AirMPA: A Meteorology-to-Pollution Adapter for Global Air Quality Forecasting

Yiheng Wang[1,2], **Kai Zheng**[2], **Yuetan Lin**[2], **Fanglu Fan**[1,2], **Chenliang Tao**[1], **Guochao Chen**[1], **Hongliang Zhang**[3,*], **Hao Li**[1,2,*]

[1]Fudan University [2]Shanghai Academy of AI for Science
[3]University of Shanghai for Science and Technology

Abstract

Accurate real-time forecasting of atmospheric pollutants is essential for reducing exposure risks and improving public health. Traditional numerical forecasting systems suffer from incomplete parameterizations, uncertain inputs, and high computational cost. Existing deep learning approaches improve efficiency, but often entangle meteorological and pollutant variables within unified forecasting frameworks, limiting flexibility and pollutant-specific representation learning. To address these challenges, we propose AirMPA, a decoupled autoregressive framework for meteorology-to-pollution forecasting. AirMPA predicts 13 atmospheric pollutants at 0.4° spatial resolution and 12-h temporal resolution using meteorological fields, emission priors, and historical pollutant concentrations. The model is built on an improved Swin-Transformer architecture with zonal release attention and dual-branch fusion, enabling more effective modeling of global transport continuity and heterogeneous pollutant dynamics. During training, AirMPA uses the fifth-generation atmospheric reanalysis from the European Centre for Medium-Range Weather Forecasts (ERA5) as meteorological forcing, while during inference it directly ingests forecast fields from external meteorological models, enabling stable autoregressive prediction up to 5 days ahead. Experiments show that AirMPA consistently outperforms Aurora and surpasses the Copernicus Atmosphere Monitoring Service (CAMS) in approximately 90% of pollutant–lead-time evaluation cases over the 5-day forecast horizon, while outperforming CAMS on all variables beyond 48 h. These results demonstrate the potential of decoupled meteorology-to-pollution forecasting for global air quality prediction.

1 Introduction

Air pollution has become a major global public health threat, causing millions of premature deaths each year and imposing substantial socioeconomic burdens [1, 2, 3]. Accurate multi-day forecasting of atmospheric pollutants is therefore essential for reducing exposure risks, supporting environmental management, and mitigating extreme pollution events[4, 5].

Air quality forecasting is intrinsically challenging because atmospheric pollutant evolution is governed by highly nonlinear interactions among meteorology, chemistry, and emissions, making it difficult to simultaneously achieve high accuracy, computational efficiency, and long-term stability [6, 7]. Traditional numerical models, such as WRF-Chem and CMAQ [8, 9], can explicitly simulate complex physical and chemical processes, but their practical use in real-time forecasting is limited by high computational cost as well as uncertainties in emission inventories and chemical parameterizations [10, 11, 12]. Data-driven approaches have substantially improved computational efficiency and shown strong potential for atmospheric forecasting [13, 14, 15, 16]. Recent studies have further demonstrated that air pollution prediction can be incorporated into unified meteorological forecasting frameworks, achieving performance competitive with traditional systems such as the Copernicus Atmosphere

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Monitoring Service (CAMS) [17, 18]. However, these methods usually couple meteorological and pollutant variables at the data level, which reduces modeling flexibility and may weaken the representation of pollutant-specific dynamics, thereby limiting forecasting accuracy and transferability across meteorological systems.

Meanwhile, the rapid advancement of meteorological foundation models provides new opportunities for addressing this challenge. High-accuracy forecasting systems include the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF) and data-driven models such as the Artificial Intelligence Forecasting System (AIFS), FuXi, Pangu-Weather, and GraphCast. These systems have achieved substantial progress in medium- to long-range weather forecasting, providing high-resolution and high-fidelity atmospheric fields. [19, 20, 21, 22, 23]. In traditional numerical models, pollutant prediction is inherently driven by meteorological processes [24, 25]. This paradigm suggests the possibility of directly using meteorological forecasts to drive downstream air pollutant prediction.

In this work, we propose AirMPA, a decoupled global air pollutant forecasting framework that explicitly separates meteorological forecasting from pollutant prediction. AirMPA produces 5-day forecasts of 13 atmospheric pollutants at 0.4° spatial resolution and 12 h temporal resolution. Our main contributions are as follows:

- We propose AirMPA, a decoupled meteorology-to-pollution forecasting framework that treats pollutant prediction as an independent downstream task driven by external meteorological forcing, producing global 12-hourly forecasts at 0.4° resolution.
- The decoupled design enables a clean transfer from reanalysis-driven training to forecast-driven inference, providing a practical pathway for using external meteorological forecasts in downstream air pollutant prediction.
- The improved Swin-Transformer architecture, equipped with zonal release attention and dual-branch historical fusion, better captures the continuity of global atmospheric pollutant transport and the heterogeneity of pollutant temporal evolution, thereby improving forecasting accuracy.
- Validation against station observations and case studies of global ozone pollution events further demonstrate the practical significance of AirMPA.

2 Related Work

2.1 Numerical Models

CAMS, operated by ECMWF, provides global atmospheric forecasts twice daily [26]. By integrating satellite observations and ground-based measurements through data assimilation, CAMS yields consistent initial states to drive physical and chemical transport models. The system outputs over 50 chemical species and 7 aerosol types (e.g., dust, black carbon, and sulfates) at a standard horizontal resolution of approximately $0.4^\circ \times 0.4^\circ$, covering a 5-day forecast horizon. These comprehensive datasets serve as a foundational benchmark for global environmental governance, supporting both real-time air quality management and long-term climate change research.

2.2 Data-Driven Deep Learning Models

Recent advances in deep learning have improved air pollution forecasting across station, regional, and global scales [15, 27, 28, 29, 30, 31, 32]. At the station scale, AirFormer combines deterministic spatiotemporal attention with stochastic latent variables for nationwide forecasting [15]. For regional gridded prediction, TopoFlow incorporates wind-guided patch reordering and topography-aware attention to represent transport and terrain effects [14]. Air-DualODE integrates open-system physical dynamics with a complementary data-driven branch, improving prediction across city and national scales [33].

At the global scale, Aurora employs a flexible encoder-decoder architecture with a 3D Swin-Transformer backbone, delivering forecasts at 0.4° resolution that outperform conventional chemical transport models in 74% of evaluation metrics [18]. AI-GAMFS provides multi-day global aerosol forecasts at 3-h resolution in under one minute, reducing dust concentration errors by 34%–74%

compared to regional physical models [17]. More recently, AIFS-COMPO jointly predicts meteorological and composition variables at approximately 80-km and 3-h resolution. It matches or improves upon CAMS for several species against independent observations, but remains less accurate for total-column ozone (O_3) and near-surface nitrogen dioxide (NO_2) against CAMS analysis [31].

While coupled global systems jointly forecast meteorology and pollutants, Fan et al. [34] explore offline meteorology–pollution coupling through bilinear pooling. AirMPA also adopts a decoupled formulation, with explicit emission conditioning, zonal release attention, and pollutant-specific historical fusion. It treats pollutant forecasting as a dedicated downstream task and can incorporate forecasts from different meteorological systems without retraining the pollutant model.

3 Meteorology to Pollutant Adapter

3.1 Problem Formulation

To improve modularity and enable direct reuse of high-quality meteorological forecasts, we formulate global air pollutant forecasting as a meteorology-driven autoregressive prediction problem. Instead of jointly modeling meteorology and atmospheric composition within a unified framework, AirMPA decouples meteorological forecasting from pollutant prediction and treats the latter as a downstream task conditioned on external meteorological fields, emission priors, and historical pollutant states.

During training, we assume access to meteorological sequences $\mathbf{M}_{0:T}$, emission sequences $\mathbf{E}_{0:T}$, and pollutant concentration sequences $\mathbf{P}_{0:T}$. Here, $\mathbf{M}_t \in \mathbb{R}^{C_m \times H \times W}$, $\mathbf{E}_t \in \mathbb{R}^{C_e \times H \times W}$, and $\mathbf{P}_t \in \mathbb{R}^{C_p \times H \times W}$ denote the meteorological fields, emission priors, and pollutant concentrations at time step t, respectively, where C_m, C_e, and C_p are the channel numbers and H, W are the spatial dimensions on the 0.4° global grid. In this study, the meteorological inputs are taken from ERA5 reanalysis during training and replaced by forecast fields from FuXi-2.0 during inference. Under this formulation, the meteorological fields at the target prediction time serve as forcing inputs for pollutant forecasting, driving transport, diffusion, and chemical evolution. To reduce error accumulation and preserve short-term pollutant memory, the prediction at time step t is conditioned on the two preceding pollutant states, the meteorological and emission inputs at time steps $t-1$ and t, and the temporal encoding of the target prediction time ($\boldsymbol{\tau}_t$):

$$\hat{\mathbf{P}}_t = f_\theta(\mathbf{P}_{t-2}, \mathbf{P}_{t-1}, \mathbf{M}_{t-1}, \mathbf{M}_t, \mathbf{E}_{t-1}, \mathbf{E}_t, \boldsymbol{\tau}_t), \tag{1}$$

Under this formulation, AirMPA serves as a plug-and-play pollution forecasting module that enables a clean transition from reanalysis-driven training to forecast-driven inference.

3.2 AirMPA Overview

AirMPA is built upon an improved Swin Transformer backbone for global atmospheric pollutant forecasting [35]. Given meteorological fields, emission priors, and historical pollutant concentrations, the model predicts future pollutant states in an autoregressive manner. As illustrated in Figure 1(a), the overall framework follows the pipeline of *input adaptation* $\rightarrow$ *cube embedding* $\rightarrow$ *spatiotemporal feature extraction* $\rightarrow$ *dual-branch pollutant fusion* $\rightarrow$ *multi-pollutant prediction*. To ensure compatibility with the backbone, the native global grid is first resized from 900×451 to 896×448 using bilinear interpolation, and the predicted fields are restored to the original resolution after decoding.

To encode short-term temporal dependencies before spatial feature extraction, AirMPA introduces a temporal cube embedding module. Let $\tilde{\mathbf{X}}_t^{(i)} \in \mathbb{R}^{C_i \times H \times W}$ denote the input variables at the i-th temporal slice. Stacking all slices and applying temporal convolution give

$$\tilde{\mathbf{X}}_t^{\text{seq}} \in \mathbb{R}^{B \times T \times C \times H \times W}, \qquad \mathbf{F}^{\text{temp}} = \text{Conv}_{3\text{D}}(\tilde{\mathbf{X}}_t^{\text{seq}}), \tag{2}$$

where B, T, and C denote the batch size, number of temporal inputs, and total number of channels, respectively.

The resulting features are then aggregated by attention pooling to obtain a compact embedded representation:

$$\mathbf{F}^{\text{emb}} = \sum_{i=1}^{T} \alpha_i \mathbf{F}_i^{\text{temp}}, \qquad \alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{T} \exp(s_j)}, \tag{3}$$

where s_i and α_i denote the learnable attention score and normalized weight of the i-th temporal slice. This module summarizes multi-source, multi-step inputs into a compact spatiotemporal representation for downstream forecasting.

The embedded features are then processed by stacked AirMPA blocks for spatiotemporal feature extraction. Built on this backbone, AirMPA incorporates two task-specific designs: (1) zonal release attention and (2) dual-branch fusion, and outputs forecasts for all pollutants at the current step.

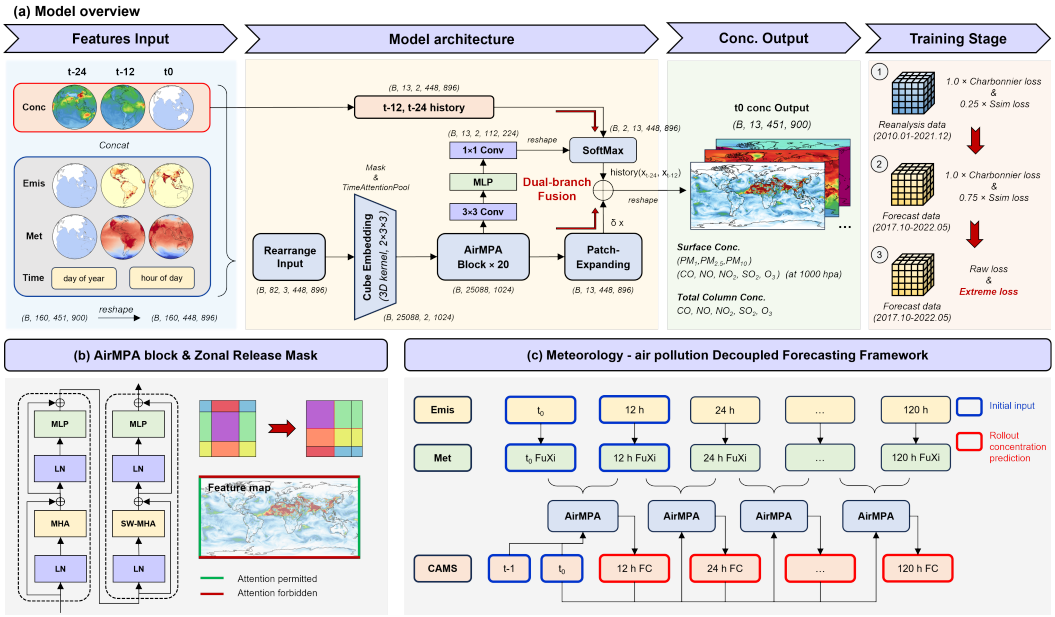

Figure 1: AirMPA architecture. (a) Overview of the AirMPA architecture and data flow. (b) Illustration of the zonal release attention mechanism, which mitigates the artificial discontinuity at the global longitude boundary. (c) Autoregressive inference pipeline for multi-step pollutant forecasting.

3.2.1 Zonal Release Attention

The backbone of AirMPA is built upon shifted-window Swin Transformer blocks, which were originally developed for image processing under a planar-grid assumption with non-periodic boundaries. However, directly applying the standard shifted-window masking strategy to global latitude–longitude grids introduces an artificial discontinuity at the longitude boundary. On the Earth system grid, the left and right boundaries are physically adjacent due to the periodic topology of the globe in longitude, whereas the original masking mechanism treats them as disconnected regions. Consequently, information exchange across the dateline is unnecessarily blocked, which may hinder the representation of zonal continuity and trans-boundary transport, especially for large-scale and long-lived pollutants.

To address this mismatch, we introduce a Zonal Release Attention mechanism (Figure 1). The key idea is to relax the boundary constraint only along the longitude dimension while preserving the standard shifted-window masking strategy along the latitude dimension. Specifically, tokens located at the two opposite longitude edges are grouped into the same attention region and computed without masking. Let $\mathbf{M}_{ij}$ denote the original shifted-window attention mask. We define a modified mask $\mathbf{M}^*_{ij}$ as

$$\mathbf{M}^*_{ij} = \begin{cases} 0, & \text{if tokens } i \text{ and } j \text{ are grouped across the longitude boundary,} \\ \mathbf{M}_{ij}, & \text{otherwise,} \end{cases} \tag{4}$$

where the zero mask indicates that no artificial separation is imposed for cross-boundary token pairs.

The resulting attention is computed as

$$\mathrm{Attn}^*(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{B} + \mathbf{M}^*\right)\mathbf{V}, \tag{5}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the query, key, and value matrices, d is the channel dimension, and $\mathbf{B}$ is the relative position bias. Compared with standard shifted-window attention, this modification does not increase the overall architectural complexity, but allows physically adjacent regions near the longitude boundaries to interact directly within the same attention window. In this way, the model better respects the periodic geometry of the Earth system and can more faithfully capture cross-boundary transport patterns in global atmospheric composition fields.

3.2.2 Dual-branch Fusion for Multi-pollutant Forecasting

Although atmospheric pollutant evolution is continuous in time, different pollutants exhibit markedly different temporal response characteristics under the 12-hour forecasting interval used in this work [36, 37, 27]. For example, O_3 and nitric oxide (NO) are strongly influenced by photochemical cycles and solar radiation, whereas fine particulate matter ($PM_{2.5}$) and sulfur dioxide (SO_2) are often more strongly constrained by persistence, transport accumulation, and emission carryover. Therefore, directly using a shared temporal fusion strategy for all pollutants may be suboptimal, because it implicitly assumes homogeneous temporal inheritance patterns across variables. To address this issue, we design a dual-branch fusion module that combines two complementary information sources: the latent spatiotemporal features extracted by the Transformer backbone and the historical pollutant reference fields that provide direct temporal guidance. Let

$$\mathbf{F}^{\text{back}} \in \mathbb{R}^{B\times C_p\times H\times W}, \qquad \mathbf{X}^{\text{ref}} \in \mathbb{R}^{B\times C_p\times T_r\times H\times W}, \tag{6}$$

denote the backbone output for pollutant prediction and the historical pollutant reference tensor constructed from previous concentration fields, respectively, where C_p is the number of target pollutants and T_r is the number of reference steps. In this study, $T_r = 2$, corresponding to $\mathbf{P}_{t-2}$ and $\mathbf{P}_{t-1}$.

Instead of assigning fixed temporal weights, the historical-reference branch learns spatially adaptive and pollutant-specific weighting coefficients for different reference states. The fused reference feature is defined as

$$R_{b,c,h,w} = \sum_{i=1}^{T_r} \left(\frac{\exp(L_{b,c,i,h,w})}{\sum_{j=1}^{T_r} \exp(L_{b,c,j,h,w})} \right) X^{\text{ref}}_{b,c,i,h,w}, \tag{7}$$

where $L_{b,c,i,h,w}$ denotes the learnable logit associated with the i-th historical reference at location (h, w) for pollutant channel c. The softmax normalization ensures that the contributions of different historical states are adaptively balanced at each spatial position and for each pollutant.

The final prediction is obtained by combining the backbone branch and the historical-reference branch:

$$\hat{\mathbf{P}}_t = \phi\left(\mathbf{F}^{\text{back}} + \mathbf{R}\right), \tag{8}$$

where $\mathbf{R}$ denotes the fused historical reference feature and $\phi(\cdot)$ is the final projection layer. This formulation can be interpreted as a residual temporal-guidance mechanism: the backbone branch captures large-scale meteorological and transport-driven dynamics, whereas the reference branch injects pollutant-specific persistence and recent-evolution information. By introducing spatially and channel-wise adaptive temporal weighting, the model can learn heterogeneous temporal inheritance patterns across pollutants rather than relying on a uniform persistence assumption. This is particularly important for multi-pollutant forecasting, where different variables are governed by distinct physical and chemical processes.

4 Experiments

4.1 Dataset

The input data consist of three components: pollutant concentrations, meteorological variables, and emission data. During training, meteorological inputs are derived from the European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) dataset. During inference, these meteorological inputs are replaced by forecast fields from FuXi-2.0, so that the pollutant prediction model is evaluated under forecast-driven conditions. Detailed information on the dataset is provided in Appendix A.

4.2 Training

4.2.1 Loss function

We optimize AirMPA with a hybrid weighted loss to jointly enforce reconstruction accuracy and spatial structural consistency. Specifically, the reconstruction term is formulated as a latitude-weighted Charbonnier loss to account for the non-uniform area of the global latitude–longitude grid, while an SSIM loss is introduced to preserve the spatial structure of pollutant fields. The detailed definitions of the loss terms are provided in Appendix B.1. The overall objective is written as

$$\mathcal{L} = \alpha \mathcal{L}_{\text{Lat-weighted Charbonnier}} + \beta \mathcal{L}_{\text{SSIM}}. \tag{9}$$

where w_i denotes the latitude-dependent area weight for the i-th latitude band, ϕ_i is the corresponding latitude, and H is the total number of latitude bands, $\mathcal{L}_{\text{Charbonnier}}$ is the Charbonnier reconstruction loss, $\mathcal{L}_{\text{SSIM}}$ is the structural similarity loss, and α and β are stage-dependent weighting coefficients.

4.2.2 Training stages

AirMPA is optimized using a three-stage progressive training strategy. The first stage performs global pre-training on CAMS reanalysis data to capture large-scale spatiotemporal dependencies among meteorology, emissions, and pollutant evolution. The second stage fine-tunes the model on CAMS forecast data to adapt the output distribution to operational forecasting products. The third stage further improves the prediction of high-concentration events through extreme-oriented optimization. Detailed loss weights and optimization settings for all stages are provided in Appendix B.2.

4.2.3 Implementation and Training Details

AirMPA is implemented in PyTorch and trained using the three-stage progressive strategy described above to ensure stable convergence across different pollutants and forecast horizons. The model adopts a patch size of 4, a Swin window size of 8, and an embedding dimension of 1024, with 20 stacked AirMPA blocks and 16 attention heads in each block. In the cube embedding module, the temporal convolution uses a kernel size of 2, and a dropout rate of 0.1 is applied throughout the network to improve generalization. Training is performed on a single NVIDIA A100 GPU with a batch size of 2 and requires approximately 72 hours in total. We optimize the model using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.99$, weight decay $= 10^{-3}$) together with a cosine annealing learning-rate schedule.

4.3 Evaluation

To ensure physically consistent evaluation on a latitude–longitude grid, we adopt the latitude-weighted RMSE as the primary metric.

$$\text{RMSE}_{\text{lat-weight}} = \sqrt{\frac{\sum_{i=1}^{N_\phi} \sum_{j=1}^{N_\lambda} \cos(\phi_i) \left(\hat{y}_{i,j} - y_{i,j}\right)^2}{\sum_{i=1}^{N_\phi} \sum_{j=1}^{N_\lambda} \cos(\phi_i)}}, \tag{10}$$

where $\hat{y}i, j$ and yi, j denote the predicted and reference values at each grid point.

To evaluate extreme-event forecasting, we adopt two metrics defined at the 90th percentile threshold, namely RMSE_{90} and CSI_{90}:

$$\text{RMSE}_{90} = \sqrt{\frac{1}{|\Omega_{90}|} \sum_{(i,j)\in\Omega_{90}} \left(\hat{y}_{i,j} - y_{i,j}\right)^2}, \qquad \text{CSI}_{90} = \frac{TP}{TP + FP + FN}, \tag{11}$$

where $\Omega_{90} = \{(i, j) \,|\, y_{i,j} \geq q_{0.9}\}$ is the set of grid points exceeding the 90th percentile threshold $q_{0.9}$, and TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

4.4 Inference

AirMPA performs multi-step autoregressive forecasting of atmospheric pollutant concentrations using external meteorological forecasts (Figure 1(c)). During inference, emission priors provide

background constraints on pollutant sources, FuXi-2.0 forecast fields serve as meteorological forcing, and pollutant concentration fields from CAMS FC at $t_0 - 12\,\mathrm{h}$ and t_0 are used to initialize the autoregressive process. Starting from these initial states, the model iteratively predicts pollutant fields at 12-hour intervals up to 120 hours (5 days), with each prediction conditioned on the two most recent pollutant fields together with the corresponding meteorological forecasts and emission priors. For fair comparison, AirMPA, Aurora, and CAMS use the same pollutant initialization at $t_0 - 12\,\mathrm{h}$ and t_0, so that performance differences arise from the forecasting framework itself and the meteorological inputs used thereafter.

5 Main Results

5.1 Model Performance

Overall latitude-weighted RMSE. Figure 2 compares the overall latitude-weighted RMSE of AirMPA (with FuXi), CAMS, and Aurora. AirMPA generally achieves lower RMSE, with larger improvements at longer lead times. We further evaluate AirMPA with ERA5 reanalysis as an offline reference. Its lower RMSE suggests that more accurate meteorological forcing could further improve pollutant forecasts; comparisons across meteorological inputs are reported in Table 2. A broader comparison (Figure 5) shows that AirMPA consistently outperforms Aurora across all 13 pollutants and surpasses CAMS on 90% of metrics over the 5-day horizon and on 100% beyond 48 hours. Similar improvements are observed over land-only regions, supporting the practical value of the proposed approach.

We additionally compare AirMPA with the offline meteorology–pollution model of Fan et al. [34] over June–November 2022. Across eight shared pollutant variables at lead times of 1, 3, and 5 days, AirMPA achieves lower RMSE in 23 of 24 cases, with a mean relative reduction of 16.65%. The only exception is total-column O_3 at day 5, for which RMSE increases by 3.88%. Detailed results are provided in Appendix F.1 (Table 9). Although the evaluation period and pollutant variables are matched, the two models use independently developed training and preprocessing pipelines. This comparison therefore assesses their predictive performance rather than isolating the effect of meteorology–pollution decoupling.

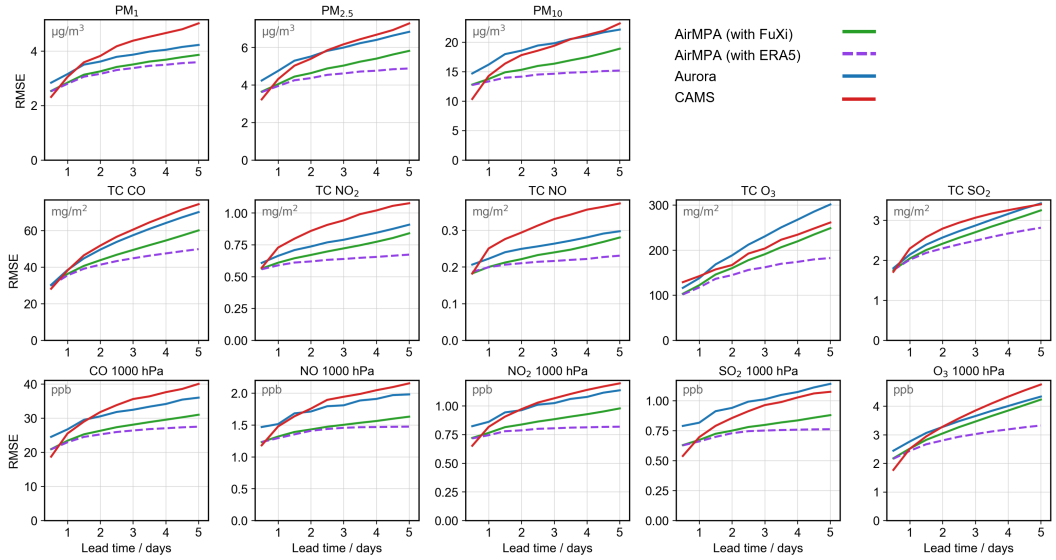

Figure 2: Comparison of forecasting performance across AirMPA (with FuXi), AirMPA (with ERA5), CAMS, and Aurora, evaluated by latitude-weighted RMSE (lower is better).

Regional variation of metrics. Figure 3 presents the spatial variation of evaluation metrics for several key pollutants, with additional results shown in Appendix Figure 6. AirMPA consistently outperforms Aurora and CAMS in terms of RMSE across lead times, variables, and most regions. The improvements are particularly evident in densely populated and heavily polluted areas, including East Asia, the Indian subcontinent, and Central Africa. As these regions are associated with higher

exposure and health risks [38, 39], the improved forecasting accuracy of AirMPA in these areas highlights its strong practical significance for air pollution prevention and public health protection.

Extreme metrics. Accurate prediction of extreme pollution events is a key objective of air quality forecasting, yet it has received limited attention in prior evaluations [40, 41]. Figure 9 shows that AirMPA consistently outperforms Aurora in both RMSE_{90} and CSI_{90} across pollutants, and achieves lower RMSE_{90} than CAMS in 83% of cases while maintaining comparable CSI_{90}. The advantage is particularly evident for $PM_{2.5}$, PM_{10}, and O_3 at longer lead times.

Observation validation metrics. We further evaluate the models using monitoring-station observations of major pollutant concentrations from China, the United States, and Europe during the evaluation period (Figure 10). For PM_{10} and $PM_{2.5}$, both AirMPA and Aurora achieve lower RMSE than CAMS across all three regions, with AirMPA performing best and showing slower error accumulation as forecast lead time increases. For O_3, AirMPA performs comparably to Aurora while still substantially outperforming CAMS. These observation-based results show that the advantages of AirMPA extend beyond gridded benchmarks to real-world monitoring data, further highlighting its practical value for operational air quality forecasting and early warning.

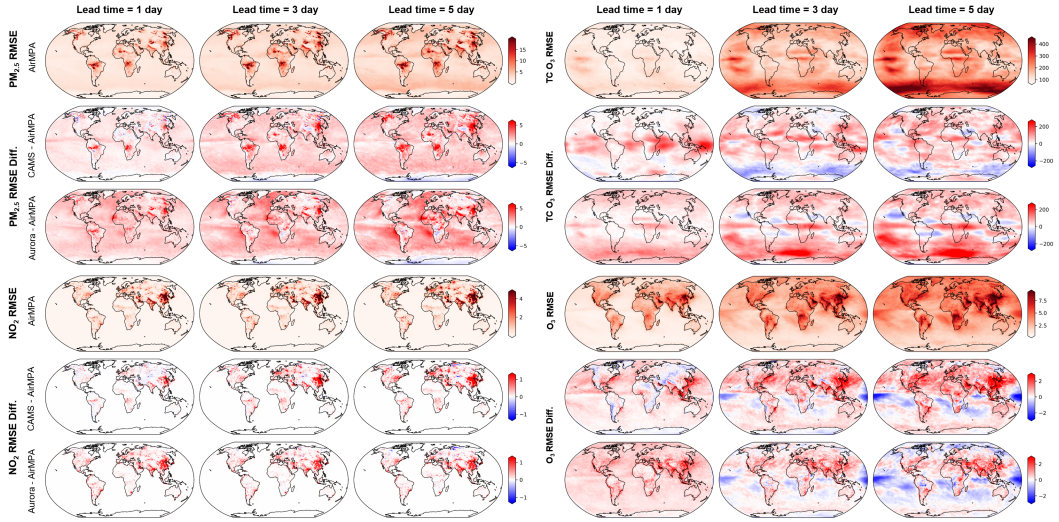

Figure 3: Spatial distribution of RMSE and RMSE differences for $PM_{2.5}$, total column O_3, surface NO_2, and surface O_3 over the evaluation period. In each panel, rows from top to bottom show AirMPA RMSE, RMSE difference ($\mathrm{CAMS} - \mathrm{AirMPA}$), and RMSE difference ($\mathrm{Aurora} - \mathrm{AirMPA}$). Red regions indicate better performance of AirMPA.

5.2 Case Study

Figure 4 presents a global O_3 pollution event at 12:00 on August 14, 2022. Forecast results from CAMS, Aurora, and AirMPA are compared at both global and regional scales (East Asia: 75°E–145°E, 5°N–55°N; North America: 130°W–60°W, 20°N–70°N). All three models capture the overall global distribution patterns. At the regional scale, the 1-day forecasts from both AirMPA and CAMS closely match the CAMS FC reference field at the valid time. In contrast, Aurora exhibits insufficient spatial detail and underestimates extreme values over southern China and central North America, with these issues becoming more pronounced at lead times of 3 and 5 days. As a numerical model, CAMS demonstrates advantages in capturing extreme values but tends to produce spurious overestimations, such as the false high O_3 concentrations over Greenland at a lead time of 5 days. Although AirMPA still shows some limitations in resolving fine-scale details and tends to underestimate certain extreme values, it overall delivers the best forecasting performance over both East Asia and North America.

5.3 Ablation Study

Effect of the decoupled formulation. We compare a decoupled Swin baseline without AirMPA-specific components against Aurora, matching its Swin architecture as closely as possible while

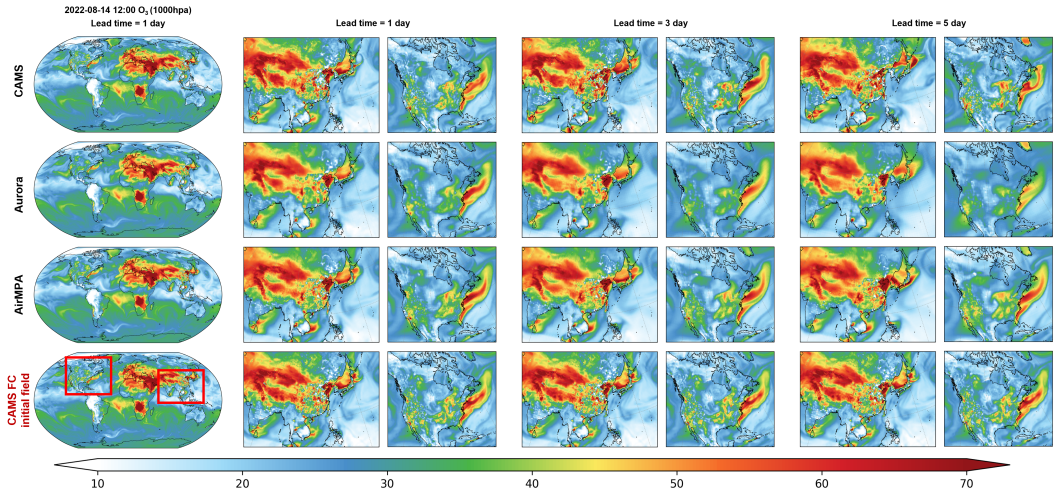

Figure 4: Comparison of surface O_3 forecasts at 12:00 on August 14, 2022. From top to bottom, the panels show results from CAMS, Aurora, AirMPA, and CAMS FC initial field at the valid forecast time.

Table 1: Decoupled Swin versus Aurora. ΔRMSE denotes the mean relative RMSE change against Aurora; negative values indicate improvement. Better cases count lower-RMSE results across 13 pollutants at each lead time.

Metric	1d	3d	5d	Overall
Better cases	11/13	11/13	10/13	32/39
Mean ΔRMSE	-5.99%	-8.17%	-7.43%	-7.20%

retaining some differences in inputs and hyperparameters (Table 1). The 280M-parameter baseline outperforms the 1.3B-parameter Aurora in 32 of 39 pollutant–lead-time cases, reducing RMSE by 7.20% on average. These results suggest that the decoupled formulation itself contributes to improved forecasting performance beyond the gains from the proposed architectural components.

Sensitivity to meteorological forcing. We evaluate AirMPA with ERA5, FuXi, and ECMWF HRES inputs without modifying or retraining the pollutant model (Table 2). FuXi and HRES yield similar performance, outperforming CAMS in 38 of 39 cases and reducing RMSE by 14.69% and 15.26% on average, respectively. ERA5 yields a larger reduction of 22.27%, with the gap widening at longer lead times. These results demonstrate that AirMPA can incorporate forecasts from different meteorological systems without retraining, while the stronger performance with ERA5 suggests that more accurate meteorological forcing could further improve pollutant forecasts.

A complete pollutant-specific comparison at lead times of 1, 3, and 5 days is provided in Appendix F.2 (Table 10). Relative to FuXi forcing, ERA5 reduces RMSE in 37 of 39 cases, with mean reductions of 1.19%, 8.42%, and 15.95%, respectively. HRES yields smaller mean reductions of 0.31%, 1.45%, and 2.35%, with the largest benefit for total-column SO_2.

Table 2: AirMPA performance under different meteorological inputs. Values are mean relative RMSE changes against CAMS across 13 pollutants; negative values indicate improvement. Overall averages the 39 pollutant–lead-time cases. ERA5 provides an offline reanalysis reference.

Input	1d	3d	5d	Overall
ERA5	-10.72%	-24.89%	-31.21%	-22.27%
FuXi	-8.63%	-17.65%	-17.80%	-14.69%
HRES	-8.81%	-18.15%	-18.80%	-15.26%

Contributions of architectural components. To validate the effectiveness of the proposed architectural components, we conduct ablation studies (Table 8) and evaluate different model variants

using the latitude-weighted RMSE over a 5-day forecasting horizon (Table 3). The results show that AirMPA achieves the best or tied-best performance across all reported variables and lead times, while different components contribute to different pollutant types in distinct ways. For emission-sensitive particulate pollutants such as $PM_{2.5}$, incorporating emission inputs consistently improves forecasting performance across the reported lead times, highlighting the role of emission priors in capturing the temporal evolution of local pollution. In contrast, near-surface O_3 and total-column NO_2 exhibit lower sensitivity to emission inputs; for long-range forecasts, adding emission inputs may even increase RMSE. The zonal release attention mechanism provides limited benefit for local pollutants such as $PM_{2.5}$, but improves the prediction of total-column NO_2 and O_3, consistent with its intended role in representing large-scale transport continuity.

Overall, introducing emission features reduces the average RMSE by approximately 0.8–2.1 units. Cube embedding and zonal release attention yield moderate gains, whereas the model with dual-branch fusion delivers the second-best performance in nearly all experiments. Relative to the original Swin baseline, the variant incorporating emission inputs and dual-branch fusion lowers the average RMSE by 7.9% for 1-day forecasts and 10.1% for 5-day forecasts, supporting its contribution to the performance of the full model.

Table 3: Ablation results of key architectural components in AirMPA (lower is better). $PM_{2.5}$ is in $\mu g/m^3$, O_3 is in ppb, and TC NO_2 is in mg/m^2. Best values are shown in bold, and second-best values are underlined. * indicates models built on the version with emission input.

Model	$PM_{2.5}$↓		O_3↓		TC NO_2↓		Average RMSE↓	
	1d	5d	1d	5d	1d	5d	1d	5d
Swin Transformer (Baseline)	4.31	6.08	2.595	4.435	0.651	0.870	17.68	32.20
w/ Emission Input	4.29	6.03	2.590	4.630	0.648	0.872	16.84	30.08
w/ Cube Embedding*	4.25	6.13	2.558	4.425	<u>0.619</u>	<u>0.845</u>	16.76	30.00
w/ Zonal Release Attention*	4.28	6.03	2.543	4.420	0.627	0.864	16.80	30.05
w/ Dual-branch Fusion*	<u>4.12</u>	<u>5.79</u>	<u>2.540</u>	<u>4.361</u>	0.622	**0.826**	<u>16.28</u>	<u>28.95</u>
AirMPA	**4.02**	**5.72**	**2.505**	**4.345**	**0.610**	**0.826**	**16.05**	**28.86**

6 Conclusions

We propose AirMPA, a decoupled autoregressive framework for global air pollutant forecasting that separates meteorological forecasting from pollutant prediction. By reformulating air quality prediction as a meteorology-driven downstream task, AirMPA establishes a flexible and scalable paradigm that can directly interface with external meteorological foundation models and can readily benefit from improvements in meteorological forecasting under the proposed decoupled formulation. Built upon an improved Swin-Transformer architecture, the model integrates zonal release attention and dual-branch fusion to effectively capture the continuity of global pollutant transport and the heterogeneous dynamics of different pollutants, resulting in consistently strong forecasting performance across variables and lead times. AirMPA provides a practical and extensible pathway for bringing meteorological foundation models into air quality forecasting, and lays a foundation for future fully AI-driven global atmospheric composition prediction systems.

References

[1] Chenliang Tao, Yuqiang Zhang, Drew Shindell, and Hongliang Zhang. Inequality in air pollution–attributable mortality by income level between and within countries. *Proceedings of the National Academy of Sciences*, 122(41):e2511394122, 2025.

[2] Chenchao Zhan, Min Xie, Hua Lu, Bojun Liu, Zheng Wu, Tijian Wang, Bingliang Zhuang, Mengmeng Li, and Shu Li. Impacts of urbanization on air quality and the related health risks in a city with complex terrain. *Atmospheric Chemistry and Physics*, 23(1):771–788, 2023.

[3] Jos Lelieveld, John S Evans, Mohammed Fnais, Despina Giannadaki, and Andrea Pozzer. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569):367–371, 2015.

[4] Adil Masood and Kafeel Ahmad. A review on emerging artificial intelligence (ai) techniques for air pollution forecasting: Fundamentals, application and performance. *Journal of cleaner production*, 322:129072, 2021.

[5] Shankar Subramaniam, Naveenkumar Raju, Abbas Ganesan, Nithyaprakash Rajavel, Maheswari Chenniappan, Chander Prakash, Alokesh Pramanik, Animesh Kumar Basak, and Saurav Dixit. Artificial intelligence technologies for forecasting air pollution and human health: a narrative review. *Sustainability*, 14(16):9951, 2022.

[6] Yansui Liu, Yang Zhou, and Jiaxin Lu. Exploring the relationship between air pollution and meteorological conditions in china under environmental governance. *Scientific Reports*, 10(1):14518, 2020.

[7] Yesol Cha, Jong-Jae Lee, Chul Han Song, Soontae Kim, Rokjin J Park, Myong-In Lee, Jung-Hun Woo, Jae-Ho Choi, Kangho Bae, Jinhyeok Yu, et al. Investigating uncertainties in air quality models used in gmap/sijaq 2021 field campaign: General performance of different models and ensemble results. *Atmospheric Environment*, 340:120896, 2025.

[8] H. Zhang, J. Wang, N. Janechek, C. Ge, M. Zhou, L. Castro García, T. Sha, Y. Wang, W. Deng, Z. Xue, C. Li, L. Chutia, Y. Wang, S. Val, J. L. McDuffie, S. Hasheminassab, S. E. Gluck, D. J. Diner, P. R. Colarco, A. M. da Silva, and J. Kim. Development of UI-WRF-Chem (v1.0) for the MAIA satellite mission: case demonstration. *Geoscientific Model Development*, 18(22):9061–9099, 2025.

[9] Ting Pan, Jiandong Wang, Dian Ding, Yuzhi Jin, Bin Wang, Zhouyang Zhang, Zhaoyang Peng, Bin Zhao, and Shuxiao Wang. Global PM2.5 exposure and attributable premature mortality in 2019: A modeling study using MPAS-CMAQ. *Atmospheric Environment*, 381:122130, 2026.

[10] N Ojha, A Sharma, M Kumar, et al. On the widespread enhancement in fine particulate matter across the indo-gangetic plain towards winter. *Scientific Reports*, 10(1):5862, 2020.

[11] Mengyuan Zhang, Shuai Wang, Wenxuan Yu, Chenliang Tao, Shengqiang Zhu, Jinlong Ma, Peng Wang, and Hongliang Zhang. The major role of anthropogenic emission underestimation in pm2. 5 estimation uncertainty over the tibetan plateau. *Geophysical Research Letters*, 52(9):e2024GL110513, 2025.

[12] K E Knowland, C A Keller, P A Wales, et al. Nasa geos composition forecast modeling system geos-cf v1. 0: Stratospheric composition. *Journal of Advances in Modeling Earth Systems*, 14(6):e2021MS002852, 2022.

[13] Ao Ding, Aoxing Zhang, Tzung-May Fu, Yuanlong Huang, Qianjie Chen, Yuyang Chen, Jiajia Mo, Wei Tao, Wai-Chi Cheng, Lei Zhu, et al. Diffusion-based probabilistic air quality forecasting with mechanistic insight. *arXiv preprint arXiv:2603.21131*, 2026.

[14] Ammar Kheder, Helmi Toropainen, Wenqing Peng, Samuel Antão, Jia Chen, Michael Boy, and Zhi-Song Liu. Topoflow: topography-aware pollutant flow learning for high-resolution air quality prediction. *npj Climate and Atmospheric Science*, 2026.

[15] Yuxuan Liang, Yutong Xia, Songyu Ke, Yiwei Wang, Qingsong Wen, Junbo Zhang, Yu Zheng, and Roger Zimmermann. Airformer: Predicting nationwide air quality in china with transformers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 14329–14337, 2023.

[16] Harna M Bodele, GM Asutkar, and Kiran G Asutkar. Advanced air quality forecasting using graph convolutional networks and spatio-temporal transformers. In *AIP Conference Proceedings*, volume 3327, page 020016. AIP Publishing LLC, 2025.

[17] Ke Gui, Xutao Zhang, Huizheng Che, Lei Li, Yu Zheng, Linchang An, Yucong Miao, Hujia Zhao, Oleg Dubovik, Brent Holben, et al. Advancing operational global aerosol forecasting with machine learning. *Nature*, pages 1–8, 2026.

[18] Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

[19] Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana CA Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, et al. Aifs–ecmwf's data-driven forecasting system. *arXiv preprint arXiv:2406.01465*, 2024.

[20] Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):190, 2023.

[21] Xiaohui Zhong, Lei Chen, Xu Fan, et al. Fuxi-2.0: Advancing machine learning weather forecasting model for practical applications. *arXiv preprint arXiv:2409.07188*, 2024.

[22] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

[23] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

[24] David C Wong, Jeff Willison, Jonathan E Pleim, Golam Sarwar, James Beidler, Russ Bullock, Jerold A Herwehe, Rob Gilliam, Daiwen Kang, Christian Hogrefe, et al. Development of the mpas-cmaq coupled system (v1. 0) for multiscale global air quality modeling. *Geoscientific Model Development Discussions*, 2024:1–18, 2024.

[25] Ruize Sun, Xiao Lu, Haipeng Lin, Tongwen Wu, Xingpei Ye, Lu Shen, Xuan Wang, Haolin Wang, Jingyu Li, Ni Lu, et al. Development of the global chemistry-climate coupled model bcc-geos-chem v2. 0: improved atmospheric chemistry performance and new capability of chemistry-climate interactions. *Geoscientific Model Development*, 19(5):2111–2136, 2026.

[26] A. Inness, M. Ades, A. Agustí-Panareda, J. Barré, A. Benedictow, A.-M. Blechschmidt, J. J. Dominguez, R. Engelen, H. Eskes, J. Flemming, V. Huijnen, L. Jones, Z. Kipling, S. Massart, M. Parrington, V.-H. Peuch, M. Razinger, S. Remy, M. Schulz, and M. Suttie. The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*, 19(6):3515–3556, 2019.

[27] Zhixin Geng, Xu Fan, Xiqiao Lu, Yan Zhang, Guangyuan Yu, Cheng Huang, Qian Wang, Yuewu Li, Weichun Ma, Qi Yu, et al. Fuxi-air: Urban air quality forecasting based on emission-meteorology-pollutant multimodal machine learning. *arXiv preprint arXiv:2506.07616*, 2025.

[28] Jindong Tian, Yuxuan Liang, Ronghui Xu, Peng Chen, Chenjuan Guo, Aoying Zhou, Lujia Pan, Zhongwen Rao, and Bin Yang. Air quality prediction with physics-guided dual neural odes in open systems. *arXiv preprint arXiv:2410.19892*, 2024.

[29] Y Abeysinghe, MA Munir, S Baliah, et al. Synergistic neural forecasting of air pollution with stochastic sampling. *arXiv preprint arXiv:2510.23977*, 2025.

[30] Vahid Khatibi and Parsa Nikpour. Advancing multi-pollutant air quality forecasting using transformer-based informer architecture. *Earth Science Informatics*, 18(3):287, 2025.

[31] Paula Harder, Johannes Flemming, Mihai Alexe, Gert Mertes, Baudouin Raoult, and Matthew Chantry. Aifs-compo: A global data-driven atmospheric composition forecasting system. *arXiv preprint arXiv:2604.03300*, 2026.

[32] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[33] Yuxuan Liang, Ronghui Xu, Peng Chen, Chenjuan Guo, Aoying Zhou, Lujia Pan, Zhongwen Rao, Bin Yang, et al. Air quality prediction with physics-guided dual neural odes in open systems. In *International Conference on Learning Representations*, volume 2025, pages 55132–55154, 2025.

[34] Xu Fan, Yuetan Lin, Bing Gong, and Hao Li. Offline meteorology-pollution coupling global air pollution forecasting model with bilinear pooling. *arXiv preprint arXiv:2503.18405*, 2025.

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[36] Suping Zhao, Ye Yu, Daiying Yin, Jianjun He, Na Liu, Jianjun Qu, and Jianhua Xiao. Annual and diurnal variations of gaseous and particulate pollutants in 31 provincial capital cities based on in situ air quality monitoring data from china national environmental monitoring center. *Environment international*, 86:92–106, 2016.

[37] Zeyu Yang, Zhanqing Li, Fan Cheng, Qiancheng Lv, Ke Li, Tao Zhang, Yuyu Zhou, Bin Zhao, Wenhao Xue, and Jing Wei. Two-decade surface ozone (o3) pollution in china: Enhanced fine-scale estimations and environmental health implications. *Remote Sensing of Environment*, 317:114459, 2025.

[38] Chi Li, Aaron van Donkelaar, Melanie S Hammer, Erin E McDuffie, Richard T Burnett, Joseph V Spadaro, Deepangsu Chatterjee, Aaron J Cohen, Joshua S Apte, Veronica A Southerland, et al. Reversal of trends in global fine particulate matter air pollution. *Nature communications*, 14(1):5349, 2023.

[39] Jun Rentschler and Nadezda Leonova. Global air pollution exposure and poverty. *Nature communications*, 14(1):4432, 2023.

[40] Chenliang Tao, Yiheng Wang, Yuhao Wang, Zhonghua Zheng, and Hongliang Zhang. Deep learning for air pollutant forecasting: opportunities, challenges, and future directions. *Frontiers of Environmental Science & Engineering*, 19(12):172, 2025.

[41] Qi Liao, Mingming Zhu, Lin Wu, Xiaole Pan, Xiao Tang, and Zifa Wang. Deep learning for air quality forecasts: a review. *Current Pollution Reports*, 6(4):399–409, 2020.

[42] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[43] Dexiang Wang, Jianlin Hu, Yong Xu, Di Lv, Xiaoyang Xie, Michael Kleeman, Jia Xing, Hongliang Zhang, and Qi Ying. Source contributions to primary and secondary inorganic particulate matter during a severe wintertime pm2. 5 pollution episode in xi'an, china. *Atmospheric environment*, 97:182–194, 2014.

Appendix

A Dataset

A.1 Air pollutant concentrations

The pollutant data used in this study comprise two CAMS products: the CAMS global reanalysis (EAC4; hereafter RA) [26] [1] and the CAMS global atmospheric composition forecasts (hereafter FC) [2]. The RA dataset provides a long-term, temporally continuous record for model pre-training, while the FC dataset is used for model adaptation and evaluation.

A.2 ERA5

This study uses the European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) reanalysis dataset [42] provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). To ensure consistency with the FuXi meteorological forecast data, the precipitation variable is processed as a 12-hour accumulated precipitation field during preprocessing.

A.3 FuXi-2.0

The FuXi-2.0 model provides global meteorological forecasts at a $0.25^\circ \times 1$-h resolution [21]. The dataset encompasses 5 upper-air variables (geopotential height, temperature, humidity, and u/v wind components) across 13 standard pressure levels (from 1000 to 50 hPa), alongside 23 surface variables covering critical atmospheric and oceanic conditions (e.g., 2-m temperature, sea level pressure, and wave parameters). For operational consistency and real-time efficiency, this study utilizes the 6-hour forecast version of FuXi-2.0. During inference, these outputs are resampled to align with the spatiotemporal resolution of the CAMS and ERA5 dataset.

A.4 Emission inventories

We adopt weekday, weekend, and hourly allocation factors from Wang et al. [43], following a temporal disaggregation approach similar to CAMS-GLOB-TEMPO [3]. The allocation is performed in local time at each grid cell, with separate normalization at the daily and hourly levels to preserve monthly emission totals.

Let $E_{g,s,m}$ denote the monthly emission total for grid cell g, emission species s, and month m after conservative remapping. For each local calendar day d in month m, we assign a factor a_d according to whether it is a weekday or a weekend. The daily emission total is calculated as

$$E_{g,s,d} = E_{g,s,m} \frac{a_d}{\sum_{d' \in \mathcal{D}_m} a_{d'}}, \tag{12}$$

where $\mathcal{D}_m$ is the set of local calendar days in month m. This normalization accounts for the actual numbers of weekdays and weekend days in each month.

The hourly allocation profile is constructed by combining source-specific activity profiles:

$$b_{g,s,d,h} = \sum_{k \in \mathcal{K}} w_{g,s,m,k}\, p_{k,d,h}, \tag{13}$$

where $\mathcal{K}$ denotes the source categories, $w_{g,s,k}$ is the contribution weight of source k, and $p_{k,d,h}$ is its activity factor at local hour $h \in \{0, \ldots, 23\}$ on day d. Hourly emissions are then obtained as

$$E_{g,s,d,h} = E_{g,s,d} \frac{b_{g,s,d,h}}{\sum_{h'=0}^{23} b_{g,s,d,h'}}. \tag{14}$$

Consequently,

$$\sum_{d \in \mathcal{D}_m} \sum_{h=0}^{23} E_{g,s,d,h} = E_{g,s,m}. \tag{15}$$

[1] https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4

[2] https://ads.atmosphere.copernicus.eu/datasets/cams-global-atmospheric-composition-forecasts

[3] https://ads.atmosphere.copernicus.eu/datasets/cams-global-emission-inventories

Local dates and hours are determined from the UTC timestamp using the time-zone offset assigned to each grid cell:

$$t_{\text{local}}(g) = t_{\text{UTC}} + \Delta_g, \tag{16}$$

where Δ_g is the local UTC offset in hours. The local calendar date determines the weekday or weekend factor, while the local hour determines the hourly activity factor. After temporal allocation, the hourly emissions are indexed in UTC and accumulated over the 12-hour interval ending at each model input time.

A.5 Data preprocess

To address unit inconsistencies among variables in the original CAMS outputs and enhance the physical interpretability of the model inputs, all variables were converted into commonly used units in atmospheric modeling. Specifically, particulate matter concentrations (PM_1, $PM_{2.5}$, and PM_{10}) were converted to $\mu g/m^3$, while total column concentrations were expressed in mg/m^2. For gas concentrations at the 1000 hPa level, mass mixing ratios (kg/kg) were converted to volume mixing ratios ($ppbv$) using the ideal gas law, as formulated in Eq. 17.

$$C_{\text{ppbv}} = C_{\text{kg/kg}} \times \frac{M_{\text{air}}}{M_{\text{gas}}} \times 10^9 \tag{17}$$

where C_{ppbv} denotes the volume mixing ratio, $C_{\text{kg/kg}}$ represents the mass mixing ratio, and M_{air} (28.97) and M_{gas} denote the mean molecular weights of air and the corresponding gas species, respectively.

To balance the training data and incorporate multi-source information, a total of 78 variables, including emissions, meteorological variables, and pollutant concentrations, were used as model inputs. All datasets were resampled to match the spatial resolution of the CAMS forecast data (0.4°). Due to the highly skewed distribution of emission variables, a log transformation was applied before model input to stabilize the data distribution:

$$\log1\text{p}(x) = \ln(1 + x) \tag{18}$$

Subsequently, all variables were standardized using z-score normalization. In addition, four temporal features—day of year and hour of day encoded using sinusoidal functions—were introduced and broadcast across the spatial domain:

$$h_{\sin} = \sin\left(\frac{2\pi h}{24}\right), \quad h_{\cos} = \cos\left(\frac{2\pi h}{24}\right) \tag{19}$$

$$d_{\sin} = \sin\left(\frac{2\pi d}{T}\right), \quad d_{\cos} = \cos\left(\frac{2\pi d}{T}\right) \tag{20}$$

where h denotes the hour of the day, d denotes the day of the year, and T represents the total number of days in a year.

A.6 Dataset partitioning

Considering the data quality and temporal stability of CAMS forecast products, and following the evaluation protocol of Aurora [18], we adopt a three-stage training strategy. Since the CAMS FC dataset provides a relatively limited number of samples and has only offered stable outputs for particulate matter, total column concentrations, and near-surface gas concentrations since October 2017, we first pre-train the model on the CAMS RA dataset from January 2010 to December 2021. This stage enables the model to learn the large-scale spatiotemporal distribution and evolution patterns of atmospheric pollutants under the influence of meteorological conditions and emission fields. The pre-trained model is then fine-tuned on the CAMS FC dataset from October 2017 to May 2022, so that the model output distribution can better align with the FC products used for downstream evaluation and comparison.

Table 4: Model Input Data

Data Type	Source	Variables	Spatial Res.	Temporal Res.	Period
Pollutants	CAMS-RA	pm1, pm2p5, pm10, tcco, tcno2, tc_no, gtco3, tcso2, co_1000, no_1000, no2_1000, so2_1000, go3_1000	$0.75^\circ \rightarrow 0.4^\circ$	12 h	2010.01–2021.12
	CAMS-FC	Same as above	0.4°	12 h	2017.10–2022.05
Meteorology	ERA5	U10m, v10m, d2m, t2m, msl, tcc, tp; z, u, v, t (at 13 levels: 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000 hPa)	$0.25^\circ \rightarrow 0.4^\circ$	12 h	2010.01–2023.12
	FuXi	U10m, v10m, d2m, t2m, msl, tcc, tp; z, u, v, t (at 13 levels)	$0.25^\circ \rightarrow 0.4^\circ$	12 h	2022.06–2022.11
Emissions	CAMS emission	ammonia, black carbon, carbon monoxide, nitrogen oxides, NMVOCs, sulfur dioxide	$0.1^\circ \rightarrow 0.4^\circ$	12 h	2010.01–2023.12

For model development, data from January 2023 to December 2023 are used as the validation set, while an independent test period from June 2022 to November 2022 is adopted for final evaluation (Table 5). Early stopping with a patience of 20 epochs is applied in Stages 2 and 3 based on the validation RMSE.

Table 5: Dataset partitioning and time ranges for different training stages.

Stage / Split	Dataset	Purpose	Period
Stage 1	CAMS-RA	Global pre-training	2010.01–2021.12
Stage 2	CAMS-FC	Operational fine-tuning	2017.10–2022.05
Test	CAMS-FC	Final evaluation	2022.06–2022.11

B Experimental Details

B.1 Loss Function

This appendix provides the detailed definitions of the two loss terms used in AirMPA, namely the latitude-weighted Charbonnier loss and the structural similarity (SSIM) loss.

B.1.1 Latitude-weighted Charbonnier loss

To account for the non-uniform area represented by the global latitude–longitude grid, we use a latitude-weighted Charbonnier loss:

$$\mathcal{L}_{\text{Lat-weighted Charbonnier}} = \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\cos(\phi_i)}{\sum_{i'=1}^{H} \cos(\phi_{i'})} \sqrt{(\hat{y}_{i,j} - y_{i,j})^2 + \epsilon^2}, \tag{21}$$

where $\hat{y}_{i,j}$ and $y_{i,j}$ denote the predicted and reference pollutant values at grid cell (i, j), respectively, ϕ_i is the latitude of the i-th row, H and W are the numbers of latitude and longitude grid cells, and ϵ is a small constant for numerical stability.

B.1.2 Structural similarity loss

To preserve the spatial structure of pollutant fields, we further introduce the SSIM loss. The structural similarity index is defined as

$$\text{SSIM}(\hat{y}, y) = \frac{(2\mu_{\hat{y}}\mu_y + C_1)(2\sigma_{\hat{y}y} + C_2)}{(\mu_{\hat{y}}^2 + \mu_y^2 + C_1)(\sigma_{\hat{y}}^2 + \sigma_y^2 + C_2)}, \tag{22}$$

where $\mu_{\hat{y}}$ and μ_y are the mean values of the predicted and reference fields, $\sigma^2_{\hat{y}}$ and σ^2_y are the corresponding variances, $\sigma_{\hat{y}y}$ is the covariance, and C_1 and C_2 are constants used to stabilize the computation.

The corresponding SSIM loss is defined as

$$\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(\hat{y}, y). \tag{23}$$

B.2 Training stages

AirMPA is optimized using a three-stage progressive training strategy, including global pre-training, operational fine-tuning, and extreme-event optimization. The detailed settings of each stage are summarized in Table 6. In the third stage, the dataset is further partitioned using pollutant-specific extreme-value thresholds, and higher loss weights are assigned to these critical samples to improve the reconstruction of high-concentration events. The manually specified extreme thresholds used for Stage 3 training and the 90th percentile thresholds used for test-period evaluation are listed in Table 7.

Table 6: Training objectives and optimization settings for the three training stages of AirMPA.

Stage	**Objective**	**Learning rate**	**Loss settings**
Stage 1	Global pre-training on CAMS reanalysis	$2 \times 10^{-4} \rightarrow 10^{-6}$	Hybrid loss: $\alpha = 1.0$, $\beta = 0.25$, $\epsilon = 10^{-3}$
Stage 2	Operational fine-tuning on CAMS forecast data	$5 \times 10^{-5} \rightarrow 10^{-6}$	Hybrid loss: $\alpha = 1.0$, $\beta = 0.75$, $\epsilon = 10^{-3}$
Stage 3	Extreme-event optimization	continued from Stage 2	Weighted RMSE loss: overall RMSE $= 0.5$, extreme-event RMSE $= 0.5$

Table 7: Extreme thresholds and 90th percentile values for different variables.

Variable	**PM_1**	**$PM_{2.5}$**	**PM_{10}**	**TC CO**	**TC NO**	**TC NO_2**	**TC SO_2**	**TC O_3**
Training extreme threshold	8.0	25.0	50.0	900	1.5	3.5	3.0	7500
Threshold (90%)	5.2323	20.8496	42.1983	811.4433	1.4372	3.2538	1.7426	7227.6772
Variable (at 1000 hpa)	**CO**	**NO**	**NO_2**	**SO_2**	**O_3**			
Training extreme threshold	135	0.05	0.8	0.5	45			
Threshold (90%)	110.1245	0.0265	0.4042	0.2312	38.8720			

B.3 Ablation setup

Table 8: Ablation study of key architectural components in AirMPA.

Model	**Zonal Release Attention**	**Cube Embedding**	**Dual-Branch Fusion**	**Emission Input**
Swin-Transformer (backbone)	×	×	×	×
w/ Emission	×	×	×	✓
w/ Zonal Release Attention	✓	×	×	✓
w/ Cube Embedding	×	✓	×	✓
w/ Dual-Branch Fusion	×	×	✓	✓
AirMPA	✓	✓	✓	✓

B.4 Compute resources

Training AirMPA required approximately 72 GPU-hours on one NVIDIA A100 GPU with 80 GB memory. Each ablation variant required approximately 72 GPU-hours. Inference over the six-month test period required approximately 1.5 GPU-hours. The processed datasets occupied approximately 2 TB of storage. Preliminary experiments are not included in these estimates.

C Limitations

Despite the strong performance, several limitations remain and point to directions for future work.

First, the model is trained on CAMS atmospheric composition reanalysis and forecast datasets, where initial conditions are derived from a combination of previous forecasts and satellite observations. The limited use of surface monitoring data may reduce consistency with real-world observations. Incorporating additional observational constraints and physical conservation laws could further improve short-term stability and physical realism.

Second, real-world emissions are highly dynamic and influenced by factors such as holidays, industrial activities, and traffic regulations. In the current framework, emission inputs are primarily based on hourly empirical factors, which are insufficient to capture rapid or abrupt emission changes. This may introduce systematic biases under scenarios with sudden emission variations, such as large-scale events or emergency control measures.

Third, due to the resolution of the training data, the model operates at a temporal resolution of 12 hours and a spatial resolution of 0.4°. Such resolution may be inadequate for resolving localized pollution events at finer regional scales. In addition, under the current 12-hour forecasting interval, the dual-branch fusion module only learns to adaptively weight the two preceding pollutant concentration fields, which may not fully exploit its potential for modeling more detailed temporal dependencies. Future work should extend the framework to finer temporal resolutions to further evaluate and enhance the effectiveness of this component.

D Broader impacts

The development of AirMPA provides a new technical paradigm for atmospheric environmental forecasting, with important implications for both scientific research and operational air quality management. By decoupling meteorological forecasting from pollutant prediction, our framework shows that high-quality outputs from meteorological foundation models can be effectively transferred to downstream air quality forecasting tasks. This modular design improves forecasting flexibility and accuracy, and provides a scalable technical basis for future AI-driven atmospheric composition prediction systems.

From a public health perspective, atmospheric pollution remains one of the most pervasive global environmental risks. The ability of AirMPA to provide more accurate global forecasts, especially for extreme pollution events, can support earlier warnings and more timely interventions. This may help reduce human exposure to harmful pollutants and improve decision-making for environmental protection and public health management.

More broadly, the significance of this work lies in improving the predictive capability of global air quality forecasting and in providing a technical foundation for future full-chain forecasting systems. In particular, the proposed framework offers a feasible pathway toward an integrated system that combines data assimilation, ensemble meteorological prediction, and downstream pollutant forecasting. Such a system could further enhance the timeliness, robustness, and operational value of air pollution early warning.

At the same time, several limitations should be noted. Since the current model is trained primarily on reanalysis and forecast products, it may inherit systematic biases between these data sources and real-world observations, which could reduce its effectiveness in practical applications. In addition, like other data-driven forecasting models, AirMPA still requires further validation under unprecedented extreme events or rapidly changing emission conditions. Future work should strengthen the integration of ground-based observations, uncertainty quantification, and physical consistency constraints, so as to improve reliability and interpretability in real-world decision-making.

E Code and data availability

The ERA5 and CAMS products used in this study were obtained from the Copernicus Climate Data Store (CDS; https://cds.climate.copernicus.eu/) and the Copernicus Atmosphere Data Store (ADS; https://ads.atmosphere.copernicus.eu/datasets), respectively, under the Copernicus CC-BY licence. In accordance with the Copernicus terms, the derived products

contain modified Copernicus Climate Change Service and Copernicus Atmosphere Monitoring Service information, and neither the European Commission nor ECMWF is responsible for any use made of the information or data contained herein. The ERA5 reanalysis data were downloaded, subsetted, and preprocessed for use in the AirMPA model. The CAMS reanalysis and forecast data, together with anthropogenic emission inventories, were used for model comparison and inference experiments.

The FuXi-2.0 forecast fields were generated from operational FuXi-2.0 outputs and used for academic research with authorization from the model developers. The corresponding source code and inference scripts are available from the official repository (https://github.com/tpys/FuXi2/), and public FuXi assets were used in accordance with their stated licence terms. The Aurora code, checkpoints, and inference scripts were obtained from the official Microsoft example (https://microsoft.github.io/aurora/example_cams.html) and used under the MIT license. For fair comparison, the same CAMS dataset covering June 2022 to November 2022 was used for Aurora inference experiments.

The AirMPA implementation is publicly available at https://github.com/yhWang0712/AirMPA.

F Additional results

F.1 Comparison with the offline model of Fan et al.

Table 9: Relative RMSE reduction (%) of AirMPA compared with the offline meteorology–pollution model of Fan et al. [34] over June–November 2022. Positive values (in bold) indicate lower RMSE for AirMPA, while negative values indicate degradation. The comparison covers eight shared pollutant variables at lead times of 1, 3, and 5 days.

Variable	Day 1	Day 3	Day 5
PM_1	**19.36**	**13.88**	**11.79**
$PM_{2.5}$	**21.89**	**15.17**	**8.10**
PM_{10}	**33.46**	**30.39**	**24.90**
TC CO	**15.38**	**13.58**	**18.14**
TC NO_2	**20.18**	**13.84**	**8.10**
TC NO	**29.33**	**24.76**	**18.13**
TC O_3	**25.79**	**8.36**	-3.88
TC SO_2	**20.16**	**7.91**	**0.95**

F.2 Pollutant-specific sensitivity to meteorological forcing

We compare ERA5, FuXi, and ECMWF HRES as meteorological inputs to the same AirMPA pollutant model, without modifying or retraining its parameters. Table 10 reports results for all 13 pollutant variables at lead times of 1, 3, and 5 days. For each variable and lead time, we compute the relative RMSE reduction with respect to FuXi forcing as

$$R_X = \frac{\mathrm{RMSE}_{\mathrm{FuXi}} - \mathrm{RMSE}_X}{\mathrm{RMSE}_{\mathrm{FuXi}}} \times 100\%, \qquad X \in \{\mathrm{ERA5}, \mathrm{HRES}\}. \tag{24}$$

Positive values indicate lower RMSE than with FuXi. We average these percentage reductions across variables, rather than averaging absolute RMSE values with different units.

ERA5 versus FuXi. ERA5 yields lower RMSE than FuXi in 37 of the 39 pollutant–lead-time cases. The mean reduction increases from 1.19% at day 1 to 8.42% at day 3 and 15.95% at day 5. The two exceptions are NO and SO_2 at 1000 hPa at day 1, for which RMSE increases by 5.80% and 4.64%, respectively. At day 5, the largest reductions occur for total-column O_3 (26.62%) and O_3 at 1000 hPa (21.44%). The reductions for PM_1, $PM_{2.5}$, and PM_{10} are 6.94%, 16.12%, and 19.66%, respectively. Total-column CO, NO_2, and NO improve by 17.06%, 19.92%, and 17.69%, while their 1000-hPa counterparts improve by 11.28%, 16.31%, and 8.12%. Total-column SO_2 and SO_2 at 1000 hPa improve by 13.39% and 12.77%, respectively.

HRES versus FuXi. HRES produces smaller mean changes relative to FuXi, reducing RMSE by 0.31%, 1.45%, and 2.35% at days 1, 3, and 5, respectively. It yields lower RMSE in 32 of the 39 cases. The largest reduction is for total-column SO_2, reaching 17.80% at day 5. For this variable, HRES also yields lower RMSE than ERA5 at all three reported lead times. In contrast, total-column CO and O_3 have slightly higher RMSE with HRES than with FuXi at all three lead times, with increases of 0.25–1.57% and 0.32–1.02%, respectively. For the remaining variables, the day-5 reductions range from 0.24% for CO at 1000 hPa to 4.46% for total-column NO.

These results show that the effect of changing meteorological forcing varies across pollutants and forecast horizons. The widening gap between ERA5 and FuXi at longer lead times suggests increasing sensitivity to the meteorological input during autoregressive forecasting. ERA5 provides a retrospective reanalysis reference, so its advantage should not be interpreted as an operationally available forecast gain.

Table 10: Pollutant-specific comparison of meteorological forcing at lead times of 1, 3, and 5 days. Values are relative RMSE reductions (%) compared with FuXi forcing. Positive values (in bold) indicate lower RMSE, while negative values indicate higher RMSE. TC denotes total column. The mean is the unweighted average across the 13 variables. ERA5 provides an offline reanalysis reference.

	Day 1		Day 3		Day 5	
Variable	ERA5	HRES	ERA5	HRES	ERA5	HRES
PM_1	**1.53**	-0.08	**3.79**	**0.17**	**6.94**	**0.38**
$PM_{2.5}$	**2.46**	**0.13**	**8.29**	**0.48**	**16.12**	**0.82**
PM_{10}	**2.99**	**0.32**	**10.44**	**0.48**	**19.66**	**1.42**
TC CO	**2.15**	-0.25	**9.22**	-0.53	**17.06**	-1.57
TC NO_2	**3.40**	**0.16**	**11.52**	**1.90**	**19.92**	**3.56**
TC NO	**0.30**	**0.35**	**9.76**	**1.92**	**17.69**	**4.46**
TC O_3	**4.00**	-0.32	**15.18**	-0.73	**26.62**	-1.02
TC SO_2	**1.47**	**2.48**	**7.82**	**11.45**	**13.39**	**17.80**
CO (1000 hPa)	**1.92**	**0.17**	**6.06**	**0.26**	**11.28**	**0.24**
NO (1000 hPa)	-5.80	**0.64**	**1.13**	**1.26**	**8.12**	**1.60**
NO_2 (1000 hPa)	**2.63**	**0.22**	**8.96**	**0.87**	**16.31**	**1.33**
SO_2 (1000 hPa)	-4.64	**0.07**	**4.61**	**0.73**	**12.77**	**1.25**
O_3 (1000 hPa)	**3.10**	**0.09**	**12.66**	**0.62**	**21.44**	**0.30**
Mean	**1.19**	**0.31**	**8.42**	**1.45**	**15.95**	**2.35**

F.3 Additional figures

To ensure a fair comparison, AirMPA, Aurora, and CAMS are evaluated from the same pollutant initialization states. Specifically, the pollutant composition fields at $t_0 - 12\,\mathrm{h}$ and t_0 are taken from CAMS FC for all methods. After initialization, the models differ in their forecasting mechanisms and meteorological inputs: AirMPA uses FuXi-2.0 forecast fields as external forcing, Aurora follows its original coupled forecasting framework, and CAMS serves as the numerical baseline. This setting is consistent with operational air-quality forecasting, where composition initial fields are typically provided by an analysis or reanalysis product and then advanced forward by the forecasting system.

Figure 5 presents a comprehensive performance evaluation across three modeling frameworks: AirMPA (driven by FuXi-2.0 forecasts), CAMS, and Aurora. Results are categorized by global and terrestrial domains, providing a rigorous assessment of model skill under both historical reanalysis and real-time operational forecasting scenarios.

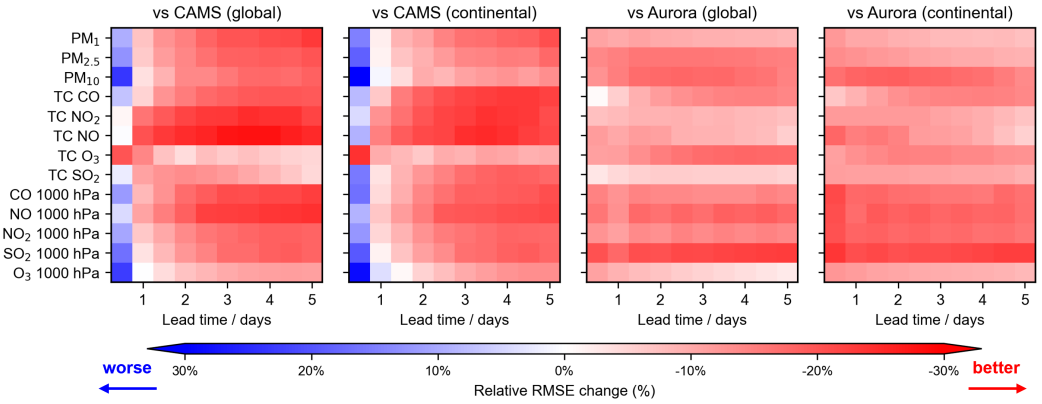

Figure 5: Comparison of forecasting performance across AirMPA (with FuXi), CAMS, and Aurora.

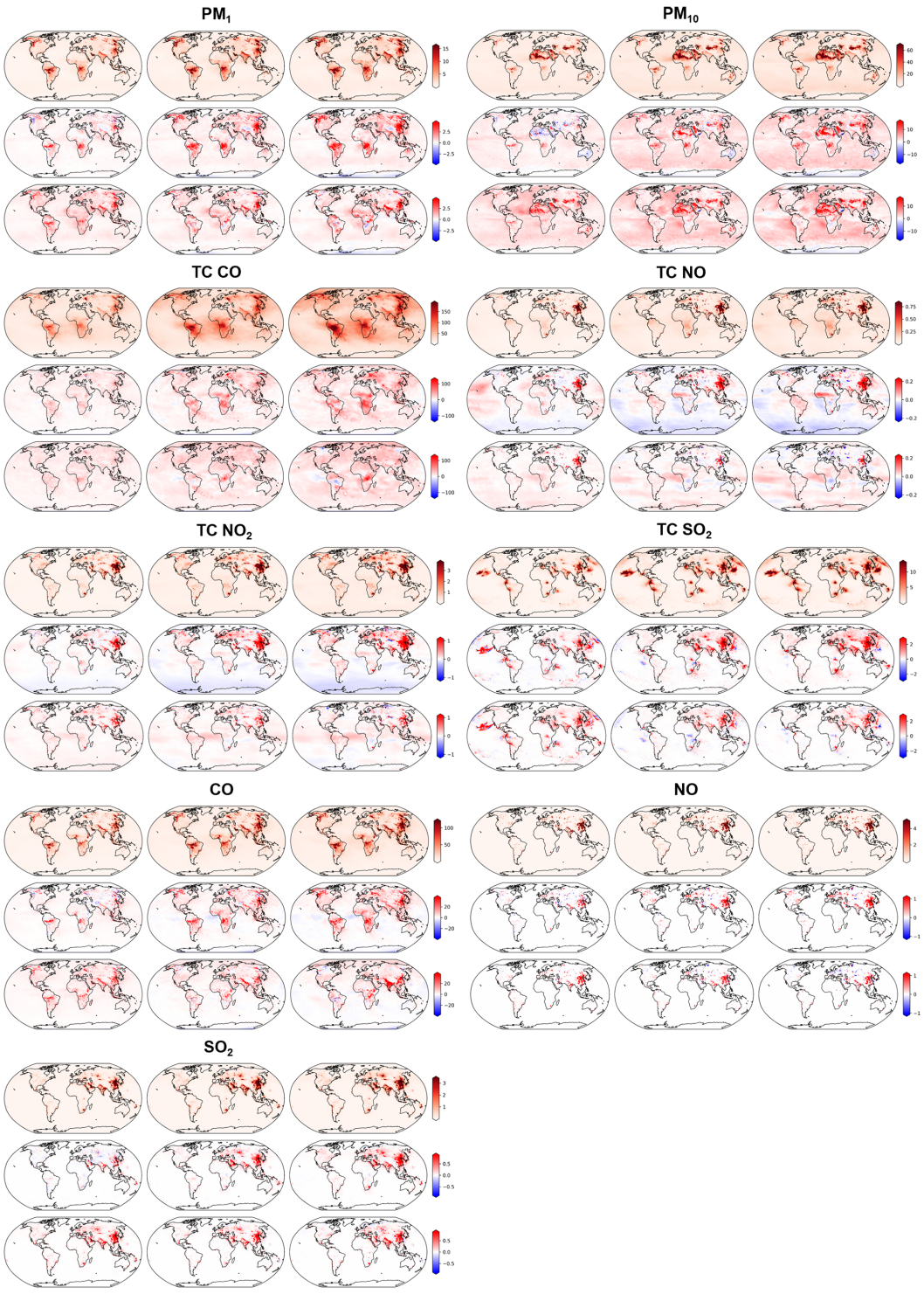

Figure 6: Same as Figure 3, but for other pollutants. The three columns in each subplot correspond to lead times of 1, 3, and 5 days, while the rows correspond to AirMPA, CAMS – AirMPA, and Aurora – AirMPA, respectively.

Figures 7, 8, and 9 illustrate the comparative predictive skill of AirMPA (driven by FuXi-2.0 forecasts), CAMS, and Aurora specifically for extreme pollution events.

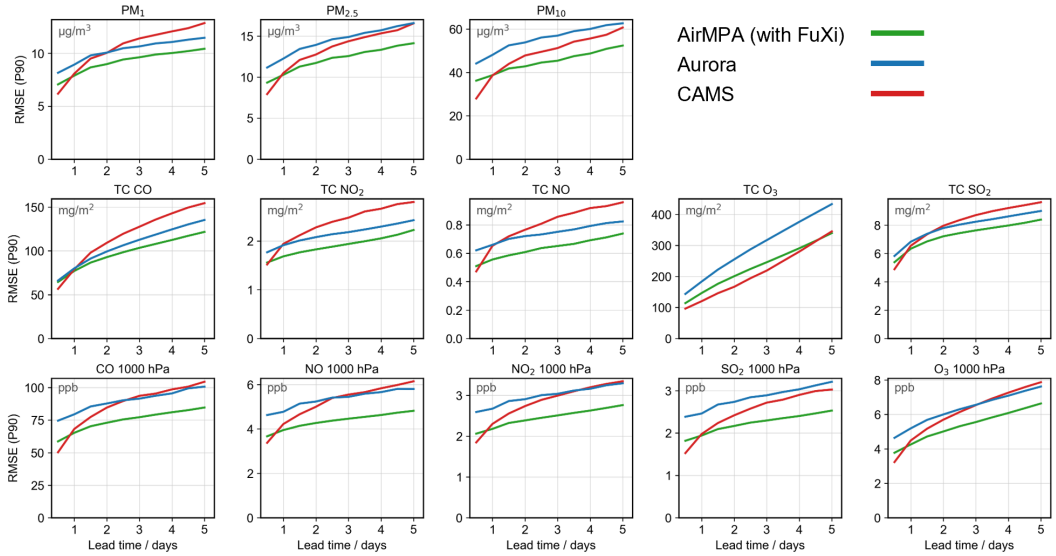

Figure 7: Comparison of extreme forecasting performance(90% RMSE during June 2022 – November 2022) across AirMPA (with FuXi), CAMS, and Aurora.

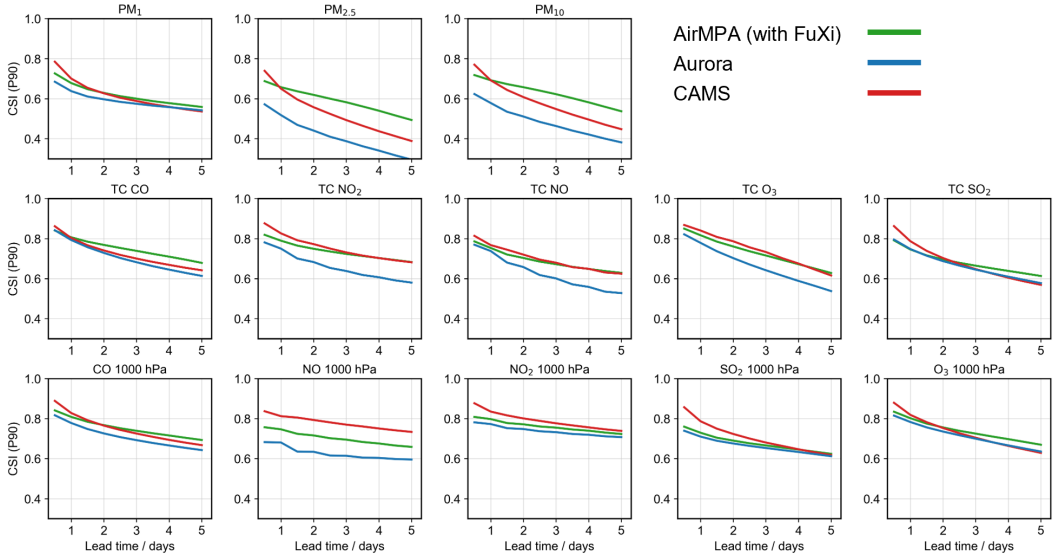

Figure 8: Comparison of extreme forecasting performance(90% CSI during June 2022 – November 2022) across AirMPA (with FuXi), CAMS, and Aurora.

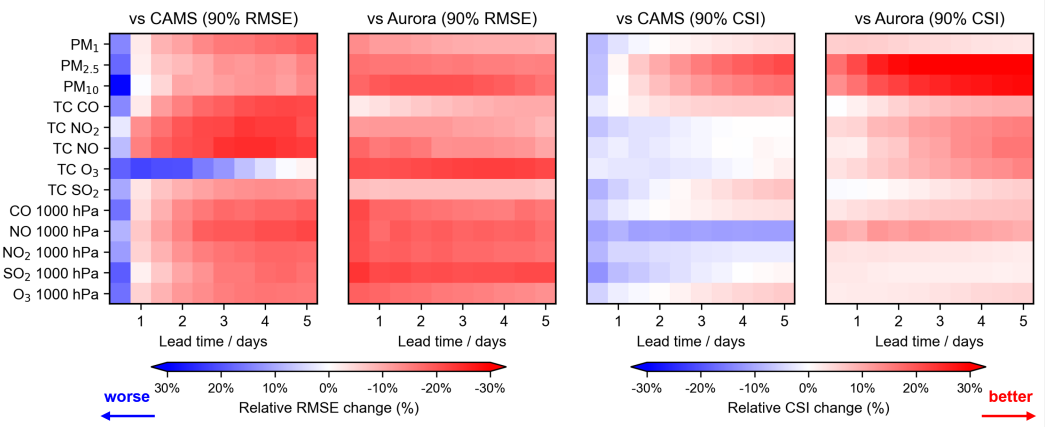

Figure 9: Comparison of the AirMPA model in extreme value forecasting metrics relative to CAMS and Aurora.

Figure 10 presents a performance comparison of three models—AirMPA, Aurora, and CAMS—at air quality monitoring stations operated by the U.S. Environmental Protection Agency (EPA)[4], the European Environment Agency (EEA)[5], and the China National Environmental Monitoring Centre (CNEMC)[6]. The spatial distribution of these stations is shown in Figure 11.

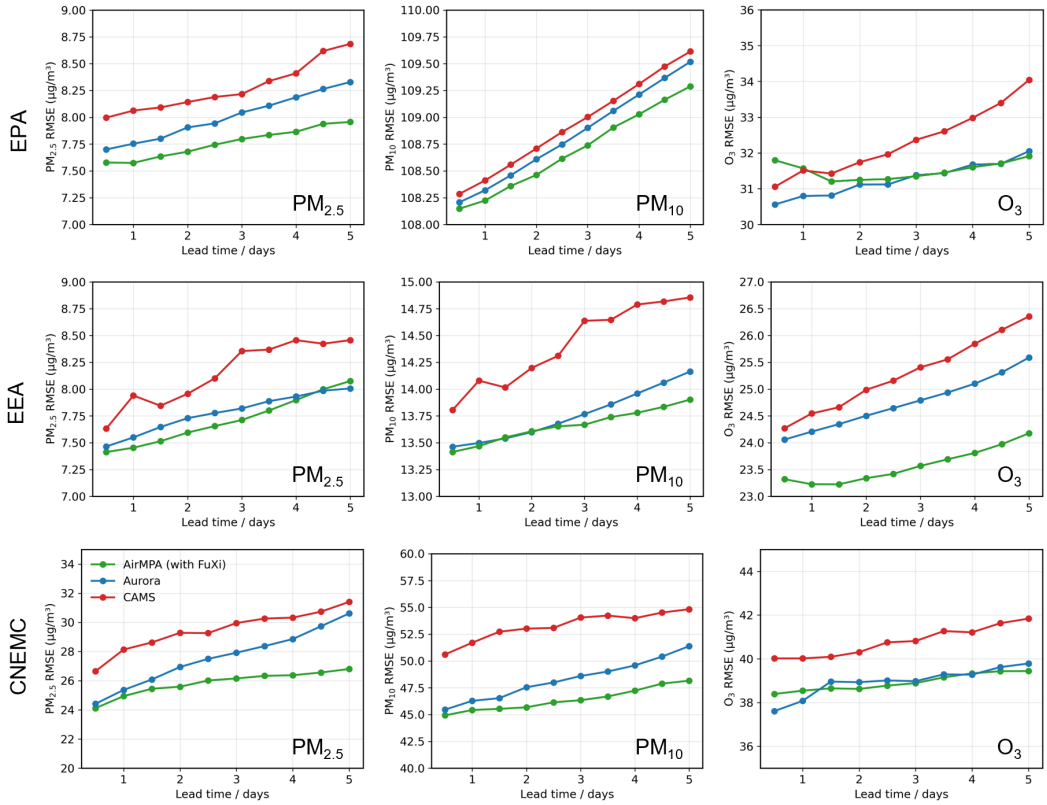

Figure 10: Comparison of forecasting performance of AirMPA, Aurora, and CAMS at monitoring stations in the United States (EPA), Europe (EEA) and China (CNEMC).

[4] https://aqs.epa.gov/aqsweb/airdata/download_files.html
[5] https://www.eea.europa.eu/en/datahub/datahubitem-view/778ef9f5-6293-4846-badd-56a29c70880d
[6] https://www.cnemc.cn/

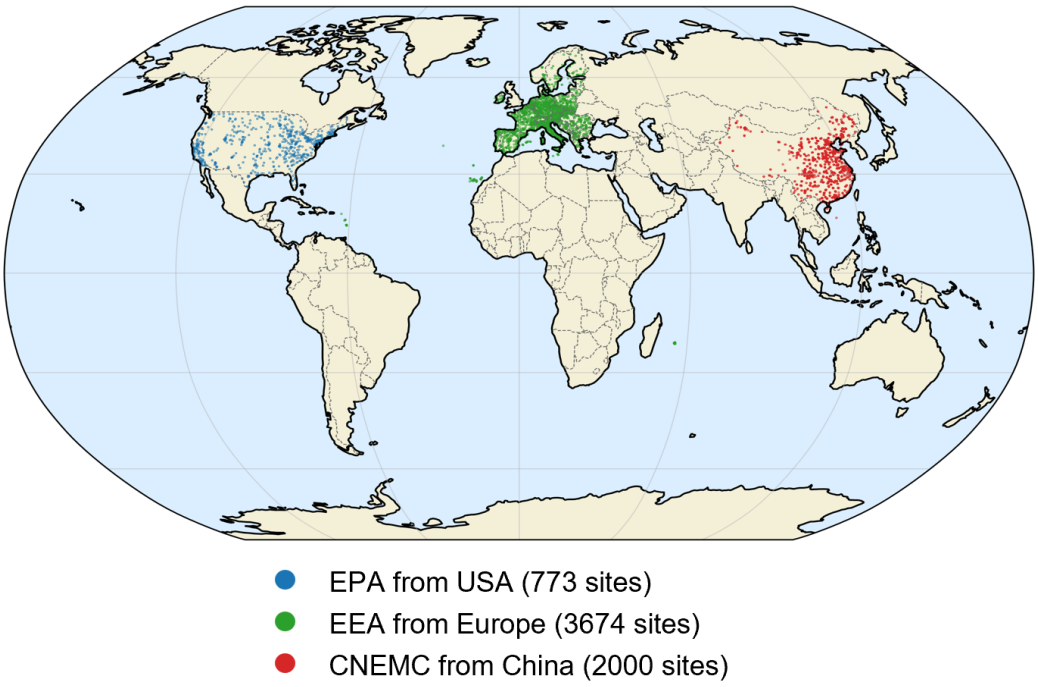

Figure 11: Distribution of monitoring stations in the United States (EPA), Europe (EEA) and China (CNEMC).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (section 1) have clearly outlined our primary contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a comprehensive discussion of the model's limitations in Appendix C

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: this paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide a comprehensive description of the model architecture and experimental pipeline in Sections 3 and 4, respectively. Furthermore, Appendix A explicitly details the data availability, including preprocessing steps and source links, ensuring that all information necessary to reproduce our main findings is fully disclosed.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The AirMPA code is publicly available at https://github.com/yhWang0712/AirMPA. Because the full source datasets are too large to redistribute with the repository, we provide a sample data subset for running a 3-day forecast. Appendix E lists the original data sources and access conditions. The sample supports an inference demonstration rather than reproduction of the full evaluation.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: As shown in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report multi-seed error bars because training global forecasting models at this scale is computationally expensive. Instead, we evaluate all methods on a continuous six-month independent test period across multiple lead times, pollutants, and spatial domains.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specify this part in Section 4.2.3 and Appendix B.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts of the paper in Section D.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: NA

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We explicitly cite and credit all third-party datasets, models, and code used in this work. ERA5 and the CAMS products used in this study, including CAMS global reanalysis, CAMS global atmospheric composition forecasts, and CAMS global emission inventories, are distributed through the Copernicus Climate Data Store and Atmosphere Data Store under the CC-BY licence, and we comply with the required Copernicus attribution, license, and disclaimer terms. FuXi-2.0 forecast fields are used for academic research with authorization from the model developers; where public FuXi assets are involved, the corresponding code and model-weight terms are respected, including the Apache-2.0 licence for the public inference/plugin code and the CC BY-NC-SA 4.0 non-commercial terms for

released FuXi model weights. Aurora is used only as a research benchmark; its source code and released checkpoints are credited to Microsoft and are distributed under the MIT license. All official repositories, data stores, and license or terms-of-use information are listed in the Data Availability section, and no third-party asset is redistributed beyond the scope permitted by its license or authorization.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release the AirMPA implementation together with a sample data subset and instructions for running a 3-day forecast. The example illustrates the required input format and inference workflow without requiring users to download the full source datasets. The model's scope and limitations are discussed in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This work does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human-subject experiments or personally identifiable data, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.