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.

Independent evaluation of Aurora air pollution forecasts beyond benchmark skill

Angelo Riccio

angelo.riccio@uniparthenope.it

Parthenope University of Naples

Article

Keywords: Earth system foundation models, air quality, particulate matter, model evaluation, predictive uncertainty, atmospheric composition

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11417641/v1

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

Additional Declarations: No competing interests reported.

Independent evaluation of Aurora air pollution forecasts beyond benchmark skill

Angelo Riccio[1*]

[1*]Department of Science and Technology, Parthenope University of Naples, Naples, Italy.

Corresponding author(s). E-mail(s): angelo.riccio@uniparthenope.it;

Abstract

We evaluate the released Aurora 0.4° Air Pollution checkpoint against CAMS, persistence and hourly background-station observations for 96 initializations spanning 2024. Against CAMS, Aurora reduces mean squared error relative to persistence by 38.1–42.6% for $PM_{2.5}$ and 49.1–52.3% for PM_{10}. At sampled station times, its mean absolute error is lower than that of the benchmark-selected CAMS fields in all six combinations of three regions and two pollutants; 95% initialization-date bootstrap intervals exclude zero. At +12 h, 6.27% of global $PM_{2.5}$ and 7.84% of PM_{10} grid-point values are negative, whereas matched CAMS states are non-negative; subsequent forecast steps use positivity clamping. The evaluated outputs provide no forecast distribution. Perturbing four static source-related inputs produces small, nearly additive regional responses, whose relationship to changes in actual emissions remains untested. Observational agreement, physical constraints and input sensitivity thus yield distinct assessments of the same forecasts.

Keywords: Earth system foundation models, air quality, particulate matter, model evaluation, predictive uncertainty, atmospheric composition

1 Introduction

Machine learning has become a serious alternative for short-range atmospheric forecasting. Systems such as GraphCast and NeuralGCM have advanced global weather prediction [1, 2], while WeatherBench 2 has made comparisons more systematic [3]. A useful benchmark score, however, says little on its own about agreement with measurements or the physical behaviour of the predicted fields.

Independent evaluation of Earth-system models requires observations, complementary diagnostics and clear limits on interpretation [4, 5]. This is especially important for atmospheric chemistry: a good reconstruction does not establish that the model represents the underlying processes [6]. Reviews of air-quality forecasting identify persistent gaps in external validation and uncertainty assessment [7, 8].

Particulate-matter concentrations depend on emissions, transport, mixing, deposition and secondary chemistry. Their relation to weather is pronounced in complex coastal terrain such as Campania, southern Italy [9]. Because an individual monitor also records variation within a model grid cell, station comparisons require an explicit spatial match [10].

We evaluate the released *Aurora 0.4° Air Pollution* checkpoint (hereafter *Aurora*), implemented as `AuroraAirPollution`. Its published evaluation used Copernicus Atmosphere Monitoring Service (CAMS) products [11]. The checkpoint was fine-tuned on CAMS analysis, which the model documentation recommends as input (https://microsoft.github.io/aurora/models.html; accessed 14 October 2026). CAMS offers a consistent benchmark; monitoring stations test agreement with independently measured concentrations. A recent evaluation in China also compared Aurora with surface observations and gridded products [12].

We evaluate all 96 sampled 2024 initializations against hourly European Environment Agency (EEA) background measurements on matched station–time records. Separately, we compare the 2023–2024 benchmark forecasts with CAMS and persistence, and check raw concentrations against simple physical constraints. Meteorological stress tests put the year-round result in context. An audit of the released output establishes what can be assessed about predictive uncertainty; controlled perturbations of four static anthropogenic inputs measure the checkpoint's response to those channels. These comparisons address different properties of this specific release, without equating input sensitivity with an emissions intervention.

2 Results

2.1 Station comparisons favour Aurora across sampled 2024 dates

Across all 96 sampled 2024 initializations (00 and 12 UTC on 48 dates), Aurora has lower station-matched MAE than CAMS at all six primary endpoints (Table 1). These endpoints are the three evaluation boxes—Europe, Italy and Campania—crossed with the two pollutants, PM_{10} and $PM_{2.5}$ (see Fig. 1 for the geographic location of these boxes). In Europe, MAE decreases from 8.32 to 7.67 $\mu g\,m^{-3}$ for PM_{10} and from 5.53 to 5.08 $\mu g\,m^{-3}$ for $PM_{2.5}$. The corresponding reductions are 9.71 to 7.96 and 5.91 to 5.16 $\mu g\,m^{-3}$ in the Italy box, and 12.52 to 9.42 and 7.00 to 5.52 $\mu g\,m^{-3}$ in Campania.

For each matched station–time–pollutant record, let A, C and O denote the Aurora value, the matched CAMS value and the observation, respectively. We define the paired absolute-error difference as $\Delta|e| = |A - O| - |C - O|$, so negative values favour Aurora. For each endpoint, the mean of these record-level differences is $\Delta\mathrm{MAE} = \mathrm{MAE}_A - \mathrm{MAE}_C$. All six 95% intervals for ΔMAE, obtained by resampling the 48 initialization dates as blocks, lie below zero (Table 1). RMSE point estimates also

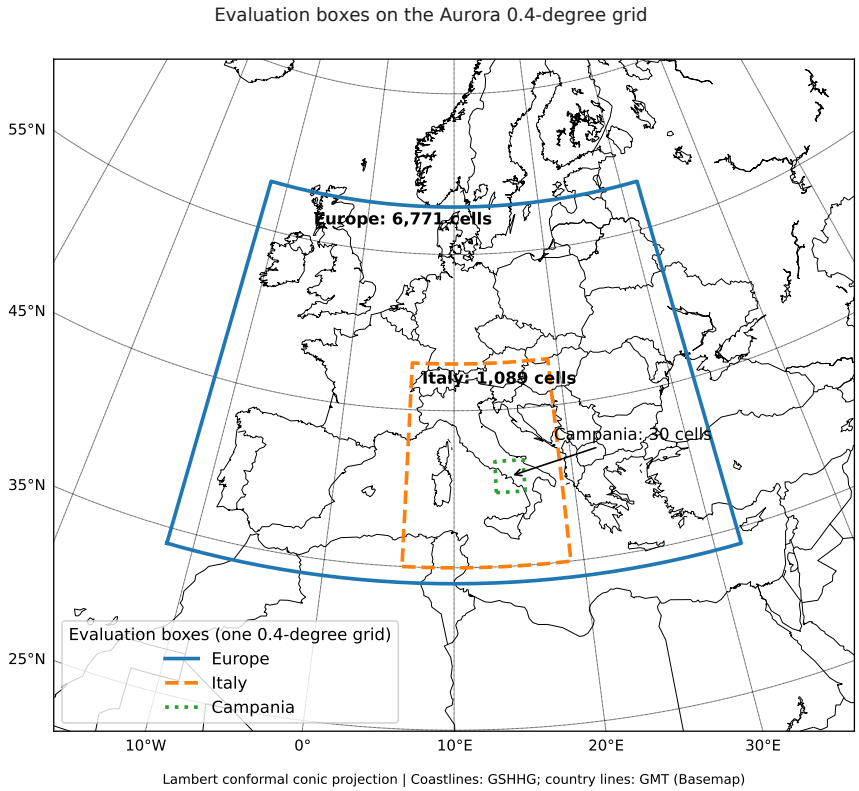


Fig. 1 Europe, Italy and Campania evaluation boxes in a Lambert conformal conic projection. Box boundaries mark scoring areas; coastlines and country outlines provide context.

favour Aurora in every comparison. Five of six intervals for the Aurora-minus-CAMS RMSE difference exclude zero; Campania $PM_{2.5}$ does not ($[-1.71, 0.12]$ $\mu g\,m^{-3}$).

The advantage persists across forecast horizons. Combining the six endpoints with the four leads (+12, +24, +36 and +48 h) gives 24 endpoint–lead comparisons, and the 95% interval for ΔMAE excludes zero in all 24. Across the same six endpoints and the four seasons, 23 of 24 seasonal ΔMAE point estimates favour Aurora; Campania $PM_{2.5}$ in December–February is the exception ($+0.05$ $\mu g\,m^{-3}$). For each initialization, we also average $\Delta|e|$ over the available station–lead records within an endpoint. This case-wise mean is negative for 71–80 of the 96 initializations (74–83%, depending on endpoint), so the sample still includes individual initializations for which CAMS agrees better with the stations.

2.2 Benchmark skill transfers across the sampled years

Aurora improves on persistence against CAMS at all six endpoints in 2024 (Fig. 2). Here persistence holds the initial CAMS concentration field fixed through verification, and MSE skill is the percentage reduction in MSE achieved by Aurora relative to that forecast, with both scored against the same CAMS verification field. $PM_{2.5}$ skill is

Table 1 Station comparison across the 96 sampled 2024 initializations. Aurora (A) and benchmark-selected CAMS states (C) are evaluated on identical quality-controlled EEA background-station records with observations O. ΔMAE is the mean record-level difference $\Delta|e| = |A - O| - |C - O|$, equivalently $\mathrm{MAE}_A - \mathrm{MAE}_C$ on identical records; negative values favour Aurora. Brackets give 95% intervals from 20,000 bootstrap resamples of 48 initialization dates, retaining the 00 and 12 UTC cycles, all four leads and stations within each date. The boxes overlap spatially; differences use unrounded scores.

Box	Pollutant	MAE A	MAE C	ΔMAE [95% CI]	RMSE A	RMSE C
		$\mu g\,m^{-3}$		$\mu g\,m^{-3}$	$\mu g\,m^{-3}$	
Europe	PM_{10}	7.67	8.32	−0.66 [−1.04, −0.32]	11.50	13.27
Europe	$PM_{2.5}$	5.08	5.53	−0.45 [−0.67, −0.26]	8.03	8.69
Italy	PM_{10}	7.96	9.71	−1.75 [−2.68, −1.03]	12.83	16.71
Italy	$PM_{2.5}$	5.16	5.91	−0.75 [−1.03, −0.51]	9.13	9.86
Campania	PM_{10}	9.42	12.52	−3.10 [−4.97, −1.69]	18.40	23.10
Campania	$PM_{2.5}$	5.52	7.00	−1.47 [−2.09, −0.95]	13.90	14.51

40.3% in Campania, 42.6% in Italy and 38.1% in Europe. The respective PM_{10} scores are 49.1%, 52.2% and 52.3%. Across the six endpoints and four forecast leads (24 lead-specific comparisons), 23 bootstrap intervals exclude zero; the exception is Campania PM_{10} at +24 h, with an interval of $[-0.52, 72.03]$%.

All six scores are positive in 2023 as well. The median change from 2023 to 2024 is +2.09 percentage points (range, −4.74 to +6.11). The geographic boxes overlap; these results chiefly establish consistency across the sampled dates in two years.

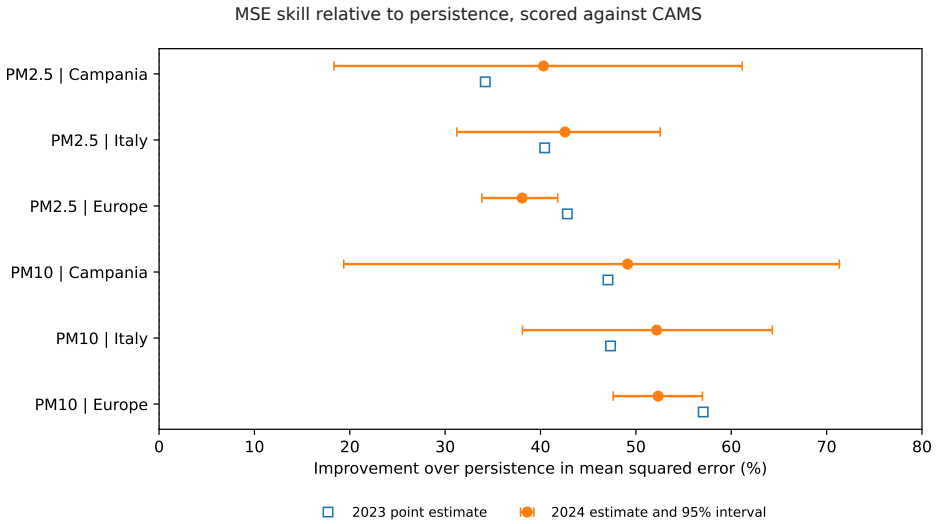


Fig. 2 MSE skill relative to persistence, using CAMS as the reference. Zero indicates equal MSE for Aurora and persistence. Open markers show 2023 point estimates; filled markers and horizontal whiskers show 2024 estimates and 95% intervals obtained by resampling the 48 initialization dates as blocks.

2.3 Meteorological stress tests contextualize the annual result

The two prespecified meteorological stress tests show how strongly absolute error can vary within the year-round sample: the 25 January 2024 00 UTC initialization was cold and stagnant over Campania, whereas the 10 August 2024 00 UTC initialization had stronger radiation and ventilation. In Campania, Aurora's cold episode has mean biases of $-28.88\ \mu g\,m^{-3}$ for PM_{10} and $-28.18\ \mu g\,m^{-3}$ for $PM_{2.5}$, with RMSE of 58.73 and 57.58 $\mu g\,m^{-3}$. CAMS shows a similar winter deficit, with RMSE of 58.49 and 56.54 $\mu g\,m^{-3}$ and biases of -32.69 and $-29.82\ \mu g\,m^{-3}$. In the warm high-radiation episode, Aurora's Campania RMSE falls to 8.07 and 4.89 $\mu g\,m^{-3}$, compared with 7.43 and 5.41 $\mu g\,m^{-3}$ for CAMS.

Ranking the 96 Campania initializations by the case-wise mean $\Delta|e|$ defined above, the cold case lies at the 19.8th and 31.3rd percentiles for PM_{10} and $PM_{2.5}$, respectively; the warm case lies at the 88.5th and 74.0th percentiles (Supplementary Fig. S2). Lower percentiles favour Aurora. By +48 h in the cold episode, Aurora's bias is 3.81 $\mu g\,m^{-3}$ more negative than CAMS for PM_{10} and 3.45 $\mu g\,m^{-3}$ more negative for $PM_{2.5}$. The year-round mean hides these differences between meteorological episodes.

2.4 Raw fields violate simple particulate matter constraints

All archived 2024 Aurora values are finite. At +12 h, 6.27% of global $PM_{2.5}$ and 7.84% of PM_{10} grid-point occurrences are negative (Fig. 3); none are negative at later leads. On the same spatial support, neither the 384 matched CAMS verification states (96 initializations at each of four leads) nor the 96 corresponding CAMS initialization states contain negative PM. The negative values therefore appear in Aurora's first forecast step. In the released implementation, positivity clamping—truncation of negative predictions at zero—is disabled at that first step by default and applied to later rollout steps; the zero fractions after +12 h must be read in that light. Negatives are much rarer in the regional boxes: absent in Campania and at most about 0.052% in Italy and Europe.

At +12 h, Aurora $PM_{2.5}$ exceeds PM_{10} in 10.47% of global grid-point occurrences, including 8.57% by more than 0.1 $\mu g\,m^{-3}$, 0.409% by more than 1 $\mu g\,m^{-3}$ and 0.0092% by more than 5 $\mu g\,m^{-3}$. At later leads, raw exceedance fractions remain approximately 6.3–6.7%. Matched CAMS verification states show only about 0.38–0.45% raw $PM_{2.5} > PM_{10}$ reversals and none above 0.1 $\mu g\,m^{-3}$ at any lead; matched initialization states show the same qualitative pattern. Aurora therefore increases both the frequency and magnitude of this ordering inconsistency relative to the matched CAMS fields.

2.5 Meteorological conditioning of forecast error

We tested whether six prespecified weather variables improve out-of-sample prediction of regional Aurora RMSE relative to a baseline that accounts for season and the combination of forecast lead and initialization hour. The regression predicts a logarithmically transformed RMSE target (Methods). At five of the six region–pollutant endpoints, adding weather lowers the held-out squared prediction error at the point-estimate level, but all six 95% date-block bootstrap intervals for the percentage reduction in

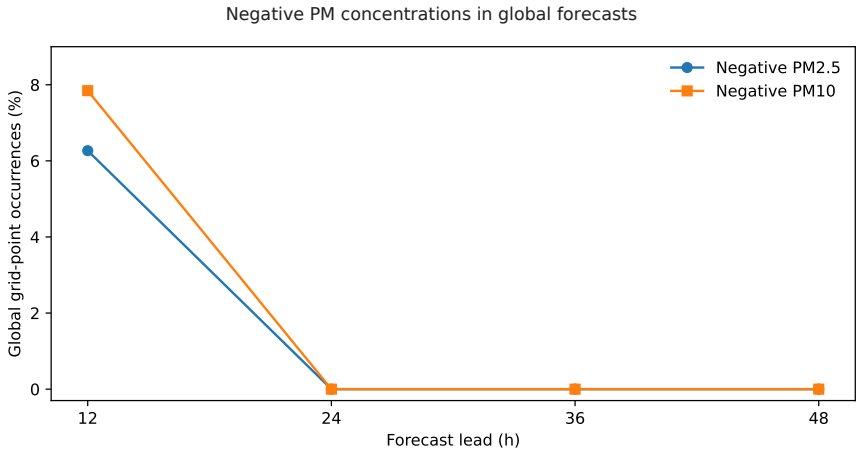


Fig. 3 Frequency of negative PM concentrations by forecast lead in the archived global fields. Percentages pool grid-point occurrences over initializations at each lead, without area weighting or additional clipping in this analysis. The released model applies its default positivity clamp after the first rollout step.

that error include zero. Across four leads per endpoint (24 lead-specific comparisons), only five intervals exclude zero. The added weather variables therefore show no clear overall predictive gain.

The fitted weather coefficients are nevertheless similar across years. For each endpoint, we compare the six weather-variable coefficients estimated in 2023 with their 2024 counterparts. The median cosine similarity between these six-coefficient vectors is 0.916, and the median Spearman correlation between their absolute-magnitude ranks is 0.829. Four endpoints retain the same sign for five of six coefficients; the other two retain all six signs. Thus a recognizable coefficient pattern persists even though the added variables do not clearly improve out-of-sample error prediction.

2.6 Predictive uncertainty exposed by the released configuration

The evaluated air-pollution runs provide no forecast probability distribution. The evaluated forward method returns a single prediction object (`Batch`), and the eight archived files from the two stress tests (four forecast leads per test) contain latitude–longitude $PM_{2.5}$ and PM_{10} fields without ensemble members, quantiles, variance, standard deviation or distribution parameters. The shared Aurora interface also has an optional stochastic setting; its air-pollution performance and calibration were not assessed here.

2.7 Sensitivity to the static anthropogenic proxies

Each of the four static anthropogenic proxies (NH_3, CO, NO_x and SO_2) is represented by a linear input field and a transformed companion channel. We perturb each proxy

pair individually and all four pairs jointly. To summarize the paired ±10% perturbations, let Y_0 denote the baseline regional-mean concentration and $Y_{+0.10}$ and $Y_{-0.10}$ the corresponding means after +10% and −10% scaling of one proxy pair. We define $D_{10} = (Y_{+0.10} - Y_{-0.10})/0.20$ and plot $0.1D_{10}$, a central linearized 10% response around the baseline rather than the directly simulated one-sided change $Y_{+0.10} - Y_0$.

The regional-mean responses are small (Figs. 4 and 5). In Italy, the $PM_{2.5}$ value of $0.1D_{10}$ for NO_x is +0.00417 $\mu g\,m^{-3}$, or 0.0436% of the 9.573 $\mu g\,m^{-3}$ baseline. The corresponding NH_3 and SO_2 values are −0.00311 and −0.00309 $\mu g\,m^{-3}$. In Campania, the joint all-four value of $0.1D_{10}$ for PM_{10} is +0.01190 $\mu g\,m^{-3}$, or 0.0660% of baseline.

Response signs vary by proxy and location. NO_x has a positive D_{10} for $PM_{2.5}$ in every box and at every lead; SO_2 has a negative one. NH_3 has a negative $PM_{2.5}$ response in Italy and Europe, while the Campania $PM_{2.5}$ interval for $0.1D_{10}$ includes zero. For PM_{10}, the central NH_3 response is negative in Italy and positive in Campania and Europe. When the 48 initialization dates are resampled as blocks, the 95% intervals for three small PM_{10} values of $0.1D_{10}$ include zero: Campania–NO_x, Italy–CO and Italy–SO_2. These values measure how Aurora responds to its static input channels.

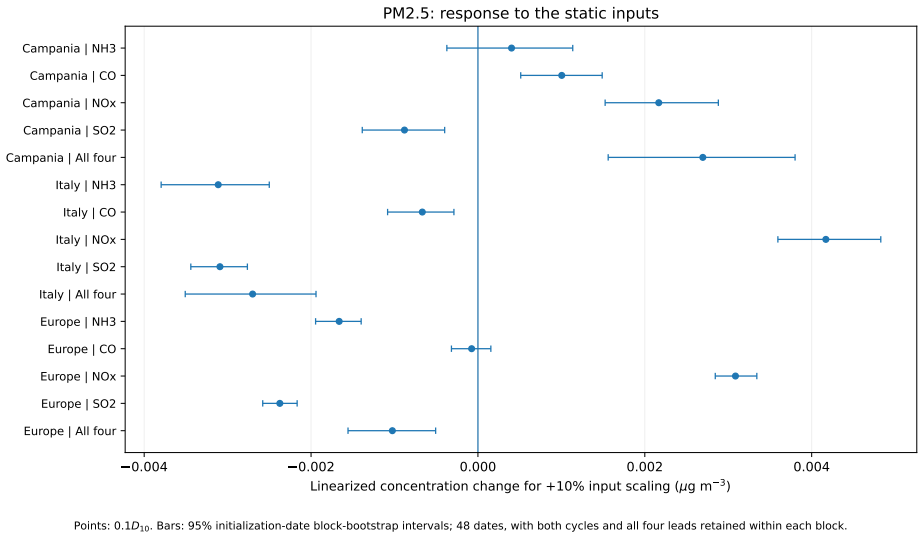


Fig. 4 Linearized regional-mean $PM_{2.5}$ response to 10% scaling of each static proxy or all four together. Points show $0.1D_{10}$, defined in the text from paired ±10% perturbations; horizontal whiskers show 95% intervals from 48 initialization-date blocks, each retaining the 00 and 12 UTC cycles and all four leads. The proxy names identify input channels, not emission sectors.

2.8 Linearity and additivity of the static-proxy response

Define the analogous 20% central sensitivity as $D_{20} = (Y_{+0.20} - Y_{-0.20})/0.40$. The values of D_{10} and D_{20} generally differ by about 1%. Let B denote the logarithmically transformed companion channel and α the fractional perturbation applied to its paired linear input ($\alpha = \pm 0.10$ or ± 0.20). Above the transformation floor, the corresponding

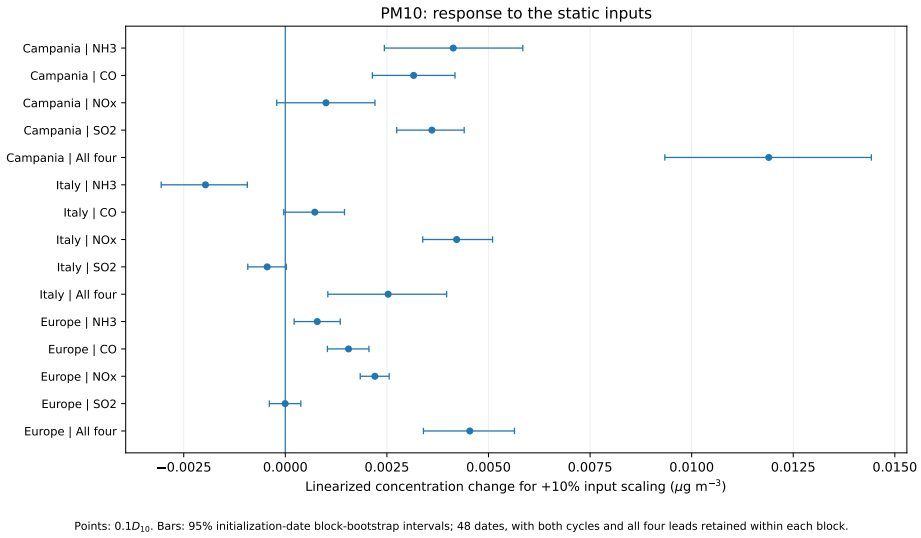


Fig. 5 Linearized regional-mean PM_{10} response to 10% static-proxy scaling, calculated as in Fig. 4. Horizontal whiskers show the same 95% initialization-date block-bootstrap intervals based on 48 initialization dates; the horizontal scale differs from that of $PM_{2.5}$.

change in the companion channel is $\Delta B = \ln(1+\alpha)/10$ (Supplementary Fig. S3). If forecast response were linear in this transformed input alone, the central-sensitivity ratio would be

$$\frac{D_{20}}{D_{10}} = \frac{\ln(1.2/0.8)/0.4}{\ln(1.1/0.9)/0.2} = 1.010275. \tag{1}$$

The observed ratios are close to this value. This correspondence partly reflects the input transformation; central differences cancel even-order response terms and cannot establish a linear one-sided forecast response.

Across the six regional endpoints (three boxes crossed with two pollutants), the mean D_{10} residual obtained by subtracting the sum of the four single-proxy sensitivities from the all-four sensitivity ranges from about -2.2×10^{-7} to 1.02×10^{-5} $\mu g\,m^{-3}$ per unit scaling. All six 95% initialization-date block-bootstrap intervals span zero (Supplementary Fig. S4). At the discrete perturbation levels themselves, the largest absolute regional-mean joint-minus-sum concentration residual is about 4.71×10^{-5} $\mu g\,m^{-3}$ for Campania PM_{10} at -20%. Regional-mean responses are approximately additive at these levels; the analysis does not resolve local interactions or test an equivalence margin.

3 Discussion

Against the benchmark-selected CAMS states, Aurora improves on persistence in both sampled years. The comparison with EEA measurements adds a separate result: Aurora agrees more closely with the stations across the six primary endpoints and four leads of the year-round 2024 sample. The station data were not the benchmark target. Still, the 48 sampled dates are a small subset of the calendar, the three boxes

overlap, and nine paired background stations lie in the Campania box. A station also measures local conditions that a 0.4° cell cannot resolve [10]. The comparison supports a paired advantage on these records, not a ranking of every CAMS product or unsampled location.

Weather regime matters despite the annual advantage. Both systems sharply underestimate PM in the cold, stagnant Campania case; errors are smaller in August, when ventilation is stronger. The deficit in January is already pronounced in the CAMS states, and the bulk PM measurements cannot distinguish the contributions of emissions, mixing and secondary production. These episodes sit in different parts of the case-wise error distribution. Meanwhile, adding six weather variables provides no clear overall gain to the retrospective Ridge error model. Visible contrasts between two episodes need not translate into robust incremental prediction of error across dates.

Raw-field checks give a different picture again. The matched CAMS fields are non-negative, but Aurora's unclamped first forecast step produces negative PM. Later steps use the release's default positivity clamp, so their absence of negative values cannot demonstrate an intrinsically non-negative prediction. Aurora also produces substantially more instances in which $PM_{2.5}$ exceeds PM_{10} than the matched CAMS fields, including reversals larger than the 0.1 $\mu g\,m^{-3}$ tolerance. These checks identify properties of the released forecast pipeline on the sampled states; the source of the PM ordering reversals remains unresolved.

The similar coefficients of the external weather-error regression across years describe that diagnostic's stability, rather than Aurora's internal reasoning. The forecasts evaluated here contain single concentration fields. Although the shared model interface accepts a stochastic setting, the present outputs offer no ensemble whose calibration or interval coverage can be assessed. Bootstrap intervals estimate variation in aggregate scores across initialization dates; they do not express the uncertainty of a forecast delivered to a user [13].

Changing static anthropogenic inputs yields small, mostly additive regional-mean responses. These channels have no documented mapping from a time-resolved, sector-specific emissions intervention, and regional averaging can conceal opposing local changes. Chemistry and policy analyses that alter actual emissions address a different perturbation [14, 15]. Here the response belongs to the model inputs as released. Across the sampled forecasts, station agreement, physical admissibility, predictive uncertainty and static-input response therefore remain distinct tests of this particular checkpoint.

4 Methods

4.1 Benchmark data, grid and evaluation regions

Each annual benchmark contains 96 initializations: 00 and 12 UTC cycles on each of 48 dates spread across the year. Predictions at +12, +24, +36 and +48 h yield 384 fields per year. The 2023–2024 comparison examines these sampled dates, not a continuous two-year forecast.

We verify forecasts against the archived CAMS PM states selected by the Aurora benchmark. The persistence forecast holds each initial CAMS concentration constant

through verification. Aurora and persistence are scored against the same CAMS verification field on identical grid points; agreement with station observations is evaluated separately.

The initial grid has 451×900 points. After Aurora crops one latitude row, forecast fields have 450×900 points. We score three overlapping geographic boxes on this 0.4° grid (Fig. 1). Their geographic bounds and grid-cell counts are summarized in Table 2. The boxes labelled Europe, Italy and Campania also contain sea and land outside those administrative areas.

Table 2 Geographic bounds and grid-cell counts of the overlapping evaluation boxes. The Campania box contains 30 grid points on the 0.4° grid (Supplementary Fig. S1).

Box	Longitude	Latitude	Grid points
Europe	12°W–32°E	34°N–58°N	6771
Italy	6°E–19°E	35°N–48°N	1089
Campania	13.5°E–15.9°E	39.7°N–41.7°N	30

4.2 Forecast metrics and physical checks

Within each box, spatial averaging uses $w_j = \cos\phi_j$, where ϕ_j is latitude. For prediction P_j and reference R_j, the spatial mean squared error is

$$\mathrm{MSE} = \frac{\sum_j w_j (P_j - R_j)^2}{\sum_j w_j}. \tag{2}$$

Overall RMSE is the square root of the aggregated MSE. We also report bias, MAE and mean spatial correlation. Skill relative to persistence is

$$S_{\mathrm{MSE}} = 100 \left[1 - \frac{\mathrm{MSE}_{\mathrm{Aurora}}}{\mathrm{MSE}_{\mathrm{persistence}}} \right]. \tag{3}$$

A score of 40% means a 40% reduction in squared error relative to persistence.

We examine archived fields for non-finite or negative concentrations, and for $PM_{2.5}$ above PM_{10} on the same reported mass and time basis. The analysis applies no additional clipping. The released `AuroraAirPollution` includes PM among its positive surface variables; its default `clamp_at_first_step=False` leaves the first rollout prediction unclamped and activates positivity clamping at later steps (https://microsoft.github.io/aurora/api.html; accessed 14 October 2026). For PM ordering, we count differences above 0.1, 1 and 5 $\mu g\,m^{-3}$. Global percentages describe *grid-point occurrences across forecasts*; no area weighting is applied. We apply the identical checks to the 384 matched CAMS verification states and the 96 CAMS initialization states. CAMS fields have 451 latitude rows; for this control the final row is removed

to reproduce Aurora's 450×900 output support before pooling the diagnostics. These checks address basic concentration properties rather than full chemical budgets.

4.3 Station evaluation and meteorological stress tests

We assembled hourly EEA background-station records before examining Aurora errors. Primary-network stations reported both PM_{10} and $PM_{2.5}$ hourly, had at least 80% annual coverage and were classified as background sites. Traffic and industrial sites are not part of the primary network. The primary hourly-comparable network contains 577 paired background stations in the Europe box, including 61 in the geographic Italy box and 9 in the Campania box. Because the Italy box includes stations in neighbouring countries, we also calculate an Italian-stations-only sensitivity check.

The observation-based evaluation uses all 96 sampled 2024 initializations and four leads. Because valid times overlap between the 00 and 12 UTC cycles, the 384 case–lead forecasts correspond to 240 unique valid times. At each case and lead, observations are matched at the exact valid time. Each station is mapped independently to its nearest Aurora and CAMS grid cells; both fields are scored on identical quality-controlled station–time–pollutant records. After observation quality control, the Europe network contributes 429,374 matched rows across the two pollutants. All six primary endpoints—three evaluation boxes crossed with two pollutants—retain all 96 cases, 48 initialization dates and 240 unique valid times.

CAMS fields follow the source-selection rules of the 2024 benchmark. The exact benchmark-selected reference file is used for each case–lead verification state, without temporal interpolation. Across the annual comparison this requires 240 unique CAMS verification states because some valid times are shared by forecasts from the two initialization cycles. We report bias, MAE and RMSE pooled over the matched station–time records within each endpoint, together with the pooled model–observation correlation.

An additional check restricted to stations within Italy gives ΔMAE values of $-2.31\ \mu g\,m^{-3}$ for PM_{10} (95% interval $[-3.43, -1.40]$) and $-0.91\ \mu g\,m^{-3}$ for $PM_{2.5}$ ($[-1.30, -0.58]$). This distinction matters because the Italy evaluation box also contains stations in neighbouring countries.

To compare Aurora and CAMS directly, we define the paired absolute-error difference

$$\Delta|e| = |A - O| - |C - O|, \tag{4}$$

where A, C and O are the Aurora value, the matched CAMS value and the observation. Negative differences favour Aurora. Because Aurora and CAMS are scored on identical records, the mean of $\Delta|e|$ over any record set equals $\Delta\text{MAE} = \text{MAE}_A - \text{MAE}_C$. Primary uncertainty is estimated with 20,000 bootstrap resamples of the 48 initialization-date blocks. Each sampled date retains the 00 and 12 UTC cycles, all four leads and all available station records, preserving dependence within a date and between overlapping valid times. Lead-specific intervals use the same date-block principle. Seasonal summaries are descriptive and have no separate intervals. The six endpoint intervals and the 24 endpoint–lead intervals (six endpoints crossed with four leads) are unadjusted for multiple comparisons, and the overlapping boxes are not independent replicates.

For case-wise robustness, we average $\Delta|e|$ across the available station–lead records within each initialization and report the fraction of the 96 cases for which this mean is negative. We also retain two meteorological stress tests selected before inspection of Aurora PM concentrations or errors: 25 January 2024 00 UTC, cold and stagnant over Campania, and 10 August 2024 00 UTC, with stronger radiation, a deeper boundary layer and better ventilation. Their positions are reported within the full 96-case distribution rather than treated as the primary observational evidence.

4.4 Meteorological error conditioning

We model the transformed regional forecast error $z = \ln[1 + \mathrm{RMSE}/(1\ \mu\mathrm{g\,m^{-3}})]$ with Ridge regression. The baseline predictors are season and the categorical interaction between forecast lead and initialization hour. The weather-augmented model adds six prespecified ERA5 variables at forecast verification time: wind speed, boundary-layer height, a stability proxy, relative humidity, precipitation and surface radiation [16]. Preprocessing uses training folds only, and nested validation leaves out each initialization date in turn; the Ridge penalty is selected separately for the baseline and weather-augmented models within the training data.

For each region–pollutant endpoint, predictive gain is the percentage reduction in held-out MSE of z for the weather-augmented model relative to the baseline model; positive values indicate lower out-of-sample error after adding weather. Descriptive stability is assessed from the similarity of the six weather-variable coefficients between 2023 and 2024. The regression describes past forecast errors rather than Aurora's internal calculations.

4.5 Audit of diagnostic stability and predictive uncertainty

For this external weather-error diagnostic, we compare coefficient vectors, coefficient ranks and coefficient signs across years. These measures do not identify Aurora's internal mechanisms.

We inspected the installed `microsoft-aurora` 2.0.1 public API, the `AuroraAirPollution` source and archived 2024 outputs. The forward method returns a prediction `Batch`. The eight archived files from the two stress tests (four forecast leads per test) examined for output structure contain $PM_{2.5}$ and PM_{10} on latitude–longitude dimensions, with no ensemble members or probabilistic fields. Although the shared API exposes a `stochastic` option (https://microsoft.github.io/aurora/api.html; accessed 14 October 2026), we did not evaluate it or infer a calibrated air-pollution distribution from its availability.

4.6 Static anthropogenic inputs and inventory context

Four static inputs represent anthropogenic NH_3, CO, NO_x and SO_2 patterns. Each has a linear channel and a transformed companion. Among the tested reconstructions of these static input patterns, the closest match to the released asset used 2019 CAMS-GLOB-ANT v4.2 data, equal weighting of 12 monthly mean fields and periodic bilinear remapping (ECCAD CAMS-GLOB-ANT release documentation, https://eccad.aeris-data.fr/2020/04/10/cams-glob-ant-v4-2-and-v4-2-r1-1-now-available/; accessed 14

October 2026). The exact preprocessing of the released asset is undocumented; [17] provide broader context on the inventory.

The inventory varies by month, but Aurora receives one spatial field through these channels for all 2024 initializations. The equally weighted monthly mean represents an input pattern, not calendar-weighted annual emissions.

The perturbation experiment starts from the released static fields. Let E_0 denote a released baseline linear proxy field and $\alpha \in \{-0.20, -0.10, +0.10, +0.20\}$ the fractional perturbation. The perturbed linear field is $E_\alpha = (1 + \alpha)E_0$. Let $B(E)$ denote the corresponding transformed companion channel, defined by the audited transformation

$$B(E) = \frac{\ln[\max(E/E_{\max,0}, 10^{-10})] - \ln(10^{-10})}{10}, \tag{5}$$

Here $E_{\max,0}$ is the maximum of the released baseline field. We hold this normalization constant and leave the unselected proxy channels unchanged. The selected linear and transformed channels change together, so the experiment measures their combined effect.

At the baseline maximum, the transformed channel equals $\ln 10 \simeq 2.302585$. Scaling by +10% or +20% raises it to about 2.312116 or 2.320817, respectively. The training distribution of these channel combinations is not documented. We therefore interpret the ±10% and ±20% perturbations as controlled model-input sensitivity probes rather than as guaranteed in-distribution interpolation or actual emissions interventions.

4.7 Response metrics and sampling uncertainty

For each of the four proxies, the four perturbation levels (−20%, −10%, +10% and +20%) give 16 single-proxy scenarios. Four additional joint scenarios perturb all four proxies together by the same fractional amount. The resulting 20 scenarios, evaluated for 96 initializations and four leads, yield 7680 perturbed forecast fields. As a control, loading the unmodified static asset through the override path reproduced four test initializations at all four leads, or 16 baseline forecasts, bit for bit.

Let Y_0 denote the baseline regional-mean concentration and Y_α the corresponding mean after a fractional perturbation α. The absolute response is $\Delta Y_\alpha = Y_\alpha - Y_0$. Central sensitivities are

$$D_h = \frac{Y_{+h} - Y_{-h}}{2h}, \qquad h \in \{0.10, 0.20\}, \tag{6}$$

and curvature $C_{20} = (Y_{+0.20} - 2Y_0 + Y_{-0.20})/0.20^2$. Here D_{10} and D_{20} denote D_h at $h = 0.10$ and 0.20, respectively. The plotted quantity $0.1D_{10}$ is a *linearized* 10% response; it is not the measured one-sided difference $Y_{+0.10} - Y_0$. For joint non-additivity, let $Y_{\text{all},\alpha}$ be the regional-mean concentration when all four proxies are perturbed together and $Y_{i,\alpha}$ the corresponding concentration when only proxy i is perturbed. We define

$$I_\alpha = (Y_{\text{all},\alpha} - Y_0) - \sum_i (Y_{i,\alpha} - Y_0), \tag{7}$$

where the sum runs over the four proxies. Regional averaging can cancel local changes of opposite sign; the reported responses describe regional means.

For sensitivity intervals, we draw 20 000 bootstrap samples from the 48 initialization-date blocks. Each sampled date retains both 00 and 12 UTC initialization cycles, all four leads and the pairing among perturbation scenarios. For each endpoint, the eight case–lead contributions within a date (two initialization cycles × four leads) are averaged before resampling. The intervals describe sampling variability in the mean regional response and are used descriptively, without a multiplicity adjustment.

4.8 Statistics and reproducibility

For benchmark scores, station comparisons and static-input responses, uncertainty is summarized by 95% intervals from 20,000 paired resamples of 48 initialization-date blocks, as detailed above. We report point estimates and intervals rather than hypothesis tests or P values. The intervals describe variation among sampled dates, conditional on the station network and evaluation boxes; they do not measure spatial representativeness or uncertainty in an individual forecast. No adjustment was applied for the multiple region–pollutant, lead or static-input comparisons.

4.9 Use of generative AI tools

OpenAI ChatGPT assisted with language editing, manuscript organization and LaTeX formatting. The numerical simulations and measured observations were not generated or altered with this tool. Responsibility for the scientific interpretations, references and submitted text rests with the author.

Data availability

The evaluated Microsoft Aurora source and model checkpoint are publicly available at https://github.com/microsoft/aurora and https://huggingface.co/microsoft/aurora. Hourly particulate-matter observations can be obtained from the European Environment Agency Air Quality data service; CAMS composition fields can be obtained from the Copernicus Atmosphere Data Store. The study-specific case manifests, station selection and mappings, quality-control metadata, derived score tables and figure data are available from Zenodo, https://dx.doi.org/10.5281/zenodo.23396315.

Code availability

Custom scripts for benchmark scoring, physical checks, meteorological error diagnostics, station matching, bootstrap analyses and static-input perturbations are available from Zenodo, https://dx.doi.org/10.5281/zenodo.23396315. The evaluated `microsoft-aurora` model code is publicly available at https://github.com/microsoft/aurora.

Acknowledgements

This study received no specific funding.

Author contributions

A.R. conceived the study, designed and performed the analyses, interpreted the results, prepared the figures, and wrote and revised the manuscript.

Competing interests

The author declares no financial or non-financial competing interests.

References

[1] Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023). https://doi.org/10.1126/science.adi2336.

[2] Kochkov, D. et al. Neural general circulation models for weather and climate. *Nature* **632**, 1060–1066 (2024). https://doi.org/10.1038/s41586-024-07744-y.

[3] Rasp, S. et al. WeatherBench 2: a benchmark for the next generation of data-driven global weather models. *J. Adv. Model. Earth Syst.* **16**, e2023MS004019 (2024). https://doi.org/10.1029/2023MS004019.

[4] Ullrich, P. A. et al. Recommendations for comprehensive and independent evaluation of machine learning-based Earth system models. *J. Geophys. Res. Mach. Learn. Comput.* **2**, e2024JH000496 (2025). https://doi.org/10.1029/2024JH000496.

[5] Zhu, X. X. et al. On the foundations of Earth foundation models. *Commun. Earth Environ.* **7**, 103 (2026). https://doi.org/10.1038/s43247-025-03127-x.

[6] Wang, N., Zhang, Y., Wei, J., Wang, Y. Y. & Sun, Y. L. Artificial intelligence for atmospheric chemistry: from observation to process understanding. *npj Clim. Atmos. Sci.* (2026). https://doi.org/10.1038/s41612-026-01559-6.

[7] Tang, D., Zhan, Y. & Yang, F. A review of machine learning for modeling air quality: overlooked but important issues. *Atmos. Res.* **300**, 107261 (2024). https://doi.org/10.1016/j.atmosres.2024.107261.

[8] Chianese, E. & Riccio, A. Air-quality forecasting across monitoring, predictor, and validation regimes: a systematic mapping review and decision framework. *Forecasting* **8**, 72 (2026). https://doi.org/10.3390/forecast8040072.

[9] Hernández-Ceballos, M.-A., Rubino, M., Sirignano, C., Chianese, E. & Riccio, A. The cause-effect relationship between synoptic and local wind patterns and

PM_{10} concentrations in the complex-orography urban area of Naples (Italy). *City Environ. Interact.* **27**, 100200 (2025). https://doi.org/10.1016/j.cacint.2025.100200.

[10] Riccio, A. & Chianese, E. Technical note: accurate, reliable, and high-resolution air quality predictions by improving the Copernicus Atmosphere Monitoring Service using a novel statistical post-processing method. *Atmos. Chem. Phys.* **24**, 1673–1689 (2024). https://doi.org/10.5194/acp-24-1673-2024.

[11] Bodnar, C. et al. A foundation model for the Earth system. *Nature* **641**, 1180–1187 (2025). https://doi.org/10.1038/s41586-025-09005-y.

[12] Li, C., Chen, Z. & Yao, J. Evaluation of the Aurora foundation model for multi-pollutant air quality forecasting over China. *Front. Environ. Sci.* **14**, 1882322 (2026). https://doi.org/10.3389/fenvs.2026.1882322.

[13] Chianese, E. & Riccio, A. Multi-model ensemble approaches in air quality prediction: a comprehensive review from chemical transport models to hybrid machine learning. *Atmosphere* **17**, 689 (2026). https://doi.org/10.3390/atmos17070689.

[14] Thunis, P. et al. Non-linear response of $PM_{2.5}$ to changes in NO_x and NH_3 emissions in the Po basin (Italy): consequences for air quality plans. *Atmos. Chem. Phys.* **21**, 9309–9327 (2021). https://doi.org/10.5194/acp-21-9309-2021.

[15] Lino, S., Toscano, D., Riccio, A., Pirovano, G. & Agresti, V. Mobility scenarios and their impact on air quality in the city of Naples. *City Environ. Interact.* **30**, 100340 (2026). https://doi.org/10.1016/j.cacint.2026.100340.

[16] Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020). https://doi.org/10.1002/qj.3803.

[17] Soulié, A. et al. Global anthropogenic emissions (CAMS-GLOB-ANT) for the Copernicus Atmosphere Monitoring Service simulations of air quality forecasts and reanalyses. *Earth Syst. Sci. Data* **16**, 2261–2279 (2024). https://doi.org/10.5194/essd-16-2261-2024.

Supplementary Files

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

- AURORAnpjCASsupplement.pdf