Overall Assessment

This preprint develops a hyperparameter-optimized Bayesian atmospheric inversion framework for urban-scale CO_2 emission estimation, addressing a long-standing limitation in traditional atmospheric inversion: error covariance matrices ***B*** for prior flux errors and ***R*** for observation errors) are mostly prescribed empirically, including manually fixed correlation-length hyperparameters. The topic is timely, the methodological framing is clear, and the two-stage evaluation design (pseudo-observation experiments with known "truth" followed by real-observation inversion over Zhengzhou) is appropriate and well conceived. Nevertheless, several scientific interpretations, mechanism explanations and supporting analyses need improvement. I recommend minor revision before it can be considered for publication.

Major Comments

1. The authors report negative sums of off-diagonal entries in the posterior covariance matrix ***A*** for the optimized method, attributing this feature to statistical negative correlation between high-emission hot-spot grids and surrounding low-flux grids. Further clarification is required. Please distinguish between *statistical artefacts from numerical optimization* versus *physically-driven flux covariance induced by atmospheric transport*. It is suggested to discuss whether this negative-sum feature remains stable under different observational network densities, to prevent readers from directly interpreting purely statistical features as real physical correlation among surface fluxes.

2. EDGAR v7.0 is adopted as the prior inventory for real-data inversion. EDGAR has relatively coarse spatial resolution and considerable local biases within urban domains. Although the authors mention inventory quality as a limiting factor for absolute accuracy, quantitative sensitivity analysis is missing. Please clarify whether the proposed hyperparameter-optimized covariance framework can compensate for large systematic biases embedded in prior inventories. It is important to explicitly state that improvements mainly originate from optimized error-covariance structures, rather than

the capability to eliminate large prior-inventory biases, so as to define the method's boundary of applicability. If possible, please also provide experiments and discussions between EDGAR and MEIC emission inventory.

3. A 24-h moving-window daily minimum is used to determine background CO_2 mixing ratios. This approach is vulnerable to advected up-wind non-local emission signals. The limitation is briefly mentioned but lacks supporting diagnostics. Please discuss how background-estimation errors propagate into posterior emissions and contribute to total relative uncertainty. State the applicable conditions of this background scheme and potential bias directions under strong long-range transport.

4. The optimal correlation-function combination varies across synthetic scenarios (BG-BG for 1-high-emission case, EN-EN for 4-high-emission case, etc.). For practical real-world applications over a new city without known "true fluxes", how should users select the best combination? Current selection relies either on synthetic truth or observation-simulation RMSE. Please note that minimal RMSE does not guarantee statistically correct covariance structure. Discuss risks associated with this selection criterion and provide practical guidance for operational usage.

5. This work builds upon the maximum-likelihood-estimation framework of Michalak et al. (2005) and Wu et al. (2013). Please more explicitly articulate the added value of this study beyond removing diagonal-matrix constraints and freeing fixed hyperparameters. Emphasize methodological improvements including the two-stage hybrid optimization (ISRES + MMA) and full preservation of both spatial and temporal correlation structures. Avoid giving readers the impression that this is merely a minor tuning of existing MLE inversion tools.

6. Even with hyperparameter-optimized covariance matrices, posterior-simulated versus observed CO_2 concentrations still show RMSE around 10 ppm in real-data runs. Please decompose major contributing error sources: WRF-STILT transport errors,

prior-inventory biases, background uncertainties and instrumental noise. Distinguish error components reducible by the present framework versus irreducible ones, and objectively evaluate the real-world performance gain of the proposed approach.

7. Section 2.1.2 states that "the full set of 25 theoretical combinations (5 × 5) were not implemented due to numerical stability issues," and that "the excluded 13 combinations produced negative values ... for the determinant." This is confusing, because the design described in the same paragraph uses three functions for R and four for B, giving 12 combinations. If 12 were retained and 13 excluded, the total is 25, which is consistent with 5 × 5, but then R and B must each have five candidate functions, contradicting the statement that only three are used for R. Please clarify the design and the exclusion logic, and specify which combinations were tested and which were not.

Minor Comments

1. Equation numbering is broken; Equation (10) is missing. Renumber all equations sequentially throughout the manuscript.
2. Figure 2 plots both RMSE and absolute daily-emission bias within each grid cell. The caption is ambiguous. Revise the figure caption to explicitly define both metrics and their corresponding units.
3. In Table A2, clearly mark units: synthetic-experiment RMSE in μmol m^{-2} s^{-1} and real-data RMSE in ppm, either in table header or footnote.
4. Double-check units in Appendix Figure captions (Fig A1-A6), especially for footprint quantity. Keep symbol formatting consistent.
5. In Section 2.1.3, note that hyperparameter bounds for correlation length ($10 < l < 80$ km) are tuned for this 40 km × 40 km inversion domain. Remind users that bounds need re-adjustment for study domains of different spatial extents.