## General Comments

This manuscript presents an urban-scale CO2 flux data assimilation and inversion system named FEISSO-Carbon, based on the FLEXPART Lagrangian transport model and the FLEXINVERT+ Bayesian framework, with applications over three Chinese cities (Weifang, Chengdu, and Xining). While the system appears technically operational, I have major concerns regarding the methodological rigor, system robustness, and scientific novelty of this work. The core modules invoked in FEISSO (FLEXPART, FLEXINVERT+) are well-established pieces of community software. The authors fail to articulate any substantive model developments or algorithmic innovations uniquely contributed by this study. Furthermore, several critical configurations of prior uncertainties and boundary conditions lack physical justification, the assimilation results present severe internal contradictions, and the open-source dataset contains noticeable anomalies. Therefore, the manuscript does not currently meet the high standards of GMD and requires a Major Revision.

## Major Comments

1. The authors claim that existing urban-scale inversions in China are heavily restricted to Observing System Simulation Experiments. This is factually incorrect. Numerous real-data inversion studies utilizing satellite and surface monitor networks have been published over major Chinese conglomerates. The literature review is inadequate.
2. Carbon flux inversion is a highly researched field, and the core models employed in this study are well-established, widely used community tools. Consequently, the specific scientific novelty of this manuscript remains unclear. The authors must explicitly articulate their original contributions, clearly detailing any methodological improvements, algorithmic advancements, or specific code developments they have made to these existing models.
3. In the methodological description, the authors artificially separate their "Bayesian framework" from EnKF and 4D-Var methods. Fundamentally, both EnKF and 4D-Var are mathematical implementations within the Bayesian inversion paradigm. The authors should avoid conceptual confusion and refrain from overstating the superiority of their method by mischaracterizing established assimilation frameworks.
4. The authors state that they used the 0.25° NCEP FNL operational analysis data to drive FLEXPART, while extracting SRRs at a refined spatial resolution of 0.1°. From a fluid dynamics perspective, this is highly problematic. Smooth 0.25°

meteorological fields cannot resolve urban-scale localized circulations or complex micro-topography. This configuration is merely a high-frequency mathematical resampling (interpolation) of air parcel trajectories rather than a representation of high-resolution physical transport. Additionally, paradoxically, the authors state later that they ran the WRF model to simulate high-resolution (0.1°) hourly meteorological fields to drive their biospheric model (VPRM). Why was this fine-scale WRF meteorological output not directly utilized to drive FLEXPART?

5. The extraction of background mixing ratios from three specific upper atmospheric pressure levels (down to 48.282 hPa, well into the stratosphere) lacks robust physical grounds for boundary layer inversions. More critically, the GEOS-Chem simulation period utilized for background fields exhibits a multi-year temporal mismatch with the study's actual assimilation window (2023). Atmospheric CO2 possesses strong interannual growth trends and pronounced vertical gradients. Forcing an unadjusted, historical, high-altitude background into a 2023 boundary-layer inversion will introduce severe, systematic biases into the posterior flux estimates. Standard protocol requires sampling the 3D background fields at the exact spatio-temporal coordinates where the backward trajectories terminate. Why was this bypassed?
6. Furthermore, please justify the scientific meaning of sub-pixel inversion, where the observation resolution is coarser than the emission grid. The 1-km spatial structures in the posterior fields are overwhelmingly dictated by the prior spatial patterns rather than any gradient constraints provided by the satellite.
7. Given that ODIAC natively provides high-resolution monthly global gridded emissions, why did the authors choose to execute a complicated spatial redistribution of EDGAR totals instead of directly deploying ODIAC as the prior emissions in the nested domain?
8. The prior flux error (FlxErr/GlobErr) was set at ~10%. This severely violates emissions inventory uncertainty standards. National-scale total inventory uncertainties frequently exceed 10%. At a hyper-local 1-km grid scale, uncertainties routinely exceed 100% or 200% due to point-source misallocations. What is the statistical or physical justification for a 10% grid-level uncertainty?
9. Within a 10-day window, valid OCO-2 overpasses occur only 1 to 3 times over a specific city. Given that urban carbon emissions and boundary layer heights undergo intense diurnal and hourly cycles, back-calculating a 10-day total budget from 1-3 instantaneous snapshot measurements is statistically unrepresentative and highly susceptible to random meteorological perturbations on the overpass days.
10. a. In Figure 4b (Chengdu), as ObsErr or GlobErr increases, the total posterior emission exhibits a non-monotonic decreasing then increasing behavior. In a

Bayesian inversion, such a parabolic response to error scaling is highly irregular. Across the entire sensitivity suite, the posterior fluxes exhibit massive volatility under different parameter combinations. Rather than demonstrating system stability, this proves that the inversion system lacks statistical robustness and is highly un-robust.

b. Regarding L410, under sparse data constraints, increasing SigTime forces the system to aggressively fit the sparse observations across adjacent days. Figure 4 shows that the posterior totals never stabilize or converge. This confirms that the system is operating in an unsaturated, unstable regime.

11. Page 18, the authors explicitly state that "Chengdu is a net sink" based on the flux distributions. However, Figure 4b clearly displays that all sensitivity runs for Chengdu's 10-day total emissions are entirely positive (ranging from 0.4 to 1.2 Mt), representing a net carbon source. This is a severe contradiction between the text and the figures.
12. Page 24, the authors claim that Chengdu's complex topography causes the forward model to overestimate atmospheric CO2 columns. In data assimilation logic, to compensate for simulated concentration overestimation, the inversion must decrease the emissions. Yet, Figure 7b shows that FEISSO's optimized emissions for Chengdu are the highest among all inventories and statistical totals. This is also a contradiction. Page 25, the authors claim that Chengdu's emissions are insensitive to observations. However, the results in Figure 4b indicate that varying the observational or prior error bounds alters the posterior total emissions by more than a factor of two.
13. The column-averaged XCO2 is a highly blended, composite signal of background air, anthropogenic plumes, and biospheric net ecosystem exchange. Without auxiliary chemical tracers (e.g., CO, NO2) or isotopic measurements, please explain how the assimilation framework cleanly separates fossil fuel signals from biospheric signals from a single integrated XCO2 value.
14. In Section 4.2, the authors list a set of recommended parameters in Table 4. Simply observing highly volatile sensitivity plots is entirely insufficient to define these parameters as optimal. The authors conducted no statistical consistency tests (e.g., Chi-square) to verify the posterior error allocations, nor did they use independent data for cross-validation. Furthermore, Equation 7 for calculating globerr on Page 22 appears entirely arbitrary, lacking any derivation, physical scaling arguments, or literature backing.
15. The posterior emission fields undergo absolutely no independent evaluation against third-party observations (e.g., surface networks, eddy covariance flux towers, or aircraft profiles). The authors argue that because their optimized inventory aligns

closer with the "local inventory," the system performance is improved. This is a circular argument: what is the spatial structure of this local inventory, and how is its accuracy established? If the local inventory contains inherent systematic biases, this closer agreement would actually indicate a failure of the inversion.

16. I strongly recommend that the authors include spatial maps of the prior emissions and the posterior-minus-prior spatial increments alongside the bar charts in Figure 7. This will allow readers to visually inspect where and how the data assimilation system adjusts the emission structures.

**Technical and Minor Comments**

1. The conceptual architecture in Figure 1 contains a topological error: the "Satellite-based remote sensing observations" box has no direct arrow pointing to the "CO2 flux distribution inversion" module.
2. The geographical boundaries and spatial extent of the "outer domain" are never specified. Please specify the exact boundaries of this outer domain.
3. Page 21, the authors state that local emissions dominate the satellite signal. This reflects a fundamental misconception of satellite remote sensing physics. The absolute magnitude of an XCO2 measurement is overwhelmingly (>99%) dominated by the global background concentration. Local emissions merely induce minute enhancements (typically 0.5 to 3 ppm) in the residual increment signal, not the total satellite signal.
4. The term "global" is loosely defined throughout the text. Please standardize the terminology.
5. Figure 5 Legibility: The text labels and axes in the flux frequency distribution histograms are extremely small and overlap with each other, rendering the plots completely unreadable
6. On Page 23, the biospheric contribution percentages are written as "-5.26%"?
7. The definition of "super observations" in Section 2.2.2 violates basic geometric logic. The physical footprint of an OCO-2 single footprint is already coarser than the target inversion grid cell. Statistically, it is impossible to have multiple independent satellite observations residing inside a single 1-km grid cell to perform an outlier filtering and extract a median. This is simple resampling/data filtering rather than true super observations which requires multi-source aggregation for noise reduction.

**Critical Data Quality Issue**

The authors deposited their demonstration dataset on Zenodo. However, an inspection of the repository reveals a substantial presence of extreme outliers and physically impossible values. The gridded spatial distributions exhibit structural chaotic noise. The authors must meticulously audit and correct their data files immediately.