Response to Referee #1 (RC1)

Referee comments are in blue. Our replies are in black. Text quoted from the revised manuscript or Supplement is in red, with its location, and is copied verbatim from the tracked-changes files. Figures and tables of the revised manuscript are reproduced under the reply where they are first needed.

Manuscript: egusphere-2026-3233, "Physics-constrained inverse neural estimation (PINE) of NOx emission spatial patterns from TROPOMI NO2 columns over the North China Plain" (title revised)

Authors: Yinan Wang, Yubing Pan, Daren Lyu

We thank the referee for a careful report. The comments on the EDGAR normalization, the redistribution metric and the independence of the checks changed how the paper is framed, and we think the revised version is clearer about what the satellite constrains and what it does not. Section, figure and table numbers refer to the revised manuscript; all changes are shown as tracked changes.

During the revision we found that the TROPOMI quality screening stated in Table 1 (qa_value ≥ 0.75, cloud radiance fraction < 0.5) had not been applied correctly when the columns were gridded. All data were re-gridded with the stated screening, all networks were retrained and all numbers were recomputed. For this reason most values differ from the discussion paper; for example, the reconstruction r on the 54 validation days is now 0.332 for EDGAR and 0.856 for PINE-NOx (0.331 and 0.871 over all 347 days). These values are not directly comparable with the submitted 0.395 and 0.837, which were computed on 80 different validation days and on the raw column; the revised values use a new fixed split and the background-removed column (Sect. 3.1). The validation days were not used to fit the weights but were used to select the checkpoint (the training epoch kept), so they are not an independent test set (Sect. 2.3).

While checking Table 1 against the files used, we also found that the processor version was given as OFFL v2.4.0 for the whole year, whereas the 2023 record contains three versions. Table 1 and Sect. 2.1 now state the three versions with their dates, and Sect. 2.1 reports what the version change does to the result (see Comment 5, last paragraph).

We also added a synthetic test with known emissions (Sect. 3.3, new Text S1 and Table S2). WRF-Chem columns simulated from EDGAR for July and January were sampled with the daily TROPOMI valid-pixel mask and inverted with the same networks, so that the true emissions equal the prior. The inversion then lowered the share of the urban–industrial cores by about as much as it does with the observed columns. We therefore describe the decrease relative to EDGAR qualitatively and no longer present its size as an observed redistribution. What we now report is the comparison with MEIC judged against this test (Comment 4).

The values below are the ones that changed and that this report refers to.

Quantity	Discussion paper	Revised paper	Reason
Reconstruction r, EDGAR prior	0.395	0.331	quality screening applied as stated; all 347 evaluated days instead of 80 validation days; skill computed on

			the background-removed column
Reconstruction r, PINE-NOx (UNet-CBAM)	0.837	0.856 (54 validation days; 0.871 over all 347 days)	quality screening applied as stated; networks retrained; new fixed validation days (54 instead of 80, also used to select the checkpoint); skill computed on the background-removed column
"Spatial redistribution relative to EDGAR"	about 11 %	not reported as a number; described qualitatively	the 11 % was $1 - r^2$ of a log-space correlation; a synthetic test shows that a change of this size also arises from the method (Comment 4)
Day-to-day variability of the domain mean	CV 29 %	withdrawn	the demeaned loss does not constrain the domain level (Comment 6)
TROPOMI processor version, Table 1	OFFL v2.4.0	v2.4.0, v2.5.0, v2.6.0	description corrected from the files used

Major comments

Comment 1

The manuscript states that PINE-NOx retrieves daily NOx emissions, but Section 2.5 clearly indicates that the domain-wide annual total is normalized to EDGAR. The Abstract should explicitly mention that the final PINE-NOx emissions are normalized to the EDGAR annual total. Without this clarification, readers may interpret the absolute emission magnitude as being independently constrained by TROPOMI, whereas the satellite observations mainly constrain spatial redistribution and temporal variability. "….. PINE-NOx retrieves daily spatial-temporal emission adjustments constrained by TROPOMI, with the annual total normalized to EDGAR."

We agree. TROPOMI constrains where the emissions are, not how large the domain total is, and the abstract now says so. "Daily" has been removed from the title, which now reads "… of NOx emission spatial patterns from TROPOMI NO2 columns …". The products are described as annual and monthly patterns, with the daily fields described as overpass-time effective fields (see also Reviewer 2, Comment 3).

The same statement is made in Sect. 2.4 (point 5), Sect. 2.5, Sect. 4.3 and the Conclusions. Sect. 2.5 also states how the monthly totals of the product are set: they follow the EDGAR v8.1 monthly profile for 2022 (built from the EDGAR sector maps and the sector-specific monthly fractions for China), so that the annual and the monthly totals come from one inventory and neither is presented as a satellite result.

Changes in the manuscript:

Title: "Physics-constrained inverse neural estimation (PINE) of NOx emission spatial patterns from TROPOMI NO2 columns over the North China Plain"

Abstract: "We retrieve annual, monthly and overpass-time daily emission patterns over the North China Plain at 0.1° for 2023, with the annual total normalized to the EDGAR inventory, so the observations constrain only the spatial pattern."

Sect. 2.4, point (5): "The same invariance means that the domain-wide emission level is not constrained by the observations at all: multiplying all emissions by a constant leaves the loss unchanged, so the level is set only by the prior term and, in the final product, by the normalization of Sect. 2.5."

Sect. 2.5: "Monthly totals of the product follow the EDGAR v8.1 monthly profile for 2022, built from the EDGAR sector maps and the sector-specific monthly fractions for China (winter-to-summer ratio 1.09), so that annual and monthly totals come from one inventory."

Table 1, Post-processing: "c = sum(E_{EDGAR}) / sum(mean$_t$[E]), E_{final} = E × c, locking the domain annual total to EDGAR; monthly totals follow the EDGAR v8.1 monthly profile"

Sect. 5, Conclusions: "With the annual total held at EDGAR, the PINE-NOx field places less of the emissions in the urban–industrial cores than EDGAR, but a synthetic test with known emissions shows that a decrease of this size also arises from the method itself, so its size is not interpreted as an observed change."

Sect. 4.3: "We emphasize that the absolute magnitude of the inversion is not determined independently by the observations: the rendering loss is demeaned log, and the domain annual total is locked to EDGAR by design, so the agreement of the annual total with EDGAR is a design constraint rather than an independent validation; what the observations constrain is the spatial pattern."

Comment 2

The manuscript normalizes the domain-wide annual total emission to EDGAR, but the rationale for choosing EDGAR rather than MEIC is not sufficiently justified. Since MEIC is available as an independent bottom-up inventory and has a lower annual total than EDGAR, the choice of EDGAR directly affects the absolute magnitude of the final emissions. The authors should explain why EDGAR is selected as the normalization target and provide a sensitivity test using MEIC-based or alternative annual-total normalization. This is particularly important because the manuscript notes that the static EDGAR 2022 inventory may not fully reflect the continued decline of NOx emissions over the North China Plain in 2023.

We chose EDGAR v8.1 for three reasons. It is a public, globally consistent 0.1° inventory, so the method can be applied elsewhere without a regional inventory. The same inventory then provides both the prior pattern and the total, which keeps the design constraint in one place. And it leaves MEIC, which the inversion never uses, available as an independent reference for the spatial pattern (Sect. 3.3, Figs. 10 and 12).

The normalization in Eq. (6) is one domain-wide constant applied after training. Normalizing to the MEIC total instead therefore multiplies every field by the same factor, MEIC/EDGAR = 0.88 for the domain, and leaves every pattern, share and correlation in the paper unchanged. This is an exact identity, so a separate run would give the same patterns; Sect. 2.5 now states it. The consequence for the absolute level is shown in the table.

Normalization target	**Domain annual total (NOx as NO2)**	**Factor relative to the EDGAR-normalized product**	**Effect on patterns, shares, correlations**
EDGAR v8.1 (2022), used here	3031 Gg yr^{-1}	1	none (reference)
MEIC 2023	2666 Gg yr^{-1}	0.88	none (one multiplicative constant)

We agree that EDGAR 2022 may overestimate the 2023 total, and the MEIC 2023 ratio of 0.88 is the best indication we have of the size of this effect. Because the demeaned loss carries no information on the domain level, the inversion cannot correct the total for the 2023 decline. Sect. 2.5 now says this, and Sect. 2.4 and Sect. 4.3 state that the domain level is not constrained by the observations.

Changes in the manuscript:

Sect. 2.5: "The lower MEIC 2023 total is consistent with the continued decline of NOx emissions after 2022, although the two inventories also differ in method; the inversion cannot correct the EDGAR total for this decline, because the domain level is not constrained by the observations."

Sect. 2.5: "Normalizing to the MEIC total instead (0.88 of the EDGAR total) would rescale every field by the same factor and leave all patterns, shares and correlations unchanged."

Sect. 2.5: "Because the normalization is a single domain-wide constant, all log-space correlations and log-RMSE values are mathematically strictly unaffected by its value (log(E c) = log(E) + log c is merely a constant shift); hence the agreement of the annual total with EDGAR is a design constraint on the absolute magnitude rather than an independent validation."

Sect. 4.3: "We emphasize that the absolute magnitude of the inversion is not determined independently by the observations: the rendering loss is demeaned log, and the domain annual total is locked to EDGAR by design, so the agreement of the annual total with EDGAR is a design constraint rather than an independent validation; what the observations constrain is the spatial pattern."

Comment 3

In Figure 11(c), the interpretation of relative differences in clean peripheral regions requires further clarification. EDGAR emissions in these regions may be extremely small or even zero, but such low values do not necessarily indicate truly negligible emissions; they may also reflect limited activity statistics or spatial allocation uncertainty in the inventory. If these grids are included in the inversion, Huber prior regularization, or annual-total normalization, the authors should clarify whether any minimum emission threshold or floor value was applied, and how relative differences were calculated for near-zero EDGAR values. A short discussion of whether low-prior grids are physically meaningful, inventory-limited, or excluded from quantitative interpretation would improve the robustness of the spatial redistribution analysis. Relevant studies that discuss low-prior emissions, prior-related uncertainty, and uncertainty-aware satellite-based NOx emission inversions may also be cited, including but not limited to: https://doi.org/10.1088/1748-9326/ac48b4; https://doi.org/10.1016/j.rse.2023.113720; https://doi.org/10.1016/j.rse.2023.113917; https://doi.org/10.5194/acp-25-2291-2025; https://doi.org/10.5194/essd-17-3329-2025

No emission floor or threshold was applied. The EDGAR v8.1 field on our 0.1° grid has no zero cells (the smallest value is about 1/40 000 of the largest), and the inversion is multiplicative around the prior, $E = E_{EDGAR} \exp(\delta)$ with δ limited to $[-3, +3]$ (Table 1, Sect. 2.3). The relative difference in the submitted Fig. 11c was computed cell by cell without masking and displayed on a ±60 % scale. One related step exists on the observation side and is now stated in Sect. 2.3: cells whose observed column does not exceed the boundary background (31 % of the valid cell-days) carry no local signal and are left out of the rendering term.

On the physical meaning of the low-prior cells, we looked at how much information a single overpass carries there. For the weakest-emitting cells, which together hold 10 % of the EDGAR total (2226 of 3621 cells), the column produced by a cell's own emissions is small compared with the TROPOMI precision: the median daily signal-to-noise ratio is 0.04 to 0.06, depending on whether pixel errors are treated as correlated or independent, and about 80 % of the local column in these cells is transported in from elsewhere. For the cells that hold the top half of the EDGAR total, the ratio is 1.1 to 2.0. The satellite therefore cannot decide, cell by cell, whether a low prior value reflects truly small emissions or a misallocation in the inventory. MEIC places 32 % of its domain total in these same cells and PINE-NOx 25 %, against 10 % for EDGAR, so inventory

allocation is a plausible part of the difference, but we do not claim missing sources. In the synthetic test described under Comment 4, the inversion also raises the share of the weakly emitting cells (the half of the cells with the lowest EDGAR emissions) when the true emissions equal EDGAR, by 9 and 12 percentage points in July and January against 8 and 11 for the observed columns, so the higher share in PINE-NOx is not, by itself, evidence of emissions missing from EDGAR. We give this analysis here and do not add it to the paper.

Cells (ranked by EDGAR emission)	Median daily signal-to-noise ratio of a cell's own column	Share of the domain total (EDGAR / PINE-NOx / MEIC)
Weakest-emitting cells holding 10 % of the EDGAR total (2226 of 3621)	0.04 to 0.06	10 % / 25 % / 32 %
Cells holding the top half of the EDGAR total	1.1 to 2.0	not evaluated

The revision handles these cells as follows. Fig. 8 (which replaces Fig. 11) shows the ratio of emission shares on a ×1/3 to ×3 scale with 0.4 % of cells outside it, and no statement relies on the ratio in individual low-prior cells: the quantitative results are given as shares of the domain total, for 0.5° blocks or for classes of cells ranked by EDGAR emission. The uncertainty budget shows that the prior term grows towards the weakly emitting cells (Fig. 9b), and Sect. 3.3 states that the observations carry least information there. We thank the referee for the suggested literature. Sect. 3.3 now cites Lu et al. (2025), whose mass-conserving inversion over China found substantial emissions in cells with low prior values, as a reminder that a low prior is not evidence of low emissions.

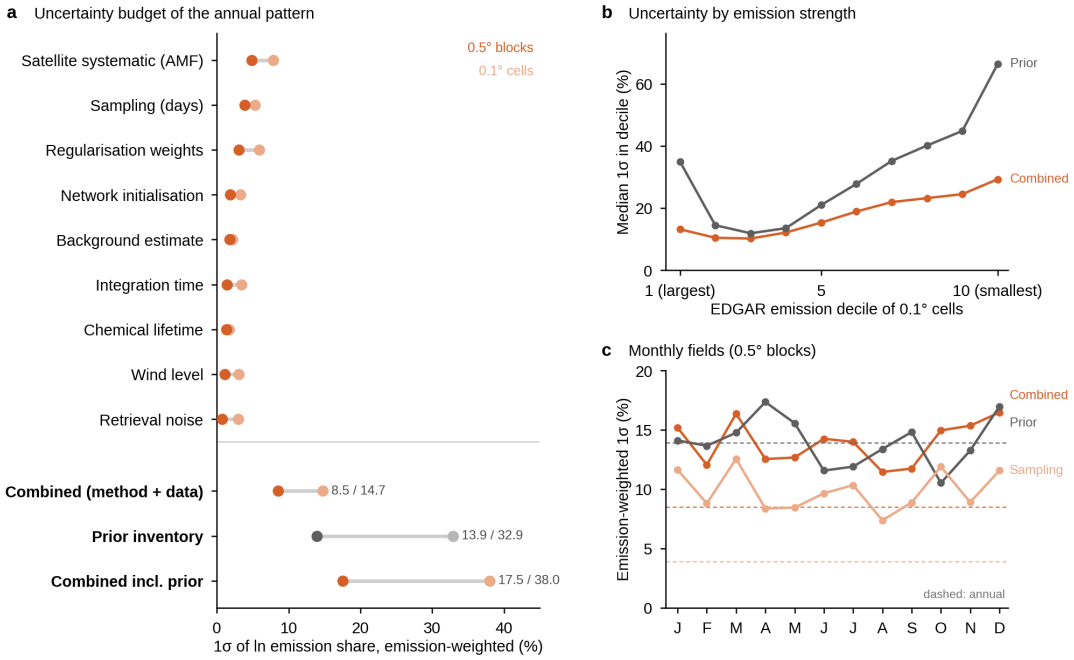


Figure 9. Uncertainty of the emission pattern, expressed as 1σ of the natural logarithm of the emission share. (a) Annual budget by source, emission-weighted over 0.5° blocks (dark) and 0.1° cells (light); combined values add the sources in quadrature in each cell and are then emission-weighted, so they differ from the quadrature sum of the weighted single-source values; the prior-inventory term (EDGAR or MEIC as prior) is shown separately. (b) Median

1σ of 0.1° cells in each EDGAR emission decile, for the combined method and data terms and for the prior term. (c) Emission-weighted 1σ of the monthly fields at 0.5° for the combined term, the prior term and day sampling; dashed lines give the annual values.

Changes in the manuscript:

Sect. 2.3: "Cells whose observed column does not exceed the boundary background (31 % of the valid cell-days) carry no local signal and are left out of the rendering term."

Table 1, Output: "$\delta = \log(E/E_{EDGAR})$, tanh-clipped to [-3, +3], $E = E_{EDGAR} \exp(\delta)$"

Sect. 3.3: "The prior term grows towards the weakly emitting cells, where the observations carry least information and where a mass-conserving inversion over China has found substantial emissions in cells with low prior values (Lu et al., 2025), and the monthly fields are 1.3 to 1.9 times as uncertain as the annual field (method-and-data term 11.5–16.5 % against 8.5 % at 0.5°), mainly because each month has fewer days."

Comment 4

The manuscript reports "about 11% spatial redistribution relative to EDGAR", but the metric is not clearly defined. Since the domain-wide annual total is normalized to EDGAR, this redistribution should also be quantified explicitly. The authors should define how the 11% value is calculated, for example whether it represents a mean absolute grid difference, an L1 redistribution fraction, or another metric. It would also be helpful to discuss whether enhanced low emissions indicate possible missing or misallocated sources in EDGAR, or simply reflect redistribution under the EDGAR-normalized total constraint.

The referee is right that the metric was not defined. The "about 11 %" was $1 - r^2$ of the log-space correlation between the PINE-NOx and EDGAR annual fields, which is not a measure of redistribution, and it has been removed. To compare two spatial patterns, the revision uses the L1 measure that the referee suggests: the fraction of the total emission placed in different grid cells, $\frac{1}{2} \Sigma_i |s_i - s'_i|$, where s_i and s'_i are the shares of the domain total in cell i (0 % for identical allocation, 100 % for no overlap). Because both fields are shares of their own totals, the EDGAR normalization does not enter it. Sect. 3.3 describes the PINE-NOx field relative to EDGAR: the 10 % of cells with the highest EDGAR emissions hold 35 % of the total in PINE-NOx and 59 % in EDGAR, and the half of the cells with the lowest EDGAR emissions hold 16 % and 6 %; the log-space correlation with EDGAR is 0.937. We no longer present the size of this change as an observed redistribution, for the following reason. In a synthetic test with known emissions (Sect. 3.3, Text S1, Table S2), WRF-Chem columns simulated from EDGAR for July and January were sampled with the daily TROPOMI valid-pixel mask and inverted with the same networks; these runs use a simpler WRF-Chem configuration than those of Sect. 3.2 (NOx-only emissions, no nudging; Text S1). Although the true emissions equal EDGAR in this test, the share of the 10 % highest-emitting cells fell by 21 percentage points in July and 26 in January, against 20 and 25 for the observed columns on the same days. A change of this size therefore arises from the method itself, and the revised text describes it qualitatively. For the same reason, the 64 of 98 0.5° blocks in which the change exceeds twice its 1σ uncertainty are no longer used as evidence, because the budget does not contain this structural term. Fig. 8 shows the pattern.

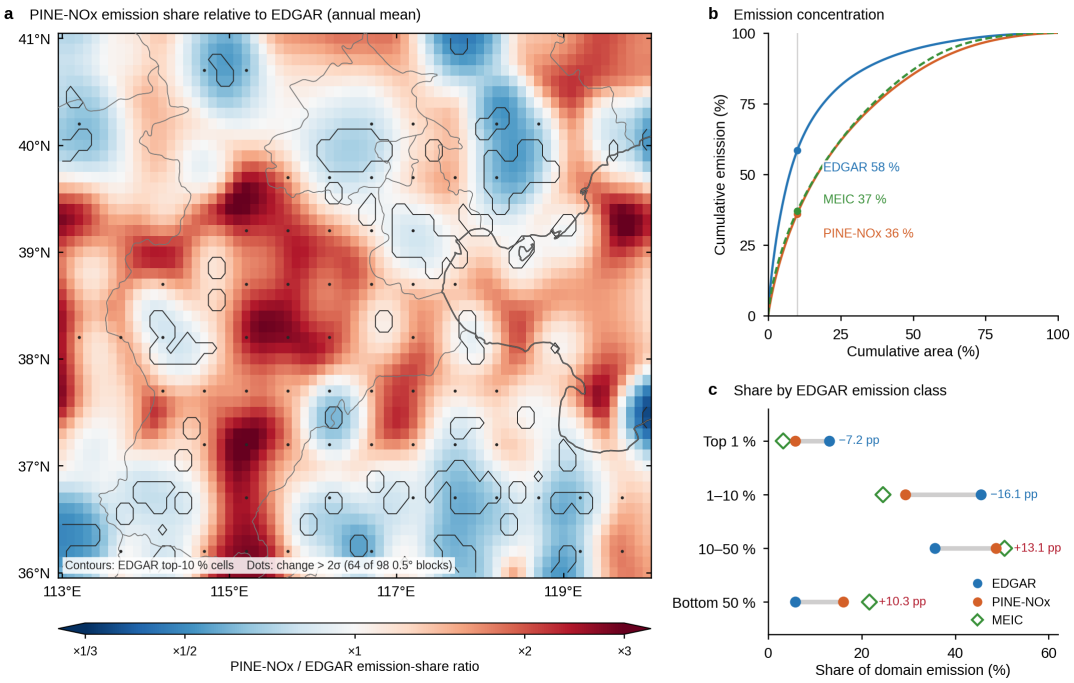


Figure 8. Annual (2023) PINE-NOx emission pattern relative to the EDGAR v8.1 prior. (a) Ratio of the PINE-NOx to the EDGAR emission share in each 0.1° cell (mean of 350 daily fields; both fields smoothed with a 2-cell Gaussian kernel for display only; colour scale ×1/3 to ×3, 0.4 % of cells outside the scale). Black contours enclose the 10 % of cells with the largest EDGAR emissions. Dots mark the 64 of 98 0.5° blocks (blocks whose 25 cells all lie in the mask of the uncertainty budget) whose change in emission share exceeds twice its 1σ uncertainty including the prior term (Fig. 9); this uncertainty does not include the structural term estimated with the synthetic test (Text S1). (b) Cumulative emission against cumulative area, cells ranked by emission, for EDGAR, PINE-NOx and MEIC 2023; values give the emission share of the 10 % of the area with the highest emission of each field (cells ranked by their own emission, unlike the EDGAR ranking used in the text and in (c)). (c) Share of domain emission held by four EDGAR rank classes in EDGAR, PINE-NOx and MEIC, with the PINE-NOx minus EDGAR change in percentage points. Statistics in (b) and (c) use unsmoothed fields.

On missing or misallocated sources: the changes take place under the fixed total, so an increase in one place is a decrease elsewhere, and, as explained under Comment 3, single low-prior cells cannot be resolved. What can be tested is whether the pattern moves towards an independent estimate. At 0.5°, 13 % of the emissions are placed in different grid cells than in MEIC for PINE-NOx, against 21 % for EDGAR. A control run fed with columns that the decoder simulates from EDGAR reaches 19 % (mean of three initialisations); for the production network and its matching control the observed columns lower the fraction by 5.1 percentage points (95 % interval 2.5 to 7.9; Sect. 3.3, Fig. 10). This control shares the decoder of the inversion and cannot contain its errors, so we also used the synthetic test with WRF-Chem columns described above. In July, 13.5 % of the emissions are placed differently from MEIC in PINE-NOx, against 18.5 % for the inversion of the synthetic columns and 20.5 % for EDGAR; in January the values are 17.1 %, 17.8 % and 21.9 %, so a gain from the observations is found in July but not clearly in January. This is consistent with misallocation in EDGAR at the 0.5° scale rather than missing sources, but MEIC has its own errors, and the comparison is not a validation (see Comment 5).

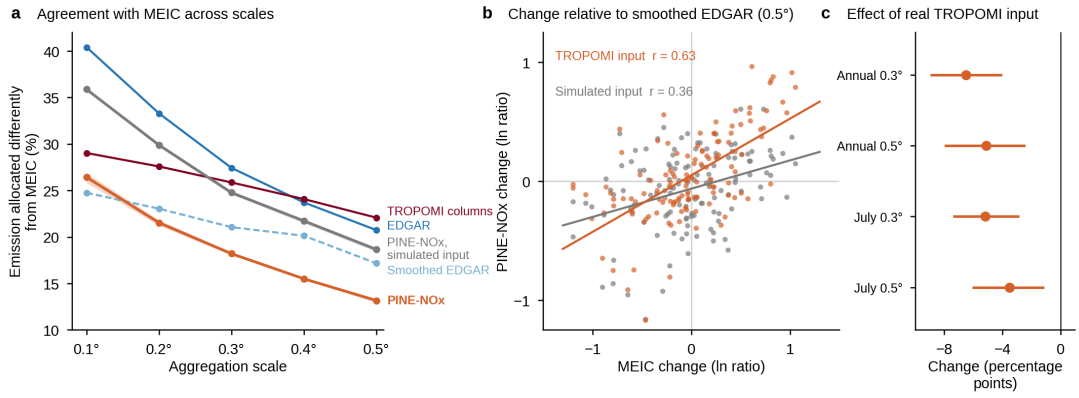


Figure 10. Agreement of the annual emission patterns with MEIC 2023, which is not used by PINE-NOx. (a) Fraction of the total emission placed in different grid cells than in MEIC, $S = \frac{1}{2} \Sigma|share_X - share_{MEIC}|$ (half the summed absolute difference of the emission shares, i.e. the total variation distance between the two maps; 0 % for identical allocation, 100 % for no overlap), for blocks of 0.1° to 0.5° (3452 to 127 blocks whose 0.1° cells all have MEIC emissions). Lines show EDGAR, EDGAR smoothed with a 2-cell Gaussian kernel, the TROPOMI column pattern treated as an emission field, PINE-NOx (four network initialisations, band = range) and the decoder control, PINE-NOx with simulated input (three initialisations). The decoder control uses the identical pipeline with the TROPOMI column replaced on every day by the column that the PINE decoder computes from EDGAR (same day's wind, valid mask and background, scaled to that day's observed column level, with noise drawn from the retrieval precision); it therefore carries no TROPOMI information on the spatial pattern of emissions. (b) Change relative to smoothed EDGAR in each 0.5° block, $\ln(X/EDGAR_s)$, for PINE-NOx with TROPOMI input and with simulated input against the corresponding MEIC change, with least-squares lines and Pearson r. (c) Paired difference in this fraction between TROPOMI input and simulated input (production network and its matching decoder control) for the annual and July fields at 0.3° and 0.5°, with 95 % block-bootstrap intervals. A stricter control with WRF-Chem columns, for July and January, is given in Table S2.

Changes in the manuscript:

Sect. 3.3: "A synthetic test with known emissions shows that a change of this size also arises from the method itself (Text S1, Table S2)."

Sect. 3.3: "In the PINE-NOx field, the 10 % of grid cells with the highest EDGAR emissions hold 35 % of the domain total, against 59 % in EDGAR, and the half of the cells with the lowest EDGAR emissions hold 16 %, against 6 %."

Sect. 3.3: "For the same reason, the 64 of the 98 0.5° blocks in which the change exceeds twice its 1σ uncertainty (dots in Fig. 8a; blocks whose 25 cells all lie in the mask of the uncertainty budget) are not evidence of an observed change, because the budget below does not contain this structural term."

Sect. 3.3: "The decoder control, i.e. the same inversion fed with columns that the PINE decoder simulates from EDGAR (simulated input), which contain the method's own smoothing but no observed pattern, reaches 19 % (mean of three initialisations, against 13 % for four PINE-NOx initialisations)."

Sect. 3.3: "For the annual fields of the production network and its matching decoder control, the observed TROPOMI columns lower the fraction by 5.1 percentage points at 0.5° (95 % interval 2.5–7.9) and by 6.5 points at 0.3° (4.1–8.9; Fig. 10c)."

Sect. 3.3: "In July, 13.5 % of the emissions are placed differently from MEIC in PINE-NOx, against 18.5 % for the inversion of the synthetic columns and 20.5 % for EDGAR. In January the values are 17.1 %, 17.8 % and 21.9 %, so no clear gain from the observations is found in that month."

The class shares behind Fig. 8c are listed here for convenience (share of the domain total, %). They describe the PINE-NOx field; as explained above, changes of this size from EDGAR also arise in the synthetic test.

EDGAR rank class (by emission)	EDGAR	PINE-NOx	MEIC
Top 1 % of cells	13.1	5.8	3.2
1 % to 10 %	45.5	29.4	24.6
10 % to 50 %	35.6	48.7	50.6
Bottom 50 %	5.8	16.1	21.6

Comment 5

The manuscript should more carefully distinguish between internal reconstruction skill and external model-consistency evaluation. For example, the Abstract states that "two independent checks corroborate the inversion," and Section 3.1 refers to external checks provided by the independent WRF-Chem simulation. However, the forward reconstruction skill is an internal consistency metric because it uses the same differentiable decoder and TROPOMI NO2 observations as in training. The WRF-Chem experiment is independent in terms of the transport-chemistry model, but it is still evaluated against the same TROPOMI NO2 observations used in the inversion. Therefore, it should be described as an external CTM-based consistency check rather than fully independent validation of the emissions. The Abstract, Section 3.1, and Conclusions should be revised accordingly to avoid overclaiming validation independence.

If the authors wish to claim independent emission validation, they should consider using observational datasets that do not enter the inversion, such as CEMS measurements, ground-based NO2 observations, aircraft/mobile measurements, or facility-level emission records. Relevant examples include, but are not limited to: https://doi.org/10.1021/acs.est.9b04488; https://doi.org/10.5194/acp-22-10875-2022; https://doi.org/10.5194/acp-23-8001-2023; https://doi.org/10.1109/TGRS.2025.3620116; https://doi.org/10.5194/acp-26-4405-2026

We agree, and the revised paper keeps the three kinds of evidence apart, as in the table.

Evidence	Uses the inversion's data or decoder?	What it can show	Where
Observation reconstruction (r = 0.856 against 0.332 for EDGAR on the validation days)	same decoder, same TROPOMI columns	internal self-consistency only; not a validation of the emissions	Sect. 3.1, Figs. 4 to 6
WRF-Chem columns (log-space r = 0.925 against 0.876; +0.049, 95 % interval 0.036 to 0.061)	independent transport model, same TROPOMI columns	consistency with the fitted observations in a different model	Sect. 3.2, Fig. 7

CNEMC surface NO2, 160 stations	not used in the inversion	all hours: +0.002 (−0.107 to 0.099); 13 to 15 CST: −0.129 (−0.187 to −0.036); no improvement over EDGAR	Sect. 3.2, Fig. 7
MEIC inventory	not used in the inversion	consistency with an independent bottom-up estimate; an inventory, not an observation	Sect. 3.3, Fig. 10

- Observation reconstruction (Sect. 3.1). The forward skill uses the same decoder and the same TROPOMI columns as the training and is described as "an internal self-consistency metric of observation reconstruction, not a validation of the emissions". Sect. 3.1 is retitled "Encoder architecture comparison and observation reconstruction".
- CTM-based consistency (Sect. 3.2, retitled "WRF-Chem consistency check"). The WRF-Chem column comparison uses an independent transport model but the same TROPOMI columns, and the text now says so in the referee's terms.
- Tests with data that do not enter the inversion. Surface NO2 at 160 CNEMC stations is the only observational one. It does not show the PINE-NOx pattern to be better: across stations the correlation is the same for all hours and lower in the 13 to 15 CST window, with all runs underestimating afternoon surface NO2 by 44 to 52 %. Compared directly with the emissions in the station cells, without a transport model, the station NO2 also correlates slightly better with EDGAR than with PINE-NOx (log-space r = 0.365 against 0.306; Fig. S6), and Sect. 3.2 now states this. We report these results as they are, in the abstract, Sect. 3.2, Sect. 4.3 and the Conclusions. The comparison with the MEIC inventory (Comment 4) is independent of the inversion but is an inventory, not an observation, and is described as a test of consistency with an independent bottom-up estimate.

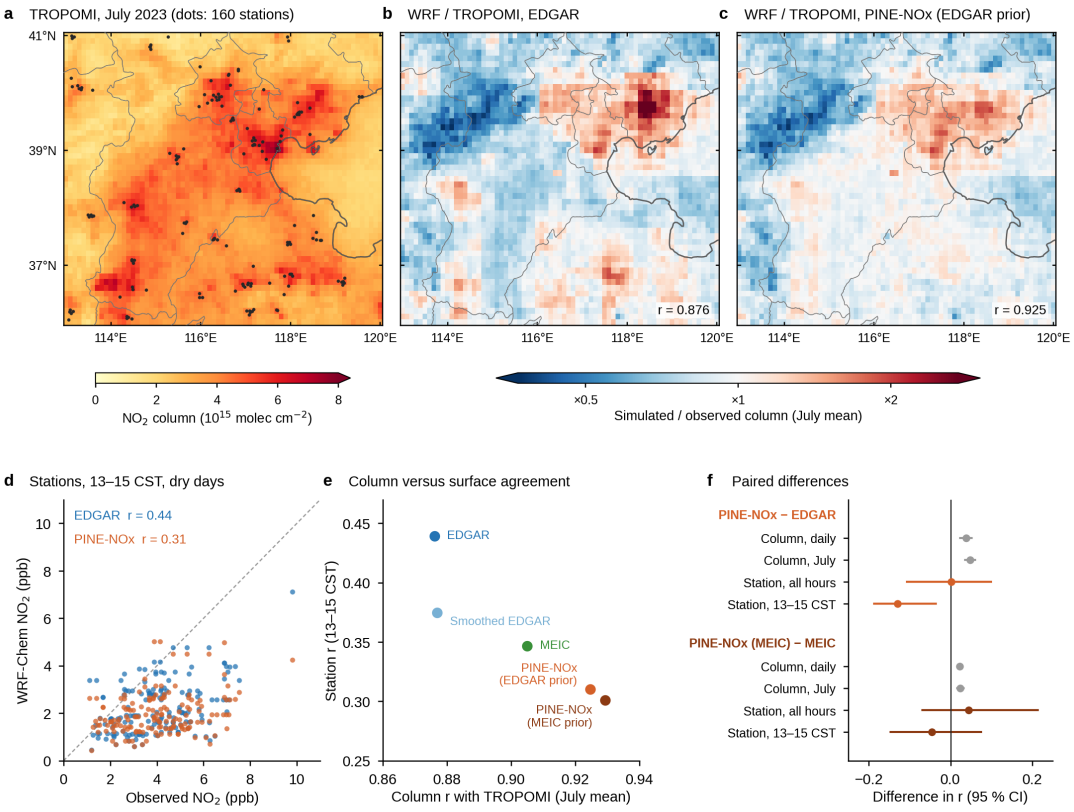


Figure 7. WRF-Chem consistency check, 8–31 July 2023. All runs use one WRF-Chem configuration with equal in-domain NOx totals; only the spatial pattern of NOx emissions differs. (a) TROPOMI July-mean NO_2 column at overpass; dots mark the 160 CNEMC stations. (b, c) Ratio of the simulated (averaging kernels applied) to the observed July-mean column for EDGAR and for PINE-NOx with the EDGAR prior, with the log-space spatial r. (d) Simulated against observed surface NO_2 at the stations, 13–15 CST means on 15 dry days, with Pearson r (one station per run above 11 ppb not shown). (e) Column r with TROPOMI against station r for EDGAR, EDGAR smoothed with a 2-cell Gaussian kernel, MEIC and PINE-NOx with the EDGAR and the MEIC prior. (f) Paired differences in r with 95 % intervals (column: bootstrap over days; stations: 0.5° spatial-block bootstrap, 2000 draws). The column comparison is not independent of TROPOMI, which PINE-NOx was fitted to; the station comparison is.

"Two independent checks corroborate the inversion" has been removed from the abstract. We have not been able to use CEMS, aircraft or flux data in this revision; Sect. 3.2 and Sect. 4.3 now state that facility-level or flux observations would be needed to settle which allocation is closer to the truth. Of the suggested references, Goldberg et al. (2019, 2022) were already cited; Goldberg et al. (2022) is cited at this point in Sect. 3.2 as an example of the kind of independent evaluation that is still missing.

The WRF-Chem configuration is now documented in full (Table S1, below).

Table S1. Configuration of the WRF-Chem runs in Sect. 3.2 and Fig. 7. All runs share this configuration; they differ only in the NOx emissions inside the study domain. The runs of the synthetic test use a simpler configuration, given in Text S1.

Item	Setting
Model	WRF-Chem v4.1.2; the ISORROPIA aerosol routine patched to exit where it would otherwise loop without end (other cases unchanged)
Domain and grid	One domain, Lambert conformal, 101 × 81 points, 27 km; 35 vertical levels, model top 50 hPa

Period	Initialised on 5 July 2023 00 UTC (chemistry from CAMS EAC4) and run to 31 July 2023 18 UTC (about 27 days); first 3 days discarded as spin-up, analysis 8–31 July (24 days)
Meteorology	NCEP GDAS/FNL analyses, 6-hourly; grid nudging of wind, temperature and humidity (3×10^{-4} s^{-1}) above the boundary layer
Physics	WSM3 microphysics; RRTM long-wave and Dudhia short-wave radiation; revised MM5 surface layer; Noah land surface; YSU boundary layer; Kain–Fritsch cumulus
Chemistry	RADM2 gas phase with MADE/SORGAM aerosol; Madronich (TUV) photolysis; Wesely dry deposition; no wet scavenging, cloud chemistry, N_2O_5 heterogeneous chemistry or convective transport of tracers
Chemical initial and boundary conditions	CAMS EAC4 reanalysis for O_3, CO, NO_2, SO_2, HCHO, HNO_3, C_2H_6, C_3H_8 and isoprene; other species from the WRF-Chem default profiles
Anthropogenic emissions	MEIC 2023 monthly emissions of all species (0.25°, five sectors); EDGAR v8.1 NOx outside the MEIC coverage; NOx emitted as 90 % NO and 10 % NO_2 (molar)
NOx inside the study domain	Replaced in the 425 model cells inside the 0.1° study domain by the July-mean pattern of the tested field (EDGAR, EDGAR smoothed with a 2-cell Gaussian kernel, MEIC, PINE-NOx with the EDGAR prior, PINE-NOx with the MEIC prior), scaled to the same NOx mass as MEIC in July; all other species unchanged
Temporal and vertical profiles	EDGAR hourly and day-of-week profiles for China, applied by sector; sector-specific vertical profiles over the lowest 8 model layers
Biogenic emissions	MEGAN v2.04, online (includes soil NO)
Not included	Lightning NOx, open fires, dust and sea salt
Comparison with TROPOMI	Each TROPOMI pixel (qa_value ≥ 0.75, cloud radiance fraction < 0.5) matched to the nearest model column, interpolated in time to the pixel; tropospheric averaging kernel applied; pixels averaged to the 0.1° grid per day
Comparison with stations	Hourly NO_2 in the lowest model layer of the cell containing each of the 160 CNEMC stations

Changes in the manuscript:

Sect. 3.1, heading: "3.1 Encoder architecture comparison and observation reconstruction"

Sect. 3.2, heading: "3.2 WRF-Chem consistency check"

Abstract: "In a one-month WRF-Chem simulation, the retrieved emissions bring simulated columns closer to TROPOMI but do not improve agreement with surface NO_2 stations."

Sect. 3.1: "We also note that the forward skill above is computed with the same differentiable decoder used in training and is therefore an internal self-consistency metric of observation reconstruction, not a validation of the emissions; consistency checks outside the training decoder use the WRF-Chem simulation in Sect. 3.2 and the MEIC inventory in Sect. 3.3."

Sect. 3.2: "The WRF-Chem experiment is an external CTM-based consistency check: it is independent of the inversion decoder but not of the observations, because the constraint reference used by WRF-Chem (TROPOMI NO_2 VCD) is the same source as the observations used in the PINE-NOx inversion, so the column comparison shows consistency with the observations the inversion was fitted to, within a different transport model; only the station comparison is independent, and in this one month and one model configuration it does not show the PINE-NOx pattern to be better than EDGAR."

Sect. 3.2: "An independent emission validation still depends on non-satellite observations such as aircraft and ground-based flux measurements, which we leave for future work (Cooper et al., 2017; Goldberg et al., 2022)."

Sect. 5, Conclusions: "In a WRF-Chem consistency check the new pattern brings the simulated columns closer to TROPOMI but does not improve the agreement with surface stations, so the allocation it proposes still needs an independent test."

Sect. 4.3: "The comparisons with independent data leave the evidence uneven: the PINE-NOx pattern is closer to MEIC (clearly in July and not in January when compared with the synthetic test), but in one month of WRF-Chem simulation it does not improve agreement with surface stations. Both MEIC and the station comparison have their own errors, and a test against facility-level or flux observations would be needed to settle which allocation is closer to the truth."

On the processor version, Sect. 2.1 now reports a sensitivity test and Table 1 lists the three versions (this correction was made while we checked every entry of Table 1 against the files used). Omitting the 33 days processed with v2.6.0 changes the annual pattern by 1.1 % of the emissions, and the December pattern, which is entirely v2.6.0, differs from the January pattern by 13.1 %, close to the 11.3 % between February and January. These tests do not show an effect of the version change on the spatial pattern in Sect. 3.3; they cannot exclude a small one.

Table 1, TROPOMI NO_2: "Sentinel-5P L2 OFFL v2.4.0 (1 January–11 March), v2.5.0 (12 March–25 November) and v2.6.0 (26 November–31 December); QC: qa_value ≥ 0.75, cloud radiance fraction < 0.5; pixels averaged in the grid cell that contains their centre"

Sect. 2.1: "Omitting the 33 retrieval days of version 2.6.0 changes the annual pattern by 1.1 % of the emissions. All December days are from version 2.6.0; the December pattern differs from the January pattern by 13.1 %, close to the 11.3 % between February and January."

Comment 6

In Table 1 ("Magnitude treatment"), Section 2.4 (lines 207-220), and Section 2.5 (lines 230-244): Since the rendering loss removes the domain-wide mean, it is unclear how domain-mean daily emission variability is constrained by observations. Please clarify how this domain-mean temporal variability is constrained under a demeaned loss formulation. Is it derived from TROPOMI spatial-pattern changes, meteorological inputs, prior regularization, annual-total normalization, or the neural-network architecture?

It is not constrained by the observations. Multiplying all emissions of a day by a constant shifts the log column by a constant, which the demeaned loss removes, so the loss is unchanged. In the network, the day-to-day domain mean is whatever the encoder outputs, given its inputs (the column pattern, the prior, the winds and the day-of-year encoding) and its architecture, held near the prior only by the Huber term on δ. The annual normalization is one constant for the whole year and does not act on individual days.

We therefore no longer report any domain-level variability. The day-to-day coefficient of variation (29 %) has been withdrawn, and the inferred winter-to-summer emission ratio of the submitted version is no longer reported as a result: Sect. 3.3 keeps the decomposition of the column seasonality only as a diagnostic, which shows that the implied ratio ranges from 1.1 to 1.8

for a ±25 % change of the winter lifetime. Sect. 2.4 (point 5) now states that "the domain-wide emission level is not constrained by the observations at all", Sect. 3.3 explains that the decoder does not constrain the seasonal amplitude of the emissions, and the monthly totals of the product follow the EDGAR monthly profile (Sect. 2.5, Fig. 11b). Seasonal information is reported only as changes of the spatial pattern, and Sect. 3.3 adds that the month-to-month changes of the pattern could not be separated from those the method produces on its own, in a test whose pass criterion was fixed before it was run. The decomposition of the column seasonality into lifetime and emission parts, which replaces the earlier winter-to-summer emission claim, is shown in Fig. S12.

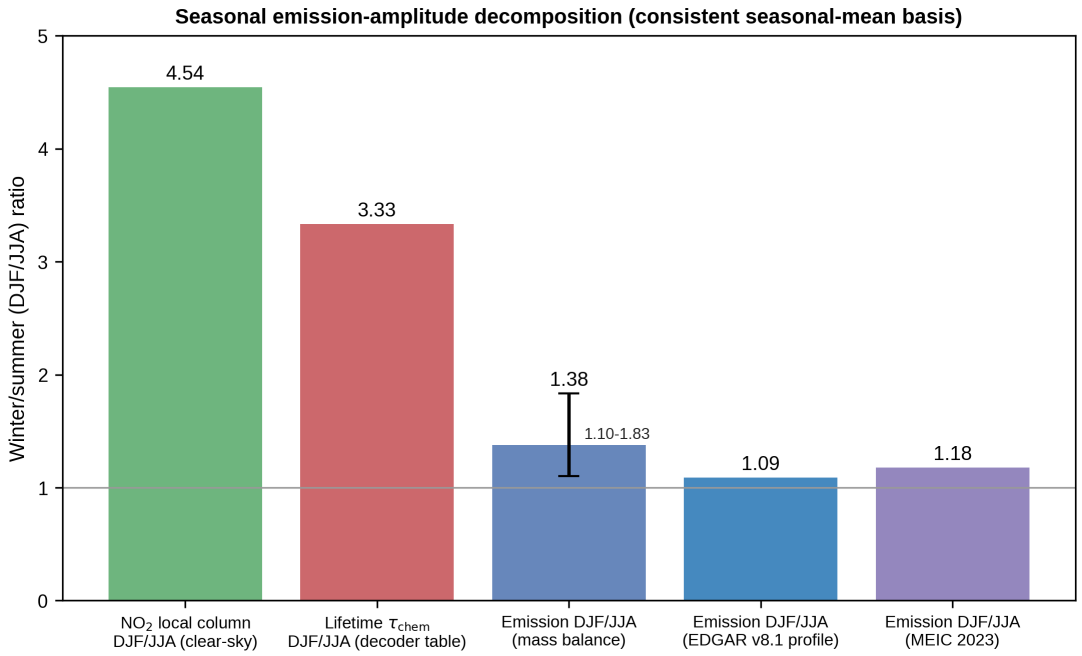


Figure S12. Decomposition of the seasonal amplitude of emissions under a consistent seasonal-mean convention. The clear-sky local NO_2 column winter/summer ratio (DJF/JJA seasonal means) is about 4.5 and the chemical lifetime τ_{chem} of the decoder table winter/summer ratio is about 3.3; dividing the two would imply an emission winter/summer ratio of about 1.4 (interval 1.10 to 1.83 allowing for a ±25% τ uncertainty), too uncertain to be reported as a result, to be compared with 1.09 for the EDGAR v8.1 2022 monthly profile used for the product totals and 1.18 for MEIC 2023. Note that the ratio of the peak month to the trough month is larger than the seasonal-mean ratio (6.5 for the local column, February/August); the two conventions differ and must not be conflated.

Changes in the manuscript:

Sect. 2.4, point (5): "The same invariance means that the domain-wide emission level is not constrained by the observations at all: multiplying all emissions by a constant leaves the loss unchanged, so the level is set only by the prior term and, in the final product, by the normalization of Sect. 2.5."

Table 1, Magnitude treatment: "Rendering loss demeaned per sample over the domain; absolute magnitude provided by annual-total normalization"

Sect. 3.3: "The column therefore constrains the seasonal amplitude of the emissions only loosely, and the decoder, whose loss removes the domain mean every day, does not constrain it at all. The monthly totals of the product are therefore taken from the EDGAR monthly profile (winter/summer 1.09; MEIC 1.18), and the monthly fields are used only for their spatial pattern."

Sect. 3.3: "The month-to-month changes of the pattern, however, could not be distinguished from those produced by the method: in a test whose pass criterion was fixed before it was run, the departures of each monthly pattern from the annual pattern correlated with the corresponding MEIC departures no better for PINE-NOx than for the decoder control (pooled over months at 0.5°, r = −0.08 to −0.05 against −0.07 to −0.06), so the monthly fields are best read as monthly realisations of the annual pattern."

Sect. 4.3: "The seasonal amplitude and the day-to-day variation of the domain total are not constrained either, and the month-to-month changes of the pattern could not be separated from those produced by the method itself (Sect. 3.3)."

Minor comments

Comment 7

In Section 3.3 and the caption of Figure 11, the interpretation of the spatial redistribution seems appear inconsistent. The text describes a "downward adjustment in pollution cores and an upward adjustment in clean peripheral regions", whereas the caption states the opposite "….. enhanced emissions in the pollution cores (red) and reduced emissions in the clean periphery (blue)". Please check Figure 11(c) and revise the text or caption accordingly.

Thank you; the caption was wrong. The data show lower emission shares in the pollution cores and higher shares in the surroundings. The figure has been replaced (now Fig. 8, shown under Comment 4), its caption describes the panels without interpretation, and Sect. 3.3 states the direction consistently with the figure, and notes that a change of this kind also arises in the synthetic test (Comment 4).

Changes in the manuscript:

Sect. 3.3: "Relative to EDGAR, PINE-NOx places less of the emissions in the urban–industrial cores and more in their surroundings (Fig. 8)."

Fig. 8 caption: "(a) Ratio of the PINE-NOx to the EDGAR emission share in each 0.1° cell (mean of 350 daily fields; both fields smoothed with a 2-cell Gaussian kernel for display only; colour scale ×1/3 to ×3, 0.4 % of cells outside the scale)."

Comment 8

In Figure 11(c), the relative difference appears to be saturated over large areas, especially in red, under the current 60% color scale. Please justify the choice of this range or provide an additional unsaturated or percentile-based version to better show spatial variability.

Fig. 8a now shows the ratio of the PINE-NOx to the EDGAR emission share on a logarithmic scale from ×1/3 to ×3; 0.4 % of the cells fall outside it, as stated in the caption (Fig. 8, shown under Comment 4). Statistics are computed from the unsmoothed fields.

Changes in the manuscript:

Fig. 8 caption: "Statistics in (b) and (c) use unsmoothed fields."

Response to Referee #2 (RC2)

Referee comments are in blue. Our replies are in black. Text quoted from the revised manuscript or Supplement is in red, with its location, and is copied verbatim from the tracked-changes files. Figures and tables of the revised manuscript are reproduced under the reply where they are first needed.

Manuscript: egusphere-2026-3233, "Physics-constrained inverse neural estimation (PINE) of NOx emission spatial patterns from TROPOMI NO2 columns over the North China Plain" (title revised)

Authors: Yinan Wang, Yubing Pan, Daren Lyu

We thank the referee for a thorough and constructive report. We agree with the main point: reconstructing the columns that the encoder receives is not evidence that the emissions are right. The revision separates the two, removes the daily-inventory framing, documents the WRF-Chem configuration and replaces the 12 to 15 % uncertainty with a budget built from separate sources, and adds a synthetic test with known emissions. Section, figure and table numbers refer to the revised manuscript; all changes are shown as tracked changes.

During the revision we found that the TROPOMI quality screening stated in Table 1 (qa_value ≥ 0.75, cloud radiance fraction < 0.5) had not been applied correctly when the columns were gridded. All data were re-gridded with the stated screening, all networks were retrained and all numbers were recomputed. Most values therefore differ from the discussion paper; the reconstruction r cited by the referee (0.837) is now 0.856 on the 54 validation days, against 0.332 for EDGAR (0.871 and 0.331 over all 347 days). These values are not directly comparable with the submitted 0.395 and 0.837, which were computed on 80 different validation days and on the raw column; the revised values use a new fixed split and the background-removed column (Sect. 3.1). The validation days were not used to fit the weights but were used to select the checkpoint (the training epoch kept), so they are not an independent test set (Sect. 2.3).

While checking Table 1 against the files used, we also found that the processor version was given as OFFL v2.4.0 for the whole year, whereas the 2023 record contains three versions. Table 1 and Sect. 2.1 now state the three versions with their dates and report what the version change does to the annual pattern.

We also added a synthetic test with known emissions (Sect. 3.3, new Text S1 and Table S2), which bears on Comments 1, 2 and 5. WRF-Chem columns simulated from EDGAR for July and January were sampled with the daily TROPOMI valid-pixel mask and inverted with the same networks, so that the true emissions equal the prior. The inversion then lowered the share of the urban–industrial cores by about as much as it does with the observed columns. We therefore describe the decrease relative to EDGAR qualitatively and no longer present its size as an observed redistribution. What we now report is the comparison with MEIC judged against this test (Comment 1).

The values below are the ones that changed and that this report refers to.

Quantity	Discussion paper	Revised paper	Reason
Reconstruction r, EDGAR prior	0.395	0.331	quality screening applied as stated; all 347 evaluated days

			instead of 80 validation days; skill computed on the background-removed column
Reconstruction r, PINE-NOx (UNet-CBAM)	0.837	0.856 (54 validation days; 0.871 over all 347 days)	quality screening applied as stated; networks retrained; new fixed validation days (54 instead of 80, also used to select the checkpoint); skill computed on the background-removed column
Overall uncertainty	12 to 15 %	per-cell budget, annual total 17.5 % at 0.5° and 38 % at 0.1° (prior included); lower bounds, structural error not included	transparent budget (Comment 5)
Day-to-day coefficient of variation; lag-1 autocorrelation	29 %; 0.249	withdrawn	domain level not constrained; daily fields are overpass-time fields (Comment 3)
Chemical lifetime, Table 1	summer 1.8 h, linear interpolation	monthly step table, 2 to 7 h	description corrected; the code was unchanged (Minor 3)
TROPOMI processor version, Table 1	OFFL v2.4.0	v2.4.0, v2.5.0, v2.6.0	description corrected from the files used

Major comments

Comment 1

The encoder receives the demeaned log TROPOMI VCD as an input and is trained to reconstruct that same VCD after passage through the decoder (Sect. 2.3). The reported held-out skill of r = 0.837 therefore measures how well an autoencoder reconstructs its own input on unseen days. Holding out days or seasons from weight optimization does not make the target independent, because the complete VCD field for each validation day is still supplied to the encoder at inference time. With millions of trainable parameters and only 284 training days, a high reconstruction correlation does not demonstrate that the latent emission field is correct or uniquely identified. The manuscript should clearly distinguish observation reconstruction, forward-model consistency, and emission validation.

We agree with every part of this comment. The revised manuscript uses the three categories the referee names.

- Observation reconstruction (Sect. 3.1, retitled "Encoder architecture comparison and observation reconstruction"). Sect. 3.1 now states that the forward skill "is computed with the same differentiable decoder used in training and is therefore an internal self-consistency metric of observation reconstruction, not a validation of the emissions". The season-withheld test (Fig. S7) is described in the same way (see Minor comment 2). On the 54 validation days the skill is 0.856, against 0.332 for EDGAR. The referee's point applies to this value in full, because the encoder still sees the columns of each validation day, and also because these days were used to select the checkpoint (the training epoch kept); Sect. 2.3 and Sect. 3.1 now state this. The 284 training days cited by the referee are those of the submitted split, a random hold-out of 80 days. With the screening applied, 350 days have at least 50 valid grid cells; days 6, 12, 18, 24 and 30 of each month (54 days) are used for validation and the other 296 for training (Table 1).
- Forward-model consistency (Sect. 3.2, "WRF-Chem consistency check"). The WRF-Chem column comparison uses an independent transport model but the same TROPOMI columns, and the text says so.
- Tests against data that do not enter the inversion. Surface NO_2 at 160 CNEMC stations, simulated with WRF-Chem, does not show the PINE-NOx pattern to be better than EDGAR (across-station r for all hours: difference +0.002, 95 % interval −0.107 to 0.099; 13 to 15 CST: −0.129, −0.187 to −0.036). Compared directly with the emissions in the station cells, without a transport model, the station NO2 also correlates slightly better with EDGAR than with PINE-NOx (log-space r = 0.365 against 0.306; Fig. S6), and Sect. 3.2 now states this. The MEIC inventory, which the inversion never uses, is closer to the PINE-NOx pattern than to EDGAR: the fraction of the total emission placed in different 0.5° grid cells than in MEIC is 13 % for PINE-NOx and 21 % for EDGAR (Sect. 3.3, Fig. 10). In the synthetic test described below, part of this gain also arises from the method; judged against that test, the gain is clear in July and not found in January. Because MEIC is an inventory with its own errors, we call this consistency with an independent bottom-up estimate, not validation.

Changes in the manuscript:

Sect. 3.1: "We also note that the forward skill above is computed with the same differentiable decoder used in training and is therefore an internal self-consistency metric of observation reconstruction, not a validation of the emissions; consistency checks outside the training decoder use the WRF-Chem simulation in Sect. 3.2 and the MEIC inventory in Sect. 3.3."

Sect. 3.2: "The WRF-Chem experiment is an external CTM-based consistency check: it is independent of the inversion decoder but not of the observations, because the constraint reference used by WRF-Chem (TROPOMI NO_2 VCD) is the same source as the observations used in the PINE-NOx inversion, so the column comparison shows consistency with the observations the inversion was fitted to, within a different transport model; only the station comparison is independent, and in this one month and one model configuration it does not show the PINE-NOx pattern to be better than EDGAR."

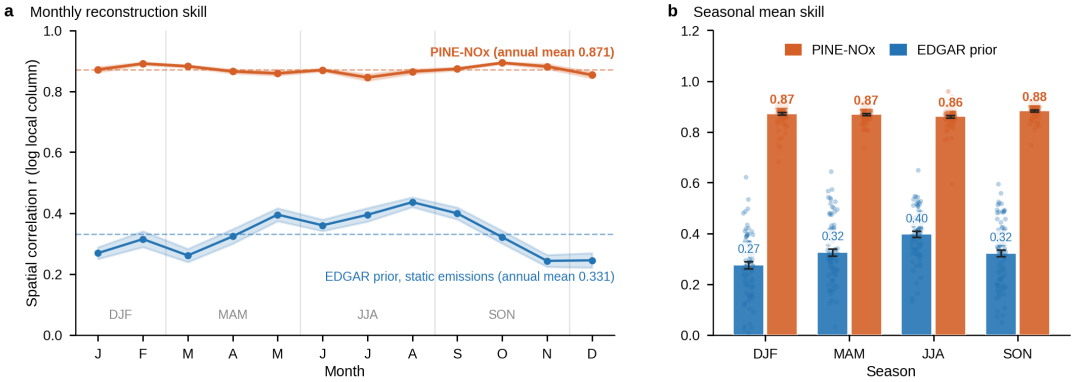


Figure 6. Forward-simulation skill of the inverted emissions over the full year (347 days). The skill r is defined as the log-space spatial Pearson correlation between the forward simulation and the background-removed TROPOMI NO_2 column, on cells above the boundary background. (a) Monthly-mean r for PINE-NOx (orange) and the EDGAR prior (blue); the shaded band is the ±1 standard error of the daily values within each month, and the dashed lines give the annual-mean skill. The EDGAR emissions are fixed; their r varies with the daily wind, the monthly lifetime and the observation sampling. (b) Season-aggregated mean skill; bars are seasonal means, whiskers are ±1 standard error, and dots are single-day values.

On unique identification: the emissions are not uniquely identified, and the paper now states which components the observations cannot identify. Sect. 2.4 explains that fine-scale structure and the domain level fall in the near null space of the decoder and are closed by the prior. Two tests quantify what this means. First, replacing the EDGAR prior by a spatially flat one leaves the reconstruction skill unchanged (0.909 against 0.871) but changes the emission pattern (log-space r = 0.83 between the two annual fields; Sect. 3.1, Fig. S9): good reconstruction does not pin down the emissions. Second, the uncertainty budget (Fig. 9) separates the spread between network initializations (1.8 % at 0.5°, emission-weighted 1σ of the log share) from the dependence on the prior inventory (14 %), so the product is conditional on the prior.

Changes in the manuscript:

Sect. 2.4, point (5): "The same invariance means that the domain-wide emission level is not constrained by the observations at all: multiplying all emissions by a constant leaves the loss unchanged, so the level is set only by the prior term and, in the final product, by the normalization of Sect. 2.5."

Sect. 2.4, point (4): "The second is the unobservable subspace, namely the fine-scale structure that falls into the approximate null space of Km together with the overall magnitude; this subspace is nearly indistinguishable in the observations and is instead closed by the prior: the total-variation prior suppresses non-physical high-frequency oscillations, the Huber prior softly anchors it to the spatial structure of EDGAR, and the overall absolute magnitude is finally calibrated by the single domain-wide scalar normalization of Sect. 2.5."

Sect. 2.4, closing paragraph: "This identifiability structure, namely spatial pattern from observations and physics and absolute magnitude from the prior constraint, is the theoretical basis on which PINE-NOx can improve the reconstruction of the observed column pattern (Sect. 3.1) without using any emission label, and it defines what the later comparisons can test: the spatial pattern, not the totals."

Sect. 3.1: "Because a uniform prior contains no spatial information, the reconstruction skill does not rest on the prior. The emission pattern itself does depend on the prior (log-space r = 0.83 between the annual fields obtained with the flat and with the EDGAR prior), and this dependence is carried as a separate term of the uncertainty budget (Sect. 3.3)."

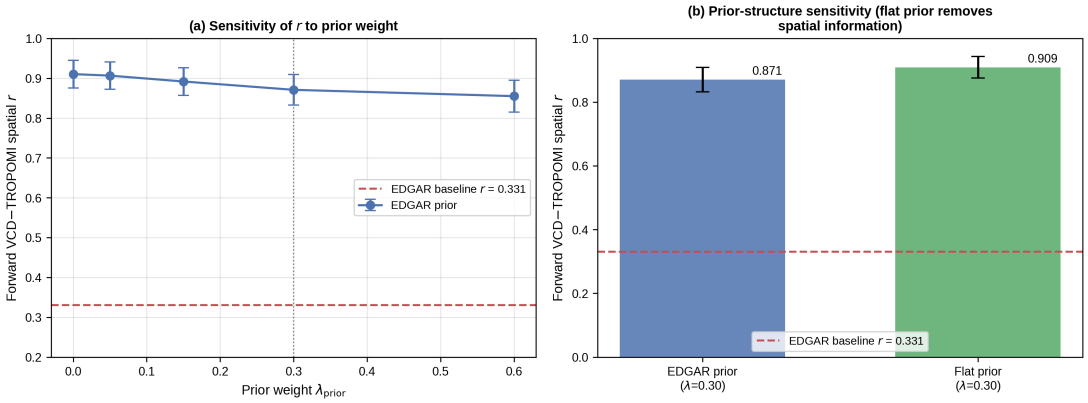


Figure S9. Sensitivity to the prior. (a) Reconstruction r against the prior weight λ_{prior} (0, 0.05, 0.15, 0.30 and 0.60; each a full retraining): 0.910, 0.907, 0.892, 0.871 and 0.855; the value used is 0.30. (b) Replacing the Huber reference prior from the real EDGAR spatial field with a uniform prior stripped of all spatial information (flat prior; domain-wide constant, equal total) and retraining within the same single-stage PINE-NOx framework: under the uniform prior the spatial correlation between the network's inverted forward VCD and TROPOMI is r = 0.909, against r = 0.871 with the EDGAR prior (EDGAR prior through the same decoder: r = 0.331). Points and bars are the mean and the standard deviation over 347 days. The reconstruction skill therefore does not rest on the prior; the emission pattern itself does depend on it (Sect. 3.1).

To test whether the spatial changes carry information from the observations rather than from the method alone, we used two controls. The first runs the identical pipeline on columns that the PINE decoder simulates from EDGAR. It reaches 19 % against MEIC (mean of three initialisations), compared with 13 % for the real inversion (mean of four); for the production network and its matching control the difference is 5.1 percentage points (95 % interval 2.5 to 7.9; Sect. 3.3, Fig. 10c). This control uses the inversion's own decoder and cannot contain the decoder's errors, which is the referee's concern under Comment 2. The second control is new: a synthetic test with known emissions, in which WRF-Chem columns simulated from EDGAR for July and January are sampled with the daily TROPOMI valid-pixel mask and inverted with the same networks (Text S1, Table S2). In this test the inversion moves emission share out of the

cores by about as much as with the observed columns (the share of the 10 % highest-emitting EDGAR cells falls by 21 and 26 percentage points in July and January, against 20 and 25 for the observations), so we no longer interpret the size of the change from EDGAR as an observed redistribution. What remains is the comparison with MEIC. In July, 13.5 % of the emissions are placed in different 0.5° cells than in MEIC in PINE-NOx, against 18.5 % in the synthetic test and 20.5 % for EDGAR, whereas in January the values (17.1 %, 17.8 % and 21.9 %) show no clear gain from the observations. The test has limits: it covers two months, its 27 km columns are smoother than TROPOMI, it simulates no air-mass-factor error, its WRF-Chem runs use a simpler configuration than those of Sect. 3.2 (NOx-only emissions, no nudging; Text S1), and MEIC is itself not the truth.

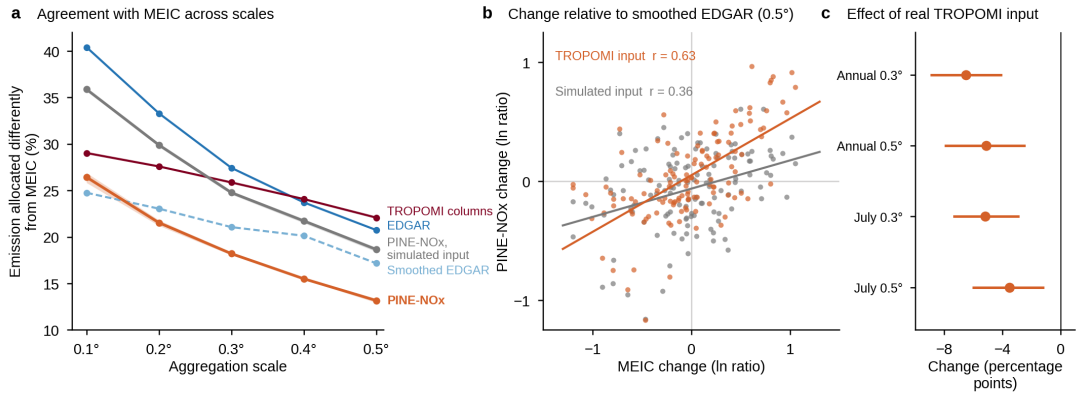


Figure 10. Agreement of the annual emission patterns with MEIC 2023, which is not used by PINE-NOx. (a) Fraction of the total emission placed in different grid cells than in MEIC, $S = \frac{1}{2} \Sigma|share_X - share_{MEIC}|$ (half the summed absolute difference of the emission shares, i.e. the total variation distance between the two maps; 0 % for identical allocation, 100 % for no overlap), for blocks of 0.1° to 0.5° (3452 to 127 blocks whose 0.1° cells all have MEIC emissions). Lines show EDGAR, EDGAR smoothed with a 2-cell Gaussian kernel, the TROPOMI column pattern treated as an emission field, PINE-NOx (four network initialisations, band = range) and the decoder control, PINE-NOx with simulated input (three initialisations). The decoder control uses the identical pipeline with the TROPOMI column replaced on every day by the column that the PINE decoder computes from EDGAR (same day's wind, valid mask and background, scaled to that day's observed column level, with noise drawn from the retrieval precision); it therefore carries no TROPOMI information on the spatial pattern of emissions. (b) Change relative to smoothed EDGAR in each 0.5° block, $\ln(X/EDGAR_s)$, for PINE-NOx with TROPOMI input and with simulated input against the corresponding MEIC change, with least-squares lines and Pearson r. (c) Paired difference in this fraction between TROPOMI input and simulated input (production network and its matching decoder control) for the annual and July fields at 0.3° and 0.5°, with 95 % block-bootstrap intervals. A stricter control with WRF-Chem columns, for July and January, is given in Table S2.

Changes in the manuscript:

Sect. 3.3: "The decoder control, i.e. the same inversion fed with columns that the PINE decoder simulates from EDGAR (simulated input), which contain the method's own smoothing but no observed pattern, reaches 19 % (mean of three initialisations, against 13 % for four PINE-NOx initialisations)."

Sect. 3.3: "For the annual fields of the production network and its matching decoder control, the observed TROPOMI columns lower the fraction by 5.1 percentage points at 0.5° (95 % interval 2.5–7.9) and by 6.5 points at 0.3° (4.1–8.9; Fig. 10c)."

Sect. 3.3: "In July, 13.5 % of the emissions are placed differently from MEIC in PINE-NOx, against 18.5 % for the inversion of the synthetic columns and 20.5 % for EDGAR. In January the values are 17.1 %, 17.8 % and 21.9 %, so no clear gain from the observations is found in that month."

Comment 2

The decoder starts from a zero column, integrates for about 4 h, uses daytime-mean 10 m winds, represents chemistry with a spatially uniform monthly lifetime, treats boundary inflow as a domain-wide scalar, and apparently converts a NOx emission flux directly into an NO2 column. This is not an "exact numerical solver" of NOx-NO2 transport and chemistry, and passing emissions through this operator does not by itself fully establish the physical reliability of the inferred emissions.

We agree, and "exact numerical solver" has been removed. Sect. 2.2 now describes the operator as a simplified transport-decay model and lists what it does not represent. The referee's reading of the conversion is correct: emissions enter as NO2-equivalent moles, so a uniform NO2/NOx ratio acts as a constant gain, which the demeaned loss removes; spatial variation of the ratio is not represented and is a structural error that we could not quantify separately. The Introduction now speaks of "a simplified numerical solver of transport and first-order decay", and the Table 3 entry now reads "Fixed, simplified two-dimensional transport–decay operator".

Changes in the manuscript:

Sect. 2.2: "The operator is thus a simplified transport–decay model: it uses one daytime-mean wind, a spatially uniform lifetime and a fixed NO_2/NOx conversion, and it does not resolve vertical mixing or nonlinear chemistry; the sensitivity of the results to the lifetime, the integration time and the wind level is part of the uncertainty budget (Sect. 3.3)."

Sect. 1: "We emphasize that our framework differs in mechanism from the classical PINN: a PINN approximates the solution of a partial differential equation (PDE) with a neural network and uses the PDE residual as a soft-constraint loss term, whereas our framework retains a simplified numerical solver of transport and first-order decay as a fixed, differentiable decoder and treats physics as a hard constraint, with the neural network performing only the amortized inverse mapping from observations to emissions (after training, a single forward pass suffices for inference, with no per-day re-optimization)."

Table 3, PINE-NOx column: "Fixed, simplified two-dimensional transport–decay operator"

Sect. 2.2: "We note that the decoder drives the transport of a single-overpass snapshot with the ERA5 daytime-mean 10 m wind during the satellite overpass window and does not explicitly represent intraday wind-direction changes; the sensitivity to the width of the background border (three and five cells instead of four) is part of the uncertainty budget (Sect. 3.3)."

Each simplification that we could vary is now part of the uncertainty budget (Sect. 3.3, Fig. 9), with the networks retrained for each variant: the lifetime (eight variants of the lifetime table (scaled by 0.7 and 1.3, without its seasonal cycle, with the winter months scaled by 0.67 and 1.5, with a seasonal shape following the O(1D) photolysis frequency with and without its latitude dependence, and with a spatial lifetime pattern diagnosed from WRF-Chem)), the integration time

(24 and 48 steps instead of 32), the wind level (10 m, 100 m and boundary-layer mean) and the background estimate (border widths of 3 and 5 cells instead of 4). For the annual pattern at 0.5°, these terms are 1.4, 1.4, 1.1 and 1.8 % (emission-weighted 1σ of the log share), small compared with the prior term (14 %). Sect. 3.3 states that structural errors of the decoder are only partly represented by these variants, so the budget values are lower bounds. The synthetic test of Comment 1 shows that these structural errors are not small: with the true emissions equal to EDGAR, the inversion moves emission share out of the cores by about as much as with the observed columns. As the referee notes, passing emissions through the operator does not establish their reliability; the evidence on reliability is the set of tests listed under Comment 1.

Sect. 3.3: "These values are lower bounds, since structural errors of the decoder are only partly represented by the variants, and the synthetic test above shows that the method alone can produce a change from EDGAR as large as the one retrieved from the observations. A narrower estimate, the spread between the four models of the season-withheld test, is shown for the daily fields in Fig. S10."

Sect. 4.3: "In addition, advection uses the ERA5 10 m daytime-mean wind to approximate column-mean transport, which may underestimate transport aloft; however, the advection ablation shows this to be a secondary contribution ($\Delta r \approx 0.09$; see Sect. 4.1), and across retrainings with the 10 m, the 100 m and the boundary-layer mean wind the annual emission shares differ by about 1 % at 0.5° and 3 % at 0.1° (1σ, emission-weighted; Fig. 9)."

Comment 3

The manuscript acknowledges the temporal sampling problem in Sect. 4.3, but still repeatedly describes the output as a daily emission inventory and interprets a day-to-day coefficient of variation of 29%. A 13:30 snapshot constrains an effective emission state over a limited preceding period, conditional on assumed lifetime and transport. It does not uniquely determine a 24 h mean emission flux or its diurnal profile. A lag-1 autocorrelation of 0.249 is not evidence that the remaining variability is genuine emission variability, because retrieval errors, clouds, sampling, and meteorology are themselves temporally correlated. The authors should either provide validation showing that the retrieved quantity represents daily mean emissions or rename and reinterpret it as an overpass-time effective emission estimate.

We accept the rename. The title no longer contains "daily"; the abstract and Sect. 4.3 describe the daily fields as overpass-time effective emissions, which should not be read as a 24 h mean emission rate. The products are annual and monthly patterns, and the daily fields are provided only as overpass-time fields. The coefficient of variation of 29 % and the lag-1 autocorrelation of 0.249 have been removed, together with their interpretation, for the reason the referee gives; in addition, the day-to-day domain mean is not constrained by the observations at all under the demeaned loss (Sect. 2.4, point 5; see also Referee 1, Comment 6).

Changes in the manuscript:

Title: "Physics-constrained inverse neural estimation (PINE) of NOx emission spatial patterns from TROPOMI NO2 columns over the North China Plain"

Abstract: "We retrieve annual, monthly and overpass-time daily emission patterns over the North China Plain at 0.1° for 2023, with the annual total normalized to the EDGAR inventory, so the observations constrain only the spatial pattern."

Sect. 4.3: "Sentinel-5P/TROPOMI provides only a single overpass per day (about 13:30 local solar time), so each retrieval is an instantaneous snapshot of the NO_2 column at a fixed hour, whereas the quantity we infer is an effective emission at overpass time. It should not be read as a 24 h mean emission rate: the diurnal cycle of emissions and of the chemical lifetime is not directly observed, and is instead absorbed into the effective quantities through the demeaned rendering loss and the monthly-mean lifetime parameterization."

Table 3, Temporal resolution: "Annual and monthly patterns; overpass-time daily fields"

For the referee's information, we also tested the daily patterns against the station data, with metrics fixed before the runs. A WRF-Chem run with the daily PINE-NOx patterns was compared with a run using the July mean of the same patterns.

Metric (daily patterns minus July-mean pattern)	**Difference**	**95 % interval**
Station-wise temporal correlation of daily means	−0.038	−0.11 to 0.05
Across-station correlation of daily means	+0.002	−0.04 to 0.04
Across-station correlation of 13 to 15 CST means	+0.046	0.014 to 0.084

The daily patterns also increased the root-mean-square error of daily means. This does not support the daily fields as daily mean emissions, consistent with the rename, and we have not added it to the paper.

Comment 4

The WRF-Chem experiment needs a complete reproducible configuration, including model version, domain and vertical grid, meteorological forcing and nudging, spin-up treatment, anthropogenic temporal profiles, biogenic and lightning emissions, boundary conditions, chemical initial conditions, model-to-satellite sampling, vertical interpolation, cloud screening, and the exact use of daily PINE emissions.

A new Table S1 in the Supplement gives the configuration of all runs in Sect. 3.2: model version (WRF-Chem 4.1.2), domain and vertical grid (27 km, 101 × 81 points, 35 levels, top 50 hPa), meteorological forcing (NCEP GDAS/FNL) and grid nudging above the boundary layer, spin-up (initialised on 5 July with chemistry from CAMS EAC4, first 3 days discarded), physics and chemistry options (RADM2 with MADE/SORGAM), chemical initial and boundary conditions (CAMS EAC4 for nine gases, WRF-Chem default profiles otherwise), anthropogenic emissions and their temporal and vertical profiles, biogenic emissions (MEGAN v2.04, including soil NO), the processes not included (lightning NOx, fires, dust, sea salt), and the sampling of the model for the comparison with TROPOMI (nearest model column interpolated to the pixel time, tropospheric averaging kernel applied, pixels averaged to the 0.1° grid) and with the stations

(lowest model layer in the station's cell). Every value was read from the namelists and input files of the runs. The table is reproduced here.

Table S1. Configuration of the WRF-Chem runs in Sect. 3.2 and Fig. 7. All runs share this configuration; they differ only in the NOx emissions inside the study domain. The runs of the synthetic test use a simpler configuration, given in Text S1.

Item	Setting
Model	WRF-Chem v4.1.2; the ISORROPIA aerosol routine patched to exit where it would otherwise loop without end (other cases unchanged)
Domain and grid	One domain, Lambert conformal, 101 × 81 points, 27 km; 35 vertical levels, model top 50 hPa
Period	Initialised on 5 July 2023 00 UTC (chemistry from CAMS EAC4) and run to 31 July 2023 18 UTC (about 27 days); first 3 days discarded as spin-up, analysis 8–31 July (24 days)
Meteorology	NCEP GDAS/FNL analyses, 6-hourly; grid nudging of wind, temperature and humidity (3×10^{-4} s^{-1}) above the boundary layer
Physics	WSM3 microphysics; RRTM long-wave and Dudhia short-wave radiation; revised MM5 surface layer; Noah land surface; YSU boundary layer; Kain–Fritsch cumulus
Chemistry	RADM2 gas phase with MADE/SORGAM aerosol; Madronich (TUV) photolysis; Wesely dry deposition; no wet scavenging, cloud chemistry, N_2O_5 heterogeneous chemistry or convective transport of tracers
Chemical initial and boundary conditions	CAMS EAC4 reanalysis for O_3, CO, NO_2, SO_2, HCHO, HNO_3, C_2H_6, C_3H_8 and isoprene; other species from the WRF-Chem default profiles
Anthropogenic emissions	MEIC 2023 monthly emissions of all species (0.25°, five sectors); EDGAR v8.1 NOx outside the MEIC coverage; NOx emitted as 90 % NO and 10 % NO_2 (molar)
NOx inside the study domain	Replaced in the 425 model cells inside the 0.1° study domain by the July-mean pattern of the tested field (EDGAR, EDGAR smoothed with a 2-cell Gaussian kernel, MEIC, PINE-NOx with the EDGAR prior, PINE-NOx with the MEIC prior), scaled to the same NOx mass as MEIC in July; all other species unchanged
Temporal and vertical profiles	EDGAR hourly and day-of-week profiles for China, applied by sector; sector-specific vertical profiles over the lowest 8 model layers
Biogenic emissions	MEGAN v2.04, online (includes soil NO)
Not included	Lightning NOx, open fires, dust and sea salt
Comparison with TROPOMI	Each TROPOMI pixel (qa_value ≥ 0.75, cloud radiance fraction < 0.5) matched to the nearest model column, interpolated in time to the pixel; tropospheric averaging kernel applied; pixels averaged to the 0.1° grid per day
Comparison with stations	Hourly NO_2 in the lowest model layer of the cell containing each of the 160 CNEMC stations

On the use of PINE-NOx emissions: inside the study domain, the NOx emissions of each run were replaced by the July-mean pattern of the tested field (EDGAR, smoothed EDGAR, MEIC, PINE-NOx with the EDGAR or the MEIC prior) and scaled to the same NOx mass, so that the runs differ only in the spatial pattern; the hourly and vertical profiles and all other species are the same in all runs (Sect. 3.2, Table S1). The run with daily PINE-NOx patterns mentioned under Comment 3 used the same set-up with the daily pattern of each day. The main-text results of these runs are in Fig. 7.

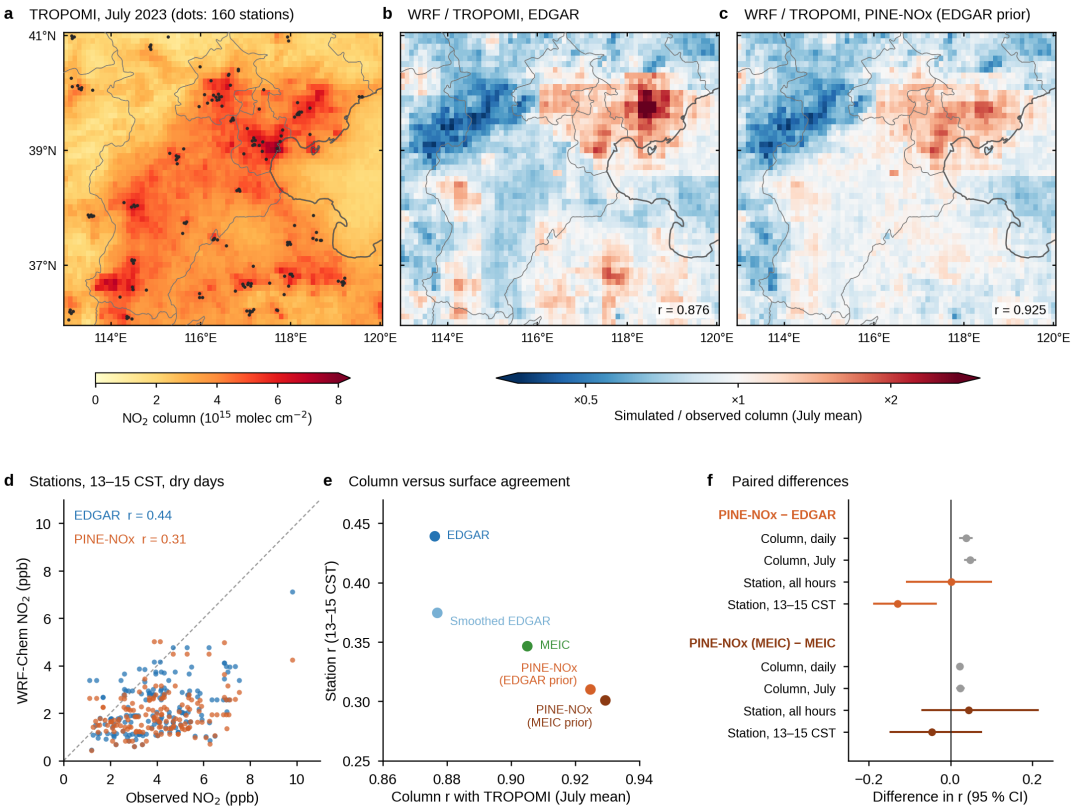


Figure 7. WRF-Chem consistency check, 8–31 July 2023. All runs use one WRF-Chem configuration with equal in-domain NOx totals; only the spatial pattern of NOx emissions differs. (a) TROPOMI July-mean NO_2 column at overpass; dots mark the 160 CNEMC stations. (b, c) Ratio of the simulated (averaging kernels applied) to the observed July-mean column for EDGAR and for PINE-NOx with the EDGAR prior, with the log-space spatial r. (d) Simulated against observed surface NO_2 at the stations, 13–15 CST means on 15 dry days, with Pearson r (one station per run above 11 ppb not shown). (e) Column r with TROPOMI against station r for EDGAR, EDGAR smoothed with a 2-cell Gaussian kernel, MEIC and PINE-NOx with the EDGAR and the MEIC prior. (f) Paired differences in r with 95 % intervals (column: bootstrap over days; stations: 0.5° spatial-block bootstrap, 2000 draws). The column comparison is not independent of TROPOMI, which PINE-NOx was fitted to; the station comparison is.

Changes in the manuscript:

Sect. 3.2: "The forward self-consistency evaluations above are all based on the same differentiable decoder used in the inversion; to provide an external reference independent of that decoder, we carried out a WRF-Chem three-dimensional forward-simulation experiment (5–31 July 2023, about 27 days simulated, of which the first 3 are spin-up and 8–31 July (24 days) are analysed; 27 km horizontal resolution, the RADM2 gas-phase chemical mechanism with the SORGAM aerosol module, chemical initial and boundary conditions from the CAMS EAC4 reanalysis; full configuration in Table S1), driven separately by the EDGAR prior and by the PINE-NOx-inverted emissions, and in addition by MEIC, by EDGAR smoothed with a 2-cell Gaussian kernel and by PINE-NOx obtained with the MEIC prior."

Sect. 3.2: "All runs share the same configuration and the same NOx total within the domain, so only the spatial pattern of the emissions differs (a comparable WRF-Chem–TROPOMI NO_2 cross-evaluation strategy is used by Poraicu et al., 2023)."

Sect. 3.2: "Taking the TROPOMI overpass time (about 13:30 local solar time) as the temporal reference, we computed the monthly-mean tropospheric NO_2 VCD over the NCP for 8–31 July 2023 and compared it with TROPOMI observations at the pixel scale, applying the TROPOMI averaging kernels (Fig. 7)."

Comment 5

The reported overall uncertainty of 12-15% is not derived transparently. Neural cross-fold spread is only one component and is not a substitute for uncertainty from the TROPOMI retrieval, sampling and clouds, background subtraction, transport winds, chemical lifetime, initial and boundary conditions, regularization strengths, EDGAR total, network initialization, and structural model error. A robust uncertainty assessment should propagate these sources to daily, monthly, grid-cell, and domain-total emissions.

We agree, and the 12 to 15 % has been removed. Sect. 3.3 and Fig. 9 now give a budget for the emission pattern, defined as the 1σ of the logarithm of each cell's emission share. Each source is estimated by retraining or re-running the inversion with the source varied, and the sources are combined in quadrature: network initialization (four seeds), regularization weights, chemical lifetime, integration time, wind level, background estimate, retrieval noise (perturbed by the TROPOMI precision), a ±30 % systematic error in the air mass factor, the sampling of days, and, separately, the prior inventory (EDGAR or MEIC as prior).

The budget for the annual pattern is listed by source below (emission-weighted 1σ of the log share, %). The combined values are emission-weighted means of the per-cell quadrature sums, so they are not the quadrature sum of the weighted means in the rows.

Source	0.5° blocks	0.1° cells
Network initialization (four seeds)	1.8	3.3
Regularization weights	3.1	5.9
Chemical lifetime	1.4	1.7
Integration time	1.4	3.4
Wind level	1.1	3.1
Background estimate	1.8	2.1
Retrieval noise	0.8	2.9
Air mass factor, ±30 % systematic	4.9	7.8
Sampling of days	3.9	5.3
Combined, method and data	8.5	14.7
Prior inventory	13.9	32.9
Combined, prior included	17.5	38.0

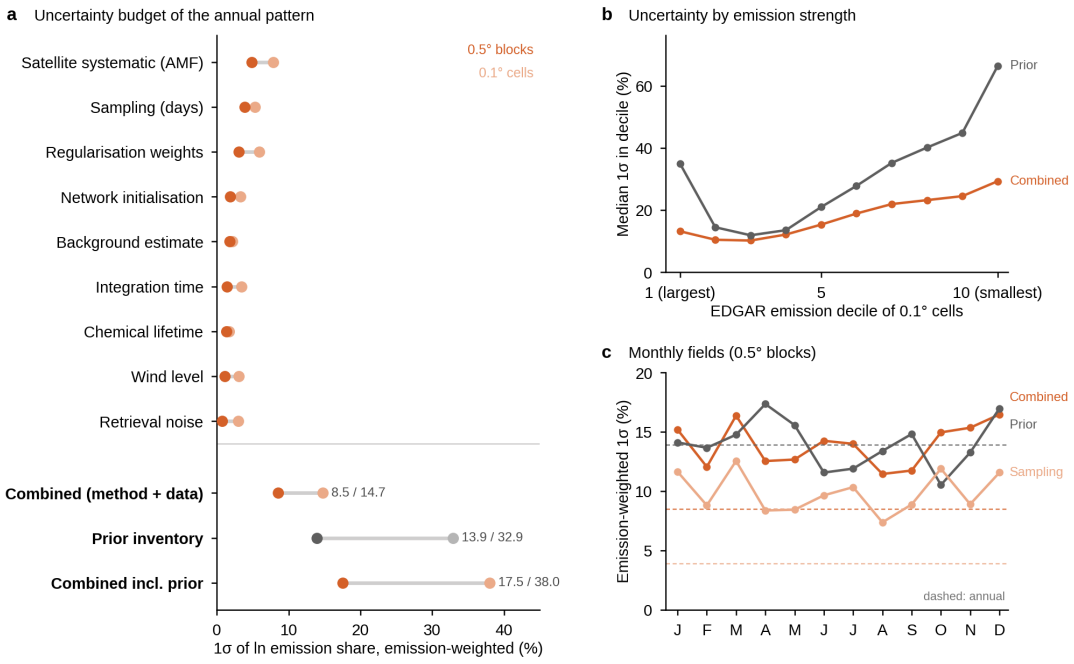


Figure 9. Uncertainty of the emission pattern, expressed as 1σ of the natural logarithm of the emission share. (a) Annual budget by source, emission-weighted over 0.5° blocks (dark) and 0.1° cells (light); combined values add the sources in quadrature in each cell and are then emission-weighted, so they differ from the quadrature sum of the weighted single-source values; the prior-inventory term (EDGAR or MEIC as prior) is shown separately. (b) Median 1σ of 0.1° cells in each EDGAR emission decile, for the combined method and data terms and for the prior term. (c) Emission-weighted 1σ of the monthly fields at 0.5° for the combined term, the prior term and day sampling; dashed lines give the annual values.

The budget is given at the scales the referee asks for, as far as the method allows. For the annual pattern, the method-and-data term is 8.5 % at 0.5° and 15 % at 0.1°; the prior adds 14 % and 33 %, giving at least 17.5 % and 38 % in total. The prior term grows towards weakly emitting cells (Fig. 9b). Monthly fields are 1.3 to 1.9 times as uncertain as the annual field (method-and-data term 11.5 to 16.5 % against 8.5 % at 0.5°; Fig. 9c). We do not give a daily uncertainty because the daily fields are no longer presented as a product (Comment 3). The domain total is set by the EDGAR normalization and carries the uncertainty of the inventory total; the budget makes no statement about it beyond the MEIC/EDGAR ratio of 0.88. Two items of the referee's list are covered only in part: initial and boundary conditions enter only through the background estimate, and the structural error of the decoder is represented only by the variants above. Sect. 3.3 therefore calls these values lower bounds. The synthetic test (Comment 1) shows that the structural term is large compared with this budget, so we no longer test the change from EDGAR in each block against it.

Changes in the manuscript:

Sect. 3.3: "The uncertainty of this pattern is summarised in Fig. 9 as the 1σ of the logarithm of each cell's emission share, combining in quadrature the spread from network initialisation, regularisation weights, chemical lifetime, integration time, wind level, background estimate, retrieval noise, a ±30 % systematic error in the satellite air mass factor and the sampling of days."

Sect. 3.3: "For the annual field this combined method-and-data term is 8.5 % at 0.5° and 15 % at 0.1° (emission-weighted). The choice of prior inventory adds a larger term (14 % and 33 %), so that the total is 17.5 % and 38 % (the combined values are emission-weighted means of the per-cell quadrature sums, so they differ from the quadrature sums of the weighted values given here)."

Sect. 3.3: "The prior term grows towards the weakly emitting cells, where the observations carry least information and where a mass-conserving inversion over China has found substantial emissions in cells with low prior values (Lu et al., 2025), and the monthly fields are 1.3 to 1.9 times as uncertain as the annual field (method-and-data term 11.5–16.5 % against 8.5 % at 0.5°), mainly because each month has fewer days."

Minor comments

Minor 1

The terms "physics-constrained", "hard constraint", "physical self-consistency", and "validated" are used too strongly.

We have narrowed these terms. "Validated" and "validation" are now used only in negations ("not an independent validation"), where we state what an independent validation would need, for the validation days of the data split and in "cross-validation" for the season-withheld test. "Physical self-consistency" has been replaced by "reconstruction" in the title of Sect. 3.1, in Sect. 2.4 and in the Conclusions. Where "forward self-consistency" remains (Sect. 3.1, Sect. 3.2 and the caption of Fig. 5), it refers to agreement with the training decoder, which Sect. 3.1 calls an internal metric and not a validation of the emissions. "Hard constraint" is kept only where it describes the structure of the method, i.e. that the decoder is fixed rather than learned (Sect. 1); the sentence attributing the seasonal robustness to it has been weakened (Sect. 3.1). "Physically interpretable" has been removed from the abstract, the Introduction, Sect. 4.2 and the Conclusions, and the Table 3 row "Physical interpretability" is now "Transport representation", with a description instead of a rating. We kept "physics-constrained" in the method name because it describes this structure: every emission field the encoder produces is evaluated through the fixed transport–decay operator. The revised text states what the operator does not represent (Sect. 2.2, Comment 2).

We also went through the text for other wording that claims more than the evidence shows. "Highly robust to the encoder architecture" now reads "insensitive to the encoder architecture", "a fair, circularity-free architecture evaluation" and "genuinely" have been removed, and "a new paradigm" has been replaced in the abstract and the Conclusions.

Changes in the manuscript:

Sect. 3.1, heading: "3.1 Encoder architecture comparison and observation reconstruction"

Sect. 3.1: "All four backbones far exceed the EDGAR prior and differ very little from one another (0.821–0.884), indicating that, under a unified random seed, training under the differentiable-physics constraint is insensitive to the encoder architecture (seasonally stratified cross-validation further supports the seasonal stability of this skill; Fig. S7)."

Sect. 5, Conclusions: "With label-free training and millisecond inference, PINE-NOx offers a practical approach to atmospheric emission inversion, provided that its departures from the prior are checked with a synthetic test of this kind."

Sect. 3.1: "This robustness is consistent with the constraint of the fixed physical decoder on the advection-chemistry processes and the seasonal chemical-lifetime parameterization, which allow the encoder to generalize without season-specific tuning."

Minor 2

Lines 286-293 infer "year-round extrapolation" from a seasonally balanced random split. Interleaved holdout is interpolation in time, not year-round extrapolation. Even the leave-one-season-out test remains reconstruction of the input VCD.

We agree. "Year-round extrapolation" has been removed (Sect. 3.1). Sect. 2.3 already noted that "the validation days are interleaved in time with the training days"; the hold-out result is therefore a test of interpolation in time, and the season-withheld test (Fig. S7) is described as a test of the reconstruction skill on a season absent from training (0.841 to 0.859), not as evidence on the emissions. On the 54 validation days the reconstruction skill is 0.856, against 0.874 on the training days and 0.871 over all 347 days, so the interleaved hold-out mainly tests interpolation in time, as the referee says.

Changes in the manuscript:

Sect. 2.3: "The validation days are interleaved in time with the training days, and the evaluation metric is the demeaned log-space correlation (characterizing the spatial pattern rather than pointwise magnitude), so the temporal autocorrelation between training and validation days has limited influence on the conclusions."

Sect. 3.1: "The seasonal means are 0.871 in winter (DJF), 0.869 in spring (MAM), 0.860 in summer (JJA) and 0.883 in autumn (SON), and when a whole season is withheld from training the skill on that season is still 0.841–0.859 (Fig. S7), so the reconstruction skill does not depend on a particular season."

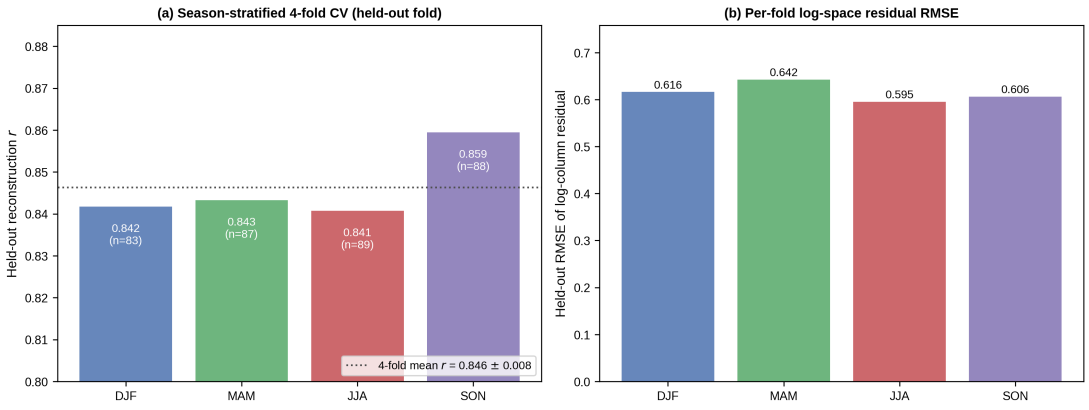


Figure S7. Seasonally stratified four-fold cross-validation (each fold holds out one complete season and retrains the same single-stage PINE-NOx model on the remaining three seasons without any emission label, then evaluates the forward-reconstruction skill on the held-out season that never participated in training). (a) The forward spatial correlation on the held-out season is 0.842 (DJF), 0.843 (MAM), 0.841 (JJA) and 0.859 (SON), with a mean of 0.846 ± 0.008, slightly below the full-set training baseline (r = 0.871). (b) Root-mean-square difference between the simulated

and the observed log local column, each demeaned per day, on the held-out season: 0.62 (DJF), 0.64 (MAM), 0.60 (JJA) and 0.61 (SON). The reconstruction skill therefore does not depend on any single season being present in the training set.

Minor 3

The definition and derivation of the monthly lifetime values should be given in the Methods, including the interpolation formula and supporting literature. The selection of E_eff/R = 3000 K also requires justification.

Sect. 2.2 now gives the table (7, 6, 5, 4, 3, 2, 2, 2, 4, 4, 5 and 7 h from January to December), states that it is spatially uniform and constant within each month (no interpolation between months), and compares it with the literature. We should be clear that the table was not fitted to these studies; it is consistent with them in magnitude and seasonal ratio. Because the values are uncertain, the lifetime is varied in the budget (Comment 2); the annual shares change by 1.4 % at 0.5° across the eight lifetime variants listed under Comment 2, and the share of the 10 % highest-emitting cells by at most 2 percentage points (Sect. 4.1).

Month	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
τ_{chem} (h)	7	6	5	4	3	2	2	2	4	4	5	7

Changes in the manuscript:

Sect. 2.2: "The effective NO_2 chemical lifetime τ_{chem} follows a monthly-mean parameterization based on the seasonal variation of OH (7, 6, 5, 4, 3, 2, 2, 2, 4, 4, 5 and 7 h from January to December, spatially uniform and constant within each month). These are daytime lifetimes near the overpass time."

Sect. 2.2: "They are of the same magnitude as satellite-derived lifetimes in China: 2.5 to 6.4 h for five North China Plain cities in May to September (Liu et al., 2016) and, at noon over Wuhan, 1.4 h in summer and 4.8 h in winter (Zhang et al., 2023). No monthly series has been published for the NCP, so the values for the transitional months were set by hand between the winter and summer values."

The 3000 K is used only in the dynamic-lifetime sensitivity test of Sect. 4.1, not in the inversion. We found no literature value for it, and it is not a physical activation energy: the rate coefficient of the main daytime sink, OH + NO2 + M, decreases slightly with increasing temperature. Sect. 4.1 now describes it as an empirical coefficient. The test changes Δr by about 1 % (Figs. S3 and S4).

Sect. 4.1: "Drawing further on observational evidence that the NOx lifetime varies with temperature and photochemical activity (Laughner and Cohen, 2019; Stavrakou et al., 2013), we introduce a pixel-level dynamic τ_{chem} driven by the ERA5 daily surface air temperature via the Arrhenius equation (E_{eff}/R = 3000 K, an empirical coefficient that varies τ_{chem} by about ±20 % for a ±5 K departure from the daily domain-mean temperature; it is not a physical activation energy, since the rate coefficient of the main daytime sink, OH + NO_2 + M, decreases slightly with increasing temperature; Seinfeld and Pandis, 2016), whose four-season spatial coefficient of variation is about 13–19% (Fig. S3); yet the improvement in Δr is only about 1 % (scalar τ: 0.535; dynamic τ: 0.542; Figs. S3–S4)."

Sect. 4.1: "Because these tests keep the emissions fixed, we also retrained the model with eight variants of the lifetime table (scaled by 0.7 and 1.3, without its seasonal cycle, with the winter months scaled by 0.67 and 1.5, with a seasonal shape following the O(1D) photolysis frequency with and without its latitude dependence, and with a spatial lifetime pattern diagnosed from WRF-Chem); across these variants the annual emission shares change by 1.4 % at 0.5° (1σ, emission-weighted), and the share of the 10 % highest-emitting cells changes by at most 2 percentage points, small compared with its difference of about 23 percentage points from EDGAR (58.6 % against 35.2 %; Sect. 3.3), a difference of the size that the method also produces in the synthetic test."

Minor 4

State whether Tg refers to Tg NO2, Tg NOx expressed as NO2, or Tg N, and use the same convention for EDGAR, MEIC, and PINE.

All NOx masses in the paper, for EDGAR, MEIC and PINE-NOx alike, are stated as NOx expressed as NO2 (NO2-equivalent mass); Sect. 2.1 now states this convention. The PINE-NOx total equals the EDGAR total by construction, and the decoder converts the mass to moles with the molar mass of NO2.

Changes in the manuscript:

Sect. 2.1: "The Multi-resolution Emission Inventory for China (MEIC; Zheng et al., 2018; Li et al., 2017) provides a bottom-up, sector-resolved monthly NOx inventory, regridded from 0.25° to the 0.1° study grid (domain annual total 2.67 Tg; NOx masses are given as NO_2-equivalent mass for all inventories in this paper); because it is constructed independently of TROPOMI and of EDGAR, it is used only to test whether the inverted spatial pattern moves towards or away from an independent bottom-up estimate (Sect. 3.3)."

Minor 5

Remove or revise the policy implication that annual inventories underestimate winter health-exposure risk. Emissions alone do not determine exposure, and the inferred winter enhancement is sensitive to the assumed chemical lifetime.

We agree. Sect. 4.4 (Implications for policy applications) has been deleted, together with the health-exposure statement. As noted under Comment 3 and in Referee 1, Comment 6, the revised paper also no longer reports a winter enhancement of the emissions as a result (Sect. 3.3 shows only that the implied winter-to-summer ratio ranges from 1.1 to 1.8 for a ±25 % change of the winter lifetime); Sect. 3.3 explains that the observed column constrains the seasonal amplitude only loosely and that the decoder does not constrain it.

Changes in the manuscript:

Sect. 3.3: "The column therefore constrains the seasonal amplitude of the emissions only loosely, and the decoder, whose loss removes the domain mean every day, does not constrain it at all."