Dear Editor and Reviewers,

We thank the Editor for the continued handling of our manuscript and the reviewers for the further comments and suggestions, which have helped improve the manuscript. We have revised the manuscript accordingly.

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.

1. Comparisons with ground-based Pandora observations are a widely accepted approach for evaluating satellite retrievals. Are additional Pandora sites available to evaluate the reconstructed product? Furthermore, why is the comparison restricted to dates when both GEMS observations and the reconstructed product are available, given that gap filling is a central objective of this product? Please also evaluate the reconstructed product against Pandora observations when GEMS observations are unavailable. How does the correlation coefficient (R) for these gap-filled observations compare with that between Pandora and GEMS when GEMS observations are available?

Thank you for this suggestion. The previous restriction to times when both GEMS and the reconstructed product were available was intended to enable comparison against Pandora using identical reference samples. However, this comparison alone did not assess reconstruction performance when GEMS observations were unavailable.

The evaluation has now been expanded from Seoul-SNU to 16 Pandora stations with usable observations within the study domains during 2019–2024. All available Pandora–reconstruction collocations were included, regardless of GEMS availability. This yielded 83,262 station-hour pairs, with an overall R of 0.644 and an RMSE of 7.200 Pmolec/cm^2. Annual and combined statistics are reported in Table 4.

For 2023–2024, the evaluation was additionally stratified by GEMS availability. On the 16,407 common valid samples, GEMS and the reconstructed product achieved R values of 0.755 and 0.712, respectively, against Pandora. When the corresponding GEMS file was present, but the collocated GEMS pixel was invalid, 21,234 Pandora–reconstruction pairs were available, yielding R = 0.612 and RMSE = 7.121 Pmolec/cm^2. The annual correlations for these gap-filled samples were 0.659 in 2023 and 0.570 in 2024, compared with Pandora–GEMS correlations of 0.804 and 0.699 on the respective valid-GEMS subsets. These results are presented together in Table 5.

The revised manuscript now describes the Pandora observations in Sect. 2.5 (Revision 1), and the results in a new Sect. 4.3 (Revision 2). The uneven station coverage and the limitations of the ground-based comparison are also stated (Revision 3&4). We have also revised the abstract.

Revision 1
2.5 Pandora ground-based NO_2 observations Pandora is a ground-based ultraviolet–visible spectrometer system that measures atmospheric trace gases, including NO_2, using direct-sun and scattered-sky observations (Herman et al., 2009). The Pandonia Global Network (PGN) coordinates the processing and distribution of Pandora measurements to support air-quality monitoring and satellite validation (Verhoelst et al., 2021; Park et al., 2026). Its high temporal resolution provides a ground-based reference for evaluating the

temporal variability of satellite observations and reconstructed products. This study uses the rnvh3p1-8 tropospheric NO_2 column product derived from sky-scanning observations. Observations available during 2019–2024 were obtained as an external reference independent of model training. Usable observations were identified at 16 stations: Beijing-RADI in NP; Busan, Daegu, Fukuoka, Incheon-ESC, Jeonju, Kobe, Seosan, Seoul, Seoul-KU, Seoul-SNU, Suwon-USW, Ulsan, and Yongin in KJ; and Haldwani-ARIES and Nainital-ARIES in NI. No eligible station was identified in CC for this evaluation. Observational coverage varied among stations and years.

The evaluation used finite positive tropospheric NO_2 columns with quality flags 10 or 11, corresponding to preliminary high- and medium-quality estimates, respectively, that had not undergone formal quality assurance. Low-quality observations (flag 12) were excluded, and no additional upper concentration threshold was applied. Column amounts were converted from mol/m^2 to Pmolec/cm^2. At Seoul-SNU, Pandora 149 was used preferentially, and Pandora 163 supplied only hours without valid observations from Pandora 149, avoiding duplicate station-hour samples. The rnvh3 tropospheric column represents approximately the surface-to-4 km column; differences in vertical sensitivity and spatial sampling relative to satellite retrievals should therefore be considered when interpreting the comparisons.

Revision 2

4.3 Evaluation against ground-based Pandora observations

For each hourly product timestamp, valid Pandora observations within $[t - 30\,min, t + 30\,min)$ were averaged and matched to the nearest 0.05° grid cell. Each hourly average required at least one valid Pandora observation, and each station contributed at most one sample per hour. Comparisons used finite positive product values. Annual and combined R, RMSE, and MAE were calculated from the collocated station-hour pairs. Reconstructed values were also evaluated separately where the corresponding GEMS scene was present but the collocated pixel lacked a valid observation.

The evaluation yielded 83,262 collocated station-hour pairs across 16 stations during 2019–2024 (Table 4). The combined R, RMSE, and MAE were 0.644, 7.200 Pmolec/cm^2, and 4.896 Pmolec/cm^2, respectively. Annual R ranged from 0.539 to 0.696, while RMSE ranged from 6.528 to 9.052 Pmolec/cm^2. These comparisons extend the ground-based assessment to the historical reconstruction period and to hours without valid GEMS observations. However, the number of contributing stations increased from two in 2019 to 16 in 2024. Differences among annual statistics therefore reflect changes in station composition and observational coverage as well as reconstruction performance, and should not be interpreted as a temporal trend in product accuracy.

Table 4. Annual and combined evaluation of the reconstructed NO_2 product against Pandora observations during 2019–2024.

Year	Stations	N	R	RMSE	MAE
2019	2	588	0.539	9.052	5.696
2020	3	2,884	0.665	8.635	5.616
2021	8	11,942	0.638	8.569	5.878
2022	9	22,139	0.617	7.208	4.950
2023	11	17,133	0.696	6.868	4.837
2024	16	28,576	0.598	6.528	4.389
2019–2024	16	83,262	0.644	7.200	4.896

For 2023–2024, GEMS and the reconstruction achieved R values of 0.755 and 0.712, respectively, against Pandora on 16,407 common-valid samples, with corresponding RMSE values of 6.186 and 6.464 Pmolec/cm^2 (Table 5). On the 21,234 samples where the collocated GEMS pixel lacked a valid observation, the reconstruction achieved R = 0.612 and RMSE = 7.121 Pmolec/cm^2. The correlations for these gap-filled samples were 0.659 in 2023 and 0.570 in 2024, compared with Pandora–GEMS correlations of 0.804 and 0.699 on the respective valid-GEMS subsets. The lower combined correlation and higher RMSE for the gap-filled samples indicate reduced agreement under these sampling conditions. However, the subsets differ in observational conditions, so these differences cannot be attributed solely to gap filling. The station-based results also do not establish uniform accuracy across all reconstructed pixels or study regions.

Table 5. Evaluation against Pandora observations during 2023–2024: (a) comparison of GEMS and the reconstruction on common-valid samples; (b) evaluation of the reconstruction where collocated GEMS pixels lack valid observations.

(a) Common-valid samples

Period	N	GEMS R	GEMS RMSE	Reconstruction R	Reconstruction RMSE
2023	6844	0.804	6.493	0.765	6.584
2024	9563	0.699	5.957	0.651	6.377
2023–2024	16407	0.755	6.186	0.712	6.464

(b) GEMS-gap samples

Period	N	Reconstruction R	Reconstruction RMSE
2023	7,693	0.659	7.204
2024	13,541	0.570	7.073
2023–2024	21,234	0.612	7.121

Revision 3

Independent evaluation of the 2019–2022 historical reconstruction remains limited by the temporal and spatial coverage of the reference observations. GOME-2C provides a satellite comparison near its approximately 09:30 LT overpass, while Pandora supplies additional hourly comparisons at the available ground stations. However, Pandora coverage is concentrated in KJ, with limited coverage in NP and NI and no eligible station in CC in this evaluation. These observations therefore do not provide comprehensive validation of daytime variability throughout all study domains.

Revision 4

Pandora observations provide additional ground-based evaluation during the historical reconstruction period, although their uneven station coverage, together with the restricted overpass sampling of GOME-2C, limits comprehensive validation of daytime variability across all study domains.

2. The response letter states that the scenes in Figure 5 were replaced with examples showing more representative GOME-2C coverage. However, in revised manuscript, only panel (a) appears to have changed. This revision does not appear sufficient to support the statement in the response letter.

Thank you for identifying this inconsistency. In the present revision, panel (a), showing NP on

15 February 2023, is retained, whereas the scenes in panels (b)–(d) have been replaced. Coverage of GOME-2C has increased in the KJ and NI examples, whereas the updated CC example provides more extensive GEMS coverage for the spatial comparison. We have revised the accompanying text to identify the selected scenes and describe both the spatial similarities and the remaining discrepancies.

Revision 5

Figure 5 presents selected examples of morning NO_2 spatial distributions across the four study regions. In the selected NP case, both the reconstruction and GOME-2C observations show an extensive area of elevated NO_2 over the North China Plain, with finer spatial structures visible in the reconstruction and available GEMS observations. In the CC case, enhanced columns around Chengdu, Chongqing, and the northeastern part of the domain are evident in both the reconstruction and satellite observations, although their magnitudes and spatial extents differ. In the KJ case, the reconstruction captures enhancements over western and southern South Korea that are also apparent in the available GEMS observations, whereas GOME-2C shows a smoother distribution and a weaker enhancement near Seoul. In the NI case, both the reconstruction and GOME-2C indicate elevated columns in the southeastern part of the domain, while differences remain in the intensity and extent of individual features. These cases illustrate the reconstruction's ability to represent broad regional pollution patterns, while also revealing differences among the reconstructed fields and satellite observations.

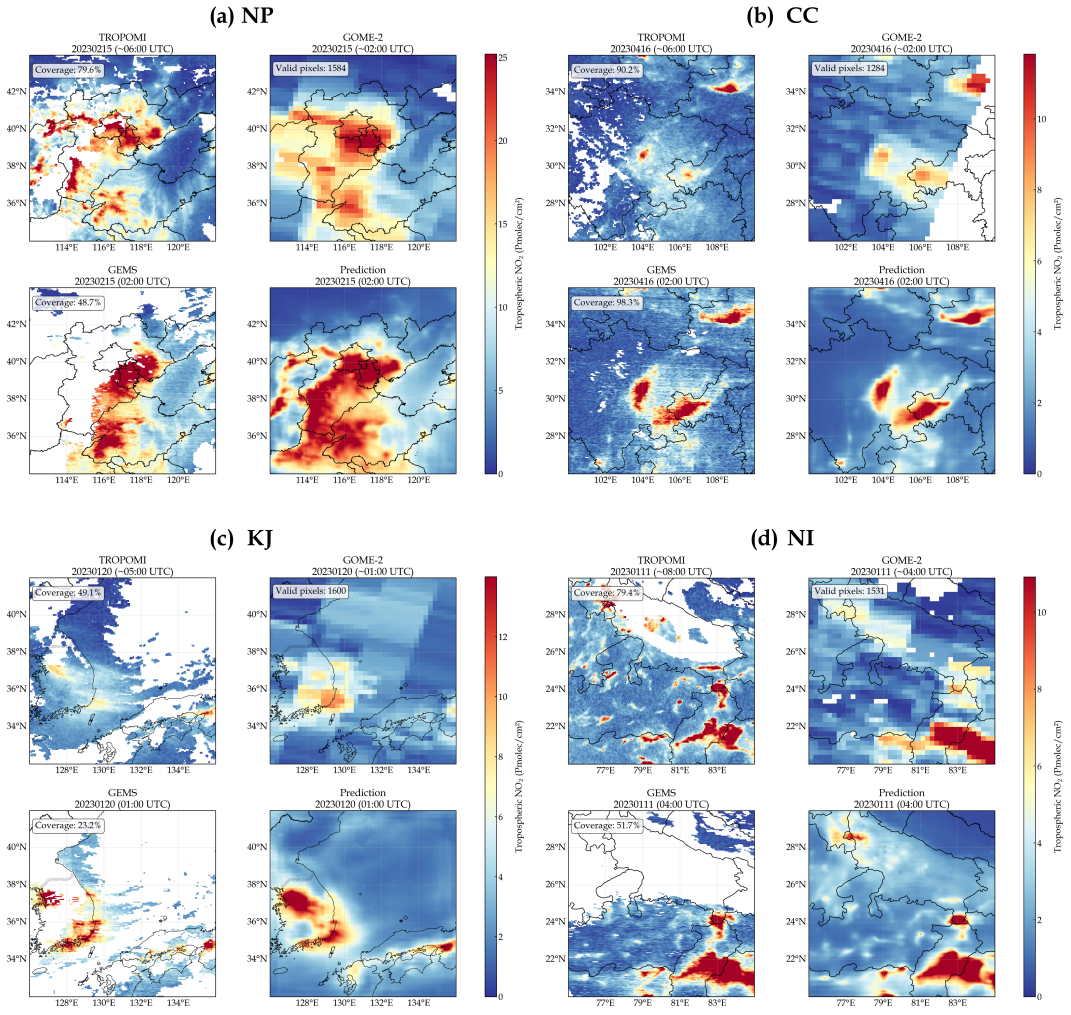

Figure 5: Comparison of tropospheric NO_2 spatial distributions from TROPOMI, GOME-2C, GEMS, and model

predictions. GEMS and predictions are shown at the hourly time corresponding approximately to the GOME-2C overpass; TROPOMI observations are shown at their own overpass times.

3. The gap-masking experiment and the evaluation for pixels with valid GEMS observations but missing TROPOMI observations provide the evidence users need to assess the product's gap-filling performance. However, these results appear only in the response letter and should be incorporated into the manuscript.

Thank you for this suggestion. The evaluation under missing TROPOMI input conditions and the controlled masking experiment have now been incorporated into Sect. 4.1 (Revision 6), for which the reconstruction achieved an R^2 of 0.801 and an RMSE of 1.235 Pmolec/cm^2 on the held-out 2023 test set. We have also incorporated the design and quantitative results of the controlled TROPOMI gap-masking experiment, including the scene selection, four masking scales, and RMSE increases relative to the corresponding unmasked predictions. The revised text explains that the smallest tested gaps had a limited effect on performance, whereas the larger gaps produced greater degradation, although the response was not monotonic with gap size. We also clarify that both evaluations assess reconstruction performance when TROPOMI inputs are missing but valid GEMS reference observations are available; they do not directly validate locations where GEMS itself is missing.

Revision 6

To assess reconstruction performance under TROPOMI input gaps, two complementary evaluations were conducted using the held-out 2023 test set. First, pixels with valid GEMS observations but missing TROPOMI inputs were evaluated, with the original TROPOMI availability mask retained during inference. The reconstruction achieved an R^2 of 0.801 and an RMSE of 1.235 Pmolec/cm^2 over this subset. Second, a controlled masking experiment assessed sensitivity to spatially contiguous input gaps using 60 temporally distributed test scenes. Because GEMS serves as the supervised target and evaluation reference rather than an inference input, artificial gaps were introduced by masking originally valid TROPOMI observations within contiguous blocks and setting the corresponding validity-mask values to indicate missing inputs, while retaining all other inputs. Four block scales were tested: 0.25°, 0.5°, 1°, and 2°. For the scenes eligible at all four scales, predictions within the artificially masked areas were evaluated against valid co-located GEMS observations. Relative to the corresponding unmasked predictions at the same locations, the RMSE increased by 0.015 Pmolec/cm^2 (1.2%), 0.853 Pmolec/cm^2 (35.6%), 1.327 Pmolec/cm^2 (64.9%), and 0.723 Pmolec/cm^2 (40.7%), respectively. These results indicate a small performance loss for the 0.25° gaps and greater degradation for the larger tested gaps, although the response was not monotonic with gap size. Together, these evaluations quantify reconstruction performance under naturally occurring and artificially imposed TROPOMI input gaps where valid GEMS references are available. Reconstructed pixels where GEMS itself is unavailable cannot be directly verified against GEMS. Therefore, these results should not be interpreted as pixelwise validation of the entire gapless product.

4. What is meant by "central bias elasticity"?

We are sorry for the confusion caused by the term "central bias elasticity". This term refers to the dimensionless sensitivity of the mean reconstructed NO_2 VCD to a relative perturbation in the EAC4 input, estimated using a central difference between the +20% and −20% perturbation experiments. Specifically,

$$E = \frac{\bar{\hat{y}}_{+20\%} - \bar{\hat{y}}_{-20\%}}{0.4\bar{\hat{y}}_{\text{baseline}}}$$

where the three means are calculated over the same GEMS-valid, TROPOMI-missing pixels under the positively perturbed, negatively perturbed, and baseline EAC4 inputs, respectively. The denominator 0.4 represents the difference between the two relative input perturbations. The reported changes of +4.02% and −4.60% therefore yield $E = [0.0402 - (-0.0460)]/0.4 \approx 0.216$. This quantity measures the response to an imposed input perturbation; it does not measure prediction bias relative to GEMS. To remove this ambiguity, we have replaced *central bias elasticity* with *normalized sensitivity to EAC4 perturbations* in Sect. 5.1 and clarified its definition and interpretation (Revision 7).

Revision 7
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 normalized sensitivity to EAC4 perturbations of 0.216 (estimated using a central difference as $E = \frac{\bar{\hat{y}}_{+20\%} - \bar{\hat{y}}_{-20\%}}{0.4\bar{\hat{y}}_{\text{baseline}}}$, where $\bar{\hat{y}}_{+20\%}$, $\bar{\hat{y}}_{-20\%}$, and $\bar{\hat{y}}_{\text{baseline}}$ denote the mean reconstructed NO_2 VCD over the same pixel subset with EAC4 increased by 20%, decreased by 20%, and unchanged, respectively). This dimensionless metric quantifies the relative response of the pooled mean reconstruction to the imposed EAC4 perturbations, rather than prediction bias relative to GEMS. Its value of 0.216 indicates that the relative output response is approximately 21.6% of the relative input perturbation, as estimated across the tested ±20% range. The region-specific normalized sensitivities 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.

References

Herman, J., Cede, A., Spinei, E., Mount, G., Tzortziou, M., and Abuhassan, N.: NO2 column amounts from ground-based Pandora and MFDOAS spectrometers using the direct-sun DOAS technique: Intercomparisons and application to OMI validation, Journal of Geophysical Research: Atmospheres, 114, https://doi.org/10.1029/2009JD011848, 2009.

Park, J.-U., Lim, S., Hanisco, T. F., Abuhassan, N., Place, B. K., Pandey, A., Cede, A., Tiefengraber, M., Gebetsberger, M., Park, J., Choi, J., Crawford, J. H., Song, C.-K., and Kim, S.-W.: Global analysis of nitrogen dioxide and formaldehyde column densities from the Pandora global network: Variability and implications for satellite validation, Remote Sensing of Environment, 335, 115249, https://doi.org/10.1016/j.rse.2026.115249, 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.