Dear Dr. Western and Reviewer

Thank you for handling our manuscript and for arranging this additional round of review. We are grateful to the editor and reviewer for a careful and constructive assessment.

The central recommendation, that the single perfect-model OSSE be expanded into a set of experiments demonstrating the influence of the error sources that affect a real inversion, is well taken, and we have substantially expanded the evaluation in response. The OSSE supplement has been reorganized so that the new error-source matrix is presented first (Sect. S1, Table S2, Figs S15–S19), followed by the retained known-truth dense-network experiment, which is presented separately as an idealized information-content diagnostic for expanded observational coverage.

In brief, we have (1) added a controlled OSSE matrix of seven experiments that introduce transport-model error, boundary-condition bias, observation-error/representativeness sensitivity, two localization radii, and their combination, each against a known-truth, error-free baseline; (2) diagnosed the adaptive inflation, ensemble-spread behavior across the OSSE matrix, and assessed consistency of ensemble size and spread under real-case configurations; (3) characterized the species-dependent recovery together with its mechanism; (4) re-grounded the real-data CH_4 attribution by imposing a boundary-condition bias of known size and reporting the resulting species-specific signature; and (5) made the scope, and the basis for each deferred choice explicit. Importantly, we now discuss the system's remaining uncertainties more explicitly and show that the current three-station network provides strongest constraints within station footprints, whereas expanded observational coverage would improve domain-wide emission constraints. We present this not as a claim of complete operational robustness, but as a transparent assessment of the present system and its most important path for improvement.

In revising the OSSE evaluation, we have used Bisht et al. (2023) and Michalak et al. (2017) as important benchmarks for the type of diagnostic coverage expected in greenhouse-gas inversion studies, and we have carried out an expanded set of controlled experiments to characterize the system. To make the comparison with that work as transparent as possible, we have aligned our terminology with theirs, describing the evaluation as observation system simulation experiments in which a known emission distribution (the truth) is retrieved from synthetic observations.

The present system is an online-coupled ensemble framework based on DART, different in scale and formulation from the global 4D-LETKF of Bisht et al. (2023). Although the generic behavior of DART/ensemble filters has been extensively studied, we do not use those studies as substitutes for evaluating the present system. Instead, we use them to justify which filter-configuration sensitivities are most relevant and then provide system-specific diagnostics for inflation and ensemble size. Therefore, we have matched their work at the level of error-source coverage rather than by replicating its exact axis list, prioritizing the transport error and observation-error/representativeness sensitivities as major contributors to regional-inversion uncertainty, adding a boundary-condition experiment motivated by this study's own CH_4 profile discrepancy, and including a localization sweep, a combined-error experiment, and a real-case comparison of three adaptive-inflation configurations. To clarify how the revised manuscript and Supplement now evaluate the system, we summarize the main evaluation components below.

<table>
<tr><th>#</th><th colspan="2">Methodology description</th><th>Results</th></tr>
<tr><td>1</td><td colspan="2">Posterior CO_2 and CH_4 concentration accuracy (bias and RMSE) at the GAW stations</td><td>Table 1</td></tr>
<tr><td>2</td><td colspan="2">Posterior CO_2 and CH_4 vertical concentration profiles compared with airborne measurements</td><td>Fig. 5</td></tr>
<tr><td>3</td><td colspan="2">Uncertainty reduction (ensemble spread) of a posterior CO_2 and CH_4 concentrations in the whole model domain</td><td>Fig. 6</td></tr>
<tr><td>4</td><td colspan="2">Uncertainty reduction (ensemble spread) of a posterior CO_2 and CH_4 emissions in the whole model domain</td><td>Fig. 7</td></tr>
<tr><td>5</td><td colspan="2">Evaluation of the annual national total emission against inventory data</td><td>Fig. 10</td></tr>
<tr><td>6</td><td rowspan="4">OSSE evaluation under prescribed error sources considering 30% uncertainty in a prior emission with a 3-site (WMO/GAW) observation network</td><td>Estimation of recovery rate</td><td>Table S2, Fig. S15, Fig. S16</td></tr>
<tr><td>7</td><td>Quantification of emissions uncertainty by transport error (PBL parameterization which has been known as a major contributor to dispersion uncertainty)</td><td>Table S2, Fig. S15, Fig. S16</td></tr>
<tr><td>8</td><td>Quantification of emissions uncertainty by CO_2 and CH_4 background concentration bias</td><td>Table S2, Fig. S15, Fig. S16</td></tr>
<tr><td>9</td><td>Quantification of emissions uncertainty by observation error</td><td>Table S2, Fig. S15, Fig. S16</td></tr>
</table>

10		Sensitivity test of localization radius	Table S2, Fig. S15, Fig. S16, Fig. S18
11		Diagnostic of adaptive prior-inflation behavior (filter stability; inflation stays modest and self-regulating)	Fig. S19
12	OSSE evaluation of expanded observing-network density (50% prior emission uncertainty) using a 50-site dense synthetic network (CO_2 only)	Comparison with the OSSE with a 3-site observation network: recovery rate	Fig. S20, Fig. S21
13		Comparison with the OSSE with a 3-site observation network: emission error	Fig. S22
14	Real case test of ensemble size sensitivity		Fig. S23
15	Real-case inflation-configuration sensitivity and RMSE/total-spread consistency		Fig. S24

Below we quote each of Reviewer #3's comments in full and respond to each in turn.

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 #3: ‘Comment on egusphere-2025-4938’, Anonymous referee #3

>> In responding to the recommendation, we have aimed to test the error sources that can be meaningfully isolated within a controlled observation system simulation experiment (OSSE) framework, in which a known emission distribution (the truth) is retrieved from synthetic observations, at the present stage of the system, to diagnose from the existing model output the filter behavior the reviewer raises, and to be explicit about the small number of tests that remain genuinely premature. The retained dense-network experiment additionally probes the information content of a denser observing system (three against fifty stations). The reviewer's comments are quoted verbatim and in the order raised.

The reviewer questioned the appropriateness of the joint state-parameter estimation method for inferring surface greenhouse gas fluxes. The authors present a strong case for their method, citing appropriate literature, and have clearly stated the assumptions and limitations of this method in the manuscript. In my opinion, they show the inversion method to be defensible and are sufficiently transparent about its strengths and weaknesses.

>> We are grateful for this assessment. The method description, its stated assumptions, and its limitations are unchanged in the present revision; our response below concerns only the OSSE evaluation that the reviewer asks us to strengthen.

The discussion of the OSSE and its interpretation, however, warrants more discussion. In the OSSE, the authors perturb the prior emissions from the fossil fuel sector by removing the power industry component. The full EDGAR prior is used to generate synthetic observations, and all other model parameters are unchanged, except that two different configurations of measurement site are used. In the first of the OSSEs, the same three-site network is used as in the real data experiment, whereas in the second, a 50-site network spanning much of the inversion domain is used. The 50-site case is an example of an idealized perfect-model diagnostic, designed primarily to verify the algorithmic behaviour of the inversion system. The authors state that "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", and repeatedly frame the OSSE as not intended to represent real-world conditions.

The reviewer questions the scope of this OSSE and its ability to determine the capacity of the system to resolve real-world emissions estimates. I tend to agree — if the experiment is not designed to test *the real-world factors that impact inversion performance, then little confidence can be placed in the results of the real-world experiment using that same inversion system. The authors make a valid point that "no single OSSE can characterize real-world performance, nor is that the standard expectation for OSSEs in the literature", and indeed, the perfect-model diagnostic is a crucial first step in model validation. Such an experiment, however, should be supported by more extensive testing. The existing OSSE does not demonstrate "real-world robustness to transport-model error, boundary-condition bias, or representativeness error", and robustness to these factors is essential for a usable model. In my view, the single OSSE presented here does not meet the GMD criterion that "sufficient verification and evaluation must be included to show that the model is fit for purpose and works as expected".*

>> As the reviewer suggested, we have responded with a set of controlled experiments rather than a single perturbation. The OSSE supplement has been reorganized so that the new error-source matrix is presented first (Sect. S1, Table S2, Figs S15–S19), followed by the retained known-truth dense-network experiment that is presented as a separate idealized information-content diagnostic for expanded observational coverage.

Each experiment introduces one prescribed error of known magnitude against a known target, so that the change in recovered emissions isolates the system's response to that error (Sect. S1.2, Table S2). We note that Bisht et al. (2023) likewise describe their evaluation not as a perfect-model exercise but as observation system simulation experiments in which a known surface emission distribution (the truth) is retrieved from synthetic observations, with error sources introduced through sensitivity tests. Our revision follows the same conception: the known-truth retrieval (CTL) provides the baseline, and the transport, boundary, representativeness, and localization perturbations are introduced as controlled departures from it. Together, these experiments demonstrate that the system transfers concentration innovations into coherent and bounded emission increments, and that it responds in the expected direction and within a bounded range to the three error sources emphasized by the reviewer as important for evaluating a usable inversion system: transport-model error, boundary-condition bias, and representativeness error, together with their combination.

While this is indeed a model description paper, and not designed to cover every possible error configuration, I would like to see a set of OSSEs testing the system's ability to recover emissions across a broader range of scenarios. This brings the study in line with standard practice in the field. In a recent GMD study, Bisht et al. (2023) use a local ensemble transform Kalman filter to estimate methane emissions. They use multiple OSSE configurations to test the performance of the model under three covariance inflation methods across sets of synthetic observations from two platforms. They also use further tests to assess the model's sensitivity to the ensemble size and the assimilation *window. I would expect the current study to test their model setup with a similar degree of rigour. Furthermore, a review by Michalak et al. (2017) highlights the role of OSSEs as being able to "quantify the magnitude and geographical distribution of uncertainty that stems from specific errors or assumptions in the inversion framework, such as transport model errors, spatiotemporal flux patterns within regions, biased priors, flux spatiotemporal resolutions, or parameter choices within computational data assimilation systems". They can also be used to determine the sensitivity of inversions to transport errors, model resolution and averaging periods. The present study makes limited use of these capabilities.*

>> We have used Bisht et al. (2023) and Michalak et al. (2017) as important benchmarks for the type of diagnostic coverage expected in greenhouse-gas inversion studies. We note, respectfully, that these works are not directly transferable as experimental templates: Bisht et al. (2023) employ global LETKF, whereas the present system is an online, sequential EAKF coupled to WRF-Chem, so several of their specific configuration choices do not map onto ours one-to-one. We therefore engage them at the level of error-source coverage and rigor, rather than by replicating their exact axis list. The matrix spans the transport and representativeness errors that Michalak et al. (2017) identify as leading contributors to regional-inversion uncertainty, together with boundary-condition error motivated by this study's own finding (Section 6.2). It also adds a localization sweep and a combined-error experiment, and is situated within the ensemble-based flux-inversion literature it builds on (Liu et al., 2019; Peng et al., 2023) and the transport-error, measurement-bias, and network studies that motivate the individual experiments (Masarie et al., 2011; Lauvaux et al., 2016; Deng et al., 2017).

The three further tests the reviewer draws from Bisht et al. (2023) - the inflation method, the ensemble size, and the assimilation window - are filter-configuration choices rather than physical error sources in the real atmosphere. We therefore treat them separately from the transport, boundary-condition, and observation-error perturbations in the OSSE matrix. For inflation, the OSSE matrix and the real-data system address complementary aspects. Within the OSSE, we diagnose the behavior of the deployed adaptive inflation under each prescribed error source: it stays modest and self-limiting in every experiment (within about 5 % of unity for CH_4 and at most about 1.19 for CO_2 with no evidence of ensemble collapse; Fig. S19). This suggests that the tested OSSEs are not strongly under-dispersive. We therefore interpret the inflation results as a filter-stability diagnostic rather than as an optimization of the inflation scheme, while recognizing from Bisht et al. (2023) that inflation sensitivity is an important component of ensemble-based greenhouse-gas inversion evaluation. In the real-data setting, we compare three configurations of the adaptive inflation implemented in DART (Sect. S3.2) and find the ensemble-spread consistency at the observation sites to be nearly identical across them (Fig. S24). We do not interpret this as an optimization of a universal inflation factor, because posterior emission increments can still depend on emission-state spread and concentration–emission cross-covariances. The ensemble size is treated in the same two ways: the localization sweep diagnoses the sensitivity of recovered emissions to covariance tapering and spurious long-range correlations, while the real-case comparison of 10-, 20-, and 30-member ensembles provides a practical diagnostic of ensemble-size sensitivity for the present observing configuration and supports the 20-member configuration as a computationally feasible compromise (Sect. S3.1, Fig. S23), with a dedicated member-count scan matched to the final observing configuration deferred until the forward operators for multi-platform data, including satellite retrievals such as those from TROPOMI, are in place. None of these is presented as fully resolved.

The assimilation window differs in kind between our system and that of Bisht et al. (2023), because the two generate and constrain their ensembles differently. In their global 4D-LETKF, the meteorology is nudged to reanalysis and lies outside the assimilation, and the ensemble is formed by perturbing only the prior flux, with no perturbation of the initial CH_4 concentration; the concentration spread therefore develops only as the perturbed fluxes propagate through transport, and the flux is constrained through the resulting time-lagged covariances accumulated over an 8-day window. The window is therefore a smoother parameter that can be varied more directly than in an online coupled meteorology–chemistry cycling system, since the meteorology is fixed by the nudging and a longer window mainly allows more flux signal to accumulate in concentration space.

Our framework has two distinct points. Its forecast ensemble is generated by perturbing the chemical initial state, the lateral boundary conditions, and the anthropogenic fluxes (Sect. 2.2) within a perturbed meteorological ensemble, and the meteorology is part of the augmented state, re-analyzed jointly with the trace gases every cycle from in situ CO_2 and CH_4 and conventional PREPBUFR observations, following the online-coupled, joint-assimilation design of Kang et al. (2011, 2012). Flow-dependent concentration spread and concentration–flux covariance are therefore present at every analysis, although a longer-window smoother could still provide additional time-lagged information under a different observing-system design. The analog of Bisht's window here is the 6-h cycling interval, tied to the update frequency of the driving meteorological analyses and to the diurnal cycle of the boundary layer, with each analysis assimilating only the observations within its ±3 h window. Because the meteorological analysis is produced on this cycle, the window is not an independent flux-smoothing parameter that can be lengthened in isolation. Widening the observation window while retaining frequent analyses would move the filter toward a smoother-like configuration and would require explicit treatment of observation reuse, temporal error correlation, and information double-counting (Liu et al., 2019, 2022). The temporal integration that a long smoother window would provide is instead accomplished in our framework by sequential accumulation across cycles, with each observation assimilated once and not reused in adjacent analyses, as reflected in the domain-wide emission convergence over the cycling window (Figs. S15, S21). We do not present this as equivalent to a multi-day smoother, which constrains within-window states retrospectively through time-lagged covariances; single-use cycling is a conservative and internally consistent design for the present coupled online system whose analysis interval is set by the meteorological update cycle. The present diagnostics suggest that observational coverage is a more immediate limitation than the temporal assimilation setup in the current three-station configuration. Recovery is limited by observational coverage rather than by the temporal assimilation setup (Sect. S1.4), and the spread-consistency diagnostics of Section S3.2 (ratios near unity) indicate no severe under-dispersion at the current setting: CH_4 is close to consistency, whereas CO_2 is mildly over-dispersive in this diagnostic.

We therefore retain the operational 6 h setting in this revision and identify a systematic window-length comparison as future work, most naturally once observation types with different temporal sampling, such as satellite retrievals, enter the observing system.

In sum, the axes raised by the reviewer are now either tested directly, diagnosed for the present configuration in the OSSE and real-case simulation, or explicitly identified as future work where a full sensitivity experiment would require a different observing-system design: transport, boundary, and representativeness error through the TR, BC, and OE experiments; the inflation through an OSSE stability diagnostic (Fig. S19) and a real-case comparison of three adaptive-inflation configurations (Sect. S3.2, Fig. S24); the ensemble size through the localization sweep and a real-case 10-, 20-, and 30-member comparison (Sect. S3.1, Fig. S23); the assimilation window through diagnosis at the operational cycling interval, justified by established online-coupled practices (e.g., Kang et al., 2011, 2012); the prior-flux specification through the two prior-error structures, a spatially uniform scaling (Sect. S1.1) and a concentrated single-sector removal (Sect. S2). Beyond that list, the matrix adds a localization-radius sweep and a combined-error experiment that appears modestly sub-additive relative to the individual perturbation (Table S2).

In general, the system described in this manuscript appears promising, and the real-world results align well with inventory data. However, the limitations described above are highlighted in the interpretation of this real-world experiment, discrepancies (e.g. in CH_4 vertical profiles) are attributed to boundary condition bias or missing emissions. Since the inversion system has not been tested against such biases, it is unclear how sensitive the system is to these factors, and whether alternative configurations could improve performance.

>> We have now tested it. The boundary-condition experiment imposes a −30 ppb CH_4 background bias of known size; it makes that experiment the weakest-recovering CH_4 experiment in the near-

station recovery, with a far weaker CO_2 response (Sect. S1.3, Fig. S17), a species-specific signature consistent with the real-data discrepancy. Because the bias magnitude is prescribed, the experiment provides a controlled scale for the sensitivity, at the level of a few percent of the prior inventory under the tested configuration. We present this as a consistency check that supports the plausibility of the boundary-condition interpretation and shows that a boundary bias of the prescribed sign and magnitude can produce a CH_4-specific signature consistent with the real-data discrepancy. We do not present this as a unique attribution, nor as excluding possible missing-emission or transport-error contributions.

The remaining comments mostly concern the interpretation of the results of the OSSE. In my opinion, the interpretation is sound, as the authors have (correctly) refrained from making general comments about the inversion to perform adequately. However, this reinforces the point that an idealised OSSE alone cannot demonstrate model suitability without additional testing. Further testing, such as those examining the impact of transport model error, boundary condition bias and incorrect prior flux specification across multiple sectors/times/regions, would alleviate these concerns.

>> We retain that posture. The matrix is a controlled verification of the assimilation's behavior under prescribed errors; it does not estimate the magnitude of those errors in the real atmosphere, and it is not generalized to other seasons, regions, inventories, or networks (Sect. S1). The recovery is strong where the network observes (recovering about 60 % of the imposed bias for CH_4 and 65 % for CO_2 within 25 km of a station) and modest in the domain mean, the latter reflecting a limitation of the present observing configuration rather than of the analysis algorithm alone. Consistent with this, degrading observation precision (the OE experiment) does not degrade the domain-mean recovery; its effect is negligible for CH_4 and marginally favorable for CO_2 (Table S2), suggesting that, under the tested observation-error perturbation, network coverage is the dominant limitation for domain-mean recovery.

On the prior-flux specification across multiple sectors, times, and regions, the supplement now tests two distinct prior-error structures: a spatially uniform scaling (the matrix) and a concentrated single-sector removal (the dense-network experiment), which bracket a smoothly biased prior and a structurally biased one. A finer sector-, region-, and time-resolved decomposition is left to future work, because spatially explicit inventories differ appreciably in how they allocate emissions, so a sector-resolved result would be specific to the inventory chosen rather than a property of the assimilation.

In conclusion, I am confident that this work could warrant publication in the future, but only following a more rigorous evaluation of this model's performance. As a result, I recommend further major revisions to this manuscript. Given that the manuscript has already been through two rounds of revision, the editor may wish to consider rejection with encouragement to resubmit.

>> We have sought to address the reviewer's suggestions to the fullest extent that is scientifically meaningful within the present online, sequential EAKF framework. The revised Supplement now includes controlled OSSEs for the dominant real-application error sources, a separate dense-network information-content experiment, and real-case diagnostics for ensemble size and inflation configuration. These experiments show that the system produces bounded and interpretable emission responses under the tested perturbations, while also identifying the conditions under which recovery is degraded and the value of expanded observation coverage. We do not present the results as proof of general operational robustness or grid-scale emission resolvability; rather, they provide a limitation-aware verification suite for the current model-description paper and define the remaining priorities for future development, including assimilation-window sensitivity and broader observing-system configurations.

We are grateful for the care the reviewer and editor have given the manuscript across these rounds.

References

Bisht, J. S. H., Patra, P. K., Takigawa, M., Sekiya, T., Kanaya, Y., Saitoh, N., and Miyazaki, K.: Estimation of CH4 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.

Deng, A., Lauvaux, T., Davis, K. J., Gaudet, B. J., Miles, N., Richardson, S. J., Wu, K., Sarmiento, D. P., Hardesty, R. M., Bonin, T. A., Brewer, W. A., and Gurney, K. R.: Toward reduced transport errors in a high resolution urban CO_2 inversion system, Elem. Sci. Anth., 5, 20, https://doi.org/10.1525/elementa.133, 2017.

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. (2012): Estimation of surface carbon fluxes with an advanced data assimilation methodology, J. Geophys. Res. Atmos., 117, D24101, doi:10.1029/2012JD018259.

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.

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

Masarie, K. A., Pétron, G., Andrews, A., Bruhwiler, L., Conway, T. J., Jacobson, A. R., Miller, J. B., Tans, P. P., Worthy, D. E., and Peters, W.: Impact of CO_2 measurement bias on CarbonTracker surface flux estimates, J. Geophys. Res.-Atmos., 116, D17305, https://doi.org/10.1029/2011JD016270, 2011.

Michalak, A. M., Randazzo, N. A., and Chevallier, F.: Diagnostic methods for atmospheric inversions of long-lived greenhouse gases, Atmos. Chem. Phys., 17, 7405–7421, https://doi.org/10.5194/acp-17-7405-2017, 2017.

Peng, Z., Kou, X. X., Zhang, M. G., Lei, L. L., Miao, S. G., Wang, H. M., Jiang, F., Han, X., and Fang, S. X.: CO_2 flux inversion with a regional joint data assimilation system based on CMAQ, EnKS, and surface observations, J. Geophys. Res.-Atmos., 128, e2022JD037154, https://doi.org/10.1029/2022JD037154, 2023.