Description of the layout

1. Black: Referee's comments
2. Blue: Author's response
3. *Blue, italic*: The revised content in the manuscript and the supplement.
4. Red: line numbers of revised content; revised tables; complemented figures.

Response to Reviewer #2

This technical note presents a new machine-learning-based method for estimating high-resolution 3D NO_2 concentrations across the atmosphere. The topic is important and has strong potential for applications in satellite retrievals, exposure assessment, and air quality research. However, several issues need clarification for better understanding.

We thank Reviewer #2 for the detailed review and constructive comments. We have revised the manuscript accordingly and provide point-by-point responses below.

Major Comments

1. Justification for the 2 km Resolution

The authors claim the model provides high-resolution estimates "up to 2 km", but the choice of 2 km resolution is not fully justified. The input features vary widely in native resolution—from ~100 m (geography) to ~25 km (meteorology). Why was 2 km chosen rather than 1 km or 500 m? The original CAMS 10 km grid is relatively coarse, and may not match well with EEA ground observations. Is 2 km sufficient to address this representativeness mismatch? Since the model applies a fine-tuning procedure assuming that EEA observations represent ground-truth conditions at the target 2 km grid, the manuscript should provide evidence or rationale demonstrating that matching EEA stations to 2 km grids is appropriate. This justification is important for establishing the validity of the downscaling strategy.

We agree with Reviewer #2 that a clearer rationale is needed for selecting 2 km as the target resolution. We chose 2 km for three main reasons. First, our intended application is to provide high-resolution a-priori NO_2 profiles for emerging satellite missions, and CO2M is expected to deliver the NO_2 product at about 2 km resolution, which is finer than current operational missions. We therefore target 2 km to align the DACNO$_2$ product with this scale.

Second, 2 km provides a practical trade-off between spatial detail and computational feasibility for the daily 3D task. If moving to 1 km or 500 m, training cost and memory use scale strongly with the number of grid cells in 3D learning. For a fixed domain and network configuration, increasing the grid resolution from 2 km to 1 km increases the number of grid

cells by about a factor of 4, and from 2 km to 500 m by about a factor of 16, which would substantially increase GPU memory demand and computational time on our current hardware.

Third, relative to the native CAMS coarse grid, a 2 km resolution is a significant improvement, allowing DACNO$_2$ to represent sharper horizontal gradients and more detailed urban-scale structures. Regarding spatial representativeness, 2 km can mitigate, but cannot fully eliminate the mismatch between model fields and point measurements because subgrid variability persists below 2 km. We mapped EEA measurements to the 2 km grid and excluded traffic stations, which are designed to represent very local conditions. The spatial representativeness analyses report indicates that NO_2 traffic sites are representative at sub-km^2 scales, whereas urban-background sites are often representative over areas on the order of several to dozens of km^2 (Kracht et al., 2017), supporting our choice of 2 km resolution. We have clarified these points in the manuscript.

We have added "*It motivates us to develop a 3D NO_2 product on a 2 km × 2 km horizontal grid (hereafter referred to as the 2 km grid) to better resolve fine-scale spatial heterogeneity and support the emerging high-resolution satellite missions (e.g., CO2M)*" in lines 57-59.

We have revised the content relative to the spatial representativeness of EEA stations to be "*EEA NO_2 was collected from background and industrial monitoring stations (European Environment Agency, 2024) and mapped onto the 2 km grid. Such stations have spatial representativeness of several to dozens of square kilometers, enabling cover our target grid size. However, traffic stations were excluded because their measurements represent a very local area (< 1 km^2), significantly smaller than the 2 km grid cells of our study (Kracht et al., 2017)*" in lines 191-195.

We have added "*Additionally, if targeting higher resolution (e.g., 1 km × 1 km or 500 m × 500 m), larger patches are required, resulting in an exponential increase in computational cost*" in lines 222-223.

2. Downscaling from 10 km to 2 km: Informational Limitations

The manuscript should more clearly explain which features actually provide meaningful spatial information for downscaling from 10 km to 2 km. Only a few indicators—such as geography, nighttime lights, and population—have resolution finer than 10 km, and these are largely time-independent. More dynamic and influential features (e.g., emissions, meteorology; shown in Figure 5) remain at 10–25 km resolution. Given this, it is unclear how the model captures high-resolution temporal variability.

Table 2 shows r/R^2, but the manuscript does not describe how r and R^2 were computed. I assume the metrics were calculated across all available records (i.e., combining space and time). To demonstrate that the model captures temporal variation—not only spatial variation—the authors should evaluate site-specific time-series performance, e.g.: compute r/R^2 for each site over time, then average these values across all sites. This would clarify whether the downscaling approach provides meaningful temporal improvements.

We agree with this suggestion to investigate how $DACNO_2$ captures temporal variability, given that the main dynamic inputs are available at coarse resolution. In the main manuscript, the Pearson correlation coefficient r and R^2 values reported in Table 2 are computed from all daily records across all EEA evaluation stations in the test year (2023), thereby accounting for both spatial and temporal variability. To specifically assess temporal performance, we conducted an additional station-specific time-series analysis as suggested. For each EEA evaluation station, r and R^2 were computed along the daily time series over 2023, and the resulting metrics were then summarized across stations. The results are provided in the added supplementary Fig. S3, including spatial distributions and boxplot statistics by station type.

This analysis shows that $DACNO_2$-Phase-3 achieves station-specific Pearson correlations comparable to CAMS-2km (CAMS-2km: r-mean = 0.85, r-median = 0.88; $DACNO_2$-Phase-3: r-mean = 0.84, r-median = 0.87), while exhibiting higher station-specific R^2 (CAMS-2km: R^2-mean = 0.09, R^2-median = 0.52; $DACNO_2$-Phase-3: R^2-mean = 0.23, R^2-median = 0.61). The improvements are most pronounced at urban stations and are accompanied by reduced very low R^2 cases at rural stations. The large difference between the mean and median R^2 is primarily driven by negative R^2 values at a subset of rural stations, possibly related to an uneven distribution of stations and to weaker, noisier signals in rural environments. Meanwhile, low R^2 stations are primarily located near boundaries and in mountainous areas, possibly due to limited spatial context for model inference and a complex environment.

Regarding the source of high-resolution temporal variability, $DACNO_2$ does not rely on fine-scale dynamic inputs. Instead, day-to-day variation is driven by coarse-resolution meteorology and temporal indicators, while fine-scale static inputs shape how this variation is distributed spatially on the 2 km grid. This interaction is learned through the phased training strategy, in which the large-scale 3D temporal behavior constrained by CAMS is preserved, and EEA constraints improve temporal consistency with surface observations on the 2 km grid. The results in Table 2 and station-specific time-series evaluation in Fig. S3 demonstrate that the framework improves spatiotemporal consistency rather than only refining spatial patterns.

In Section 3.1, we have clarified the calculation of Table 2 by adding "*The performance results were calculated across all paired measurements and model estimations. The station-specific time-series consistency analyses are provided in Fig. S3, where the results for each EEA evaluation station were calculated along the daily time series independently*" in lines 387-389.

We have also added "*Such improvement is consistent with station-specific time-series consistency analysis (Fig. S3). It indicates that $DACNO_2$-Phase-3 achieves station-specific Pearson correlations comparable to CAMS-2km (CAMS-2km: r-mean = 0.85, r-median = 0.88; $DACNO_2$-Phase-3: r-mean = 0.84, r-median = 0.87), while exhibiting higher station-specific R^2 (CAMS-2km: R^2-mean = 0.09, R^2-median = 0.52; $DACNO_2$-Phase-3: R^2-mean = 0.23, R^2-median = 0.61). The R^2 improvements are attributed to more high-R^2 sites at urban stations and fewer very low-R^2 sites at rural stations. The large difference between the mean*

and median R^2 is primarily caused by negative R^2 values at a subset of rural stations, likely due to uneven station distribution across the network and the challenge of modeling weaker, noisier signals in rural environments. Overall, these results suggest that the DACNO$_2$ model improves agreement with the independent EEA evaluation stations relative to the CAMS-2km baseline across station types" in lines 419-428.

We have also added "*In addition, Fig. S3 shows that DACNO$_2$-Phase-3 achieves better station-specific agreement in the central domain than near the boundaries. This may be due to boundary areas that lack sufficient spatial context and have complex mountainous terrain*" in lines 430-432.

In Section 4.1, we have added "*Overall, the DACNO$_2$ model is developed by combining multi-scale inputs and multi-source constraints. The fine-scale spatial structure on the 2 km grid is primarily informed by high-resolution emission-related proxies and geographic features, whereas large-scale spatiotemporal variation and vertical structure are driven by meteorological variables and temporal indicators. Through the phased training strategy, the CAMS constraint transfers large-scale spatiotemporal variation to the DACNO$_2$ model, and the EEA constraint guides the model to use fine-scale static inputs to shape this variation on the 2 km grid spatially*" in lines 563-568.

In Conclusions and Outlook, we have added "*At the 2 km grid resolution, most spatial detail is provided by high-resolution, time-independent geographic data and emission-related proxies. Meanwhile, large-scale variability is driven primarily by meteorological variables and temporal indicators at coarse scales. The DACNO$_2$ model learns, through a phased training strategy, how these dynamic coarse-scale drivers interact with fine-scale spatial inputs to improve the spatiotemporal representation of the NO$_2$ variability*" in lines 713-717.

The added figures in the Supplementary Materials are shown below:

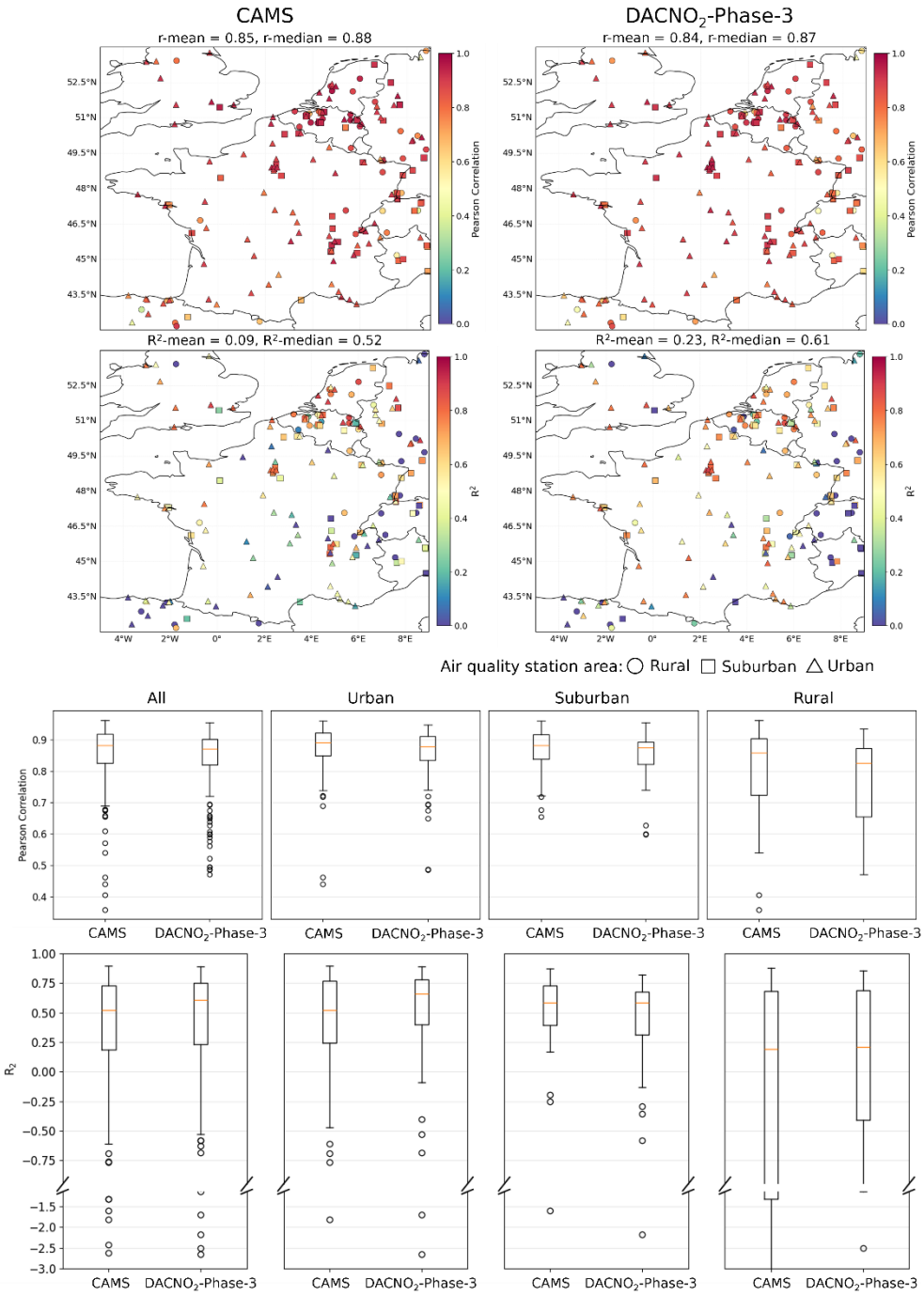

Figure S3. Station-specific performance of CAMS and DACNO$_2$-Phase-3 against EEA measurements in 2023. Model performance is evaluated at individual EEA stations using daily surface NO$_2$ measurements. For each station, statistics are calculated from the paired daily time series of measurements and model estimates over 2023. The upper panels show the spatial distribution of Pearson correlation coefficient (r) and coefficient of determination (R^2) for CAMS and DACNO$_2$-Phase-3. Station types (urban, suburban, rural) are indicated by different marker shapes. The lower panels summarize the same station-specific statistics using boxplots, allowing comparison of the distribution of performance metrics across different station groups. In each boxplot, the horizontal line indicates the median value. This figure characterizes how model skill varies geographically and across station environments at the individual-station scale.

3. Model Generalization and the Role of Fine-Tuning

For the best-performing model (Phase 3), the approach resembles a machine-learning data fusion method, since it directly fine-tunes using current observations. This raises concerns about whether it remains a fair comparison to Phase 1 and Phase 2, which do not use test data for training. The loss function appears unconstrained, relying heavily on ground measurements. This may lead to limited generalization, especially with sample-imbalance—EEA sites are often concentrated in high-pollution areas. I recommend analyzing the EEA observation distribution relative to the full domain, including: concentration distributions, vertical profiles, spatial representativeness.

Because ground stations are mostly located in urban or high-NO_2 regions, the fine-tuning may bias the model toward overestimating suburban and rural concentrations. This may explain: the higher Phase-2 biases in suburban and rural areas (Table 2), and why $DACNO_2$ performs worse than CAMS-2km in these regions (Tables 2 and 3). Addressing this issue may require physical constraints, emission priors, or sample rebalancing strategies. Even if not feasible within the current work, it warrants further discussion.

We appreciate this comment. The first question about a fair comparison between phases suggests that the current manuscript may not be fully clear. The $DACNO_2$-Phase-3 is the final model, whereas the other two phase models are included to investigate how the model evolves across three-phase training strategies. We have clarified it by adding "*Phases 1–3 represent successive development stages of the $DACNO_2$ model. The phase-to-phase comparison in Table 2 is used to quantify the incremental effect of adding constraints and the final adaptation step. In Phase-3, the fine-tuning step uses EEA observations from the training stations in 2023. All reported EEA-based metrics are computed on the held-out evaluation stations*" in lines 397-400.

We follow the suggestion to complement the analyses of the EEA station distribution by examining relative distribution density and the spatial relationship with average NO_2 concentration maps. The results are shown in Supplementary Fig. S1. We have also examined the interannual NO_2 variation from EEA measurements (Fig. S7). In addition, we examine the average time-series consistency between models and EEA NO_2, as shown in Fig. S4. It is observed that the NO_2 levels in 2023 are lower than in the Phase-2 training years (2019, 2021, 2022), which should primarily account for the overestimate of $DACNO_2$-Phase-2 for EEA NO_2 in 2023. The $DACNO_2$-Phase-3 reduces this bias, highlighting the importance of the adaptive fine-tuning step. Nevertheless, Fig. S4 indicates that $DACNO_2$-Phase-3 still slightly overestimates rural sites. This might be related to the uneven distribution of EEA stations, where higher-density coverage in high-NO_2 regions (Fig. S1) leads the model to fit the high-NO_2 situation and to overestimate at rural sites. This might be addressed by balancing the sample, and it warrants further investigation in future work.

In Section 2.2.2, we have added "*The spatial distribution of training and evaluation stations is shown in Fig. S1, along with the distribution density and the average surface NO_2 concentration map*" in lines 201-203.

In Section 3.1, we have added "*The average time-series consistency between models and EEA NO_2 is shown in Fig. S4*" in line 390, and "*Moreover, Table 2 and Fig. S4 show a positive bias for DACNO$_2$-Phase-2 in 2023. This offset is consistent with the fact that the NO_2 level in 2023 is lower than in the Phase-2 training years (2019, 2021, 2022), as illustrated in Fig. S7. DACNO$_2$-Phase-3 reduces the effect of the interannual variation while maintaining the temporal correlation, highlighting the role of the adaptive fine-tuning step*" in lines 407-411.

We have also added "*In addition, Fig. S3 shows that DACNO$_2$-Phase-3 achieves better station-specific agreement in the central domain than near the boundaries. This may be due to boundary areas that lack sufficient spatial context and have complex mountainous terrain. Additionally, a slight overestimation of DACNO$_2$-Phase-3 at EEA rural stations persists despite adaptive fine-tuning. A possible reason is the imbalance in the EEA constraint. Fig. S1 shows that most stations are located in urban and suburban areas with relatively higher NO_2 concentrations, whereas fewer stations are in rural areas. This may lead to positive bias in the model's estimates for rural areas, and the solutions require further investigation, including sample rebalancing strategies, expanding the study region to include more rural sites, and additional constraints. Meanwhile, given the R^2 definition, positive prediction bias at rural stations may be influential, as these stations generally have low NO_2 standard deviations and a smaller tolerance for prediction bias*" in lines 430-439.

We acknowledge that the current constraint strategy relies heavily on surface EEA measurements, which may leave the model not well constrained during future adaptive fine-tuning. In Conclusions and Outlook, we have added "*The constraint strategy still needs improvement, as the model's fine-tuning currently relies heavily on surface EEA measurements, which are biased due to uneven distribution, measurement methods, and spatial representativeness. Future development of DACNO$_2$ could incorporate constraints above the surface, such as integrating high-resolution 3D process-based NO_2 fields from models (e.g., WRF-Chem) and column observations from satellites, and embedding additional physical constraints into the loss function*" in lines 752-756.

The added figures in the Supplementary Materials are shown below:

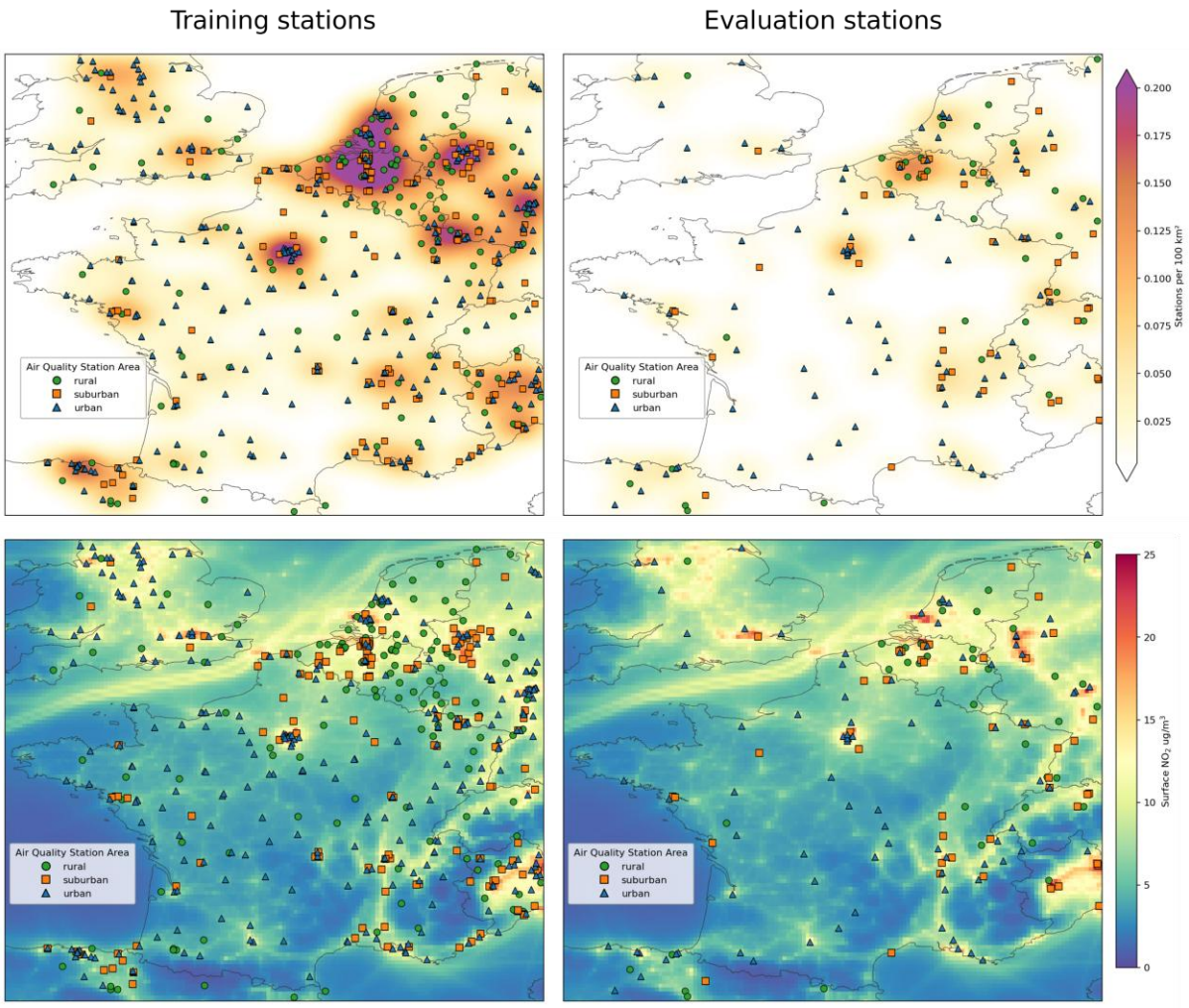

Figure S1. Spatial distribution of EEA surface NO_2 monitoring stations over the study domain, with station density and background surface NO_2 levels for context. EEA air-quality monitoring stations used in this work are shown by station area type (rural, suburban, urban). The upper panels show the spatial station density (stations per 100 km^2), estimated by smoothing station locations using a Gaussian kernel density estimator (KDE) on a 2 km grid with a bandwidth of 35 km. The lower panels show the CAMS 2023 mean surface NO_2 concentration as background context, allowing the station distribution to be interpreted relative to typical pollution levels across the domain.

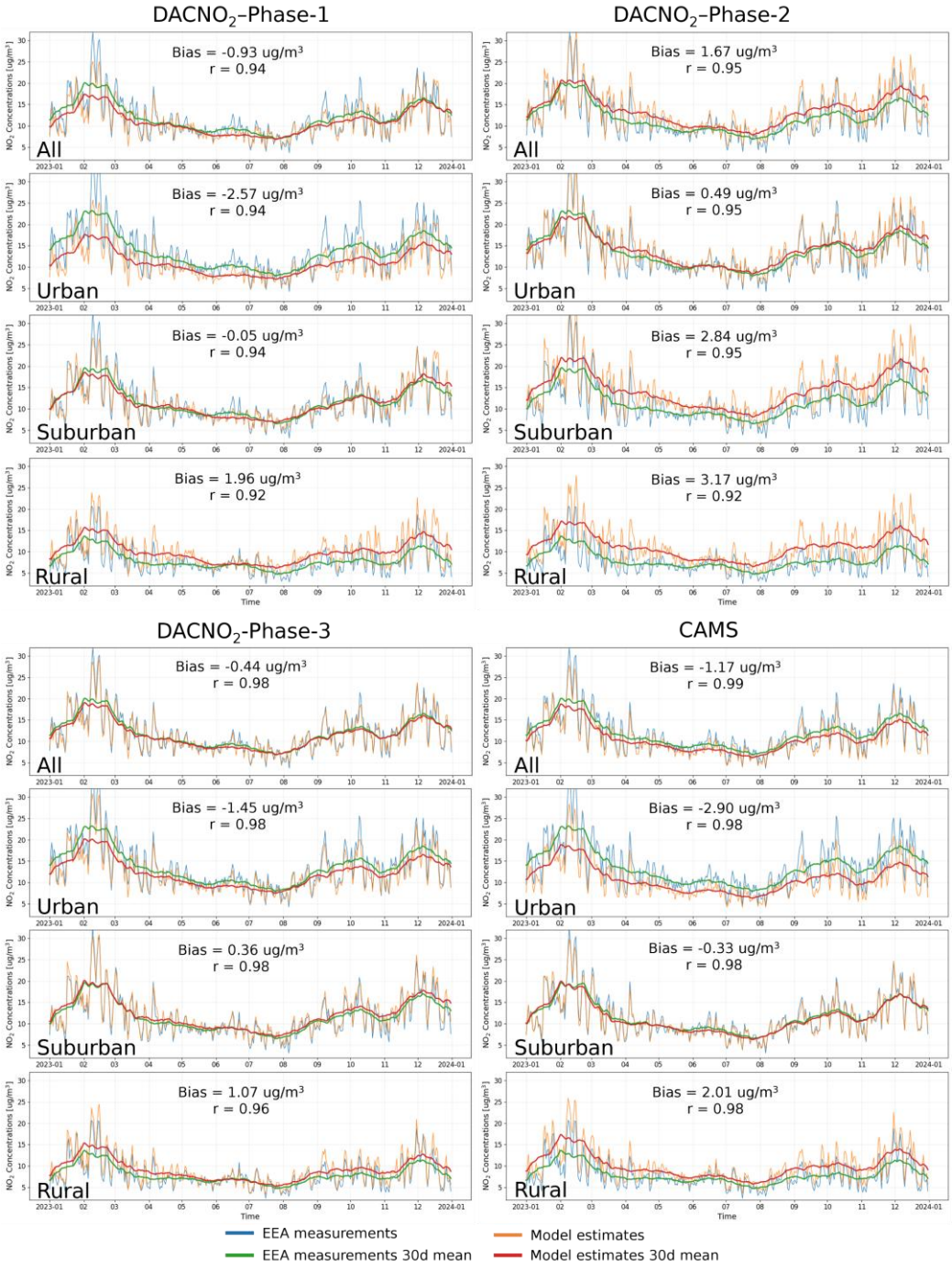

Figure S4. Daily comparison of surface NO_2 measurements and model estimates in 2023. Daily mean surface NO_2 time series for 2023 are shown for EEA measurements and for model estimates from CAMS and the three DACNO$_2$ phases. Results are presented for all stations and separately for urban, suburban, and rural stations according to EEA metadata. For each day, two station-network means are computed independently for measurements and model estimates by averaging over the stations where both values are available on that day. Thin lines show the resulting daily network means, and thick lines indicate the 30-day moving average to emphasize seasonal variability. Performance metrics (bias and Pearson correlation coefficient r) are calculated from the paired daily network-mean time series over the full year.

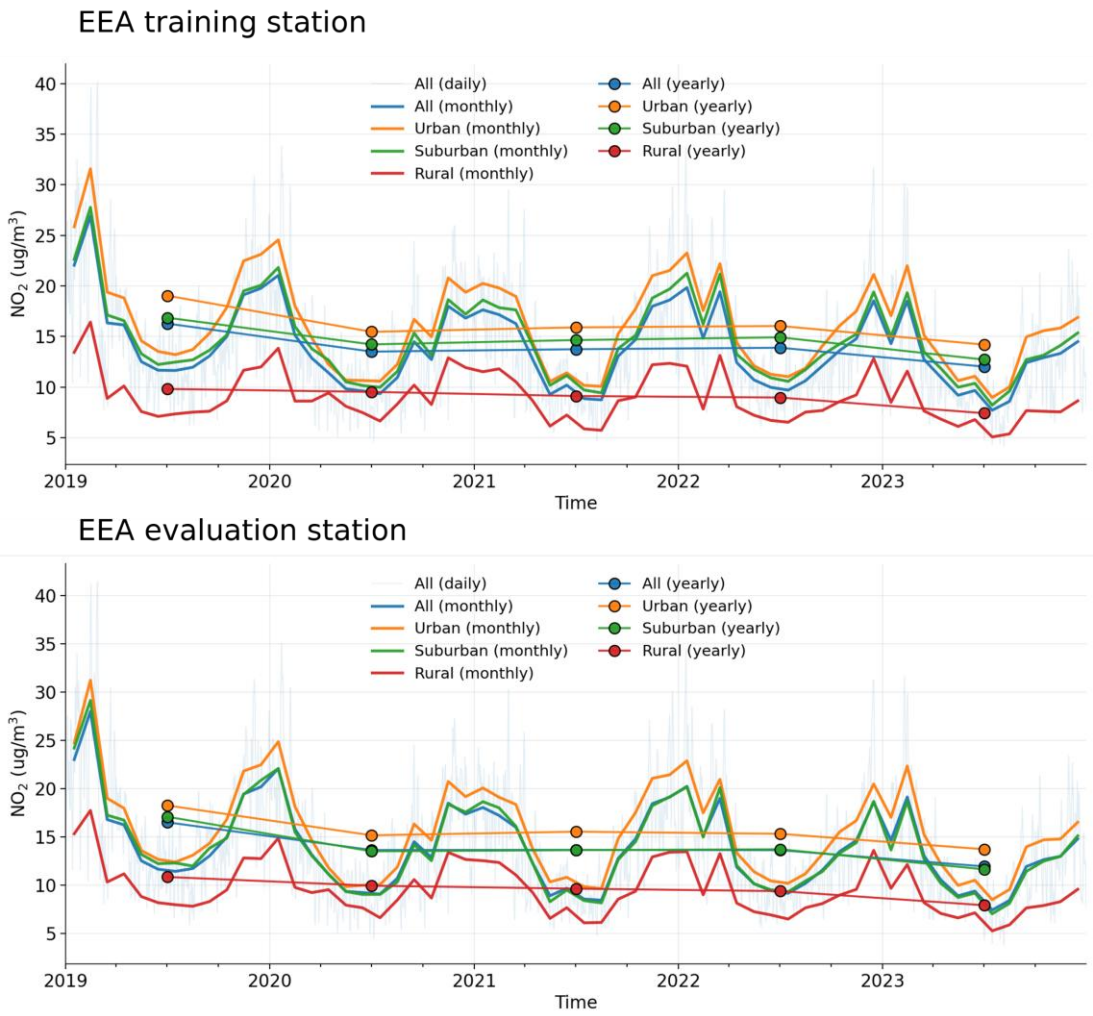

Figure S7. Long-term surface NO_2 measurements at EEA training and evaluation stations (2019–2023). Time series of measured NO_2 concentrations from EEA monitoring stations between 2019 and 2023. The upper panel shows stations used for model training, and the lower panel shows independent evaluation stations. The light blue background line represents the daily network-mean concentration, computed as the average across all available stations for each day. Colored curves indicate monthly averages for all, urban, suburban, and rural stations, while filled markers denote the corresponding yearly average. This figure characterizes the temporal variability and relative concentration levels across station types used in the study.

4. Vertical Profiles and 3D Validation

Since the main goal is to develop a 3D NO_2 field, the lack of vertical validation is a concern. Are there any available observed vertical profiles (e.g., MAX-DOAS, aircraft, lidar) that could be used? The manuscript should include a comparison of vertical profiles, not only absolute values but also relative changes across urban, suburban, and rural environments. Physically, one would expect that enhanced horizontal resolution should: preserve regional total columns, but increase near-surface NO_2 in urban areas, and decrease upper-layer NO_2 in rural areas due to plume dynamics. However, Figure 6 shows DACNO$_2$ (which version?) with substantially higher concentrations across all vertical levels over Paris. Has the vertical profile been influenced by data fusion with observations? Clarifying the source of this behavior is important for evaluating the scientific validity of the 3D fields.

The DACNO$_2$ model in Fig. 6 is DACNO$_2$-Phase-3, and we have clarified it in the figure caption. We agree that a more comprehensive examination of 3D NO$_2$ fields is necessary. We have conducted additional analyses of relative profile changes between phases and CAMS across environments, and the results are shown in Supplementary Fig. S10. Furthermore, we have calculated the relative column changes between DACNO$_2$-Phase-3 and CAMS in the study domain (Fig. S11).

In Section 4.2, we have added "*To assess how vertical profiles differ between the two models across environments, we analyzed the mean DACNO$_2$-to-CAMS profile ratio across the entire study region in urban, suburban, rural, and uninhabited environments classified based on population density (Fig. S10) and the urbanization definition (Dijkstra et al., 2021). The results indicate that the DACNO$_2$-Phase-3 adjustment is not a uniform scaling of the CAMS field. Instead, near the surface, DACNO$_2$-Phase-3 shows higher concentrations relative to CAMS in urban regions (about 6%) and lower concentrations in other areas (from about -20% to -1%). In the boundary layer (1000 m), the NO$_2$ concentrations are systematically higher in DACNO$_2$-Phase-3 compared to CAMS (from 10% to 30%), except in the uninhabited area (remains the same). At higher layers, DACNO$_2$-Phase-3 values converge to a lower ratio (about -22% to -15%) at 5000 m for the entire region. This behavior is also reflected in the layer-integrated column diagnostics shown in Fig. S11, which indicate near preservation of the regional column (0 – 5000 m), accompanied by a significant redistribution in the lower (0 – 1000 m) and conservative adjustment in upper (1000 – 5000 m) layers. Together, these results suggest that the DACNO$_2$-Phase-3 primarily redistributes NO$_2$ within the lower layers, enhancing horizontal contrast linked to human activity and emission strength, while maintaining consistently low estimates in the upper layers*" in lines 594-607.

We have also added "*Additionally, we assessed the profile ratio across the three phases (Fig. S10). The results indicate that applying EEA constraints almost systematically increases NO$_2$ estimates in DACNO$_2$-Phase-2 relative to DACNO$_2$-Phase-1, likely because the CAMS data used for pretraining in Phase-1 underestimates NO$_2$ at EEA measurement stations. In contrast, the EEA constraint reduces NO$_2$ estimates in DACNO$_2$-Phase 3 relative to DACNO$_2$-Phase-2, consistent with the lower surface NO$_2$ levels observed in 2023 compared with the training years (2019, 2021, 2022, Fig. S7). However, the boundary-layer NO$_2$ estimates exhibit different trends that do not align with phase-dependent changes, which warrants further investigation*" in lines 609-615.

We acknowledge that vertical validations of the DACNO$_2$ 3D NO$_2$ structure are necessary. However, this validation work would require daytime data rather than the daily average and an investigation of the appropriate process for vertical measurement data. Given that this work would significantly expand the manuscript and may dilute the focus on model development, we therefore place it in the scope for future work. We have added "*In this work, the vertical structure of DACNO$_2$ is assessed through comparison with CAMS. Independent evaluation against vertically resolved observations, such as MAX-DOAS or aircraft measurements, would be the next step in future work. Such an analysis would require an*

hourly version of the DACNO$_2$ fields to provide daytime data and the application of appropriate observation operators to ensure comparability between the model and the observational data" in lines 617-621.

The added figures in the Supplementary Materials are shown below:

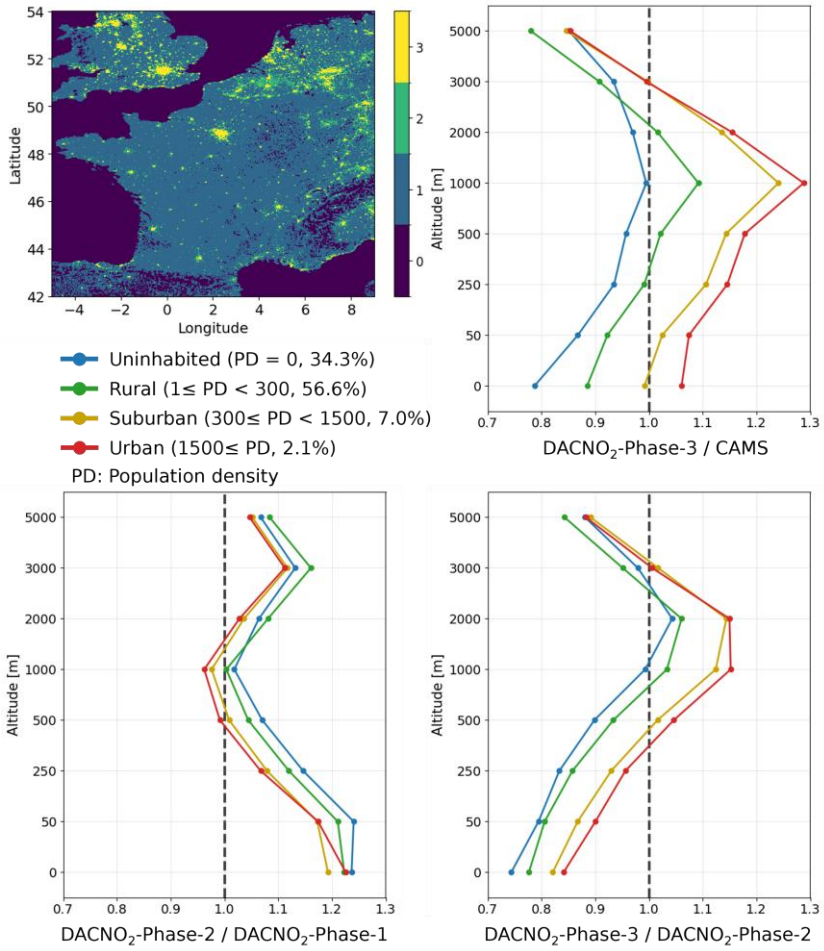

Figure S10. Vertical profile relative changes across urban, suburban, rural, and uninhabited environments. Relative changes of vertical NO_2 profiles between DACNO$_2$ model phases and CAMS across different environments. The upper-left panel shows the spatial distribution of four environment types derived from population density (0-uninhabited, 1-rural, 2-suburban, 3-urban) based on the JRC-GEOSTAT 2018 gridded population dataset (Silva et al., 2021) and the urbanization definition of Dijkstra et al. (2021). The remaining panels present mean vertical profile ratios for each environment type: DACNO$_2$-Phase-3 relative to CAMS (upper right), DACNO$_2$-Phase-2 relative to DACNO$_2$-Phase-1 (lower left), and DACNO$_2$-Phase-3 relative to DACNO$_2$-Phase-2 (lower right). Profiles are averaged across all grid cells within each class. The dashed vertical line denotes a ratio of 1.0, indicating no change between the two compared datasets.

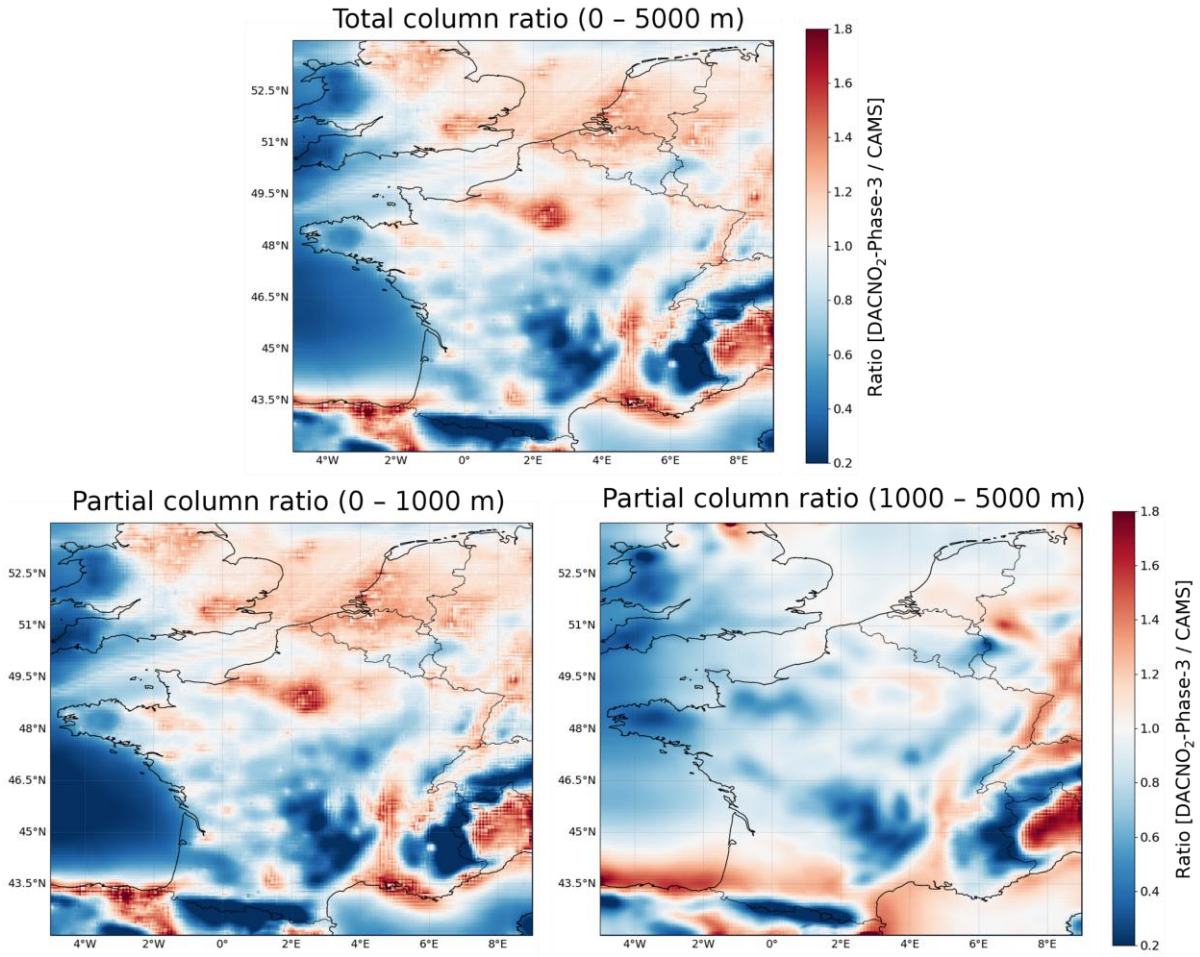

Figure S11. Comparison of vertically integrated NO_2 between $DACNO_2$-Phase-3 and CAMS. Ratios between $DACNO_2$-Phase-3 and CAMS for vertically integrated NO_2 over the study domain. The upper panel shows the full-column layer-integrated proxy, while the lower panels show integrations over 0–1000 m and 1000–5000 m. The integrated quantity is obtained by summing concentration multiplied by layer thickness on the common model levels, providing a consistent vertical mass proxy for relative comparison between the two models. A diverging color scale centered at 1 highlights relative enhancement and reduction. Area-weighted domain-mean ratios are 0.995 for the full column, 1.000 for 0–1000 m, and 0.949 for 1000–5000 m, indicating that the total column is largely preserved despite redistribution.

Minor Comments

(Line 21) Abstract: "Applying $DACNO_2$ a-priori profiles to TROPOMI retrievals increases tropospheric NO2 columns by 3% on average over those using European CAMS profiles," would benefit from a brief explanation. Is the increase primarily due to improved spatial resolution, enhanced vertical accuracy, or another factor?

The primary cause is that $DACNO_2$ profiles increase NO_2 levels in the emission region and decrease them in the low-emission region, a typical benefit of higher resolution. The details have been described in the response for the major comment "4. Vertical Profiles and 3D

Validation" above. We have added a brief explanation in the Abstract, and that sentence now reads "*Applying DACNO$_2$ a-priori profiles to TROPOMI retrievals increases tropospheric NO$_2$ columns by 3% on average over those using European CAMS profiles, with enhanced contrast between low- and high-NO$_2$ regions, primarily attributable to improved resolution.*" (lines 21-23)

(Line 156) "Notably, satellite-derived NO_2 products were deliberately excluded from the input features for two key reasons." Since satellite-derived NO_2 products are excluded, the training effectively relies on learning relationships of from emissions and meteorology to concentrations, similar to process-based CTMs. Please clarify whether there are sufficient training samples to support this, and discuss how the performance under this setup is validated.

In DACNO$_2$, satellite NO_2 products are excluded from the input features. Supervision in Phase-1 and Phase-2 is provided by the spatially continuous daily 3D NO_2 fields from the CAMS reanalysis, and Phase-2 to Phase-3 additionally introduces surface constraints from EEA stations. With the patching scheme (12 samples per day), the Phase-1 and Phase-2 training uses 2019, 2021, and 2022, yielding 13,140 patch samples, split 80% for training and 20% for validation, which is described in lines 329-330 (Section 2.4.1). Model performance under this setup is evaluated on the independent year 2023, using held-out EEA evaluation stations and CAMS fields, with metrics and discussion provided in Section 3.1 (Table 2). The results show that DACNO$_2$ is consistent with CAMS in the 3D structure across the three phases and improves agreement with independent EEA evaluation stations.

(Line 176) "CAMS NO_2 was processed by averaging hourly data to daily values and by bilinearly interpolating its horizontal resolution from 10 km to 8 km to match the model's scaling scheme." to 2km? Is bilinear interpolation to 2 km scientifically reasonable here? This approach appears to rely solely on mathematical interpolation without additional physical or high-resolution informational support. Please justify this choice.

The CAMS fields used as training targets are regridded from the native 10 km grid to the 8 km grid via bilinear interpolation to ensure grid alignment. This is an implementation choice to align with the factor-of-two scaling used in DACNO$_2$ and to enable loss computation during training. CAMS-2km is used solely for evaluation against EEA stations as a benchmark. We do not use CAMS-2km as a training target because the interpolated field would not provide additional physical information and could bias the target. To avoid confusion, we have added the sentence "*This regridding is used for grid alignment only and supports the computation of the loss function during training*" in lines 182–183.

(Line 320) "DACNO$_2$-Phase-2 was fine-tuned using only EEA NO$_2$ data from training stations during the test period (2023 in this study)". This raises concerns regarding information leakage and fairness when comparing Phase-2 to other models that do not use test-year data for training. Clarification or additional justification is needed.

DACNO$_2$-Phase-3 is the final model in this study. The comparison of three-phase models aims to assess the incremental effect of adding constraints and fine-tuning adaptation. To clarify, we have refined this sentence to be "*DACNO$_2$-Phase-3 is initialized from the DACNO$_2$-Phase-2 weights and further fine-tuned using EEA NO$_2$ data from the training stations during the test period (2023 in this study) to reflect a typical application scenario*" in lines 356-358, and have added "*Phases 1–3 represent successive development stages of the DACNO$_2$ model. The phase-to-phase comparison in Table 2 is used to quantify the incremental effect of adding constraints and the final adaptation step. In Phase-3, the fine-tuning step uses EEA observations from the training stations in 2023. All reported EEA-based metrics are computed on the held-out evaluation stations*" in lines 397-400.

Tables 2-3 are difficult to interpret due to their complexity. Consider simplifying the presentation, e.g., using a comparison bar chart or reorganizing the table to more clearly highlight performance differences across models.

We have revised the layout and captions of Table 2 to make comparisons clearer. The modifications include (1) separating the EEA-based evaluation (2 km) and the CAMS-based evaluation (10 km) into two distinct panels, (2) clarifying the evaluation resolution and baseline in the table note, and (3) highlighting the best-performing values within each row using boldface and an asterisk (*). We hope these changes reduce the effort needed to identify the main performance differences while keeping the full set of metrics. The same formatting and clarification were applied to Table 3 and Table S1.

References

Dijkstra, L., Florczyk, A. J., Freire, S., Kemper, T., Melchiorri, M., Pesaresi, M., and Schiavina, M.: Applying the Degree of Urbanisation to the globe: A new harmonised definition reveals a different picture of global urbanisation, Journal of Urban Economics, 125, 103312, 10.1016/j.jue.2020.103312, 2021.

AirBase - The European air quality database: https://eeadmz1-downloads-webapp.azurewebsites.net/, last access: June 2024.

Kracht, O., Santiago, J. L., Martin, F., Piersanti, A., Cremona, G., Righini, G., Vitali, L., Delaney, K., Basu, B., Ghosh, B., Spangl, W., Brendle, C., Latikka, J., Kousa, A., Pärjälä, E., Meretoja, M., Malherbe, L., Letinois, L., Beauchamp, M., Lenartz, F., Hutsemekers, V., Nguyen, L., Hoogerbrugge, R., Eneroth, K., Silvergren, S., Hooyberghs, H., Viaene, P., Maiheu, B., Janssen, S., Roet, D., and Gerboles, M.: Spatial representativeness of air quality monitoring sites – Outcomes of the FAIRMODE/AQUILA intercomparison exercise, Publications Office of the European Union, 10.2760/60611, 2017.

Silva, F. B. e., Poelman, H., and Dijkstra, L.: JRC-GEOSTAT 2018 [dataset], 2021.