Reviewer Comments

This manuscript presents NeuPlume, a neural-physical probabilistic framework for estimating point-source emissions and associated source and meteorological parameters from sparse atmospheric observations. The work addresses an important methodological gap. Under limited near-field sampling, emission rate, effective release height, wind speed, turbulence, and plume morphology can be strongly coupled, whereas conventional methods generally return a single estimate and cannot fully represent this non-uniqueness. The combination of a Lagrangian stochastic model, conditional neural fields, latent diffusion, and observation-constrained inversion is technically interesting and potentially valuable for atmospheric emission quantification.

I appreciate that the authors generally acknowledge the physical limitations of the current implementation and interpret the coal-mine application as an illustrative consistency check rather than a strict field validation. Nevertheless, the present evidence remains insufficient to fully support several of the main claims regarding inversion accuracy, uncertainty quantification, and applicability to real observations. The principal concern is not the novelty of the framework, which is clear, but the robustness and structure of the validation. I believe the manuscript could make a useful contribution after substantial revision, particularly if the authors strengthen the numerical experiments, clarify the probabilistic interpretation, and more clearly demonstrate the added value of the diffusion component.

Major comments

1. The validation remains too closely coupled to the forward model used for training.

The training library, the six main synthetic test cases, and the 30 uncertainty-calibration cases appear to be generated using the same Lagrangian stochastic model and within the same parameter distribution. The current synthetic experiments therefore mainly demonstrate that NeuPlume can learn and invert concentration fields generated by its own physical simulator. This is an important initial test, but it does not establish robustness to transport-model error, which is likely to be a major uncertainty in real applications.

I encourage the authors to add an independent or out-of-distribution validation. For example, NeuPlume could be trained with the current LSM library and tested using concentration fields generated by a different Lagrangian parameterization, a Gaussian-puff model, LES, or CFD. Perturbation experiments involving atmospheric stability, wind shear, plume rise, or turbulence parameterization would also help evaluate the sensitivity to model mismatch.

The real-data application is useful, but the authors correctly note that the ventilation-shaft plume may involve buoyancy, initial momentum, non-neutral stability, and local-flow effects that are absent from the training library. A controlled-release experiment with a known emission rate would substantially strengthen the study. If such data are not currently available, the authors should expand the cross-model synthetic validation and further moderate any claims regarding field accuracy.

2. The synthetic experiments should be more statistically robust and more representative of actual UAV sampling.

The principal accuracy results are based on only six synthetic cases, with Q fixed at 1 kg s-1, and the synthetic inversions use a single fixed random seed. In addition, the observations are generated by randomly sampling points throughout a three-dimensional concentration field. This protocol provides useful volumetric information, but it differs substantially from actual UAV measurements, which are sequential, spatially correlated, restricted to a small number of flight transects, and often affected by incomplete plume interception.

The authors could strengthen this section by repeating each physical case with multiple independent sampling masks and reporting the distribution of errors rather than only the mean over six cases. Different emission strengths, signal-to-noise ratios, flight geometries, and sampling distances should also be considered. Realistic synthetic observations should include at least some combination of instrument noise, background uncertainty, wind-direction error, location uncertainty, and temporal mismatch.

It would also be helpful to use actual UAV trajectories as observation operators for the synthetic plumes. This would provide a more direct connection between the numerical and field experiments.

The manuscript should additionally clarify the apparent difference between the reported six-case mean Q error of 10.0% and the mean error of 26.4% reported at 1% observation coverage, although the principal synthetic configuration also appears to correspond to approximately 1% of the computational grid. This difference may arise from alternative sampling masks or inversion settings, but it should be explicitly explained.

3. The statistical interpretation and calibration of the approximate posterior require further justification.

The manuscript appropriately states that the resulting distribution is not a strict Bayesian posterior. However, the current construction combines a Gaussian observation equation, a Huber reconstruction loss, least-squares or robust estimation of Q, the minimum reconstruction loss among diffusion trajectories, and Boltzmann weighting with an empirically selected temperature. These elements do not yet appear to form a fully consistent probabilistic model.

In particular, using the minimum loss among multiple diffusion trajectories as a surrogate for the negative log-likelihood may make the resulting distribution dependent on the number of trajectories. Increasing the number of samples will generally reduce the minimum loss even when the underlying probability model is unchanged. The authors should explain the statistical justification for this operation or consider an alternative approximation, such as averaging, log-sum-exp marginalization, or another method that integrates over latent-field realizations.

The influence of the parameter grid should also be discussed. The coarse grid is non-uniform, and the fine grid is generated only around the best coarse-grid cell. Without accounting for grid-cell volume, denser regions may receive greater posterior mass. Local refinement around one optimum may also omit secondary modes. The authors should define the parameter prior explicitly and demonstrate that the posterior distributions are stable under alternative grid resolutions and search strategies.

Most importantly, the uncertainty-calibration experiment evaluates H, U, and I, but not the primary target variable Q. Since uncertainty-aware emission quantification is a central contribution, the authors should report empirical coverage, interval width, and preferably proper probabilistic scores for Q. Quantile-based or highest-density intervals may also be more appropriate than symmetric intervals for bounded, skewed, or multimodal distributions.

The high coverage values at the 90% and 95% levels should be interpreted together with posterior sharpness. The reported effective sample size of 70.7 out of 75 at tau = 1 suggests relatively flat importance weights. Reporting posterior contraction relative to the prior would help demonstrate how much information is actually provided by the observations.

4. The specific contribution of diffusion posterior sampling has not yet been clearly demonstrated.

The ablation experiment shows that including the observation-likelihood gradient in the reverse diffusion process does not reduce the emission-rate error and slightly worsens the mean result. The authors also state that the dominant constraint in the present implementation comes from explicit parameter-grid enumeration and loss comparison.

This result raises an important methodological question: does the performance improvement mainly originate from the LSM-based physical library, the conditional neural-field surrogate, the diffusion prior, or the coarse-to-fine grid search?

I recommend adding closer methodological baselines, such as:

- deterministic CNF reconstruction followed by the same grid search;
- direct interpolation or nearest-neighbour inversion using the LSM library;
- the same surrogate model combined with conventional Bayesian sampling or MCMC;
- a conditional auto-decoder without the diffusion prior.

These comparisons would clarify whether diffusion improves parameter accuracy, full-field reconstruction, posterior calibration, or robustness under sparse observations. This is particularly important because the current inverse parameter space is low-dimensional, while one coarse-to-fine inversion requires approximately 3 GPU-hours.

Additional comments

1. The real-data preprocessing should be described in greater detail, including background determination, concentration-GPS time alignment, instrument response, flight duration, wind variability, coordinate rotation, source-location uncertainty, and the criteria used to select the four flights.

2. The blind-mode wind speed should be interpreted cautiously. In the field cases, it may act as an effective transport or model-compensation parameter rather than an independently retrieved physical wind speed. It may absorb errors associated with plume rise, atmospheric stability, background estimation, sampling geometry, and transport-model mismatch.

3. Reconstruction of the complete three-dimensional concentration field is presented as an important contribution, but the current evaluation appears mainly qualitative. Quantitative field-level metrics should be reported, including normalized RMSE, spatial correlation, plume-centerline displacement, plume-width and vertical-distribution errors, peak-concentration bias, and predictive performance at unobserved locations.

4. The physical interpretation of the generated field ensemble should be clarified. Each parameter combination appears to correspond to one time-averaged LSM field. Repeated independent LSM realizations at fixed (H, U, I) would help determine whether the diffusion ensemble represents atmospheric variability, finite-particle noise, interpolation uncertainty, or generative-model stochasticity.

5. The manuscript should harmonize the descriptions of Q and I estimation. The main text describes analytical least-squares estimation of Q, whereas Appendix B includes optimization of log k and a weighted robust/least-squares estimate. Similarly, the main text describes grid search over (H, U, I), while Appendix B states that I is optimized continuously.

6. More complete implementation details are needed for the LSM, including the definitions of directional velocity variances, the Lagrangian time scale, averaging duration, grid resolution, particle-number convergence, and the relationship between Iref and the turbulence statistics.

7. Given the complexity of the workflow, public release of the code, trained models, synthetic datasets, and sampling masks would substantially improve reproducibility.

Overall assessment

NeuPlume addresses a meaningful problem and presents an original neural-physical framework with clear potential. The manuscript is strongest in its identification of the non-uniqueness of point-source inversion and its attempt to represent a family of data-compatible source and plume configurations rather than a single deterministic estimate. However, the current validation remains largely a proof of concept within the same simulation environment used for training.

The manuscript would be considerably strengthened by more realistic and statistically robust synthetic experiments, independent or controlled validation, a clearer probabilistic formulation, and direct evidence that the diffusion component contributes beyond the physical surrogate and parameter search. I therefore encourage the authors to undertake a substantial revision before the principal performance and applicability claims can be fully supported.