Response to Reviewer 1

We thank the reviewer for the careful and constructive assessment of NeuPlume. The comments helped us sharpen the distinction between performance within the specified LSM family and robustness to transport-model mismatch, strengthen the statistical and observational design of the validation, and clarify the contribution and evidential scope of the generative inversion framework. In response, we made the 30 training-excluded synthetic cases the primary validation set; added controlled strong-shear and effective-plume-rise experiments, three UAV-derived observation geometries, a matched comparison of direct LSM search, CNF library search, and full NeuPlume, quantitative full-field reconstruction, and open-boundary diagnostics; and expanded the real-data and LSM descriptions. We also revised the statistical formulation throughout the manuscript: NeuPlume now reports the single minimum-loss candidate as a point estimate, and we withdrew the posterior, credible-interval, calibration, temperature, effective-sample-size, and parameter-uncertainty claims that were not supported by the present design. The revised manuscript states the remaining evidence boundaries explicitly, including same-simulator validation, the absence of controlled-release field truth, and the passive, low-height, neutral, flat-terrain scope of the current implementation. Our point-by-point responses below identify the corresponding manuscript text, tables, and figures.

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

Response. We thank the reviewer for identifying the important distinction between learning the specified physical model family and robustness to transport-model error. Sharing the LSM forward model between training and validation is an intentional design principle of NeuPlume: the framework first uses LSM as a physics-based approximation of atmospheric transport and observation generation, and then trains the CNF and diffusion components to learn the structure of this physically generated field family. The 30 validation fields, including the six representative cases, were excluded from CNF and diffusion training. This design tests whether the learned representation and candidate-generation machinery recover the governing physical relationship from observations, while keeping the validation samples separate from the training samples.

To address the reviewer's request for cross-model synthetic validation, we added controlled transport-mismatch experiments. The six representative cases were evaluated with the same observations and inversion protocol after generating new truth fields with (i) the matched simple-transport model, (ii) a strong-shear model with power-law exponent 0.30, and (iii) an effective-plume-rise model with an unrepresented vertical mean velocity. All mismatch fields were inverted using CNF and diffusion models trained only on the simple-transport library. Strong shear increased mean errors in effective plume height and wind speed to 3.44% and 27.41%, compared with 2.46% and 25.63% for the matched baseline; mean normalized RMSE and spatial correlation remained similar (0.308 and 0.953 versus 0.322 and 0.957). Effective plume rise increased mean height and wind-speed errors to 22.72% and 35.12%, raised normalized RMSE to 0.518, and lowered spatial correlation to 0.821. These results show that transport mismatch has mechanism-dependent effects: shear primarily shifts parameter identification, whereas plume rise degrades both parameter and field recovery.

Data provenance and mismatch protocols are defined in Sect. 3. Results are reported in Sect. 4.1 and Appendix Figs. F1 and G1--G2, with interpretation in Sects. 5 and 6.

Revised manuscript text.

Sect. 3, Experimental setup:

> The transport-mismatch stress test uses the same six off-grid parameter settings, 1\% random-volume observation masks, and inversion protocol. New truth fields are generated with the simple transport model, a strong-shear model with power-law exponent $\alpha = 0.30$, and an effective-plume-rise model with prescribed vertical mean velocity $w(x) = 0.6\,\exp(-x/80)\ \mathrm{m\,s^{-1}}$ for $x \geq 0$. All fields are inverted with the CNF and diffusion models trained on the simple transport fields. The parameter settings remain within the trained support, whereas the strong-shear and plume-rise fields contain transport mechanisms absent from the training library.

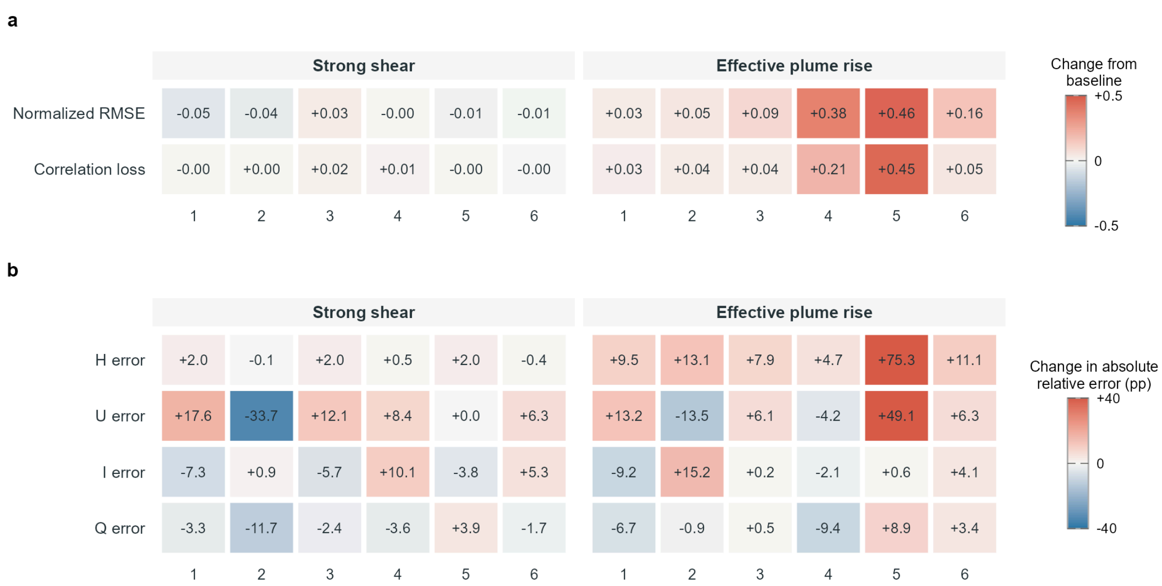
Sect. 4.1, Synthetic data validation:

> Under strong shear, mean H and U errors increase from 2.46\% and 25.63\% to 3.44\% and 27.41\%, whereas mean I and Q errors decrease slightly to 7.20\% and 9.40\%. Mean full-field normalized RMSE changes from 0.322 to 0.308 and spatial correlation from 0.957 to 0.953. Effective plume rise produces a different response: mean H and U errors increase to 22.72\% and 35.12\%, mean I and Q errors are 8.78\% and 11.83\%, normalized RMSE increases to 0.518, and spatial correlation decreases to 0.821.

Sect. 5, Physical scope and boundary effects:

> The present implementation and validation cover passive, low-height, neutral releases over flat terrain. The current forward model uses open boundaries. Within this parameterization, H is an effective release height that can absorb unrepresented vertical transport. Flights #11 and #12 select heights near or above the 50~m training-grid ceiling after continuous refinement, marking the edge of the trained height support. The present observations do not resolve which omitted transport process produces these boundary solutions.
>
> The two transport stress tests show why field agreement and parameter identification must be assessed separately. Strong shear changes the selected simple-model parameters without a corresponding mean loss of full-field similarity, whereas effective plume rise degrades both parameter and field recovery. In the plume-rise test, H absorbs part of the missing vertical transport and functions as an effective height rather than a physical outlet elevation.

Relevant figure.



> **Appendix Fig. F1.** Controlled transport-mismatch stress tests across six representative conditions within the trained parameter support. Rows show changes in normalized RMSE, correlation loss, and the absolute relative errors of H, U, I, and Q; columns identify Cases~1--6 for each mismatch condition. Changes are relative to each case's simple-transport baseline under the same 1\% random-volume observation mask. Positive values denote worse performance, with correlation loss defined as baseline minus tested spatial correlation. Cell labels report untruncated changes; parameter-error colors are truncated at ± 40 percentage points to preserve contrast.

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

Response. We thank the reviewer for recommending a broader statistical evaluation and more realistic UAV-representative sampling. We expanded the principal synthetic evaluation to 30 held-out cases. Emission rates span 0.75--1.25 kg s-1, and each case uses a random 1% mask containing 1,000 observations. Mean absolute relative errors are 2.39% for effective plume height, 23.13% for wind speed, 5.74% for source location, and 13.59% for emission rate. Mean full-field spatial correlation is 0.942 and mean normalized RMSE is 0.345.

Transformed trajectories from the real UAV flights define complete, partial, and off-center observation geometries, each with 1,000 points. Their mean normalized RMSE values are 0.744, 0.565, and 0.611, compared with 0.322 for random-volume sampling; mean spatial correlations are 0.622, 0.803, and 0.817, compared with 0.957. These differences quantify protocol-specific plume interception.

The evaluation covers case-specific random masks and three UAV-derived geometries, with fixed measurement-noise, background, location-error, and wind-direction settings.

The 30-case design and trajectory protocols are described in Sect. 3. Results are reported in Sect. 4.1, Fig. 2, Appendix Table B1, Fig. 3, and Appendix Figs. G1--G2. Measurement-design implications are discussed in Sect. 5.

Revised manuscript text.

Sect. 3, Experimental setup:

> The primary synthetic validation uses 30 concentration fields generated with the same neutral-stability LSM procedure as the training library and held out from conditional neural field and diffusion training. The validation conditions cover $H = 10$--40~m, $U_{\mathrm{ref}} = 3$--$12\ \mathrm{m\,s^{-1}}$, $I_{\mathrm{ref}} = 0.08$--0.27, and five emission rates from 0.75 to $1.25\ \mathrm{kg\,s^{-1}}$. Each time-averaged field contains 10^5 evaluation points in the $520 \times 200 \times 100\ \mathrm{m}^3$ domain, and 1\% of these points (1,000 observations) are selected by a case-specific random volumetric mask.
>
> The UAV trajectories and methane measurements used in the observation-geometry stress test and field experiment come from the publicly available AirCore-UAV data for flights #11, #12, #14, and #15 above shaft V of the Pniowek coal mine in Poland \cite{Andersen2023}. For the observation-geometry stress test, each simple-transport truth field is sampled with four equal-budget protocols. The random-volume protocol selects 1,000 of the 10^5 evaluation points. The complete, partial, and off-center protocols transform trajectories from flight #11 or #12, #14, and #15 into the synthetic domain and select the 1,000 evaluation points nearest each trajectory. The partial protocol retains one crosswind portion of its source trajectory; the off-center protocol applies a 60~m crosswind shift and domain

> truncation. These compound protocols test three realistic trajectory patterns rather than individual geometric factors.

Sect. 4.1, Synthetic data validation:

> Across the 30 synthetic validation cases, the mean absolute relative errors were 2.39\% for effective release height H, 23.13\% for wind speed U, 5.74\% for turbulence intensity I, and 13.59\% for emission rate Q. The corresponding medians were 0.75\%, 16.96\%, 5.41\%, and 10.72\%, respectively.
>
> With 1,000 random-volume observations per case, the mean errors in H, U, I, and Q are 2.46\%, 25.63\%, 7.31\%, and 12.51\%, respectively; mean full-field normalized RMSE is 0.322 and spatial correlation is 0.957. The complete, partial, and off-center trajectory protocols use the same point budget. Their mean normalized RMSE values are 0.744, 0.565, and 0.611, while mean spatial correlations are 0.622, 0.803, and 0.817. Mean I error exceeds 23\% for all three trajectory protocols. The ordering varies by metric: the complete trajectory has the largest mean H error and normalized RMSE, the partial trajectory has the largest mean U and Q errors, and the off-center trajectory has the largest mean I error.

Appendix B, Table B1:

Metric	Mean	Median	Range	Unobserved-point mean
H error (\%)	2.390	0.752	0.075--10.032	n/a
U error (\%)	23.128	16.962	0.296--67.262	n/a
I error (\%)	5.742	5.414	0.196--14.597	n/a
Q error (\%)	13.590	10.716	1.819--35.923	n/a
Spatial correlation	0.942	0.954	0.859--0.981	0.942
Normalized RMSE	0.345	0.320	0.216--0.543	0.345
Centerline displacement (m)	0.782	0.680	0.144--1.954	0.783
Crosswind width bias (m)	-1.202	-0.904	-2.530---0.133	-1.202
Vertical width bias (m)	-0.764	-0.532	-2.191--0.024	-0.765

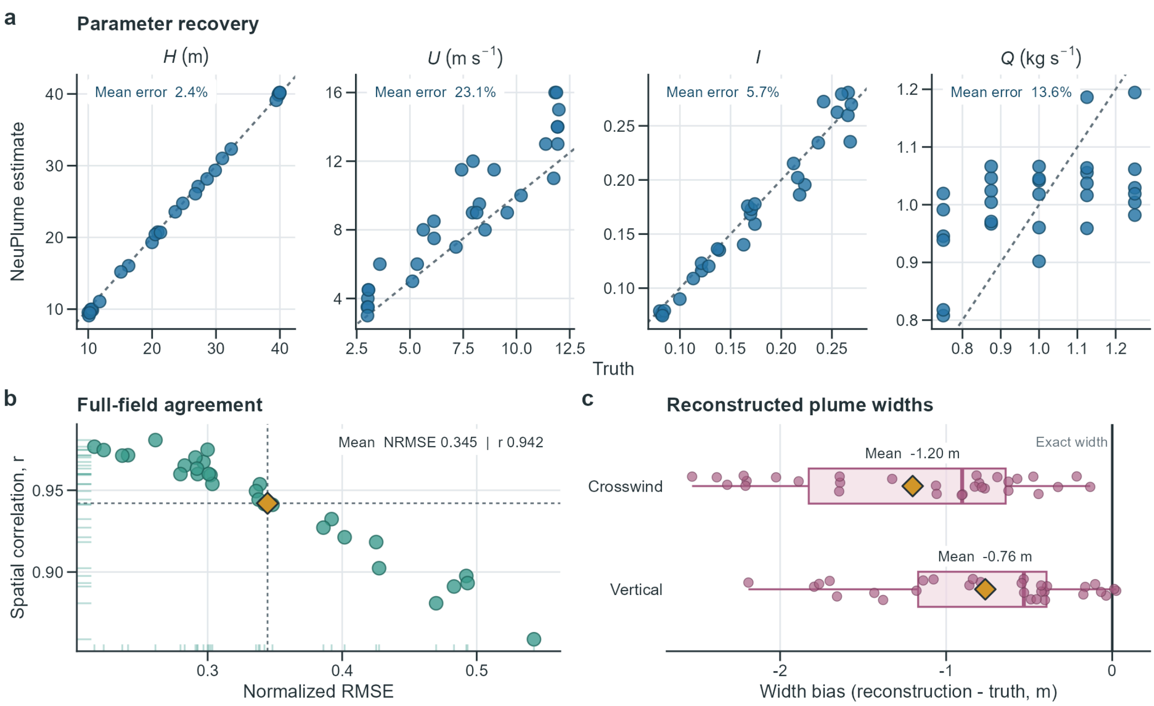
> **Table B1 caption.** Performance across the 30-case synthetic validation set. Parameter rows report absolute relative error. Field rows compare the selected concentration field with the LSM truth. Width bias is reconstruction minus truth, so negative values indicate a narrower reconstructed plume.

Sect. 5, Measurement geometry, wind representativeness, and identifiability:

> The controlled geometry test shows that point count alone does not determine performance. Concentrating the matched point budget near transformed UAV trajectories degrades parameter and full-field recovery on average relative to random-volume sampling. Each trajectory protocol represents a compound sampling pattern, but their common

> contrast with the matched random-volume protocol supports a clear measurement objective: sample multiple downwind positions, obtain crosswind and vertical contrast, and record wind over the plume-transport scale.

Relevant figures.



> **Fig. 2.** Synthetic validation across 30 held-out cases. (a) Each point is one synthetic case. Estimated values are plotted against the true H, U, I, and Q; dashed diagonal lines denote exact recovery and inset labels give mean absolute relative errors. (b) Joint full-field normalized RMSE and spatial correlation, for which the upper-left direction indicates lower error and stronger spatial agreement; dashed crosshairs and the diamond mark the two sample means. (c) Crosswind and vertical width biases; negative values indicate a narrower reconstruction. Points are cases, boxes show medians and interquartile ranges, and diamonds mark means. Quantitative field metrics use all 100,000 evaluation points.

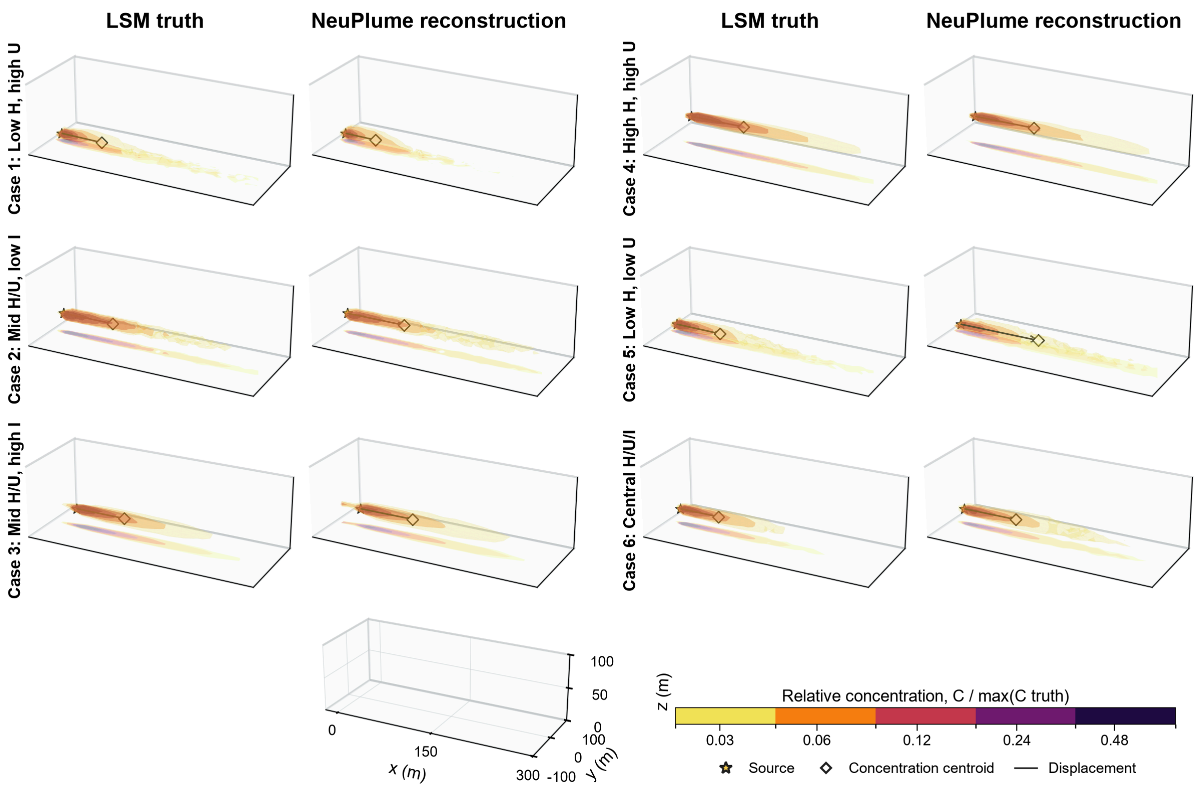
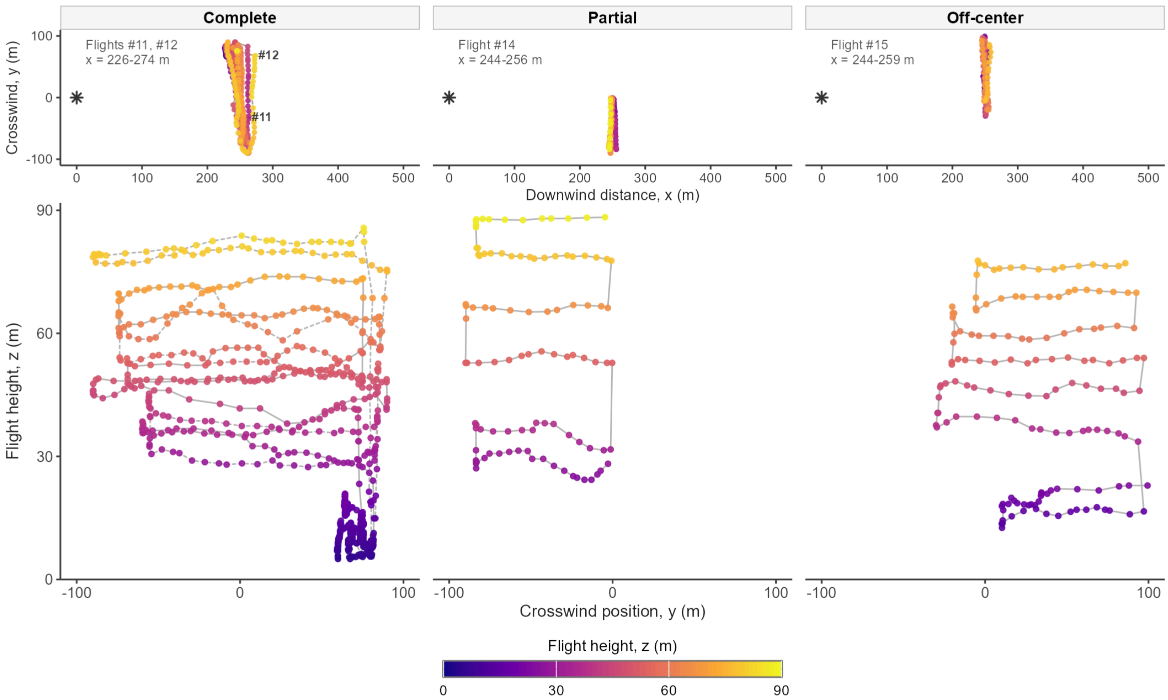
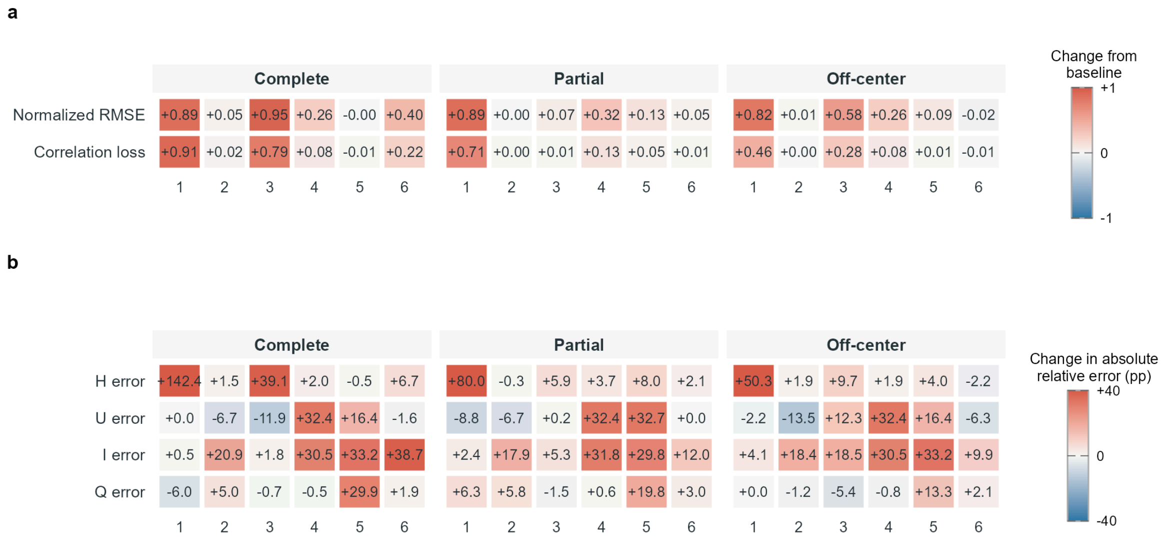


Fig. 3. Three-dimensional plume comparisons for the six representative cases. The first two columns compare the LSM truth and NeuPlume reconstruction for Cases 1--3: Case 1 (low H, high U), Case 2 (mid H/U, low I), and Case 3 (mid H/U, high I). The last two columns show the same comparison for Cases 4--6: Case 4 (high H, high U), Case 5 (low H, low U), and Case 6 (central $H/U/I$). Nested isosurfaces show the three-dimensional concentration field, with maximum-concentration projections on the ground and rear planes. Color and opacity denote $C/\max(C_{\mathrm{truth}})$ within each case; values below 0.02 are omitted for visibility. Stars mark the source location at the case-specific effective release height, diamonds mark the concentration-weighted centroid of the displayed volume bins, and lines show their displacement. The display uses $52 \times 30 \times 20$ bins reconstructed from all 100,000 evaluation points.



Appendix Fig. G1. Complete, partial, and off-center transformed UAV sampling trajectories used in the controlled observation-geometry stress test. The upper strips locate the trajectories in the downwind--crosswind plane; stars mark the source and annotations identify the source flights and sampled downwind ranges. The lower panels show the crosswind--height sampling curtains. Color denotes flight height. Each protocol uses a different source flight and transformation and therefore represents a compound sampling geometry.



> **Appendix Fig. G2.** Case-wise degradation under three transformed UAV sampling protocols across six representative conditions. Rows show changes in normalized RMSE, correlation loss, and the absolute relative errors of H, U, I, and Q; columns identify Cases 1--6 for each protocol. Changes are relative to each case's matched 1,000-point random-volume baseline. Positive values denote worse performance, with correlation loss defined as baseline minus tested spatial correlation. Cell labels report untruncated changes; parameter-error colors are truncated at ± 40 percentage points to preserve contrast. Each protocol combines a different source flight and transformation and therefore represents a compound geometry.

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

Response. We thank the reviewer for requesting clarification of the statistical interpretation. NeuPlume produces a point estimate: DPS generates candidate fields, and the reported output is the parameter vector and decoded concentration field that minimize the observation-mismatch loss. Each benchmark run evaluates one candidate trajectory at each grid node, followed by explicit coarse-to-fine minimum-loss selection. Emission rate is evaluated as a point estimate across all 30 validation cases.

The argmin estimand is defined in Sect. 2.1; candidate generation and selection are defined in Sect. 2.4. Appendix Table A1 lists the grid and optimization settings. Point estimates are reported throughout the manuscript and appendices.

Revised manuscript text.

Sect. 2.1, NeuPlume model architecture:

> Let $C(\mathbf{x} \mid \boldsymbol{\theta})$ denote a unit-emission concentration field and let $\mathcal{H}$ map the complete field to the measurement locations. The observations satisfy
>
> $$\mathbf{y} = Q\,\mathcal{H}[C(\mathbf{x} \mid \boldsymbol{\theta})] + \boldsymbol{\epsilon},$$
>
> where $\boldsymbol{\epsilon}$ collects measurement and representation residuals, and linear scaling by Q follows from unit-rate normalization. NeuPlume evaluates a finite candidate set produced by the conditional generative model and coarse-to-fine grid search. Its formal output is

> $$(\hat{\boldsymbol{\theta}}, \hat{Q}, \hat{C}) = \underset{(\boldsymbol{\theta}_k, Q_k, C_k) \in \mathcal{G}}{\arg\min} \mathcal{L}_{\text{obs}}(\boldsymbol{\theta}_k, Q_k, C_k),$$
>
> where $\mathcal{G}$ is the generated candidate set.

Sect. 2.4, Point-source emission inversion:

> DPS supplies observation-guided candidate fields for grid-node evaluation, and the minimum-loss rule produces a point estimate.
>
> The coarse stage evaluates a broad grid of H, U_{ref}, and I_{ref}. At each node, NeuPlume generates a candidate field, fits Q, and evaluates the observation mismatch. The node with the smallest loss defines the center of the fine grid. The fine stage evaluates local offsets and returns the lowest-loss fine-grid candidate with its fitted Q and decoded concentration field.

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

Response. We thank the reviewer for asking us to isolate the contribution of diffusion and DPS. We added and report a matched three-method comparison using the same six representative cases, 1,000 observations per case, parameter grid, and loss: direct LSM library search, CNF library coarse-to-fine search, and full NeuPlume.

Relative to CNF library search, full NeuPlume reduces mean errors in effective plume height, wind speed, source location, and emission rate from 3.21%, 49.36%, 22.10%, and 18.54% to 2.46%, 25.63%, 7.31%, and 12.51%. Mean normalized RMSE decreases from 0.438 to 0.322, and spatial correlation increases from 0.910 to 0.957. This matched comparison quantifies the benefit of diffusion-based candidate generation.

Direct LSM search is a finite-library reference with mean normalized RMSE 0.178 and spatial correlation 0.981. Its candidates are restricted to the 1,000-member physical library, whereas the validation cases are off-grid. The comparison measures candidate-generation performance under the same observation and loss protocol.

Results and interpretation are given in Sect. 4.2, Fig. 4, and Sects. 5 and 6.

Revised manuscript text.

Sect. 4.2, Matched component comparison:

> We compare full NeuPlume with two references using the same six representative cases, 1,000 random-volume observations per case, and loss definition. The CNF library search removes diffusion and evaluates learned latent fields directly on the same coarse-to-fine parameter grid. Relative to this reference, full NeuPlume reduces the mean errors in H, U, I, and Q from 3.21\%, 49.36\%, 22.10\%, and 18.54\% to 2.46\%, 25.63\%, 7.31\%, and 12.51\%, respectively. Mean normalized RMSE decreases from 0.438 to 0.322, and spatial correlation increases from 0.910 to 0.957. These matched results show that diffusion improves parameter selection and full-field reconstruction relative to direct CNF library search under the tested protocol.
>
> Direct LSM search evaluates the 1,000-member physical library against the off-grid representative truth fields and provides a finite-library lookup reference. It attains a mean normalized RMSE of 0.178 and a mean spatial correlation of 0.981. Its remaining parameter errors (H: 1.62\%, U: 17.79\%, I: 11.41\%, Q: 22.91\%) quantify discretization and parameter ambiguity in this same-simulator lookup.

Relevant figure.

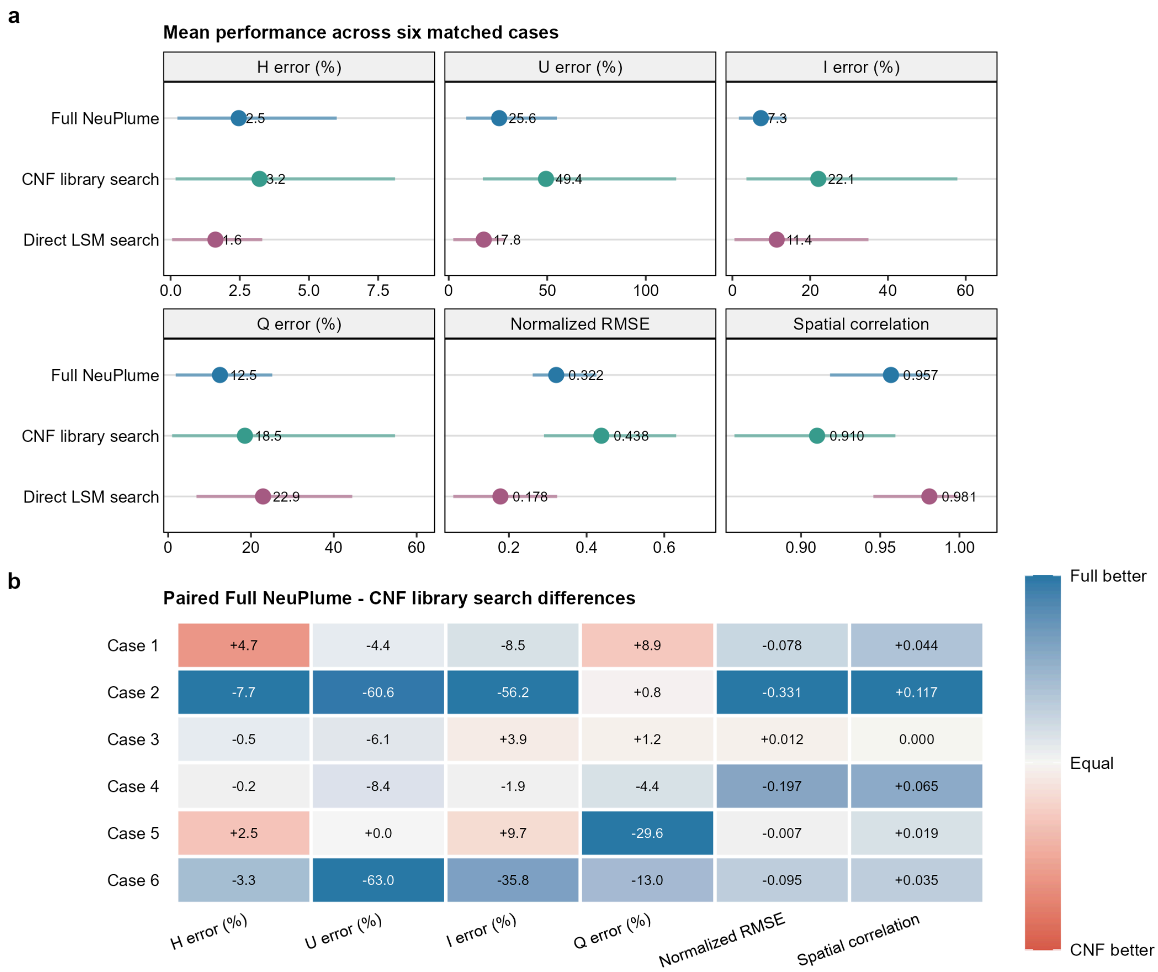


Fig. 4. Matched accuracy comparison for the six representative cases. (a) Mean absolute relative errors for effective release height H, wind speed U, turbulence intensity I, and emission rate Q, together with mean full-field normalized RMSE and spatial correlation; horizontal intervals span the minimum to maximum across the six cases. (b) Case-wise paired differences between Full NeuPlume and CNF library search. Cell labels report the raw Full NeuPlume minus CNF library search difference in the units of each metric. Colour direction is oriented so that blue denotes better Full NeuPlume performance and orange denotes better CNF library-search performance, with intensity scaled separately within each metric. Lower values indicate better performance for the four parameter errors and normalized RMSE; higher values indicate better spatial correlation. Direct LSM search is a finite-library lookup reference whose 1,000 candidates do not contain the six truth conditions.

Additional comment 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.

Response. We thank the reviewer for requesting a transparent account of the real-data preprocessing. We now describe how the inversion reverses and aligns the analyzer sequence to GPS time, obtains methane enhancement after background subtraction, expresses coordinates relative to the source, filters observations to the 520 x 200 x 100 m model domain, and associates them with the corresponding flight locations. Flights 11, 12, 14, and 15 in the Pniowek V source dataset provide 362, 214, 108, and 130 in-domain observations, respectively. Meteorological wind speed is 6--7 m s-1.

Preprocessing and data scope are described in Sect. 3. Observation counts and known-wind/blind settings are reported in Sect. 4.3 and Table 2.

Revised manuscript text.

Sect. 3, Experimental setup:

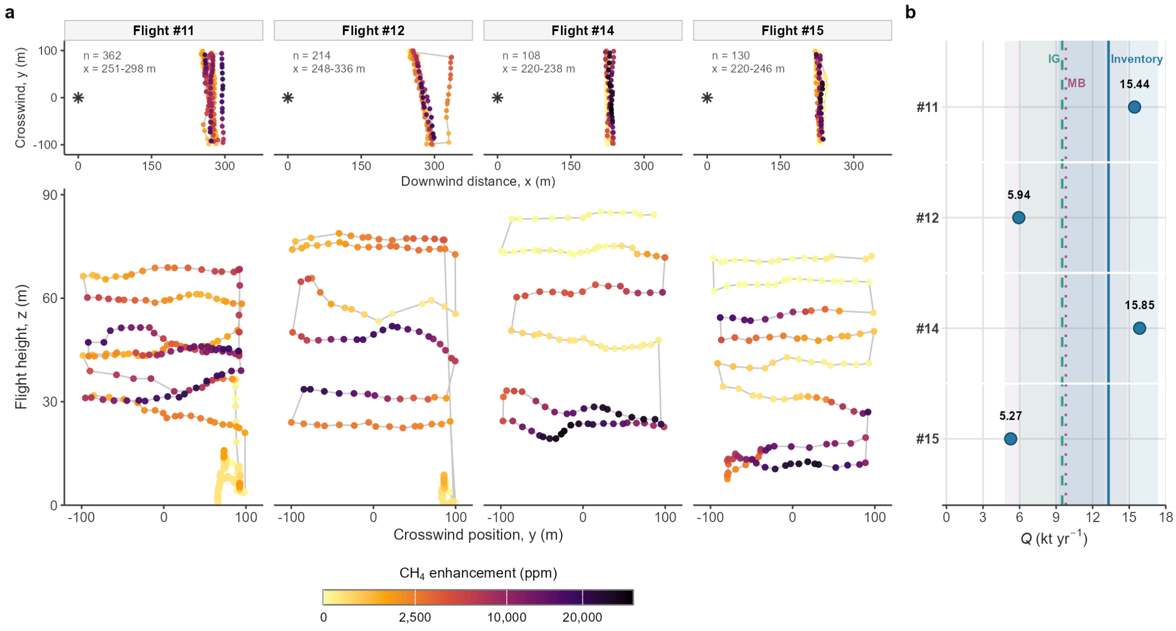
> The field experiment applies NeuPlume to the methane measurements from these four flights. Ground-station wind speed was 6--7~$\mathrm{m\,s^{-1}}$. Preprocessing reverses the analyzer sequence, aligns it with GPS time, and assigns each methane observation to the corresponding flight location. From the same study, descriptive external references are the same-shaft hourly inventory (13.3 ± 4.1 kt yr^{-1}) and the campaign-level integrated Gaussian (IG, 9.5 ± 3.5 kt yr^{-1}, $n = 13$) and mass-balance (MB, 9.8 ± 5.0 kt yr^{-1}, $n = 14$) estimates.

Sect. 4.3, Real-data case study:

> We evaluate the Pniowek V flights in known-U and blind modes. The same-shaft hourly inventory (13.3 ± 4.1 kt yr^{-1}) provides the primary descriptive reference. Campaign-level IG (9.5 ± 3.5 kt yr^{-1}) and MB (9.8 ± 5.0 kt yr^{-1}) estimates provide additional references \cite{Andersen2023}. Table 2 and Fig. 5 report NeuPlume's selected minimum-loss estimates.

Flight	N_{obs}	U_{meteo} ($\mathrm{m\,s^{-1}}$)	Known-U: $\hat{Q}$ (kt yr^{-1})	Blind: $\hat{Q}$ (kt yr^{-1})	Blind: $\hat{U}$ ($\mathrm{m\,s^{-1}}$)
#11	362	7.0	15.44	5.17	3.0
#12	214	7.0	5.94	3.49	2.5
#14	108	6.0	15.85	2.84	1.3
#15	130	6.0	5.27	13.26	14.0

Relevant figure.



> **Fig. 5.** Real-data inversion for the Pniowek V ventilation shaft. (a) UAV sampling geometry for each flight. The upper strips locate the observations in the downwind--crosswind plane; stars mark the shaft. The lower panels show the crosswind--height sampling curtains, with grey lines tracing the flight paths and color denoting methane enhancement. Labels report the number of observations and sampled downwind range. (b) NeuPlume known-U minimum-loss Q estimates (points). The same-shaft inventory and published IG and MB values retain their reported external ranges.

Additional comment 2: The blind-mode wind speed should be interpreted cautiously.

Response. We thank the reviewer for encouraging a cautious interpretation of blind-mode wind speed. Flights 11, 12, and 14 select winds below the station value, whereas Flight 15 selects 14.0 m s-1, more than twice the 6.0 m s-1 station value. This pattern demonstrates emission-rate--wind-speed coupling under field-model mismatch. We therefore describe wind speed in blind inversion as an effective transport parameter rather than as an independently validated local wind measurement.

This interpretation appears in Sects. 4.3, 5, and 6.

Revised manuscript text.

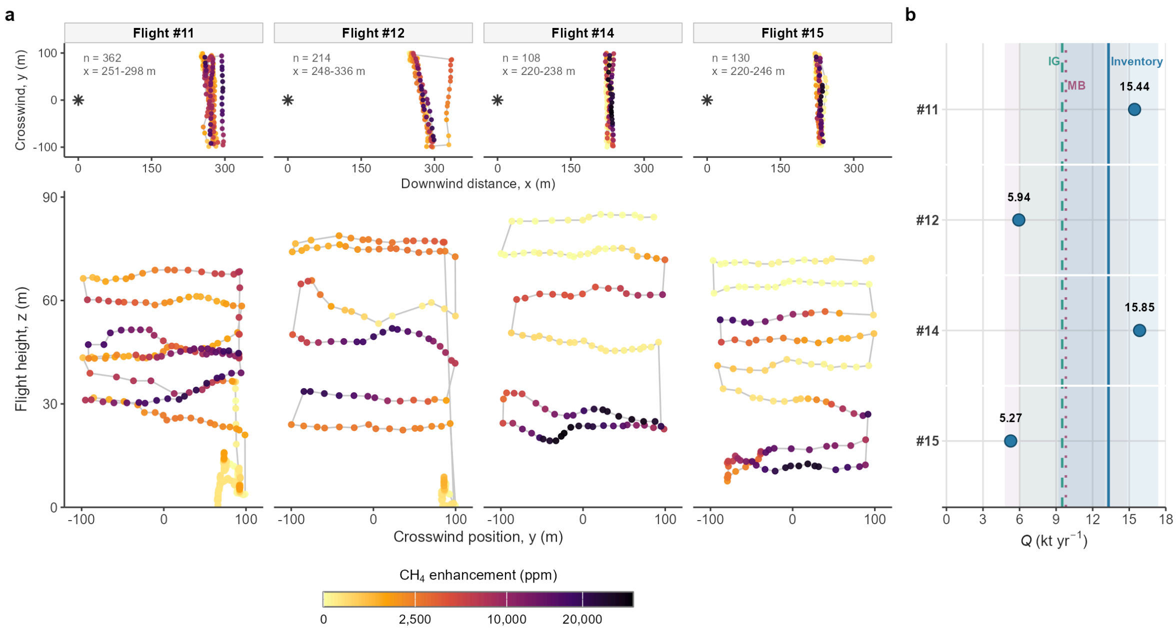
Sect. 4.3, Real-data case study:

> Blind inversion selects lower wind speeds for flights #11, #12, and #14 and correspondingly lower Q values. Flight #15 selects $U = 14.0\ \mathrm{m\ s^{-1}}$, more than twice the station value, and $Q = 13.26\ \mathrm{kt\ yr^{-1}}$. Across all four flights, the selected U and Q move in the same direction relative to the known-U solution.

Sect. 5, Measurement geometry, wind representativeness, and identifiability:

> The field inversions expose strong coupling between wind-speed representativeness and the selected emission rate. Because the CNF represents unit-emission fields, the fitted source strength scales with effective transport speed. Across all four flights, the blind solutions move U and Q in the same direction relative to the known-U solutions. This behavior is consistent with wind sensitivities reported in cross-sectional-flux and controlled-release studies \cite{Sandro2026, Eastwood2025, Conrad2023, Zhonghua2024}.

Relevant figure.



> **Fig. 5.** Real-data inversion for the Pniowek V ventilation shaft. (a) UAV sampling geometry for each flight. The upper strips locate the observations in the downwind--crosswind plane; stars mark the shaft. The lower panels show the crosswind--height sampling curtains, with grey lines tracing the flight paths and color denoting methane enhancement. Labels report the number of observations and sampled downwind range. (b) NeuPlume known-U minimum-loss Q estimates (points). The same-shaft inventory and published IG and MB values retain their reported external ranges.

Additional comment 3: Reconstruction of the complete three-dimensional concentration field is presented as an important contribution, but the current evaluation appears mainly qualitative.

Response. We thank the reviewer for asking for quantitative three-dimensional reconstruction evidence. We added full-field evaluation across all 30 validation cases, reporting normalized RMSE, spatial correlation, concentration-weighted centerline displacement, horizontal and vertical width bias, and the same metrics on the 99% unobserved locations. The 30-case means are 0.345 normalized RMSE, 0.942 correlation, and 0.78 m centerline displacement. Reconstructed crosswind and vertical widths are narrower than truth by 1.20 and 0.76 m on average. Unobserved-point values are nearly identical to full-field values. The main paper now displays paired LSM truth and NeuPlume reconstructions for all six representative cases.

Definitions are in Sect. 3; results are in Sect. 4.1, Table 1, Figs. 2--3, and Appendix Tables B1--B2.

Revised manuscript text.

Sect. 3, Experimental setup:

> For synthetic cases, parameter accuracy is evaluated using the absolute relative error of the selected Q, H, U, and I. Full-field reconstruction is evaluated using normalized root-mean-square error (RMSE divided by the root-mean-square truth concentration), Pearson spatial correlation, concentration-weighted centerline displacement, and horizontal and vertical width bias. The same metrics are also computed after excluding the 1\% observation mask.

Sect. 4.1, Synthetic data validation:

> The selected three-dimensional fields had a mean spatial correlation of 0.942 and a mean normalized RMSE of 0.345 against the LSM truth. Their concentration-weighted centerlines were displaced by 0.78~m on average. The reconstructed crosswind and vertical widths were narrower than the truth by 1.20 and 0.76~m on average. Metrics computed only at the 99\% unobserved points were nearly identical: the mean spatial correlation was 0.942, normalized RMSE was 0.345, and centerline displacement was 0.78~m. Figure 2 relates each estimate directly to its truth value and shows the joint full-field agreement and width biases.
>
> Figure 3 provides the corresponding three-dimensional comparison for all six preselected representative cases. Across these cases, spatial correlation ranges from 0.918 to 0.981 and centerline displacement ranges from 0.18 to 1.68~m. The complete 30-case summary and the physical conditions of the six cases are reported in Appendix B.

Appendix B, Table B1:

Metric	Mean	Median	Range	Unobserved-point mean
H error (\%)	2.390	0.752	0.075--10.032	n/a
U error (\%)	23.128	16.962	0.296--67.262	n/a
I error (\%)	5.742	5.414	0.196--14.597	n/a
Q error (\%)	13.590	10.716	1.819--35.923	n/a
Spatial correlation	0.942	0.954	0.859--0.981	0.942
Normalized RMSE	0.345	0.320	0.216--0.543	0.345
Centerline displacement (m)	0.782	0.680	0.144--1.954	0.783
Crosswind width bias (m)	-1.202	-0.904	-2.530---0.133	-1.202
Vertical width bias (m)	-0.764	-0.532	-2.191--0.024	-0.765

> **Table B1 caption.** Performance across the 30-case synthetic validation set. Parameter rows report absolute relative error. Field rows compare the selected concentration field with the LSM truth. Width bias is reconstruction minus truth, so negative values indicate a narrower reconstructed plume.

Relevant figures.

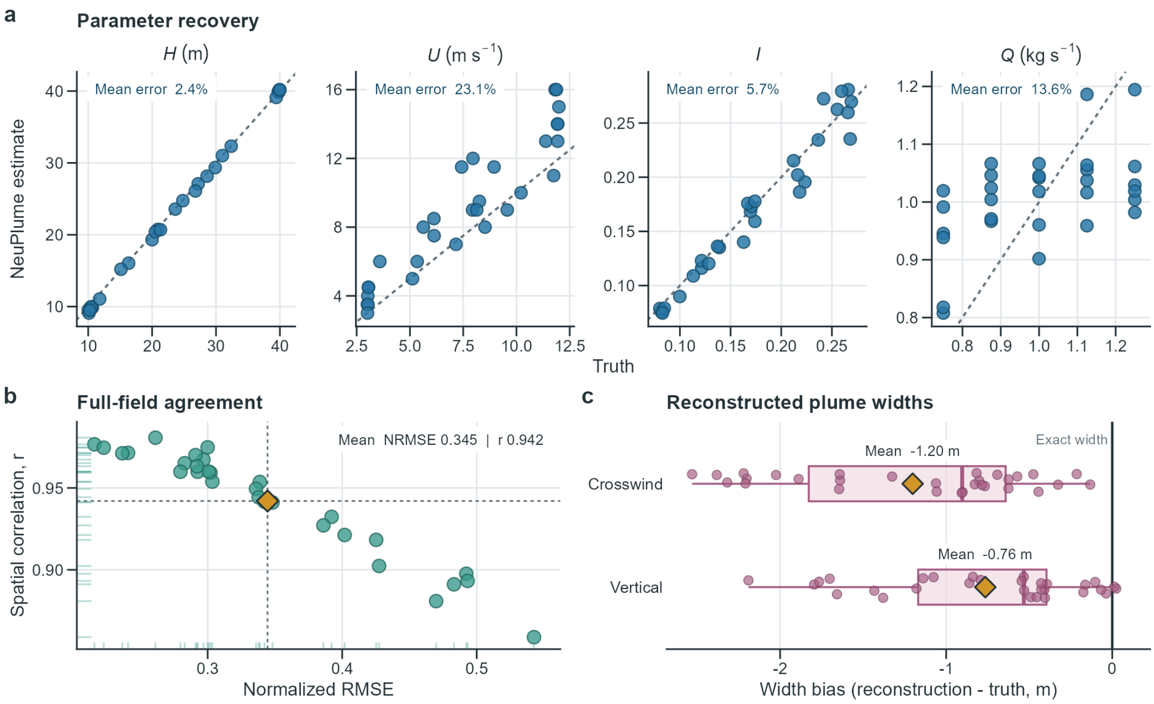


Fig. 2. Synthetic validation across 30 held-out cases. (a) Each point is one synthetic case. Estimated values are plotted against the true H, U, I, and Q; dashed diagonal lines denote exact recovery and inset labels give mean absolute relative errors. (b) Joint full-field normalized RMSE and spatial correlation, for which the upper-left direction indicates lower error and stronger spatial agreement; dashed crosshairs and the diamond mark the two sample means. (c) Crosswind and vertical width biases; negative values indicate a narrower reconstruction. Points are cases, boxes show medians and interquartile ranges, and diamonds mark means. Quantitative field metrics use all 100,000 evaluation points.

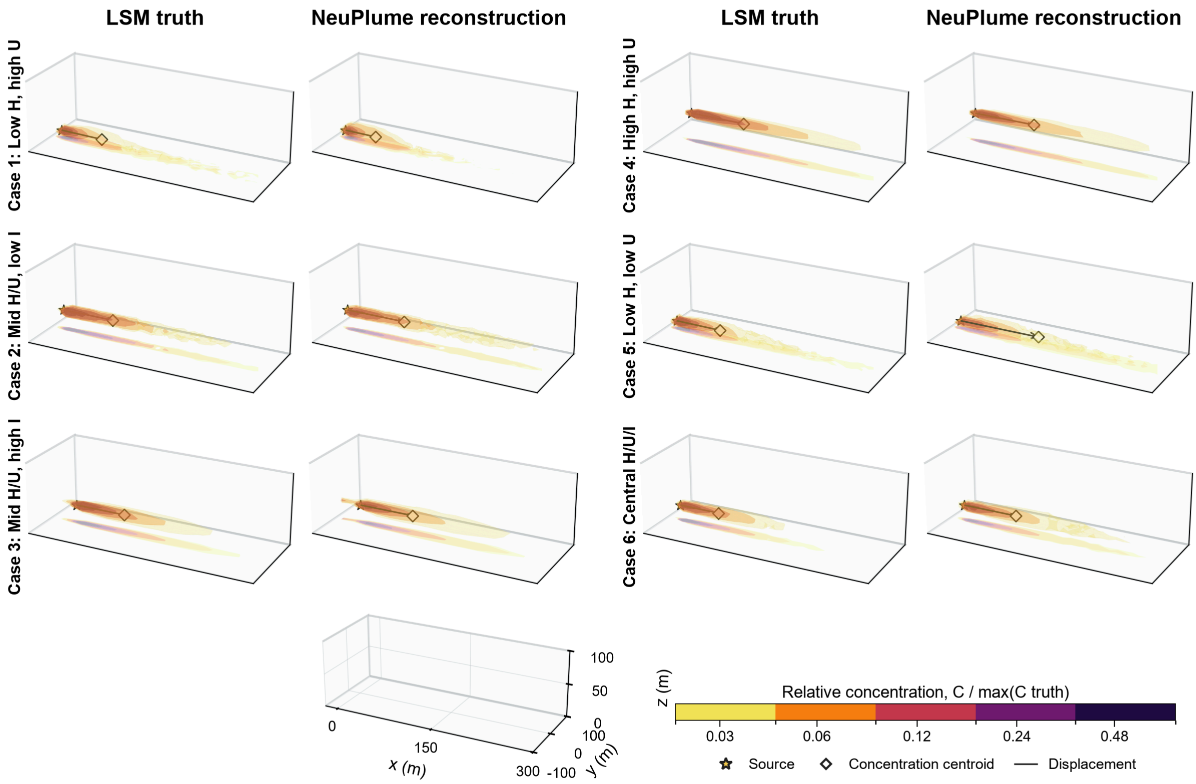


Fig. 3. Three-dimensional plume comparisons for the six representative cases. The first two columns compare the LSM truth and NeuPlume reconstruction for Cases 1--3: Case 1 (low H, high U), Case 2 (mid H/U, low I), and Case 3 (mid H/U, high I). The last two columns show the same comparison for Cases 4--6: Case 4 (high H, high U), Case 5 (low H, low U),

> and Case 6 (central $H/U/I$). Nested isosurfaces show the three-dimensional concentration field, with maximum-concentration projections on the ground and rear planes. Color and opacity denote $C/\max(C_{\text{truth}})$ within each case; values below 0.02 are omitted for visibility. Stars mark the source location at the case-specific effective release height, diamonds mark the concentration-weighted centroid of the displayed volume bins, and lines show their displacement. The display uses $52 \times 30 \times 20$ bins reconstructed from all 100,000 evaluation points.

Additional comment 4: The physical interpretation of the generated field ensemble should be clarified.

Response. We thank the reviewer for asking us to clarify the physical interpretation of the generated fields. Each training condition has one time-averaged LSM field and one learned latent code. The reported result is the selected minimum-loss candidate field; it is not presented as a posterior ensemble of repeated atmospheric realizations.

The point-output definition is given in Sects. 2.1 and 2.4.

Revised manuscript text.

Sect. 2.1, NeuPlume model architecture:

> The framework has three stages. Stage 1 uses a Lagrangian stochastic model (LSM) to generate concentration fields at a unit emission rate $Q = 1\,\mathrm{kg\,s^{-1}}$. Each field is indexed by $\boldsymbol{\theta} = (H, U, I)$, where H is effective release height, U reference wind speed, and I turbulence intensity. Stage 2 uses Hash-INR to compress each field into a 192-dimensional latent code $\mathbf{z}$, and a denoising diffusion probabilistic model (DDPM) learns $p(\mathbf{z} \mid \boldsymbol{\theta})$. In Stage 3, sparse observations guide diffusion posterior sampling (DPS) at each candidate $\boldsymbol{\theta}_k$. NeuPlume decodes a candidate unit-emission field, fits Q by linear scaling, evaluates the observation mismatch, refines the search around the lowest-loss coarse-grid node, and returns $(\hat{\boldsymbol{\theta}}, \hat{Q}, \hat{C})$ from the minimum-loss fine-grid candidate.

Sect. 2.4, Point-source emission inversion:

> NeuPlume searches a finite source and meteorological parameter grid. For each candidate $\boldsymbol{\theta}$, the conditional diffusion model generates a concentration field in the Hash-INR latent space. The algorithm estimates the emission rate Q from the linear relation between the predicted unit field and the observations, evaluates the observation mismatch, and returns the lowest-loss candidate field and parameter vector.

Additional comment 5: The manuscript should harmonize the descriptions of Q and I estimation.

Response. We thank the reviewer for pointing out the need to harmonize the estimation rules. We now state consistently that effective plume height, wind speed, and source location are evaluated on explicit coarse and local fine grids. Emission rate is initialized analytically from linear scaling of the generated unit-emission field and refined during the reverse process. Appendix Table A1 describes source location as a discrete grid variable and lists the fine-grid offsets.

The common estimation rule appears in Sect. 2.4 and Appendix A, Table A1.

Revised manuscript text.

Sect. 2.4, Point-source emission inversion:

> For a generated unit field $\boldsymbol{c}(\boldsymbol{\theta})$ sampled at M observation locations, the predicted observations are $\hat{\boldsymbol{y}} = Q\,\mathcal{H}[\boldsymbol{c}(\boldsymbol{\theta})]$. Candidate quality is measured with a smoothed L_1 observation-mismatch objective:
>
> $$\mathcal{L}_{\text{obs}}(\boldsymbol{\theta}, Q) = \frac{1}{M}\sum_{j=1}^{M} h\big(\hat{y}_j - y_j^{\text{obs}}\big), \quad h(e) = \begin{cases} \frac{1}{2}e^2 & |e| \le 1, \\ |e| - \frac{1}{2} & \text{otherwise.} \end{cases}$$
>
> The Huber form limits the influence of large residuals. The field of a passive tracer scales linearly with source strength, so Q is initialized by the least-squares scaling between the generated unit field and the observations and may be refined during the reverse process.
>
> The coarse stage evaluates a broad grid of H, U_{ref}, and I_{ref}. At each node, NeuPlume generates a candidate field, fits Q, and evaluates the observation mismatch. The node with the smallest loss defines the center of the fine grid. The fine stage evaluates local offsets and returns the lowest-loss fine-grid candidate with its fitted Q and decoded concentration field.

Appendix A, Table A1 (reported search settings):

Component	Parameter	Value
Coarse grid search	H grid	2, 5, 10, 15, 20, 30, 40, 50 m
Coarse grid search	U_{ref} grid	0.8--15.0 $\mathrm{m\,s^{-1}}$ (11 levels)
Coarse grid search	I_{ref} grid	0.05--0.30 (coarse-node values)
Coarse grid search	Candidate nodes	88
Fine grid search	ΔH offsets	$\{-5, 0, +5\}$ m
Fine grid search	ΔU_{ref} offsets	$\{-1.0, -0.5, 0, +0.5, +1.0\}\ \mathrm{m\,s^{-1}}$
Fine grid search	ΔI_{ref} offsets	$\{-0.10, -0.05, 0, +0.05, +0.10\}$
Fine grid search	Candidate nodes	75
Emission-rate refinement	k learning rate	5×10^{-4}
Emission-rate refinement	k mixing weight	$Q = 0.7k_{\text{robust}} + 0.3k_{\text{LS}}$

Additional comment 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.

Response. We thank the reviewer for requesting the additional LSM implementation details. We expanded the description to define the Langevin equation, directional velocity variance, Lagrangian time scale, neutral 1/7 power-law wind profile, reflective ground and open lateral, top, and downwind boundaries, unit-emission scaling, parameter support, computational domain, 0.5 s time step, 250 emitted particles per step, and approximately 100,000 active particles at steady state. In the six representative conditions, 92.0% of emitted particles leave the domain by the end of the diagnostic simulation; 96.8% of exits are downwind on average, while side and top exits account for 3.2% on average and at most 10.2%.

Implementation and assumptions are in Sect. 2.2. Boundary results are reported in Sect. 2.2, Sect. 5, and Appendix Table E1.

Revised manuscript text.

Sect. 2.2, Lagrangian stochastic forward simulation:

> The model continuously releases particles from a point source at height H above the ground. Each particle carries position $\vec{x}$ and turbulent velocity fluctuation $\vec{u}'$, which evolves by a Langevin-type stochastic differential equation:
>
> $$\mathrm{d}u_i' = -\frac{u_i'}{T_\mathrm{L}}\,\mathrm{d}t + \sqrt{\frac{2\,\sigma_i^2}{T_\mathrm{L}}}\,\mathrm{d}W_i, \qquad i \in \{x, y, z\},$$
>
> where T_L is the Lagrangian integral time scale, σ_i is the directional velocity standard deviation, and $\mathrm{d}W_i$ denotes an independent Wiener increment. The mean wind field follows the power-law profile $\bar{U}(z) = U_\mathrm{ref}(z/z_\mathrm{ref})^\alpha$. We set $\alpha = 1/7 \approx 0.143$ for neutral atmospheric stability and define U_ref at $z_\mathrm{ref} = 10$ m. An elastic ground boundary reverses the vertical velocity component, and particles leaving the domain are removed. After a flushing period for statistical steadiness, we bin particle positions onto a regular three-dimensional grid and divide by grid-cell volume to obtain the concentration field $c(\vec{x})$.
>
> The current LSM instance represents a passive inert tracer without initial source momentum, explicit buoyant plume rise, wind-direction shear, convergence, or divergence. It uses flat terrain, neutral stability, and an open single-source domain. Particles that cross the lateral, top, upwind, or downwind boundaries are removed, whereas the ground boundary is reflective.
>
> With the unit-rate convention above, we constructed the training dataset required by the conditional neural field. The training set covers the three-dimensional conditional parameter space spanned by effective release height $H \in [0, 50]$~m, reference wind speed $U_\mathrm{ref} \in [1, 15]$~$\mathrm{m\,s^{-1}}$, and turbulence intensity $I_\mathrm{ref} \in [0.05, 0.3]$. Here, H is the effective release height represented in the training library; it is not necessarily identical to the physical outlet height when initial momentum or buoyancy produces plume rise. These ranges cover near-surface passive point-source monitoring conditions, from calm and weak-turbulence scenarios to strong-wind and high-turbulence cases. Atmospheric stability is neutral in all simulations.
>
> We uniformly sample 1,000 parameter combinations from this space. For each combination, the LSM is integrated over a $520 \times 200 \times 100$~$\mathrm{m}^3$ computational domain ($x \in [-20, 500]$, $y \in [-100, 100]$, $z \in [0, 100]$ m), with time step $\Delta t = 0.5$ s, 250 particles emitted per step, and approximately 10^5 active particles at steady state. The resulting 1,000 concentration fields are paired with their conditional parameter triplets $(H, U_\mathrm{ref}, I_\mathrm{ref})$ to form the supervised training set for neural field compression.

Appendix E, Table E1:

Metric	Mean	Minimum	Maximum
Particle exit fraction	0.920	0.913	0.928
Retained particle fraction	0.080	0.072	0.087

Metric	Mean	Minimum	Maximum
Side/top share of exits	0.032	1.4×10^{-5}	0.102
Downwind share of exits	0.968	0.898	1.000
Side/top share of emitted particles	0.029	1.3×10^{-5}	0.095

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

Response. We thank the reviewer for emphasizing reproducibility. We have documented the evaluation cases, LSM and observation-mask random seeds, trained model parameters, grid definitions, and source data for all reported figures. Code and data are available from the first author upon reasonable request.

The availability statement is in the Code and data availability section. Fixed seeds and protocols are described in Sect. 3 and the appendices.

Revised manuscript text.

Code and data availability:

> The code and data supporting this study are available from the first author upon reasonable request.