Response to Reviewer 2

We thank the reviewer for the detailed and constructive evaluation of the statistical interpretation, component contributions, baseline comparisons, field-reconstruction evidence, and atmospheric measurement implications of NeuPlume. These comments led us to define the method's output more precisely and to reorganize the validation around claims that the current experiments can support. The revised manuscript defines NeuPlume as a finite-candidate, minimum-Huber-loss point estimator, with DPS serving as the observation-guided candidate-generation procedure; unsupported posterior, credible-interval, and calibration claims have been removed. We added a matched comparison of direct LSM search, CNF library search without diffusion, and full NeuPlume; a Gaussian-plume control that separates implementation validity from forward-model and sampling mismatch; quantitative field metrics across 30 training-excluded cases; paired three-dimensional truth--reconstruction displays; and controlled UAV-geometry, transport-mismatch, and boundary diagnostics. We also tie the GP and mass-balance results to the evaluated LSM-generated fields and random-volume observation protocol, and we present the measurement-design findings as conditional results for the tested domain and sampling configurations rather than universal thresholds. Our point-by-point responses below provide the relevant revised manuscript text, tables, and figures.

Major comment 1: meaning of the reported probability distribution

Response. We thank the reviewer for asking us to clarify what probability distribution is represented by the method. NeuPlume returns the candidate that minimizes a Huber observation-mismatch objective over a finite generated set. The observation equation and argmin estimand are given explicitly. DPS is the reverse-diffusion algorithm that supplies observation-guided candidate fields. The reported output is a point estimate.

The training data contain one time-averaged field and one learned latent code per parameter condition. Latent variation quantifies encoding variability; physical plume variability and model uncertainty are outside the estimand.

The estimand and output definition are given in Sects. 2.1 and 2.4, Eqs. (1)--(5). Point-estimation statements are used throughout the manuscript.

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

Sect. 2.4, Point-source emission inversion:

> NeuPlume uses diffusion posterior sampling (DPS), which introduces the gradient of the observation-mismatch objective into the reverse process. 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 2: isolate diffusion and DPS

Response. We thank the reviewer for encouraging a direct isolation of diffusion and DPS. We added a matched component experiment comparing full NeuPlume with direct LSM library search and CNF library search without diffusion. 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%, lowers normalized RMSE from 0.438 to 0.322, and raises spatial correlation from 0.910 to 0.957. The comparison quantifies the benefit of diffusion-based candidate generation under a matched 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.

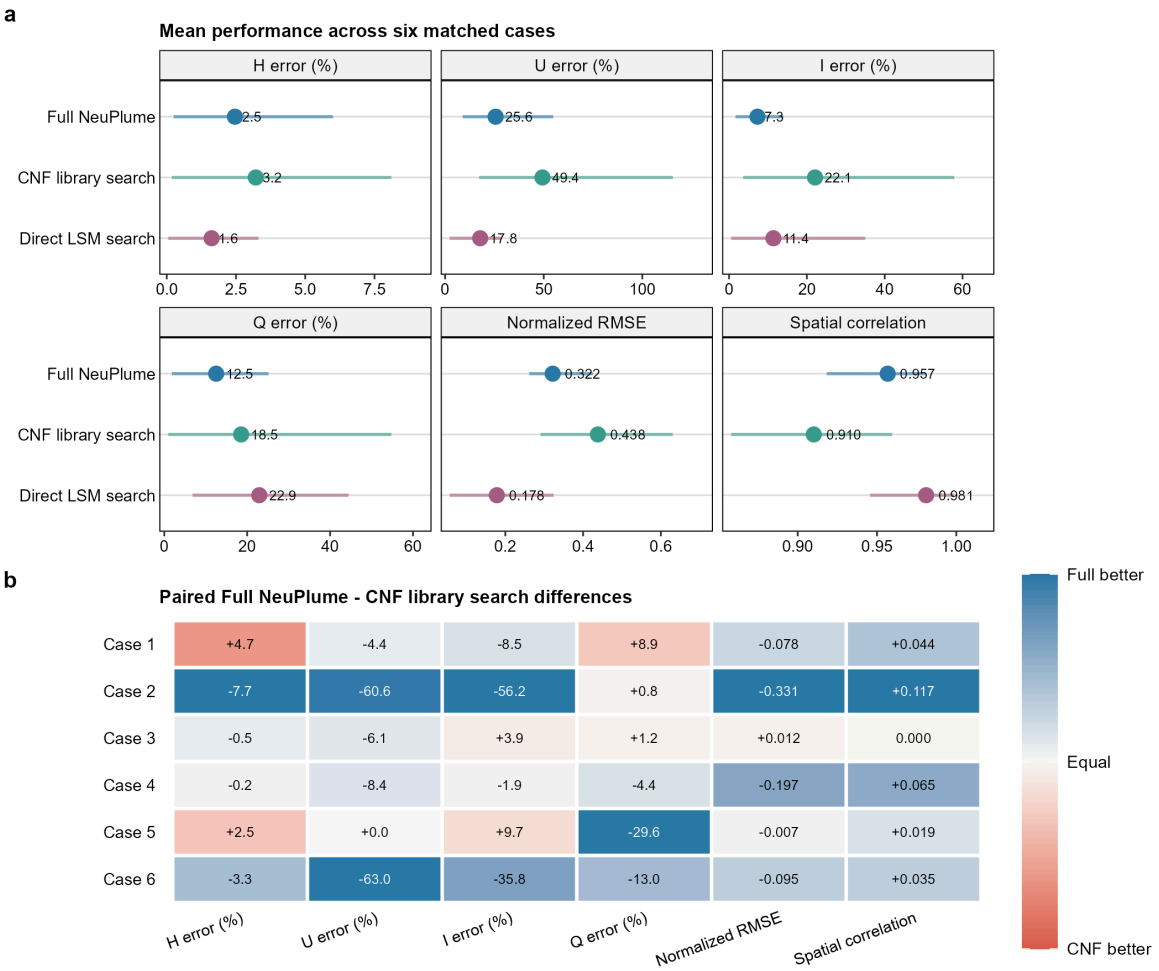


Fig. 4. Matched accuracy comparison for the six representative cases. (a) Mean absolute relative errors for effective release height H, wind speed U, turbulence intensity I, and emission rate Q, together with mean full-field normalized RMSE and spatial correlation; horizontal intervals span the minimum 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.

Major comment 3: baseline comparison mixes method, model, and geometry

Response. We thank the reviewer for pointing out that baseline performance can be confounded by method assumptions, forward-model matching, and observation geometry. We therefore made the operating assumptions explicit and used the same six representative cases and observations for the GP, mass-balance, direct LSM, CNF, and NeuPlume comparisons. Under random-volume sampling, mean emission-rate errors are 1,134.5% for blind GP, 329.3% for known-wind mass balance, and 12.5% for NeuPlume. GP and mass-balance methods are formulated for plume-crossing sections; their performance here reflects the evaluated random-volume geometry.

The matched Gaussian control verifies the GP implementation under its corresponding field model: known-wind GP recovers emission rate and plume height with errors below 0.001%, while blind GP retains emission-rate--wind-speed non-identifiability. Direct LSM and CNF searches provide simulation-based references. These controls distinguish implementation validity, forward-model matching, and random-volume performance.

The protocol is described in Sect. 3. The comparison and Gaussian control are reported in Sect. 4.1, Table 1, and Appendix Table C1; method positioning is discussed in Sects. 5 and 6.

Revised manuscript text.

Sect. 3, Experimental setup:

> Six representative validation cases span low and high release heights, wind speeds, and turbulence intensities within the validation design; their emission rates range from 0.75 to 1.25~$\mathrm{kg\,s^{-1}}$. The baseline and component comparisons use the same 1\% random-volume mask for each case. This geometry differs from plume-crossing transects and defines the interpretation of the GP and MB comparisons.
>
> NeuPlume is compared with GP and MB baselines using the same observations and meteorological inputs. GP minimizes residual sum of squares with differential evolution followed by L-BFGS-B. It is evaluated in blind and known-U modes. MB interpolates observations onto a cross-wind section and integrates concentration flux using the specified wind speed, so it is evaluated only in known-U mode. NeuPlume also uses blind and known-U modes; the latter fixes wind speed during candidate generation and grid search.

Sect. 4.1, Six-case benchmark results:

> The six representative cases support the GP and MB comparisons, Gaussian-plume control, and coarse-to-fine analysis. Across these cases, NeuPlume's mean absolute relative errors are 12.5\% for Q, 2.5\% for H, 25.6\% for U, and 7.3\% for I.

Case	NeuPlume Q	NeuPlume H	NeuPlume U	NeuPlume I	GP blind Q	GP blind U	GP known-U Q	GP known-U H	MB known-U Q
Case 1	9.8	4.9	14.3	12.6	62.6	54.4	51.9	0.4	146.4
Case 2	17.0	0.4	54.9	1.7	43.4	54.8	143.1	35.4	529.6
Case 3	6.2	0.2	15.0	7.6	81.3	88.5	30.8	7.5	299.2
Case 4	15.1	0.4	8.9	5.3	198.7	16.6	87.0	14.2	207.3
Case 5	25.1	6.0	47.3	14.0	559.4	440.9	234.9	49.3	450.1
Case 6	1.8	2.8	13.4	2.7	5861.9	48.2	278.3	1.6	343.3
Mean	**12.5**	**2.5**	**25.6**	**7.3**	**1134.5**	**117.2**	**137.6**	**18.1**	**329.3**
Median	**12.4**	**1.6**	**14.7**	**6.4**	**140.0**	**54.6**	**115.1**	**10.8**	**321.3**

> **Table 1 caption.** Inversion results for the six representative cases. Values are absolute relative errors (\%) against true values. NeuPlume, GP, and MB use the same 1,000 random-volume observations.

Appendix C, Gaussian-plume control:

> Table C1 reports the Gaussian-plume control for the six representative source conditions. The analytical Gaussian-plume model generated the observations used in this control.

Case	GP known-U Q error (\%)	GP known-U H error (\%)	GP blind inversion Q error (\%)	GP blind inversion U error (\%)
Case 1	< 0.001	< 0.001	43.49	43.49
Case 2	< 0.001	< 0.001	145.04	145.04
Case 3	< 0.001	< 0.001	137.83	137.83
Case 4	< 0.001	< 0.001	66.93	66.93
Case 5	< 0.001	< 0.001	228.09	228.09
Case 6	< 0.001	< 0.001	136.44	136.44
Mean	< 0.001	< 0.001	**126.30**	**126.30**

> **Table C1 caption.** Gaussian-plume control on analytical GP synthetic data. Values are absolute relative errors. Known-U errors below 0.001\% show parameter recovery when the plume model and wind input match the data generator.

Sect. 5, Method positioning and validation:

> NeuPlume targets near-field settings where a simulation-trained representation can describe non-Gaussian structure and several source and environmental parameters must be searched jointly. GP and MB remain efficient choices when their plume and sampling assumptions hold \cite{Shah2019, Krings2013, Conley2017}. The field comparison here is therefore descriptive: without a flight-specific certified emission rate, the spread among NeuPlume, GP, and MB estimates cannot establish absolute accuracy. A controlled-release experiment with a known reference emission rate for each flight would provide the decisive test of NeuPlume's absolute emission-rate accuracy under field conditions \cite{Morales2022, Vollrath2026}. Such validation should span wind regimes and observation geometries while retaining the minimum-loss selection rule used here.

Relevant figure.

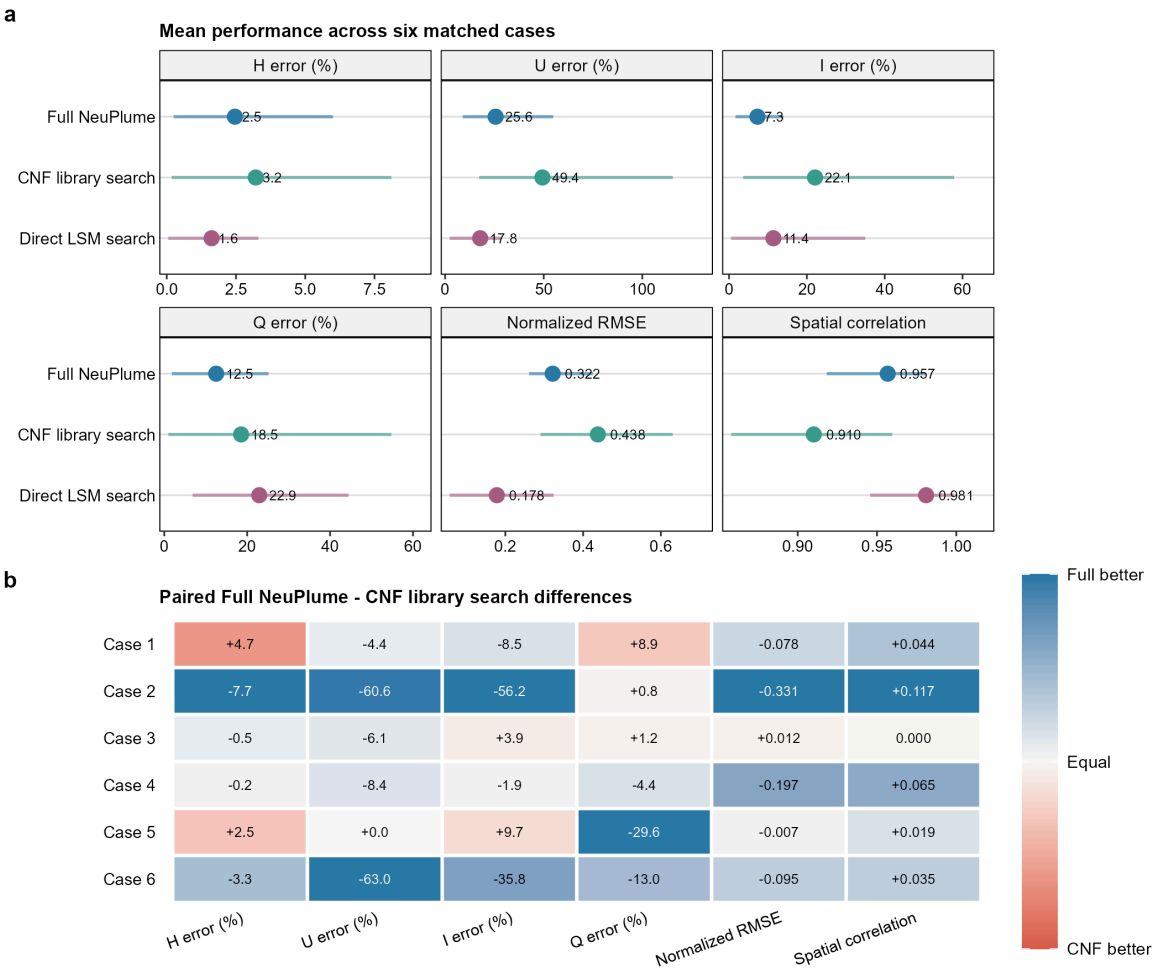


Fig. 4. Matched accuracy comparison for the six representative cases. (a) Mean absolute relative errors for effective release height H, wind speed U, turbulence intensity I, and emission rate Q, together with mean full-field normalized RMSE and spatial correlation; horizontal intervals span the minimum 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 4: bring field reconstruction into the main paper

Response. We thank the reviewer for recommending that complete field reconstruction be brought into the main paper. We added quantitative field metrics over 30 validation cases and now display paired LSM truth and NeuPlume reconstructions for all six representative cases in Fig. 3. The three-dimensional isosurfaces and ground and rear-plane projections use coordinate units and a common within-case concentration normalization. The six-case parameter errors are reported in Table 1.

Results are given in Sect. 4.1, Table 1, Figs. 2--3, and Appendix Tables B1--B2.

Revised manuscript text.

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.

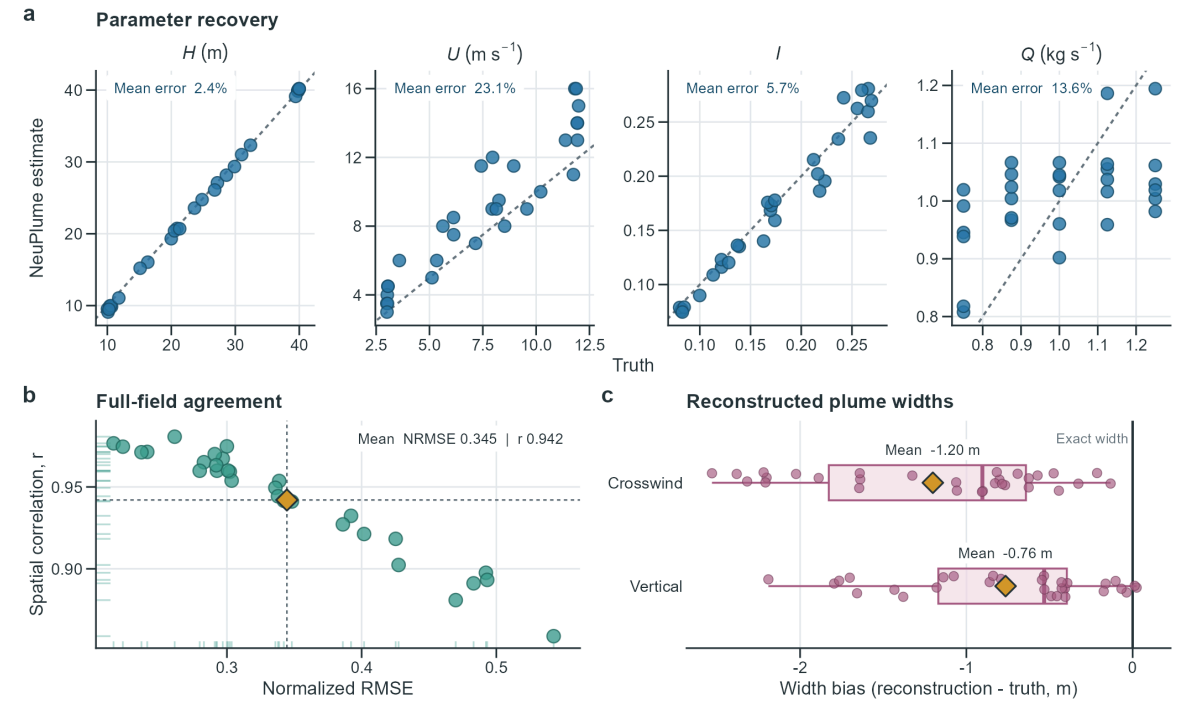


Fig. 2. Synthetic validation across 30 held-out cases. (a) Each point is one synthetic case. Estimated values are plotted against the true H, U, I, and Q; dashed diagonal lines denote exact recovery and inset labels give mean absolute relative errors. (b) Joint full-field normalized RMSE and spatial 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.

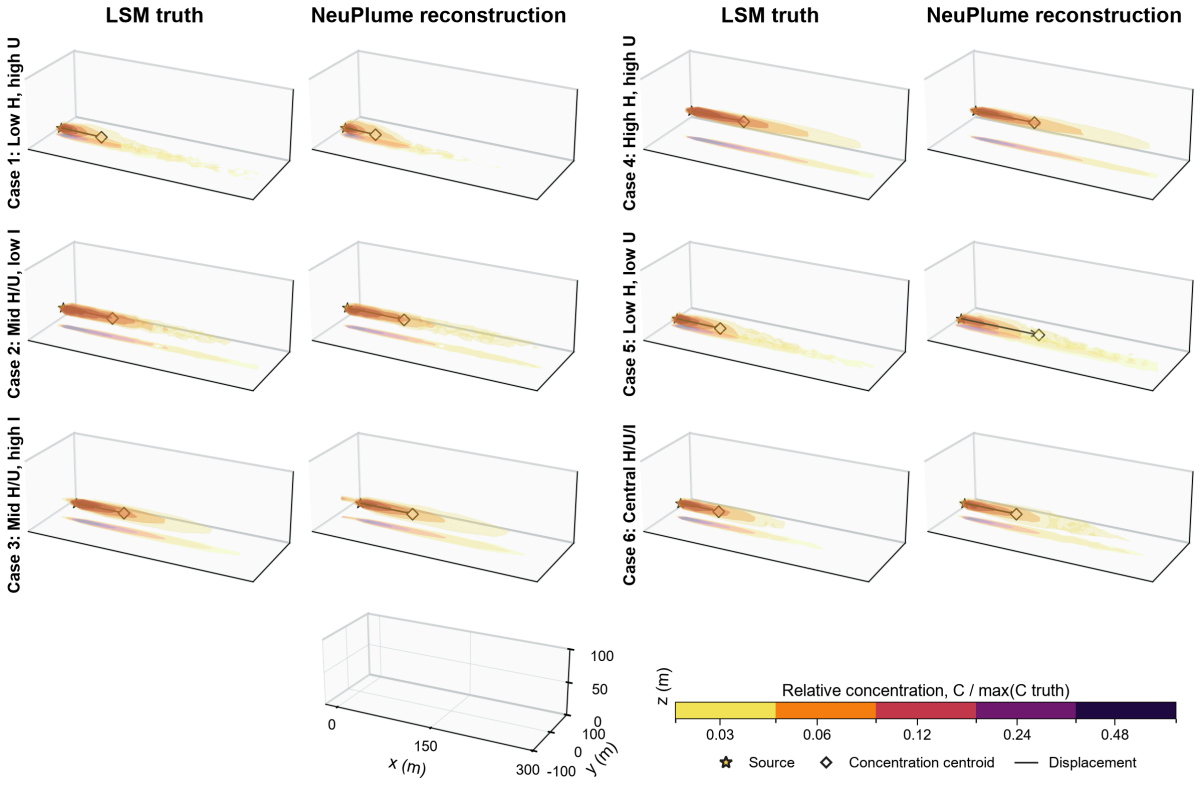


Fig. 3. Three-dimensional plume comparisons for the six representative cases. The first two columns compare the LSM truth and NeuPlume reconstruction for Cases 1--3: Case 1 (low H, high U), Case 2 (mid H/U, low I), and Case 3 (mid H/U, high I). The last two columns show the same comparison for Cases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

Response. We thank the reviewer for raising the question of uncertainty around calibration estimates. We clarified that the method specifies a point estimand and therefore reports point-estimation error distributions across the 30 validation cases, rather than posterior uncertainty intervals.

The 30-case point and field metrics are reported in Sect. 4.1, Fig. 2, and Appendix Table B1.

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. 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.

Relevant figure.

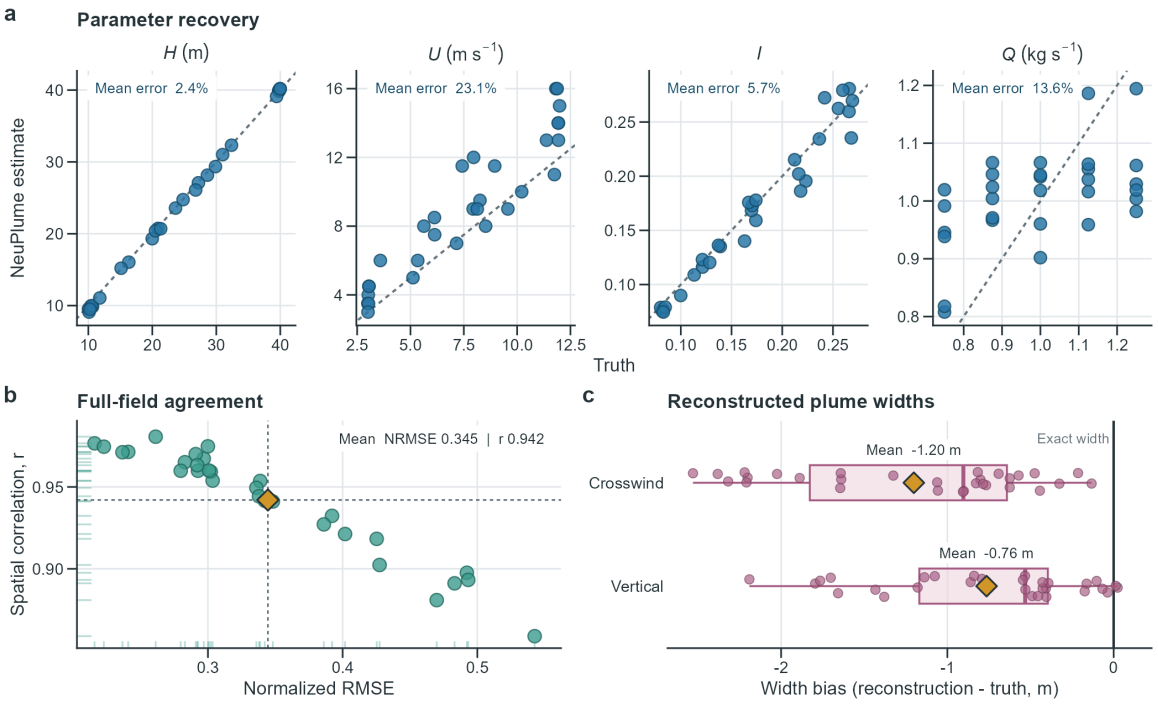


Fig. 2. Synthetic validation across 30 held-out cases. (a) Each point is one synthetic case. Estimated values are plotted against the true H, U, I, and Q; dashed diagonal lines denote exact recovery and inset labels give mean absolute relative errors. (b) Joint full-field normalized RMSE and spatial 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.

Additional comment 6: atmospheric measurement implications

Response. We thank the reviewer for asking us to connect the synthetic experiments to atmospheric measurement design. We added a joint analysis of measurement geometry, wind representativeness, transport mismatch, and boundary losses. Equal-budget trajectory sampling has higher mean normalized RMSE and lower mean spatial correlation than random-volume sampling in the six-case comparison. Strong shear changes selected parameters while preserving broad field metrics, whereas effective plume rise degrades height recovery and field reconstruction. Side and top boundary exits average 3.2% of exits and reach 10.2% in the most restrictive representative case.

These results support spatially complementary sampling and representative wind measurements for the 520 x 200 x 100 m domain, 1,000-point budget, transformed trajectories, and evaluated forward models. The 1% coverage and approximately 250 m downwind distance define the experimental design.

Results are reported in Sect. 4.1, Appendix Table E1, and Figs. F1 and G1--G2. Interpretation is discussed in Sects. 5 and 6.

Revised manuscript text.

Sect. 4.1, Synthetic data validation:

> 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.
>
> 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.

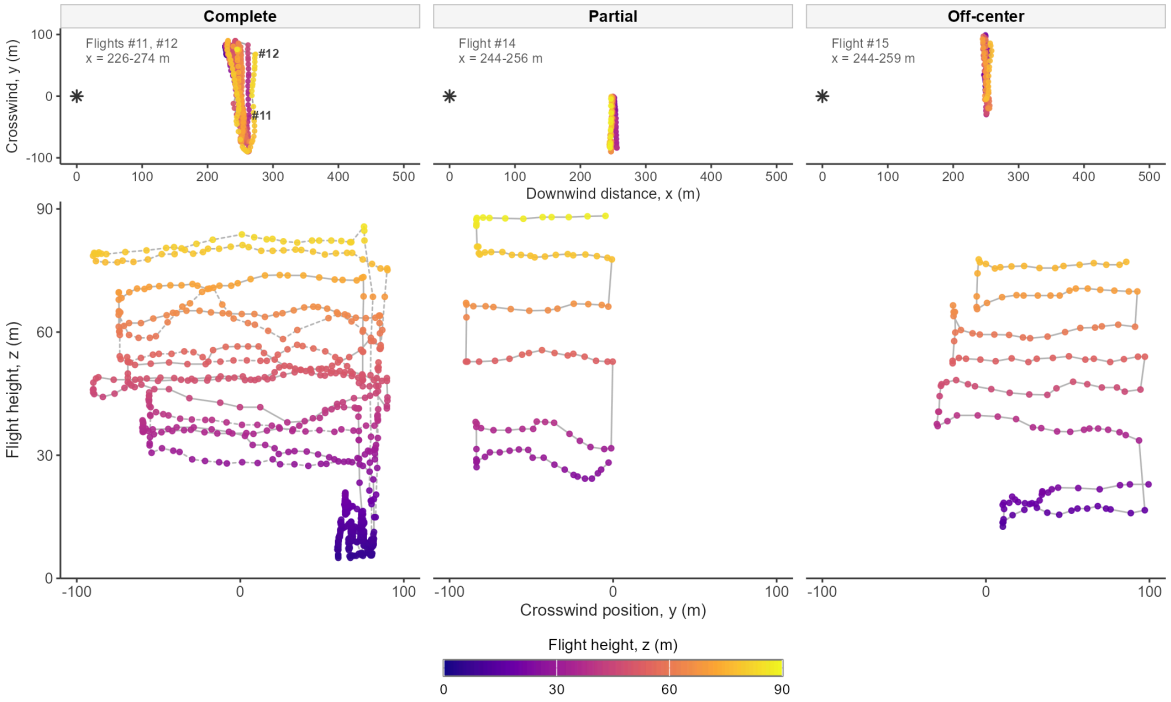
Sect. 5, Measurement geometry, wind representativeness, and identifiability:

> 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.

Appendix E, Table E1:

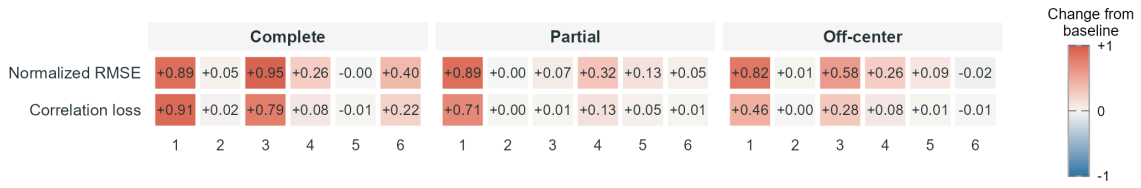
Metric	Mean	Minimum	Maximum
Particle exit fraction	0.920	0.913	0.928
Retained particle fraction	0.080	0.072	0.087
Side/top share of exits	0.032	1.4×10^{-5}	0.102
Downwind share of exits	0.968	0.898	1.000
Side/top share of emitted particles	0.029	1.3×10^{-5}	0.095

Relevant figures.

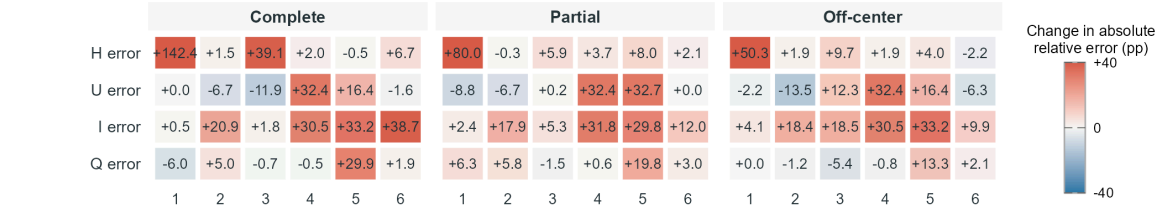


Appendix Fig. G1. Complete, partial, and off-center transformed UAV sampling trajectories used in the controlled observation-geometry stress test. The upper strips locate the trajectories in the downwind--crosswind plane; stars mark the source and annotations identify the source flights and sampled downwind 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.

a



b



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.

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.