Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An operational data-driven chemical-weather model for skillful multi-pollutant forecasting

Junting Zhong
Chinese Academy of Meteorological Sciences
Xiaoye Zhang
xiaoye@cma.gov.cn

Chinese Academy of Meteorological Sciences
Zijiang Zhou
National Meteorological Information Centre, China Meteorological Administration (CMA)
Zhiquan Liu
National Meteorological Information Centre, China Meteorological Administration (CMA)
Tao Zhang
National Meteorological Information Centre, China Meteorological Administration (CMA)
Wenjie Zhang
Chinese Academy of Meteorological Sciences
Deying Wang
Chinese Academy of Meteorological Sciences
Zhisen Zhang
National Meteorological Information Centre, China Meteorological Administration (CMA)
Wenhui Xu
National Meteorological Information Centre, China Meteorological Administration (CMA)
Lipeng Jiang
CMA Earth System Modeling and Prediction Centre
Yike Zhou
National Meteorological Information Centre, China Meteorological Administration (CMA)
Yaping Ma
National Meteorological Information Centre, China Meteorological Administration (CMA)
Lifeng Guo
Chinese Academy of Meteorological Sciences
Wenxing Jia
Chinese Academy of Meteorological Sciences
Jing Sun
National Meteorological Information Centre, China Meteorological Administration (CMA)
Yuhang Xin

National Meteorological Information Centre, China Meteorological Administration (CMA)

Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11562604/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Numerical chemical-weather models provide guidance, yet forecasts remain challenged by uncertainties in meteorology, emissions, nonlinear chemistry and boundary inflow. Whether learned models can jointly evolve particles, reactive gases and weather for 5–10 days without lateral conditions remains unresolved. Here we present CWAI, an operational data-driven model advancing a 166-variable atmospheric state on a 15-km grid at 6-h intervals for 10 days. Across 62 initializations in 2020, CWAI reduced root-mean-square error for six pollutants by 15.8–25.8% relative to the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem). Improvements persisted to day 10 and through $PM_{2.5}$ and O_3 profiles from 30 to 3,000 m; particulate and SO_2 advantages also persisted at the domain edge without lateral forcing. Across 60 forecasts in July 2025 and January 2026, gridded errors for $PM_{2.5}$, PM_{10}, O_3, NO_2 and CO fell by 16.4–33.7% relative to China Meteorological Administration Chemistry–Weather Coupled Model (CMA-CW), while independent-station errors for all six fell by 35.1–67.8%. Since May 2026, CWAI has operated twice daily. Post-startup inference and output take 1.5 min and the active analysis-to-product chain takes ~ 70 min. CWAI shows that a learned joint atmospheric state can provide rapid operational guidance for multiple pollutants while retaining three-dimensional chemical and meteorological structure at medium range.

Introduction

Air pollution evolves through a coupled system of emissions, advection, turbulent mixing, deposition, photochemistry, aerosol thermodynamics, and meteorological feedbacks. Forecasts therefore need to resolve not only the movement of existing pollution but also the production and loss of chemically reactive species under changing weather. Online coupled models such as the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem) represent these interactions in a physically interpretable framework and remain central to atmospheric-composition research and forecasting[1–5]. Their skill, however, is sensitive to meteorological initial conditions, emission inventories, chemical boundary conditions, data assimilation and parameterized sub-grid processes[6–8]. The computational cost of repeatedly integrating these components also constrains forecast resolution, horizon and ensemble size.

The lateral boundary is a defining constraint of limited-area prediction. Conventional regional models require time-evolving meteorological inflow, and boundary errors can propagate into the interior as integration proceeds[9]. Atmospheric composition adds a second dependence because transported background concentrations and episodic external plumes can strongly affect regional ozone (O_3) and carbon monoxide (CO) [10]. A recent learned limited-area weather system reported only 12 h of effective forecasting without imposed lateral conditions and used global artificial-intelligence (AI) forecasts to extend its rollout[11], while Zeeman supplies three-dimensional boundary values to regional atmospheric-chemistry forecasts[12]. Whether a learned joint chemical-weather state can sustain useful continental-scale open-boundary behavior without externally supplied future lateral forcing has not been established.

Data-driven models provide another approach for Earth-system prediction[13,14]. Pangu-Weather, GraphCast and FuXi demonstrated that learned models can advance high-dimensional atmospheric states with competitive medium-range skill[15–17], while WeatherBench and WeatherBench 2 established common evaluation frameworks for data-driven weather models[18,19]. These advances show that multiscale neural representations can learn coherent spatial evolution and that autoregressive models can retain useful information over many forecast steps. Atmospheric composition presents an additional challenge because pollutant distributions are more intermittent and emission-sensitive than many meteorological variables, while reactive gases and particles are coupled through processes that are not directly observed at every grid point.

Recent studies have begun to address this gap. Regional deep-learning systems have improved 3-day forecasts of particulate matter with aerodynamic diameter ≤ 2.5 μm ($PM_{2.5}$) when supplied with numerical weather predictions and preceding observations[20]. The machine-learning-driven Global Aerosol–Meteorology Forecasting System (AI-GAMFS) provides 5-day global forecasts of aerosol optical and surface components jointly with meteorology and has been evaluated against independent aerosol networks[21]. FuXi-Air predicts six pollutants at monitoring sites for 72 h by combining emissions, meteorology, and station observations[22], while Zeeman extends regional atmospheric-chemistry forecasts to 5 days[12]. SynCast further demonstrates probabilistic multi-pollutant forecasting through stochastic neural sampling[23]. However, no regional learned system has yet been evaluated for joint 10-day forecasts of inhalable particulate matter (PM_{10}), ozone, nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and carbon monoxide over a national domain, including behavior near an unforced lateral boundary and transfer across seasons, numerical comparators and verification sources.

Answering this question requires more than comparison with the gridded product from which a model state was derived. A learned model can appear skillful because it reproduces the statistical properties of a reference field while shared errors remain hidden. Numerical baselines also differ in initialization, chemistry, resolution, and operational maturity. We therefore evaluated CWAI in three settings: a consistently configured WRF-Chem benchmark in winter and summer 2020, contemporary forecasts from the China Meteorological Administration Chemistry–Weather Coupled Model (CMA-CW) in July 2025 and January 2026, and independent national surface observations. Statistical inference is based on forecast initializations rather than the much larger number of correlated grid cells or station-hours.

Here we test the hypothesis that jointly evolving meteorology and three-dimensional composition allows a learned atmospheric state to preserve chemical-weather predictability beyond the horizons commonly studied in regional air-quality machine learning, including near an unforced lateral boundary. Chemical Weather Artificial Intelligence (CWAI) uses two consecutive states, emissions and static surface descriptors, in a hierarchical encoder-decoder based on shifted-window Transformers and their scaled Swin Transformer V2 formulation[24,25]. Forty recursive 6-h steps produce a 10-day forecast. We evaluate horizontal pollution events, vertical chemical structure, distance from the computational boundary, two numerical comparators, independent monitoring stations, and high-pollution regimes. The same model

has also progressed from retrospective evaluation to twice-daily forecast production. This design tests when a learned joint atmospheric state remains predictive across forecast ranges, seasons and verification sources, and whether that skill can be used in routine operations.

Results

A learned joint state carries chemical weather to 10 days

CWAI treats chemical weather as a single evolving atmospheric state rather than a collection of independent pollutant regressions (Fig. 1). On a 15-km East Asian grid, it jointly advances meteorology, particles and reactive gases in 166 channels. Eleven variables extend through 14 model-height levels from 30 to 7,500 m, while additional channels resolve surface meteorology and composition. Two consecutive states provide the recent atmospheric tendency, and emissions and static surface descriptors condition each forecast step on source and land-surface context.

A hierarchical shifted-window Transformer compresses this state across spatial scales and reconstructs the atmosphere 6 h later. Applying the same frozen model recursively produces 40 steps to 240 h, allowing pollutant evolution and meteorological transport to remain coupled throughout the forecast.

We tested this joint-state forecast in three complementary settings. A 62-start WRF-Chem comparison spans winter and summer 2020; contemporary gridded CMA-CW forecasts provide a second numerical benchmark; and national monitoring stations provide an observation-based test across 60 starts in July 2025 and January 2026. Their differences in period, comparator and verification source make convergence among them more informative than any single pooled score.

Skill persists across pollutants and lead times

CWAI produced lower initialization-weighted root-mean-square error (RMSE) than WRF-Chem for all six evaluated surface pollutants across the 62 January and July 2020 starts (Fig. 2). In the paired combined-period analysis, reductions were 15.8% for $PM_{2.5}$ (95% confidence interval, 12.3–18.8%), 23.1% for PM_{10} (18.1–27.0%), 22.3% for O_3 (21.2–23.5%), 25.8% for NO_2 (21.9–29.0%), 20.1% for SO_2 (17.0–23.0%) and 19.4% for CO (16.0–22.0%). Across all six pollutants, CWAI had lower initialization RMSE in 94–100% of starts.

The advantage persisted across the 6–240-h forecast sequence rather than arising from one short lead band (Fig. 2d-i). Lead-time curves retained pronounced sub-daily structure, particularly for ozone and the meteorological variables, indicating that the lower error was not produced by an artificially flat climatological forecast. The size of the reduction varied by species and season. PM_{10} and NO_2 gains were especially large in July, whereas winter $PM_{2.5}$ was more difficult. This heterogeneity is expected for pollutants controlled by different combinations of primary emissions, photochemistry and boundary-layer mixing.

The 2020 benchmark covers January and July and uses the validation year; the 2025–2026 experiments provide a separate test of temporal transfer. Meteorological performance was mixed (Supplementary Fig. 1). CWAI improved water vapor, 10-m wind speed, downward shortwave radiation, boundary-layer height and winds at 100 and 150 m, but 2-m temperature RMSE was 18.9% higher than WRF-Chem in January and 4.1% higher in July. Sparse meteorological assimilation constraints, particularly aloft, and the limited training archive are plausible contributors, but establishing causality will require controlled retraining experiments.

Pollution episodes retain horizontal and vertical structure

Domain-average statistics can conceal a forecast that has the right mean concentration but places a plume in the wrong region or layer. We therefore examined severe $PM_{2.5}$ and ozone episodes at fixed + 24-h and + 30-h leads, selecting cases from the upper decile of reference pollution without using either model's performance. In the $PM_{2.5}$ event, CWAI reduced domain RMSE from 44.4 to 31.2 μg m^{-3}; in the ozone event, it reduced RMSE from 57.8 to 42.6 μg m^{-3} (Fig. 3a–j). The maps retain the dominant pollution gradients and event locations, while their error fields reveal local displacement and amplitude errors that are invisible in a domain mean.

Because CWAI evolves a three-dimensional chemical state, the same events provide a stronger test than surface maps alone. Meridional sections show the reference near-surface $PM_{2.5}$ maximum and the deeper ozone structure extending through the lower troposphere (Fig. 3k–m,o–q). CWAI reproduces these dominant vertical gradients, although neither event is locally exact. Across all 62 starts and 6–240-h leads, its mean RMSE was lower than WRF-Chem at every evaluated height from 30 to 3,000 m, with reductions of 10.4–16.3% for $PM_{2.5}$ and 8.5–22.3% for O_3 (Fig. 3n,r). The evidence therefore extends beyond surface concentration skill, but remains verification against a gridded reference rather than independent vertical observations.

To assess whether these two cases were representative, spatial RMSE was aggregated at + 24, +72, + 120, +168 and + 240 h over all 62 starts showed that most, but not all, grid cells improved (Fig. 4). Across all six pollutants, median grid-cell reductions ranged from 5.8% for NO_2 to 10.1% for PM_{10} in January and from 3.0% for SO_2 to 16.3% for PM_{10} in July. The improved-area fraction ranged from 59.6% to 69.7% in January and from 52.4% to 69.0% in July.

The July NO_2 and SO_2 maps contained coherent regions of degradation. We therefore emphasize median grid-cell skill and the improved-area fraction rather than the spatial mean of percentage skill, which can be unstable where baseline RMSE approaches zero. Six predeclared regional aggregates provided an additional test (Supplementary Fig. 5). All 60 region-month-pollutant summaries had positive median grid-cell skill in the 2020 archive, but the station evidence was less uniform: 26 of 30 regional intervals were above zero, two crossed zero, and Guangdong $PM_{2.5}$ (− 5.5%, 95% interval − 11.0 to − 0.4%) and O_3 (− 14.7%, − 21.8 to − 8.1%) were negative. The aggregate advantage is therefore geographically broad but not universal.

Learned states partially sustain open-boundary behavior

Unlike a limited-area numerical integration, the CWAI rollout receives no time-evolving meteorological or chemical fields at its lateral boundaries. Every grid cell, including the outermost cells, is advanced recursively from the initialized joint state. We therefore stratified the 2020 evaluation domain by distance to the nearest computational boundary and compared CWAI and WRF-Chem at daily leads to 240 h (Supplementary Fig. 7).

When all daily leads were pooled, CWAI reduced RMSE within the outermost 300 km by 10.1% for $PM_{2.5}$ (95% interval 7.5–12.7%), 12.4% for PM_{10} (10.1–14.5%) and 16.9% for SO_2 (12.1–21.0%). The corresponding intervals for O_3, NO_2 and CO crossed zero. At day 10, the edge-band reductions remained 8.2% for $PM_{2.5}$ (3.2–13.0%), 10.7% for PM_{10} (6.2–15.0%) and 16.4% for SO_2 (8.7–22.5%). In contrast, O_3 and NO_2 degraded by 5.2% (2.0–8.7%) and 14.9% (6.1–25.0%), respectively, while CO remained inconclusive. Beyond 600 km from the boundary, day-10 O_3 skill was again positive, and most NO_2 and CO distance bands also favored CWAI.

Error did not increase uniformly at the domain edge, although the forecasts were not boundary independent. The initialized joint state retained effective open-boundary behavior more consistently for particles and sulfur dioxide than for reactive and transported gases. The sharper winter edge degradation for O_3 and NO_2 further indicates that this capacity depends on chemical regime rather than distance alone.

Independent stations expose species-specific limits

National monitoring stations provide a verification target independent of either gridded forecast system (Fig. 5 and Supplementary Fig. 2). The July 2025 and January 2026 cohorts contain more than 1.6 million unique station-hour records from over 1,000 sites in each month.

Across 60 forecast starts, CWAI reduced station-verified RMSE relative to CMA-CW by 48.4% for $PM_{2.5}$ (95% interval 37.0–59.2%), 35.3% for PM_{10} (26.4–42.9%), 42.6% for O_3 (39.6–45.5%), 35.1% for NO_2 (30.7–39.9%), 57.6% for SO_2 (52.3–62.2%), and 67.8% for CO (62.9–72.0%). Between 98% and 100% of starts had lower CWAI RMSE, depending on pollutant. The independent observations therefore support improvement across all six pollutants, while the magnitude remains pollutant-dependent. Both forecast systems report gas mass concentrations calculated with real-time model temperature and are compared on this common ambient-state basis. Taken together, the independent-station reductions span 35.1–67.8% across all six pollutants, matching the cross-tier synthesis in Fig. 6c.

Skill transfers but remains regime dependent

On matched Chinese-land gridded fields across the same 60 contemporary starts, CWAI RMSE reductions relative to CMA-CW were 33.7% for $PM_{2.5}$ (25.7–40.3%), 33.6% for PM_{10} (27.1–38.5%), 22.4% for O_3 (20.4–24.4%), 18.8% for NO_2 (16.6–21.2%), − 1.0% for SO_2 (− 9.0–6.8%) and 16.4% for corrected

CO (14.1–18.5%) (Fig. 6b). The confidence intervals were above zero for five pollutants, whereas the gridded sulfur dioxide difference was unresolved. Grid and station evidence agreed for the other species, but the much stronger station sulfur dioxide result shows that conclusions can depend on the verification target. Station and grid reductions are not expected to be identical because spatial representativeness, interpolation and observation error differ.

Skill weakened with range but remained positive at the cross-pollutant median in all three evidence tiers. From 6–48 h to 126–228 h, median RMSE reduction changed from 23.1% to 20.1% in the retrospective benchmark, from 31.8% to 15.5% on contemporary grids and from 56.0% to 39.4% at independent stations (Fig. 6d). Start-wise consistency was 94–100% retrospectively and 98–100% at stations. On contemporary grids it was also 98–100% for five pollutants, while SO_2 improved in only 62% of starts (Fig. 6e), identifying winter SO_2 as the principal exception rather than a general loss of forecast skill.

Seasonal transfer was broad but not uniform: every pollutant-tier pair except gridded SO_2 retained positive RMSE reduction in both July 2025 and January 2026 (Fig. 6f). Skill for particulate matter strengthened in January in both verification tiers, whereas relative gains for O_3, NO_2 and CO generally weakened; SO_2 diverged, improving further at stations but reversing on gridded fields. The weak combined SO_2 estimate concealed a marked seasonal and lead-time transition. Gridded SO_2 RMSE fell by 29.0% in July but increased by 14.6% in January. January skill remained positive through 48 h (25.8%) before a growing positive CWAI bias produced a 34.7% degradation at 126–228 h. Parameter mapping, units, UTC validity, and common-mask counts were independently checked, supporting winter autoregressive drift rather than a processing error as the source of the reversal.

Two stricter tests examined performance under more demanding conditions. During station observations at or above pollutant- and cohort-specific P90 thresholds, CWAI reduced RMSE by 42.9% for $PM_{2.5}$, 24.2% for PM_{10}, 34.8% for NO_2 and 33.5% for SO_2. Extreme ozone improvement was only 4.8% (1.9–7.7%); it was positive in July 2025 but negative in January 2026 (− 9.1%). CWAI increased ozone critical success index from 0.250 to 0.267, showing a small event-detection gain rather than a decisive resolution of extreme ozone (Supplementary Fig. 4).

Finally, we compared CWAI with observation persistence on a strict common mask that also required the initialization-time observation to be present. Combined-period RMSE was lower than persistence for PM_{10}, O_3 and NO_2, with paired intervals above zero. $PM_{2.5}$ (− 0.3%), SO_2 (− 1.5%) and CO (8.3%, 95% interval − 0.8 to 16.9%) were inconclusive, and January 2026 SO_2 was 20.6% worse than persistence (95% interval − 42.2 to − 0.1%) (Supplementary Fig. 6). CWAI therefore provided dynamic forecast value for several pollutants, but did not consistently outperform persistence for every species and season.

From learned predictability to operational forecasting

CWAI was subsequently integrated into a repeated production cycle. It now ingests the latest chemical-weather analysis and issues forecasts twice daily for 00:00 and 12:00 UTC initializations. Between 6 May

and 29 July 2026, all 3,220 retained outputs carried the expected UTC validity. The local conversion-and-inference workflow completed in a median of 9.7 min (10th–90th percentile, 3.7–16.7 min). The logged inference stage, including process and Open Neural Network Exchange (ONNX) session initialization, required a median of 99 s; after session initialization, inference and output required 92 s (60–194 s) across the retained phase-resolved logs.

The operational latency is dominated by production of the input analysis rather than neural forecasting. Regional analysis fields became available a median of 59.8 min after their nominal valid time. Combining that production time with the measured local workflow gives an active computational chain of 69.7 min. Under the present conservative schedule, the local workflow begins 85 min after initialization, leaving an approximately 25-min buffer after the input field is available; complete 5-day products were consequently issued after a median 95.2 min (88.9–102.0 min). The current release latency could therefore approach 70 min if the workflow were triggered directly by input availability, without changing the forecast model.

The 120-h production horizon is a deliberate service configuration chosen to balance present user requirements, storage and routine computational demand; it is not an architectural or learned-predictability limit. The same autoregressive model was evaluated for 40 steps to 240 h. The operational record is independent of the retrospective skill comparisons and demonstrates repeated forecast production rather than prospective forecast skill. A normalized speed comparison with WRF-Chem or CMA-CW would require matched hardware and product definitions.

Discussion

Across the three evaluation settings, CWAI retained useful multi-pollutant skill from the surface through the lower troposphere. It produced lower errors than WRF-Chem in the winter and summer 2020 archive, transferred to contemporary CMA-CW comparisons, agreed with independent surface observations, and reproduced coherent fine-particle and ozone profiles through 3 km. Agreement across these tests makes a comparator-, season- or verification-specific artefact less likely.

The model is also part of a monitored, twice-daily forecast-production chain rather than solely an offline experiment. Once an analysis is available, the neural forecast is not the dominant source of latency; most elapsed time lies in regional analysis production, conversion and a conservative scheduling buffer. Triggering the workflow when its input becomes available could reduce the present 95-min median issue latency to approximately 70 min without changing the learned model. The current 5-day product horizon reflects operational demand, storage and routine computational constraints, whereas the experiments evaluate the same model to 10 days.

CWAI differs from station-only air-quality predictors because it advances a complete gridded atmospheric state. It also differs from aerosol-focused systems by including reactive gases and their vertical structure. AI-GAMFS has established the feasibility of global 5-day aerosol-meteorology prediction[21], while FuXi-Air, Zeeman and SynCast demonstrate high-resolution multi-pollutant, regional-

chemistry or probabilistic skill over shorter horizons[12,22,23]. CWAI occupies a complementary regime: a 10-day, 6-hourly, national-scale regional forecast with particles, gases and meteorology evaluated against two coupled numerical systems and surface observations.

Several architectural and data choices plausibly contribute to the retained skill. The joint state allows meteorological transport and pollutant evolution to share a multiscale representation. Two consecutive states provide local temporal tendency, while emission and static channels condition the forecast on source and surface context. The hierarchical encoder-decoder can combine broad transport structures with regional gradients.

The boundary experiment probes a basic constraint of regional forecasting. Numerical weather and chemistry systems ordinarily import time-varying lateral fields, and learned limited-area systems have generally retained that dependence[9–12]. CWAI instead maintained lower edge error than WRF-Chem for particulate matter and SO_2 through day 10 using only its initialized atmospheric states and prescribed auxiliary fields. A plausible interpretation is that the large domain, two initial states and jointly evolved meteorology and composition encode enough transport tendency and conditional background structure to emulate part of the required inflow. This does not imply that lateral influence has disappeared: the initial states and training targets already contain the effects of their parent numerical system, and the day-10 O_3 and NO_2 degradation shows that learned conditional inflow cannot fully replace evolving external information. The distance bands also pool inflow and outflow sectors and diagnose error growth rather than isolating a boundary mechanism. Unexpected transboundary events and controlled forecasts with alternative boundary treatments should provide decisive future tests. The result indicates that learned joint states can partially relax, but do not remove, the lateral-boundary constraint.

Mean RMSE alone did not capture important aspects of chemical-weather performance. Spatial maps reveal degradation that disappears in national averages. The Guangdong $PM_{2.5}$ and O_3 results and the January 2026 SO_2 persistence comparison identify species-region-season failures. Extreme ozone provides a second boundary: mean ozone forecasts improve strongly, but upper-decile RMSE gains are small and reverse sign in one month. Severe pollution is shaped by nonlinear secondary formation and meteorological accumulation[26], while deterministic systems trained with mean-squared error (MSE) tend to favor central distributions. Probabilistic objectives, tail-aware losses and ensemble generation may improve forecasts in these regimes[23].

The numerical comparisons should not be interpreted as proof that data-driven models can replace coupled chemistry models. WRF-Chem and CMA-CW encode mass transport, chemical mechanisms, deposition and aerosol feedbacks that support diagnosis, emission scenarios and scientific attribution. CWAI currently provides a deterministic forecast conditioned on historical relationships and prescribed auxiliary fields; it does not enforce elemental conservation or expose reaction pathways. CWAI is therefore best used as complementary guidance: for rapid deterministic updates, ensemble members, quality control, or as a learned forecast component combined with numerical analyses.

Future work should prioritize prospective probabilistic forecasts, broader observational constraints, and controlled ablations of previous-state and emission inputs. Direct comparisons of speed and energy require normalized hardware and forecast-product definitions. The present experiments nevertheless show that learned joint atmospheric states can extend multi-pollutant chemical-weather guidance to 10 days. Their value is most clearly established by independent, species-resolved evaluation that also reports the conditions under which skill degrades.

Methods

CWAI state representation

The model grid contains 300 × 360 cells. The 166 predicted channels comprise 11 variables at 14 model-height levels and 12 surface variables. Heights are 30, 50, 100, 150, 300, 500, 750, 1,000, 1,500, 2,000, 3,000, 4,000, 5,000 and 7,500 m. Three-dimensional variables are pressure, air temperature, water-vapor mixing ratio, zonal wind, meridional wind, $PM_{2.5}$, PM_{10}, O_3, SO_2, NO_2 and CO. Surface variables are 2-m temperature, 2-m water-vapor mixing ratio, surface pressure, 10-m wind speed, downward shortwave radiation, planetary boundary-layer height and surface $PM_{2.5}$, PM_{10}, O_3, SO_2, NO_2 and CO.

Model architecture and inference

The model uses a three-dimensional convolutional patch embedding with kernel and stride 2 × 4 × 4. After collapsing the two-time dimension, a three-stage Swin Transformer V2 encoder applies local-window self-attention[24,25] with depths 12, 6 and 2 and 8, 16 and 24 heads. Convolutional residual downsampling doubles channel width between stages. A symmetric decoder applies transposed-convolution upsampling, concatenates encoder skip features, and recovers the forecast through a 4 × 4 transposed convolution followed by residual convolutions. The embedding dimension is 1,152, the local attention window is 7, the multilayer-perceptron ratio is 4, and the drop-path rate is 0.1.

The evaluated model contains approximately 2.56 billion parameters. It predicts the atmospheric state 6 h ahead; this prediction and the latest input state then condition the next recursive step. Forty steps yield a 240-h forecast. Contemporary comparisons use leads through 228 h, the common horizon available from CMA-CW.

Lateral-boundary treatment and distance analysis

No forecast-time meteorological or chemical lateral fields are supplied during recursive inference. Predicted values at the domain edge are neither replaced, nudged nor blended with an external forecast; every cell in the 300 × 360 domain is advanced by the same learned mapping. The two initialized atmospheric states nevertheless contain the accumulated influence of the numerical system from which they were derived, and the training targets reflect that system’s boundary forcing. We therefore describe the rollout as lacking externally supplied time-evolving lateral forcing, not as physically boundary independent.

For the boundary-distance analysis, each cell was assigned its center-to-edge distance to the nearest side of the rectangular computational grid using the 15-km spacing. Cells were grouped into 0–300, 300–600, 600–900, 900–1,200 and greater than 1,200-km bands. CWAI and WRF-Chem were evaluated against the same gridded reference and common finite mask for six surface pollutants across all 62 January and July 2020 initializations at daily leads from 24 to 240 h. Squared errors and valid counts were pooled within each initialization and distance band. Skill was expressed as the CWAI-to-WRF-Chem RMSE ratio and its relative reduction; uncertainty used the paired month-stratified circular-block bootstrap described below.

Operational implementation

CWAI was integrated into an automated forecast-production system initialized at 00:00 and 12:00 UTC. The service performs 20 recursive steps to produce a 120-h forecast, a horizon selected to balance current user requirements, storage, and routine computational demand. Operational continuity was assessed from retained products issued between 6 May and 29 July 2026. Of 169 scheduled cycles, all 3,220 retained fields carried the expected UTC validity.

Runtime was separated into input-analysis production, local format conversion, ONNX session initialization, recursive inference, and product output. Successful local workflows had a median duration of 582 s (10th–90th percentile, 224–1,002 s). The logged inference stage, which launches a new process and ONNX session for each cycle, had a median duration of 99 s (67–363 s). In 82 logs with phase-resolved session timestamps, session initialization required a median of 7.1 s, and the subsequent inference-and-output phase required 91.8 s (60.1–194.5 s).

Input readiness was determined from the modification time of the regional analysis field used to initialize each forecast. Among all successful routinely scheduled cycles, the field became available 3,590 s after nominal initialization, while the sum of input-production and local-workflow durations was 4,184 s. Because the production workflow was scheduled to begin approximately 5,101 s after initialization, complete products were issued after a median of 5,715 s. These records characterize observed production latency and continuity, not independent forecast skill or hardware-normalized throughput.

Training data and optimization

Training samples span 4 February 2013 to 14 December 2019 and 1 January 2022 to 14 June 2023. The year 2020 was reserved for validation and 2021 for testing. Training used two consecutive atmospheric states to predict a 6-h target, with a channel-weighted mean-squared-error objective, AdamW optimization, and cosine learning-rate scheduling.

WRF-Chem baseline

The 2020 baseline used WRF-Chem v4.5.2[1–3]. A single Lambert-conformal domain contained 360 × 300 mass-grid cells at 15-km spacing, with 50 atmospheric levels, a 1,000-Pa model top, a 60-s integration

step, and 6-hourly meteorological boundary input. Forecasts began at 00:00 UTC and extended 10 days with hourly output. The same configuration was used across all 62 January and July initializations.

The chemical configuration combined Carbon Bond Mechanism Z (CBM-Z) gas chemistry[27] with the Model for Simulating Aerosol Interactions and Chemistry (MOSAIC) using four aqueous aerosol bins[28]. Anthropogenic emissions were based on the Multi-resolution Emission Inventory for China (MEIC)[6,7], and biogenic emissions were calculated with the Model of Emissions of Gases and Aerosols from Nature (MEGAN)[29]. Meteorological initial and lateral boundary conditions were derived from the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) at 6-h intervals. Gas and aerosol deposition, wet scavenging, cloud chemistry, vertical mixing, convective transport, and aerosol-radiation feedback were enabled. Physical parameterizations comprised Lin microphysics, Rapid Radiative Transfer Model for General Circulation Models (RRTMG) radiation, the revised fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5) surface layer, Noah land surface, Yonsei University (YSU) boundary layer and Grell three-dimensional ensemble cumulus.

CMA-CW baseline

CMA-CW forecasts were collected for daily 00:00 UTC initializations in July 2025 and January 2026. The archive uses a 0.1° regular grid with 851 × 551 points and extends to 228 h. The product specification reports $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3 in μg m^{-3}; CO was converted to mg m^{-3} for comparison. Gas mass concentrations are calculated using time-varying model temperature and were therefore treated as ambient-state concentrations.

CMA-CW couples the China Meteorological Administration Mesoscale model (CMA-MESO) v5.1 and the Chinese Unified Atmospheric Chemistry Environment (CUACE) model online[30–32]. The published system includes Piecewise Rational Function Method (PRM) tracer transport, Carbon Bond 2005 chemistry with chlorine chemistry (CB05CL), the Canada Aerosol Module (CAM), ISORROPIA II thermodynamics, aerosol-cloud-radiation feedback, the CMA-CUACE-EMISSION (CCES) v1.0 inventory, and ensemble optimal interpolation (EnOI) or three-dimensional variational (3DVAR) chemical assimilation.

Gridded evaluation

The retrospective comparison includes 31 January and 31 July initializations at 6-h leads through 240 h. Pollutant squared-error sums and common finite-sample counts were calculated from CWAI, WRF-Chem, and reference fields over 41,173 China-domain grid cells. These quantities were pooled across all 40 leads within each initialization before RMSE calculation.

For the 2025/2026 comparison, CMA-CW and CWAI were aligned by initialization, lead, and valid UTC time. Both forecasts were scored against the same gridded reference and identical finite mask for every variable and pair. Additive squared error and valid count were accumulated, enabling exact initialization-

level RMSE reconstruction. Sixty starts passed the common-data protocol. Forecast starts were equally weighted.

Station observations and collocation

National air-quality observations provided 907,391 station-hour records from 1,264 stations for the July 2025 cohort and 781,726 records from 1,066 stations for January 2026. Cohorts were defined by forecast-initialization month, with verification extending to 9 August and 9 February, respectively. All station-time keys within these windows were unique.

WRF-Chem gas mixing ratios were converted to mass concentrations using local air temperature and pressure. CWAI inherits this ambient-state convention from its training targets, and CMA-CW production information confirms the use of time-varying model temperature. Both forecasts were therefore compared as delivered on a common ambient-state basis. The observation-processing record indicates a 298.15-K reporting convention after September 2018, but collocated site-level temperature and pressure were unavailable; no asymmetric transformation was applied to either forecast system. WRF-Chem aerosol mass mixing ratios are dry-air based. A sensitivity correction using archived 30-m water-vapor mixing ratio changed headline PM reductions by at most 0.49 percentage points and did not alter conclusions.

Extreme events, regions and persistence

Extreme thresholds were calculated separately for each pollutant and evaluation cohort from unique observations at all valid times entering the matched analysis, including leads crossing month boundaries. Samples at or above P90 were then scored on an identical forecast-observation mask. Hits, misses and false alarms used the same fixed threshold to derive probability of detection, false-alarm ratio and critical success index.

Six non-overlapping province unions were declared before regional analysis: Beijing-Tianjin-Hebei, Yangtze River Delta, Guangdong, Sichuan-Chongqing, Northeast China and Northwest China. Spatial skill used median grid-cell RMSE reduction and improved area fraction. Station regional skill used the same initialization-level aggregation as the national analysis.

Persistence held the observation at initialization fixed at each station for future leads. Persistence, CMA-CW, CWAI, and verification observations were all required to be finite, producing a stricter mask than the main station analysis. The persistence result is therefore a sensitivity benchmark and does not replace the paired CMA-CW comparison.

Statistical analysis

The primary metric was relative reduction in equal-initialization-weighted RMSE,

$$100 \times \frac{RMSE_{baseline} - RMSE_{CWAI}}{RMSE_{baseline}}$$

Uncertainty was estimated with 5,000 paired circular-block bootstrap resamples. Blocks contained 10 consecutive daily initializations, matching the approximate overlap imposed by a 10-day forecast, and resampling was stratified by month. Two-sided 95% percentile intervals are reported; an independent and identically distributed (IID) initialization bootstrap was retained as a sensitivity analysis. The start-wise consistency statistic in Figs. 2c and 6e is the percentage of forecast initializations for which CWAI had lower RMSE than the paired baseline after pooling all retained leads within that initialization. It is a win fraction, not a reliability or calibration score. For headline comparisons, one-sided non-positive probabilities used a plus-one correction and Benjamini-Hochberg adjustment within each evidence tier. Effect sizes and intervals are primary; adjusted probabilities are secondary. Grid cells, station-hours and observations reused by overlapping forecasts were not treated as independent replicates.

Declarations

Competing interests

The authors declare no competing interests.

Author Contribution

X.Z. and J.Z. conceived and designed the study. J.Z. led the development, training and operational deployment of CWAI, conducted the experiments and analyses, prepared the figures, and drafted the manuscript. X.Z. supervised the study, guided the research design and interpretation of the results, and revised the manuscript. Z.J.Z., Z.L., T.Z., Z.S.Z., W.X., L.J., Y.Z. and Y.M. contributed to the development and operational production of CWAI and provided data and technical support. W.Z., D.W., L.G. and W.J. provided the CMA-CW forecast data and contributed to the comparative evaluation and interpretation of the results. J.S. and Y.X. provided computational resources and technical support. All authors discussed the results, reviewed and edited the manuscript, and approved the final version.

Acknowledgements

Supported by State Key Laboratory of Severe Weather Meteorological Science and Technology(2026QZA17)and National Natural Science Foundation of China (42090030)

Data availability

The regional chemical-weather fields used to train and evaluate CWAI were from National Meteorological Information Centre, CMA. The archive contains the meteorological, chemical, emission-related and static geographical variables described in the Methods and Supplementary Information. Meteorological initial and lateral boundary conditions used for the WRF-Chem simulations were obtained from ECMWF IFS. Anthropogenic emissions were obtained from MEIC. Contemporary CMA-CW forecasts for July 2025 and January 2026 were obtained from the operational archive of CMA. Surface observations of $PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2 and CO were obtained from the national air-quality monitoring network operated by the China National Environmental Monitoring Centre of the Ministry of Ecology and Environment of China.

Code availability

The CWAI inference code, frozen model weights, model configuration, normalization metadata and example workflows required to reproduce the forecasts will be released in a versioned public repository upon publication.

References

1. Grell, G. A. *et al.* Fully coupled “online” chemistry within the WRF model. *Atmos. Environ.* 39, 6957–6975 (2005).
2. Fast, J. D. *et al.* Evolution of ozone, particulates, and aerosol direct radiative forcing in the vicinity of houston using a fully coupled meteorology-chemistry-aerosol model. *J. Geophys. Res.: Atmos*. 111, D21305 (2006).
3. Powers, J. G. *et al.* The weather research and forecasting model: overview, system efforts, and future directions. *Bull. Am. Meteorol. Soc.* 98, 1717–1737 (2017).
4. Baklanov, A. *et al.* Online coupled regional meteorology chemistry models in Europe: current status and prospects. *Atmos. Chem. Phys.* 14, 317–398 (2014).
5. Zhang, Y. Online-coupled meteorology and chemistry models: history, current status, and outlook. *Atmospheric Chemistry and Physics* 8, 2895–2932 (2008).
6. Li, M. *et al.* MIX: a mosaic asian anthropogenic emission inventory under the international collaboration framework of the MICS-asia and HTAP. *Atmos. Chem. Phys.* 17, 935–963 (2017).
7. Zheng, B. *et al.* Trends in China’s anthropogenic emissions since 2010 as the consequence of clean air actions. *Atmos. Chem. Phys.* 18, 14095–14111 (2018).
8. Bocquet, M. *et al.* Data assimilation in atmospheric chemistry models: current status and future prospects for coupled chemistry meteorology models. *Atmos. Chem. Phys.* 15, 5325–5358 (2015).
9. Warner, T. T., Peterson, R. A. & Treadon, R. E. A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bull. Am. Meteorol. Soc.* 78, 2599–2618 (1997).
10. Tang, Y. *et al.* Influence of lateral and top boundary conditions on regional air quality prediction: a multiscale study coupling regional and global chemical transport models. *J. Geophys. Res.: Atmos.*

112, D10S18 (2007).

11. Xu, P. *et al.* An artificial intelligence-based limited area model for forecasting of surface meteorological variables. *Commun. Earth Environ.* 6, 372 (2025).
12. Pang, M. *et al.* Zeeman: a deep learning framework for regional atmospheric chemistry forecasting. *Geophys. Res. Lett.* 53, e2025GL117459 (2026).
13. Reichstein, M. *et al.* Deep learning and process understanding for data-driven earth system science. *Nature* 566, 195–204 (2019).
14. Schultz, M. G. *et al.* Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379, 20200097 (2021).
15. Bi, K. *et al.* Accurate medium-range global weather forecasting with 3D neural networks. *Nature* 619, 533–538 (2023).
16. Lam, R. *et al.* Learning skillful medium-range global weather forecasting. *Science* 382, 1416–1421 (2023).
17. Chen, L. *et al.* FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.* 6, 190 (2023).
18. Rasp, S. *et al.* WeatherBench: a benchmark data set for data-driven weather forecasting. *JAMES* 12, e2020MS002203 (2020).
19. Rasp, S. *et al.* WeatherBench 2: a benchmark for the next generation of data-driven global weather models. *JAMES* 16, e2023MS004019 (2024).
20. Qiu, Y. *et al.* Regional aerosol forecasts based on deep learning and numerical weather prediction. *npj Clim. Atmos. Sci.* 6, 71 (2023).
21. Gui, K. *et al.* Advancing operational global aerosol forecasting with machine learning. *Nature* 651, 658–665 (2026).
22. Geng, Z. *et al.* FuXi-air: air quality forecasting based on emission-meteorology-pollutant multimodal machine learning. *npj Clean Air* 2, 21 (2026).
23. Abeysinghe, Y. *et al.* Synergistic neural forecasting of air pollution with stochastic sampling. *npj Clean Air* 2, 36 (2026).
24. Liu, Z. *et al.* Swin transformer: hierarchical vision transformer using shifted windows. in 9992–10002 (IEEE, 2021). doi:10.1109/ICCV48922.2021.00986.
25. Liu, Z. *et al.* Swin transformer V2: scaling up capacity and resolution. in 11999–12009 (IEEE, 2022). doi:10.1109/CVPR52688.2022.01170.
26. Huang, R.-J. *et al.* High secondary aerosol contribution to particulate pollution during haze events in China. *Nature* 514, 218–222 (2014).
27. Zaveri, R. A. & Peters, L. K. A new lumped structure photochemical mechanism for large-scale applications. *J. Geophys. Res.: Atmos.* 104, 30387–30415 (1999).
28. Zaveri, R. A., Easter, R. C., Fast, J. D. & Peters, L. K. Model for simulating aerosol interactions and chemistry (MOSAIC). *J. Geophys. Res.: Atmos*. 113, D13204 (2008).

29. Guenther, A. B. *et al.* The model of emissions of gases and aerosols from nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions. *Geosci. Model Dev.* 5, 1471–1492 (2012).
30. Liu, C. *et al.* Impact of aerosols on weather forecasts in China during winter 2016–2017. *JAMES* 17, e2024MS004696 (2025).
31. Wang, H. *et al.* The CMA chemistry–weather coupled model and its operational forecast of fog–haze and sand/dust storms. *J. Meteorolog. Res.* 40, 1077–1098 (2026).
32. Zhuang, Z. *et al.* Development and evaluation of a novel 3D variational assimilation framework for regional chemistry–weather forecasting. *J. Meteorolog. Res.* 40, 902–920 (2026).

Figures

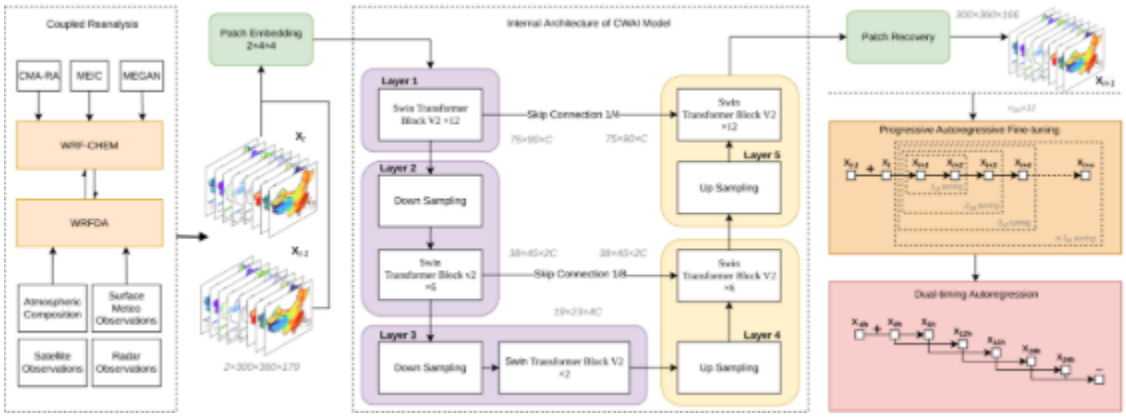


Figure 1

CWAI framework for chemical-weather forecasting. Meteorological analyses, emissions and atmospheric-composition observations are integrated into coupled chemical-weather states. Two consecutive 300 × 360 states are mapped by a 2 × 4 × 4 patch embedding into a hierarchical Swin Transformer V2 encoder-decoder; skip connections and patch recovery reconstruct the 166-channel state 6 h later. Recursive application of the same model yields forecasts to 10 days. The developmental schematics at right place this forecast configuration within the broader model-development strategy; all quantitative results use the 6-h model recursively.

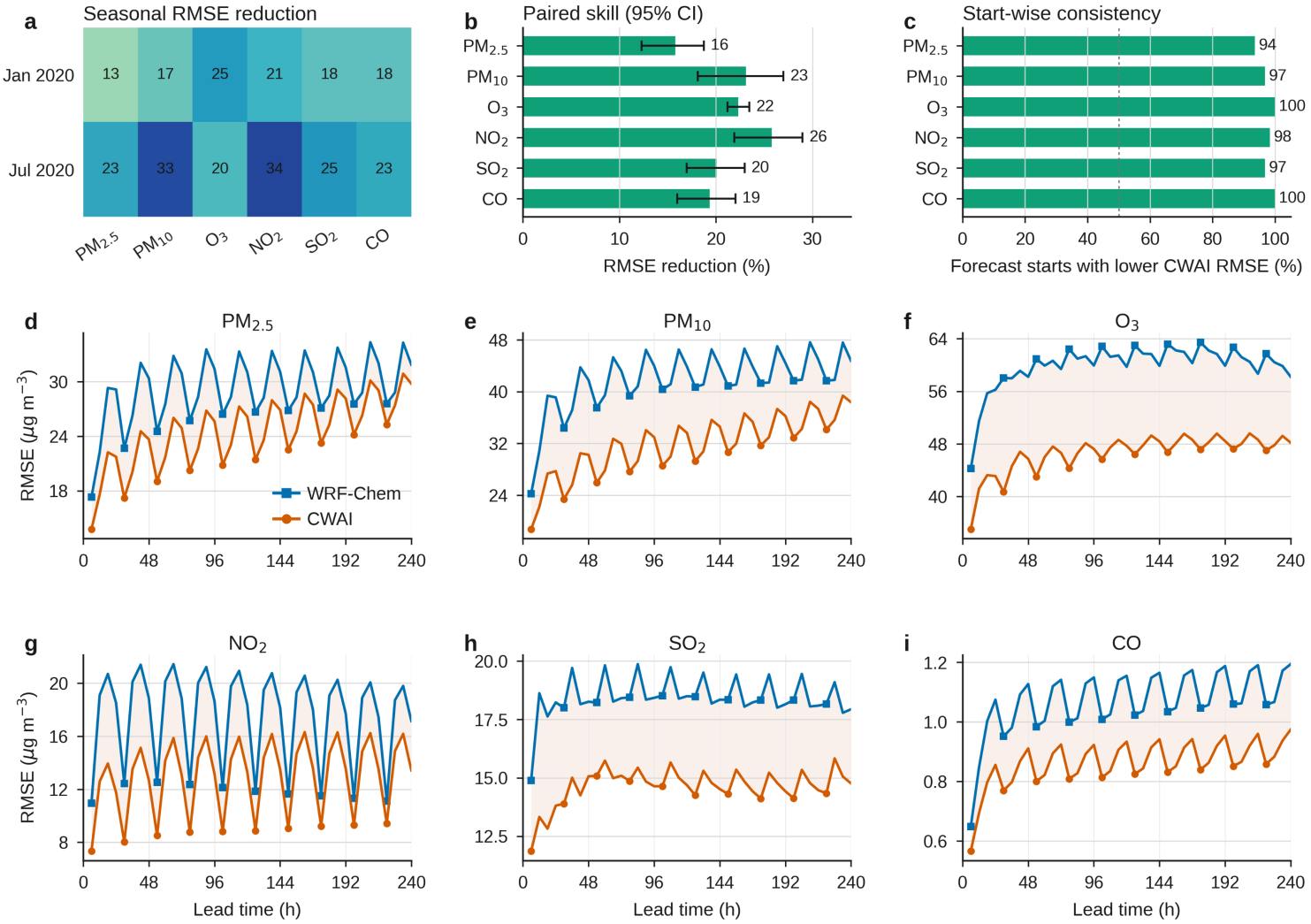


Figure 2

Historical benchmark against WRF-Chem in winter and summer 2020. CWAI was evaluated against WRF-Chem for 62 forecast starts in January and July 2020. **a**, Initialization-weighted RMSE reduction for six surface pollutants in each evaluated month. **b**, Combined-period RMSE reduction with 95% confidence intervals from 5,000 paired bootstrap resamples at the forecast-initialization level. **c**, Fraction of starts with lower CWAI initialization RMSE. **d–i**, Lead-time RMSE curves for $PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2 and CO. Positive reduction indicates lower CWAI RMSE.

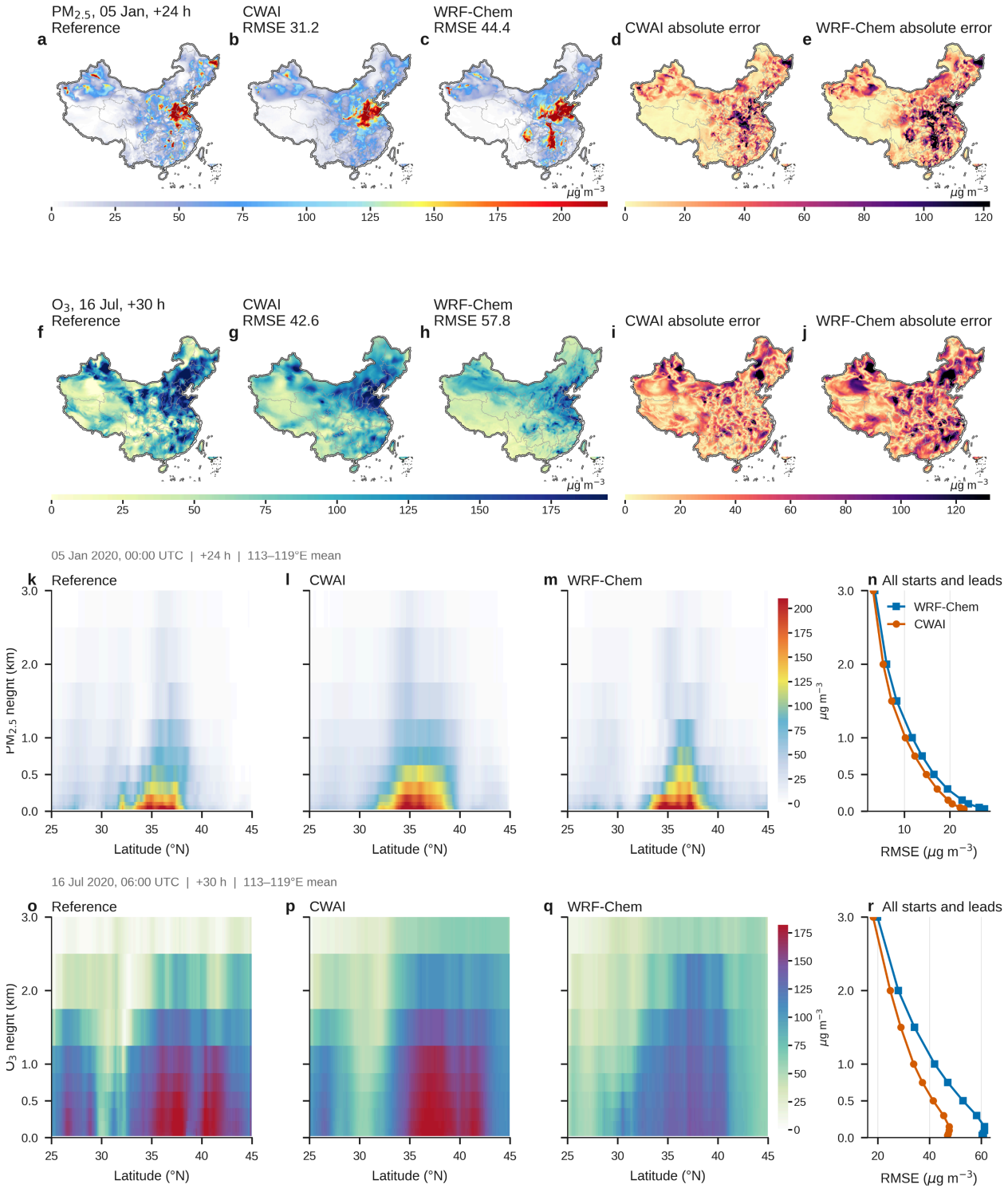


Figure 3

Horizontal and vertical structure during severe pollution episodes. Cases were selected from forecasts whose China-domain reference concentration exceeded the 90th percentile (P90) at fixed leads without using forecast performance. **a–e**, Reference $PM_{2.5}$, CWAI, WRF-Chem and absolute errors at +24 h. **f–j**, Corresponding O_3 fields at +30 h. **k–m,o–q**, Reference, CWAI and WRF-Chem meridional-height sections for the same events, averaged across 113–119° E from 30 to 3,000 m. **n,r**, Mean height-dependent RMSE

across all 62 starts and 6–240-h leads. Concentration and error maps use separate shared scales within each row.

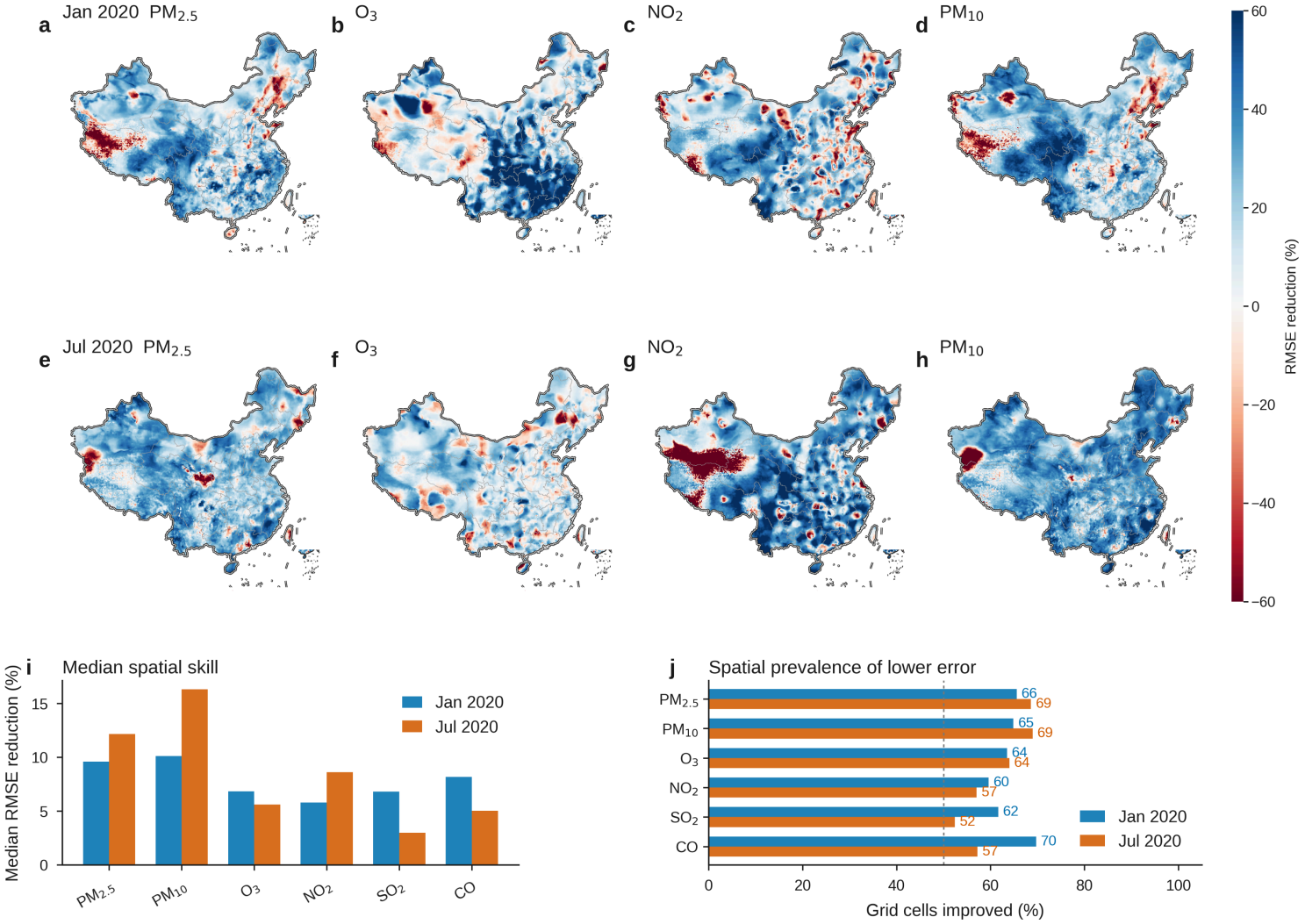


Figure 4

Spatial heterogeneity of pollutant forecast skill. Spatial RMSE reduction was calculated using common finite samples over 62 January and July 2020 starts at +24, +72, +120, +168 and +240 h. **a–h**, Spatial RMSE reduction for $PM_{2.5}$, O_3, NO_2 and PM_{10} in January and July. Blue indicates lower CWAI RMSE and red indicates degradation relative to WRF-Chem. **i**, Month-specific median grid-cell RMSE reduction for six pollutants. **j**, Month-specific fraction of grid cells with lower CWAI RMSE. Medians and improved-area fractions show broad but non-uniform improvement.

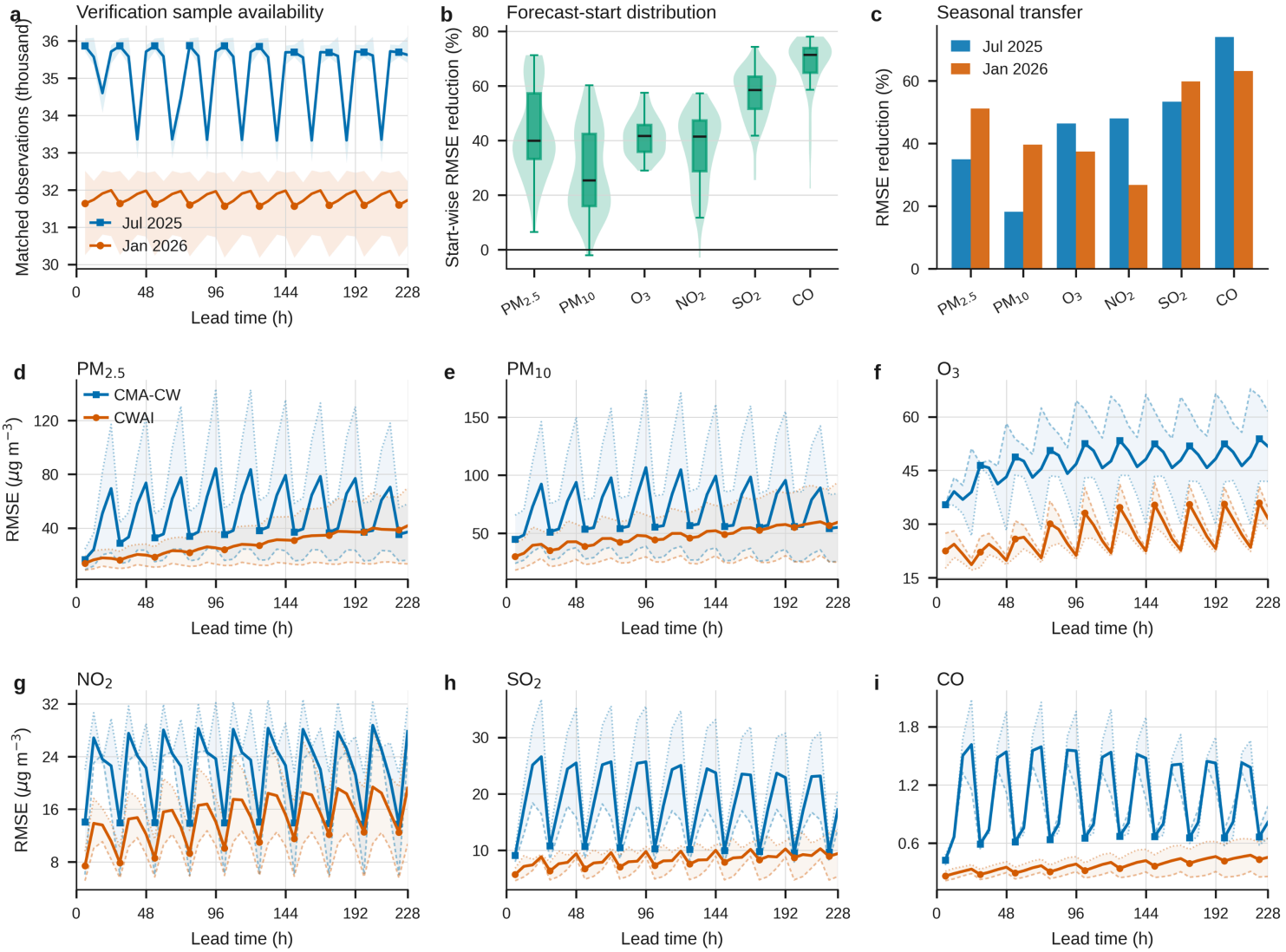


Figure 5

Independent station validation of pollutant forecasts. CWAI and CMA-CW were evaluated against national monitoring-station observations for 60 starts in July 2025 and January 2026. **a**, Mean number of matched observations at each forecast lead; shading spans the pollutant-specific range. **b**, Distribution of paired RMSE reduction across forecast starts. **c**, Month-specific initialization-weighted RMSE reduction. **d–i**, Lead-time RMSE for $PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2 and CO under identical observation masks. Thick lines show the two-period mean; faint lines and envelopes show the July 2025 and January 2026 cohorts. Timestamps were matched directly in UTC.

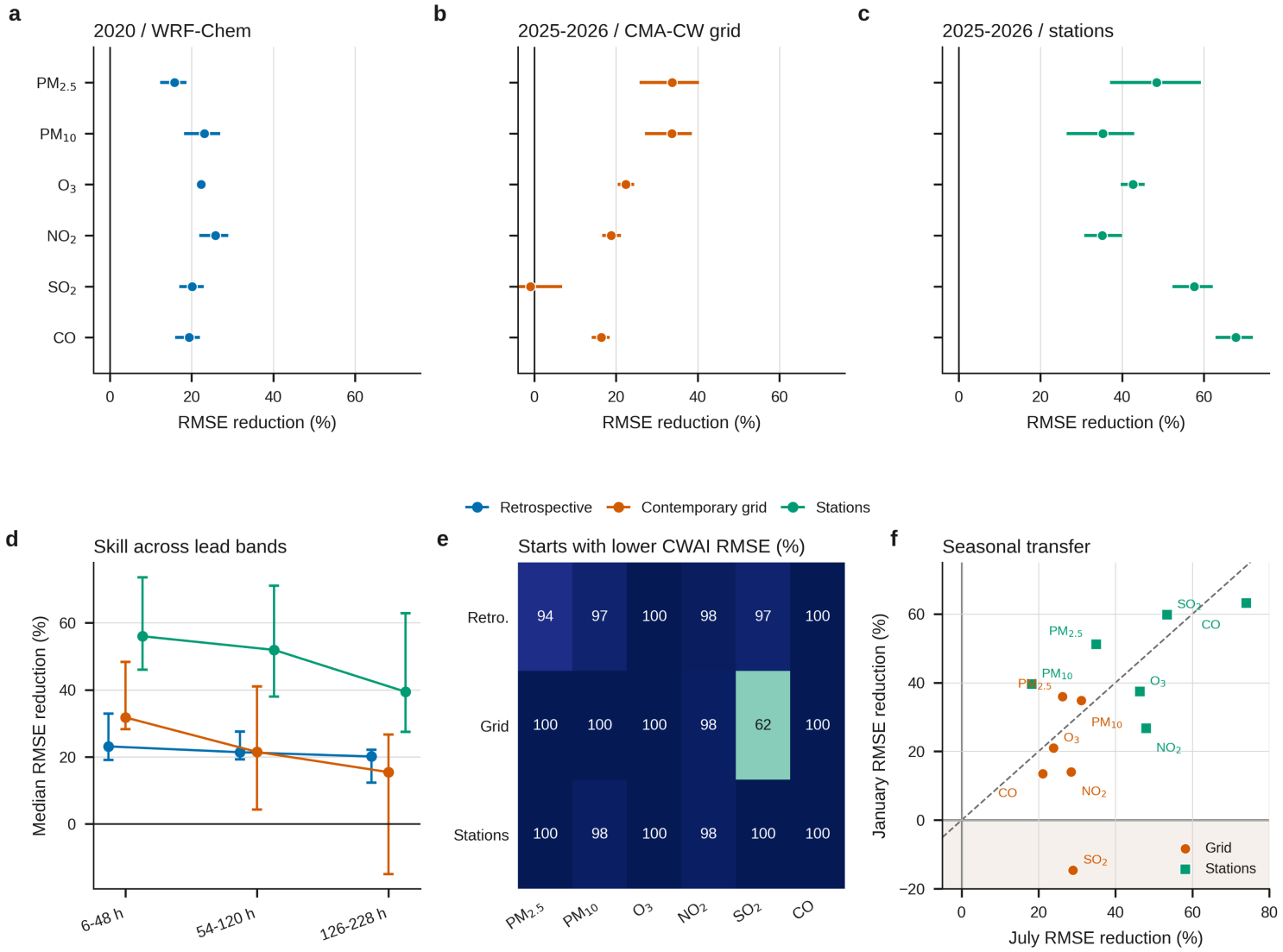


Figure 6

Cross-tier synthesis of pollutant forecast skill. Six pollutants are synthesized across the 2020 WRF-Chem benchmark, the 2025/2026 CMA-CW grid benchmark and independent stations. **a–c**, Initialization-weighted RMSE reductions and 95% paired-bootstrap confidence intervals for the retrospective grid, contemporary grid and station tiers, respectively. **d**, Median reduction across pollutants in short, medium and long lead bands; ranges span pollutants rather than statistical uncertainty. **e**, Percentage of starts with lower CWAI RMSE. **f**, January versus July RMSE reduction for contemporary grids and independent stations. The dashed diagonal indicates unchanged skill between seasons; points below zero on the vertical axis indicate winter degradation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CWAIsupplementaryinformation.docx