Dear Dr. Western,

Thank you for handling our manuscript and for coordinating the review process. We are grateful for the time invested by both referees, and we have prepared a detailed point-by-point response addressing all comments raised.

In response to the reviews and your editorial guidance, we have substantially revised the manuscript. Most importantly, we reduced the prominence of the OSSE in the main manuscript and moved the detailed OSSE design and results to Supplement S2. The OSSE is now framed explicitly as an idealized information-content diagnostic under perfect-model assumptions and a large prescribed prior-truth difference, not as evidence of real-world robustness, operational flux accuracy, or grid-scale emission resolvability. We also added a 3-site reference OSSE under matched filter settings to provide context for the dense-network experiment and re-ran the 50-site OSSE under the updated localization setting.

We further clarified the timing and interpretation of the augmented state vector, corrected the OH-related CH_4 interpretation, and revised the Lagrangian–Eulerian discussion to avoid any implication of demonstrated superiority over footprint-based approaches. We also removed comparative wording in the Conclusions that was not directly supported by the experiments presented.

We have revised the manuscript to clarify the scope of our claims: the system is not presented as a perfect or universally robust inversion framework, nor as fully resolving grid-scale emissions under the present sparse network. Instead, we clarify that, under the stated assumptions and available observations, the framework can provide meaningful national-scale top-down constraints and diagnostics, as well as a reproducible basis for future regional GHG inversion applications.

We emphasize that the manuscript is intended as a GMD model-description paper presenting a reproducible WRF-Chem/DART regional dual-species data-assimilation framework, not as a comparative benchmarking study against other inversion classes. We now present Eulerian, Lagrangian, variational, analytical Bayesian, and sequential ensemble approaches as complementary methodologies with different assumptions, strengths, and trade-offs.

We have addressed all referee comments in the point-by-point response and respectfully request the Editor's independent assessment of the revised manuscript within the scope of a GMD model-description paper.

We thank you for your careful editorial consideration.

Yours sincerely,

Jinkyu Hong, Ph.D.
Professor, Department of Atmospheric Sciences
Yonsei University
Corresponding author on behalf of all co-authors

Report #1: 'Comment on egusphere-2025-4938', Anonymous referee #1, 25 Feb 2026

In response to the reviewers' previous comments, the manuscript has been substantially improved. The authors have addressed many of the earlier shortcomings by performing OSSE experiments, providing a more detailed description of the inversion system, and expanding the interpretation and discussion of the results. Although this manuscript is primarily a model description paper, it is essential to demonstrate that the model produces reasonable results when applied in practice. Adequate interpretation of the outcomes and clear justification of the chosen configurations are necessary to build confidence in the developed system and to support its future application. In this regard, the revisions made so far represent an important and necessary step.

Provided that the following revisions are addressed, I believe that the manuscript is suitable for publication.

Specific comments

Lines 12 and 14 (Abstract):

The term greenhouse gas (GHG) is repeated twice in close succession. Please retain the full term at its first occurrence and use GHG alone in line 14.

>> We thank the reviewer for catching this. The redundant expansion of "greenhouse gas (GHG)" on line 14 has been removed; the acronym is now defined at its first occurrence (line 12) and used consistently thereafter. We also checked the full abstract for other acronym inconsistencies.

Line 33:

In the phrase "residual CH4 discrepancies consistent with background uncertainties", the meaning of background is ambiguous. Please clarify whether this refers to boundary conditions or another specific background component.

>> We agree that "background uncertainties" is ambiguous here. The sentence has been revised to: "…indicate residual CH_4 discrepancies consistent with prior emission and boundary condition uncertainties," explicitly identifying the two dominant error sources in this system.

Line 36:

Please clarify that ROK-BTR stands for Biennial Transparency Report of the Republic of Korea, or alternatively use the full name or refer to it as the Republic of Korea inventory. In addition, the phrase

"CO2 total is broadly consistent" is vague and not sufficiently scientific; consider replacing broadly with a quantitative metric or a more precise description.

>> We have spelled out the acronym as 'Republic of Korea Biennial Transparency Report (ROK-BTR)' at first use. We also replaced 'broadly consistent' with quantitative comparisons: the posterior national CO_2 total (620 ± 45 Mt yr^{-1}) is consistent with the ROK-BTR estimate (624 Mt yr^{-1}), and the posterior CH_4 total (54.7 ± 5.2 Mt CO_2eq yr^{-1}) exceeds the ROK-BTR estimate (35.5 Mt CO_2eq yr^{-1}) by 19.2 Mt CO_2eq yr^{-1}.

Line 215:

Additional explanation is needed regarding the default observation error specifications used in DART.

>> We have added clarification. For PREPBUFR meteorological observations, DART assigns observation error standard deviations from the NCEP 2005 GFS error tables. Errors are platform- and pressure-dependent: radiosonde (rawin) wind errors range from 1.4 m s^{-1} near the surface (~1000 hPa) to 3.2 m s^{-1} near the tropopause (~250 hPa); land station wind errors are 3.5 m s^{-1} (constant); marine wind errors are 2.5 m s^{-1} (constant); radiosonde temperature errors range from 0.8 to 1.5 K depending on level; and surface pressure errors are 1.0 hPa. These are standard operational values applied when no observation-specific error is present in the PREPBUFR record. A summary has been added to Section 3.

Lines 243–245:

The sentence "Main emission source regions in the model domain are boxed: (SMA: the Seoul Metropolitan Area, MWI: Mid-Western Industrial Area, SEI: SouthEastern Industrial Area, SCI: South Coast Industrial Area, CLA: Central Livestock Area)" mixes colons and parentheses inconsistently. Please revise for consistent punctuation.

>> We have revised the sentence to use consistent punctuation throughout. The abbreviations now follow a uniform parenthetical format, and "SouthEastern" has been corrected to "Southeastern" for consistent capitalization.

Lines 258–259:

In the comparison between modeled 10-m winds and KMA observations, the analysis is currently limited to the GSN and ULD sites. Please consider including results for the AMY site as well. For ULD, the discrepancies between observations and model results are relatively large (Table S1, Figs.

S1 and S2). A discussion is needed to explain the possible reasons for these differences and their potential impact on the greenhouse gas assimilation results.

>> We thank the referee for these two constructive requests.

(1) We have extended the meteorological evaluation to include the AMY site. Since no KMA ASOS station is co-located with AMY, we used Seosan (ASOS station 129), the nearest KMA ASOS station to AMY (~32 km NNE on the mainland), for the wind comparison. Table S1 and Figs. S1–S2 have been updated to include this additional comparison together with GSN and ULD.

(2) For ULD, the elevated wind errors are primarily attributable to the limited ability of the 9 km model grid to resolve the complex volcanic terrain of Ulleungdo Island. Two independent diagnostics support this attribution. First, we added a comparison against the Marine meteorological observation buoy ~18 km east of ULD over open ocean (ULD Buoy in Table S1), which samples the surrounding marine wind field rather than the orographically modulated land-station environment. The model agrees substantially better with the buoy data than with the ASOS station (Table S1), indicating that a non-trivial fraction of the ASOS discrepancy reflects local representativeness errors over the island rather than systematic model bias in the transport fields. Second, we performed an internal sensitivity experiment in which our team's additional near-surface parameterizations (topo-wind, slope effects, roughness sublayer scheme, LiDAR-derived canopy height, etc.) were disabled. Relative to this sensitivity run, the standard configuration reduces the wind-speed MBE from +2.90 to +0.72 m s^{-1} and the RMSE from 4.25 to 2.97 m s^{-1}, confirming that these treatments improve near-surface wind simulation in this complex-terrain setting, consistent with our earlier studies cited in Section 3 (e.g., Lee et al., 2015; Lee and Hong, 2016; Lim et al., 2019; Lee et al., 2020; Lee et al., 2023; Kim et al., 2024; Jo et al., 2025; WMO, 2025).

Lines 296–298:

These sentences require revision for clarity.

>> We have revised these sentences for clarity and corrected the typographical error ('is uses' → 'it relies on').

Lines 317–320:

While the text states that the TIMES factor was applied to CO2 emissions, it is unclear how temporal variability in CH4 emissions was treated. Please clarify this point.

>> We thank the referee for this clarification request. We have added explicit text clarifying that for CH_4 anthropogenic emissions, no diurnal TIMES disaggregation is applied; temporal variation is

retained at the monthly level. This simplification reflects the weaker diurnal cycling of the dominant Korean anthropogenic CH_4 source sectors (waste management, livestock) compared to combustion-driven CO_2. We acknowledge that incorporating sector-specific diurnal profiles for CH_4 would improve accuracy and identify this as a target for future work.

Line 359:

Remove the period following WRF-Chem.

>> We have removed the extraneous period following 'WRF-Chem' and also removed the unnecessary article 'the' before 'WRF-Chem'. A full manuscript search was conducted to correct any similar punctuation errors.

Line 440:

There is an extra period within the sentence; please correct this.

>> We have corrected the punctuation error at L440. The sentence fragment introduced by the erroneous period has been resolved by integrating the clause with the preceding sentence.

Lines 528–529:

The improvement in CH4 at the GSN site is attributed to the "resolution of the latitudinal OH gradient". Please clarify whether this means that data assimilation effectively enhances the representation of the OH gradient, leading to a significant reduction in CH4 errors.

>> We thank the referee for requesting this clarification. We have revised the text to make clear that the DA system does not directly modify OH fields. Rather, it adjusts CH_4 concentrations to match observations, correcting the latitudinal CH_4 gradient inherited from the CAMS EGG4 boundary conditions. The original phrase 'resolution of the latitudinal OH gradient' was imprecise and has been replaced with a mechanistically accurate description.

Line 545 (Figure 5 caption):

The caption refers to filled symbols, but only open symbols are visible in the figure. Please check and revise accordingly.

>> We thank the referee for catching this inconsistency. We have revised the caption to correctly describe the symbols as 'Symbols', instead of 'Filled symbols'.

Line 547 (Table 1 caption):

A closing parenthesis is missing after RMSE.

>> We have added the missing closing parenthesis after 'RMSE' in the Table 1 caption (corrected to 'root-mean-square error (RMSE) of the top-down estimates') and checked all table captions for similar parenthetical formatting inconsistencies.

Line 557 and Figure S9:

Please replace Figure S9 with a higher-resolution version. In addition, the caption should refer to Fig. 6 instead of Fig. 5. Currently, only CH4 trajectory plots are shown; adding corresponding CO2 trajectories would improve clarity for readers.

>> We thank the referee for these suggestions. We have (1) regenerated Fig. S9 at 300 dpi, (2) cross-reference in the caption now correctly cites Fig. 5, and (3) added CO_2 flight-track concentration panels alongside the existing CH_4 panels, using the same CM-01 aircraft data sampled at 5-min intervals.

Figure 6:

Including prior mean vertical profiles in addition to the posterior would allow readers to more easily assess the improvement achieved through data assimilation.

>> We thank the referee for this suggestion and considered it carefully in the revision process. We have ultimately retained Figure 6 in its original posterior-and-observation form, for the following reason rooted in the cycling-ensemble-DA framework: the cycled prior (6-h ensemble forecasts) in the middle of DA cycles is itself an analysis product (it incorporates every previous cycle's analysis), not an independent uninformed baseline. The visual prior-to-posterior difference at any cycle reflects only the per-cycle increment, which is small for passive tracers (CO_2, CH_4) because the ensemble cross-covariance between the full concentration fields and unobserved emissions over a 6-hour window is not strong; emission information accumulates across cycles rather than within any single cycle.

The convention for real-data validation in cycling ensemble DA is posterior fit to independent (non-assimilated) observations. Prior-vs-posterior comparison is informative when the prior is a static, uninformed baseline (such as in single-shot 4D-Var or batch Bayesian inversion); in our cycling framework, the prior at each cycle is itself part of an analysis product, not a baseline. Cumulative DA-driven learning evidence for the system is provided in emission space (50-site OSSE recovery against known truth) with the cycled-prior-to-truth tracking shown in Fig. S16 of the Supplement S2.

Line 683:

Typographical error; please correct.

>> We have corrected the typographical error ('constrainaed' → 'constrained') at L683 and performed a spell-check of the full manuscript.

Lines 711–712:

The statement comparing national posterior totals with the inventory may be misleading. The relatively good agreement applies to CO2, whereas CH4 shows larger discrepancies. Please explicitly indicate that this comparison refers to CO2 only.

>> We thank the referee for pointing out this ambiguity. The referee is correct that CO_2 and CH_4 show different levels of agreement with the inventory. We have revised the text to explicitly separate the CO_2 and CH_4 comparisons and to note the larger discrepancy for CH_4, consistent with evidence of national inventory underestimation in agricultural and waste sectors.

Report #2: 'Comment on egusphere-2025-4938', Anonymous referee #2, 30 Mar 2026

The authors have considerably extended the manuscript, compared to the initial submission. The methodological description is somewhat clearer, although it can still be improved. The results analysis and interpretation has also been improved. However, they have added a new OSSE experiment, which is terribly designed, and whose results are extremely poorly interpreted. I have some theoretical reservations on the adequacy of their inversion approach, now that I understand it better, but with a well designed experiment, they could help lift these. Unfortunately, their (single) OSSE is designed to "showcase" their setup, rather than help determine its strengths and weaknesses.

I think that, beyond the poor quality of at least some parts of the revision, the problem is that the scope of the paper is not well defined. They resolve a joint meteo-tracer-emission state, with two tracers, using an unpublished 6-hourly sequential data assimilation framework, and make quite a few bold claims about the advantages of their system. It's a lot of things to prove in a single paper.

The manuscript hasn't really gotten closer to publication-ready since the first revision, I doubt that the comments below can be addressed in just one or two extra rounds of revisions (even sections 2 and 7, which are now in a better shape, still need significant improvements). For this reason, I recommend a rejection, with an encouragement to resubmit a new manuscript, with a better focus and experimental design.

>> The substantive concerns are addressed individually in the specific responses below. This overview addresses four framing issues that recur across the report. Figure numbers are based on the revised manuscript.

Scope. The manuscript is a model description paper for GMD describing the WRF-Chem/DART regional EAKF framework as a multivariate, high-resolution joint meteorology-tracer-emission assimilation platform. It documents the framework architecture in Section 2, presents the controlled twin OSSE as an auxiliary perfect-model diagnostic in Supplement S2, and evaluates the 2020 real-data application in Section 6. It is not a comparative benchmarking study against alternative inversion classes (Lagrangian footprint-based, batch 4D-Var, offline Eulerian). Such comparisons would each constitute separate studies, and any comparative claims that exceeded this scope in the original Conclusions have been removed in the present revision (see L778–803).

Published methodological basis: The report characterizes the manuscript as describing *"an unpublished 6-hourly sequential data assimilation framework."* We respectfully clarify the description of the framework as 'unpublished'. The underlying components - DART/EAKF for ensemble Kalman filter data assimilation (Anderson, 2001, 2003, 2009; Anderson and Lei, 2013) and

WRF-Chem for online-coupled chemistry transport (Grell et al., 2005; Skamarock et al., 2008) - are long-established. The integration of WRF-Chem with DART for cycling chemical and trace-gas data assimilation is a documented published lineage spanning chemical reanalysis, NO_x assimilation, anthropogenic emission adjustment, and OCO-2-based CO_2 flux inversion (Mizzi et al., 2016, 2018; Gaubert et al., 2016; Liu et al., 2017; Ma et al., 2019, 2020; Zhang et al., 2021a, b). Ensemble-based simultaneous state–parameter estimation for regional CO_2 flux inversion specifically has been developed and published in multiple recent frameworks (Chatterjee et al., 2012; Kang et al., 2011, 2012; Peng et al., 2015, 2023; Liu et al., 2016, 2019, 2022; Kou et al., 2017; Chen et al., 2019, 2023; Khade et al., 2021). 6-hourly sequential analysis cycling is standard practice in operational and research data assimilation, including operational reanalysis systems (NCEP, ECMWF). The framework presented in this manuscript builds on this published lineage and documents the system integration; the framing of the framework as 'unpublished' mistakes the integration of published methods for the absence of underlying methodology.

OSSE design. Following the editor's guidance, we have reduced the prominence of the OSSE in the main manuscript and moved the detailed design, figures, and results to Supplement S2. The OSSE is now framed as an auxiliary, idealized information-content diagnostic under perfect-model assumptions, not as evidence of real-world robustness, operational flux accuracy, or grid-scale emission resolvability. We also added a 3-site reference OSSE under matched filter settings to contextualize the dense-network experiment and added an explicit posterior emission error map in Supplement S2 (Fig. S17), defined as posterior minus truth, to address the reviewer's request for a direct truth-based OSSE diagnostic.

Evaluation standards. We could not identify a specific factual or logical error that cannot be revised for several judgements (e.g., "terribly designed", "extremely poorly interpreted", "showcase rather than test", and the rejection recommendation). We have revised the relevant text to clarify the limited role of the OSSE, to avoid any implication of operational robustness, and to place the primary practical evaluation on the 2020 real-data application and independent aircraft comparison.

Main remarks

Inversion technique

As far as I could understand, during an assimilation cycle, the data assimilation system computes an analyzed state that comprise meteorological variables, the atmospheric tracer (CO2 and CH4) concentrations, and emissions which are in fact "surface forcing parameters for the subsequent forecast". Only the meteo and concentration are observed, so the emission forcings are constrained indirectly, through their (ensemble-based) correlations with the concentrations. The inferred atmospheric state is then used to initialize the next 6-hourly forecast, wich runs with the emission forcings optimized in the previous analysis cycle.

If I understand correctly, this means that there is no causal relationship (in the model or in the real world) between the emissions and the observations that are used to optimize them. This approach is justified in an atmospheric composition forecasting application, where the aim is to obtain the most realistic forecast of the atmospheric composition, and adjusting the emissions during the forecast cycle p based on the analysis of cycle p-1 is an efficient way to do that: if the emissions are overestimated at p-1, they will probably also be over-estimated at p, though it's not guaranteed and so it may occasionally degrade the forecast.

I am far less convinced about the appropriateness of this approach for GHGs inverse modeling, where the focus is on the accuracy of the emission estimate, and not on that of the 4D concentration fields: it certainly works, to a degree, but is it really better than making use of the actual, modelable causal relationships between the observations and the emissions anterior to them? It implicitly relies on the fact that emission errors are correlated in time, which is often the case, but not always. These considerations are not stated in the paper.

Another issue with this approach is that the short assimilation window (6 hours) means that a given observation will never be informative of emissions further back in time. This may be justified with short lived tracers (e.g. NO2), but with stable tracers such as CO2 or CH4, emission signals are carried over long distances and time. It may be that the information is still passed, indirectly, through the propagation of the optimized concentrations, but this is not discussed.

>> The reviewer's questions about the inversion approach are addressed in detail at specific comments for L195, L482, L502–504, L505, L506–508, and L511–512. Several of the framings in this overview appear to interpret the cycling mechanism differently from how it is implemented in our framework and in many similar DA studies based on WRF-Chem. We address the main methodological questions here and provide detailed responses to the corresponding line-specific comments below.

"No causal relationship between emissions and observations." We respectfully clarify this point because this appears to reflect a misunderstanding of the framework. In the EAKF, observations constrain emissions through the ensemble forecast covariance: each ensemble member's emission perturbation propagates through WRF-Chem to a concentration response, and the resulting cross-covariance between simulated concentrations at observation locations and the underlying emission parameters provides the regression that maps innovations to emission analysis increments. The causal pathway is the model itself (emission perturbation → forecast concentration response → ensemble covariance → analysis increment). This is the standard mechanism for joint state–parameter estimation in ensemble Kalman filters (Aksoy et al., 2006; Kang et al., 2011, 2012; Liu et al., 2019, 2022; Anderson, 2001, 2003), and is documented in Section 2.

"Implicitly relies on the fact that emission errors are correlated in time." The framework does not prescribe an explicit temporal error-correlation model for emissions (not 4d-VAR). However, as in most sequential state–parameter estimation systems, the updated emission field persists into the subsequent forecast window. Each analysis cycle's emission update is derived from the current ensemble's flow-dependent forecast covariance, which is computed dynamically rather than assumed. The cycled prior carries forward the previous posterior mean and the inflation-managed ensemble spread, but no assumption of temporal correlation in the emission *errors* is required. The method is therefore most appropriate for systematic or slowly varying emission errors over the cycling period, and we have clarified this assumption and limitation in Section 2.2.

"A given observation will never be informative of emissions further back in time." This conflates sequential cycling DA with single-shot retrieval. In a cycling system, the emission field is persistent across cycles, and the analysis at any cycle is conditioned on all prior observations via the cycled prior. Information propagates forward through this persistent emission field; an observation does not need to "look backward" because the cycled prior already integrates all prior constraint cumulatively. The reviewer's own subsequent remark, *"It may be that the information is still passed, indirectly, through the propagation of the optimized concentrations,"* correctly identifies this mechanism. It is documented in Section 2 and addressed substantively at specific comments for L195, L506–508, and L511–512.

"This is not discussed." We discussed this issue clearly in Section 2 in the previous revision (cycling architecture and joint state–parameter estimation). We have expanded the discussion in Section 2.2 to state how information propagates through the cycled prior and how emission updates are carried into subsequent forecast windows. Supplement S2 provides an idealized diagnostic of this behavior under known-truth conditions.

Stable-tracer concern. For passive tracers such as CO_2 and CH_4, the per-cycle emission information is small relative to the per-cycle concentration information because the ensemble covariance between surface concentration observations and emissions over a 6 h window is weak. Emission information therefore accumulates across cycles through the persistent emission field. We now state this explicitly as both a design feature and a limitation of the sequential state-parameter formulation.

OSSE experimental design

I don't think that the remarks above regarding the inversion method should be necessarily blocking. It may be that they don't matter much in practice, or just that I didn't understand the approach (if so, please clarify it). Regardless, since it is a new inversion setup, its performance should be tested, and that's the role of OSSEs.

So, in principle, adding an OSSEs section to the paper was a good idea, the problem is that the single one that was added is really poorly designed. They have created a known truth and a perturbed prior, by removing one entire emission sector, representing >40% of the anthropogenic emissions over Korea, so it's a massive signal, that will have a major impact on the fit to observations. They have then added 50 sites with synthetic observations, distributed downwind of urban areas … closely following the geographical extent of the emission perturbation, and comprehensively covering it (at the part inside South Korea). For reference, 50 sites is on the same order of magniture that the entire GHG monitoring networks in the US or in Europe.

Besides the ensemble perturbations, the only differences between the true and prior simulations is this emission perturbation. The boundary condition is the same, the transport model is the same, the C fluxes in other sectors (including natural) are the same. This can be a useful technical test, but it doesn't provide any meaningful insights on the performance of the DA, so I don't understand the rationale for presenting it.

The stated aim is to provide a "technical illustration of the framework behaviour under substantially expanded observational coverage". But it doesn't even do that: first, a reference experiment, without expanded network, would be necessary. Such an experiment, with much fewer sites, would likely be configured with longer localization lengths and eventually … lead to almost exactly the same result, since the perturbation is spatially homogeneous (at least in proportion of the prior). It is heterogeneous at the pixel scale (point sources), but that's not something that even their 50-sites network can resolve anyway.

Regardless, the objective of an OSSE is not to be a "show-case", but to test a refutable hypotheses, and, ideally, in this case, determine what are the limits of the system in terms of its capacity to resolve emission errors, at different spatial and temporal scales. Doing a sensitivity test with a few extra sites to test the impact of the network density can be a good idea … covering the domain of observations not so much.

>> The reviewer's specific concerns about the OSSE design are addressed at L440, L452, L482, L485, L502–504, L513–515, and L527–535. The reviewer's specific concerns about the OSSE design contradicts the original request (e.g., different request at L440 between the first and second revision). In response to these concerns and the Editor's guidance, we have moved the detailed OSSE design and results to Supplement S2, narrowed its interpretation, added a 3-site reference OSSE, and added a posterior emission error map. We now present the OSSE as an auxiliary perfect-model information-content diagnostic, rather than as a showcase of real-world robustness or operational performance.

OSSE class and rationale. The OSSE is an idealized perfect-model diagnostic (Errico et al., 2013; Hoffman and Atlas, 2016), in which a known imposed signal is recovered from synthetic observations under controlled conditions to verify the algorithmic behavior of the assimilation system. Holding transport, boundary-condition treatment, and non-target sectors fixed is intentional because it isolates the response of the implemented state-augmentation machinery to a known prior-emission perturbation. We have moved the detailed OSSE to Supplement S2 and now describe it only as an information-content and algorithmic-consistency diagnostic, not as an operational performance assessment.

The imposed perturbation is not spatially homogeneous, on either reading. The reviewer claims that *"the perturbation is spatially homogeneous (at least in proportion of the prior)"* but the perturbation is not spatially homogeneous. *In absolute terms*, the removed EDGAR power-industry sector is concentrated at discrete point sources in industrial corridors and zero elsewhere; spatial homogeneity does not describe this distribution (Fig. S15a). *In proportion-to-prior terms*, the same heterogeneity holds: the power-industry contribution as a fraction of the total prior emission varies sharply across cells. It is dominant at power-plant cells, modest at urban background cells, and vanishing in rural and ocean cells, because the spatial pattern of the power sector does not follow the spatial pattern of total anthropogenic emissions. We have revised the wording from 'highly heterogeneous' to 'spatially structured' to avoid overstatement.

Reference experiment now provided. In direct response to the reviewer's call for a reference experiment, the 3-site reference OSSE has been added in Supplement S2 under filter settings matched to the 50-site experiment except for the number of observation sites. The 3-site and 50-site experiments use matched filter settings, with the number of synthetic observing sites as the main controlled difference. The resulting recovery time series and recovery metrics are now provided in Supplement S2 (Fig. S16).

OSSE results analysis

Even omitting the design flaws of the OSSE experiment presented, its results interpretation needs to be improved. The inversion recovers about half of the prior error, at the national level, and “the posterior emission increment exhibits a distribution that is broadly consistent with the true difference …”: this goes in the right direction, but the aim is not just to “detect a signal”, but use it to accurately quantify the missing emissions. Therefore the success criteria cannot just be that “broad consistency”, it needs to be quantified. There is no presentation or dicussion of the fit residuals to the observations. There is no presentation of the posterior emission error (since the truth is known, that would be a much more useful metric to present than the posterior uncertainty!).

If only about 50% of the missing emissions are recovered, does that mean that the observations are, somehow, not sensitive to the remaining 50%? (with such an extensive coverage)? Is it that adjusting the tracer fields directly offers degrees of freedom that allow fitting the observations without changing the emissions (that wouldn’t be good, in a twin experiment where only the emissions are perturbed!)? Or is it that the observations are improperly fitted? (again, not great ...)

The authors make a lot of claim about the supposed advantages of their system compared to “Lagrangian footprint-based inversions”, both in the introduction and in the discussion. But any not-too-poorly designed Lagrangian inversion should be able to recover the missing emissions here, with such an observational coverage. It may be that there are valid reasons why this is not the case here, but they should be discussed more than in vague terms.

>> The premise of this section requires correction on two points, and one assertion in the report deserves direct engagement.

The Lagrangian assertion. We do not use the OSSE to claim superiority over Lagrangian or other inversion frameworks. The reviewer’s statement about how a Lagrangian inversion would perform may be plausible in some configurations, but it is not tested in this manuscript. In response to the reviewer and editor, we have removed language that could be read as claiming demonstrated superiority over footprint-based approaches. The revised manuscript now presents Lagrangian, Eulerian, variational, analytical Bayesian, and sequential ensemble methods as complementary frameworks with different assumptions, error structures, and trade-offs. A rigorous intercomparison would require a dedicated experiment and is beyond the scope of this GMD model-description paper. The report states that "*any not-too-poorly designed Lagrangian inversion should be able to recover the missing emissions here, with such an observational coverage.*" This claim is asserted without any supporting reference, comparative experiment, or quantitative basis. We additionally note that the manuscript does not claim demonstrated comparative superiority over Lagrangian inversions, and any phrasing that previously approached such a claim has been removed in the present revision (see L778–803). The framework is presented on its own merits as a multivariate, high-resolution joint state–parameter assimilation platform, not as a competitor to footprint-based methods. The GHG inverse modelling literature is methodologically pluralistic: Lagrangian footprint-based, online-coupled Eulerian, offline Eulerian, batch variational, and sequential ensemble methods each carry distinct strengths and trade-offs and address different scientific questions and observational regimes. The diversity of applications across these classes is a feature of the field, not a deficiency to be normalized. A model description paper for one regional online-coupled Eulerian system does not require, and should not require, demonstrated benchmarking against any specific alternative class. The reviewer's framing of the Lagrangian class as an implicit reference standard, against which other systems must demonstrate parity, reflects a narrower methodological scope than the GHG inversion literature itself supports.

Recovery rate. In the present revision, the OSSE has been revised in response to the reviewer’s concern at L482 in two ways.

(i) The localization half-width has been increased from 0.01 rad (~64 km) to 0.025 rad (~159 km), consistent with the Section 6 real-data configuration. This wider localization allows observations to influence a broader upwind region, as requested for passive tracers such as CO_2.

(ii) The prior emission uncertainty has been increased to 50 % so that the ensemble can better encompass the deliberately imposed truth–baseline emission gap. The previous 30 % spread did not provide sufficient ensemble spread for the missing power-industry sector, which accounts for approximately 45 % of South Korean anthropogenic CO_2 emissions in this OSSE. In the present revision, both OSSE configurations are evaluated under the same internally consistent settings: a 0.025 rad localization half-width and a 50 % prior emission spread.

Under this revised configuration, the 50-site OSSE recovers 86.5 % of the imposed truth–baseline emission gap over South Korea, whereas the 3-site reference experiment recovers 25.5 % under otherwise identical settings (Fig. S16 in Supplement S2).

We now report the recovery metrics transparently for both networks and explicitly state that the sector-removal perturbation is deliberately large and is not intended to represent typical real-world prior uncertainty. To further use the known truth in the OSSE evaluation, we also added a posterior emission error map in Supplement S2 (Fig. S17), defined as posterior minus truth.

Other concerns. We agree that comparison with the known truth is essential in an OSSE. In the revised manuscript, we evaluate this comparison using emission-space diagnostics at the regional and national scales: the South Korea-integrated recovery rate, the spatial correspondence between the imposed truth–baseline difference and posterior increments, prior uncertainty, posterior uncertainty reduction, and the recovery time series (Fig. S16). This considers a property of Kalman/ensemble filtering theory, not a deficiency of our implementation. This is presented together with an explicit posterior emission error map for both the 50-site and 3-site experiments (Fig. S17), defined as posterior minus truth. These diagnostics are interpreted at regional and pattern scales under perfect-model assumptions, not as evidence of grid-cell-scale resolvability or real-world robustness.

Specific comments

The line numbers refer to the track-changed document.

L73: "Atmospheric inverse modeling frameworks for GHGs typicall employ either Lagrangian or Eulerian methodologies. Lagrangian-based … depend on external meteorological fields … Conversely, Eulerian frameworks solve advection-diffusion equations directly on fixed spatial grids ... Notable examples of operational Eulerian systems include the NOAA CarbonTracker ...": So as an example of a Eulerian model that "solve advection-diffusion equations directly on fixed spatial grids", you chose a model that precisely doesn't do that (TM5 is an offline transport model, it relies on ERA5 meteorological forcings … exactly the same way than would do a Lagrangian model).

>> We agree that CarbonTracker/TM5 is an offline transport model driven by prescribed reanalysis meteorology, and our original example therefore did not support the accompanying claim about "a physically consistent coupling of meteorology and tracer transport." That coupling property belongs to online-coupled regional Eulerian systems such as WRF-Chem, not to offline Eulerian systems. We have removed the TM5 example from the passage and reframed it around how each class of framework numerically represents tracer transport, rather than around a binary Lagrangian-versus-Eulerian contrast. Lagrangian approaches track air parcels along trajectories computed from prescribed wind fields; global offline Eulerian systems solve grid-based continuity using reanalysis wind, convective mass-flux, and boundary-layer depth fields supplied at each integration step; online-coupled regional Eulerian systems such as WRF-Chem integrate meteorology and tracer fields through a single numerical framework, applying the same advection operators, turbulence closures, and planetary-boundary-layer scheme to both.

We respectfully note that the reviewer's further characterization of an offline Eulerian model as operating "*exactly the same way*" as a Lagrangian model is not accurate in a numerical sense. An offline Eulerian model such as TM5 solves the continuity equation on a fixed Eulerian grid, whereas Lagrangian models such as FLEXPART (Pisso et al., 2019) track parcel trajectories backward from receptors. The two representations have distinct numerical properties regarding mass conservation, numerical diffusion, and sub-grid mixing, regardless of whether the driving meteorology is prescribed or generated internally. We revised introduction to better clarify these points.

L195: "The analysis directly updates the prognostic meteorological variables and 3D tracer concentration fields each cycle, while emission adjustments are estimated as surface forcing parameters for the subsequent forecast": so, emissions at cycle c+1 are influenced by the model data mismatches at time c. Then, at time c+1, a new model-data comparison will be performed (with the tracer and meteo data at time c+1), and the concentrations will be adjusted: in a perfect world, where the only error comes from the emissions (which is the case in your OSSE), adjusting the emissions should be sufficient to bring the model in agreement with the observations … which is impossible in this inversion approach… Or am I missing something?

>> We appreciate the reviewer's careful reading and the direct question "*Or am I missing something?*" The answer is yes, and we are glad to clarify. The reviewer is correct that emission sensitivity within one cycle is weaker for passive tracers (CO_2, CH_4) than for species with short atmospheric lifetimes. Within a single 6 h window, each analysis step applies a smaller correction to the emission field than to the concentration field, because the ensemble covariance between surface concentration observations and emissions is weak at that timescale. However, the conclusion that emission estimation is "impossible" does not follow. Emissions are static surface forcing parameters, not prognostic variables, so corrections applied at each cycle accumulate rather than being overwritten. Each corrected emission field enters the subsequent forecast, modifies the simulated concentrations, generates new innovations, and drives further emission updates. This is the standard mechanism for joint state and parameter estimation in ensemble Kalman filters (Aksoy et al., 2006; Kang et al., 2011, 2012; Liu et al., 2019, 2022).

Our OSSE results demonstrate this accumulation directly in emission space: the emission recovery time series (Fig. S16) shows the domain-total posterior emission tracking from the initial baseline toward the truth over the cycling window, the spatial structure of posterior emission increments is compared with the known true difference in Fig. S15, and the residual posterior emission error is shown explicitly in Fig. S17. A revised text in Section 2.2 makes this mechanism and its empirical verification explicit.

L199-207: I guess that what you mean by "correcting the pronostic tracer concentration state at each analysis time step" means just to initialize the next forecast with the last timestep of the previous analysis? (lots of jargon for saying something simple …). Note that the studies cited here concern NO2, which is a really different tracer (short-lived, highly reactive, very uncertain emissions), whereas CO2 and CH4 are passive and long-lived tracers, I'm not so sure if the justification for NO2 is equally valid for CO2 and CH4 …

>> We agree that the passage was unnecessarily complex and have simplified the language in the revised manuscript.

Regarding the cited studies: Liu et al. (2017) describes the general WRF-Chem/DART cycling framework that our system builds upon, and we retain this citation. The reviewer is correct that Hsu et al. (2024) addresses NOx, a short-lived reactive species. However, this distinction strengthens the justification for concentration state correction. The passage concerns why the concentration field must be updated at each analysis step, i.e., to prevent transport-driven biases from persisting and being aliased into emission estimates. For NOx, such biases are partly self-correcting through rapid photochemical loss (lifetime of hours). For CO_2 and CH_4, chemical loss is negligible at the cycling timescale, so transport-driven biases accumulate without meaningful attenuation. If concentration correction is warranted for NOx, where biases can self-correct through rapid chemistry, it is a fortiori necessary for CO_2 and CH_4, where chemistry is too slow to provide meaningful self-correction within the cycling window.

To complement this reasoning with direct evidence for passive tracers, we have added citations to studies that demonstrate joint state–flux estimation for CO_2 in ensemble DA systems: Kang et al. (2011, 2012) for CO_2 flux estimation with LETKF, and Peng et al. (2015) for regional CO_2 data assimilation in East Asia.

Equation 1: To be clear, do I underestand correctly that x_met and x_chem are at analysis cycle i, whereas x_flux is at i+1?

>> Not exactly. All three components (x_met, x_chem, and x_flux) are updated simultaneously at analysis cycle i through the EAKF. The distinction is in how they are used afterward. The updated x_met and x_chem fields directly initialize the WRF-Chem forecast from DA cycle step i to i+1. The updated x_flux, being a surface forcing parameter rather than a prognostic variable, is additionally propagated to the intra-cycle emission files via a multiplicative scaling (posterior/prior ratio) so that the corrected emission rates are applied throughout the subsequent forecast window. The revised passage (see Section 2.2; *"The updated emission fields are prescribed as static surface inputs that persist during the subsequent forecast window, rather than being advected as prognostic variables. Thus, the emission corrections obtained at each analysis cycle propagate forward through the forecast into the next analysis cycle."*) makes this timing explicit, with an added cycle-to-cycle

propagation sentence that also addresses the reviewer's related comment at L250 regarding "time-indexed forcing parameters".

L247: “H has no sensitivity to flux measurements” you mean it has no sensitivity to fluxes?

>> We note that the reviewer quotes "*H has no sensitivity to flux measurements.*" The manuscript text (L247) reads "*H has no sensitivity to flux elements*"; the phrase "*flux measurements*" does not appear in the manuscript. We have nevertheless revised the wording to "*fluxes*" for clarity.

L250: “these flux adjustments are estimated as time-indexed forcing parameters rather than being propagated as a dynamic prognostic state”: Sentence is unclear. What do you mean by “time indexed forcing parameters”?

>> We agree the terminology was unclear. "*Time-indexed forcing parameters*" refers to the fact that emission fields are applied as surface boundary forcings at each model time step within the forecast window, rather than being advected as prognostic 3-D variables. We have revised the sentence to state this plainly. We believe this will also address the comment for Eq. (1)

L440: What is the interest of doing a single OSSE with such an unrealistic amount of sites? What can you possibly conclude on the real-world capacity of your system to estimate emissions (with the real existing network)? Especially since you don’t provide a reference experiment to compare it to ...

>> The OSSE presented in Supplement S2 was added in direct response to the reviewer's Round 1 Comment 2, which stated: "*Although the study focuses on developing an inversion framework, it lacks prior validation of its performance. An OSSE experiment using synthetic observations would allow the authors to quantitatively assess the accuracy and robustness of the system before applying it to real observations.*" The 50-site experiment addresses this request by checking, under perfect-model assumptions, whether the augmented-state EAKF can transfer concentration innovations into physically interpretable emission increments when observational coverage is dense. We now frame this OSSE as an auxiliary diagnostic of algorithmic consistency and observing-network information content, not as evidence of real-world robustness. Furthermore, to characterize performance under the current sparse network, we have added a 3-site reference OSSE (Supplement S2) using the current operational WMO/GAW ground network under identical settings. Following the editor’s guidance, we have moved the detailed OSSE design and results to Supplement S2 and reframed the experiment

as an auxiliary perfect-model diagnostic. Readers can now directly compare constraint strength between dense and sparse networks within the same perfect-model framework.

L452: “the corresponding inversion (prior) configuration uses EDGAR … with the power-industry sector removed”: so, in other words, you impose a uniform(ish) bias to the anthropogenic emissions (in proportion of the prior).

>> The perturbation is not uniform. If a spatially uniform perturbation were intended, a simple multiplicative scaling factor (e.g., ×0.7) applied to total emissions would suffice. The sector-removal approach was chosen precisely because it produces a spatially heterogeneous signal: the removed power-industry sector is concentrated at discrete large point sources (coal-fired power plants and industrial complexes) in specific corridors (SMA, MWI, SEI, SCI), with zero perturbation in rural, forested, and agricultural areas. The sector accounts for 36.5% of domain-total and 44.7% of South Korea CO_2 emissions, but this domain-averaged fraction masks substantial grid-cell-level heterogeneity. Furthermore, only the power-industry sector is removed; transportation, residential, agricultural, and other sectors remain at their prior values. The perturbation is therefore heterogeneous in both space and sector.

L457: “By removing such a major component … detection problem”: That can be a useful test, but what does that tell you, in the end, of the capacity of your system to resolve real-world emission estimates? The real errors are both of much smaller magnitude (for the anthropogenic component), with more complex patterns, and entangled with errors in other emission categories.

>> We fully understand that no single OSSE can fully characterize real-world performance, nor is this the standard expectation for OSSEs in the literature (Hoffman and Atlas, 2016). We also do say that the OSSE is information-content behavior under controlled assumptions not real-world capacity, as the reviewer mentioned in the first revision. As we mentioned before, the OSSE under perfect-model assumption with large prior-emission uncertainties cannot characterize full real-world performance.

As noted in our response to the reviewer's L440 comment, the OSSE was added at the reviewer's explicit Round 1 request to "*quantitatively assess the accuracy and robustness of the system.*" The reviewer now asks what this tells us about real-world capacity - a different and broader question than was originally posed.

We clarify the purpose, scope, and limitations of our OSSE in this revised manuscript. This manuscript is submitted to GMD as a model description paper and our focus is to describe the inversion framework and demonstrate its functionality, not to provide an exhaustive characterization of real-world emission estimation under all possible error structures. However, we tried our best to follow up the reviewers' comments and added relevant information in the revised manuscript. As we repeatedly mentioned, our OSSE provides algorithmic consistency: the system detects and recovers a spatially heterogeneous emission signal under controlled conditions. Real-world complexity (smaller errors, multi-sectoral patterns, model and boundary condition errors) is addressed in the real-data demonstration (Section 6) using actual observations. The new 3-site reference OSSE bridges these by testing fewer-sites performance under controlled conditions.

L482: By applying such a short localization, you effectively prevent observations to constrain emissions a few hundreds of kms upwinds. With such a dense observation network, it might not be needed, but is this really what you want to do? CO_2 and CH_4 being stable tracers, long range transport should be modelled reliably enough. Or you don't trust enough the long-range flux concentration correlations in your ensemble, but then this is going to be a problem in real inversions, when you have only a handful of sites….

>> Both the 50-site OSSE and the 3-site reference OSSE have been re-run with the localization half-width widened from 0.01 rad / ~64 km to 0.025 rad / ~159 km (Supplement S2). The 3-site reference OSSE was newly added in this revision specifically to test the system under the current operational WMO/GAW ground network the reviewer flags. The resulting recovery rates and time series are reported in Supplement S2 (Fig. S16)

The real-data experiments in Section 6 use the 0.025 rad localization, providing a single localization choice across all three experiments (50-site OSSE, 3-site reference OSSE, real-data). Within the synthetic-OSSE pair, the only varying configuration parameter is the number of observation sites, directly attributing recovery differences to coverage. Localization remains essential at $N = 20$ to suppress sampling-noise contamination at the smallest separations (Anderson and Lei, 2013; Kondo and Miyoshi, 2016; Chang et al., 2022); the 0.025 rad half-width sits within the localization-required regime while extending the upwind reach. The variable-localization scheme applied to the augmented emission state aligns with Roh et al. (2015).

L483: "This configuration mimics a realistic operational capabilities assessment" No it doesn't! Nothing is close to realistic in that experiment.

>> We agree that the phrase "*mimics a realistic operational capabilities assessment*" overstated the experiment's scope and have removed it. However, we do not agree with the claim that "*nothing is close to realistic*". The filter configuration - observation error specification, localization settings, inflation parameters, and ensemble size - uses settings consistent with the real-data configuration, not idealized perfect-knowledge settings. The 50-sites OSSE is the controlled variable being tested to understand impacts of high observation network in larger a prior emission uncertainties. The revised text describes the experiment as characterizing system behavior under enhanced observational density with settings consistent with the real-data DA filter configuration

L485: "This ensures that the evaluation isolates the impact of expanded network density, rather than the effects of perfect parameter optimization": how can you claim to isolate anything, without having a reference experiment to compare to?
>> The 50-site OSSE is the algorithmic-auxiliary diagnostic of the framework. Under sufficient observational coverage, the system recovers a known spatially heterogeneous emission signal, which is the controlled consistency check the reviewer's Round 1 Comment 2 explicitly requested. The 3-site reference OSSE (with the emissions recovery time series in the Supplement) introduced in response to the localization concern at L482 (see response to R1-10) repeats the same setup with fewer observation sites and yields reduced recovery under otherwise identical settings (observation error, ensemble size, localization, cycling interval). The contrast under matched configuration attributes the recovery gap to the number of observation sites, isolating the effect the manuscript text describes.

L488-499: Your inversion infers flux adjustments where there are observations, and where there are prior errors (there are locations with prior errors and no observations, e.g. in China, but no locations with observations and no prior error). That is certainly a good start, but that's not a proper evaluation of your results: Figure 4 shows the location of prior errors (4a) and analysis increments (4b), but doesn't show the posterior errors, although based on 4a and 4b, I imagine that you get a degradation of the fit to the true emissions almost everywhere, with even some emission increments over the sea (East of Seoul), where there should not be any emission or emission uncertainty. How do you define, and test, the expectations from such an inversion? With such a dense network, it should be quite high, and needs to be tested through a much more complex prior error structure.
>> We want to clarify that Fig. 4d that is exactly what the reviewer wanted (i.e., posterior errors). The reviewer's comment raises four points. We address each.

First, on the absence of a posterior error map (truth minus posterior): pixel-wise agreement with truth is not a deliverable of any data assimilation system operating with finite observations, a finite ensemble, and imperfect error characterization. The analysis carries non-zero posterior covariance by mathematical construction, and an N=20 ensemble Kalman filter with localization at 9 km resolution is rank-deficient in the directions along which posterior covariance can be reduced. The reviewer's own observation at L513–515 ("*we indeed cannot expect the inversions to recover the pixel-by-pixel emissions*") is consistent with this. The framework characterizes emission recovery at regional and national scales rather than at individual grid cells, and the diagnostics matched to those scales are (i) the spatial agreement between the imposed target (Fig. S15a) and the analysis increment (Fig. S15d) on logarithmic color scales over the same domain, supplemented by the prior uncertainty structure (Fig. S15b) and the posterior uncertainty reduction (Fig. S15c); (ii) the time-averaged scalar emission-space recovery rate over South Korea (defined in our response to L502–504); and (iii) the emission timeseries (Fig. S16).

Second, on emission increments over the sea: as already stated in Section 3 (L320), "*Over ocean grid cells, EDGAR emissions are near zero except along major shipping lanes.*" Our multiplicative prior emission perturbation scheme scales the prior ensemble spread proportionally with the prior magnitude, so prior uncertainty over ocean is correspondingly near-zero (Fig. S15b) and ocean analysis increments are bounded proportionally (negligible in absolute terms; Fig. S15d). We clarified this point in the figure caption and surrounding text.

Third, on expectations from the inversion: a single pre-defined success threshold is not a meaningful test for cycling ensemble DA because the recovery rate is a deterministic function of design choices (window length, ensemble size, observational coverage, observation error, and the rate-limiting effect of multiplicative prior perturbation). The full argument is given in our response to L502–504.

Fourth, on testing under "*a much more complex prior error structure*": the Round 1 Comment 2 request for "*an OSSE experiment using synthetic observations to quantitatively assess the accuracy and robustness of the system*" is delivered by the 50-site experiment; the 3-site reference OSSE with the posterior emissions time series in the Supplement S2 provides the complementary sparse-network baseline. W agree that multi-sector, smaller-amplitude, and boundary/transport-error OSSEs would be valuable future work. The present OSSE is intentionally narrower and is now framed as an idealized information-content diagnostic complementing the real-data evaluation, rather than as a complete robustness assessment.

L502-504: So the recovery is only about 50% of the error: What recovery rate should be expected to consider the experiment successful? Are the posterior emissions consistent with the observations (once propagated with the transport model, i.e. in a copy of your "nature" run, but with the posterior emissions)?

>> A predefined success threshold for OSSE recovery is not meaningful. The recovery rate is a deterministic function of design choices. First, we chose 15 days, which yields 60 cycles (as opposed to monthly-to-annual as shown in Kang et al., 2011, 2012). Second, we applied the filter configuration that uses settings consistent with the real-data configuration, rather than parameters tuned to the known truth. Third, in our inversion system for passive tracers, per-cycle emission corrections are small relative to concentration corrections, allowing the rate-limiting emission-information to accumulate. Finally, we chose the prior emission ensemble spread to scale with the prior emission magnitude, similar to the characteristic of perturbation schemes designed to preserve emission positivity. The per-cycle analysis increment is bounded proportionally by the prior, so cells whose true emission greatly exceeds the prior admit only correspondingly small per-cycle corrections.' The informative diagnostics are the spatial structure of the recovery, and the fact that these properties emerge under settings consistent with the real-data configuration without tuning to the known truth.

We agree with the reviewer's underlying concern that instantaneous concentration-state updates and emission updates should not be conflated. In an online cycling system, however, a strict replay with "posterior emissions only and all other components held at the prior" is not directly equivalent to the system actually evaluated, because meteorology, tracer concentrations, and emissions co-evolve through repeated analyses. We therefore address the concern using the controlled OSSE, where the emission truth is known. The OSSE provides emission-space diagnostics, including the integrated recovery rate and recovery time series, which directly quantify whether the cycling system transfers information into the emission field rather than only correcting concentrations.

These measures are reported in the paper: the emission-space recovery rate of 86.5% for CO_2 over South Korea, with 25.5% under the operational-network configuration in the 3-site reference OSSE (Supplement S2), and the spatial-structure alignment between posterior emission increments and the known true difference (Figure S15). Together, they characterize the accuracy of the posterior emission field sequence without requiring an additional forward-replay experiment.

L505: "These recovery rates should be interpreted as a conservative lower bound": How so? You put the model in an un-realistically easy situation, you have a complete observational coverage of the prior error, you have no model error, no boundary condition error, a huge, easily detectable prior error. This is the polar opposite of "conservative"!!!

>> We agree that "*conservative lower bound*" could be misread, particularly with respect to the perfect-model and synthetic-observation aspects of the experiment, and we have removed this phrase.

The OSSE is now described as an idealized information-content diagnostic. The perfect-model setup is used to isolate the response of the implemented state-augmentation system to a known prior-emission perturbation, not to represent real-world performance.

We respectfully note, however, that the reviewer's enumeration characterizes the experiment as lacking several uncertainty sources that are in fact explicitly included and described in the manuscript. As stated in Supplement S2, "*ensemble perturbations for emissions and IC/BCs are generated following the exact strategy used in the real-data simulations,*" with the perturbation scheme detailed in Section 2.2. Observation errors are introduced as unbiased Gaussian noise parameterized at 5% of the sampled mole fraction, dominated by a representativeness term. The perfect-model design eliminates systematic truth-versus-forecast divergence introduced by forward-model physics; it does not eliminate the prior-concentration spread or the observational uncertainty, both of which are sampled at every assimilation cycle. We now state these exclusions explicitly in Supplement S2 and in the main text.

Beyond these explicitly included uncertainty sources, the filter configuration uses settings consistent with the real-data configuration rather than parameters tuned to the known truth. The 5% observation-error specification amounts to approximately 20 ppm CO_2 at ambient concentrations, roughly 180 times the ~0.11 ppm instrument-only measurement uncertainty reported for the Korean WMO/GAW reference stations (Lee et al., 2019); the excess reflects the representativeness component that dominates the obs-error budget. The ensemble size (N = 20), the localization half-width (~159 km) chosen to suppress sampling noise at this ensemble size (see our response to Comment L482), and the 15-day assimilation window (60 cycles) are all values used in the real-data configuration. Within the perfect-model setup, tuning these choices to the known truth (for example, by increasing the ensemble size, lowering the assumed observation error, or extending the assimilation window) would be expected to raise the recovered fraction. The reported recovery rates therefore reflect the behavior of a DA configuration setting consistent with the real-data configuration rather than the intrinsic skill ceiling of the WRF-Chem/DART system under these conditions.

L506-508: “The short duration of the total assimilation period (14 days) relative to typical flux convergence timescales”: what is this “flux convergence timescale”? You have a comprehensive hourly observational coverage of your emission error, that’s more than sufficient to constrain the emissions over a two weeks periods. If you system cannot do this, it’s a design flaw.

>> We agree that the phrase "flux convergence timescale" was imprecise and have replaced it. However, the assertion that comprehensive hourly observations should be sufficient to fully constrain emissions within two weeks reflects a misunderstanding of how observations inform emissions in a

cycling ensemble filter for passive tracers (see our response to Comment L195 for the full mechanism). Hourly observations directly constrain concentrations. The correction applied to the emission field each cycle is inherently small, because the ensemble covariance between surface concentration observations and emission perturbations over a 6 h window is weak for passive tracers such as CO_2 and CH_4. Emission information accumulates across cycles and is not instantaneously available regardless of observation frequency. This is not a design flaw; it is a fundamental property of joint state and parameter estimation for passive tracers (Aksoy et al., 2006; Kang et al., 2011, 2012; Liu et al., 2019).

L507-508: "and the use of an operationally representative, conservative filter configuration …": but in a operational context, you would use these error configurations, and you would have far far fewer observations, so you would have a much much worse error recovery rate.

>> The L507–508 critique addresses the same underlying concern raised at L505 (that OSSE recovery is not representative of real-world conditions), focused here on one specific factor: observation density. The specific observation-density question is directly tested in this revision by the 3-site reference OSSE we have added (Supplement S2). The 3-site experiment uses the current operational WMO/GAW ground network and matches the 50-site experiment in observation error, ensemble size, inflation, cycling interval, localization half-width, and prior emission spread; the observation network is the only configuration that varies between the two synthetic experiments. The resulting reduction in recovery, reported in our response to L502–504, empirically quantifies the observation-density effect the reviewer asks about. The reviewer's argument that fewer observations would yield lower recovery presupposes that recovery is observation-driven; the controlled comparison confirms this scaling and locates the binding constraint at observational coverage rather than at system design. By the same scaling, the framework is positioned to deliver increasing recovery as observing systems mature toward multivariate, high-density configurations, which is the expansion trajectory the manuscript explicitly motivates (Supplement S2).

L508-510: "by avoiding idealized tuning …": You already have an idealized perturbation, with an idealized network and an idealized model. So maybe you should in fact start with an idealized tuning, as a consistency check to verify that, at least in an idealized case, you recover most of the error. And from there on move towards more complex cases to try identifying the limits of your setup. Your current setup is all but cautious!

>> The reviewer's Round 1 Comment 2 requested "*an OSSE experiment using synthetic observations to quantitatively assess the accuracy and robustness of the system.*" As we described in our replies and manuscript, the OSSE should not be evaluated as though it should demonstrate superiority, optimality, or operational readiness. Questions of optimal tuning, operational performance, and progressive stress-testing were not requested by the reviewer originally but will be valuable future work starting from the developed system.

We note inconsistency between the reviewer's comments on the OSSE, and we revised our manuscript to frame the OSSE more narrowly. The 50-site OSSE is an identical-twin / perfect-model OSSE, a specific class whose appropriate uses are clearly delineated in the OSSE methodology literature. Errico and Privé (2018, NASA TR GSFC-E-DAA-TN69069) state that "*an identical twin OSSE nonetheless may be used successfully for certain theoretical experiments or carefully designed proof of concept trials, but the perfect model setup should be avoided for any reliable quantitative evaluation of new observing systems*" (see also Errico et al., 2013). This exactly matches the manuscript's framing: the 50-site OSSE is presented as an algorithmic consistency check under perfect-model assumptions - the recommended use for this OSSE class - not as a quantitative evaluation of real-network operational performance, which the methodology literature explicitly cautions against for identical-twin frameworks. The reviewer's demand would push the experiment into exactly the regime the methodology literature warns against.

The OSSE in this manuscript tests the specific GHG configuration, not the underlying filter algorithm. We note that the manuscript addresses three distinct questions through three experiments, which the reviewer's comments at times conflate:
• Does the system recover known spatial structure? → 50-site OSSE
• How does sparse vs. dense coverage affect recovery? → 3-site reference OSSE
• Does the system perform under real observational constraints? → real-data demonstration
Applying real-world performance criteria to the first experiment, or algorithmic-consistency criteria to the third, crosses categories that the manuscript deliberately separates. Regarding "all but cautious": as detailed in our response to Comment L505, the observation errors (~180× instrumental precision; Lee et al., 2019; see response to L505), ensemble size (N=20), and localization (~159 km) are all conservative relative to what a perfect-model environment permits. We revised our manuscript for this clarification.

L511-512: I guess we are just supposed to believe your word regarding these expectations?

>> We have added an emission timeseries (Fig. S16) showing that the posterior emission converges from the initial baseline toward the truth over the cycling window in the OSSE experiments. The spatial structure of the posterior emission increment matches the imposed signal (Figure S15 A). These diagnostics replace the qualitative language in the original passage.

L513-515: “The imposed power-industry signal is highly heterogeneous”: is it, really? It’s a unidirectional bias, likely quite uniform in proportion to the overall anthropogenic emissions signal, and to where your observation sites are (conveniently) located. We indeed cannot expect the inversions to recover the pixel-by-pixel emissions, but the expectation should be much higher in such an idealized case than in a more realistic scenario, such as those you discuss in Section 7, and you never even define what these expectations should be.

>> On heterogeneity: see our response to Comment L452. The perturbation is spatially heterogeneous; concentrated at discrete point sources with zero perturbation elsewhere. If uniform scaling were intended, a multiplicative factor would suffice.

On "conveniently located" sites: the 50-site network is designed to provide dense coverage of source regions. This is the stated purpose of the experiment; testing algorithmic consistency under sufficient coverage. The 3-site reference OSSE (Supplement) uses the operational WMO/GAW stations, whose locations were not selected for this study.

On expectations: see our responses to Comments L502–504 and L511–512. A single pre-defined success threshold is not a meaningful test for cycling ensemble DA because the recovery rate is a deterministic function of design choices. The emission timeseries (Fig. S16) provides the empirical characterization that replaces the qualitative assertions in the original passage. We have revised "highly heterogeneous" to "spatially structured" to avoid overstating.

L527-535: These whole two paragraphs are complete nonsense, unsubstantiated blabber.

>> This comment does not identify a specific factual or logical error in either paragraph, nor does it suggest an actionable revision. We have nevertheless revised the relevant text to clarify the scope of the OSSE and to avoid any implication of operational robustness.

The first paragraph states that the OSSE used settings consistent with the real-data configuration rather than configuration optimized to the known truth, and that the recovered spatial structure reflects observational coverage rather than tuning advantages. The first claim is documented in Supplement S2: observation errors are specified at ~20 ppm (~180× instrumental precision; Lee et al., 2019; see

response to L505), with N=20 ensemble members and 0.025 rad (~159 km) localization (matched to the Section 6 real-data setting; see response to L482). The second is now directly supported by the 3-site reference OSSE, which provides the baseline comparison needed to attribute constraint strength to the number of observation sites (see response to L485). We have tempered the language: "underscores the robustness" → "is consistent with expected behavior."
The second paragraph discusses the framework's potential for observing-network design assessment. We kindly refer the reviewer to the well-populated sub-literature applying OSSEs to this exact problem class. Two same-journal (GMD) precedents are directly on point: Bisht et al. (2023) evaluate CH_4 emission estimation under varying observational coverage with 4D-LETKF, and Vardag and Maiwald (2024) optimize urban CO_2 measurement network configurations with a high-resolution OSSE. For anthropogenic CO_2 monitoring specifically, Santaren et al. (2021) develop a CHIMERE-based inversion system at 2 km resolution controlling emissions from ~300 cities and power plants for future satellite imagery assessment, and Broquet et al. (2018) present an OSSE testing 2–10 km CO_2 imagery configurations for monitoring spatially integrated city emissions. The underlying OSSE design methodology is formally documented by Errico et al. (2013); see response to L508–510 for extended discussion of OSSE-class appropriateness. We have tempered "powerful potential" → "potential" and shortened the passage to state only what the OSSE demonstrates.

Figure 4: The whole interest of doing an OSSE is that you can use the truth to validate your results. However, a comparison of the analysis to the truth is precisely what you are not showing here. You are also not showing the prior uncertainty structure, so it's impossible to tell how consistent are these analysis increments with the prior uncertainty. I find it quite worrying to see significant emission increase over the sea (where, if I trust Figure 6 of the original manuscript, there shouldn't be any emissions).
>> This comment substantively overlaps with Comment L488–499. We address the same three issues here: the absence of a pixel-level posterior error map, the absence of prior uncertainty structure, and ocean emission increments. We addressed all three in our response to that comment and refer the reviewer and editor there for the full discussion. Figure S15 now includes (a) the imposed target (the "true difference" in CO_2 emissions, defined as EDGAR full minus EDGAR without the power-industry sector), (b) the prior emission uncertainty (time-averaged prior ensemble spread), (c) the posterior uncertainty reduction (percent decrease in ensemble spread relative to the prior), and (d) the analysis increment (posterior minus prior). Panels (a) and (d) on a matched logarithmic color scale provide the pattern-level truth comparison the reviewer requests; panel (b) shows the prior uncertainty structure; panel (c) quantifies the posterior reduction in that uncertainty. Ocean increments remain

bounded proportionally to the near-zero ocean prior under our multiplicative perturbation scheme, as EDGAR ocean emissions are near-zero except along major shipping lanes (Section 3, L320).

In addition, Fig. S17 now provides the requested truth-based posterior emission error diagnostic, defined as posterior minus truth, for both the 50-site and 3-site OSSEs. Equivalently, this error is the posterior-minus-prior analysis increment minus the imposed true difference. This directly addresses the reviewer's request for a posterior-error comparison against the known truth. Ocean increments remain small in absolute terms because EDGAR ocean emissions are near zero except along major shipping lanes and the multiplicative perturbation scheme scales prior spread with prior magnitude. We interpret the posterior-error map together with the integrated recovery rate and recovery time series, rather than as a claim of grid-cell-scale emission resolvability with the reasons discussed before. As the reviewer mentioned as "we indeed cannot expect the inversions to recover the pixel-by-pixel emissions", we carefully revised our manuscript to explicitly bring up the OSSE design and its limitations under perfect-model assumptions.

L551: "With the controlled proof-of-concept established in the previous section …": but yet, you use a completely different observation network, different localization settings, etc., so what does the setup describe in Section 6 tell you about the results presented in Section 7? (not much …).
>> We are saying that the OSSE is algorithmic consistency check under perfect-model assumptions, not the controlled proof-of-concept. The characterization "*completely different...* etc." is overstated. The two synthetic OSSEs are configured identically across all cycling parameters in Supplement S2: observation error specification, inflation, ensemble size (N=20), cycling interval, localization half-width (0.025 rad, ~159 km), and emission prior spread (50%). The only difference is observation density (50 vs 3 sites), enabling the controlled comparison the reviewer's Comment L485 specifically requested. Section 6 (real-data) shares the filter architecture, localization, and ensemble configuration of the OSSEs; the emission prior spread is 30% in Section 6, calibrated to expected real-world emission uncertainty rather than to the OSSE-imposed truth–prior gap (see Comment L452). The 3-site reference OSSE uses the operational AMY–ULD–GSN network, bridging the controlled OSSE setting and the real-data experiments.

Section 7.1: Here I guess what you call "posterior concentrations" are the concentrations after the analysis step. However, you should also isolate the impact of the emission adjustment. You can do that via a new "nature" run driven by the inversion-optimized emissions, and with everything else kept the same as the prior. In a setup where the DA can adjust the concentration fields directly, I think this is critical.

>> Posterior concentrations refer to the post-analysis 3-D tracer field. This was documented in Section 2.

The proposed test (forward simulation with optimized emissions while holding IC, BC, and meteorology at prior) presupposes that emissions are the only optimized component while meteorology and tracer fields are externally prescribed and unchanged by the inversion. This is the architecture of offline-coupled inversion systems where a chemical transport model is driven by reanalysis meteorology and the augmented state contains only emissions. Our framework is online-coupled: the augmented state contains meteorology, 3-D tracer concentration, and emission scaling jointly, all updated at every analysis cycle and co-evolving through cycling. There is no "prior" meteorological or tracer-concentration trajectory that the system was operating from at any later cycle; every state component has analysis history attached. The "*everything else at prior*" reference trajectory therefore does not exist as a coherent dynamical state in our framework, and the proposed swap-and-rerun comparison cannot be interpreted as it would be in an offline-coupled system.

Direct concentration adjustment in the analysis step is a defining feature of atmospheric data assimilation, not an unusual or problematic property (Houtekamer and Zhang, 2016 and among others). We are not aware of this being a standard diagnostic for this class of online-coupled cycling systems. Please provide citation establishing the proposed test as a standard procedure for an online-coupled joint state-parameter inversion class, or any quantitative or theoretical basis for its claimed criticality.

The legitimate concern motivating the request (whether observation fit might be driven by direct concentration updates rather than by emission corrections) is real, but the proposed test does not answer it. The cycling-DA-appropriate diagnostic is direct truth comparison in the emission space, which the OSSE provides (see response to L502–504): if concentration updates were shortcutting the emission inversion, the recovery rate would be near zero. The emission-space recovery rate and time series (Fig. S16), together with the spatial increment diagnostics (Fig. S15) and posterior emission error map (Fig. S17), evaluate whether information is transferred into the emission field rather than only into the tracer state.

The truth-comparison the reviewer's test would interpret is moreover unavailable in the real-data setting (Section 6) regardless of architecture: the resolved emission field is precisely the quantity the inversion is designed to estimate. Bottom-up inventories (EDGAR) are themselves estimates with their own uncertainty (~20–30 %) and are not independent at the 9 km, 6-hourly scale; concentration observations sample mole fractions, not fluxes. This is the standard motivation for OSSE-based validation (see response to L508–510). The truth-comparison evidence for the framework therefore lives in Supplement S2, where the truth is by construction known.

L571: OH plays absolutely no role at these timescales (the residence time of air over your regional domain being a few days, at most). An incorrect representation of OH in the global model that serves as BC may be one of the drivers for latitudinal bias, but this is just a BC bias among others.

>> The manuscript does not claim that OH chemistry plays a role at regional timescales. As stated in our Round 1 response to Comment 10: "*the chemical lifetime of CH_4 against OH is on the order of years (about 9 years), whereas typical air-mass residence times across our regional domain are <1 day. The implied fractional loss is therefore ≈ 0.03%, which is negligible.*" This was also incorporated into the revised manuscript at Section 2.1. The reviewer's statement repeats a point we have already addressed and agreed with.

The sentence at L571 describes an observational result (the largest CH_4 improvement occurs at GSN) and attributes it to the correction of the boundary-condition-inherited latitudinal CH_4 bias. We have revised the phrasing to remove the ambiguous "OH gradient" reference.

L582-583: yet, it's not as pronounced in the observations, and in your "posterior" (whatever definition of it is used) concentrations. So your inversion corrects it somehow, and I would like to know how much of it is obtained by adjusting the emissions, and how much is obtained by adjusting the boundary condition (however that is done …). Here the extra diagnostic requested in my comment on Section 7.1 would be necessary.

>> The reviewer asks how much improvement is obtained by "*adjusting the boundary condition.*" The DA system does not re-estimate or optimize the external boundary-condition source: EGG4 fields are used as the prescribed inflow product throughout cycling and are not updated by the analysis step. Our Round 1 response to Comment 7 and the revised Section 2.2 state: "*Time-varying lateral boundary conditions are prescribed from the external product… the DART 'update_wrf_bc' procedure is used to apply the analysis increments consistently near the lateral boundaries to avoid spurious discontinuities, rather than to 'optimize' the external boundary forcing itself.*" Prior BC uncertainty is nonetheless explicitly represented: initial and lateral-boundary tracer inflows are perturbed across ensemble members at every cycle for prior-covariance sampling (see our response to Comment L505).

The reviewer's decomposition question therefore reduces to attributing the posterior improvement between emission corrections, the reduction of sampled prior IC/BC concentration uncertainty by the analysis update, and meteorological state corrections from PREPBUFR assimilation. Such a post hoc attribution is ill-defined in cycling ensemble DA, for the reasons detailed in our response to Comment Sec 7.1, where we also describe the cycling-DA-appropriate emission-space evidence.

Figure 6: where is the prior? And again, would be good to isolate the impact of the emission correction (since this is your key target).

>> Regarding the absence of a prior profile in Figure 6 (currently Fig. 5): in cycling sequential ensemble DA, the cycled prior at any cycle is itself an analysis product (it incorporates every previous cycle's analysis), not an independent baseline. Visual prior-to-posterior comparison at a single cycle therefore shows the per-cycle increment, which is relatively small for passive tracers (CO_2, CH_4) by the rate-limiting mechanism described in our response to L195 and L506–508; it does not show cumulative DA-driven learning.

The convention for real-data validation in cycling ensemble DA is posterior fit to independent (non-assimilated) observations. Prior-vs-posterior comparison is informative when the prior is a static, uninformed baseline (such as in single-shot 4D-Var or batch Bayesian inversion); in our cycling framework, the prior at each cycle is itself an analysis product, not a baseline. Figure 6 (currently Fig. 5) follows the cycling-DA convention: cycled DA CO_2 and CH_4 vertical profiles are validated against independent CM-01 aircraft observations not used in the assimilation. The CM-01 mole fractions are never used in any analysis update (the DA-technical sense of "independent"), and the vertical-profile structure they probe is distinct from the surface boundary-layer values that the WMO/GAW sites constrain through assimilation. For the three WMO/GAW assimilation sites, the corresponding prior-posterior-observation comparison is presented in Figure 4 (and Table 1) in the conventional observation-space fit-to-obs format, which is the appropriate diagnostic at locations where the system actively assimilates the observations. The known-truth emission-space diagnostic is provided in Supplement S2, with spatial diagnostics in Fig. S15 and recovery time series in Fig. S16.

Regarding "isolating the impact of the emission correction": we address this concern collectively in our response to Comment Sec 7.1.

Figure 9: Why showing only the posterior flux, which is still largely dominated by emission patterns present in the prior? Why not the difference with the prior?

>> Figure 9 (currently Fig. 8) shows annual mean surface concentrations, not fluxes, as stated in the figure caption. The reviewer's comment refers to content that does not appear in the figure being discussed. Even on the generous reading where the reviewer's intended question was about emission

fluxes (specifically, the assimilation increment, or "difference with the prior", that distinguishes the posterior estimate from prior emission patterns), the manuscript already provides this diagnostic: Figure 9 (Section 6.4) presents the emission assimilation increment (posterior minus prior) for both CO_2 (panel b) and CH_4 (panel d). The corresponding prior concentration comparison for Figure 8 is provided in Figure S15 of the Supplement, presenting the EGG4 annual mean surface concentrations. We also revised the surrounding text to distinguish concentration-field representation from emission-increment diagnostics.

L666-679: "Figure 9 … Figs 510-113": all of that is not really an outcome of your inversion, it's just features of your (forward) modelling setup.

>> We note that the figure reference "*Figs 510-113*" does not appear anywhere in the manuscript. The passage at L666–679 references Figures S10–S13 in the Supplement. We are unable to determine what "Figs 510-113" refers to.

The reviewer states that the results are "*not really an outcome of your inversion*" but "*just features of your (forward) modelling setup.*" This conflates two distinct capabilities: high-resolution forward transport and high-resolution inversion. Any mesoscale model can simulate fine-scale concentration gradients in forward mode. This is not what the manuscript claims as a contribution. The contribution is that the inversion system operates at 9 km resolution, meaning that observation-driven emission corrections are estimated, applied, and propagated at that resolution through 1,460 sequential analysis cycles. The posterior concentration field reflects the cumulative effect of these emission corrections on the high-resolution transport. It is not reproducible by a forward simulation with prior emissions, regardless of the model resolution. Dismissing the inversion output as "*just features of the forward model*" disregards the fundamental distinction between prior and posterior in a data assimilation system.

L679-688: "However, the structural differences ...sub-national patterns": If this is an inverse modeling setup, as in, you are trying to resolve the emissions, representing the tracer concentration at a finer resolution is not necessarily a benefit (it may help, depending on the circumstances, but it shouldn't be an aim in and of itself. This is a very long paragraph to just explain that using a high resolution transport model leads to finer representation of concentration patterns …

>> The 50-site OSSE (added at the reviewer's own Round 1 request) directly demonstrates that the high-resolution inversion framework recovers spatially structured emission signals when observational coverage is sufficient. This tells us that high resolution inversion system is beneficial

for emission estimation when the network supports it particularly for large prior-emission uncertainties.

This manuscript describes a framework. The 9 km resolution is a documented design capability, not a claim that the current 3-station network fully exploits it. The framework is designed to support future integration of denser observing systems where representation error at sub-grid scales directly affects the observation operator. Resolution effects on regional CO_2 inversion are documented in the literature, including transport-error magnitudes at fine scales (Kim et al., 2021) and resolution-induced biases in regional flux partitioning at the global scale (Liu et al., 2024). We have revised the text to separate design capability from what is constrained by the present network

L688-693: “These features … heterogeneity”: yes, coarser resolution model will smooth out fine scale features. But the question is whether it matters or not in your case. For assimilating TROPOMI data, which have a super high resolution, it certainly does (that’s probably what the authors of the studies you cited meant). For your case, with just 3 observation sites … maybe it still matters if you can show that the representation error becomes to large to efficiently assimilate these data (but you don’t step in that direction).

>> We agree that representation error is more critical for high-resolution satellite retrievals than for the current 3-station in situ network, and we appreciate the reviewer acknowledging that high resolution "certainly does" matter for TROPOMI assimilation. This is precisely the design motivation: the framework is built at 9 km to support future integration of satellite column observations (e.g., TROPOMI CH_4 at ~7 km footprint) where sub-grid gradients directly affect the averaging kernel forward operator (see Broquet et al., 2018; Santaren et al., 2021, for OSSEs explicitly developed for the satellite-CO_2 assimilation use case). The current 3-station deployment is the first demonstration, not the design target. The revised passage at L688–693 accurately attributes the fine-scale heterogeneity observations to the Korean satellite-based studies (Shim et al., 2019; Moon et al., 2024), and the adjacent revision at L693–696 introduces the forward-looking framing of the 9-km resolution together with an explicit scope caveat for the present 3-station deployment.

L693-696: “These results illustrate …”: This sentence is completely unsubstantiated. All you have shown so far in this section is that your model is capable of simulating fine scale atmospheric gradients in CO2 and CH4. But you haven’t shown whether this leads to higher quality emission increments than what you would obtain with a coarser resolution inversion (you haven’t actually shown emission increments at all, at this stage …).

>> We agree that the original sentence could be read as extending too far from concentration gradients to emission fidelity. We have revised it to state only what the section demonstrates, and we now refer separately to the emission increment results in Section 6.4 and Fig. 9. A formal multi-resolution emission comparison is beyond the scope of this model description paper.

We do not agree with the reviewer's claim that "*you haven't actually shown emission increments at all*". Previous version clearly showed the assimilation increments and Figure 9 (in the revised version) presents assimilation increments (posterior minus prior) for both CO_2 (panel b) and CH_4 (panel d) in Section 5.4. Figure 7 (in the revised version) presents flux-uncertainty reduction maps in Section 6.3.

L778-782: "By coupling meteorology and tracer transport … terrain": This statement reflects seems to reflect the thought process behind the design choices of this inversion framework, but I don't think it's supported by facts. At least it's not demonstrated in the paper that updating both the meteorology and transport provides any meaningful advantage compared to simply using prescribed meteorological data from a reanalysis (or from a reanalysis ensemble). Perhaps there are practical reasons behind this choice, and it probably doesn't really hurt, although I can imagine it somewhat limits the number of degrees of freedom of the assimilation, since it has to fit jointly meteo and tracer concentrations. Likewise, the choice of a dual-tracer inversion (CO2 + CH4) is not really justified.

>> We agree that the previous wording could be read as implying a general advantage over footprint-based Lagrangian inversions, which was not our intent and was not demonstrated by our experiments. We have therefore removed the comparative framing and revised the conclusion to present these methods as complementary. The revised text now describes the DART/EAKF properties relevant to our own framework-flow-dependent covariance, adaptive inflation, and joint cycling-while explicitly acknowledging trade-offs, including computational cost and dependence on meteorological and sub-grid transport parameterizations. The joint meteorology-tracer update is a design choice inherited from the WRF-Chem/DART architecture: DART updates all state variables in the augmented vector simultaneously, and meteorological observations are assimilated to maintain transport consistency. Regarding dual-tracer inversion: CO_2 and CH_4 are updated independently (no cross-species covariances are applied, as stated in Section 2.2). The "dual-tracer" configuration means both species are included in the state vector and cycled simultaneously, a practical design choice that avoids running separate inversion systems for species sharing the same transport, not a scientific claim requiring comparative experiments.

Formal comparison of alternative configurations would each constitute separate studies beyond the scope of this model description paper. We have revised the language to present these as design choices rather than demonstrated advantages.

We also note an unresolved editing artifact in the reviewer's comment at L778 ("*reflects seems to reflect*"), which we have interpreted according to its evident intent.

L792-803: This entire paragraph is a long list of theoretical qualities that this inversion setup would have and that would constitute advantages over “footprint-based Lagrangian inversions”, but if it’s not supported by any facts and demonstrations … it’s just pedantic. Especially since it denotes a certain lack of understanding of the GHG inverse modelling field in general (which is quite diverse and not limited to using Lagrangian models, which themselves can be employed in a wide range of manners). If you think that solving jointly meteo and tracer concentrations has an advantage over using prescribed meteorology, prove it, you have the right tool for it! If the DART inflation framework provides a quantifiable advantage over some other mechanism that would not be ″directly coupled to sequential filter dynamics”, please set up an experiment to demonstrate it, otherwise this sentence has nothing to do in the conclusions

>> We agree that the previous wording could be read as implying a general advantage over footprint-based Lagrangian inversions, which was not our intent and was not demonstrated by our experiments. We have therefore removed the comparative framing and revised the conclusion to present these methods as complementary. The revised text now describes the DART/EAKF properties relevant to our own framework-flow-dependent covariance, adaptive inflation, and joint cycling-while explicitly acknowledging trade-offs, including computational cost and dependence on meteorological and sub-grid transport parameterizations (e.g., Anderson, 2001, 2003, 2009). We also address the concern using the controlled OSSE, where the emission truth is known. The OSSE provides emission-space diagnostics, including the integrated recovery rate and recovery time series, which directly quantify whether the cycling system transfers information into the emission field rather than only correcting concentrations

References

Aksoy, A., Zhang, F., and Nielsen-Gammon, J. W.: Ensemble-based simultaneous state and parameter estimation with MM5, Geophys. Res. Lett., 33, L12801, https://doi.org/10.1029/2006GL026186, 2006.

Anderson, J. L.: An Ensemble Adjustment Kalman Filter for Data Assimilation, Mon. Weather Rev., 129, 2884–2903, https://doi.org/10.1175/1520-0493(2001)129%253C2884:AEAKFF%253E2.0.CO;2, 2001.

Anderson, J. L.: A Local Least Squares Framework for Ensemble Filtering, Mon. Weather Rev., 131, 634–642, https://doi.org/10.1175/1520-0493(2003)131%253C0634:ALLSFF%253E2.0.CO;2, 2003.

Anderson, J. L.: Spatially and temporally varying adaptive covariance inflation for ensemble filters, Tellus A, 61, 72–83, https://doi.org/10.1111/j.1600-0870.2008.00361.x, 2009.

Anderson, J. L. and Lei, L.: Empirical Localization of Observation Impact in Ensemble Kalman Filters, Mon. Weather Rev., 141, 4140–4153, https://doi.org/10.1175/MWR-D-12-00330.1, 2013.

Bisht, J. S. H., Patra, P. K., Takigawa, M., Sekiya, T., Kanaya, Y., Saitoh, N., and Miyazaki, K.: Estimation of CH_4 emission based on an advanced 4D-LETKF assimilation system, Geosci. Model Dev., 16, 1823–1838, https://doi.org/10.5194/gmd-16-1823-2023, 2023.

Bréon, F. M., Broquet, G., Puygrenier, V., Chevallier, F., Xueref-Remy, I., Ramonet, M., Dieudonné, E., Lopez, M., Schmidt, M., Perrussel, O., and Ciais, P.: An attempt at estimating Paris area CO_2 emissions from atmospheric concentration measurements, Atmos. Chem. Phys., 15, 1707–1724, https://doi.org/10.5194/acp-15-1707-2015, 2015.

Broquet, G., Bréon, F.-M., Renault, E., Buchwitz, M., Reuter, M., Bovensmann, H., Chevallier, F., Wu, L., and Ciais, P.: The potential of satellite spectro-imagery for monitoring CO_2 emissions from large cities, Atmos. Meas. Tech., 11, 681–708, https://doi.org/10.5194/amt-11-681-2018, 2018.

Brunner, D., Suter, I., Bernet, L., Constantin, L., Grange, S. K., Rubli, P., Li, J., Chen, J., Bigi, A., and Emmenegger, L.: Building-resolving simulations of anthropogenic and biospheric CO_2 in the city of Zurich with GRAMM/GRAL, Atmos. Chem. Phys., 25, 14387–14410, https://doi.org/10.5194/acp-25-14387-2025, 2025.

Bukosa, B., Mikaloff-Fletcher, S., Brailsford, G., Smale, D., Keller, E. D., Baisden, W. T., Kirschbaum, M. U. F., Giltrap, D. L., Liáng, L., Moore, S., Moss, R., Nichol, S., Turnbull, J., Geddes, A., Kennett, D., Hidy, D., Barcza, Z., Schipper, L. A., Wall, A. M., Nakaoka, S.-I., Mukai, H., and Brandon, A.: Inverse modelling of New Zealand's carbon dioxide balance estimates a larger than expected carbon sink, Atmos. Chem. Phys., 25, 6445–6471, https://doi.org/10.5194/acp-25-6445-2025, 2025.

Chang, M., Zuo, H., and Duan, J.: Modifying Covariance Localization to Mitigate Sampling Errors from the Ensemble Data Assimilation, Adv. Meteorol., 2022, 6101721, https://doi.org/10.1155/2022/6101721, 2022.

Chatterjee, A., Engelen, R. J., Kawa, S. R., Sweeney, C., and Michalak, A. M.: Toward reliable ensemble Kalman filter estimates of CO_2 fluxes, J. Geophys. Res.-Atmos., 117, D22306, https://doi.org/10.1029/2012JD018176, 2012.

Chen, H. W., Zhang, F., Lauvaux, T., Davis, K. J., Feng, S., Butler, M. P., and Alley, R. B.: Characterization of regional-scale CO_2 transport uncertainties in an ensemble with flow-dependent transport errors, Geophys. Res. Lett., 46, 4049–4058, https://doi.org/10.1029/2018GL081341, 2019.

Chen, H. W., Zhang, F., Lauvaux, T., Scholze, M., Davis, K. J., and Alley, R. B.: Regional CO_2 inversion through ensemble-based simultaneous state and parameter estimation: TRACE framework and controlled experiments, J. Adv. Model. Earth Syst., 15, e2022MS003208, https://doi.org/10.1029/2022MS003208, 2023.

Dekker, I. N., Houweling, S., Aben, I., Röckmann, T., Krol, M., Martínez-Alonso, S., Deeter, M. N., and Worden, H. M.: Quantification of CO emissions from the city of Madrid using MOPITT satellite retrievals and WRF simulations, Atmos. Chem. Phys., 17, 14675–14694, https://doi.org/10.5194/acp-17-14675-2017, 2017.

Errico, R. M., Yang, R., Privé, N. C., Tai, K.-S., Todling, R., Sienkiewicz, M. E., and Guo, J.: Development and validation of observing-system simulation experiments at NASA's Global Modeling and Assimilation Office, Q. J. Roy. Meteor. Soc., 139, 1162–1178, https://doi.org/10.1002/qj.2027, 2013.

Errico, R. M. and Privé, N. C.: Some General and Fundamental Requirements for Designing Observing System Simulation Experiments (OSSEs), NASA Tech. Rep. GSFC-E-DAA-TN69069, NASA Technical Reports Server, available at: https://ntrs.nasa.gov/citations/20190025338, 2018.

Gaubert, B., Arellano, A. F., Barré, J., Worden, H. M., Emmons, L. K., Tilmes, S., Buchholz, R. R., Vitt, F., Raeder, K., Collins, N., Anderson, J. L., Wiedinmyer, C., Martínez Alonso, S., Edwards, D. P., Andreae, M. O., Hannigan, J. W., Petri, C., Strong, K., and Jones, N.: Toward a chemical reanalysis in a coupled chemistry-climate model: An evaluation of MOPITT CO assimilation and its impact on tropospheric composition, J. Geophys. Res.-Atmos., 121, 7310–7343, https://doi.org/10.1002/2016JD024863, 2016.

Gaudet, B. J., Davis, K. J., Pal, S., Jacobson, A. R., Schuh, A., Lauvaux, T., Feng, S., and Browell, E. V.: Regional-Scale, Sector-Specific Evaluation of Global CO_2 Inversion Models Using Aircraft Data From the ACT-America Project, J. Geophys. Res.-Atmos., 126, e2020JD033623, https://doi.org/10.1029/2020JD033623, 2021.

Henne, S., Brunner, D., Oney, B., Leuenberger, M., Eugster, W., Bamberger, I., Meinhardt, F., Steinbacher, M., and Emmenegger, L.: Validation of the Swiss methane emission inventory by atmospheric observations and inverse modelling, Atmos. Chem. Phys., 16, 3683–3710, https://doi.org/10.5194/acp-16-3683-2016, 2016.

Hoffman, R. N. and Atlas, R.: Future Observing System Simulation Experiments, Bull. Am. Meteorol. Soc., 97, 1601–1616, https://doi.org/10.1175/BAMS-D-15-00200.1, 2016.

Houtekamer, P. L. and Zhang, F.: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation, Mon. Weather Rev., 144, 4489–4532, https://doi.org/10.1175/MWR-D-15-0440.1, 2016.

Hsu, C.-H., Henze, D. K., Mizzi, A. P., González Abad, G., He, J., Harkins, C., Naeger, A. R., Lyu, C., Liu, X., Chan Miller, C., Pierce, R. B., Johnson, M. S., and McDonald, B. C.: An Observing System Simulation Experiment Analysis of How Well Geostationary Satellite Trace-Gas Observations Constrain NOx

Emissions in the US, J. Geophys. Res.-Atmos., 129, e2023JD039323, https://doi.org/10.1029/2023JD039323, 2024.

Kang, J.-S., Kalnay, E., Liu, J., Fung, I., Miyoshi, T., and Ide, K.: "Variable localization" in an ensemble Kalman filter: Application to the carbon cycle data assimilation, J. Geophys. Res.-Atmos., 116, D09110, https://doi.org/10.1029/2010JD014673, 2011.

Kang, J.-S., Kalnay, E., Miyoshi, T., Liu, J., and Fung, I.: Estimation of surface carbon fluxes with an advanced data assimilation methodology, J. Geophys. Res.-Atmos., 117, D24101, https://doi.org/10.1029/2012JD018259, 2012.

Khade, V. M., Hansen, J. A., Reid, J. S., and Westphal, D. L.: Ensemble Kalman filter assimilation of OCO-2 column-averaged CO_2 retrievals: a regional carbon data assimilation system demonstration, Geosci. Model Dev., 14, 1265–1284, https://doi.org/10.5194/gmd-14-1265-2021, 2021.

Kim, J., Polavarapu, S. M., Jones, D. B. A., Chan, D., and Neish, M.: The Resolvable Scales of Regional-Scale CO_2 Transport in the Context of Imperfect Meteorology: The Predictability of CO_2 in a Limited-Area Model, J. Geophys. Res.-Atmos., 126, e2021JD034896, https://doi.org/10.1029/2021JD034896, 2021.

Kondo, K. and Miyoshi, T.: Impact of Removing Covariance Localization in an Ensemble Kalman Filter: Experiments with 10 240 Members Using an Intermediate AGCM, Mon. Weather Rev., 144, 4849–4865, https://doi.org/10.1175/MWR-D-15-0388.1, 2016.

Kou, X., Tian, X., Zhang, M., Peng, Z., and Zhang, X.: Accounting for CO_2 variability over East Asia with a regional joint inversion system and its preliminary evaluation, J. Meteorol. Res., 31, 834–851, https://doi.org/10.1007/s13351-017-7066-6, 2017.

Lauvaux, T., Miles, N. L., Deng, A., Richardson, S. J., Cambaliza, M. O., Davis, K. J., Gaudet, B., Gurney, K. R., Huang, J., O'Keefe, D., Song, Y., Karion, A., Oda, T., Patarasuk, R., Razlivanov, I., Sarmiento, D., Shepson, P., Sweeney, C., Turnbull, J., and Wu, K.: High-resolution atmospheric inversion of urban CO_2 emissions during the dormant season of the Indianapolis Flux Experiment (INFLUX), J. Geophys. Res.-Atmos., 121, 5213–5236, https://doi.org/10.1002/2015JD024473, 2016.

Lee, H., Han, S.-O., Ryoo, S.-B., Lee, J.-S., and Lee, G.-W.: The measurement of atmospheric CO_2 at KMA GAW regional stations, its characteristics, and comparisons with other East Asian sites, Atmos. Chem. Phys., 19, 2149–2163, https://doi.org/10.5194/acp-19-2149-2019, 2019.

Liu, J., Bowman, K. W., and Lee, M.: Comparison between the Local Ensemble Transform Kalman Filter (LETKF) and the 4D-Var in atmospheric CO_2 flux inversion with the Goddard Earth Observing System–Chemistry transport model and the observation impact diagnostics from the LETKF, J. Geophys. Res.-Atmos., 121, 13066–13087, https://doi.org/10.1002/2016JD025100, 2016.

Liu, X., Mizzi, A. P., Anderson, J. L., Fung, I. Y., and Cohen, R. C.: Assimilation of satellite NO_2 observations at high spatial resolution using OSSEs, Atmos. Chem. Phys., 17, 7067–7081, https://doi.org/10.5194/acp-17-7067-2017, 2017.

Liu, Y., Kalnay, E., Zeng, N., Asrar, G., Chen, Z., and Jia, B.: Estimating surface carbon fluxes based on a local ensemble transform Kalman filter with a short assimilation window and a long observation window:

an observing system simulation experiment test in GEOS-Chem 10.1, Geosci. Model Dev., 12, 2899–2914, https://doi.org/10.5194/gmd-12-2899-2019, 2019.

Liu, Z., Zeng, N., Liu, Y., Kalnay, E., Asrar, G., Wu, B., Cai, Q., Liu, D., and Han, P.: Improving the joint estimation of CO_2 and surface carbon fluxes using a constrained ensemble Kalman filter in COLA (v1.0), Geosci. Model Dev., 15, 5511–5528, https://doi.org/10.5194/gmd-15-5511-2022, 2022.

Liu, Z., Zeng, N., Han, P., and Cai, Q.: Impact of the horizontal resolution of GEOS-Chem on land–ocean and tropic–extratropic partitioning and seasonal cycle in CO_2 inversion, Environ. Res. Lett., 19, 114007, https://doi.org/10.1088/1748-9326/ad7870, 2024.

Ma, C., Wang, T., Mizzi, A. P., Anderson, J. L., Zhuang, B., Xie, M., and Wu, R.: Multiconstituent data assimilation with WRF-Chem/DART: Potential for adjusting anthropogenic emissions and improving air quality forecasts over eastern China, J. Geophys. Res.-Atmos., 124, 7393–7412, https://doi.org/10.1029/2019JD030421, 2019.

Ma, C., Wang, T., Jiang, Z., Wu, H., Zhao, M., Zhuang, B., Li, S., Xie, M., Li, M., Liu, J., and Wu, R.: Importance of bias correction in data assimilation of multiple observations over eastern China using WRF-Chem/DART, J. Geophys. Res.-Atmos., 125, e2019JD031465, https://doi.org/10.1029/2019JD031465, 2020.

Mizzi, A. P., Arellano Jr., A. F., Edwards, D. P., Anderson, J. L., and Pfister, G. G.: Assimilating compact phase space retrievals of atmospheric composition with WRF-Chem/DART: a regional chemical transport/ensemble Kalman filter data assimilation system, Geosci. Model Dev., 9, 965–978, https://doi.org/10.5194/gmd-9-965-2016, 2016.

Mizzi, A. P., Edwards, D. P., and Anderson, J. L.: Assimilating compact phase space retrievals (CPSRs): comparison with independent observations (MOZAIC in situ and IASI retrievals) and extension to assimilation of truncated retrieval profiles, Geosci. Model Dev., 11, 3727–3745, https://doi.org/10.5194/gmd-11-3727-2018, 2018.

Moon, J., Shim, C., Seo, J., and Han, J.: Evaluation of Korean methane emission sources with satellite retrievals by spatial correlation analysis, Environ. Monit. Assess., 196, 296, https://doi.org/10.1007/s10661-024-12449-w, 2024.

Nalini, K., Lauvaux, T., Abdallah, C., Lian, J., Ciais, P., Utard, H., Laurent, O., and Ramonet, M.: High-Resolution Lagrangian Inverse Modeling of CO_2 Emissions Over the Paris Region During the First 2020 Lockdown Period, J. Geophys. Res.-Atmos., 127, e2021JD036032, https://doi.org/10.1029/2021JD036032, 2022.

Peng, Z., Zhang, M., Kou, X., Tian, X., and Ma, X.: A regional carbon data assimilation system and its preliminary evaluation in East Asia, Atmos. Chem. Phys., 15, 1087–1104, https://doi.org/10.5194/acp-15-1087-2015, 2015.

Peng, Z., Lei, L., Liu, Z., Liu, J., Sun, K., Wang, J., Lin, J., Zou, M., Hu, M., and Wang, Z.: CO_2 flux inversion with a regional joint data assimilation system based on CMAQ, EnKS, and surface CO_2 observations, J. Geophys. Res.-Atmos., 128, e2023JD038834, https://doi.org/10.1029/2023JD038834, 2023.

Pisso, I., Sollum, E., Grythe, H., Kristiansen, N. I., Cassiani, M., Eckhardt, S., Arnold, D., Morton, D., Thompson, R. L., Groot Zwaaftink, C. D., Evangeliou, N., Sodemann, H., Haimberger, L., Henne, S., Brunner, D., Burkhart, J. F., Fouilloux, A., Brioude, J., Philipp, A., Seibert, P., and Stohl, A.: The Lagrangian particle dispersion model FLEXPART version 10.4, Geosci. Model Dev., 12, 4955–4997, https://doi.org/10.5194/gmd-12-4955-2019, 2019.

Roh, S., Jun, M., Szunyogh, I., and Genton, M. G.: Multivariate localization methods for ensemble Kalman filtering, Nonlin. Processes Geophys., 22, 723–735, https://doi.org/10.5194/npg-22-723-2015, 2015.

Santaren, D., Broquet, G., Bréon, F.-M., Chevallier, F., Siméoni, D., Zheng, B., and Ciais, P.: A local- to national-scale inverse modeling system to assess the potential of spaceborne CO_2 measurements for the monitoring of anthropogenic emissions, Atmos. Meas. Tech., 14, 403–433, https://doi.org/10.5194/amt-14-403-2021, 2021.

Shim, C., Han, J., Henze, D. K., and Yoon, T.: Identifying local anthropogenic CO_2 emissions with satellite retrievals: a case study in South Korea, Int. J. Remote Sens., 40, 1011–1029, https://doi.org/10.1080/01431161.2018.1523585, 2019.

Sijikumar, S., Raju, A., Valsala, V., Tiwari, Y., Girach, I. A., Jain, C. D., and Ratnam, M. V.: High-Resolution Bayesian Inversion of Carbon Dioxide Flux Over Peninsular India, Atmos. Environ., 308, 119868, https://doi.org/10.1016/j.atmosenv.2023.119868, 2023.

Super, I., Denier van der Gon, H. A. C., van der Molen, M. K., Sterk, H. A. M., Hensen, A., and Peters, W.: A multi-model approach to monitor emissions of CO_2 and CO from an urban–industrial complex, Atmos. Chem. Phys., 17, 13297–13316, https://doi.org/10.5194/acp-17-13297-2017, 2017.

Vardag, S. N. and Maiwald, R.: Optimising urban measurement networks for CO_2 flux estimation: a high-resolution observing system simulation experiment using GRAMM/GRAL, Geosci. Model Dev., 17, 1885–1902, https://doi.org/10.5194/gmd-17-1885-2024, 2024.

Zhang, Q., Li, M., Wei, C., Mizzi, A. P., Huang, Y., and Gu, Q.: Assimilation of OCO-2 retrievals with WRF-Chem/DART: A case study for the Midwestern United States, Atmos. Environ., 246, 118106, https://doi.org/10.1016/j.atmosenv.2020.118106, 2021a.

Zhang, Q., Li, M., Wang, M., Mizzi, A. P., Huang, Y., Wei, C., Jin, J., and Gu, Q.: CO_2 flux over the contiguous United States in 2016 inverted by WRF-Chem/DART from OCO-2 XCO_2 retrievals, Remote Sens., 13, 2996, https://doi.org/10.3390/rs13152996, 2021b.