ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Global and regional performance of the aurora foundation model for surface $PM_{2.5}$ prediction

To cite this article: Delin Li *et al* 2026 *Environ. Res. Lett.* **21** 154018

View the article online for updates and enhancements.

ENVIRONMENTAL RESEARCH
LETTERS

LETTER

OPEN ACCESS

RECEIVED
11 April 2026

REVISED
22 June 2026

ACCEPTED FOR PUBLICATION
27 July 2026

PUBLISHED
4 August 2026

Global and regional performance of the aurora foundation model for surface $PM_{2.5}$ prediction

Delin Li[1,2], Zhiqiang Chen[1,2,*], Gang Huang[3,4] and Jianjun Xu[1,2]

1 College of Ocean and Meteorology, Guangdong Ocean University, Zhanjiang, People's Republic of China
2 Shenzhen Institute of Guangdong Ocean University, Shenzhen, People's Republic of China
3 State Key Laboratory of Earth System Numerical Modeling and Application, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, People's Republic of China
4 University of Chinese Academy of Sciences, Beijing, People's Republic of China
* Author to whom any correspondence should be addressed.

E-mail: czqiang@gdou.edu.cn

Keywords: air quality forecasting, Aurora foundation model, surface $PM_{2.5}$, transfer learning, machine learning

Supplementary material for this article is available online

Abstract

Accurate global forecasts of surface fine particulate matter ($PM_{2.5}$) are important for environmental early warning, but operational atmospheric composition models remain computationally expensive. Although deep learning foundation models offer a potentially efficient alternative, their performance for $PM_{2.5}$ prediction remains insufficiently explored. Here, we evaluate Aurora, an Earth system foundation model fine-tuned for air quality forecasting, through year-long retrospective forecasts of global surface $PM_{2.5}$ for 2025. We assess its performance using Copernicus Atmosphere Monitoring Service (CAMS) analysis and forecast products, independent surface observations from 3540 monitoring stations compiled from OpenAQ and the China National Environmental Monitoring Centre, and the Global Earth Observing System Composition Forecast analysis. We find that Aurora is highly competitive: it reproduces the broad annual evolution of global surface $PM_{2.5}$ and agrees closely with the CAMS analysis. Relative to the operational CAMS forecast, Aurora shows smaller mean bias during the first 72 h and a slower increase in error with lead time. Against independent observations, Aurora captures regional and seasonal variability across North America, Europe, East Asia, and India. Over China, Aurora achieves skill comparable to, or locally even exceeding all the numerical products, exhibits low station-wise bias, and better represents the evolution of a persistent winter haze episode. These results show that transfer learning can adapt a weather foundation model to global air quality forecasting with competitive skill and high computational efficiency, supporting rapid early warning and future ensemble applications.

1. Introduction

Fine particulate matter ($PM_{2.5}$) is a pervasive air pollutant that affects Earth's radiative balance (Heald *et al* 2014, Wang *et al* 2015, Dang and Liao 2019, Li *et al* 2022), reduces visibility (Chen *et al* 2014, Pui *et al* 2014, Luan *et al* 2018), and drives haze formation (Huang *et al* 2014, Wang *et al* 2016, An *et al* 2019). Exposure to $PM_{2.5}$ is associated with adverse respiratory and cardiovascular outcomes (Chen and Hoek 2020, Manisalidis *et al* 2020, Thangavel *et al* 2022), emphasizing the need for timely and accurate forecasts to support public health responses and environmental regulation (Jin *et al* 2016, Wei and Kao 2023). At present, global atmospheric composition prediction is dominated by numerical systems, such as the Copernicus Atmosphere Monitoring Service (CAMS) (Inness *et al* 2019, Peuch *et al* 2022), which couples the European Centre for Medium Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) with an atmospheric composition component (Huijnen *et al* 2016). By assimilating satellite and *in situ* observations (Garrigues *et al* 2022, 2023), CAMS routinely provides global composition forecasts out to five days (Bozzo *et al* 2020). These systems explicitly simulate emissions, chemical transformation and transport, and wet and dry deposition (Manders et al 2017, Aleksankina et al 2018). However, these systems are computationally demanding, even exceeding the cost of global numerical weather prediction (NWP) (Zhang et al 2012, Stein et al 2015). This computational burden constrains operational choices in spatial resolution, ensemble size, and update frequency, all of which are critical for timely $PM_{2.5}$ early warning applications (Sokhi et al 2022).

The advances in machine/deep learning (ML/DL) have enabled a new generation of data-driven global artificial intelligence (AI) weather models, including Pangu-Weather (Bi et al 2023), FuXi (Chen et al 2023a, 2024), and GraphCast (Lam et al 2023). Running four to five orders of magnitude faster than traditional NWP, these models have achieved forecast skill comparable to, or even exceeding the state-of-the-art global medium range systems (de Burgh-day and Leeuwenburg 2023, Chen et al 2023b). They have also shown strong capability in predicting high impact events such as tropical cyclones, heatwaves, and extreme precipitation and flooding (Chen et al 2020, Brunner et al 2021, Salcedo-Sanz et al 2024). Recently, the global aerosol forecasting system with machine learning (AI-GAMFS) has demonstrated strong predictive skill for air quality (Gui et al 2026). At present, major operational centers (e.g. ECMWF and NOAA) have increasingly incorporated their AI forecasts as a complement to physics-based models (Lang et al 2024, Pan et al 2025). However, a key limitation remains: most existing AI models are developed for specific targets and therefore lack the flexibility needed for cross domain applications. Adapting them to new tasks often requires training from scratch, making end-to-end retraining highly computationally demanding (Schultz et al 2021, Charlton-Perez et al 2024, Yu et al 2024). Consequently, repeating this process is inefficient and may hinder broader adoption and scalability of AI across the Earth system (Reichstein et al 2019, Gettelman et al 2022).

Recently, a growing direction in AI for Earth science is to support cross domain prediction within a single modeling framework (Irrgang et al 2021, Yu and Ma 2021). For example, Aurora, a deep learning foundation model explicitly developed for the Earth system (Bodnar et al 2025), is built on a three-dimensional (3D) Swin Transformer architecture and pretrained on more than one million hours of diverse weather and climate simulation data, enabling it to learn the governing dynamics of the atmosphere. Using a parameter efficient fine-tuning technique, the model can be rapidly adapted with task specific data for a range of downstream applications, including global air quality prediction, ocean surface wave forecasting, tropical cyclone tracking, and high-resolution forecasting (Bodnar et al 2025). Compared with training separate large models from scratch, this fine-tuning paradigm substantially reduces data and compute requirements, accelerates transfer to new tasks, and enables comparisons across applications within a unified modeling framework (Pan and Yang 2010, Safonova et al 2023). This transition signals a broader shift in AI-driven modeling paradigms (Jakubik et al 2023). Although the pre-training and fine-tuning paradigm has been highly successful in other domains (e.g. large language models, LLMs), its potential in Earth science has not been fully explored (Ma et al 2024). In particular, it remains unclear to what extent a foundation model trained primarily for global weather prediction can be repurposed for atmospheric composition (Ma et al 2019, Méndez et al 2023).

Building on the fine-tuned Aurora foundation model for global air quality forecasting, we conduct year-long retrospective five-day forecasts of surface $PM_{2.5}$ for 2025. Using the CAMS analysis as a consistent reference and the operational CAMS forecast as a benchmark, we quantify global forecast skill and its dependence on lead time. We further validate the predictions against 3540 independent surface monitoring stations compiled from OpenAQ and the China National Environmental Monitoring Center (CNEMC), enabling a robust evaluation across regions and seasons. In addition, we assess model performance over China, including its representation of a high impact haze episode. Overall, this study assesses the potential of transferring weather foundation model to air quality forecasting and provides a useful reference for future data-driven model development.

2. Data, model and methods

2.1. Data collection

CAMS Products: Gridded surface $PM_{2.5}$ fields were obtained from the CAMS. The CAMS analysis was used as the reference dataset baseline, and the CAMS forecast was used as a benchmark. The CAMS analysis for the full year of 2025 was downloaded at the two synoptic times each day (00:00 and 12:00 UTC). For the CAMS forecast, forecast cycles initialized at 12:00 UTC were selected every five days (starting from 31 December 2024 at 12:00 UTC), and forecast outputs were retained at 12 h intervals from 12 to 120 h lead time, corresponding to a five-day prediction horizon.

OpenAQ and CNEMC: OpenAQ and CNEMC. Surface $PM_{2.5}$ observations were compiled from OpenAQ and the CNEMC. For OpenAQ, we used only measurements from government reference monitoring stations and applied quality control based on annual data completeness. Stations in North America, Europe, and East Asia were retained only when annual data coverage was at least 95%. For India, this criterion retained only one station. Therefore, a relaxed threshold of 80% was applied, which retained 204 of

306 stations with at least 292 valid observation days for 2025. For China, where OpenAQ coverage was absent, city-level $PM_{2.5}$ observations were obtained from CNEMC, and stations with at least 95% annual coverage were retained. Overall, 3540 surface monitoring sites were included across the four analysis regions and used as the independent observational reference for surface $PM_{2.5}$ evaluation.

ERA5: Meteorological fields were obtained from the ERA5 reanalysis, which provided a globally complete and physically consistent depiction of the atmosphere through data assimilation (Hersbach *et al* 2020). In our study, boundary layer height, 2 m temperature, total precipitation, 10 m winds and variables at 1000 hPa including: geopotential, temperature, zonal and meridional winds, specific humidity, were used on a 0.25° × 0.25° latitude–longitude grid for meteorological diagnosis.

GEOS-CF: Gridded surface $PM_{2.5}$ fields were obtained from the NASA Global Earth Observing System Composition Forecast (GEOS-CF). GEOS-CF combines the GEOS weather analysis and forecasting system with the GEOS-Chem chemistry module and provides global atmospheric composition fields on a 0.25° × 0.25° latitude–longitude grid (Keller *et al* 2021). In this study, we used the GEOS-CF replay/analysis archive for the full year of 2025. Specifically, 15 min instantaneous surface $PM_{2.5}$ fields at 00:00 and 12:00 UTC were used to match the CAMS output times used for comparison.

2.2. Model and setup

Aurora is a 1.3-billion-parameter AI foundation model for Earth system. It is first pretrained on more than one million hours of historical data spanning multiple sources and resolutions, including weather analysis/reanalysis products and numerical simulations from weather and climate models, to learn a transferable representation of global 3D atmospheric variability. This pretrained backbone is primarily aimed at global medium range weather prediction and can then be adapted to other Earth science forecasting tasks via parameter efficient fine-tuning (Low-rank adaptation; LoRA), such as global air quality prediction, ocean surface wave forecasting, tropical cyclone tracking, and higher resolution weather prediction.

Architecturally, Aurora follows an encoder-processor-decoder paradigm: a Perceiver-based encoder maps multi-variable inputs into a universal latent representation. A multiscale 3D Swin Transformer U-Net evolves this latent state forward in time, and a Perceiver-based decoder translates the updated representation back to physical fields on a latitude–longitude grid. Multi-lead forecasts are produced autoregressively by recursively feeding model predictions back as inputs. For atmospheric composition, Aurora was fine-tuned on CAMS analysis fields (data are from 2015 to 2022) to emulate global air quality forecasting at 0.4° resolution, including key gaseous pollutants (CO, NO, NO_2, SO_2, and O_3) and particulate matter (PM_1, $PM_{2.5}$, and PM_{10}).

In our study, we used the publicly released Aurora air quality fine-tuned checkpoint and performed deterministic inference throughout 2025. Each forecast was initialized from the CAMS analysis and rolled out to 120 h, with outputs archived every 12 h on the global 0.4° × 0.4° latitude–longitude grid consistent with the CAMS products. Although Aurora produces a broad set of composition fields, we evaluate surface $PM_{2.5}$ as the primary target variable in this work.

2.3. Metrics and data processing

Model performance was quantified using the root mean squared error (RMSE), normalized RMSE (NRMSE), percentage error (PE), and the Pearson correlation coefficient (R). For paired samples (reference, prediction) indexed by $i = 1, \ldots, N$, RMSE was defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2} \tag{1}$$

where y_i denotes the reference (e.g. CAMS analysis or observations) and $\hat{y}_i$ denotes the prediction (e.g. Aurora or CAMS forecast). NRMSE was computed by normalizing RMSE by the mean of the reference values,

$$\mathrm{NRMSE}\,\frac{\mathrm{RMSE}}{\bar{y}}, \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{2}$$

The percentage error (PE, %) was computed as the relative mean bias,

$$\mathrm{PE} = 100 \times \frac{\bar{\hat{y}} - \bar{y}}{\bar{y}}, \quad \bar{\hat{y}} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i. \tag{3}$$

The Pearson correlation coefficient was computed as

$$\mathrm{R} = \frac{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}}. \tag{4}$$

Gridded model and reanalysis fields were spatially collocated with surface monitoring sites before station-level statistics were calculated. For Aurora and CAMS on the 0.4° grid, and GEOS-CF and ERA5 on the 0.25° grid, values at each station were obtained by bilinear interpolation from the surrounding latitude–longitude grid cells. Before interpolation, latitude and longitude coordinates were consistently ordered, and station longitudes were converted to match the longitude convention of each model grid when necessary. The interpolated values were then paired with the corresponding surface observations at the same valid UTC time. Only valid paired samples were used to calculate RMSE, NRMSE, PE, and R.

3. Results

3.1. Global model skills

The global daily mean surface $PM_{2.5}$ (figure 1(a)) shows broad agreement between Aurora and the CAMS analysis. The CAMS forecast is systematically lower than the analysis throughout the year, whereas Aurora exhibits a seasonally varying bias: it is higher than the analysis in winter and spring but drops sharply in summer (July–September) and becomes lower than the analysis. The global spatial variability, quantified by the spatial standard deviation (figure 1(b)), is consistently smaller in Aurora than in the CAMS products, with the largest discrepancy during mid-year. The difference then narrows toward late autumn and winter.

We further assess lead-time dependence over a 120 h horizon using 73 forecast initializations in 2025. The (PE; figure 1(c)) shows opposite systematic tendencies: Aurora overestimates the global mean $PM_{2.5}$ (positive PE) with an increasing magnitude toward longer lead times, while the CAMS forecast underestimates it (negative PE). For lead times up to ~72 h, Aurora exhibits a smaller bias relative to CAMS. The normalized RMSE (NRMSE; figure 1(d)) increases approximately linearly with lead time for both systems. Although CAMS maintains lower NRMSE than Aurora across all lead times, Aurora shows a smaller fitted NRMSE growth rate (2.4×10^{-3}) than the CAMS forecast (3.3×10^{-3}), indicating a slower degradation of skill with increasing lead time.

$PM_{2.5}$ has large variations across different seasons and regions. To assess Aurora's performance across diverse environments, we validate surface $PM_{2.5}$ model predictions against observations from 3540 monitoring stations worldwide, compiled from OpenAQ and CNEMC. We focus on four regional domains: North America (NA), Europe (EU), East Asia (EA), and India (IN) (figure 2(a)).

Seasonal density scatterplots (figure 2(b)) show that Aurora broadly reproduces the observed variability, with pronounced regional and seasonal dependence. In North America, Aurora attains relatively modest errors (e.g. PE = 5.8% and NRMSE = 0.55 in October–December), whereas Europe exhibits strong seasonal contrasts in bias, with marked underestimation in winter (PE = −29.8% in January–March) and overestimation in summer (PE = 28.2% in July–September). Across regions, regression lines lie below the 1:1 line at higher concentrations, indicating a tendency to underpredict extremes. In East Asia, regression fits for Aurora and the CAMS products lie close to the 1:1 line across seasons. Aurora captures broad $PM_{2.5}$ variability particularly during the cold season ($R = 0.82$ and NRMSE = 0.55 in January–March), but performance degrades substantially in summer (PE = 54.4% and NRMSE = 0.94 in July–September). Over India, the number of observations is relatively small and the scatter points are widely distributed. Nevertheless, Aurora generally shows regression fits closest to the 1:1 line among the three products, suggesting skill comparable to an operational system when evaluated against independent surface observations.

3.2. The overall performance over China

We next examine model performance over China, focusing on northern, central, and eastern China, where heavy haze episodes occur frequently. Because Aurora is fine-tuned on CAMS analysis, we include NASA GEOS-CF as an additional independent product for comparison. For annual evolution of surface mean, all four gridded products reproduce the broad seasonal cycle, with high concentrations in winter, a pronounced minimum in summer, and a renewed increase in late autumn and winter (figure 3(a)). CAMS analysis agrees best with the CNEMC mean time series (NRMSE = 0.24, $R = 0.90$). Aurora closely tracks the observed temporal variability and yields skill comparable to the CAMS products (NRMSE = 0.25, $R = 0.86$), slightly better than the CAMS forecast in NRMSE (0.26). Notably, GEOS-CF exhibits larger errors and weaker correlations, indicating systematic differences from the CAMS-based products and the observations. The station-wise distributions of mean bias and NRMSE (figure 3(b)) shows the smallest overall positive bias (3.7 μgm^{-3}) for Aurora, lower than CAMS analysis (4.2 μgm^{-3}), CAMS forecast (5.0 μgm^{-3}), and GEOS-CF (11.1 μgm^{-3}), while its NRMSE distribution remains close to that of CAMS analysis and clearly lower than those of CAMS forecast and GEOS-CF.

The spatial distribution of station-level NRMSE (figure 3(c)) shows that CAMS analysis captures the broad regional pattern of skill, whereas GEOS-CF exhibits systematically larger errors across much of eastern China. Relative to CAMS analysis, the CAMS forecast shows mostly weak positive NRMSE differences, indicating modest degradation from the analyzed state. Aurora, in contrast, exhibits predominantly negative NRMSE differences, with reduced errors over many stations in central and eastern China, suggesting performance comparable to, and locally better than, the numerical products.

3.3. Performance during an extreme winter haze episode

Assessing model skill is important not only for overall performance, but also for the representation of extreme pollution episodes. We therefore further examine model behavior during a persistent winter haze episode over northern and eastern China. This event spans 13–21 December 2025 (Beijing local time) and affects northern China as well as adjacent central and eastern regions. Figure 4(a) summarizes the temporal evolution at 12 h intervals by comparing the CNEMC station mean (black line) and the

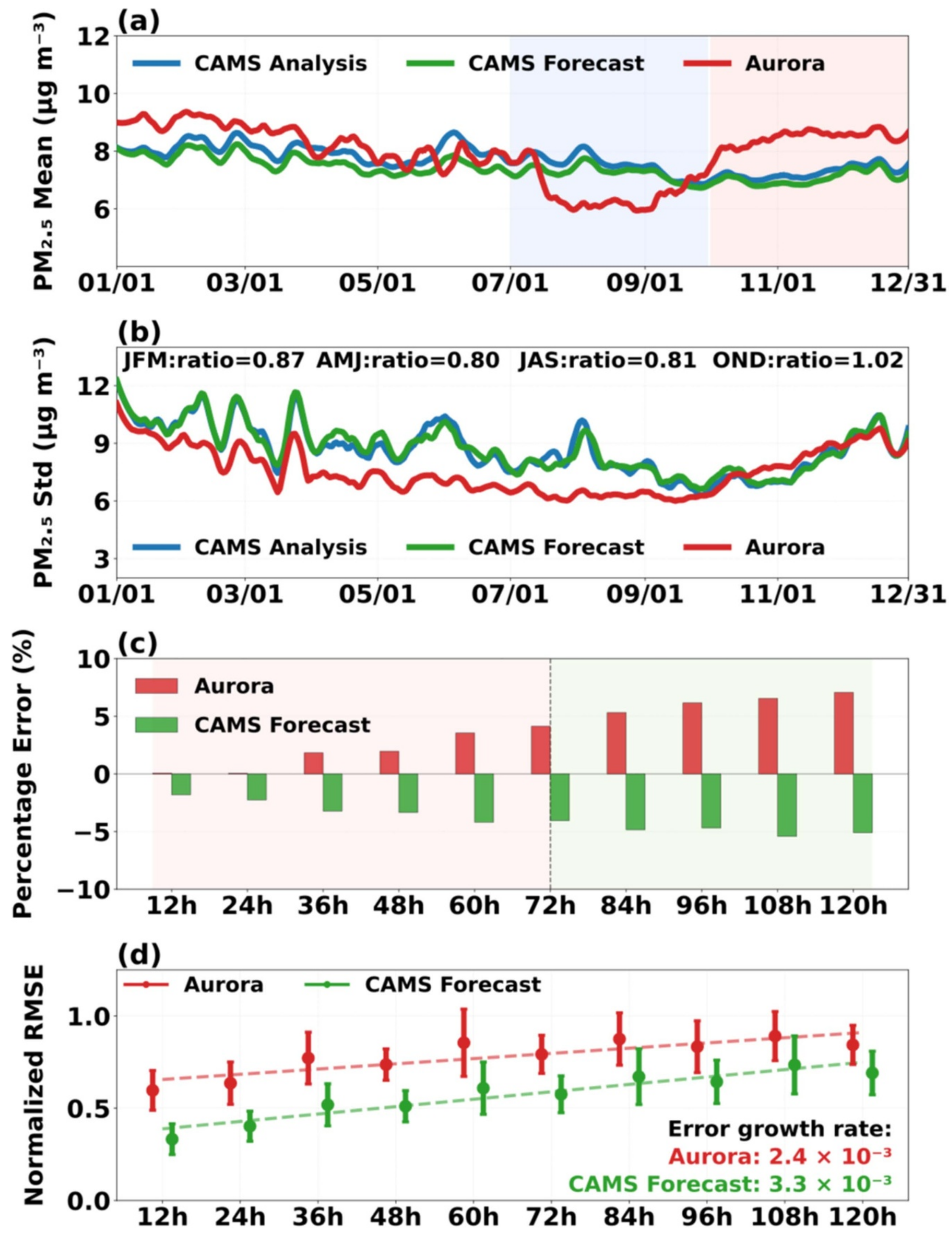

Figure 1. Temporal evolution and lead-time dependence of global surface $PM_{2.5}$ predictions in 2025. (a) Daily global mean surface $PM_{2.5}$ from the CAMS analysis, the CAMS forecast and Aurora. Shaded intervals highlight periods with distinct seasonal divergence between the predictions and the analysis. (b) Daily time series of the global spatial standard deviation of surface $PM_{2.5}$, the seasonal ratios between Aurora and CAMS Analysis are also shown. (c) Percentage error (PE, %) of the global mean $PM_{2.5}$ relative to the CAMS analysis as a function of lead time (12–120 h). The dashed vertical line marks 72 h. (d) Normalized RMSE (NRMSE) relative to the CAMS analysis versus lead time (12–120 h). Error bars denote the standard deviation across 73 forecast initializations. Dashed lines show linear fits used to estimate error growth rates.

observed mean-to-maximum range across stations (shading) with Aurora, the CAMS analysis, and the CAMS forecast. For each time step, bars indicate the collocated model mean (bar base) and maximum (bar top), enabling a simultaneous assessment of background conditions and extremes. Across most time steps, Aurora is closest to both the observed mean and the station-maximum, highlighting the best overall performance for this episode.

The spatial evolution at four representative times is broadly consistent with the temporal diagnostics (figure 4(b)). CNEMC observations show a

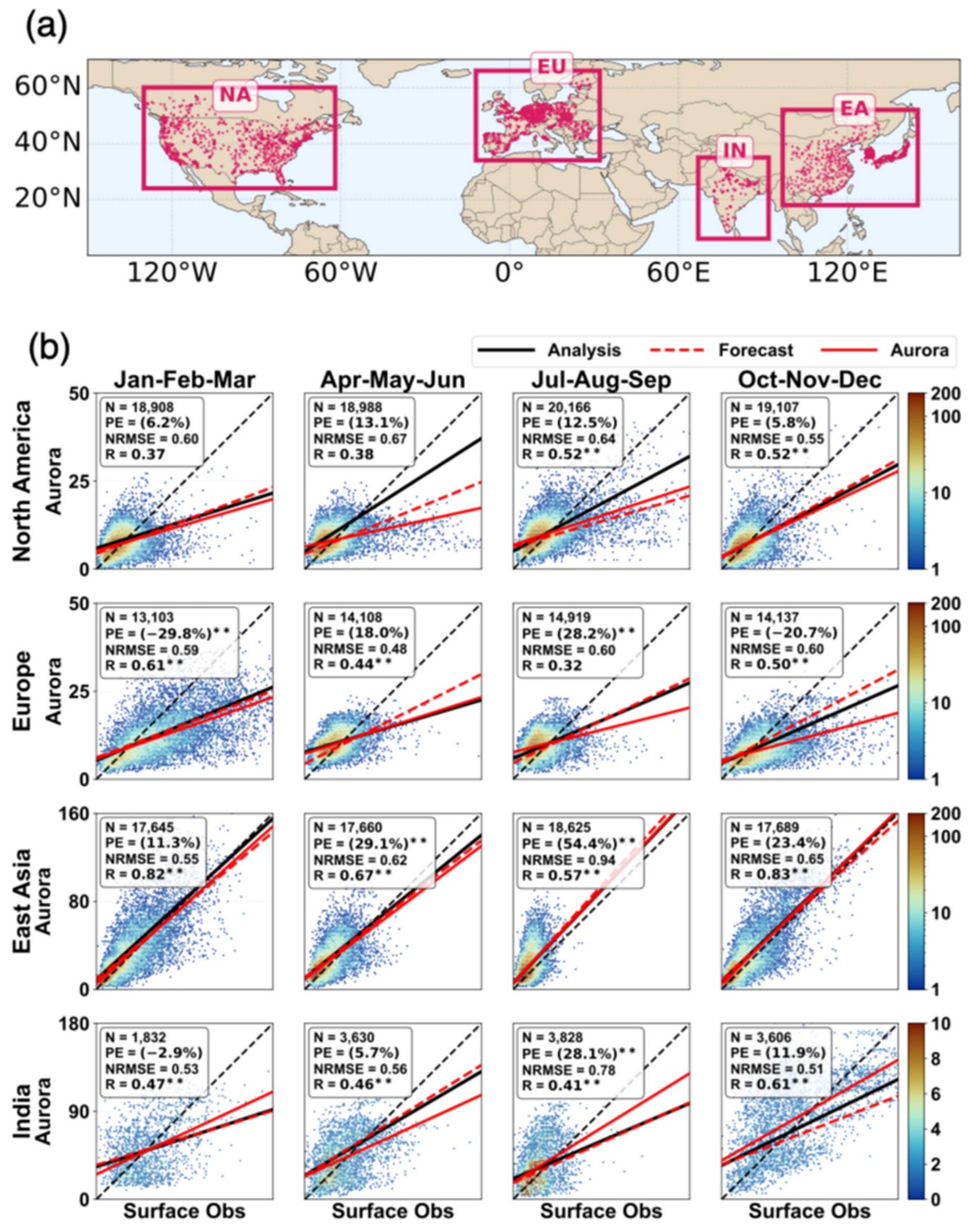

Figure 2. Validation of Aurora surface $PM_{2.5}$ against surface monitoring stations in 2025. (a) Locations of the 3540 stations (red dots) used for evaluation, compiled from OpenAQ and CNEMC. Red boxes delineate the four regional domains: North America (NA), Europe (EU), India (IN), and East Asia (EA). (b) Seasonal density scatterplots of predicted versus observed surface $PM_{2.5}$ for 2025, organized by region (rows) and season (columns). Points show Aurora predictions against surface observations. Colors indicate point density on a logarithmic scale. The black dashed line denotes the 1:1 line. Solid black, red dashed, and solid red lines show linear regression fits for the CAMS analysis, CAMS forecast, and Aurora, respectively. Double asterisks (**) indicate statistically significant at the 95% confidence level ($p < 0.05$).

pronounced pollution core over the North China Plain that subsequently expands southeastward toward central and eastern China. Aurora broadly reproduces the location of the high pollution center and its downstream extension, indicating that it captures the main spatiotemporal structure of this persistent haze episode. Although local discrepancies in intensity and fine scale structure remain, this case study further suggests that the fine-tuned Aurora model provides skill comparable to, and at some times better than, the operational numerical products during extreme haze conditions.

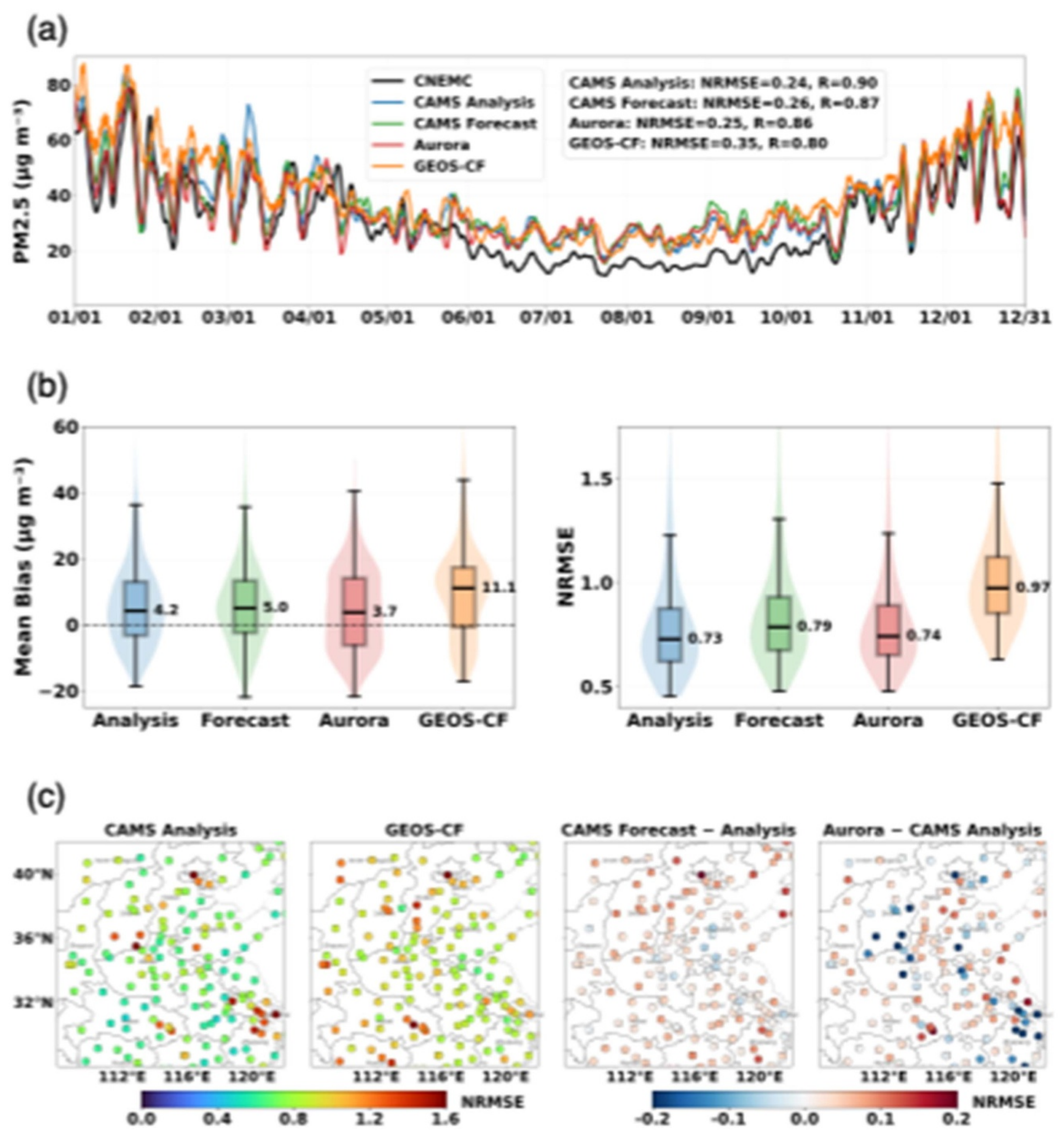

Figure 3. Evaluation of surface $PM_{2.5}$ over China against CNEMC observations in 2025. (a) Time series of mean surface $PM_{2.5}$ from CNEMC and from the collocated CAMS analysis, CAMS forecast, Aurora, and GEOS-CF products over China. (b) Violin-box distributions of station-wise mean bias and NRMSE for the four products across CNEMC sites in 2025. (c) Station-level spatial distributions of NRMSE for CAMS analysis and GEOS-CF over China, together with NRMSE differences for CAMS forecast minus CAMS analysis and Aurora minus CAMS analysis. Negative values in the difference maps indicate smaller errors than the CAMS analysis, whereas positive values indicate larger errors.

4. Discussion

The relatively large positive bias of Aurora over East Asia in summer is obvious (figure 2(b)). This bias is not unique to Aurora. A similar systematic overestimation is also evident in the CAMS analysis and forecast (figure S2), suggesting that part of the bias may originate from the underlying CAMS model. In addition, although the absolute bias of Aurora is moderate (6.79 μg m^{-3}), the observed summer mean $PM_{2.5}$ concentration is relatively low (only 12.49 μg m^{-3}). This low observed baseline acts as the denominator in the PE calculation and therefore amplifies the PE. Notably, the widespread precipitation associated with the East Asian summer monsoon during its active phase (June–September), can substantially enhance atmospheric cleaning through wet deposition (figures S2(g) and (h)). If the CAMS numerical system underrepresents wet deposition or aerosol removal with precipitation, such deficiencies may contribute to the systematic positive biases observed in both CAMS and Aurora.

Aurora shows a systematic wintertime underestimation over Europe (figure 2(b)). This bias also may be partly inherited from the CAMS Analysis used for fine-tuning, as CAMS Analysis exhibits a similar negative bias and its spatial pattern is broadly consistent with regions of higher heating degree days (HDD) in central and eastern Europe (figures S3(a) and (b)). In winter, residential heating, especially biomass and

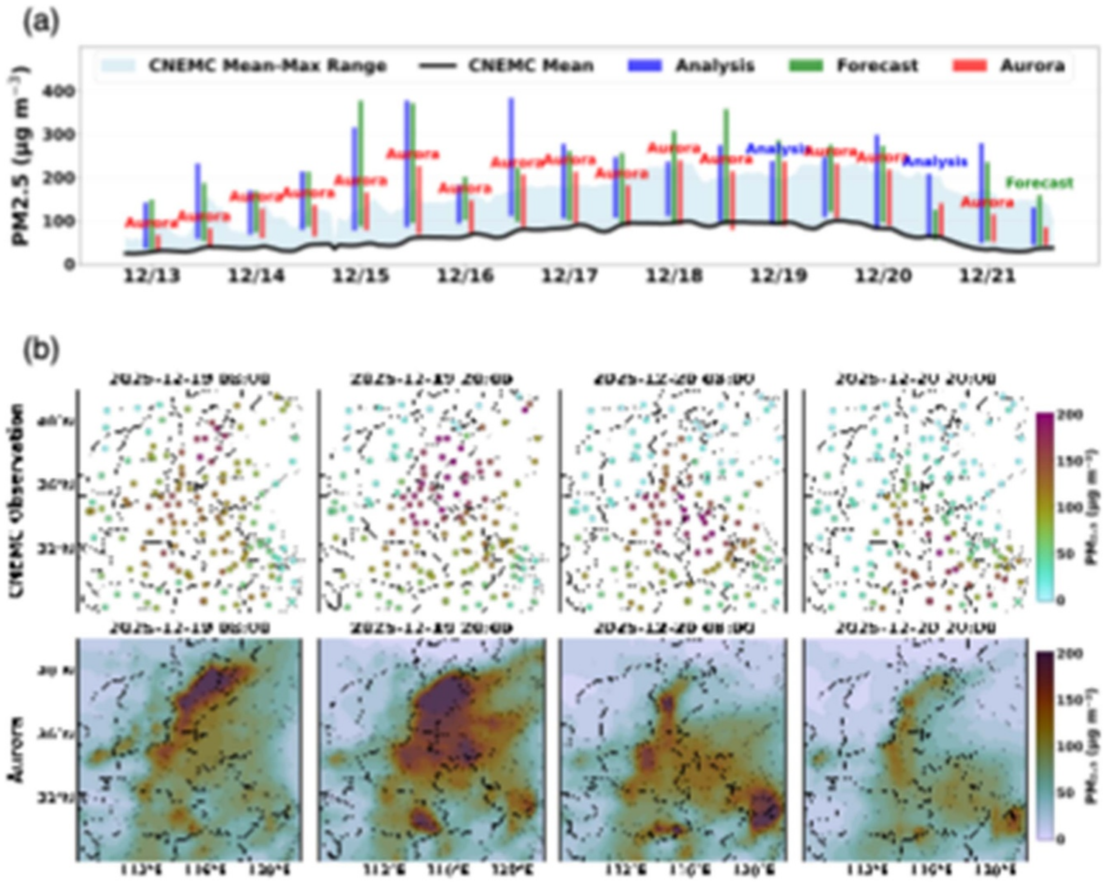

Figure 4. Case study of a persistent winter haze episode over China (13–21 December 2025, Beijing local time) evaluated against CNEMC observations. (a) Regional time series of surface $PM_{2.5}$ over the impacted domain at 12 h intervals. The black line shows the CNEMC station mean, and the shaded area indicates the observed mean-maximum range across stations. Colored bars summarize each model's regional distribution at the same times, with bar bases denoting the mean and bar tops denoting the maximum (Aurora, CAMS analysis, and CAMS forecast). Text labels indicate the model that is closest to both the observed station-mean and station-maximum at each time step. (b) Evolution at four peak times during the episode. The top row shows CNEMC station $PM_{2.5}$, the bottom row shows Aurora surface $PM_{2.5}$.

wood combustion, represents an important but spatially heterogeneous and uncertain source of $PM_{2.5}$ emissions, which can introduce errors into air quality models (Denier van der Gon *et al* 2015, Lacima *et al* 2023). In addition, CAMS Analysis is slightly warmer than ERA5 over Europe in January–March (+0.34 °C on average), implying a possible underestimation of HDD demand (figure S3(c)). Shallow boundary layers and weak near surface winds in winter may further enhance $PM_{2.5}$ accumulation (figure S3(d)).

Aurora produces smoother $PM_{2.5}$ variations than CAMS, presenting a weaker ability to capture observed concentration peaks (figure 1(b)). We further examined this behavior across regions and seasons (table T1). Similar smoothing is evident at the regional scale, although its magnitude varies by region and season. The underestimation of variability is particularly pronounced in North America and Europe, which may be partly related to their lower regional mean $PM_{2.5}$ levels.

In the extreme haze case study, Aurora performed better than the CAMS products in capturing both mean and peak $PM_{2.5}$ concentrations. The event developed over the North China Plain and then extended into central and eastern China, highlighting the important role of meteorological control in regional pollutant transport. During this episode, a shallow boundary layer and strong near surface temperature inversion favored pollutant accumulation (figure S1(a)). Under these conditions, Aurora showed better agreement with ERA5 in the 1000 hPa low-level meteorological fields (figure S1(b)), suggesting a more realistic representation of the meteorological environment. Nevertheless, Aurora still showed limited skill in reproducing extreme $PM_{2.5}$ peaks. We further analyzed the winter period from October 2023 to April 2024, Aurora generally agreed well with surface observations (figure S4), but it underestimated three high-pollution episodes to varying degrees over the North China Plain and eastern China (figure S5). This further emphasizes the need to improve the representation of localized extreme pollution peaks in future model development.

Computational efficiency should also be considered. In our inference setup, Aurora required approximately 120 s to complete a five-day global forecast on a single NVIDIA A6000 GPU. Harder *et al* (2026) reported that the newly developed data-driven model AIFS-COMPO generated a 3 hourly, five-day global atmospheric composition forecast in about 50 s on a single GPU, whereas the corresponding operational IFS-COMPO forecast required roughly 1000 s on about 8000 CPUs. In terms of hardware-time consumption, the IFS-COMPO estimate corresponds to approximately 2.2×10^3 CPU-h, whereas Aurora requires only about 0.03 GPU-h per five-day forecast. Although CPU-hours and GPU-hours are not directly interchangeable, this comparison indicates that Aurora has a substantially smaller operational hardware footprint.

5. Conclusion

This study evaluated the Aurora Earth system foundation model for five-day global surface $PM_{2.5}$ prediction using retrospective forecasts for 2025. Aurora closely reproduced the large scale temporal evolution of surface $PM_{2.5}$ and showed a smaller global mean bias than the operational CAMS forecast during the first 72 h, although CAMS retained lower normalized RMSE when evaluated against its own analysis. Comparisons with independent 3540 surface monitoring sites further showed that Aurora captured major regional and seasonal $PM_{2.5}$ variability across North America, Europe, East Asia, India, and China. Over China, Aurora achieved skill comparable to CAMS products and better than GEOS-CF in annual station level statistics, and it reproduced the main spatiotemporal structure of a persistent winter haze episode over northern and eastern China. These results indicate that transfer learning from a weather foundation model can provide a computationally efficient route toward global air quality forecasting. Future work should focus on regional fine-tuning with dense surface observations and higher resolution emission and meteorological information, together with probabilistic or ensemble designs to better represent localized extremes and forecast uncertainty.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (72293604, 42206253, 42205025, 42530606 and 42261144687), and the program for scientific research start-up funds of Guangdong Ocean University (060302032305).

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: CAMS global atmospheric composition analysis and forecast fields were obtained from the Copernicus Atmosphere Data Store (ADS) (https://ads.atmosphere.copernicus.eu/datasets/cams-global-atmospheric-composition-forecasts?tab=overview) (CAMS 2021). Meteorological reanalysis fields were taken from the Copernicus Climate Data Store (CDS) ERA5 products on single levels (https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview) (CDS *et al* 2020) and pressure levels (https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels?tab=overview) (C3S 2018). Ground-based air quality observations were obtained from the OpenAQ platform (https://openaq.org) (OpenAQ 2015) and the China National Environmental Monitoring Centre (CNEMC) data service (www.cnemc.cn) (CNEMC 2026). NASA GEOS-CF model output data can be accessed at https://gmao.gsfc.nasa.gov/gmao-products/geos-cf/data-access_geos-cf/ (GMAO 2019).

Supplementary Figures and Table available at: https://doi.org/10.1088/1748-9326/ae90c2/data1.

Conflict of interest

The authors declare no competing interests.

Author contributions

Delin Li 0000-0002-9471-0358
Conceptualization (lead), Data curation (lead), Formal analysis (lead), Funding acquisition (equal), Investigation (equal), Methodology (equal), Project administration (equal), Resources (equal), Software (equal), Supervision (lead), Validation (equal), Visualization (equal), Writing – original draft (lead), Writing – review & editing (lead)

Zhiqiang Chen
Conceptualization (lead), Data curation (lead), Formal analysis (lead), Funding acquisition (equal), Investigation (equal), Methodology (lead), Project administration (equal), Resources (lead), Software (equal), Supervision (lead), Validation (lead), Visualization (lead), Writing – original draft (lead), Writing – review & editing (equal)

Gang Huang 0000-0002-8692-7856
Conceptualization (equal), Data curation (supporting), Formal analysis (supporting), Funding acquisition (supporting), Investigation (equal), Methodology (equal), Project administration (supporting), Resources (supporting), Software (supporting), Supervision (supporting), Validation (supporting), Visualization (supporting), Writing – original draft (supporting), Writing – review & editing (supporting)

Jianjun Xu 0000-0002-6219-0976
Conceptualization (equal), Data curation (supporting), Formal analysis (supporting), Funding acquisition (lead), Investigation (equal), Methodology (supporting), Project administration (equal), Resources (supporting), Software (supporting), Supervision (supporting), Validation (supporting), Visualization (supporting), Writing – original draft (supporting), Writing – review & editing (supporting)

References

Aleksankina K, Heal M R, Dore A J, Van Oijen M and Reis S 2018 Global sensitivity and uncertainty analysis of an atmospheric chemistry transport model: the FRAME model (version 9.15.0) as a case study *Geosci. Model Dev.* **11** 1653–64

An Z *et al* 2019 Severe haze in northern China: a synergy of anthropogenic emissions and atmospheric processes *Proc. Natl Acad. Sci. USA* **116** 8657–66

Bi K, Xie L, Zhang H, Chen X, Gu X and Tian Q 2023 Accurate medium-range global weather forecasting with 3D neural networks *Nature* **619** 533–8

Bodnar C *et al* 2025 A foundation model for the Earth system *Nature* **641** 1180–7

Bozzo A, Benedetti A, Flemming J, Kipling Z and Rémy S 2020 An aerosol climatology for global models based on the tropospheric aerosol scheme in the integrated forecasting system of ECMWF *Geosci. Model Dev.* **13** 1007–34

Brunner M I, Slater L, Tallaksen L M and Clark M 2021 Challenges in modeling and predicting floods and droughts: a review *WIREs Water* **8** e1520

CAMS: Copernicus Atmosphere Monitoring Service 2021 CAMS global atmospheric composition forecasts *Atmosphere Data Store (ADS)* (available at: https://ads.atmosphere.copernicus.eu/datasets/cams-global-atmospheric-composition-forecasts?tab=overview)

Charlton-Perez A J *et al* 2024 Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciaran *npj Clim. Atmos. Sci.* **7** 93

Chen J and Hoek G 2020 Long-term exposure to PM and all-cause and cause-specific mortality: a systematic review and meta-analysis *Environ. Int.* **143** 105974

Chen J, Qiu S, Shang J, Wilfrid O M F, Liu X, Tian H and Boman J 2014 Impact of relative humidity and water soluble constituents of $PM_{2.5}$ on visibility impairment in Beijing, China *Aerosol. Air Qual. Res.* **14** 260–8

Chen L *et al* 2024 A machine learning model that outperforms conventional global subseasonal forecast models *Nat. Commun.* **15** 6425

Chen L, Han B, Wang X, Zhao J, Yang W and Yang Z 2023b Machine learning methods in weather and climate applications: a survey *Appl. Sci.* **13** 12019

Chen L, Zhong X, Zhang F, Cheng Y, Xu Y, Qi Y and Li H 2023a FuXi: a cascade machine learning forecasting system for 15-day global weather forecast *npj Clim. Atmos. Sci.* **6** 190

Chen R, Zhang W and Wang X 2020 Machine learning in tropical cyclone forecast modeling: a review *Atmosphere* **11** 676

CNEMC 2026 China National Air Quality Monitoring Data *China National Environmental Monitoring Centre (CNEMC)* (available at: www.cnemc.cn)

Copernicus Climate Change Service (C3S) 2018 ERA5 hourly data on pressure levels from 1940 to present *Climate Data Store (CDS)* (available at: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels?tab=overview)

Copernicus Climate Change Service (C3S) 2020 ERA5 hourly data on single levels from 1940 to present *Climate Data Store (CDS)* (available at: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview)

Dang R and Liao H 2019 Radiative forcing and health impact of aerosols and ozone in China as the consequence of clean air actions over 2012–2017 *Geophys. Res. Lett.* **46** 12511–9

de Burgh-day C O and Leeuwenburg T 2023 Machine learning for numerical weather and climate modelling: a review *Geosci. Model Dev.* **16** 6433–77

Denier van der Gon H A C, Bergström R, Fountoukis C, Johansson C, Pandis S N, Simpson D and Visschedijk A J H 2015 Particulate emissions from residential wood combustion in Europe-revised estimates and an evaluation *Atmos. Chem. Phys.* **15** 6503–19

Garrigues S *et al* 2022 Monitoring multiple satellite aerosol optical depth (AOD) products within the Copernicus atmosphere monitoring service (CAMS) data assimilation system *Atmos. Chem. Phys.* **22** 14657–92

Garrigues S *et al* 2023 Impact of assimilating NOAA VIIRS aerosol optical depth (AOD) observations on global AOD analysis from the Copernicus atmosphere monitoring service (CAMS) *Atmos. Chem. Phys.* **23** 10473–87

Gettelman A, Geer A J, Forbes R M, Carmichael G R, Feingold G, Posselt D J, Stephens G L, van den Heever S C, Varble A C and Zuidema P 2022 The future of Earth system prediction: advances in model-data fusion *Sci. Adv.* **8** eabn3488

Gui K *et al* 2026 Advancing operational global aerosol forecasting with machine learning *Nature* **651** 658–65

Harder P, Flemming J, Alexe M, Mertes G, Raoult B and Chantry M 2026 AIFS-COMPO: a global data-driven atmospheric composition forecasting system (arXiv: 2604.03300)

Heald C L, Ridley D A, Kroll J H, Barrett S R H, Cady-pereira K E, Alvarado M J and Holmes C D 2014 Contrasting the direct radiative effect and direct radiative forcing of aerosols *Atmos. Chem. Phys.* **14** 5513–27

Hersbach H *et al* 2020 The ERA5 global reanalysis *Q. J. R. Meteorol. Soc.* **146** 1999–2049

Huang R-J *et al* 2014 High secondary aerosol contribution to particulate pollution during haze events in China *Nature* **514** 218–22

Huijnen V *et al* 2016 C-IFS-CB05-BASCOE: stratospheric chemistry in the integrated forecasting system of ECMWF *Geosci. Model Dev.* **9** 3071–91

Inness A *et al* 2019 The CAMS reanalysis of atmospheric composition *Atmos. Chem. Phys.* **19** 3515–56

Irrgang C *et al* 2021 Towards neural Earth system modelling by integrating artificial intelligence in Earth system science *Nat. Mach. Intell.* **3** 667–74

Jakubik J *et al* 2023 Foundation models for generalist geospatial artificial intelligence (arXiv: 2310.18660)

Jin Y, Andersson H and Zhang S 2016 Air pollution control policies in China: a retrospective and prospects *Int. J. Environ. Res. Public Health* **13** 1219

Keller C A *et al* 2021 Description of the NASA GEOS composition forecast modeling system GEOS-CF v1.0 *J. Adv. Model. Earth Syst.* **13** e2020MS002413

Lacima A, Petetin H, Soret A, Bowdalo D, Jorba O, Chen Z, Mendez Turrubiates R F, Achebak H, Ballester J and Perez Garcia-pando C 2023 Long-term evaluation of surface air pollution in CAMSRA and MERRA-2 global reanalyses over Europe (2003–2020) *Geosci. Model Dev.* **16** 2689–718

Lam R *et al* 2023 Learning skillful medium-range global weather forecasting *Science* **382** 1416–21

Lang S *et al* 2024 AIFS–ECMWF's data-driven forecasting system (arXiv: 2406.01465)

Li J *et al* 2022 Scattering and absorbing aerosols in the climate system *Nat. Rev. Earth Environ.* **3** 363–79

Luan T, Guo X, Guo L and Zhang T 2018 Quantifying the relationship between $PM_{2.5}$ concentration, visibility and planetary boundary layer height for long-lasting haze and fog-haze mixed events in Beijing *Atmos. Chem. Phys.* **18** 203–25

Ma J, Cheng J C P, Lin C, Tan Y and Zhang J 2019 Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques *Atmos. Environ.* **214** 116885

Ma Y, Chen S, Ermon S and Lobell D B 2024 Transfer learning in environmental remote sensing *Remote Sens. Environ.* **301** 113924

Manders A M M *et al* 2017 Curriculum vitae of the LOTOS-EUROS (v2.0) chemistry transport model *Geosci. Model Dev.* **10** 4145–73

Manisalidis I, Stavropoulou E, Stavropoulos A and Bezirtzoglou E 2020 Environmental and health impacts of air pollution: a review *Front. Public Health* **8** 14

Méndez M, Merayo M G and Núñez M 2023 Machine learning algorithms to forecast air quality: a survey *Artif. Intell. Rev.* **56** 10031–66

NASA Global Modeling and Assimilation Office (GMAO) 2019 GEOS-CF (GEOS Composition Forecast) *NASA GMAO Data Portal* (available at: https://gmao.gsfc.nasa.gov/gmao-products/geos-cf/data-access_geos-cf/)

OpenAQ 2015 Air Quality Data *OpenAQ* (available at: https://openaq.org)

Pan L *et al* 2025 Comparative analysis of precipitation forecasts between the ECMWF artificial intelligence system (AIFS) and its integrated forecast system (IFS) *Weather Forecasting* **40** 1653–70

Pan S J and Yang Q 2010 A survey on transfer learning *IEEE Trans. Knowl. Data Eng.* **22** 1345–59

Peuch V-H *et al* 2022 The Copernicus atmosphere monitoring service: from research to operations *Bull. Am. Meteorol. Soc.* **103** E2650–68

Pui D Y H, Chen S-C and Zuo Z 2014 $PM_{2.5}$ in China: measurements, sources, visibility and health effects, and mitigation *Particuology* **13** 1–26

Reichstein M *et al* 2019 Deep learning and process understanding for data-driven Earth system science *Nature* **566** 195–204

Safonova A *et al* 2023 Ten deep learning techniques to address small data problems with remote sensing *Int. J. Appl. Earth Obs. Geoinf.* **125** 103569

Salcedo-Sanz S *et al* 2024 Analysis, characterization, prediction, and attribution of extreme atmospheric events with machine learning and deep learning techniques: a review *Theor. Appl. Climatol.* **155** 1–44

Schultz M G *et al* 2021 Can deep learning beat numerical weather prediction? *Phil. Trans. R. Soc. A* **379** 20200097

Sokhi R S *et al* 2022 Advances in air quality research—current and emerging challenges *Atmos. Chem. Phys.* **22** 4615–703

Stein A F *et al* 2015 NOAA's HYSPLIT atmospheric transport and dispersion modeling system *Bull. Am. Meteorol. Soc.* **96** 2059–77

Thangavel P, Park D and Lee Y-C 2022 Recent insights into particulate matter ($PM_{2.5}$)-mediated toxicity in humans: an overview *Int. J. Environ. Res. Public Health* **19** 7511

Wang G *et al* 2016 Persistent sulfate formation from London fog to Chinese haze *Proc. Natl Acad. Sci. USA* **113** 13630–5

Wang H *et al* 2015 Mesoscale modelling study of the interactions between aerosols and PBL meteorology during a haze episode in China Jing-Jin-Ji and its near surrounding region—Part 2: aerosols' radiative feedback effects *Atmos. Chem. Phys.* **15** 3277–87

Wei C-C and Kao W-J 2023 Establishing a real-time prediction system for fine particulate matter concentration using machine-learning models *Atmosphere* **14** 1817

Yu M, Huang Q and Li Z 2024 Deep learning for spatiotemporal forecasting in Earth system science: a review *Int. J. Digit. Earth* **17** 2391952

Yu S and Ma J 2021 Deep learning for geophysics: current and future trends *Rev. Geophys.* **59** e2021RG000742

Zhang Y, Bocquet M, Mallet V, Seigneur C and Baklanov A 2012 Real-time air quality forecasting, part I: history, techniques, and current status *Atmos. Environ.* **60** 632–55