u>	<u>0.43</u>
	48	0.26**	0.46	<u>0.23</u>	0.48	0.14	0.53	0.10	0.53	0.19	0.49	<u>0.23</u>	<u>0.47</u>
	72	<u>0.20</u>	0.48	0.00	0.52	0.07	0.54	0.02	0.55	0.21	0.49	0.22	<u>0.48</u>
	96	0.20**	0.47**	0.11	<u>0.50</u>	0.06	0.55	0.05	0.56	0.13	0.54	<u>0.14</u>	0.51
	120	0.13	0.51**	0.05	0.55	0.04	0.57	0.03	0.58	<u>0.09</u>	0.60	<u>0.09</u>	<u>0.54</u>
	144	0.09	0.51**	-0.10	<u>0.57</u>	0.01	0.58	-0.01	0.59	<u>0.09</u>	0.58	0.06	<u>0.57</u>
	168	<u>0.09</u>	0.53*	-0.02	0.58	0.01	0.58	-0.02	0.59	0.11	0.57	0.08	<u>0.55</u>
	336	<u>0.02</u>	<u>0.60</u>	-0.11	<u>0.60</u>	-0.05	0.57	-0.09	0.63	0.04	0.62	-0.01	0.64
	720	<u>-0.13</u>	0.71	-0.55	<u>0.68</u>	-0.19	0.59	-0.24	0.69	-0.09	0.69	-0.28	0.80
	AVG	0.13*	0.52**	-0.01	<u>0.55</u>	0.04	<u>0.55</u>	0.01	0.58	<u>0.11</u>	0.56	0.09	<u>0.55</u>
	1 st Count	5	7	0	0	0	2	0	0	3	0	1	0

Table C.5

Multi-horizon forecasting performance for NO_x and PM₁₀ at Folkungagatan station, averaged over five independent runs. The best result for each horizon is shown in **bold**, and the second-best is underlined. “AVG” denotes the average performance across all horizons. “1_{st} Count” indicates how many times each model achieved the best result across all horizons for the corresponding pollutant. The units for MAE and MSE are $\mu\text{g}/\text{m}^3$ and $(\mu\text{g}/\text{m}^3)^2$, respectively.

Models	Metric	Ours		Crossformer		TimeXer		iTransformer		Transformer		LSTM	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
NO _x	24	9.38	232.00	9.41	230.79	11.07	329.75	10.76	346.74	<u>9.10</u>	221.71	9.07	<u>229.50</u>
	48	<u>10.03</u>	245.81	11.74	311.18	11.39	355.34	11.41	372.69	11.47	306.65	9.75	<u>245.96</u>
	72	9.84	252.74	10.55	267.59	11.55	339.53	11.36	346.86	10.14	261.22	<u>10.06</u>	<u>260.60</u>
	96	9.99	<u>254.52</u>	<u>10.10</u>	252.49	11.23	314.94	11.50	358.59	10.95	285.84	10.67	286.46
	120	10.21	262.35	<u>10.46</u>	<u>270.88</u>	11.00	300.51	11.79	374.44	11.62	307.43	10.64	281.61
	144	<u>10.45</u>	269.69	10.38	268.26	10.99	298.47	11.97	386.04	12.06	324.39	10.91	281.50
	168	10.53	267.26	<u>10.55</u>	<u>271.29</u>	11.02	298.67	12.28	390.13	10.92	294.77	11.10	282.67
	336	10.66	<u>281.34</u>	<u>10.70</u>	278.55	11.13	306.31	12.61	380.87	11.10	300.56	12.15	319.88
	720	10.90	290.97	<u>11.08</u>	<u>294.22</u>	12.39	357.88	14.73	461.49	11.95	346.25	12.84	354.29
	AVG	10.22	261.85	<u>10.55</u>	<u>271.69</u>	11.31	322.38	12.05	379.76	11.03	294.31	10.80	282.50
	1 st Count	6	5	1	3	0	0	0	0	0	1	2	0
PM ₁₀	24	3.83	29.54	4.02	<u>31.09</u>	4.19	32.56	4.24	33.09	4.30	34.62	<u>3.92</u>	31.36
	48	4.09	33.37	4.30	<u>34.50</u>	4.48	36.42	4.48	36.26	4.37	35.81	<u>4.26</u>	36.09
	72	4.21	34.50	5.21	49.15	4.56	37.44	4.58	37.23	4.35	<u>34.93</u>	<u>4.24</u>	35.23
	96	4.20	33.76	4.65	39.84	4.59	37.55	4.69	38.51	4.60	38.27	<u>4.37</u>	<u>37.03</u>
	120	4.34	35.67	4.76	40.17	4.69	38.81	4.65	37.87	4.66	37.81	<u>4.37</u>	<u>36.55</u>
	144	4.42	36.28	5.31	49.69	4.69	38.48	4.75	39.15	4.62	<u>37.61</u>	<u>4.51</u>	37.77
	168	4.42	<u>36.02</u>	4.89	42.07	4.72	38.68	4.76	39.33	4.46	35.06	<u>4.44</u>	37.07
	336	4.47	34.91	4.90	41.74	4.83	39.92	4.76	37.98	<u>4.55</u>	<u>35.27</u>	4.68	37.93
	720	4.58	34.88	5.54	53.09	4.88	40.48	4.95	39.64	<u>4.72</u>	<u>36.44</u>	<u>4.93</u>	<u>39.85</u>
	AVG	4.28	34.33	4.842	42.37	4.63	37.81	4.65	37.67	<u>4.52</u>	<u>36.20</u>	4.41	36.54
	1 st Count	9	8	0	0	0	0	0	0	0	1	0	0

Table C.6

Multi-horizon forecasting performance for NO_x and PM₁₀ at Folkungagatan station, averaged over five independent runs. For each forecast horizon, the best-performing model is highlighted in bold, and the second-best is underlined. The ‘AVG’ reports the mean performance across all horizons, while “1_{st} Count” indicates the number of first-place ranks a model achieved across all forecast horizons.

Models	Metric	Ours		Crossformer		TimeXer		iTransformer		Transformer		LSTM	
		R ²	MAPE	R ²	MAPE	R ²	MAPE	R ²	MAPE	R ²	MAPE	R ²	MAPE
NO _x	24	0.43	0.49	<u>0.44</u>	0.50	0.19	0.59	0.15	0.54	0.46	0.49	<u>0.44</u>	0.48
	48	0.39	0.58	0.23	0.72	0.12	0.62	0.08	0.57	0.24	0.73	<u>0.39</u>	0.55
	72	0.37	<u>0.55</u>	0.34	0.60	0.16	0.65	0.14	0.60	<u>0.35</u>	0.57	0.35	0.54
	96	<u>0.36</u>	0.58	0.37	0.58	0.21	0.66	0.10	0.62	0.28	0.68	0.28	0.59
	120	0.33	<u>0.60</u>	<u>0.31</u>	0.59	0.24	0.65	0.05	0.65	0.22	0.76	0.28	0.59
	144	0.31	0.64	<u>0.31</u>	0.60	0.23	0.66	0.01	0.66	0.17	0.82	0.28	<u>0.62</u>
	168	0.31	<u>0.64</u>	<u>0.30</u>	0.63	0.23	0.66	-0.01	0.70	0.24	0.66	0.27	0.65
	336	<u>0.24</u>	<u>0.65</u>	0.25	0.63	0.17	0.66	-0.03	0.73	0.19	0.66	0.14	0.72
	720	0.19	<u>0.66</u>	<u>0.19</u>	0.63	0.01	0.75	-0.28	0.92	0.04	0.75	0.02	0.76
AVG		0.33	0.60	<u>0.30</u>	<u>0.61</u>	0.17	0.66	0.02	0.67	0.24	0.68	0.27	<u>0.61</u>
PM ₁₀	1 st Count	6	1	2	5	0	0	0	0	1	0	0	4
	24	0.34	0.49	<u>0.30</u>	<u>0.53</u>	0.27	0.58	0.26	0.58	0.22	0.61	0.29	0.55
	48	0.24	0.56	<u>0.21</u>	<u>0.56</u>	0.17	0.66	0.17	0.62	0.18	0.63	0.18	0.60
	72	0.20	0.53	-0.15	<u>0.55</u>	0.13	0.65	0.13	0.63	<u>0.18</u>	0.64	<u>0.18</u>	0.61
	96	0.19	0.53	0.05	<u>0.55</u>	0.10	0.66	0.08	0.66	0.09	0.66	<u>0.11</u>	0.61
	120	0.13	0.59	0.02	<u>0.60</u>	0.05	0.67	0.08	0.69	0.08	0.71	<u>0.11</u>	0.61
	144	0.10	0.57	-0.23	<u>0.58</u>	0.05	0.69	0.03	0.68	<u>0.07</u>	0.71	<u>0.07</u>	0.66
	168	<u>0.10</u>	0.59	-0.05	0.63	0.03	0.69	0.02	0.68	0.12	0.68	0.07	<u>0.61</u>
	336	0.03	0.70	-0.16	0.71	-0.11	<u>0.70</u>	-0.05	0.72	<u>0.02</u>	0.76	-0.05	0.75
	720	-0.08	0.83	-0.65	<u>0.76</u>	-0.25	0.72	-0.23	0.85	<u>-0.13</u>	0.86	-0.24	0.92
AVG		0.14	0.60	-0.07	<u>0.61</u>	0.05	0.67	0.05	0.68	<u>0.09</u>	0.70	0.08	0.66
1 st Count		8	8	0	0	0	1	0	0	1	0	0	0

Table C.7

Multi-horizon forecasting performance for NO_x and PM₁₀ at Hornsgatan station, averaged over five independent runs. The best result for each horizon is shown in **bold**, and the second-best is underlined. “AVG” denotes the average performance across all horizons. “1_{st} Count” indicates how many times each model achieved the best result across all horizons for the corresponding pollutant. The units for MAE and MSE are $\mu\text{g}/\text{m}^3$ and $(\mu\text{g}/\text{m}^3)^2$, respectively.

Models	Metric	Ours		Crossformer		TimeXer		iTransformer		Transformer		LSTM	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
NO _x	24	<u>14.88</u>	428.18	14.92	430.60	16.59	548.23	16.19	518.15	15.37	455.26	14.77	416.49
	48	<u>15.44</u>	<u>448.97</u>	16.08	480.70	17.82	610.52	17.04	550.33	15.77	470.75	15.18	433.27
	72	<u>15.56</u>	<u>450.03</u>	17.85	575.55	18.01	605.93	17.57	579.26	16.62	516.50	15.43	447.26
	96	15.96	471.34	17.90	567.15	18.18	605.49	17.31	566.12	17.68	576.54	<u>16.26</u>	<u>496.60</u>
	120	16.67	508.53	19.03	627.13	18.12	598.58	17.79	593.57	18.42	602.16	<u>17.39</u>	<u>559.65</u>
	144	17.09	527.67	18.09	<u>581.88</u>	18.33	604.89	18.22	617.39	18.52	605.16	<u>17.92</u>	583.76
	168	17.52	550.72	18.20	<u>582.46</u>	18.35	608.02	18.43	626.59	18.70	613.99	<u>17.89</u>	584.77
	336	17.84	575.94	19.60	653.42	<u>18.70</u>	645.06	19.27	671.66	19.23	<u>641.01</u>	19.02	656.38
	720	18.90	613.48	<u>19.32</u>	<u>627.58</u>	19.86	723.21	19.96	710.41	20.47	697.34	21.59	836.62
AVG		16.65	508.32	17.89	569.61	18.22	616.66	17.98	603.72	17.86	575.41	<u>17.27</u>	<u>557.20</u>
PM ₁₀	1 st Count	6	6	0	0	0	0	0	0	0	0	3	3
	24	4.32	74.98	4.98	83.46	4.87	86.60	4.97	91.39	5.37	88.69	<u>4.66</u>	<u>80.99</u>
	48	4.93	89.18	5.24	<u>90.15</u>	5.38	97.93	5.81	112.97	5.30	91.98	<u>5.00</u>	91.59
	72	5.02	91.99	5.90	<u>102.86</u>	5.62	104.28	6.14	125.57	<u>5.22</u>	<u>94.04</u>	5.27	96.37
	96	5.03	91.83	<u>5.48</u>	<u>96.97</u>	5.61	103.77	5.78	107.99	5.63	98.07	5.51	103.82
	120	5.29	98.47	5.90	<u>103.03</u>	5.66	106.43	6.00	111.69	6.15	104.52	5.70	104.53
	144	5.36	101.60	6.55	119.05	5.85	110.85	6.16	117.38	5.73	<u>100.10</u>	<u>5.66</u>	104.74
	168	5.47	<u>101.33</u>	6.22	110.28	5.79	111.29	6.05	116.08	5.78	98.58	<u>5.70</u>	101.50
	336	5.47	87.00	5.81	96.86	5.53	93.01	6.47	105.80	5.74	<u>88.84</u>	<u>5.70</u>	89.82
	720	<u>5.67</u>	<u>70.41</u>	7.03	108.37	5.74	78.47	6.35	87.47	5.54	69.08	6.76	85.11
AVG		5.18	89.64	5.90	101.23	5.56	99.18	5.97	108.48	5.60	92.66	<u>5.55</u>	<u>95.39</u>
1 st Count		8	7	0	0	0	0	0	0	1	2	0	0

Table C.8

Multi-horizon forecasting performance for NO_x and PM₁₀ at Hornsgatan station, averaged over five independent runs. For each forecast horizon, the best-performing model is highlighted in bold, and the second-best is underlined. The ‘AVG’ reports the mean performance across all horizons, while “1_{st} Count” indicates the number of first-place ranks a model achieved across all forecast horizons.

Models	Metric	Ours		Crossformer		TimeXer		iTransformer		Transformer		LSTM	
		R ²	MAPE	R ²	MAPE	R ²	MAPE	R ²	MAPE	R ²	MAPE	R ²	MAPE
NO _x	24	<u>0.52</u>	1.91	0.51	1.93	0.38	<u>1.79</u>	0.42	1.75	0.49	2.24	0.53	1.98
	48	<u>0.49</u>	2.07	0.45	2.19	0.30	<u>1.91</u>	0.37	<u>1.91</u>	0.46	2.02	0.51	1.86
	72	<u>0.48</u>	2.21	0.34	2.45	0.30	1.95	0.33	2.11	0.41	2.39	0.49	<u>2.05</u>
	96	0.45	2.29	0.34	2.35	0.30	1.89	0.34	2.18	0.33	2.50	<u>0.42</u>	<u>2.01</u>
	120	0.41	2.29	0.27	2.42	0.30	1.95	0.31	2.32	0.30	2.46	<u>0.35</u>	<u>2.30</u>
	144	0.38	<u>2.26</u>	<u>0.32</u>	2.33	0.29	2.01	0.28	2.45	0.29	2.30	0.32	2.37
	168	0.35	2.27	<u>0.31</u>	<u>2.19</u>	0.28	2.13	0.26	2.62	0.28	2.61	<u>0.31</u>	2.25
	336	0.31	2.60	<u>0.22</u>	<u>2.62</u>	0.23	2.63	0.20	2.97	<u>0.23</u>	3.32	0.22	2.71
	720	0.26	1.88	<u>0.24</u>	1.78	0.12	<u>1.90</u>	0.14	1.97	0.15	2.12	-0.02	2.05
AVG		0.41	2.20	0.33	2.25	0.28	2.02	0.29	2.25	0.33	2.44	<u>0.35</u>	<u>2.17</u>
PM ₁₀	1 st Count	6	1	0	1	0	5	0	1	0	0	3	1
	24	0.38	0.37	0.31	0.45	0.28	0.44	0.24	0.44	0.26	0.46	<u>0.33</u>	<u>0.39</u>
	48	0.25	<u>0.44</u>	<u>0.24</u>	0.48	0.17	0.48	0.05	0.53	0.22	0.46	0.23	0.43
	72	0.20	0.46	0.11	0.55	0.10	0.51	-0.09	0.57	<u>0.19</u>	0.45	0.17	<u>0.45</u>
	96	0.19	0.47	<u>0.15</u>	0.52	0.08	0.51	0.05	0.54	0.14	0.52	0.08	<u>0.50</u>
	120	0.12	0.49	<u>0.08</u>	0.56	0.05	<u>0.54</u>	0.00	0.58	0.07	0.61	0.07	<u>0.54</u>
	144	<u>0.09</u>	0.48	-0.07	0.63	0.00	0.57	-0.06	0.60	0.10	0.56	0.06	<u>0.55</u>
	168	0.08	0.54	0.00	0.61	-0.01	<u>0.56</u>	-0.05	0.59	0.11	0.58	<u>0.08</u>	<u>0.57</u>
	336	0.04	0.60	-0.07	<u>0.59</u>	-0.03	0.55	-0.17	0.67	<u>0.01</u>	0.61	0.00	0.62
AVG		0.14	0.50	0.00	0.57	0.05	<u>0.53</u>	-0.04	0.58	<u>0.11</u>	0.54	0.08	0.54
1 st Count		6	5	0	0	0	2	0	0	3	1	0	1

Table C.9

Multi-horizon forecasting performance for NO_x and PM₁₀ at Sveavägen station, averaged over five independent runs. The best result for each horizon is shown in **bold**, and the second-best is underlined. “AVG” denotes the average performance across all horizons. “1_{st} Count” indicates how many times each model achieved the best result across all horizons for the corresponding pollutant. The units for MAE and MSE are $\mu\text{g}/\text{m}^3$ and $(\mu\text{g}/\text{m}^3)^2$, respectively.

Models	Metric	Ours		Crossformer		TimeXer		iTransformer		Transformer		LSTM	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
NO _x	24	7.79	135.39	8.22	143.91	10.12	218.54	9.59	197.97	8.24	150.22	8.41	150.51
	48	8.08	140.62	8.89	<u>158.59</u>	10.87	243.92	9.85	208.45	8.91	163.62	<u>8.72</u>	160.63
	72	8.32	148.89	9.36	174.21	11.04	249.06	10.22	220.75	9.61	185.48	<u>9.03</u>	<u>170.40</u>
	96	8.95	168.46	<u>9.25</u>	<u>171.31</u>	10.84	234.11	10.11	210.61	10.67	222.14	9.50	190.62
	120	9.41	183.82	10.22	<u>201.94</u>	10.75	228.83	<u>10.16</u>	209.98	11.25	236.01	10.27	212.90
	144	10.03	197.58	<u>10.17</u>	<u>200.62</u>	10.74	228.13	10.28	214.70	10.97	232.13	10.24	213.77
	168	10.07	202.68	10.86	<u>219.56</u>	10.71	227.67	<u>10.48</u>	219.92	11.01	231.54	10.66	223.03
	336	10.53	217.38	11.36	242.33	<u>10.65</u>	<u>228.05</u>	10.69	230.85	11.53	249.12	11.07	233.32
	720	10.78	230.72	<u>10.86</u>	<u>228.50</u>	11.24	259.80	10.92	238.93	12.30	273.52	11.92	262.63
AVG		9.33	180.62	<u>9.91</u>	<u>193.44</u>	10.77	235.34	10.26	216.91	10.50	215.98	9.98	201.98
PM ₁₀	1 st Count	9	9	0	0	0	0	0	0	0	0	0	0
	24	<u>3.91</u>	27.66	4.02	29.15	4.37	34.40	4.31	33.29	4.05	30.28	3.86	<u>28.48</u>
	48	<u>4.11</u>	<u>30.80</u>	4.25	32.31	4.75	40.18	4.71	39.19	4.51	36.60	4.06	30.42
	72	4.47	34.09	4.78	40.75	4.95	43.31	4.85	41.42	<u>4.16</u>	<u>31.78</u>	4.04	29.78
	96	<u>4.36</u>	<u>33.53</u>	4.52	36.77	4.94	42.41	4.87	41.52	<u>4.51</u>	35.70	4.30	33.17
	120	4.61	36.54	4.86	40.40	4.90	41.66	4.87	41.13	4.70	<u>37.42</u>	<u>4.69</u>	37.92
	144	<u>4.73</u>	<u>38.22</u>	4.97	42.02	4.95	42.26	4.95	41.93	4.68	37.37	4.86	39.45
	168	<u>4.71</u>	38.26	4.98	41.88	4.89	41.11	4.97	42.18	4.67	37.03	4.76	37.58
	336	<u>4.86</u>	39.32	5.11	43.63	4.74	39.59	4.88	40.77	4.70	36.63	4.89	<u>38.70</u>
AVG		4.58	36.02	4.76	39.53	4.80	40.67	4.82	40.45	4.55	35.81	<u>4.56</u>	<u>35.90</u>
1 st Count		1	2	0	0	1	0	0	0	3	4	4	3

Table C.10

Multi-horizon forecasting performance for NO_x and PM₁₀ at Sveavägen station, averaged over five independent runs. For each forecast horizon, the best-performing model is highlighted in bold, and the second-best is underlined. The ‘AVG’ reports the mean performance across all horizons, while “1_{st} Count” indicates the number of first-place ranks a model achieved across all forecast horizons.

Models	Metric	Ours		Crossformer		TimeXer		iTransformer		Transformer		LSTM	
		R ²	MAPE	R ²	MAPE	R ²	MAPE	R ²	MAPE	R ²	MAPE	R ²	MAPE
NO _x	24	0.61	0.42	<u>0.58</u>	<u>0.44</u>	0.37	0.55	0.43	0.51	0.56	<u>0.44</u>	0.56	0.48
	48	0.59	0.45	<u>0.53</u>	0.50	0.28	0.60	0.39	0.54	0.52	0.51	<u>0.53</u>	<u>0.49</u>
	72	0.56	0.47	<u>0.48</u>	0.53	0.26	0.62	0.35	0.57	0.45	0.56	<u>0.50</u>	<u>0.52</u>
	96	0.50	0.52	<u>0.49</u>	<u>0.52</u>	0.31	0.64	0.38	0.59	0.34	0.68	0.43	0.55
	120	0.45	0.56	<u>0.40</u>	0.61	0.32	0.64	0.37	<u>0.60</u>	0.30	0.75	0.37	0.62
	144	0.41	0.63	<u>0.40</u>	0.60	0.32	0.64	0.36	<u>0.61</u>	0.30	0.71	0.36	0.62
	168	0.39	0.61	<u>0.34</u>	0.70	0.31	0.64	<u>0.34</u>	<u>0.63</u>	0.30	0.69	0.33	0.66
	336	0.32	0.67	<u>0.25</u>	0.76	<u>0.29</u>	<u>0.63</u>	<u>0.28</u>	0.62	0.22	0.78	0.27	0.73
	720	<u>0.26</u>	0.68	0.27	0.69	0.17	0.61	0.24	<u>0.65</u>	0.13	0.87	0.16	0.80
	AVG	0.45	0.56	<u>0.42</u>	0.60	0.29	0.62	0.35	<u>0.59</u>	0.35	0.67	0.39	0.61
	1 st Count	8	6	1	1	0	1	0	1	0	0	0	0
PM ₁₀	24	0.36	<u>0.37</u>	0.33	<u>0.37</u>	0.21	0.42	0.23	0.42	0.30	<u>0.37</u>	0.34	0.35
	48	<u>0.28</u>	0.38	0.25	0.41	0.07	0.45	0.09	0.45	0.15	0.39	0.29	<u>0.38</u>
	72	0.20	0.45	0.05	0.45	-0.01	0.47	0.03	0.47	<u>0.26</u>	<u>0.39</u>	0.30	0.38
	96	<u>0.21</u>	0.43	0.13	<u>0.42</u>	-0.00	0.48	0.02	0.47	0.16	0.45	0.22	0.42
	120	<u>0.13</u>	0.46	0.04	0.50	0.01	0.48	0.02	0.48	<u>0.11</u>	<u>0.47</u>	0.10	<u>0.47</u>
	144	0.08	<u>0.48</u>	-0.01	0.49	-0.01	0.49	-0.00	0.50	0.10	0.46	0.05	0.49
	168	0.08	0.46	-0.01	0.50	0.01	0.49	-0.02	0.50	0.11	<u>0.46</u>	<u>0.10</u>	0.49
	336	0.01	0.50	-0.10	0.52	-0.00	0.46	-0.03	<u>0.49</u>	0.07	0.50	<u>0.02</u>	0.54
	720	-0.23	0.61	-0.31	0.53	<u>-0.10</u>	0.43	-0.14	<u>0.51</u>	-0.06	0.55	-0.28	0.64
	AVG	0.13	<u>0.46</u>	<u>0.04</u>	0.47	0.02	<u>0.46</u>	0.02	0.48	0.13	0.45	0.13	<u>0.46</u>
	1 st Count	2	3	0	0	0	2	0	0	4	1	3	3

C.3. Statistical significance tests

Table C.11

Paired *t*-test results (p-values) between our model and baselines for NO_x and PM₁₀, averaged across all monitoring stations over five independent runs. “AVG” denotes the average result across all horizons. Entries marked with “***” indicate $p < .01$, and “**” indicates $p < .05$.

Models	Metric	Crossformer				TimeXer				iTransformer				Transformer				LSTM			
		MAE	MSE	R ²	MAPE	MAE	MSE	R ²	MAPE	MAE	MSE	R ²	MAPE	MAE	MSE	R ²	MAPE	MAE	MSE	R ²	MAPE
NO _x	24	0.06	0.18	0.10	0.37	**	**	**	*	**	**	**	0.25	**	**	**	*	0.40	0.70	0.20	*
	48	*	0.05	0.05	*	**	**	**	0.10	**	**	**	0.24	**	**	**	*	0.89	0.85	*	**
	72	*	*	*	0.07	**	**	**	0.60	**	**	**	0.48	**	**	**	*	0.06	0.10	*	0.14
	96	**	**	**	0.44	**	**	**	*	**	**	**	0.90	**	**	**	*	**	**	**	0.02
	120	**	*	*	*	**	**	**	*	**	**	**	*	**	**	**	*	**	**	**	0.24
	144	0.06	*	0.05	0.94	**	**	**	*	**	**	**	*	**	**	**	*	*	**	*	0.15
	168	*	*	0.08	0.71	**	**	**	*	**	**	**	*	**	**	**	*	*	*	*	0.52
	336	**	**	**	0.42	**	**	**	0.58	**	**	**	*	**	**	**	*	**	**	**	*
	720	0.32	0.58	0.77	0.02	**	**	**	0.45	**	**	**	*	**	**	**	*	**	**	**	*
	AVG	**	**	**	*	**	**	**	*	**	**	**	*	**	**	**	*	**	**	**	0.20
PM ₁₀	24	**	**	*	*	**	**	**	*	**	**	**	*	**	**	**	*	*	*	*	*
	48	**	0.29	0.22	0.24	**	**	**	*	**	**	**	*	**	**	**	0.09	0.45	0.29	0.28	0.55
	72	**	**	**	0.14	**	**	**	*	**	**	**	*	0.98	0.95	0.67	0.67	0.53	0.62	0.57	0.74
	96	**	**	*	0.34	**	**	**	*	**	**	**	*	**	**	**	*	*	0.05	*	*
	120	*	0.06	*	*	**	**	**	*	**	**	**	*	**	0.15	0.08	**	**	0.07	0.06	0.06
	144	**	**	**	*	**	**	**	*	**	**	**	*	*	0.97	0.78	**	*	0.19	0.12	*
	168	**	*	*	*	**	**	**	*	**	**	**	*	0.11	0.07	0.02	*	0.09	0.74	0.58	*
	336	*	**	**	0.82	0.36	**	**	0.20	**	**	**	0.29	0.58	0.93	0.56	0.45	0.21	0.22	0.14	0.26
	720	**	**	**	0.29	0.13	*	*	*	0.09	**	*	0.10	0.06	0.13	0.13	0.00	**	**	**	*
	AVG	**	**	**	*	**	**	**	*	**	**	**	*	**	**	**	*	**	**	**	*

Table C.12

Effect size (Cohen's d) results between our model and baselines for NO_x and PM_{10} , averaged across all monitoring stations over five independent runs. "AVG" denotes the overall effect size across all horizons.

Models	Metric	Crossformer				TimeXer				iTransformer				Transformer				LSTM			
		MAE	MSE	R^2	MAPE	MAE	MSE	R^2	MAPE	MAE	MSE	R^2	MAPE	MAE	MSE	R^2	MAPE	MAE	MSE	R^2	MAPE
NO_x	24	1.15	0.73	0.96	0.45	24.25	35.55	49.34	4.21	14.11	23.38	24.60	-0.60	3.84	2.92	2.26	2.12	0.42	-0.19	0.69	1.56
	48	1.34	1.23	1.23	1.25	31.70	29.24	54.03	0.94	27.89	24.27	26.70	-0.62	9.41	12.07	11.71	1.66	-0.07	0.09	1.41	-3.01
	72	1.58	1.37	1.64	1.11	68.89	38.94	35.39	-0.25	29.31	25.02	25.90	0.35	4.78	3.44	4.97	1.47	1.15	0.95	1.46	-0.84
	96	2.26	2.53	2.15	0.39	36.78	70.24	48.00	-4.39	16.49	18.22	20.78	-0.06	7.55	7.48	10.28	3.00	2.29	2.27	4.19	-1.69
	120	2.11	1.54	1.58	1.41	10.58	16.89	13.59	-2.37	10.65	14.76	14.41	1.30	9.67	10.12	9.52	3.44	3.44	6.46	5.84	0.62
	144	1.19	1.60	1.23	0.04	5.34	7.31	6.44	-2.22	6.85	9.39	9.80	2.52	5.18	5.95	5.84	1.63	1.30	2.15	1.66	0.80
	168	1.50	1.31	1.07	-0.18	6.02	16.31	14.22	-3.05	9.73	12.36	19.79	8.61	3.94	3.86	4.07	3.45	1.59	1.43	1.92	-0.31
	336	2.71	2.58	2.25	0.41	4.86	8.16	9.06	0.27	11.51	11.17	13.29	4.57	3.89	3.79	4.23	5.55	3.40	2.38	2.98	1.94
	720	0.51	0.27	0.14	-1.66	2.55	4.74	6.65	0.37	3.99	5.22	8.41	2.42	3.17	3.25	3.75	3.25	4.31	6.43	6.08	3.07
AVG		1.16	1.02	1.01	0.41	2.05	2.32	2.19	0.43	4.40	6.62	6.54	0.79	2.29	2.34	2.32	1.83	0.98	0.88	1.07	0.19
PM_{10}	24	3.03	2.75	1.41	3.02	4.65	4.25	4.36	12.65	5.59	5.81	5.60	9.26	3.58	2.55	2.16	4.80	1.63	1.97	1.72	1.25
	48	2.19	0.55	0.65	0.62	5.37	6.16	10.03	3.70	6.16	10.37	12.21	3.51	3.20	2.44	2.67	1.01	0.37	0.54	0.56	0.29
	72	7.31	5.66	5.85	0.83	3.04	3.63	3.11	2.34	4.48	5.12	3.81	2.50	0.01	-0.03	-0.21	0.20	-0.31	0.24	-0.28	-0.16
	96	2.26	2.27	1.41	0.49	10.54	8.03	8.06	6.23	21.94	12.54	10.48	4.86	3.52	2.20	2.81	4.90	1.37	1.22	1.55	1.29
	120	1.94	1.13	1.52	3.44	5.81	4.02	5.68	2.52	9.98	5.92	8.17	7.33	2.47	0.81	1.02	10.15	2.20	1.11	1.15	1.15
	144	2.59	2.08	2.64	2.49	3.35	3.83	2.66	6.55	4.72	5.49	3.34	6.96	1.38	0.02	0.13	3.84	1.75	0.70	0.88	1.58
	168	2.35	1.87	1.53	1.41	10.96	7.72	6.96	4.80	7.14	7.41	9.03	3.99	0.92	-1.11	-1.58	1.95	0.98	0.16	0.27	1.80
	336	1.29	2.38	3.24	-0.11	0.46	2.96	2.75	-0.68	2.44	5.71	4.62	0.55	0.27	-0.05	-0.29	0.38	0.66	0.65	0.83	0.59
	720	3.36	4.95	6.25	-0.55	-0.85	1.53	1.26	-4.96	1.01	2.33	1.84	-0.95	-1.14	-0.86	-0.84	-2.61	2.11	2.14	2.32	2.68
AVG		1.74	1.26	1.16	0.42	1.38	2.60	2.60	0.46	2.80	2.95	3.20	1.38	0.79	0.41	0.34	0.95	0.82	0.83	0.73	0.85

C.4. Accuracy in peak concentrations

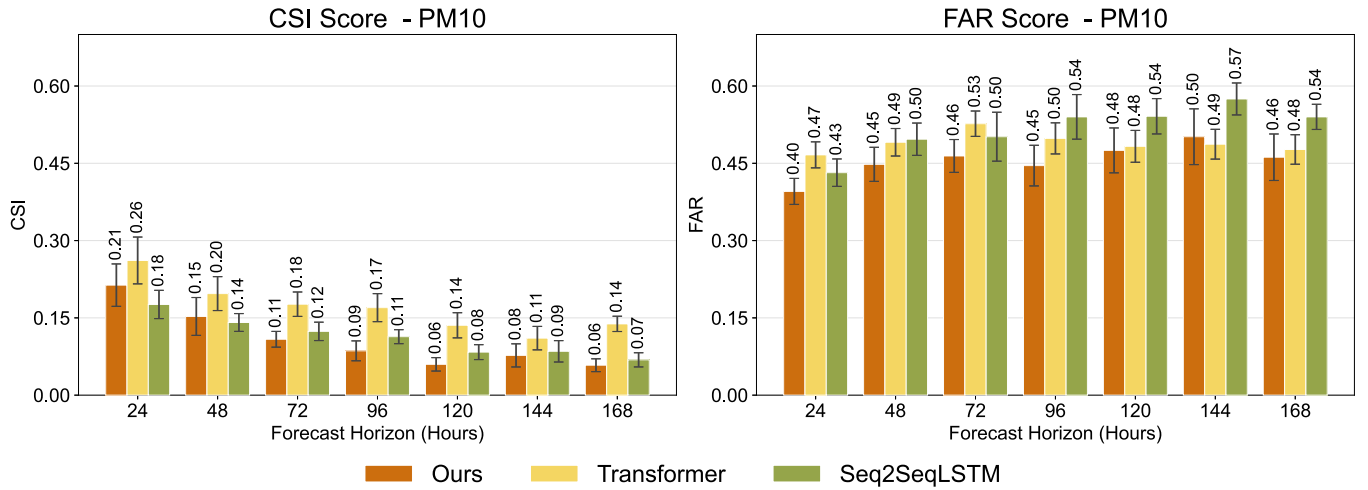


Fig. C.16. Detection accuracy for peak PM_{10} concentrations measured by CSI and FAR scores. Results show the average across all stations and five independent runs. Corresponding mean values are labeled on each bar.

Appendix D. Generalization Study at Suburban Site

D.1. Site description

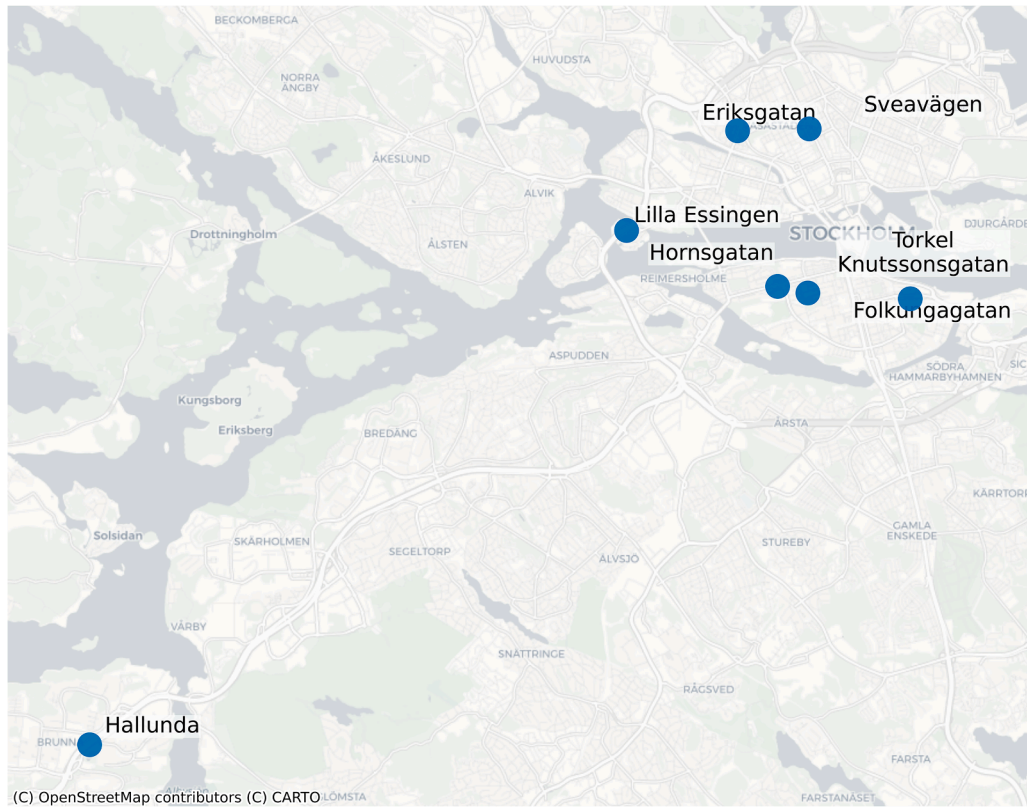


Fig. D.17. Geographical distribution of the monitoring stations. The suburban Hallunda station is located adjacent to a major highway, geographically isolated from inner-city street canyon stations.

We evaluate the distributional and temporal discrepancies between the suburban Hallunda highway site and three urban street canyon stations. The Kolmogorov-Smirnov (KS) tests confirm statistically significant distributional shift between the open highway environment and the canyon sites for both pollutants ($p < .001$). Furthermore, the diurnal profiles reveal divergent aerodynamic behaviors: the Hallunda site exhibits superior ventilation and dispersion, characterized by a rapid attenuation of NO_x concentrations between AM and PM traffic peaks. In contrast, the street canyon stations show a distinct “trapping effect” with prolonged pollutant accumulation. Similarly, PM_{10} levels at the highway site demonstrate higher volatility during daylight hours compared to the more stable patterns within the urban canyons.

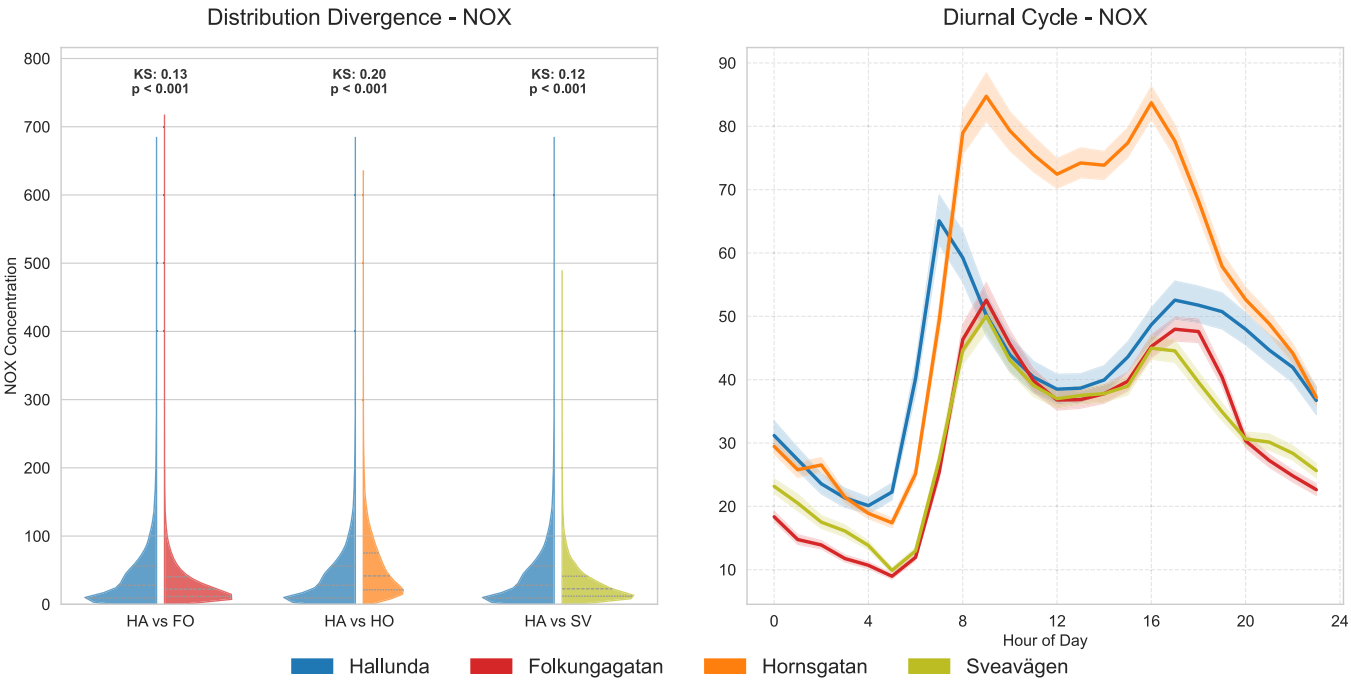


Fig. D.18. Distributional divergence and diurnal variations of NO_x comparing the Hallunda highway site with three street canyon stations.

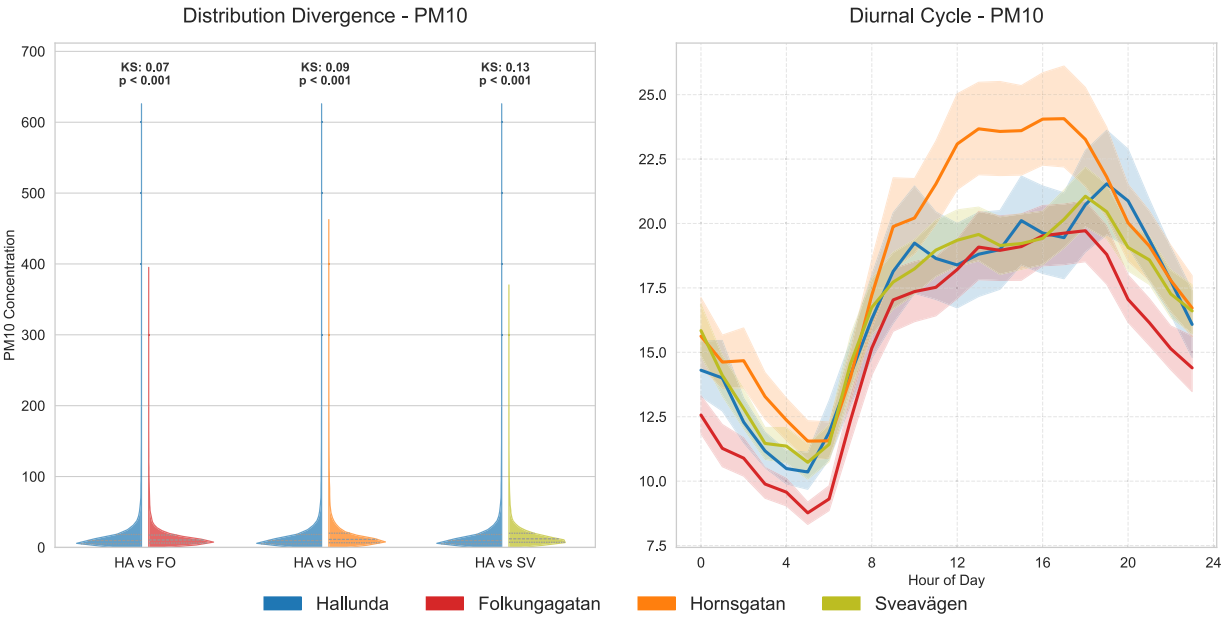


Fig. D.19. Distributional divergence and diurnal variations of PM₁₀ comparing the Hallunda highway site with three street canyon stations.

D.2. Model performance and significance tests

Table D.13

Multi-horizon forecasting performance for NO_x and PM_{10} at Hallunda, averaged over five independent runs. For each forecast horizon, the best-performing model is highlighted in bold, and the second-best is underlined. The ‘AVG’ reports the mean performance across all horizons, while ‘1st Count’ indicates the number of first-place ranks a model achieved across all forecast horizons.

Models	Metric	Ours				Crossformer				Transformer				LSTM			
		MAE	MSE	R^2	MAPE	MAE	MSE	R^2	MAPE	MAE	MSE	R^2	MAPE	MAE	MSE	R^2	MAPE
NO_x	24	12.84	<u>349.47</u>	<u>0.43</u>	0.74	14.16	386.35	0.36	0.87	16.95	522.99	0.14	1.06	<u>13.24</u>	343.43	0.44	<u>0.78</u>
	48	13.68*	370.48	0.39	0.82*	15.28	430.93	0.29	0.97	16.60	489.00	0.19	1.08	<u>14.48</u>	<u>383.64</u>	<u>0.37</u>	<u>0.87</u>
	72	13.70	373.29	0.38	0.82	15.33	436.06	0.28	0.98	16.46	485.48	0.19	1.07	<u>14.43</u>	<u>377.31</u>	<u>0.37</u>	<u>0.88</u>
	96	15.18	435.42	0.27	<u>0.96</u>	15.67	442.82	0.26	1.03	17.26	506.86	0.15	1.19	<u>15.46</u>	<u>439.47</u>	0.27	0.94
	120	15.38*	445.39*	0.26*	1.02	16.52	484.92	0.19	1.12	17.95	541.80	0.09	1.29	<u>16.11</u>	<u>468.11</u>	0.22	<u>1.04</u>
	144	16.01	486.48	0.18	1.06	16.85	509.08	0.14	1.14	17.55	531.10	0.11	1.22	<u>16.79</u>	<u>499.59</u>	<u>0.16</u>	<u>1.10</u>
	168	16.40	<u>502.14</u>	<u>0.15</u>	1.08	17.64	537.04	0.09	1.25	18.06	552.65	0.07	1.29	<u>16.57</u>	482.23	0.19	<u>1.09</u>
	336	17.25	<u>552.53</u>	<u>0.06</u>	1.14	19.13	615.98	-0.04	1.38	<u>18.04</u>	545.54	0.08	1.26	18.27	563.61	0.05	<u>1.24</u>
	720	17.46*	530.19*	0.10*	1.20	19.95	638.03	-0.08	<u>1.51</u>	20.06	645.17	-0.09	<u>1.51</u>	<u>19.58</u>	<u>631.68</u>	<u>-0.07</u>	1.39
AVG		15.32**	449.49*	0.25*	0.98**	16.73	497.91	0.17	1.14	17.66	535.62	0.10	1.22	<u>16.10</u>	<u>465.45</u>	0.22	<u>1.04</u>
1 st Count		9	6	6	8	0	0	0	0	0	1	1	0	0	2	3	1
PM_{10}	24	3.84**	31.96*	0.57*	0.44	4.24	36.67	0.51	<u>0.47</u>	4.26	35.32	0.53	0.50	<u>3.93</u>	<u>32.31</u>	0.57	0.48
	48	<u>4.38*</u>	39.67*	0.45*	0.53	4.52	40.10	0.44	<u>0.51</u>	4.87	44.61	0.38	0.59	4.26	<u>39.77</u>	<u>0.44</u>	0.49
	72	4.20**	37.07	0.45	0.47**	5.83	61.77	0.09	0.74	<u>4.62</u>	<u>41.54</u>	<u>0.39</u>	0.55	<u>4.62</u>	<u>42.88</u>	<u>0.37</u>	<u>0.59</u>
	96	4.49**	42.75	0.33	0.53**	5.57	59.91	0.06	0.66	5.05	46.34	0.27	0.66	<u>4.86</u>	<u>46.05</u>	<u>0.28</u>	<u>0.63</u>
	120	4.56**	41.83*	0.31*	0.52**	5.34	51.73	0.14	0.64	5.44	48.43	<u>0.20</u>	0.76	5.36	50.45	0.16	<u>0.75</u>
	144	4.74**	43.46	0.25	0.56**	5.86	57.22	0.01	<u>0.78</u>	5.50	48.21	<u>0.17</u>	0.79	5.57	51.41	0.11	<u>0.79</u>
	168	4.78**	42.28**	0.24**	0.60**	5.29	48.98	0.12	<u>0.69</u>	5.52	<u>47.61</u>	<u>0.15</u>	0.80	5.37	50.85	0.09	0.74
	336	4.89*	41.49*	0.11*	0.67**	5.92	56.95	-0.22	<u>0.86</u>	<u>5.75</u>	<u>49.40</u>	<u>-0.06</u>	0.89	6.81	64.81	-0.39	1.10
	720	<u>6.29</u>	<u>53.84</u>	<u>-0.40</u>	1.06	7.15	87.77	-1.28	<u>1.04</u>	6.11	52.80	-0.37	1.03	7.92	87.92	-1.28	1.33
AVG		4.69**	41.59**	0.26**	0.60**	5.53	55.68	-0.01	<u>0.71</u>	<u>5.24</u>	<u>46.03</u>	<u>0.18</u>	0.73	5.41	51.83	0.04	0.77
1 st Count		7	8	8	7	0	0	0	0	2	5	1	2	1	0	1	1

*/**: Ours outperforms the second-best model at $p < .05/0.01$ using two-sided paired t -test. Second-best models: LSTM for NO_x and Transformer for PM_{10} .

Table D.14

Paired t -test results (p-values) between our model and baselines for NO_x and PM_{10} at Hallunda, averaged over five independent runs. ‘AVG’ denotes the average result across all horizons. Entries marked with ‘***’ indicate $p < .01$, and ‘**’ indicates $p < .05$.

Models	Metric	Crossformer				Transformer				LSTM			
		MAE	MSE	R^2	MAPE	MAE	MSE	R^2	MAPE	MAE	MSE	R^2	MAPE
NO_x	24	*	0.11	0.11	*	**	**	**	**	0.21	0.61	0.61	0.13
	48	*	*	*	*	**	**	**	**	*	0.12	0.12	*
	72	*	*	*	*	**	**	**	**	0.06	0.57	0.57	0.23
	96	0.28	0.70	0.70	0.22	**	*	*	**	0.72	0.88	0.88	0.81
	120	**	*	*	**	**	**	**	**	*	0.09	0.09	0.40
	144	0.14	0.36	0.36	0.13	**	**	**	**	0.17	0.51	0.51	0.44
	168	0.13	0.20	0.20	0.10	**	**	**	*	0.73	0.12	0.12	0.96
	336	*	*	*	*	*	0.59	0.59	*	0.18	0.70	0.70	0.22
	720	*	**	**	*	**	**	**	**	*	*	*	0.05
AVG		**	**	**	**	**	**	**	**	**	*	*	**
PM_{10}	24	*	**	**	0.44	**	*	*	0.08	0.10	0.73	0.73	0.20
	48	0.53	0.76	0.76	0.82	*	*	*	0.12	0.60	0.97	0.97	0.55
	72	**	**	**	**	**	0.09	0.09	**	*	**	**	**
	96	*	*	*	*	**	0.42	0.42	**	*	0.05	0.05	0.09
	120	*	*	*	0.05	**	*	*	**	*	*	*	*
	144	0.08	0.10	0.10	0.12	**	0.06	0.06	**	*	*	*	**
	168	*	0.10	0.10	0.11	**	**	**	**	*	*	*	*
	336	*	*	*	*	*	*	*	**	**	**	**	**
	720	0.42	0.22	0.22	0.91	0.54	0.77	0.77	0.60	*	*	*	0.07
AVG		**	**	**	**	**	**	**	**	**	**	**	**

Table D.15

Effect size (Cohen's d) results between our model and baselines for NO_x and PM_{10} at Hallunda, averaged over five independent runs. "AVG" denotes the overall effect size across all horizons.

Models	Metric	Crossformer				Transformer				LSTM			
		MAE	MSE	R^2	MAPE	MAE	MSE	R^2	MAPE	MAE	MSE	R^2	MAPE
NO_x	24	1.33	0.91	0.91	1.62	26.76	21.55	21.55	12.55	0.66	-0.25	-0.25	0.85
	48	1.93	1.71	1.71	2.02	3.47	5.85	5.85	3.18	1.59	0.89	0.89	1.45
	72	1.84	1.53	1.53	1.55	5.26	5.27	5.27	4.50	1.20	0.28	0.28	0.63
	96	0.55	0.19	0.19	0.66	2.33	2.00	2.00	2.43	0.17	0.07	0.07	-0.11
	120	2.31	1.61	1.61	2.16	4.66	4.33	4.33	4.43	1.82	1.00	1.00	0.42
	144	0.82	0.46	0.46	0.85	4.44	2.53	2.53	3.79	0.75	0.32	0.32	0.39
	168	0.86	0.68	0.68	0.94	2.25	3.12	3.12	1.68	0.17	-0.89	-0.89	0.03
	336	2.01	1.30	1.30	1.89	1.31	-0.26	-0.26	1.54	0.73	0.19	0.19	0.66
	720	1.79	2.43	2.43	1.51	4.63	5.10	5.10	3.57	1.37	1.30	1.30	1.24
AVG		1.30	1.02	1.02	1.18	2.18	1.59	1.60	2.54	0.71	0.31	0.31	0.44
PM_{10}	24	1.59	2.08	2.08	0.39	2.71	1.42	1.42	1.07	0.95	0.17	0.17	0.69
	48	0.31	0.15	0.15	-0.11	1.69	1.35	1.35	0.87	-0.26	0.02	0.02	-0.29
	72	2.65	2.24	2.24	2.13	2.43	1.00	1.00	3.00	1.74	2.27	2.27	2.43
	96	1.41	1.25	1.25	1.42	2.20	0.40	0.40	2.35	1.85	1.22	1.22	1.00
	120	1.68	1.34	1.34	1.22	4.26	1.93	1.93	4.25	1.59	1.65	1.65	1.62
	144	1.06	0.95	0.95	0.88	2.37	1.15	1.15	3.29	2.93	1.99	1.99	3.53
	168	1.28	0.94	0.94	0.94	5.06	2.75	2.75	6.06	1.71	1.80	1.80	1.52
	336	1.86	1.82	1.82	1.57	2.02	1.86	1.86	2.14	7.19	3.40	3.40	28.88
	720	0.41	0.65	0.65	-0.05	-0.30	-0.14	-0.14	-0.26	1.27	1.49	1.49	1.10
AVG		0.91	0.69	0.69	-	1.30	0.87	0.87	-	0.91	0.77	0.77	-

Appendix E. Additional Model Interpretability

E.1. PM10 Attribution analysis

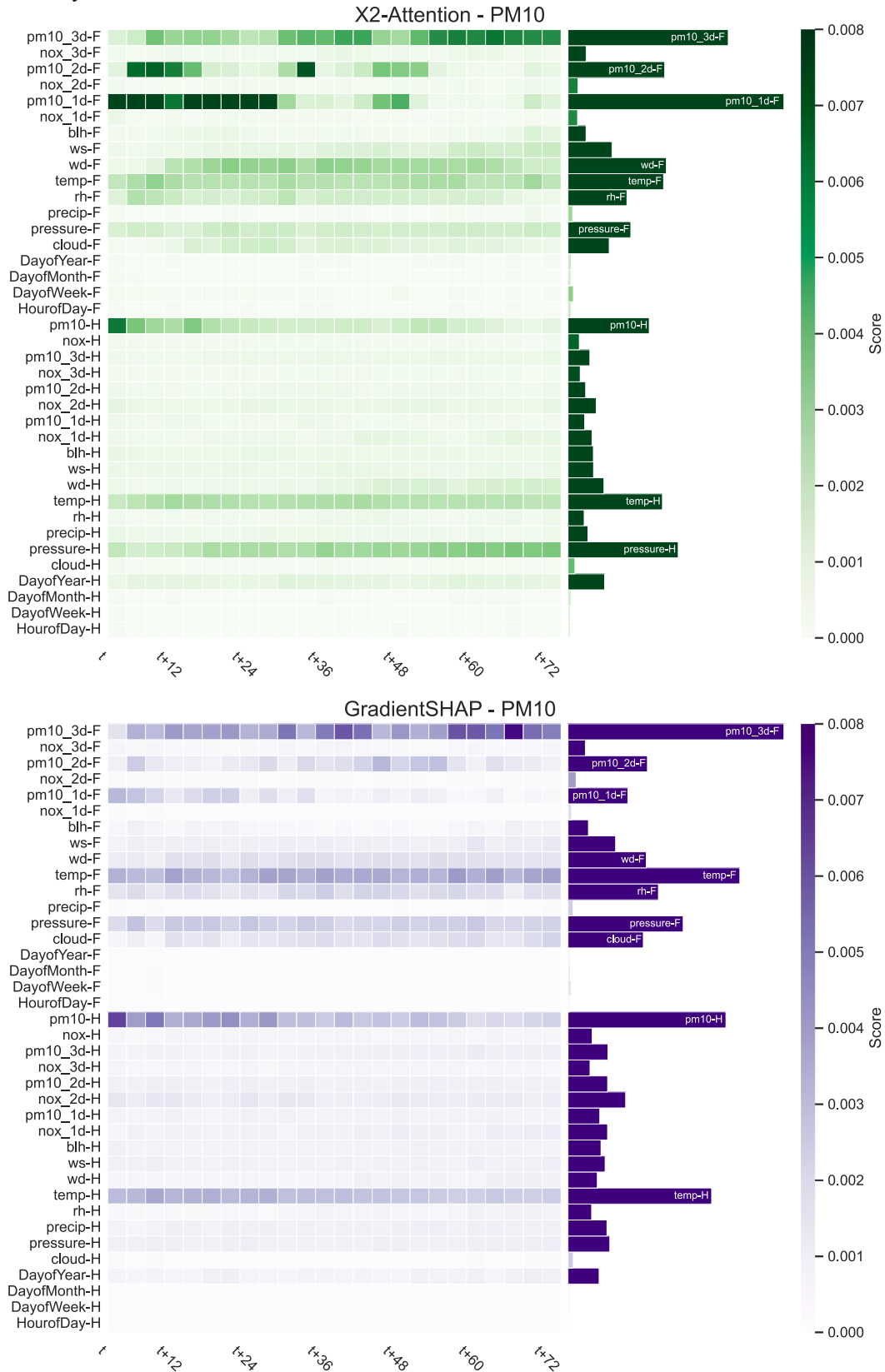


Fig. E.20. Attribution scores and feature rankings for PM₁₀ forecasting. The upper panel shows results obtained with X2-Attention, and the lower panel shows results using GradientSHAP. For both, the heatmap illustrates how feature importance evolves across lead times (1–72 hours), with darker colors indicating higher importance. The accompanying bar chart presents the overall feature ranking aggregated across all forecast steps, with labels indicating the top 10 features.

References

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*.
- Appel, K. W., Napelenok, S. L., Foley, K. M., Pye, H. O. T., Hogrefe, C., Lueken, D. J., Bash, J. O., Roselle, S. J., Pleim, J. E., Foroutan, H. et al. (2017). Description and evaluation of the community multiscale air quality (CMAQ) modeling system version 5.1. *Geoscientific Model Development*, 10(4), 1703–1732.
- Arndt, S., Turvey, C., & Andreasen, N. C. (1999). Correlating and predicting psychiatric symptom ratings: Spearman's r versus Kendall's tau correlation. *Journal of Psychiatric Research*, 33(2), 97–104.
- Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*.
- Bertrand, J.-M., Meleux, F., Ung, A., Descombes, G., & Colette, A. (2023). Improving the european air quality forecast of the copernicus atmosphere monitoring service using machine learning techniques. *Atmospheric Chemistry and Physics*, 23(9), 5317–5333.
- Bi, J., Knowland, K. E., Keller, C. A., & Liu, Y. (2022). Combining machine learning and numerical simulation for high-resolution PM2.5 concentration forecast. *Environmental Science & Technology*, 56(3), 1544–1556. <https://doi.org/10.1021/acs.est.1c05578>
- Brasseur, G. P., & Jacob, D. J. (2017). Modeling of atmospheric chemistry. Cambridge University Press.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- CAMS, C. A. M. S. (2017). Meteo-france: Regional production, description of the operational models and of the ENSEMBLE system. https://atmosphere.copernicus.eu/sites/default/files/2018-02/CAMS50_factsheet_201610_v2.pdf.
- Casalicchio, G., Molnar, C., & Bischl, B. (2019). Visualizing the feature importance for black box models. In *Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2018, dublin, ireland, september 10–14, 2018, proceedings, part i 18* (pp. 655–670). Springer.
- Chen, Y., Céspedes, N., & Barnaghi, P. A closer look at transformers for time series forecasting: Understanding why they work and where they struggle. In *Forty-second international conference on machine learning*.
- Denby, B. R., Sundvor, I., Johansson, C., Pirjola, L., Ketzler, M., Norman, M., Kupiainen, K., Gustafsson, M., Blomqvist, G., & Omstedt, G. (2013). A coupled road dust and surface moisture model to predict non-exhaust road traffic induced particle emissions (NOR-TRIP). part 1: Road dust loading and suspension modelling. *Atmospheric Environment*, 77, 283–300.
- Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutarakis, P., Lyapustin, A. et al. (2019). An ensemble-based model of PM2.5 concentration across the contiguous united states with high spatiotemporal resolution. *Environment International*, 130, 104909.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*.
- Engardt, M., Bergström, S., & Johansson, C. (2021). Luftens du andas – nu och de kommande dagarna: Utveckling av ett automatiskt prognosystem för luftföroreningar och pollen. Technical Report SLB 36:2021. SLB-analys, Stockholm Stad. In Swedish.
- Feng, Y., Kim, J.-S., Yu, J.-W., Ri, K.-C., Yun, S.-J., Han, I.-N., Qi, Z., & Wang, X. (2023). Spatiotemporal informer: a new approach based on spatiotemporal embedding and attention for air quality forecasting. *Environmental Pollution*, 336, 122402. <https://doi.org/10.1016/j.envpol.2023.122402>
- Gerges, F., Llaguno-Munitxa, M., Zondlo, M. A., Boufadel, M. C., & Bou-Zeid, E. (2024). Weather and the city: machine learning for predicting and attributing fine scale air quality to meteorological and urban determinants. *Environmental Science & Technology*, 58(14), 6313–6325. <https://doi.org/10.1021/acs.est.4c00783>
- Gidhagen, L., Johansson, C., Langner, J., & Foltescu, V. L. (2005). Urban scale modeling of particle number concentration in stockholm. *Atmospheric Environment*, 39(9), 1711–1725.
- Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., & He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv:1706.02677*.
- Gu, J., Yang, B., Brauer, M., & Zhang, K. M. (2021). Enhancing the evaluation and interpretability of data-driven air quality models. *Atmospheric Environment*, 246, 118125.
- Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv:1606.08415*.
- Hooker, G., Mentch, L., & Zhou, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6), 82.
- Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32.
- Hou, L., Dai, Q., Song, C., Liu, B., Guo, F., Dai, T., Li, L., Liu, B., Bi, X., Zhang, Y., & Feng, Y. (2022). Revealing drivers of haze pollution by explainable machine learning. *Environmental Science & Technology Letters*, 9(2), 112–119. <https://doi.org/10.1021/acs.estlett.1c00865>
- Houdou, A., El Badisy, I., Khomsi, K., Abdala, S. A., Abdulla, F., Najmi, H., Obtel, M., Belyamani, L., Ibrahim, A., & Khalis, M. (2024). Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review. *Aerosol and Air Quality Research*, 24(1), 230151.
- Hu, K., Che, J., Xia, W., Xu, Y., & Li, Y. (2025). Enhancing multi-step air quality prediction with deep learning using residual neural network and adaptive decomposition-based multi-objective optimization. *Expert Systems with Applications*, 273, 126969.
- Karner, A. A., Eisinger, D. S., & Niemeier, D. A. (2010). Near-roadway air quality: Synthesizing the findings from real-world data. *Environmental Science & Technology*, 44(14), 5334–5344.
- Kilbo Edlund, K., Kisieli, M. A., Asker, C., Segersson, D., Bennet, C., Spanne, M., Gustafsson, S., Lindvall, J., Eneroth, K., Tondel, M. et al. (2024). High-resolution dispersion modelling of PM2.5, PM10, NOx and NO2 exposure in metropolitan areas in sweden 2000–2018—large health gains due to decreased population exposure. *Air Quality, Atmosphere & Health*, 17(8), 1661–1675.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv:1412.6980*.
- Laudan, J., Banzhaf, S., Khan, B., & Nagel, K. (2025). Air quality-driven traffic management using high-resolution urban climate modeling coupled with a large traffic simulation. *Atmosphere*, 16(2), 128.
- Li, R., Zeng, D., Li, T., Ti, B., & Hu, Y. (2023). Real-time prediction of SO2 emission concentration under wide range of variable loads by convolution-LSTM VE-transformer. *Energy*, 269, 126781. <https://doi.org/10.1016/j.energy.2023.126781>
- Li, Z., Bi, J., Liu, Y., & Hu, X. (2025). Forecasting o3 and NO2 concentrations with spatiotemporally continuous coverage in southeastern china using a machine learning approach. *Environment International*, 195, 109249.
- Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748–1764.
- Liu, X., Pu, X., Lu, C., Zhang, H., Li, T., Grieneisen, M. L., Li, J., Ma, N., Yan, C., Zhan, Y., & Yang, F. (2024). Interpretable regional meteorological feature extraction enhances deep learning for extended 120-h PM2.5 forecasting. *Journal of Cleaner Production*, 483, 144287. <https://doi.org/10.1016/j.jclepro.2024.144287>
- Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2023). Itransformer: Inverted transformers are effective for time series forecasting. *arXiv:2310.06625*.
- Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv:1705.07874*.
- Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review*, 56(9), 10031–10066. <https://doi.org/10.1007/s10462-023-10424-4>
- MET Norway (2025). Norwegian meteorological institute API documentation. <https://docs.api.met.no/doc/>. Last accessed: 15 June 2025.
- Molnar, C. (2020). Interpretable machine learning. Lulu. com.
- Munkhdalai, L., Munkhdalai, T., Pham, V.-H., Li, M., Ryu, K. H., & Theera-Umporn, N. (2022). Recurrent neural network-augmented locally adaptive interpretable regression for multivariate time-series forecasting. *IEEE Access*, 10, 11871–11885.
- Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2022). A time series is worth 64 words: Long-term forecasting with transformers. *arXiv:2211.14730*.
- Ottosen, T.-B., Kakosimos, K. E., Johansson, C., Hertel, O., Brandt, J., Skov, H., Berkowicz, R., Ellermann, T., Jensen, S. S., & Ketzler, M. (2015). Analysis of the impact of inhomogeneous emissions in the operational street pollution model (OSPM). *Geoscientific Model Development*, 8(10), 3231–3245.
- Qi, Y., Li, Q., Karimian, H., & Liu, D. (2019). A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory. *Science of the Total Environment*, 664, 1–10. <https://doi.org/10.1016/j.scitotenv.2019.01.333>
- Qiu, Y., Feng, J., Zhang, Z., Zhao, X., Li, Z., Ma, Z., Liu, R., & Zhu, J. (2023). Regional aerosol forecasts based on deep learning and numerical weather prediction. *npj Climate and Atmospheric Science*, 6(1), 71.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K.-R. (2016). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11), 2660–2673.
- Sayed, A., Choi, Y., Eslami, E., Jung, J., Lops, Y., Salman, A. K., Lee, J.-B., Park, H.-J., & Choi, M.-H. (2021). A novel CMAQ-CNN hybrid model to forecast hourly surface-ozone concentrations 14 days in advance. *Scientific Reports*, 11(1), 10891. <https://doi.org/10.1038/s41598-021-90446-6>
- Scornet, E. (2023). Trees, forests, and impurity-based variable importance in regression. In *Annales de l'institut henri poincaré (b) probabilités et statistiques* (pp. 21–52). Institut Henri Poincaré (vol. 59).
- Sedgwick, P. (2012). Pearson's correlation coefficient. *BMJ (Clinical Research Ed.)*, 345.
- Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 180–186).
- SMHI (2025). Swedish meteorological and hydrological institute open data portal. <https://www.smhi.se/data/oppna-data>. Last accessed: 15 June 2025.
- Tang, D., Zhan, Y., & Yang, F. (2024). A review of machine learning for modeling air quality: Overlooked but important issues. *Atmospheric Research*, 300, 107261.
- Tian, X., Zhang, C., Liu, H., Zhang, B., Lu, C., Jiao, P., & Ren, S. (2024). Research on air quality in response to meteorological factors based on the informer model. *Sustainability*, 16(16), 6794. <https://doi.org/10.3390/su16166794>
- Tuccella, P., Curci, G., Visconti, G., Bessagnet, B., Menut, L., & Park, R. J. (2012). Modeling of gas and aerosol with WRF/chem over europe: Evaluation and sensitivity study. *Journal of Geophysical Research: Atmospheres*, 117(D3).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wang, H., Zhang, L., & Wu, R. (2023). MSAFormer: A transformer-based model for PM2.5 prediction leveraging sparse autoencoding of multi-site meteorological features in urban areas. *Atmosphere*, 14(8), 1294. <https://doi.org/10.3390/atmos14081294>
- Wang, Y., Wu, H., Dong, J., Qin, G., Zhang, H., Liu, Y., Qiu, Y., Wang, J., & Long, M. (2024). Timexer: Empowering transformers for time series forecasting with exogenous variables. *arXiv:2402.19072*.
- Wilks, D. S. (2011). Statistical methods in the atmospheric sciences (vol. 100). Academic press.

- WHO, W. H. O. (2024). Exposure & health impacts of air pollution. <https://www.who.int/teams/environment-climate-change-and-health/air-quality-energy-and-health/health-impacts/exposure-air-pollution#:~:text=An%20estimated%204.2%20protect%20leavevmode@ifvmode\kern+.1667em\relax%20million%20deaths,cancer%20and%20acute%20respiratory%20infections>
- Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34, 22419–22430.
- Yang, W., Li, H., Wang, J., & Ma, H. (2024). Spatio-temporal feature interpretable model for air quality forecasting. *Ecological Indicators*, 167, 112609.
- Yu, C., Wang, F., Wang, Y., Shao, Z., Sun, T., Yao, D., & Xu, Y. (2025). Mgsformer: A multi-granularity spatiotemporal fusion transformer for air quality prediction. *Information Fusion*, 113, 102607.
- Yu, M., Masrur, A., & Blaszcak-Boxe, C. (2023). Predicting hourly PM_{2.5} concentrations in wildfire-prone areas using a spatiotemporal transformer model. *Science of the Total Environment*, 860, 160446.
- Zeng, Y., Chen, J., Jin, N., Jin, X., & Du, Y. (2022). Air quality forecasting with hybrid LSTM and extended stationary wavelet transform. *Building and Environment*, 213, 108822. <https://doi.org/10.1016/j.buildenv.2022.108822>
- Zhang, A., Fu, T.-M., Feng, X., Guo, J., Liu, C., Chen, J., Mo, J., Zhang, X., Wang, X., Wu, W., Hou, Y., Yang, H., & Lu, C. (2023a). Deep learning-based ensemble forecasts and predictability assessments for surface ozone pollution. *Geophysical Research Letters*, 50(8), e2022GL102611. <https://doi.org/10.1029/2022GL102611>
- Zhang, Y., & Yan, J. (2023). Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*.
- Zhang, Z., Johansson, C., Engardt, M., Stafoggia, M., & Ma, X. (2024). Improving 3-day deterministic air pollution forecasts using machine learning algorithms. *Atmospheric Chemistry and Physics*, 24(2), 807–851.
- Zhang, Z., Ma, X., Johansson, C., Jin, J., & Engardt, M. (2023b). A meta-graph deep learning framework for forecasting air pollutants in stockholm. In *2023 IEEE 9th world forum on internet of things (WF-IoT)* (pp. 01–06). IEEE.
- Zhang, Z., Ma, X., & Schlesinger, D. (2025). Interpretable air pollution forecasting by physics-guided spatiotemporal decoupling. [arXiv:2511.20257](https://arxiv.org/abs/2511.20257).
- Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11106–11115). (vol. 35).
- Zhou, Y., Chang, F.-J., Chang, L.-C., Kao, I.-F., & Wang, Y.-S. (2019). Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. *Journal of Cleaner Production*, 209, 134–145.