Response to Reviewers' Comments

Dear Editor,

Thank you very much for your efforts in handling and evaluating our submission.

We sincerely thank the reviewers' careful evaluation of our manuscript and for their constructive comments and suggestions, which have greatly contributed to improving the quality and clarity of the manuscript. We have prepared a detailed, point-by-point response to all comments and revised the manuscript accordingly by carefully considering each suggestion. We believe that the revised manuscript has been substantially improved and is now suitable for publication in *Earth System Science Data*.

The major revisions are summarized as follows:

(1) Additional descriptions of the observational data used for model training and evaluation have been included to better characterize the spatial distribution of the observations.

(2) The performance of the overall model across different continents has been presented to provide a clearer assessment of the dataset's spatial applicability.

(3) The manuscript text has been revised to better clarify the applicability of the dataset, emphasizing its robustness in regions with relatively dense monitoring coverage.

(4) Some figures have been improved to enhance readability and clarity.

(5) Minor revisions have been made to the manuscript text and formatting.

More details about these revisions can be found in the revised manuscript and in our point-by-point responses to the reviewers below.

Best regards,

Yuqiang Zhang on behalf of all the co-authors

The reviewer's comments are listed below in *black italics*, and our responses and changes in the manuscript are highlighted in blue and red, respectively.

Response to Reviewer #1:

Comments:

In this study, the authors developed an advanced Air Transformer deep learning model and constructed a continuous global daily NO2 concentration dataset for the period of 2005 to 2023 based on satellite observations, ground site monitoring, meteorological reanalysis, land-use information, and auxiliary geophysical variables. The dataset has exhibited robust performance with improved spatial consistency. NO2 concentration trends in key regions globally and their crucial drivers were analyzed and properly reasoned. The manuscript is well organized and written. Overall in my opinion, the manuscript is acceptable for publication on Earth System Science Data after minor revision.

We sincerely thank the reviewer for the positive and encouraging assessment of our work. We greatly appreciate the recognition of the novelty of the Air Transformer framework, the construction of the long-term global daily NO_2 dataset, and the evaluation of its spatial consistency and performance. We are also grateful for the reviewer's acknowledgment of the organization and clarity of the manuscript.

Major comments:

Introduction: The spatiotemporal resolution is considered to be the main novelty in this study. Therefore, the resolution and accuracy of the current state-of-the-art methods for developing NO2 concentration datasets is suggested to be introduced in this section.

Response: We thank the reviewer for the suggestion. We have added a concise description in the Introduction summarizing the spatial and temporal resolution, as well as the reported accuracy, of existing NO_2 datasets. The added text is inserted after line 56 and provided below.

Revision: In recent years, several global or regional NO_2 products developed using these approaches have achieved spatial resolutions ranging from ~0.25° to 0.1°, with temporal resolutions from annual or monthly averages to limited daily estimates, and reported performance of $R^2 \approx 0.6–0.8$ (Larkin et al., 2023; Long et al., 2022; Shao et al., 2023; Sun et al., 2024; Wei et al., 2022; 2023).

Section 3.1: More concrete information on key indicators should be provided for the comparison between this study and previous studies. Also, is the LUR model the most widely accepted in generating NO2 datasets so far? If not, more studies could be included in the comparison.

Response: We thank the reviewer for this thoughtful comment. We have added more quantitative performance indicators when comparing our results with previous studies. Regarding the use of LUR for comparison, we selected the SoGA LUR-based NO_2 product because it has been adopted in the "State of Global Air" assessment for estimating global NO_2-related health burdens, and therefore represents a policy-relevant and authoritative benchmark dataset. We have clarified this rationale in the revised manuscript.

Revision:

Line 228:

"Overall, the performance metrics obtained from both validation strategies demonstrate that the AiT-NO_2 model performs consistently well across multiple temporal scales and validation settings. Compared with previous studies, our framework not only provides improvements in spatial resolution (0.1°) and temporal coverage (daily estimates from 2005–2023), but also achieves higher or comparable predictive performance. For example, reported random cross-validation R^2 values in previous regional NO_2 modeling studies typically range from 0.70 to 0.88, with RMSE values between approximately 3–6 ppbv depending on region and temporal aggregation (Chan et al., 2021; Shao et al., 2023; Wong et al., 2021). In contrast, our model achieves an R^2 of 0.91 and an RMSE of 2.32 ppbv for daily predictions under random cross-validation at the global scale. Notably, many earlier studies were conducted at regional or national scales, often benefiting from dense monitoring networks and region-specific tuning, whereas globally consistent daily surface NO_2 datasets with comparable resolution remain relatively limited. Furthermore, The Spatial cross-validation results are generally comparable to, or exceed, those reported in previous global and regional NO_2 modeling studies (Wei et al., 2019; 2023), underscoring the robustness and reliability of the AiT-based framework for constructing spatially continuous, long-term NO_2 datasets."

Line 245:

"As we mentioned earlier, the SoGA2024 for the first time included NO_2 in its report, as well as the annual NO_2 concentration datasets they generated using a global LUR model (Anenberg et al., 2022; Larkin et al., 2023). The LUR-based dataset was selected for comparison due to its adoption in the State of Global Air (SoGA) assessment for global NO_2 health burden estimation, where it functions as an authoritative and policy-relevant exposure benchmark."

Section 3.2: The drivers of the Northern and Southern Hemisphere trends could be discussed more in detail, which can be linked to the discussion in Section 3.3.

Response: We thank the reviewer for this valuable suggestion. In the revised manuscript, we have expanded the discussion of the contrasting NO_2 trends between the Northern and Southern Hemispheres by incorporating more detailed explanations of emission structure, economic development pathways, regulatory interventions, and the relative influence of natural sources.

Revision:

Line 285: "This upward phase coincides with rapid industrial expansion, increasing fossil fuel consumption, and accelerated urbanization in several emerging economies, which contributed to sustained growth in anthropogenic NO_x emissions during this period. Subsequently, despite a decrease to 3.73 ppbv by 2019, concentrations remained relatively high. The post-2015 decline is broadly consistent with the implementation of stricter emission standards, large-scale installation of pollution control technologies, and structural shifts in energy systems in major emitting regions. These policy-driven reductions partially offset emission growth from ongoing economic activities. In 2020, due to the significant reduction in global economic activities and transportation caused by the COVID-19 pandemic, NO_2 concentrations notably dropped to 3.42 ppbv. This abrupt decline highlights the strong sensitivity of surface NO_2 to short-term changes in transportation intensity and industrial output, underscoring the dominant contribution of anthropogenic combustion sources. From 2021 to 2023, with the gradual recovery of economic activities, NO_2 concentrations rebounded slightly in 2021 but then declined again in 2022 and 2023, reaching 3.38 ppbv in 2023. The post-pandemic evolution suggests that while mobility and economic activity resumed, ongoing structural adjustments in emission sources and continued regulatory efforts may have moderated the rebound in NO_2 concentrations."

Line 298: "Figures S5b and S5c show that the NO_2 concentrations in the Northern Hemisphere and tropic regions exhibited similar trends. This similarity reflects the dominance of anthropogenic NO_x emissions in the Northern Hemisphere, which accounts for the majority of global industrial production, transportation activity, and energy consumption. As a result, large-scale emission control policies and economic transitions in this hemisphere exert a disproportionate influence on global NO_2 variability."

Line 312: "Compared with the Northern Hemisphere, the Southern Hemisphere is characterized by lower anthropogenic emission densities, different energy consumption patterns, and a relatively larger influence of natural sources such as biomass burning and lightning-produced NO_x. In particular, interannual variability associated with fire activity and meteorological conditions may play a more prominent role in shaping regional NO_2 patterns. Moreover, the smaller extent of highly industrialized urban clusters limits sustained long-term anthropogenic-driven growth in NO_2 concentrations."

Line 301: How is "high-income" defined as a region? The definition should be proposed upon its first occurrence.

Response: We thank the reviewer for this helpful comment. Here, "high-income" refers to the high-income super-regions defined in the Global Burden of Disease (GBD) regional classification framework, in which countries are grouped according to similar cause-of-death patterns and epidemiological characteristics. We define the several super regions as used in GBD from line 348, as well as a map provided in the Supplementary Information (Fig. S2) for reference.

Section 3.3: Are there any new findings from this study based on the more refined NO_2 concentration dataset?

Response: We thank the reviewer for the comment. In this study, the primary objective is to develop and validate an advanced, spatially and temporally refined global NO_2 dataset. Accordingly, Section 3.3 focuses on describing the fundamental spatiotemporal distribution patterns and regional evolution revealed by the dataset, rather than conducting an in-depth mechanistic or causal analysis. The key contribution here lies in demonstrating how the high-resolution daily dataset enables a more detailed characterization of geographic heterogeneity, regional hotspots, and population-weighted exposure patterns at the global scale.

More comprehensive analyses on population exposure, environmental inequality, and associated health implications based on this dataset are being prepared for subsequent dedicated studies.

Section 5: The advantages of the Transformer-based model compared to common machine learning models could be further addressed in the last section.

Response: We thank the reviewer for this valuable suggestion. In the revised manuscript, we have further elaborated the methodological advantages of the Transformer-based framework compared to conventional machine learning models in the final section.

Revision: Line 407: "Compared with conventional machine learning approaches, the Transformer-based AiT framework offers advantages in modeling long-range spatial and temporal dependencies through its attention mechanism. This structure allows the model to dynamically weight multi-source inputs and capture complex non-linear interactions among meteorological, geophysical, and satellite-derived variables. Such capability enhances the stability of predictions across heterogeneous regions and improves spatial transferability relative to traditional regression-based or tree-based models."

Response to Reviewer #2:

Major Comments:

This dataset developed by Mu et al. presents an exciting potential improvement to global estimates of daily surface-level NO_2. Global surface-level NO_2 products have historically been unable to capture localized impacts due to reliance on coarser satellite-derived data. The Air Transformer approach is especially promising, and the authors have done a nice job of presenting their findings in a clear coherent way. I especially thought that the figures were well conceived and interesting to interpret.

Response: We sincerely thank the reviewer for the careful and thoughtful evaluation of our manuscript. We greatly appreciate the recognition of our methodological innovation, the potential impact of the dataset, and the clarity of presentation. We also acknowledge the concerns raised regarding specific aspects of the evaluation and interpretation. We have carefully considered all comments and substantially revised the manuscript to address these issues. We believe that these revisions have significantly strengthened the rigor, clarity, and transparency of the study.

Although I generally think this is a strong paper that presents some exciting potential high impact advances, there are a number of issues that I have identified that make me unable to recommend this dataset for publication at this time. If these issues are resolved appropriately, I believe that this dataset could be a nice addition to ESSD.

First, the major issue I had with this work is that throughout the paper the authors make claims that their dataset is improved at predicting in "diverse regions" or in quantifying "localized impacts"; however, I do not believe they have done sufficient evaluation to evidence these claims. There are no summary statistics regarding the regional number and values of the training data. Additionally, there is no evaluation in regions that are historically undermonitored such as many countries in the Global South and rural areas. It would also be interesting to compare performance in these regions to past studies.

Response: We thank the reviewer for raising this point regarding regional representativeness and performance evaluation. For this study, our model is trained and evaluated at the gridded level, where observations are harmonized and aggregated to a common spatial resolution prior the modeling. To make it clear, firstly, in the revised manuscript, we report summary statistics regarding the regional number and values of the training data at the gridded level (Tab. S1). This provides a view of data

density and value ranges across regions. These additions clarify differences in observational density across regions.

Secondly, we now report continent-specific performance metrics derived from the existing global random cross-validation framework. Specifically, the model was trained at the global scale, and out-of-fold predictions from random cross-validation were retained. We then stratified these held-out predictions by continent based on geographic coordinates and computed R^2, RMSE, MAE for each continent (Tab. S2).

Third, we acknowledge that many countries in the Global South and rural areas remain comparatively under-monitored. In such regions, predictions rely more heavily on spatial covariates and large-scale drivers rather than dense local measurements. Nevertheless, the model is trained globally using diverse climatic, emission, and land-use conditions, enabling it to learn generalized relationships between NO_2 and its governing factors. As a result, predictions in sparsely monitored regions exhibit reasonable magnitude ranges, even where local observations are limited. While independent spatial validation is inherently constrained by data scarcity in these regions, the stability of model performance across well-monitored continents supports the reliability of the framework for generating spatially continuous global estimates.

Revision: Line 83: "Finally, daily observations were spatially aggregated to unique grid locations based on geographic coordinates, and the distribution of sample counts and NO_2 descriptive statistics is summarized in Table S2."

Line 223: "In addition to the global evaluation, continental-scale performance is summarized in Table S2. Under random cross-validation, the model maintains high predictive accuracy across Europe, Asia, North America, and South America (R^2 = 0.89–0.92). Spatial cross-validation results show greater variability across continents, with R^2 values ranging from 0.54 to 0.68 in regions with dense monitoring coverage. Performance in Africa and Australia is substantially lower under spatial cross-validation, which is attributable to the limited number of available monitoring grids in this region."

Second, while it is nice to see that the authors compared to a past estimate from Anenberg et al. 2022, I suggest they compare to the newer product from Larkin et al. 2023 (cited below) as well as the surface-level estimate from Cooper et al. 2022. The latter is slightly different from this study in that it estimates NO2 at 1pm (TROPOMI overpass time) but it would still be interesting to compare the statistical performance of this dataset to that one.

Response: Thanks for the suggestion. We carefully reviewed the manuscript of Larkin et al. 2023 and

noted that NO_2 dataset from this study was not publicly available. However, Larkin et al. (2023) also reported model performance metrics from cross-validation. Specifically, their global daily model captured 47% of variation ($R^2 \approx 0.47$) with a daily RMSE of 6.8 ppb, and monthly and annual R^2 values of 0.59 and 0.63, respectively. In comparison, our AiT-NO_2 model achieves higher predictive performance under random cross-validation, while also maintaining stable performance across temporal aggregations.

Besides, we compared our dataset with the surface NO_2 estimates from Cooper et al. (2022). It should be noted that the Cooper et al. product represents NO_2 concentrations at approximately 13:00 local time corresponding to the OMI/TROPOMI overpass, whereas our dataset represents daily mean surface NO_2 concentrations. Because of the strong diurnal variability of NO_2, due to the emission and atmospheric photochemistry, afternoon concentrations are generally lower than daily averages, and therefore the absolute values are not directly comparable. When comparing the interannual variability at the continental scale, our dataset generally exhibits smoother temporal evolution, whereas the Cooper et al. estimates show more abrupt year-to-year fluctuations in several regions. For most continents (Africa, Europe, North America, Oceania, and South America), the two datasets show broadly comparable long-term tendencies despite differences in short-term variability. For Asia, Cooper et al. dataset shows that NO_2 concentrations drop sharply between 2014 and 2015 with a magnitude that is substantially larger than typical year-to-year variability. Such an abrupt change is unlikely to reflect a realistic continental-scale interannual variation. A detailed examination suggests that this anomaly is attributable to the use of an incorrect emission inventory for China in 2015 (Zhang et al., 2021), which disproportionately biases the regional aggregation. And the year in which NO_2 decreased in the Cooper dataset was earlier than ours. In contrast, the AiT dataset represents daily mean surface NO_2 concentrations constrained by multiple geophysical and atmospheric predictors, resulting in a more temporally consistent representation of large-scale NO_2 evolution.

Revision:

Line 265: “In addition, we further compared our product with another independent NO_2 dataset (Copper et al., 2023) to provide a broader evaluation of model performance (Fig. S10).”

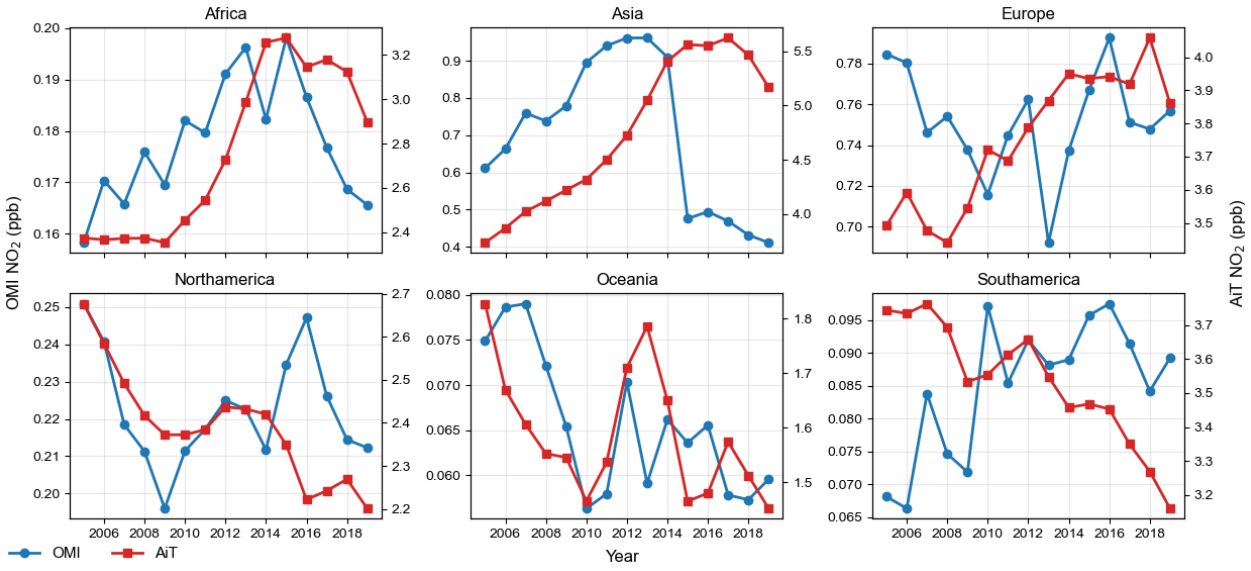

Figure S10. Comparison of continental NO_2 trends between the AiT dataset and Cooper et al. (2022)

Third, I wonder if the authors could consider a few additional predictor variables specifically tied to (1) anthropogenic features (e.g., road systems, built environment, etc.), and (2) wildfire activity. I guess TROPOMI partially captures the latter, but I am curious if including some marker of wildfires would have a notable effect on the model predictions and/or performance.

Response: We appreciate the reviewer's suggestions to include more localized predictors. We would like to clarify that most of the anthropogenic and physical features suggested by reviewer (e.g., road systems, built environment, etc.) have already been incorporated into our model as input features. We apologize if this was not sufficiently clear in the text. Specifically, we included ***road network density*** to represent infrastructural influences. We used ***population density (WorldPop)*** and ***land use data (MCD12C1)*** to characterize human activity and urban morphology. We used ***Digital Elevation Model (DEM)*** data to account for the impact of terrain on NO_2 dispersion. We included ***NDVI*** to account for vegetation cover. These predictors collectively represent built environment intensity, human activity, vegetation distribution, and infrastructure density, which are closely aligned with the anthropogenic features suggested by the reviewer. At the same time, regarding wildfire activity, we agree that explicit fire indicators (e.g., burned area or burning emission) could potentially provide additional information in fire-prone regions. In the current framework, wildfire signals are partially reflected through satellite-derived atmospheric variables (including TROPOMI observations), as well as through land cover, NDVI dynamics, and meteorological covariates that respond to fire events. Therefore, wildfire effects are not entirely absent from the predictor space. We acknowledge that incorporating explicit wildfire markers may further refine predictions in regions where biomass burning is a dominant source.

However, systematically integrating global fire datasets would require re-training and re-tuning the full modeling pipeline. Given the scope of the current study and the already strong cross-validated performance, we consider this an important direction for future model extensions. We have clarified relevant outlook in the revised manuscript.

Revision: Line 425: "Future work should focus on expanding data sources, such as emissions inventories, traffic data, and other dynamic activity indicators, to further improve the model's accuracy, especially in less urbanized regions. Incorporating additional episodic drivers (e.g., wildfire-related products in fire-prone regions) may provide incremental improvements where such processes substantially influence NO_2 variability. These refinements would enable more comprehensive global assessments of NO_2 pollution and contribute to the development of more targeted and effective air quality management strategies."

Lastly, it is good that both a spatial and random cross-validation was performed but I am curious if you could perform a temporal cross-validation especially given that "annualizing" the data appears to have only a minor effect on performance for the spatial cross-validation. I also wonder if instead of doing a random spatial cross-validation, you could assess performance in specific regions (maybe the GBD super regions) to address the first point I brought up.

Response: We thank the reviewer for this thoughtful suggestion. We agree that temporal cross-validation can be informative, particularly if the primary goal is forecasting to future years or ensuring uniform performance across macro-regions. In this study, however, the main objective is spatially continuous high-resolution mapping; therefore, we prioritized spatially CV to directly quantify out-of-location generalization, and used random CV only as a reference.

Implementing temporal CV would require re-running the complete training and tunning process under time-based splits, which is computationally intensive for our gridded dataset. The relatively minor improvement in performance after annual aggregation under spatial CV reflects the dominant role of spatial generalization error. Spatial CV evaluates model performance at previously unseen locations, where errors are primarily driven by spatial heterogeneity in emissions, land use, and regional characteristics rather than short-term temporal variability. Thus, we don't add additional temporal CV.

Annual aggregation reduces high-frequency meteorological noise, but it does not mitigate spatial extrapolation uncertainty. This limited improvement from daily to annual scales is also consistent with the intrinsic characteristics of NO_2 as a chemically reactive pollutant with a relatively short atmospheric lifetime. Its concentrations are strongly influenced by local emissions and rapidly varying meteorological conditions, leading to substantial temporal variability that is not fully suppressed by

temporal aggregation. Similar behavior has been reported in previous studies (Wei et al., 2022; 2023), where aggregation to longer timescales yields only modest improvements in model performance for short-lived pollutants such as NO_2. Therefore, only modest gains in R^2 are observed when moving from daily to annual timescales under spatial CV. This result suggests that model performance is constrained primarily by spatial transferability rather than temporal noise.

In addition, we explored the feasibility of region-based CV (i.e., leaving out entire regions). However, the limited number and uneven distribution of monitoring sites across regions lead to insufficient training samples and unstable evaluation results under such a scheme. Constructing multiple balanced region-based splits is also not practically feasible and would require substantial additional computational cost. Moreover, coordinating multiple alternative CV strategies (e.g., temporal, regional, and spatial CV) would introduce considerable computational burden without necessarily providing proportionate gains in interpretability. Notably, most previous studies (Tao et al., 2024) have predominantly adopted random site-based CV rather than region-based CV, largely due to similar constraints in monitoring site density and spatial representativeness, which further supports our choice of validation strategy.

Overall, these considerations indicate that model performance is primarily constrained by spatial transferability rather than temporal variability, and spatial CV remains the most appropriate validation framework for the objectives of this study.

Regarding region-specific evaluation, we have now included continent-level performance metrics derived from the global cross-validation framework, as presented in Tab. S2.

Revision:

"Table S2. Continental evaluation of the final NO_2 model under random and spatial cross-validation schemes (R^2, RMSE, and MAE).

	Random CV			Spatial CV		
Continents	R^2	RMSE	MAE	R^2	RMSE	MAE
Europe	0.92	4.12	2.82	0.54	7.30	5.16
Asia	0.89	6.08	4.20	0.59	8.41	5.63
North America	0.90	4.43	2.99	0.68	6.06	4.29
South America	0.89	6.10	4.14	0.58	8.92	6.11
Africa	0.78	4.90	2.92	0.04	10.29	6.39
Oceania	-	-	-	-	-	-

Note: Results for Australia are not reported due to the limited number of samples, which precludes statistically robust evaluation."

Minor Comments:

L29-32: Agreed that the anthropogenic sources dominate emissions; however, given that you are

developing a dataset that includes rural / non-urban estimates, I think it is important to mention the natural sources (e.g., soil NOx, lightning NOx, wildfire NOx) as well.

Revision: We have added the following sentence in the Introduction:

Line 33: "In addition to anthropogenic sources, natural processes such as soil NO_x emissions, lightning-produced NO_x, and biomass burning also contribute to background NO_2 levels, particularly in rural and remote regions where monitoring coverage is sparse (Fei et al., 2016; Hall et al., 1996; Xing et al., 2018)."

L39-40: And also, primarily in the Global North as NO2 monitoring infrastructure in e.g., Africa and much of Latin America is lacking. Calls into question that uncertainty is likely higher around NO2 in developing regions.

Revision: Following the reviewer's comment, we revised the relevant description:

Line 43: "Ground-based monitoring networks, while essential, are unevenly distributed, with most stations concentrated in high-income urban areas (Huang et al., 2023; Cooper et al., 2022; Di et al., 2019; Huang et al., 2018), whereas monitoring infrastructure in regions such as Africa and much of Latin America remains severely lacking. This spatial imbalance leads to greater uncertainty in NO_2 estimates for developing regions."

L54: And also, a well-documented inability to capture NO2 in rural areas and areas with low concentrations that has implications for estimating background concentrations (see DOI: 10.1029/2021GL092783)

Revision: Thanks for the comments. The new sentence is below:

Line 59: "However, existing approaches often suffer from several limitations, such as lacking the spatial and temporal resolution needed to accurately reflect local variations, a documented difficulty in accurately representing NO_2 concentrations in rural and low-concentration regions, and failing to incorporate essential geophysical and atmospheric parameters (Di et al., 2019; Qu et al., 2021)."

L57-58: "the model overcomes retrieval uncertainties and better captures local variations in NO2 concentrations" Curious to hear how you will quantify this, I will read on.

Revision: Thanks for the comments. We acknowledge that "overcomes" might be too strong a term, and we have revised it to better maintain the scientific rigor.

Line 64: "By taking advantage of the unique strengths of AiT in predicting atmospheric pollutant

concentrations which will be discussed later, and incorporating important geophysical, atmospheric, and socio-economic parameters, the model mitigates the retrieval uncertainties and better captures local variations in NO_2 concentrations, resulting in a more accurate and comprehensive understanding of NO_2 dynamics on a global scale."

Introduction: Somewhere in the introduction I think it is worth writing a sentence or two more specifically on the associated health effects of NO2, especially given this dataset's relevance for long-term epidemiological cohort studies.

Revision: Thanks for pointing this out. We expand the description of the NO_2 effect.

Line 27: "As an important precursor of ground-level ozone (O_3) and secondary fine particulate matter ($PM_{2.5}$), NO_2 strongly influences atmospheric oxidation processes and regional pollution levels (Li et al., 2019; Xue et al., 2014), contributing to respiratory and cardiovascular diseases while disproportionately affecting vulnerable populations (WHO, 2021; Freire et al., 2010; Sentís et al., 2017; Kim et al., 2014; Schmidt, 2019; Chowdhury et al., 2021)."

L65-68: Can you include a map and a summary statistics table of regional-level information from these monitors either in the main or supplement. I am curious to see the number of monitors and observations as well as statistics (e.g., mean, standard deviation, minimum, maximum, and percentage of observations that were removed from the QAQC). Especially given your claim in the introduction that the model better captures local variation I think it is necessary to understand where (and what time periods) the data are available for in historically undermonitored areas in the Global South.

Revision: Thanks for the suggestion. We included a Table that shows the summary statistics of the training dataset by continent, including sample counts, unique spatial grid counts, and mean surface NO_2 ($\mu g/m^3$) concentrations.

"Table S1. Summary statistics of the training dataset by continent, including sample counts, unique spatial grid counts, and mean surface NO_2 ($\mu g/m^3$) concentrations.

Continents	Samples	Grids	Mean	Min	Max	Std
Europe	8506985	2728	18.8	0	221.5	14.3
Asia	3812568	2559	27.2	0	253.1	18.2
North America	3176171	990	16.3	0	214.6	14.3
South America	248215	102	25.8	0	239.7	18.0
Africa	68707	102	12.2	0	192.2	10.7
Oceania	61340	54	9.10	0	155.0	8.3

"

L84: I am curious if you might also need some markers for anthropogenic activity (e.g., roadways)

and wildfire activity given that you are creating a daily estimate product the latter is potentially significant.

Response: We thank the reviewer for this suggestion. Anthropogenic activity indicators (including road density and related socio-economic proxies) are already incorporated among the predictor set used in the AiT framework. These variables are described in the Methods section. In the current framework, wildfire signals are partially reflected through satellite-derived atmospheric variables (including TROPOMI observations), as well as through land cover, NDVI dynamics, and meteorological covariates that respond to fire events. Therefore, wildfire effects are not entirely absent from the predictor space. We have clarified relevant outlook about adding wildfire characterization features in the revised manuscript.

Revision: Line 425: "Future work should focus on expanding data sources, such as emissions inventories, traffic data, and other dynamic activity indicators, to further improve the model's accuracy, especially in less urbanized regions. Incorporating additional episodic drivers (e.g., wildfire-related products in fire-prone regions) may provide incremental improvements where such processes substantially influence NO_2 variability. These refinements would enable more comprehensive global assessments of NO_2 pollution and contribute to the development of more targeted and effective air quality management strategies."

L90: Nice idea to correct the OMI data here. I don't think you have defined "n" in this equation, does this mean the total number of years in the overlap period (4)? If so maybe just put 4 there. While I think this is appropriate to correct for the coarser resolving pattern of OMI it is a potential source of uncertainty given that the 2019-2022 period is not necessarily representative of the period prior to 2019 (especially given the influence of COVID-19 lockdowns).

Response: To enable consistent long-term analysis and minimize discrepancies arising from differences in sensor characteristics, retrieval algorithms, horizontal resolution, and overpass time between OMI and TROPOMI, we applied a seasonal, grid-specific adjustment to OMI products using TROPOMI as a reference during their overlap period (2019–2022, n = 4 years). We acknowledge that the 2019–2022 period includes the COVID-19 lockdown years. However, the adjustment is designed to correct for **systematic sensor-related biases** rather than absolute concentration levels. Because the correction is calculated at the seasonal and grid-specific scale using multi-year averages, it primarily addresses structural inter-sensor differences while minimizing sensitivity to short-term emission anomalies.

Revision: "$\Delta\Omega_{weights}(i,m) = \frac{1}{4}\sum_{yr=2019}^{yr=2022}(\Omega_{TROPOMI}(i,yr,m) - \Omega_{OMI}(i,yr,m)$ "

L159-160: Nice to see both a random-based and spatial-based cross validation. I am curious if you additionally considered a temporal-based cross validation?

Response: Thanks for this suggestion. Our primary focus is on evaluating the spatial generalization ability of the model, as the dataset aims to provide reliable estimates in regions without monitoring coverage. Therefore, spatial cross-validation is particularly important. Temporal variability is already represented in the training and evaluation samples, and implementing an additional temporal cross-validation would substantially increase computational cost for the global model. Therefore, we did not implement an additional temporal cross-validation scheme.

L203-209: A few comments on the spatial cross-validation. First, it is interesting that you see only a marginal improvement for the monthly and annual predictions compared to the daily. I generally tend to think that averaging the data tends to significantly improve the correlation but in this case, it is a minor improvement. This points to the potential need to conduct a temporal cross-validation. Second, is it possible to perform a region-specific cross-validation? I am curious of the model performance in Africa (where observations are generally limited) compared to Europe and/or China.

Response: We thank the reviewer's comments. Regarding the relatively minor improvement from daily to monthly and annual aggregation under spatial cross-validation, this reflects that model performance is primarily constrained by spatial generalization rather than short-term temporal variability. While temporal aggregation reduces high-frequency noise, it does not substantially reduce spatial extrapolation errors at previously unseen locations. This limited improvement from daily to annual scales is also consistent with the intrinsic characteristics of NO_2 as a chemically reactive pollutant with a relatively short atmospheric lifetime. Its concentrations are strongly influenced by local emissions and rapidly varying meteorological conditions, leading to substantial temporal variability that is not fully suppressed by temporal aggregation. Similar behavior has been reported in previous studies (Wei et al., 2022; 2023), where aggregation to longer timescales yields only modest improvements in model performance for short-lived pollutants such as NO_2. Therefore, only modest gains in R^2 are observed when moving from daily to annual timescales under spatial CV. Therefore, the gains from temporal averaging remain limited under the spatial CV framework.

Regarding region-specific evaluation, we have now included continent-level performance metrics derived from the global cross-validation results (Tab. S2).

Revision:

"Table S2. Continental evaluation of the final NO_2 model under random and spatial cross-validation schemes (R^2, RMSE, and MAE).

	Random CV			Spatial CV		
Continents	R^2	RMSE	MAE	R^2	RMSE	MAE
Europe	0.92	4.12	2.82	0.54	7.30	5.16
Asia	0.89	6.08	4.20	0.59	8.41	5.63
North America	0.90	4.43	2.99	0.68	6.06	4.29
South America	0.89	6.10	4.14	0.58	8.92	6.11
Africa	0.78	4.90	2.92	0.04	10.29	6.39
Oceania	-	-	-	-	-	-

Note: Results for Australia are not reported due to the limited number of samples, which precludes statistically robust evaluation."

L211-215: Could you also compare to other popular global NO2 datasets such as the Larkin et al. 2023 dataset used in the GBD study (https://doi.org/10.3389/fenvs.2023.1125979) and Cooper et al. 2022 (https://doi.org/10.1038/s41586-021-04229-0). Ah I see you compared to Larkin later on, but this is, as I understand, not the most recent version. It might also be worth comparing in non-urban areas as the Larkin product has a known high bias in rural areas.

Response: We thank the reviewer's comments. We have already responded in Major Comments part.

Figure 1: Can you also include some metric of bias in the figures? It looks like there is a potential low bias in the estimates especially for the spatial cross-validation, but it is difficult to tell without a statistic.

Revision: Thanks for the comments. The new Figure 1 is below. The spatial cross-validation does have a low bias, with the NMB values of -1.5%, -1.4%, -1.4% for Spatial daily, monthly and yearly, respectively.

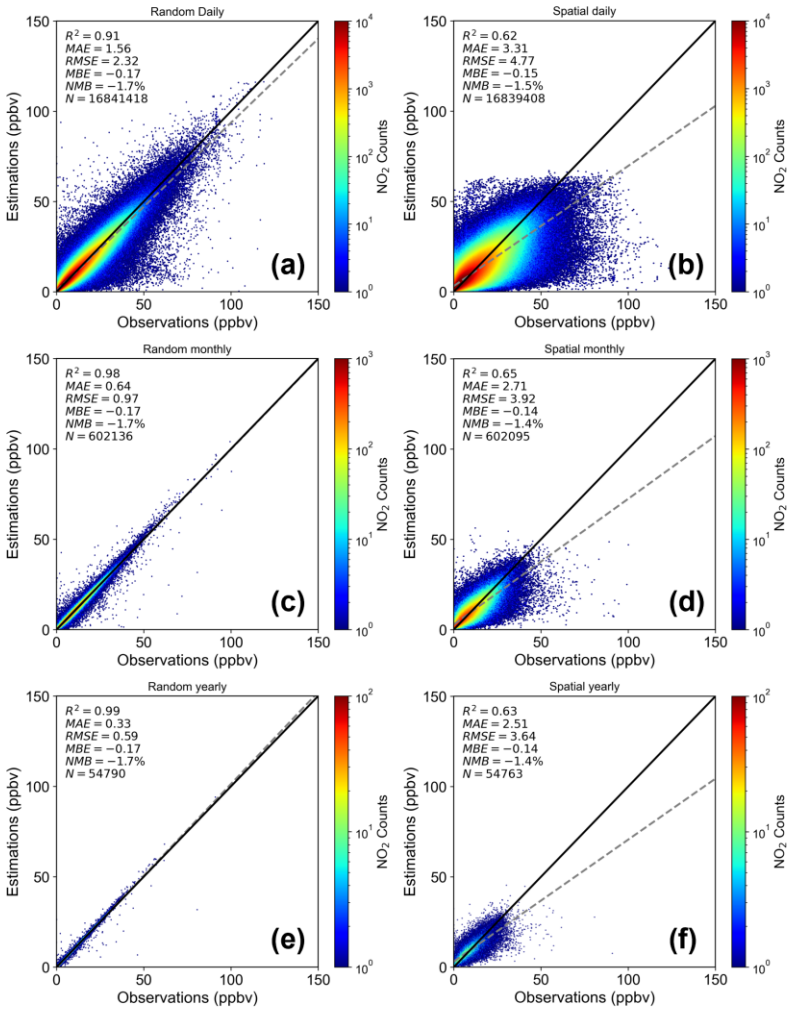

Figure 1: Model validation and uncertainties from the Random-based (left) and Spatial-based (right) cross-validation. Black lines are 1:1 lines, and grey dashed lines are best-fit lines from linear regression. Additional statistical metrics given are the correlation coefficient (R^2), mean absolute error (MAE), root-mean-square error (RMSE), mean bias error (MBE), normalized mean bias (NMB), and sample number.

L235: Can you be more specific here? I think you mean they use OMI only (not TROPOMI) in terms of the satellite data but unsure what you mean for the land-use variables.

Response: Thank you for the suggestion. We have clarified the description of the predictors used in the LUR model. Specifically, the LUR dataset relies on spatial predictor variables derived from land-use and environmental datasets, including population density, impervious surface area, road networks, vegetation indices (e.g., NDVI and tree cover), elevation, and OMI satellite-based NO_2 observations (Larkin et al. 2017).

Revision: Line 262: "In contrast, the LUR model relies on a set of spatial predictor variables derived from land-use and environmental datasets, including population density, impervious surface area, road networks, vegetation indices (e.g., NDVI and tree cover), elevation, and satellite-based NO_2 observations from OMI. As a result, the LUR framework is less capable of fully reflecting high NO_2

concentration areas, especially in urban settings."

L239: I don't think you can conclusively say that the spatial validation is "effective in generalizing to new locations" without testing accuracy in diverse regions. I am curious how well the model does in regions with limited monitors (e.g., Africa, Latin-America, Oceania) and also non-urban scenes. Otherwise, I don't think you can make this claim but rather need to caveat it with "generalizing to new urban locations in regions with strong monitoring infrastructure".

Revision: Thanks for pointing this out. Due to the limited availability of ground observations in regions such as Africa, and Oceania, the evaluation of model performance in these areas remains more uncertain. So to be accurate, we now have revised the text based on the reviewer's suggestion:

Line 269: "The Random validation highlights the model's accuracy across diverse data points, while the Spatial validation underscores its effectiveness in generalizing to new urban locations in regions with strong monitoring infrastructure."

L244: Again, this needs an important caveat that the LUR is likely conservative in urban areas but an underestimate in rural areas.

Revision: Thanks for your comment. We have revised relevant description.

Line 274: "While the LUR model captures the general trends in NO_2 variation, it may provide relatively conservative estimates in urban areas and may underestimate concentrations in rural environments in some countries."

Figure S3: These are annual values, right? Did you also compare at the daily or monthly timescale?

Response: Thank you for the comment. Yes, the values shown in Figure S3 are annual averages. The comparison is conducted at the annual scale because the LUR dataset provides only annual mean NO_2 estimates and does not include monthly or daily variability.

Figure S4: These are somewhat strange "regions" to separate the data into. Could you instead (or in addition) group these by continent? Or by GBD super region as you do in Figure 2?

Response: Thank you for the suggestion. The regions shown in Figure S4 (global, Tropics, Northern Hemisphere, and Southern Hemisphere) were selected to provide a broad-scale comparison between the LUR and AiT datasets. A more detailed regional comparison using GBD super regions is already presented in Figure S3.

Figure 2: These are very cool figures, thank you for sharing. I Think currently panels b and d are

difficult to read, I wonder if you can redesign the layout of the subplots to make these appear larger?

Revision: Thanks for your comment. We have redesigned the figure layout and increased the size of these panels to improve clarity.

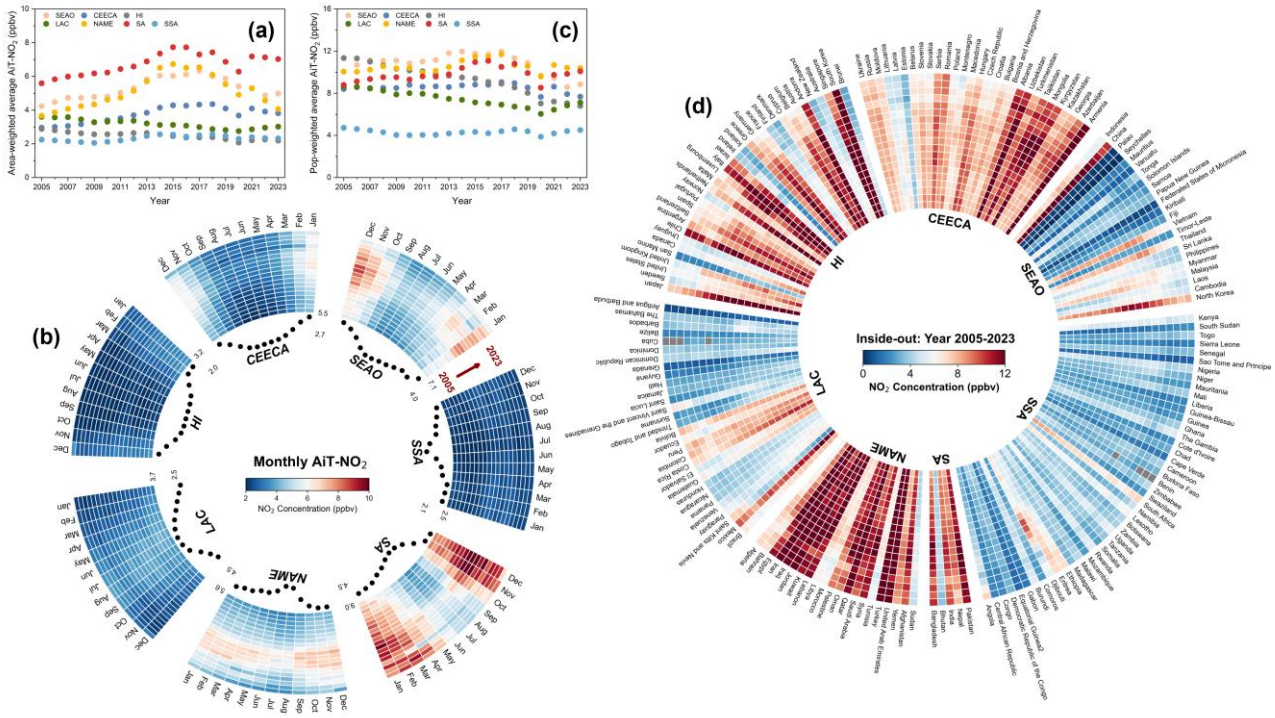

Figure 2: Temporal trends of annual NO_2 concentration from 2005 to 2023 in seven GBD super-regions for a) area-weighted average; c) population-weighted average. b) Heatmap of monthly area-weighted average, with scatter plots indicating monthly averages over the 19-year period. d) Heatmap of annual population-weighted average NO_2 concentrations for each country within the super-regions from 2005 to 2023. The seven super-regions defined in GBD are South-East Asia, East Asia & Oceania (SEAO), Central Europe, Eastern Europe & Central Asia (CEECA), High-income (HI), Latin America & Caribbean (LAC), North Africa & Middle East (NAME), South Asia (SA), and Sub-Saharan Africa (SSA).

L359: I take issue with the claim that the dataset can "assess localized impacts" without qualifying that these are only evaluated primarily in urban areas in the Global North. For example, if a small city in Uganda uses this dataset to characterize local NO2 do we anticipate this estimate can accurately characterize its "localized impacts". I think either more investigation is needed (as indicated above) or some of the claims need to be better qualified throughout the whole paper.

Response: Thank you for raising this important point. We agree that because monitoring networks are denser in urban regions as well as in the Global North, model performance is expected to be more robust in these areas than in regions with sparse observations. Nevertheless, the spatial cross-validation results ($R^2 \approx 0.6$) indicate that the model retains a reasonable ability to generalize to new locations, suggesting that the framework can capture large-scale spatial patterns of NO_2 variability. We have revised the text to clarify that the dataset primarily supports analyses of spatial NO_2 variability and exposure assessments in regions with relatively dense monitoring coverage, particularly urban areas.

In our manuscript, we also state that future incorporation of additional observations from currently under-monitored regions could further improve the prediction accuracy in these areas.

Revision: "The implications of this work extend far beyond conventional air quality monitoring. The dataset provides a valuable resource for analysing spatial patterns of anthropogenic NO_2 emissions, including those associated with industrial production and urban development, as well as for evaluating the effectiveness of pollution control measures. Its fine spatial resolution enables improved characterization of NO_2 variability, particularly in urban regions and in areas with relatively dense monitoring coverage."

Line 425: "Future work should focus on expanding data sources, such as emissions inventories, traffic data, and other dynamic activity indicators, to further improve the model's accuracy, especially in less urbanized regions."

Reference:

Zhang, Y., Shindell, D., Seltzer, K., Shen, L., Lamarque, J. F., Zhang, Q., Zheng, B., Xing, J., Jiang, Z. and Zhang, L.: Impacts of emission changes in China from 2010 to 2017 on domestic and intercontinental air quality and health effect, *Atmos. Chem. Phys.*, 21(20), 16051-16065, https://doi.org/10.5194/acp-21-16051-2021, 2021.

Wei, J., Li, Z., Chen, X., Li, C., Sun, Y., Wang, J., Lyapustin, A., Brasseur, G. P., Jiang, M., and Sun, L.: Separating daily 1 km PM2.5 inorganic chemical composition in China since 2000 via deep learning integrating ground, satellite, and model data, *Environ. Sci. Technol.*, 57, 18282–18295, https://doi.org/10.1021/acs.est.3c00272, 2023.

Wei, J., Liu, S., Li, Z., Liu, C., Qin, K., Liu, X., Pinker, R., Dickerson, R., Lin, J., Boersma, K., Sun, L., Li, R., Xue, W., Cui, Y., Zhang, C., and Wang, J.: Ground-level NO2 surveillance from space across China for high resolution using interpretable spatiotemporally weighted artificial intelligence, *Environ. Sci. Technol.*, 56, 9988–9998, https://doi.org/10.1021/acs.est.2c03834, 2022.

Tao, C., Peng, Y., Zhang, Q., Zhang, Y., Gong, B., Wang, Q., and Wang, W.: Diagnosing ozone–NO_x–VOC–aerosol sensitivity and uncovering causes of urban–nonurban discrepancies in Shandong, China, using transformer-based estimations, *Atmos. Chem. Phys.*, 24, 4177–4192, https://doi.org/10.5194/acp-24-4177-2024, 2024.

Larkin, A., Geddes, J. A., Martin, R. V., Xiao, Q., Liu, Y., Marshall, J. D., Brauer, M., and Hystad, P.:

Global land use regression model for nitrogen dioxide air pollution, *Environ. Sci. Technol.*, 51(12), 6957–6964, https://doi.org/10.1021/acs.est.7b01148, 2017.