Dear Editor and Reviewers,

We thank the Editor for handling our manuscript and the reviewers for their careful reading and constructive comments. We have revised the manuscript and prepared a point-by-point response to each comment. The comments have helped us clarify the scope of the dataset, explain methodological choices more fully, and state the remaining limitations more explicitly.

For clarity, the original comments are shown in **bold black type**. Our responses are shown in blue. Text revised or added to the manuscript is reproduced in black-bordered boxes, with the changed wording highlighted in red. References cited in this response are listed at the end of the document.

Reviewer 1:

1. The abstract should be clearer. It is not clear from the abstract which variables were used as model input and which were used as the training target; please state this explicitly. Please also make clear that the dataset covers daytime hours only, and define what the EAC4 chemical prior is.

Thank you for pointing out these ambiguities in the abstract. The previous version did not clearly distinguish the model inputs from the supervised target and could also have been read as implying continuous 24 h coverage.

We have revised the abstract (Revision 1) to explicitly state that the model inputs comprise the 0.05° TROPOMI tropospheric NO_2 VCD and its validity mask, together with the 0.25° predictors derived from ERA5 meteorological fields, EAC4 total-column NO_2, land cover, elevation, and the cosine of the solar zenith angle. We now clearly identify the hourly GEMS tropospheric NO_2 VCD from 2023 as the supervised training target. We have also clarified that the released dataset contains hourly fields during daytime hours only. Accordingly, we revised Sect. 3.1 (Revision 2).

Revision 1
High-frequency observations of tropospheric nitrogen dioxide (NO_2) support regional air quality assessment and analysis of its photochemical evolution. Although the Geostationary Environment Monitoring Spectrometer (GEMS) and other new-generation geostationary satellites provide hourly observations during daytime, their spatial coverage is often severely compromised by cloud obstruction and inherent retrieval limitations. Furthermore, their short observation history hinders long-term, cross-regional environmental assessments. To address this, this study presents a gapless, 0.05° daytime-hourly tropospheric NO_2 vertical column density (VCD) dataset for representative hotspots in Asia spanning 2019 to 2024, reconstructed using a physics-aware deep learning framework. The framework integrates partial convolutions to handle irregularly missing satellite observations and proposes a Physics-aware Normalization (PhysNorm) module. Specifically, the model inputs comprise the 0.05° TROPOMI tropospheric NO_2 VCD and its validity mask, together with 0.25° predictors including ERA5 meteorological variables; EAC4 total-column NO_2; land cover; elevation; and the cosine of the solar zenith angle. The supervised training target is the hourly GEMS tropospheric NO_2 VCD from 2023. PhysNorm dynamically modulates 0.05° high-resolution feature maps using 0.25° low-resolution physical features, thereby ensuring rigorous physical continuity while filling data gaps. Validation results show that the model performs exceptionally on the test set

(R^2 = 0.889) and maintains high generalization stability in an independent validation on 2024 data, which was not used for training. Cross-validation against an independent polar-orbiting satellite instrument (GOME-2C) confirms the reliability of the dataset in reconstructing pollution hotspots and its applicability to periods without GEMS observations (2019–2022). Using this reconstructed dataset, this study finely delineates the periodic diurnal characteristics of NO_2 in typical Asian regions and accurately captures the inter-annual concentration gradients driven by public health events and emission reduction policies over the past six years. This newly generated dataset provides a spatiotemporally continuous daytime record, overcoming the limitations of cloud cover and short observational histories, thereby facilitating long-term, high-resolution air quality assessments and epidemiological studies. This dataset is available at https://doi.org/10.5281/zenodo.20427767 (Gao et al., 2026).
Revision 2
This study integrates multi-source datasets with heterogeneous origins, spatiotemporal resolutions, and physical units, constructing an input tensor system comprising two 0.05° high-resolution input channels (the TROPOMI tropospheric NO_2 VCD and its validity mask), 33 low-resolution predictor channels at 0.25° resolution, and one supervised target channel consisting of the quality-controlled 0.05° hourly GEMS tropospheric NO_2 VCD. Table 2 systematically summarizes the source, original resolution, physical significance, and normalization strategy of all input variables. All reconstructed hourly fields correspond to daytime GEMS observation hours.

2. Section 3.1, line 248: how did you treat negative values in TROPOMI/GEMS/GOME-2? Negative retrievals are part of the random error distribution and cancel against positive noise during spatial or temporal averaging; discarding them gives an overall high bias. If negative retrievals were removed, please justify the filtering and quantify the resulting bias in both the training labels and the final product.

Retrievals with tropospheric NO_2 VCD values less than or equal to zero were excluded during the preprocessing of TROPOMI, GEMS, and GOME-2 data, and this treatment is now explicitly documented in Sects. 2.2–2.4 (Revision 3–5) of the revised manuscript.

These retrievals were treated as missing values rather than replaced by zero. In particular, invalid GEMS pixels were excluded from the supervised loss and therefore did not serve as zero-valued training labels. The exclusion of negative satellite NO_2 retrievals has also been reported in related studies (Hoinaski et al., 2024; Gödeke et al., 2025). This processing choice was made because the model was formulated to reconstruct a nonnegative tropospheric NO_2 column field and because the logarithmic normalization used in the framework requires nonnegative input and target values.

We agree with the reviewer that negative retrievals form part of the random retrieval-error distribution and may compensate for positive noise during spatial or temporal averaging. Consequently, excluding them may introduce an upward conditional-sampling bias, particularly under low-column or background conditions. We have therefore added this issue to Sect. 5.2 (Revision 6). We have also clarified that the reported model evaluations and cross-satellite comparisons characterize the positive-retrieval subset and that the reconstructed product should not be interpreted as an unbiased estimate near the satellite retrieval detection limit. We have added that future work will investigate sign-preserving preprocessing, alternative transformation strategies, and corresponding sensitivity analyses to quantify the influence of negative-retrieval treatment on both the training labels and the reconstructed product.

Revision 3
This study utilizes the GEMS Level 2 NO_2 tropospheric VCD v3.0 product. To ensure data quality, quality control was applied, retaining only pixels where the VCD is greater than zero, the cloud fraction is less than 0.3, and the solar zenith angle is less than 70°. Thus, negative and zero GEMS retrievals were treated as missing values rather than being replaced with zero. This criterion was adopted because the supervised target was defined as a nonnegative tropospheric NO_2 column field and because the subsequent logarithmic transformation requires nonnegative input values. Pixels removed by this criterion were excluded from the training loss and did not act as zero-valued training labels.
Revision 4
For TROPOMI data, a quality assurance flag (qa_value) greater than 0.75 was used as the filtering criterion. The filtered data were subsequently regridded to the same 0.05° × 0.05° grid using the HARP toolkit (https://github.com/stcorp/harp). In addition to the qa_value threshold, TROPOMI retrievals with tropospheric NO_2 VCD values less than or equal to zero were treated as missing. This filtering maintained consistency with the nonnegative model-input domain and the logarithmic normalization used in the framework.
Revision 5
The data were regridded to a 0.25° × 0.25° resolution using the HARP toolkit, and pixels with a cloud fraction greater than 0.3 were filtered out. For consistency with the GEMS and TROPOMI processing, GOME-2 retrievals with tropospheric NO_2 VCD values less than or equal to zero were also excluded.
Revision 6
The training domains were selected to emphasize representative high-emission and complex-terrain regions, which increases the representation of polluted and high-gradient conditions but limits model generalizability to very clean backgrounds, such as remote oceanic regions. Uncertainty under low-column conditions also arises from treating satellite NO_2 VCD retrievals less than or equal to zero as missing values. Although this treatment is consistent with the nonnegative model domain and the logarithmic transformation, negative retrievals form part of the random retrieval-error distribution, and their exclusion may introduce an upward conditional-sampling bias near the satellite retrieval detection limit. The reported evaluations therefore characterize the positive-retrieval subset, and the product should not be interpreted as an unbiased estimate under ultra-low background conditions. Future work will examine adaptive sampling, loss weighting, sign-preserving preprocessing, and alternative transformation strategies.

3. Regarding the gap filling: all your evaluation, including that for 2024, is performed over pixels where observations exist, so the gap-filled values themselves are never evaluated. This limitation should be stated explicitly in the text, as the reported metrics do not support claims about the quality of the dataset as a whole. In addition, EAC4 is the only NO2 input with complete coverage, so the gap-filled values presumably depend largely on it. Please quantify how strong this dependence is, discuss what happens if EAC4 is biased, and explain why a user should prefer your product over EAC4 itself in these regions.

The reviewer is correct that direct evaluation against GEMS necessarily requires a valid GEMS retrieval. Therefore, reconstructed pixels at locations or times where GEMS itself is unavailable cannot be directly verified against GEMS. However, this does not mean that predictions under gap conditions were not evaluated. The 2023 test dates were withheld from model training and were randomly partitioned by date. During evaluation, metrics were calculated at all valid GEMS pixels without requiring the collocated TROPOMI observations to be valid, and the original TROPOMI availability mask was retained during inference. Consequently, the evaluation includes naturally occurring TROPOMI gaps, where the collocated high-resolution satellite observation is absent and the model reconstructs NO_2 using the surrounding available information and the other predictors. For the held-out 2023 test set, a separate evaluation over GEMS-valid but TROPOMI-missing pixels yielded an R^2 of 0.801 and an RMSE of 1.235 Pmolec/cm^2, providing direct evidence of model performance under this input-gap condition. We have also stated in Sect. 4.1 that pixels lacking a valid GEMS reference cannot be directly evaluated and that the reported metrics should therefore not be interpreted as pixelwise validation of the entire gapless product (Revision 7).

To quantify the model's dependence on EAC4, we conducted an inference-only sensitivity experiment using all 1,450 scenes in the held-out 2023 test set, rather than selecting representative months or scenes. The trained model and all other inputs were kept unchanged, while the de-standardized EAC4 input was independently perturbed by +20% and −20%. At the GEMS-valid pixels where TROPOMI was missing, the baseline reconstruction achieved R^2=0.801 and an RMSE of 1.235 Pmolec/cm^2. The +20% and −20% EAC4 perturbations changed the pooled mean reconstruction by +4.02% and −4.60%, respectively, yielding a central bias elasticity of 0.216 (defined as $E = \frac{\bar{\hat{y}}_{+20\%} - \bar{\hat{y}}_{-20\%}}{0.4\bar{\hat{y}}_{\text{baseline}}}$, where $\bar{\hat{y}}_{\text{baseline}}$ denotes the mean reconstructed NO_2 VCD). The corresponding regional elasticities were 0.134 for CC, 0.269 for KJ, 0.220 for NI, and 0.227 for NP. These results show that EAC4 is an important large-scale chemical constraint, but its variations are substantially attenuated rather than proportionally copied into the reconstructed product. We have revised Sect. 5.1 accordingly (Revision 8).

Thus, the reconstructed product is not interchangeable with EAC4 itself. EAC4 supplies a spatially complete but coarse-resolution total-column NO_2 prior, whereas our product provides 0.05° daytime-hourly tropospheric NO_2 estimates and combines EAC4 with TROPOMI, ERA5 meteorology, solar geometry, land cover, elevation, and learned high-resolution spatial information. The product is therefore intended for applications requiring fine-scale, target-aligned tropospheric NO_2 information, while EAC4 remains the underlying large-scale chemical prior.

Revision 7
To characterize model performance on held-out dates within the 2023 training year, Figure 2 presents the overall scatter density plots for the training, validation, and test sets. The predicted dataset achieved high prediction accuracy across all three sets, with R^2 values of 0.939, 0.882, and 0.889, respectively (all p-values < 0.01). Furthermore, during the training and testing process involving tens of millions of samples, the performance metrics on the validation and test sets were highly consistent, with no significant signs of overfitting. This indicates that the designed PhysNorm-Net model can effectively learn the non-linear mapping relationship between multi-source physical drivers and pollutant concentrations. Although this evaluation directly examines model performance under naturally occurring TROPOMI gaps wherever valid GEMS references are available, reconstructed pixels for which GEMS itself is unavailable cannot be directly verified against GEMS. Therefore, the reported metrics should not be interpreted as a pixelwise validation of the entire gapless product.
Revision 8

During the above feature-importance analysis, EAC4 was identified as an important chemical prior. We therefore further quantified the dependence of the reconstructed values on EAC4 using all 1,450 scenes in the held-out 2023 test set. With the model weights and all other inputs fixed, the de-standardized EAC4 field was perturbed by ±20%, and the results were pooled over GEMS-valid pixels for which the TROPOMI input was missing. The baseline reconstruction for this subset achieved R^2=0.801 and an RMSE of 1.235 Pmolec/cm^2. Increasing and decreasing EAC4 by 20% changed the pooled mean reconstruction by +4.02% and −4.60%, respectively, corresponding to a central bias elasticity of 0.216 (defined as $E = \frac{\bar{\hat{y}}_{+20\%} - \bar{\hat{y}}_{-20\%}}{0.4\bar{\hat{y}}_{\mathrm{baseline}}}$, where $\bar{\hat{y}}_{\mathrm{baseline}}$ denotes the mean reconstructed NO_2 VCD). The region-specific elasticities were 0.134, 0.269, 0.220, and 0.227 for CC, KJ, NI, and NP, respectively. This attenuated and regionally variable response indicates that EAC4 provides an important large-scale chemical constraint but is not transferred proportionally to the model output.

4. Lines 483-485: you attribute the drop in model performance in 2024 to the model being unable to capture the inter-annual shifts in emissions. If so, how can we trust your predictions prior to 2023, when real emission changes were occurring?

Thank you for pointing out the ambiguity in our original explanation of the 2024 results. The previous wording could have suggested that the model was unable to capture inter-annual emission changes. We have revised this passage to avoid this interpretation (Revision 9). The modest decrease is now discussed as a reasonable outcome of a more demanding independent-year validation, in which the model encounters broader inter-annual variability in anthropogenic activities, emission-control policies, meteorology, extreme events, and satellite retrieval availability than is represented by the within-year test set. Importantly, the model still retained high predictive performance in 2024, including R^2 values above 0.81 in NP and KJ, demonstrating substantial temporal robustness and transferability to an unseen year.

The reliability of the predictions prior to 2023 is supported by two additional considerations. First, the historical reconstructions are not generated by applying a fixed 2023 climatological pattern. For each historical year, the model uses the corresponding time-varying TROPOMI observations, EAC4 chemical prior, and ERA5 meteorological fields. Therefore, year-specific variations contained in these inputs can be propagated into the reconstructed NO_2 fields.

Second, and more importantly, the pre-2023 predictions were independently evaluated against GOME-2C observations in Sect. 4.2 and Fig. 6. For the four regions combined, the annual correlations between the reconstructed dataset and GOME-2C remained within 0.76–0.79 throughout 2019–2024, while the corresponding RMSE and MAE remained within 3.00–3.25 and 1.80–1.90 Pmolec/cm^2, respectively. The performance during the retrospective period of 2019–2022 was comparable to that during 2023–2024, with no systematic degradation toward the earlier years. Thus, confidence in the pre-2023 reconstruction is based not only on the model's successful forward transfer to 2024 but also on direct independent comparisons with GOME-2C during the retrospective years.

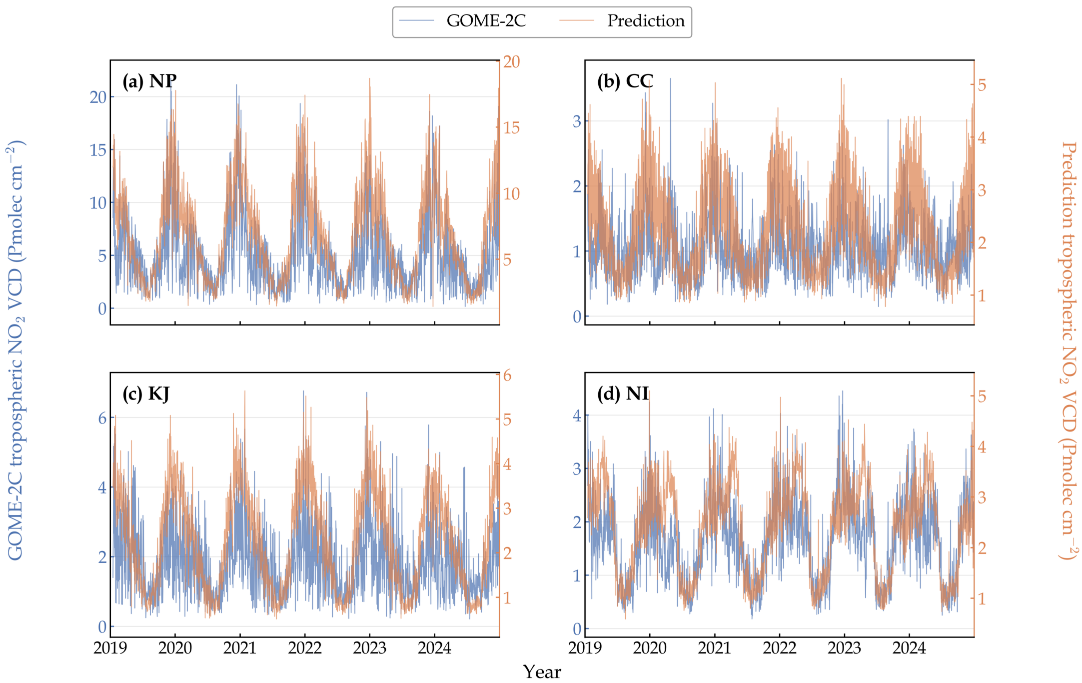
We acknowledge that the GOME-2C evaluation is limited to its overpass time and spatial sampling and therefore does not constitute a complete validation of every daytime-hourly pixel. Nevertheless, it provides independent evidence that the reconstructed dataset captures the major spatial and inter-annual variations in NO_2 during 2019–2022.

Revision 9

The 2024 evaluation produced lower metrics than the internal 2023 test set but retained predictive skill in all four regions, with R^2 values above 0.81 in NP and KJ. These results indicate that the learned relationships remain partly transferable to an unseen year, although performance varied by region and decreased most in NI. The comparatively larger decrease in NI may be partly associated with the smaller number of valid GEMS labels available for this region, which provides less information for model training. Overall, the stable performance in 2024 demonstrates the temporal robustness and generalization capability of the proposed framework, supporting its extension to longer-term historical NO_2 reconstruction.

5. Section 4.2: since GOME-2 provides long-term observations, please add an interannual evaluation against it. Does GOME-2 show trends and interannual changes similar to those in your product (at 09:30 LT from 2019 to 2024, Section 4.3)?

We performed an additional daily comparison between GOME-2C and the reconstructed Prediction from 2019 to 2024 at the nominal GOME-2 overpass time of 09:30 LT and provide the result as Response Fig. R1. The response-only figure indicates broad correspondence in their temporal variability, including the recurring seasonal cycle, although the daily regional means exhibit substantial short-term fluctuations and temporal gaps.



Response Fig. R1. Daily regional mean tropospheric NO_2 VCDs from GOME-2C and the reconstructed Prediction over the four study regions from 2019 to 2024 at the nominal GOME-2 overpass time of 09:30 LT. Daily means were calculated using co-located valid pixels, and no temporal smoothing was applied. GOME-2C and Prediction correspond to the left and right y axes, respectively.

The co-located evaluation already presented in Section 4.2 quantifies the overall consistency between GOME-2C and the Prediction using pixels with matched spatial and temporal sampling. This evaluation fulfills the principal role of GOME-2C in our study as an independent long-term observational reference for assessing whether the reconstructed fields remain consistent with another satellite product during periods and at locations where valid observations are available.

GOME-2C has its own spatial resolution, retrieval characteristics, sampling frequency, and cloud-related data gaps. It is therefore most appropriately used as an independent reference under co-located sampling conditions rather than as an uncertainty-free representation of the complete tropospheric NO_2 field. The correlation statistics reported in Section 4.2 offer complementary evidence for the reliability of the reconstructed product.

The daily time series and the collocated evaluation address complementary aspects of cross-satellite consistency. The inter-annual and diurnal statistics in Sect. 4.3 are calculated from the gapless reconstructed dataset on a fixed spatial grid with continuous daytime-hourly coverage, whereas the GOME-2C regional means are derived from a varying number and spatial distribution of cloud-free, quality-screened pixels at a single morning overpass. Consequently, the unsmoothed GOME-2C daily means combine atmospheric variability with day-to-day changes in spatial sampling. The collocated analysis in Sect. 4.2 provides the quantitative assessment under matched spatial and temporal sampling, while Response Fig. R1 provides a complementary qualitative view of longer-term temporal variability.

6. Section 4.3: you attribute all the interannual changes in the dataset solely to emission changes. Yet you also attribute the drop in model performance in 2024 to the model not capturing inter-annual shifts in emissions (Comment #4). These two statements are in conflict. Regarding these emission changes, do you have supporting evidence from other sources, such as government reports, bottom-up inventories, or surface NO2 observations, to support your statement. For example, you state that NP and KJ underwent abrupt and widespread emission reductions in 2024 compared to 2023 – is there independent evidence for this? Could meteorology (e.g., transport) also play a role?

As explained in our response to Comment 4, we have revised Sect. 4.1 so that the modest decrease in model performance in 2024 is no longer attributed to an inability of the model to capture inter-annual emission shifts. Also, we have revised Sect. 4.3 to provide a more balanced interpretation of the inter-annual variations in the reconstructed dataset, including the discussions of the representative cities and all four study regions (Revision 10).

We have revised Section 4.3 in three respects. First, we now explicitly state that changes in tropospheric NO_2 VCD are not direct estimates of changes in NO_x emissions. The reconstructed columns are jointly influenced by emissions, atmospheric chemistry, boundary-layer evolution, and regional transport, whose relative contributions vary with meteorological conditions and time of day. Statements attributing the changes solely to emission reductions have been removed.

Second, we added published evidence to place the overall 2019–2024 patterns in context. Huber et al. (2026) analyzed TROPOMI NO_2 VCDs over the same period and reported substantial decreases in North and East China, pronounced decreases in South Korea, and heterogeneous changes across India. Their country-level urban population-weighted trends were -6.0 ± 1.0 % yr^{-1} for China, -8.74 ± 0.9 % yr^{-1} for South Korea, and -4.1 ± 0.6 % yr^{-1} for Japan. For India, Delhi showed a negative absolute change, although the 2019–2024 trend was not statistically significant, while increases occurred near several coal-mining and power-generation areas.

We additionally examined official annual mean surface NO_2 records for 2019–2024. The reported values were 37, 29, 26, 23, 26, and 24 μg/m^3 for Beijing and 42, 37, 35, 30, 28, and 24 μg/m^3 for Chengdu, respectively. In South Korea, the corresponding values were 0.028, 0.024, 0.024, 0.021, 0.020, and 0.019 ppm for Seoul; 0.024, 0.020, 0.021, 0.019, 0.019, and 0.018 ppm for Incheon;

and 0.024, 0.020, 0.019, 0.018, 0.017, and 0.015 ppm for Gyeonggi. These independent surface records are directionally consistent with the longer-term decreases reconstructed for NP, CC, and KJ. However, surface concentrations and tropospheric columns represent different atmospheric quantities, and their year-to-year variations are not identical. We therefore use these records only as independent directional evidence rather than as quantitative validation of the reconstructed VCD changes. In particular, they do not by themselves establish an abrupt and spatially widespread reduction in emissions from 2023 to 2024, and we have removed that claim from the manuscript.

Third, we refocused the discussion of Figures 9 and 10 on the principal added value of the dataset. Polar-orbiting satellites observe NO_2 at a single local time of day, whereas our hourly daytime product enables the 2019–2024 changes to be separated into Morning, Midday, and Afternoon periods. The revised analysis shows that the decreases in NP and KJ occurred throughout the daytime, that the six-year decrease in CC was strongest in the afternoon, and that NI exhibited a 2022 enhancement and spatially compensating changes that differed among daytime periods. These time-of-day-resolved results demonstrate that an All-day mean or a single satellite overpass may not fully characterize the inter-annual evolution of daytime NO_2. We have nevertheless retained a cautious interpretation because emissions, chemistry, boundary-layer development, and transport can all contribute to these temporal differences.

Revision 10
Beyond the regional macroscopic patterns, Figure 8 focuses on four representative cities, including Beijing, Chengdu, Seoul, and Delhi, to reveal fine-grained diurnal structures and their inter-annual evolution under the combined influence of anthropogenic activity, atmospheric chemistry, and meteorological conditions. The diurnal amplitude in Beijing is higher in winter, whereas strong photochemical loss and deeper boundary-layer mixing suppress the overall NO_2 levels and weaken the amplitude in summer. From 2019 to 2024, the seasonal diurnal curves in Beijing show a generally parallel downward shift, indicating a sustained decrease in the reconstructed NO_2 column burden across daytime hours. Chengdu exhibits a relatively flat diurnal variation, which may be associated with the enclosed topography of the Sichuan Basin, restricted horizontal ventilation, and limited boundary-layer development. The comparatively smaller decrease in Chengdu and the temporary rebound in 2021 may reflect the combined effects of local anthropogenic activity, regional transport, and inter-annual variations in basin meteorology. Seoul exhibits an afternoon peak in winter and spring, indicating that its NO_2 burden may be influenced by regional transport from surrounding industrial areas in addition to local traffic emissions. Seoul shows the largest six-year decrease among the four cities, consistent with the declining trend recorded by the official surface NO_2 monitoring network in the Seoul metropolitan area. Delhi is strongly modulated by the South Asian monsoon system. Its diurnal amplitude is largest during the pre-monsoon spring, with high morning values followed by a sharp afternoon decrease. During the summer monsoon, daytime concentrations decline to their annual minimum under the combined effects of convective mixing, wet removal, atmospheric transport, and seasonal changes in anthropogenic and natural sources. It should be emphasized that the inter-annual differences in the reconstructed tropospheric NO_2 VCDs should not be interpreted as direct estimates of changes in NO_x emissions. Tropospheric NO_2 columns are jointly affected by emissions, chemical lifetime, boundary-layer mixing, horizontal and vertical transport, and meteorological factors such as wind, temperature, humidity, and solar radiation (Seo et al., 2024; Yang et al., 2024). Accordingly, all the analysis describes inter-annual increases and decreases in the reconstructed NO_2 column burden rather than inferring direct changes in NO_x

emissions. Socioeconomic activities and emission-control policies are discussed only as possible contextual factors and are not treated as quantitatively isolated causes. Therefore, the reconstructed dataset alone cannot isolate the effects of emission-control policies from those of regional transport and meteorological variability.

Figures 9 and 10 use the hourly structure of the reconstructed dataset to compare inter-annual changes during All-day (08:00–17:00 LT), Morning (08:00–10:00 LT), Midday (11:00–13:00 LT), and Afternoon (14:00–17:00 LT) periods. This time-of-day-resolved analysis provides information that cannot be obtained from polar-orbiting satellite observations made at a single local overpass time. The broad 2019–2024 patterns are consistent with a recent TROPOMI-based analysis reporting substantial NO_2 VCD decreases in North and East China and South Korea, together with more heterogeneous changes across India (Huber et al., 2026).

The NP region exhibited an overall decrease in reconstructed NO_2 VCD from 2019 to 2024 across the morning, midday, and afternoon periods. The 2019–2020 decrease was slightly larger in the morning, whereas the pronounced decrease from 2023 to 2024 occurred during all three periods. Figure 10 further shows that the strongest six-year decreases extended from the Beijing–Tianjin–Hebei urban corridor towards Shandong, with a broadly consistent spatial footprint throughout the daytime.

In the CC region, the decrease in 2020 and the subsequent rebound in 2021 occurred during all three daytime periods. Spatially, the principal urban centers, including Chengdu, Chongqing, and Xi’an, showed localized decreases, whereas several surrounding industrial and urban areas exhibited increases. In contrast to the relatively consistent daytime decrease in NP, the six-year decrease in CC was most pronounced during the afternoon.

The KJ region showed an overall decrease from 2019 to 2024 during the morning, midday, and afternoon periods, together with a modest rebound in 2022 and a pronounced decrease in 2024. Figure 10 indicates that the six-year decreases were concentrated over South Korea during all three periods, although localized inter-annual variations remained near Incheon, coastal areas, and the surrounding waters.

The NI region exhibited the most spatially heterogeneous pattern among the four study regions. Regional mean reconstructed NO_2 VCD increased markedly in 2022 during all three daytime periods and subsequently returned to approximately its 2019 level by 2024. Figure 10 reveals a complex mixture of increases and decreases along the Indo-Gangetic Plain, with clear differences in their spatial distribution among the morning, midday, and afternoon periods.

7. The groups in Fig. 13 contain very different numbers of channels: 12 for static geography, 9 for transport, 8 for thermodynamics, but only one each for EAC4, SZA and TROPOMI. Since the zero-out perturbation removes a whole group at once, larger groups will produce larger deviations, so the finding that the three largest groups rank in the top three in every region may be an artifact of aggregation. This also explains the apparent conflict with Fig. 12, where EAC4 ranks in the top three for winter NP and for most months in KJ. Please consider normalizing by channel count or reporting per-channel importance, and revise Section 5.1 accordingly.

Thank you for drawing attention to the effect of unequal group sizes on the aggregated importance scores. We clarify that Fig. 13 was not generated by simultaneously zeroing all channels within a physical group. Instead, each input channel was perturbed individually, and the resulting channel-level importance scores were subsequently summed within each group. Nevertheless, we agree with the reviewer's central concern: summing importance scores across groups containing different numbers of channels can systematically favor larger groups and can obscure the importance of individual variables, such as EAC4. This aggregation also contributed to the apparent inconsistency between Figs. 12 and 13.

We have therefore removed Fig. 13 and all associated group-level rankings, percentages, and interpretations from Sect. 5.1. The revised discussion (Revision 11) is now based exclusively on individual-channel results, including the prominence of EAC4 during winter in NP and during many months in KJ. We have also added an explicit caution that zero-out importance measures model dependence rather than causal physical contribution, and that importance scores from correlated channels should not be summed directly.

Revision 11

Building on this diagnostic foundation, to further unveil the model's decision-making mechanism, we quantitatively analyzed the relative contributions of each input variable during daytime hours in 2023 using a zero-out perturbation method (Zhang and Gao, 2020). This method calculates a variable's importance score by zeroing out its specific channel and measuring the resulting deviation from the baseline output, quantified by MAE. Figure 12 reports the three highest-ranked individual input channels for each bin. The dominant variable combinations differ significantly across the four regions, indicating that the model does not simply rely on a fixed set of factors but instead forms region-specific control structures.

Because several input variables are physically and statistically correlated, such as specific humidity and temperature across multiple pressure levels, the individual-channel scores are not additive and should not be interpreted as unique causal contributions. When correlated variables contain overlapping information, the model may compensate for the perturbation through the remaining channels, or the perturbation may disrupt information shared across several channels. Therefore, the resulting rankings represent conditional model sensitivity under the applied perturbation, rather than independent explanatory importance; and the rankings in Fig. 12 are interpreted as diagnostic evidence of relative model sensitivity under the sampled 2023 conditions, and are considered together with the physical context of each region and season.

As shown in Figs. 12, in the NP region, the most stable and prominent variables during winter and late autumn are EAC4 and surface pressure, with the third position often supplemented by near-surface or mid-tropospheric thermodynamic variables. This channel-level pattern has a clear physical meaning: in East Asian winters, frequent temperature inversions, stable weather conditions, and high relative humidity collectively suppress vertical mixing and promote the accumulation of NO_2 in the near-surface layer (Chai et al., 2014; Xu et al., 2011). Concurrently, the increased importance of EAC4 during certain winter periods indicates greater model sensitivity to the large-scale chemical background represented by EAC4. In summer (JJA), SZA becomes more prominent and shows a strong daytime peak around noon. This aligns with the physical processes of enhanced convective mixing and accelerated photochemical decomposition in the warm season (Liu et al., 2022).

In the CC region, surface pressure ranks first during most daytime hours throughout the year, and the corresponding group consistently occupies the top three throughout the year. This result is highly consistent with the fundamental understanding of air pollution distribution in a basin: within an enclosed topography, surface pressure not only reflects the state of synoptic-scale circulation but also indirectly indicates ventilation capacity and pollution trapping conditions (Sun et al., 2020). DEM and individual land-use channels also frequently appear among the top three, indicating that terrain differences and land-cover types provide persistent spatial constraints on the modeled NO_2 distribution. During summer, near-surface thermodynamic variables become more prominent. The importance of SZA also increases in spring and summer and reaches the highest rank in some region–month–hour bins. This is consistent with the meteorological background of the Sichuan Basin in summer, which is characterized by the development of thermal lows, valley winds, and enhanced convective mixing (Feng et al., 2022).

In the KJ region, the water-body land-cover channel is the top-ranked variable during almost all months and daytime hours. This implies that in a peninsula-archipelago system with complex coastlines and strong land-sea segmentation, the water body variable effectively encodes the regional land-sea distribution, which in turn determines the spatial extent of emission sources and sea breeze influence. EAC4 also appears more frequently among the top-ranked channels during the cold season and near the beginning and end of the daytime observation window, indicating stronger model dependence on the large-scale chemical background under these conditions. Overall, the channel-level results indicate the combined influence of land–sea distribution, the EAC4 chemical prior, and seasonally varying radiation and thermodynamic conditions, without assigning an aggregate importance rank to any physical group.

The NI region exhibits a strong seasonal switch. During winter and late autumn, surface pressure is almost consistently ranked first, indicating that NO_2 distribution in this region is primarily controlled by boundary layer stability and ventilation conditions during the cold and transitional seasons. Concurrently, the land use types grasslands/cereal crops and broadleaf crops become prominent in winter. These agricultural land types are widespread in the Indo-Gangetic Plain, and their high importance reflects the impact of agricultural burning—the widespread burning of crop residues in this region is a major seasonal source of NO_x (Cusworth et al., 2018; Jethva et al., 2018). Around the monsoon season, the dominant variables shift to specific humidity, temperature, and SZA, with these thermodynamic and radiation-related channels appearing more frequently among the top-ranked variables, whereas the prominence of individual land-use channels decreases. This reflects the profound influence of monsoon dynamics on vertical mixing, wet scavenging, and the large-scale redistribution of pollutants (Ravindra et al., 2003).

The rankings in Fig. 12 quantify the model's dependence on individual input channels rather than the total contribution of a physical-process group. Furthermore, the zero-out perturbation method measures model dependency rather than strict physical causality. Therefore, the importance of static variables such as land use and DEM should be interpreted as the model's reliance on spatial encoding and regional templates.

8. Last paragraph of the Introduction: please clarify the spatial and temporal coverage of the dataset.

We have revised the final paragraph of the Introduction to define the spatial and temporal coverage of the generated dataset more explicitly (Revision 12). The revised text now states that the dataset has a spatial resolution of 0.05°, covers the period from 1 January 2019 to 31 December 2024, and provides hourly records during daytime hours only. It covers four 10° × 10° Asian hotspot domains, namely the North China Plain, the Sichuan Basin and its surroundings, Korea and Japan, and northern India, with the geographical boundaries of each domain now specified. We have also replaced hourly with daytime-hourly in the contribution statement to avoid implying 24-hour coverage.

Revision 12

To address this, this study proposes PhysNorm-Net, a physics-aware network to reconstruct a gapless 0.05° daytime-hourly tropospheric NO_2 dataset covering the period from 1 January 2019 to 31 December 2024 over four 10° × 10° Asian hotspot domains: the North China Plain (NP; 34–44° N, 112–122° E), the Sichuan Basin and its surroundings (CC; 26–36° N, 100–110° E), Korea and Japan (KJ; 32–42° N, 126–136° E), and northern India (NI; 20–30° N, 75–85° E).

9. Line 158: what is the spatial resolution of the GOME-2 NO_2

We have revised Section 2.4 to distinguish the native spatial footprint of the GOME-2C Level 2 observations from the regridded resolution used in this study (Revision 13). The nominal native ground-pixel size of GOME-2C is 80 km × 40 km for forward-scan observations and 240 km × 40 km for back-scan observations. After quality control, the Level 2 observations were regridded to a 0.25° × 0.25° latitude–longitude grid using the HARP toolkit. This distinction has now been stated explicitly in the revised manuscript.

Revision 13

This study utilizes the GOME-2C Level 2 daily tropospheric NO_2 product provided by EUMETSAT AC-SAF (Hassinen et al., 2016). The nominal native ground-pixel size of GOME-2C is 80 km × 40 km for forward-scan observations and 240 km × 40 km for back-scan observations (Chan et al., 2023).

10. Figure 11: which day is shown? It would be better if it were from 2024 instead of the training sample.

The original figure presents the KJ region at 05:00 UTC on 2 January 2023. We have replaced the original training-period example with an example from 2024, which was not used for model training. Specifically, the revised Fig. 11 presents the KJ region at 04:00 UTC on 11 November 2024. The corresponding date and time have also been explicitly stated in the figure caption.

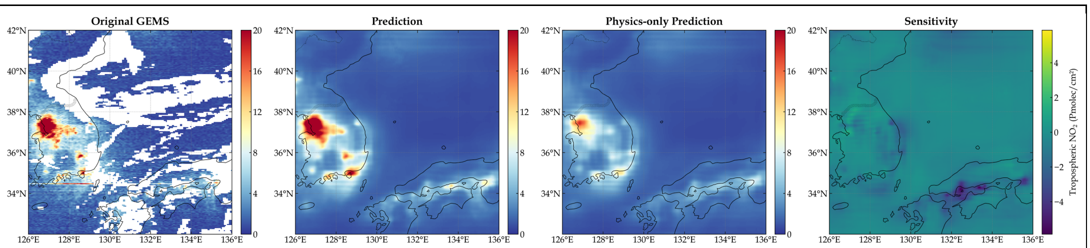


Figure 11: Functional diagnostic experiment of the PhysNorm-Net dual-branch architecture. The example shown corresponds to the KJ region at 04:00 UTC on 11 November 2024.

11. Figure 11: the prediction shows an anomalous feature along the border between South Korea and North Korea. Why?

We understand that the anomalous feature refers to the apparently sharp north–south gradient near the border between South Korea and North Korea in the Prediction panel.

We re-examined the plotting procedure, model inputs, and contemporaneous observations. The national boundary was added only during visualization using the Cartopy border layer and was not supplied to the model as an input. The model therefore cannot directly impose a discontinuity based on country identity or an administrative boundary.

The selected scene corresponds to 2 January 2023 at 05:00 UTC. The available observations already show a strong contrast between the Seoul metropolitan area and the region to its north. Within 126–130°E, the mean GEMS NO_2 VCDs over the 37–38°N and 38–39°N latitude bands were approximately 10.48 and 3.16 Pmolec/cm^2, respectively. The EAC4 prior and urban land-cover predictor also contain a broad south–north contrast associated with the Seoul metropolitan area. This explains why the feature remains visible in the Physics-only Prediction.

We therefore interpret the feature primarily as the reconstructed representation of the strong observed NO_2 gradient between the Seoul metropolitan area and the lower-column region to its north, rather than as a discontinuity imposed by the political boundary. Its visual sharpness is enhanced by the superimposed border line and the fixed color range used in the figure.

Reviewer 2:

1. It is unclear how the predictions are validated over the regions where GEMS observations are absent (the "gaps"), which is precisely where the reconstruction adds value. I suggest a dedicated holdout experiment: artificially mask out blocks of valid GEMS pixels at inference, and examine whether the model's filled values agree with the withheld observations, as a function of gap size. This would directly quantify the accuracy of the gap-filled portion of the dataset.

We acknowledge that the values reconstructed at locations where GEMS is unavailable cannot be directly validated against GEMS, because no co-located GEMS reference exists at those locations. As clarified in our response to the related comment (Revision 7), the reported evaluation metrics should therefore be interpreted as performance statistics conditional on the availability of a valid GEMS reference, rather than as pixelwise verification of every value in the gapless product.

GEMS is used only as the supervised target and validation reference; it is not an input during inference. Therefore, artificially masking valid GEMS pixels at inference would not alter the model prediction, and such an experiment would not provide an effective holdout test for the present model architecture. The operational prediction relies on TROPOMI and its validity mask, together with EAC4, ERA5 meteorological fields, land cover, elevation, and solar-geometry information; GEMS is not required when generating the reconstructed product.

To examine the model's sensitivity to spatially contiguous gaps in the high-resolution observational input, we performed a controlled masking experiment. In 60 temporally distributed scenes from the independent 2023 test set, originally valid contiguous TROPOMI blocks and their validity masks were artificially removed, while all other model inputs were retained. Valid co-located GEMS pixels were then used only as an independent reference for evaluation. Among the 50 scenes eligible at all four scales, the RMSE relative to GEMS increased by 0.015 Pmolec/cm^2 (1.2%) for 0.25° blocks, 0.853 Pmolec/cm^2 (35.6%) for 0.5° blocks, 1.327 Pmolec/cm^2 (64.9%) for 1° blocks, and 0.723 Pmolec/cm^2 (40.7%) for 2° blocks, compared with the corresponding unmasked predictions at the same locations.

This diagnostic indicates that the model retains reconstruction capability when small areas of high-resolution TROPOMI information are unavailable, while performance decreases as the contiguous input gaps become larger. The experiment characterizes sensitivity to artificially enlarged TROPOMI gaps as a function of gap size. Because collocated GEMS observations are unavailable in naturally GEMS-missing areas, the experiment does not provide a direct reference for those locations and is interpreted specifically as an input-gap sensitivity analysis.

2. No GEMS observations exist before 2023, so GOME-2C is the only independent check on the 2019–2022. The sub-daily variability in the pre-GEMS years is never validated at any hour other than mid-morning (9:30). The manuscript should state this limitation explicitly in the abstract and conclusions, and ideally support the pre-GEMS diurnal cycles with independent hourly observations (e.g., Pandora/MAX-DOAS stations in the KJ and NP domains). Also, the agreement in Figure 5 is not visually convincing, in Figure 5a in particular, it is hard to conclude that the predicted spatial pattern is consistent with GOME-2C.

We recognize the reviewer's concern that the pre-GEMS period is not independently evaluated at all daytime hours. GOME-2C provides an independent satellite reference only near its approximately 09:30 LT overpass, and therefore cannot directly validate the reconstructed sub-daily variability at other hours during 2019–2022. We have now explicitly stated in Sect.5.2 (Revision 14) and Conclusions (Revision 15) that, for 2019–2022, independent satellite evaluation is limited to the approximately 09:30 LT GOME-2C overpass and that the reconstructed variability at other daytime hours has not been independently assessed throughout all study domains.

Revision 14
The independent evaluation of the pre-GEMS period is limited by temporal sampling. For 2019 to 2022, GOME-2C provides an external satellite comparison only near its approximately 09:30 LT overpass, and the reconstructed variability at other daytime hours has not been independently evaluated across all study domains. Cross-sensor retrieval differences add further uncertainty because the TROPOMI input, GEMS training target, and GOME-2C external reference differ in their retrieval configurations, observation times, and spatial sampling. Differences in AMF formulation, a priori NO_2 profiles, cloud treatment, viewing geometry, and spatial footprints may affect both the relationship learned between TROPOMI and GEMS and the discrepancies observed in the GOME-2C evaluation. Future work will assess cross-sensor consistency using retrieval-harmonized products or comparison approaches that explicitly account for these differences.
Revision 15
Quantitative assessments confirm that the reconstructed dataset achieves extremely high fitting accuracy on the test set (R^2 = 0.889) and can effectively recover morning and evening pollution hotspots missed by polar-orbiting instruments like TROPOMI due to their spatiotemporal coverage limitations. More importantly, the long-term retrospective validation against GOME-2C data (R = 0.780) not used in the PhysNorm-Net training demonstrates that the dataset possesses robust temporal stability and high fidelity in historical periods (2019–2022) prior to the operational deployment of GEMS. A limitation is that, for 2019–2022, this independent assessment is confined to the approximately 09:30 LT GOME-2C overpass; the reconstructed variability at other daytime hours has not been independently evaluated across all study domains.

To provide an additional consistency check, we conducted a collocated comparison using the official Pandora tropospheric NO_2 VCD product at the Seoul-SNU station in the KJ domain. The commonly used Pandora direct-sun product is primarily applied to total-column NO_2 comparisons because it observes the direct solar beam and has relatively limited dependence on radiative-transfer assumptions. In contrast, the tropospheric product used here is derived from multi-axis measurements of scattered skylight at several elevation angles using the MAX-DOAS approach and is therefore more sensitive to retrieval assumptions and local spatial representativeness. Measurements with quality flags 0, 1, 10, and 11 were retained, nonpositive and nonfinite retrievals were excluded, and the remaining observations were averaged to hourly UTC values. These observations were then matched to the nearest 0.05° grid cell of the reconstructed product and the original GEMS field.

To ensure a direct comparison between the reconstruction and the original GEMS product, we restricted the analysis to hours for which Pandora, GEMS, and the reconstructed product were all simultaneously available. The resulting statistics are summarized below:

Period	**Product**	**N**	**R**	**RMSE (Pmolec/cm^2)**	**MAE (Pmolec/cm^2)**
2023	GEMS vs. Pandora	1000	0.706	9.757	7.751
2023	Prediction vs. Pandora	1000	0.720	9.313	7.493
2024	GEMS vs. Pandora	687	0.683	9.785	7.918
2024	Prediction vs. Pandora	687	0.594	10.116	8.297
2023–2024	GEMS vs. Pandora	1687	0.696	9.769	7.819
2023–2024	Prediction vs. Pandora	1687	0.676	9.648	7.820

On the common collocated samples from 2023–2024, the reconstructed product and the original GEMS product show comparable consistency with the Pandora tropospheric observations. The correlation coefficient differs by only 0.020, while the RMSE and MAE are nearly identical. This result suggests that the reconstruction does not introduce a substantial degradation relative to the original GEMS product at this station and during these common observation hours.

We nevertheless do not treat the Pandora multi-axis sky-scan tropospheric product as an absolute or universal reference for the entire dataset. Unlike the direct-sun product, which is primarily used for total-column NO_2, the rnvh3 tropospheric product is retrieved from scattered-skylight observations at multiple elevation angles. The latter involves scattered-light radiative transfer and air-mass-factor assumptions and is affected by vertical-profile sensitivity and the difference between the point-scale station footprint and the 0.05° satellite grid. Previous validation studies have also emphasized that Pandora and MAX-DOAS comparisons may contain substantial representativeness and retrieval-related differences, particularly for spatially heterogeneous tropospheric NO_2 distribution (Verhoelst et al., 2021).

In addition, the available long-term record suitable for this analysis was limited to one station in the KJ domain, with no equivalent station-level record included for the NP domain. The number of common observations was also constrained by the availability of valid GEMS retrievals. We therefore regard this analysis as a supplementary, station-level consistency check rather than a domain-wide validation of all daytime hours.

Regarding the visual agreement in Figure 5, we replaced the original scenes with new examples selected to provide more representative valid coverage from GOME-2C.

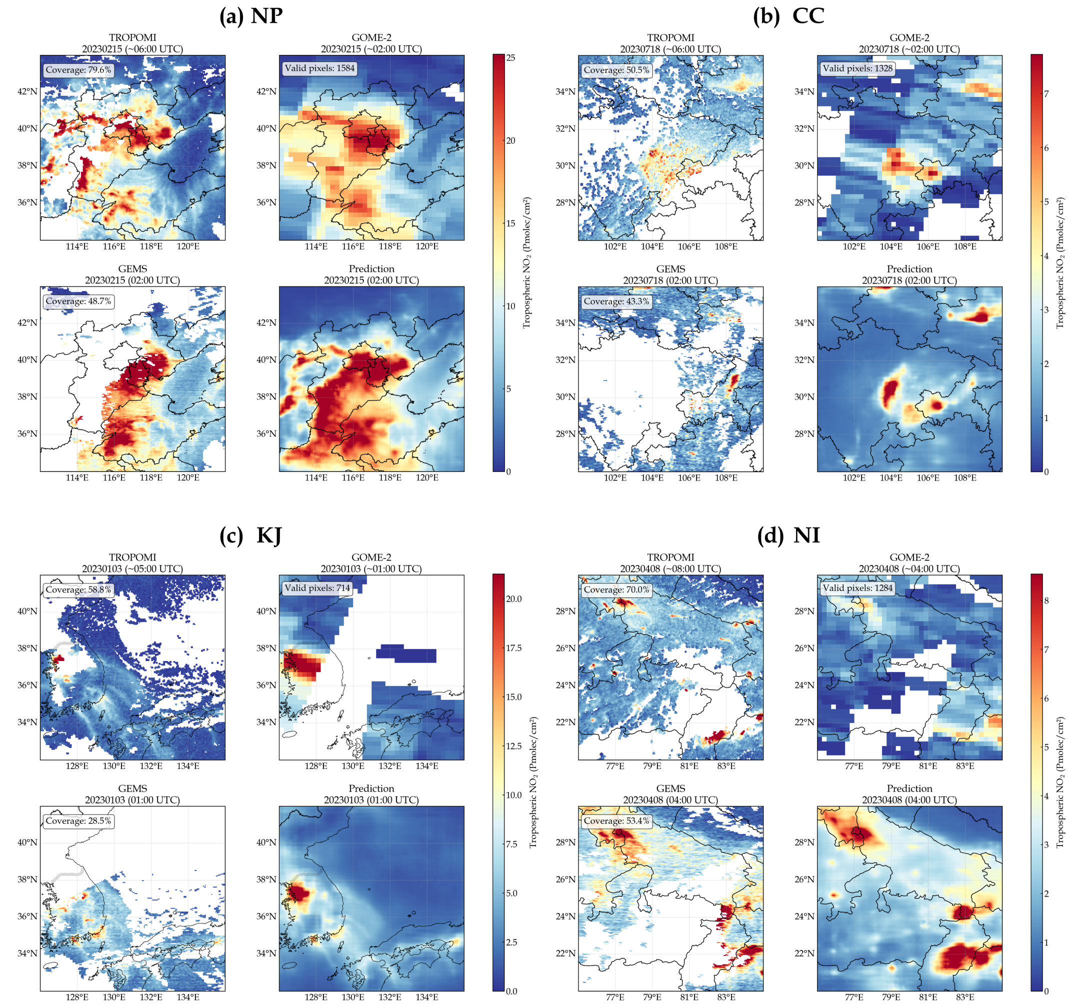


Figure 5: Comparison of tropospheric NO_2 spatial distributions from TROPOMI, GOME-2, GEMS, and model predictions at the GOME-2 overpass time for the four study regions.

3. The 2023 data are randomly partitioned by date. Because both NO_2 and the meteorological priors are strongly autocorrelated at the synoptic scale, a random day-level partition places temporally adjacent, dynamically similar days on opposite sides of the split, so the test set is not independent of the training set in the sense the reported metric implies.

Thank you for identifying the temporal dependence introduced by the random day-level split. A within-2023 split does not provide a strictly independent temporal test because NO_2 and the meteorological predictors may remain autocorrelated across adjacent days.

We have revised the manuscript to distinguish the purposes of the two evaluations. The random 2023 date split was used for model development and to characterize performance on held-out dates within the same year. We no longer describe the resulting test-set statistics as evidence of strict temporal extrapolation or temporal independence. This clarification has been added to Sect. 3.2 (Revision 16) and Sect. 4.1 (Revision 17).

The full 2024 dataset was retained outside the 2023 training, validation, and test split and was not used for model training. We therefore use the 2024 results as the more stringent independent-year assessment of temporal transfer. Despite the expected increase in difficulty relative to the within-year split, the model retained R^2 values of 0.855, 0.771, 0.817, and 0.650 in NP, CC, KJ, and NI, respectively. These results provide evidence that the model retains substantial skill when applied to an unseen calendar year with different meteorological conditions, emission patterns, and retrieval availability.

Revision 16
To ensure numerical stability during neural network training and eliminate the influence of dimensional discrepancies on gradient propagation, differentiated normalization strategies were tailored to the characteristics of each variable type. Prior to normalization, the 2023 dataset was randomly partitioned by date into training (256 days), validation (54 days), and test (55 days) sets. This split was used for model development and for evaluating performance on held-out dates within 2023. Because randomly selected dates may remain synoptically related to temporally adjacent training dates, the 2023 test set is not interpreted as a strictly temporally independent evaluation. The complete 2024 dataset was retained outside the 2023 data split and used for independent-year validation.
Revision 17
To characterize model performance on held-out dates within the 2023 training year, Figure 2 presents the overall scatter density plots for the training, validation, and test sets. The predicted dataset achieved high prediction accuracy across all three sets, with R^2 values of 0.939, 0.882, and 0.889, respectively (all p-values < 0.01).

4. The systematic underestimation of the high tail concerns exactly the regime that matters most for the advertised applications, such as pollution episodes, health exposure, and policy assessment. Yet it remains unquantified.

We appreciate the reviewer's emphasis on the high-value tail. We agree that this regime is relevant to pollution episodes and some exposure- or policy-oriented applications. We therefore conducted an additional value-stratified diagnostic using the existing predictions and valid GEMS labels from the fixed 55-day 2023 test set.

The results show a concentration-dependent behavior rather than a uniform high-value failure. Across the observed 5–40 Pmolec/cm^2 range, which represents 21.733% of valid test pixels, the aggregated NMB is −0.044, with an MAE of 1.994 Pmolec/cm^2 and an RMSE of 2.753 Pmolec/cm^2. In the elevated 20–40 Pmolec/cm^2 range, the mean bias is −3.239 Pmolec/cm^2 (NMB = −0.126; MAE = 4.672 Pmolec/cm^2; N = 317,663). These results indicate that the reconstruction retains useful skill for identifying and characterizing the spatial extent and evolution of elevated pollution conditions.

The underestimation becomes more pronounced only for the most extreme observed values above 40 Pmolec/cm^2. This subset contains 15,037 valid pixels, representing 0.062% of the test set. The mean bias in this tail is −12.203 Pmolec/cm^2 (NMB = −0.264; MAE = 12.442 Pmolec/cm^2; RMSE = 14.201 Pmolec/cm^2), and 96.482% of these pixels are underestimated.

We have also clarified the interpretation of the product. The dataset can support analyses of the occurrence, spatial extent, and temporal evolution of moderate-to-high pollution conditions, including regional episode tracking. However, applications that require the absolute magnitude of individual extreme peaks, such as peak-specific exposure estimates or event-scale policy assessments, should account for the quantified negative bias. The high-tail diagnostic is based on valid GEMS pixels and thus quantifies performance where reference observations are available; it does not independently verify extreme predictions inside GEMS gaps. All the revisions have been incorporated into the manuscript (Revision 18).

Revision 18
Notably, the predicted dataset does not show a systematic underestimation or overestimation trend, except for a slight underestimation in regions with high tropospheric NO_2 column densities (>40 Pmolec/cm^2). A value-stratified evaluation of valid GEMS pixels in the fixed 2023 test set indicates that the reconstruction remains well constrained across the main moderate-to-high concentration range. For observed VCDs between 5 and 40 Pmolec/cm^2, which account for 21.73% of valid test pixels, the aggregated NMB is −0.044, with an MAE of 1.994 Pmolec/cm^2 and an RMSE of 2.753 Pmolec/cm^2. Within the 20–40 Pmolec/cm^2 range, the mean bias is −3.239 Pmolec/cm^2 (NMB = −0.126; N = 317,663), showing that the model retains useful discrimination of elevated pollution conditions. A more pronounced compression occurs only in the extreme tail above 40 Pmolec/cm^2, which contains 15,037 valid pixels (0.062% of the test set): the mean bias is −12.203 Pmolec/cm^2 (NMB = −0.264; MAE = 12.442 Pmolec/cm^2). This phenomenon is relatively common in current deep learning-based atmospheric remote sensing retrieval studies (Wei et al., 2021). The main reasons are twofold: first, the L1 loss functions used for model optimization tend to fit the global mean, with a relatively insufficient penalty weight for extreme high values; second, high-concentration samples constitute a very small fraction of the overall data distribution and often correspond to local, sudden emission events or adverse stagnant meteorological conditions, which low-resolution data cannot fully capture. Nevertheless, considering that the core concentration range of tropospheric NO_2 accounts for the vast majority of samples, the dataset demonstrates high reliability in this range, thereby supporting regional air quality assessment and spatiotemporal evolution analysis, including the characterization of the occurrence, spatial extent, and temporal evolution of elevated pollution conditions. However, analyses requiring the absolute magnitude of individual extreme peaks should account for the quantified negative bias.

5. "Hourly" is in fact daytime hourly. This should be clarified in the title, the abstract, and the dataset documentation.

The temporal coverage has now been clarified consistently throughout the manuscript and dataset documentation. The title has been revised to replace hourly with daytime-hourly. In the abstract, we now explicitly state that the dataset provides hourly fields (Revision 1). We have also replaced hourly with daytime-hourly in other statements referring specifically to the released dataset, including the Introduction (Revision 19), Conclusions (Revision 20) and Data availability (Revision 21). These changes prevent the dataset from being interpreted as providing continuous 24 h coverage.

Revision 19
To address this, this study proposes PhysNorm-Net, a physics-aware network to reconstruct a gapless 0.05° daytime-hourly tropospheric NO_2 dataset covering the period from 1 January 2019 to 31 December 2024 over four 10° × 10° Asian hotspot domains: the North China Plain (NP; 34–44° N, 112–122° E), the Sichuan Basin and its surroundings (CC; 26–36° N, 100–110° E), Korea and Japan (KJ; 32–42° N, 126–136° E), and northern India (NI; 20–30° N, 75–85° E). This architecture is designed to fuse multi-source heterogeneous satellite data with physical and meteorological priors, achieving high-precision, spatiotemporally continuous tropospheric NO_2 vertical column density (VCD) reconstruction. The value and major contributions of this dataset are summarized as follows:

1) High spatiotemporal completeness and resolution: It provides a seamless hourly NO_2 records during daytime at 0.05° resolution covering both pre- and post-GEMS periods, effectively mitigating data gaps caused by cloud contamination and observation limitations. 2) Physical consistency: Instead of simple statistical interpolation, the dataset generation is strictly constrained by meteorological and chemical priors, ensuring that the reconstructed concentration fields adhere to atmospheric physical continuity. Independent validation confirms its robust inter-annual stability. 3) Enhanced utility for long-term environmental analysis: The dataset enables fine-grained delineation of previously obscured periodic diurnal characteristics and accurately captures inter-annual concentration gradients driven by socioeconomic changes over the past six years.
Revision 20
This study presents the first seamless, 0.05° daytime-hourly tropospheric NO_2 vertical column density dataset covering typical Asian hotspots from 2019 to 2024. To overcome the core challenges of spatial discontinuity and short historical records inherent in high-frequency geostationary satellite data, this dataset was reconstructed using an advanced physics-aware deep learning framework (PhysNorm-Net).
Revision 21
TROPOMI (https://doi.org/10.5270/S5P-9bnp8q8) and MCD12C1 (https://doi.org/10.5067/MODIS/MCD12C1.061) data used in this study are available on the EARTHDATA platform. GEMS data can be accessed at https://nesc.nier.go.kr/en/html/datasvc/data.do?pageIndex=1&outputInnb=64&atrb=NO2_Trop. The GOME-2 data are provided by the EUMETSAT AC-SAF as described in https://acsaf.org/products/nto_no2.php. ERA5 on single and pressure levels can be accessed at https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview and https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels?tab=overview, respectively. EAC4 data can be accessed at https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4?tab=overview. ETOPO data can be found at https://www.ncei.noaa.gov/products/etopo-global-relief-model. The gapless 0.05° daytime-hourly tropospheric NO_2 dataset (2019–2024) produced in this study is available at https://doi.org/10.5281/zenodo.20427767 (Gao et al., 2026).

6. EAC4 tcno2 is a total column, while the target is tropospheric. In addition, EAC4 is 3-hourly and linearly interpolated to 1-hourly, which smooths exactly the diurnal chemistry the model is meant to reconstruct. The paper should discuss the expected consequences of both choices.

Thank you for pointing out the two consequences of using EAC4 as a model input. We have revised Sects. 2.6 (Revision 22) and 5.2 (Revision 23) to explain both the mismatch between EAC4 total-column NO_2 and the tropospheric target and the temporal smoothing introduced by interpolating the 3-hourly EAC4 field to hourly resolution.

First, EAC4 tcno2 is a total-column quantity, whereas the supervised target is the GEMS tropospheric NO_2 VCD. We have clarified that EAC4 is not treated as a target-equivalent predictor. Instead, it serves as a spatially complete, coarse-resolution chemical background prior. The model learns the relationship between this prior and the tropospheric target from the hourly GEMS-supervised training data, together with the other physical and satellite predictors. Nevertheless, the total-column-to-tropospheric mismatch means that EAC4 background or stratospheric biases can influence the reconstruction, particularly when the high-resolution observational constraint is weak.

Second, we now explicitly state that linearly interpolating the 3-hourly EAC4 data to hourly resolution preserves large-scale chemical evolution but cannot create independent information on sub-3-hour timescales. Thus, rapidly evolving diurnal NO_2 features may be smoothed when the model relies strongly on the physical-prior branch, especially within TROPOMI gaps. This limitation is partly mitigated by the use of hourly ERA5 fields, solar geometry, and available TROPOMI observations, as well as by supervision against hourly GEMS data, but it cannot be eliminated entirely.

The revised manuscript therefore frames the product as an hourly tropospheric reconstruction constrained by, rather than identical to, the EAC4 chemical field. We also identify higher-temporal-resolution and vertically resolved tropospheric chemical priors as a priority for future work.

Revision 22
This study employs the total column NO_2 (tcno2) variable from EAC4 as an input feature. Because tcno2 includes both tropospheric and stratospheric NO_2, it is not directly equivalent to the GEMS tropospheric VCD used as the supervised target. EAC4 is therefore used as a spatially complete, coarse-resolution chemical background prior rather than a target-equivalent representation of tropospheric NO_2. Although it differs from the satellite-observed tropospheric NO_2 concentration, this product provides a spatiotemporally complete a priori estimate of the NO_2 field, thereby enriching valuable physical prior information available to the model.
Revision 23
At the data-product level, integrating the 0.25° physical prior fields with the 0.05° observations may introduce spatial smoothing. Although this integration can suppress retrieval noise, it may also weaken localized gradients around individual emission sources. Related uncertainty arises from the EAC4 chemical prior because EAC4 tcno2 is a coarse-resolution, 3-hourly total-column quantity, whereas the target is an hourly tropospheric VCD. The column-type mismatch may transmit EAC4 background or stratospheric biases into the reconstruction, while linear temporal interpolation may smooth rapidly evolving chemical features between EAC4 analysis times. These effects are expected to be most relevant where TROPOMI information is absent or sparse. Future work will examine higher-temporal-resolution and vertically resolved tropospheric chemical priors and methods that explicitly represent uncertainties associated with temporal interpolation and column-type mismatch.

7. The zero-out perturbation used for feature importance conflates shared variance among correlated inputs (e.g., q and T at four pressure levels each), so the reported rankings should be interpreted cautiously.

Thank you for pointing out how correlations among the input variables affect the interpretation of the zero-out analysis. Following a related concern raised during review, we have removed Fig. 13 and its group-level aggregation of importance scores, because summing individual-channel scores across groups with unequal numbers of variables could obscure the interpretation of individual predictors.

We have further revised Sect. 5.1 to clarify that several input variables are physically and statistically correlated (Revision 24), including specific humidity and temperature at multiple pressure levels. Consequently, perturbing one channel may be partly compensated for by correlated channels, or may disrupt information jointly represented by several channels. The resulting scores are therefore not additive and should not be interpreted as unique causal contributions or independent explanatory power.

Accordingly, the rankings retained in Fig. 12 are interpreted as conditional model-sensitivity diagnostics under the applied zero-out perturbation and the sampled 2023 daytime conditions. They indicate the variables to which the model output is relatively sensitive, and are interpreted together with the regional and seasonal physical context.

Revision 24
Building on this diagnostic foundation, to further unveil the model's decision-making mechanism, we quantitatively analyzed the relative contributions of each input variable during daytime hours in 2023 using a zero-out perturbation method (Zhang and Gao, 2020). This method calculates a variable's importance score by zeroing out its specific channel and measuring the resulting deviation from the baseline output, quantified by MAE. Figure 12 reports the three highest-ranked individual input channels for each bin. The dominant variable combinations differ significantly across the four regions, indicating that the model does not simply rely on a fixed set of factors but instead forms region-specific control structures. Because several input variables are physically and statistically correlated, such as specific humidity and temperature across multiple pressure levels, the individual-channel scores are not additive and should not be interpreted as unique causal contributions. When correlated variables contain overlapping information, the model may compensate for the perturbation through the remaining channels, or the perturbation may disrupt information shared across several channels. Therefore, the resulting rankings represent conditional model sensitivity under the applied perturbation, rather than independent explanatory importance; the rankings in Fig. 12 are interpreted as diagnostic evidence of relative model sensitivity under the sampled 2023 conditions, and are considered together with the physical context of each region and season.

8. Cross-sensor consistency is not discussed: the labels are GEMS, the high-resolution observational input is TROPOMI (~13:30 LT), and the historical validation is GOME-2C (~09:30 LT). These retrievals differ in AMF formulation, a priori profiles, and resolution, and those systematic differences are silently absorbed into the learned mapping. A brief inter-sensor bias assessment over coincident scenes would substantially strengthen the validation chapter.

Thank you for pointing out that the original manuscript did not adequately discuss cross-sensor consistency. In this study, TROPOMI provides the high-resolution observational input, GEMS provides the training target, and GOME-2C serves as the external reference for the historical evaluation. Because the three products differ in AMF formulation, a priori profiles, cloud treatment, viewing geometry, spatial resolution, and sampling time, their retrieved tropospheric NO_2 VCDs may contain systematic differences.

We have revised Sect. 4.2 (Revision 25) to clarify that, although GOME-2C is independent of the model training, it is not a retrieval-harmonized reference. Accordingly, the GOME-2C comparison is interpreted primarily as evidence of consistency in the spatial patterns and variability captured by the reconstructed product, rather than as an estimate of an absolute sensor-independent NO_2 scale. We have also added a corresponding limitation in Sect. 5.2 (Revision 14), stating that inter-sensor differences may affect both the learned TROPOMI-GEMS relationship and the discrepancies in the GOME-2C evaluation, in addition to real atmospheric variability. Future work will conduct retrieval-harmonized inter-sensor evaluations that explicitly account for the respective a priori profiles, AMFs, cloud screening, observation geometry, and spatial footprints.

Revision 25
In order to evaluate the hotspot reconstruction capability and historical prediction accuracy of the dataset predicted by PhysNorm-Net, this section introduces the GOME-2 instrument as a completely independent reference object. Although GOME-2C is independent of model training, it is not a retrieval-harmonized reference. Differences in AMF formulation, a priori NO_2 profiles, cloud treatment, viewing geometry, and spatial sampling among GOME-2C, GEMS, and TROPOMI may introduce systematic differences among the retrieved tropospheric NO_2 VCDs. The following cross-satellite comparisons are therefore interpreted primarily as an assessment of consistency in spatial patterns and variability, rather than as an estimate of an absolute sensor-independent NO_2 scale.

References

Gödeke, J., Richter, A., Lange, K., Maaß, P., Hong, H., Lee, H., and Park, J.: Hourly surface nitrogen dioxide retrieval from GEMS tropospheric vertical column densities: benefit of using time-contiguous input features for machine learning models, Atmospheric Measurement Techniques, 18, 3747–3779, https://doi.org/10.5194/amt-18-3747-2025, 2025.

Hoinaski, L., Will, R., and Ribeiro, C. B.: Brazilian Atmospheric Inventories – BRAIN: a comprehensive database of air quality in Brazil, Earth System Science Data, 16, 2385–2405, https://doi.org/10.5194/essd-16-2385-2024, 2024.

Huber, D. E., Kerr, G. H., Nawaz, M. O., Runkel, S., Anenberg, S. C., and Goldberg, D. L.: Global NO_2 changes between 2019 and 2024 as observed by TROPOMI in urban areas and emerging hotspots, Atmospheric Chemistry and Physics, 26, 3783–3803, https://doi.org/10.5194/acp-26-3783-2026, 2026.

Verhoelst, T., Compernolle, S., Pinardi, G., Lambert, J.-C., Eskes, H. J., Eichmann, K.-U., Fjæraa, A. M., Granville, J., Niemeijer, S., Cede, A., Tiefengraber, M., Hendrick, F., Pazmiño, A., Bais, A., Bazureau, A., Boersma, K. F., Bognar, K., Dehn, A., Donner, S., Elokhov, A., Gebetsberger, M., Goutail, F., Grutter de la Mora, M., Gruzdev, A., Gratsea, M., Hansen, G. H., Irie, H., Jepsen, N., Kanaya, Y., Karagkiozidis, D., Kivi, R., Kreher, K., Levelt, P. F., Liu, C., Müller, M., Navarro Comas, M., Piters, A. J. M., Pommereau, J.-P., Portafaix, T., Prados-Roman, C., Puentedura, O., Querel, R., Remmers, J., Richter, A., Rimmer, J., Rivera Cárdenas, C., Saavedra de Miguel, L., Sinyakov, V. P., Stremme, W., Strong, K., Van Roozendael, M., Veefkind, J. P., Wagner, T., Wittrock, F., Yela González, M., and Zehner, C.: Ground-based validation of the Copernicus Sentinel-5P TROPOMI NO_2 measurements with the NDACC ZSL-DOAS, MAX-DOAS and Pandonia global networks, Atmospheric Measurement Techniques, 14, 481–510, https://doi.org/10.5194/amt-14-481-2021, 2021.