

NeuPlume: ~~Probabilistic~~ Generative inversion of atmospheric point-source emissions from sparse observations

Lei Wang¹ and Xin Ma²

¹School of Earth and Space Science and Technology, Wuhan University, Wuhan 430079, China

²State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Correspondence: Xin Ma (maxinwhu@whu.edu.cn)

Abstract. Inverting point-source emissions from sparse atmospheric observations is ~~a central challenge because different combinations of~~ difficult because emission rate, release height, wind speed, turbulence ~~intensity~~, and plume morphology can ~~produce similar measurements. This paper presents~~ compensate for one another. We present NeuPlume, a neural-physical probabilistic inversion framework that characterizes the ensemble of plume fields ~~compatible with the data rather than returning only a point estimate~~ framework that formulates inversion as observation-guided generation and selection of complete three-dimensional concentration fields. A Lagrangian stochastic model ~~constructs the forward mapping from physical parameters to concentration fields;~~ builds a scenario-specific forward library, a conditional neural field compresses the simulated fields, and a latent diffusion model ~~then learn the feasible field prior~~ learns their conditional distribution. During inversion, diffusion posterior sampling incorporates sparse observations and jointly infers ~~emission rate, release height, wind speed, turbulence intensity, and the full concentration field~~. ~~Across six~~ generates candidate fields on a coarse-to-fine parameter grid, and NeuPlume returns the parameters and field that best match the observations. Across 30 synthetic cases, the ~~mean errors for emission rate, effective release height, and turbulence intensity are 10.0%~~ selected fields achieve a mean spatial correlation of 0.942 and a mean emission-rate error of 13.6%. Matched stress tests show that both observation geometry and transport-model fidelity govern inversion quality: trajectory-concentrated sampling weakens full-field reconstruction, while unrepresented transport alters parameter identification and, 1.4%, and 6.1%, respectively, lower than Gaussian plume and mass-balance baselines under the same volumetric observation protocol. On 30 holdout cases, the nominal 68% credible interval attains 73.3% empirical coverage. In a real-data application, NeuPlume analyzes four unmanned aerial vehicle (UAV) for effective plume rise, also degrades field reconstruction. Applied to four UAV methane transects above a coal-mine ventilation shaft. Using the same shaft hourly emission inventory of 13.3 ± 4.1 kt/yr as the comparator, posterior estimates from three flights fall within the inventory range; the remaining lower estimate is attributed to unrepresentative wind speed. NeuPlume therefore provides uncertainty-aware source-parameter constraints under near-field non-Gaussian conditions while propagating uncertainty to meteorological states, NeuPlume yields known- U minimum-loss emission estimates of $5.27\text{--}15.85$ kt yr⁻¹ and exposes wind-speed and height-boundary sensitivities under model mismatch. NeuPlume thus provides an extensible methodology for jointly estimating source parameters and concentration fields; additional transport regimes can be incorporated through regime-specific simulation ensembles and neural-model retraining.

1 Introduction

Quantifying greenhouse-gas emissions from individual point sources ~~is a central task in atmospheric monitoring, because such estimates support national supports~~ emission inventories, regulatory ~~compliance, and the verification of climate-mitigation commitments. Current approaches fall mainly into two classes: assessment, and verification of mitigation commitments.~~ Gaussian plume (GP) models ~~infer estimate~~ emission rate by fitting analytical dispersion kernels to downwind concentration measurements Shah et al. (2019). ~~The mass-balance~~ Mass-balance (MB) ~~method estimates emission rate by integrating the advective flux of a tracer methods integrate tracer flux~~ across one or more cross-wind planes Conley et al. (2017); Varon et al. (2018). Both ~~paradigms have been deployed on platforms ranging from ground-based sensors to satellites and, increasingly, approaches are used with ground, airborne, satellite, and~~ unmanned aerial vehicle (UAV) ~~platforms, as comprehensively reviewed by Shaw et al. (2021) observations Shaw et al. (2021).~~

However, under sparse Sparse near-field observations, point estimates of emission rate convey more certainty than the data can support. When the wind field, dispersion morphology, or observation cross-section is prescribed externally, sources of uncertainty that the data could otherwise constrain are removed from the inversion measurements challenge these methods because source parameters, meteorology, observation geometry, and plume morphology remain coupled. GP methods ~~also rely~~ on far-field assumptions, including commonly assume steady conditions, a spatially homogeneous wind field, and Gaussian cross-wind and vertical ~~concentration profiles. These assumptions break down in the near-field region within approximately 1 km of the source, where profiles. Near-field~~ turbulence and plume meandering produce non-Gaussian ~~concentration structures Yuvaraj et al. (2026) and structures Yuvaraj et al. (2026), while~~ rapidly varying winds violate the steady-state assumption Jia et al. (2025). ~~The MB method relaxes the Gaussian-shape assumption by directly integrating observed concentrations, but it remains MB methods avoid prescribing a Gaussian plume shape but remain~~ sensitive to incomplete plume interception, ~~nonuniform~~ background subtraction, and the ~~representativeness of the~~ wind speed used to convert ~~column concentrations concentration~~ into flux. UAV and controlled-release experiments ~~both~~ show that observation coverage, ~~sampling cross-sections, and wind-field inputs can cross-section geometry, and wind inputs~~ substantially affect methane quantification Mohammadloo et al. (2025); Liu et al. (2024). ~~In this setting, a point estimate of emission rate merely summarizes the sparse concentration information available and cannot characterize the structural uncertainty jointly determined by plume morphology, observation coverage, and wind-speed representativeness.~~

Beyond the modeling assumptions embedded in GP and MB frameworks, the treatment of wind speed forms an equally important source of error. Because the advective flux of a passive tracer is linearly proportional to local wind speed, wind-speed bias is transmitted directly to the inferred emission rate. A campaign using the AVIRIS-4 airborne imaging spectrometer ~~demonstrated this coupling through a formal uncertainty decomposition. Wind-speed uncertainty representativeness is especially important because source strength and effective advective speed are coupled in steady passive-tracer transport. In an AVIRIS-4 campaign, wind-speed uncertainty~~ contributed up to 99.4% of the ~~total~~ variance in cross-sectional-flux ~~inversion of methane~~ emission rate, far exceeding spectral-retrieval noise and plume-segmentation uncertainty Meier et al. (2026). Eastwood et al. (2025) directly measured plume advection velocity from sequential airborne imaging flights using optical flow; in their controlled-release

experiments, methane inversion Meier et al. (2026). Controlled-release experiments using sequential airborne imagery found that ERA5 reanalysis winds underestimated the winds underestimated effective plume transport velocity by approximately 28%. Crucially, existing GP, MB, and related inversion frameworks usually specify wind speed from meteorological observations, reanalysis products, or in-situ anemometers, and wind speed is rarely estimated jointly from the concentration data. Even if the forward dispersion model is accurate, emission estimates can inherit systematic biases from external wind-speed inputs.

Airborne methane detection and coal-mine remote-sensing studies of coal-mine point sources likewise identify wind-field error as a key source of quantification uncertainty Conrad et al. (2023); He et al. (2024). Likewise identify wind information as a major source of emission-quantification error Conrad et al. (2023); He et al. (2024). Jointly searching wind speed with source parameters provides a direct diagnostic of this coupling, although sparse observations may leave several parameter combinations with similar mismatch.

Point-source emission inversion under limited measurements is intrinsically underdetermined. Because sparse concentration samples constrain only local features of the plume field, emission rate, release height, wind speed, turbulence intensity, and field morphology remain coupled. Bayesian inversion provides a natural framework for handling this one established treatment of non-uniqueness by combining physical prior information with observational likelihoods to characterize a family of data-compatible solutions rather than returning a single optimum. In point-source monitoring, this probabilistic description matches the information content of the observations: the same UAV trajectory may support multiple source-height, wind-speed, and source-strength combinations, and the posterior distribution gives their relative credibility and correlations. a parameter prior and an observational likelihood. Markov chain Monte Carlo Xu et al. (2025), variational data assimilation Aghdasi et al. (2025), regional-network atmospheric inversion Whites et al. (2019), and ensemble-based ensemble source reconstruction De Meutter et al. (2021) have all been applied to emission-parameter inversion. Recent atmospheric source estimation; recent reviews compare these methods with optimization-based alternatives Xie et al. (2026). However, classical Bayesian implementations usually depend on manually specified priors, and their computational cost rises rapidly. Their computational cost grows with parameter dimension and state-space complexity, limiting applicability to joint inversion of state complexity, particularly when source terms, meteorological parameters, and concentration fields spatial concentration fields are inferred jointly.

Deep generative models provide another route: learning a offer a complementary route by learning low-dimensional prior representations from physical simulations and introducing limited measurements as likelihood constraints during sampling using measurements to guide candidate generation. Conditional diffusion models can generate calibrated posterior distributions of spatially heterogeneous permeability fields in Darcy-flow inverse problems Liu et al. (2025b), and score-based posterior sampling can reconstruct seismic wavefields from incomplete and noisy measurements Meng et al. (2024). More closely related to this work, conditional have been used for inverse reconstruction of permeability and seismic wavefields Liu et al. (2025b); Meng et al. (2024). Conditional neural field (CNF) latent diffusion frameworks use CNF compression, a latent diffusion prior, and diffusion combine field compression with diffusion posterior sampling (DPS) to generate and reconstruct spatiotemporal turbulent fields Du et al. (2024); Liu et al. (2025a). These studies show that complex physical fields can be encoded in a compact latent space and recovered under observational constraints. In the reconstruct turbulent fields from incomplete observations Du et al. (2024); Liu et al. (2025a).

95 . In atmospheric point-source setting, Zhao et al. (2026) applications, learned plume representations can preserve non-Gaussian structure that analytical plume models omit. Zhao et al. (2026), for example, combined a U-Net encoder with inverse Gaussian fitting to reduce single-source emission rate error from UAV data, but that framework still targets a point estimate and does not provide a joint posterior distribution over source terms, wind speed, and concentration fields. These advances suggest that point-source inversion can learn a plume prior from high-fidelity physical simulations, rather than prescribing plume morphology as Gaussian. During inversion, incomplete concentration data concentrates the posterior on control parameters and concentration fields compatible with both the learned prior and the measurements, characterizing not only the emission rate but also the compatible meteorological conditions, plume structure, and their joint uncertainty. UAV emission rate error.

This paper presents NeuPlume, a three-stage probabilistic-generative inversion framework for atmospheric point-source emissions. In the first stageFirst, a Lagrangian stochastic model generates a library of concentration fields covering physically relevant ranges and establishes the forward mapping from indexed by source and meteorological parametersto concentration fields. In the second stage, a conditional neural field (CNF). Second, a CNF compresses each field into a low-dimensional latent representation, and a denoising diffusion probabilistic model learns a field prior conditioned on the source and meteorological parameter vector. In the third stage, NeuPlume performs the conditional distribution of those representations. Third, NeuPlume runs observation-guided diffusion posterior sampling (DPS) DPS on a coarse-to-fine parameter grid. It generates concentration field samples for each grid node, analytically, estimates the emission rate , and weights the nodes by reconstruction loss to form a discrete approximate posterior over source parameters at each node, and returns the candidate with the lowest observation mismatch. The main contributions of this paper are:

1. We formulate near-field point-source emission inversion as joint posterior reconstruction of physically feasible plumes and their controlling parameters, thereby incorporating non-uniqueness under sparse observations into the inversion resultsource estimation as generative candidate-field inversion, jointly returning a minimum-loss parameter vector and its complete three-dimensional concentration field.
2. We provide a neural-physical implementation path that is not tied to a specific forward simulator: as long as the target scenario can be represented by a forward model mapping parameters to concentration fields, the resulting simulation ensemble can be used to learn the corresponding field prior and infer the parameter posterior under observational constraintsdefine a simulator-independent implementation path in which a forward model maps scenario parameters to the concentration-field library used to train a new CNF and diffusion model.
3. We evaluate inversion accuracy on six synthetic cases, showing that NeuPlume outperforms Gaussian plume and mass-balance baselines under the tested near-neutral conditions, and assess posterior calibration on a holdout set of minimum-loss parameter and field accuracy on 30 casesheld-out synthetic cases and use six representative off-grid cases within the trained parameter support for controlled baseline and component comparisons.
4. We test the framework onapply the framework to public UAV methane observations above a coal-mine ventilation shaft , compare posterior emission estimates with the same shaft hourly inventory, and use ablation experiments to identify

the observation-coverage conditions under which inversion quality degrades, and use matched stress tests to quantify sensitivity to observation geometry and unrepresented transport mechanisms.

130 The framework and its evaluation are presented in Sections 2–6.

2 Methods

2.1 NeuPlume model architecture

NeuPlume learns plume morphologies from forward simulations and ~~constrains the resulting physical prior with field measurements to produce a joint posterior over the three-dimensional concentration field and its controlling parameters~~ uses sparse measurements to guide the generation and selection of candidate concentration fields.

Computationally, NeuPlume organizes point-source emission inversion as a three-stage probabilistic framework that jointly estimates source parameters, meteorology, emission rate, and the complete concentration field from incomplete observations. The framework has three stages (Fig. 1). Stage 1 uses a Lagrangian stochastic model (LSM) to generate concentration fields. ~~Each field is simulated at a unit emission rate $Q = 1 \text{ kg s}^{-1}$ and indexed by $\theta = (H, U, I)$. Each field is indexed by~~ $\theta = (H, U, I)$, where H is the effective release height, U the reference wind speed, and I the turbulence intensity; together, these simulations define the physically plausible fields for the target scenario turbulence intensity. Stage 2 ~~trains a conditional neural field (CNF) to represent each concentration field as $f_\psi(\mathbf{x}; \theta)$. This implicit field representation provides a low-dimensional state space for the generative model. A uses Hash-INR to compress each field into a 192-dimensional latent code $\mathbf{z}$, and a denoising diffusion probabilistic model learns the distribution of these representations under training conditions and supplies a generative field prior. (DDPM) learns $p(\mathbf{z} | \theta)$.~~ In Stage 3 performs observation-guided diffusion sampling over candidate θ values and forms a discrete approximate posterior through analytical, sparse observations guide diffusion posterior sampling (DPS) at each candidate θ_k . NeuPlume decodes a candidate unit-emission field, fits Q estimation and reconstruction loss weighting by linear scaling, evaluates the observation mismatch, refines the search around the lowest-loss coarse-grid node, and returns $(\hat{\theta}, \hat{Q}, \hat{C})$ from the minimum-loss fine-grid candidate.

150 The corresponding inversion problem can be written as the following observation model.

Let $C(\mathbf{x} | \theta)$ denote the steady-state concentration field produced by the LSM under condition θ , $C(\mathbf{x} | \theta)$ denote a unit-emission concentration field and let $\mathcal{H}$ be the observation operator mapping the ~~map the~~ complete field to measurement locations, such as points along a UAV flight path. The observed concentration vector $\mathbf{y}$ follows the measurement locations. The observations satisfy

$$155 \quad \mathbf{y} = Q \mathcal{H}[C(\mathbf{x} | \theta)] + \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (1)$$

where ~~the ϵ collects measurement and representation residuals, and~~ linear scaling by Q follows from unit-rate normalization. ~~The goal is to recover the posterior jointly determined by θ , Q , and the underlying concentration field. The current implementation samples concentration fields from the generative prior at a given θ and represents parameter-level inversion~~

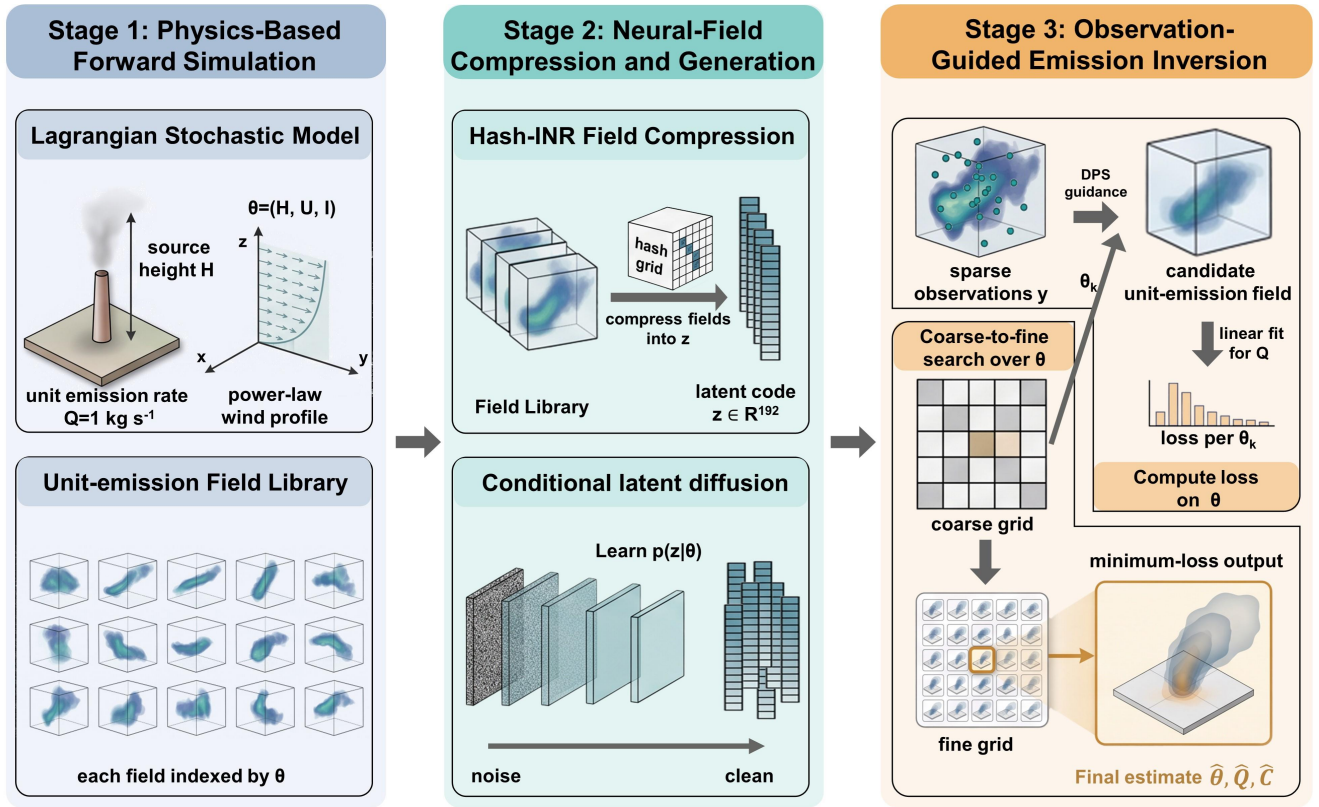


Figure 1. Overview of the three-stage NeuPlume framework. Stage 1 $\div$ the constructs a unit-emission field library with a Lagrangian stochastic model generates training concentration fields for 1000 parameter combinations (H, U, I) under a unit emission rate. Stage 2 $\div$ compresses the fields with Hash-INR conditional neural field compresses each field into a latent code, and the diffusion model learns the latent code their conditional latent distribution. Stage 3 $\div$ combines observation-guided diffusion posterior sampling combines the learned prior with sparse point observations to invert source parameters coarse-to-fine parameter search and reconstruct emission-rate fitting to obtain the complete three-dimensional concentration field minimum-loss reconstruction.

with source-parameter marginal posterior $p(\theta, Q | y)$. By Bayes' theorem, NeuPlume evaluates a finite candidate set produced by the conditional generative model and coarse-to-fine grid search. Its formal output is

$$p(\hat{\theta}, \hat{Q} | y, \hat{Q}, \hat{C}) \propto \arg \min_{(\theta_k, Q_k, C_k) \in \mathcal{G}} p(y | \theta_k, Q_k, C_k) \mathcal{L}_{\text{obs}}(\theta_k, Q_k, C_k), \quad (2)$$

NeuPlume implements this decomposition in two components: the learned implicit-field model defines a generative prior over physically consistent concentration fields, and the observation operator and noise model define the likelihood $p(y | \theta, Q)$. Computationally, DPS injects the observation-likelihood gradient into the reverse diffusion process, so generated samples match both the learned field distribution and the measurement data. The following subsections describe the LSM forward

simulator (Section 2.2), the CNF and diffusion training procedure (Section 2.3), and the inversion algorithm (Section 2.4).
 where $\mathcal{G}$ is the generated candidate set.

2.2 Lagrangian stochastic forward simulation

NeuPlume first takes source and meteorological parameters as inputs to generate concentration fields for neural field compression
 170 ~~and diffusion prior learning.~~ The forward simulator uses a Lagrangian stochastic model (LSM) to generate concentration fields
from source and meteorological parameters. LSMs represent turbulent dispersion with stochastic particle trajectories and must
 satisfy a stationary phase-space distribution, Eulerian-equation compatibility, and the well-mixed condition Thomson (1987).
 Established codes such as FLEXPART and STILT apply Lagrangian particle models to point-source dispersion, near-field
 source-receptor relationships, and flux inversion Stohl et al. (2005); Lin et al. (2003). Microscale comparisons of forward-
 175 in-time and backward-in-time LSMs ~~further confirm their suitability~~ support their use for local-scale dispersion and source-
 receptor sensitivity problems Li and Du (2020). Relative to the Gaussian plume formula, an LSM ~~can represent~~ represents
 near-field dispersion and moderate-turbulence non-Gaussian structures at a computational cost several orders of magnitude
 below large-eddy simulation Katata et al. (2015).

The model continuously releases particles from a point source at height H above the ground. Each particle carries position
 180 $\mathbf{x}$ and turbulent velocity fluctuation $\mathbf{u}'$, which evolves by a Langevin-type stochastic differential equation:

$$du'_i = -\frac{u'_i}{T_L} dt + \sqrt{\frac{2\sigma_i^2}{T_L}} dW_i, \quad i \in \{x, y, z\}, \quad (3)$$

where T_L is the Lagrangian integral time scale, σ_i is the directional velocity standard deviation, and dW_i denotes an in-
 dependent Wiener increment. The mean wind field follows the power-law profile $\bar{U}(z) = U_{\text{ref}}(z/z_{\text{ref}})^\alpha$; ~~under~~. We set
 $\alpha = 1/7 \approx 0.143$ for neutral atmospheric stability ~~, $\alpha = 1/7 \approx 0.143$, where and define~~ U_{ref} is the reference wind speed
 185 ~~at $z_{\text{ref}} = 10$ m. The $1/7$ exponent is often used as a neutral condition approximation, but the actual wind-shear exponent~~
~~varies with atmospheric stability and surface roughness. The exponent usually increases under stable conditions and decreases~~
~~under unstable conditions Newman and Klein (2014). Non-neutral boundary layers also alter turbulence statistics, drift terms,~~
~~and implementation of the well-mixed constraint Cassiani et al. (2014). Thus, extending NeuPlume to stable or unstable~~
~~stratification requires more than replacing the power-law exponent; the LSM training library must be regenerated under the~~
 190 ~~corresponding stability conditions and the CNF retrained.~~ 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 $c(\mathbf{x})$.

Because the steady-state concentration field of an inert tracer is linear in source strength, all training simulations set a unit
 emission rate $Q = 1 \text{ kg s}^{-1}$. Thus, the LSM, CNF, and diffusion model learn only the unit-rate concentration distribution deter-
 195 mined by $(H, U_{\text{ref}}, I_{\text{ref}})$. During inversion, a candidate emission rate Q linearly scales this unit field to give the corresponding
 concentration prediction.

With the unit-rate convention above, we constructed the training dataset required by the conditional neural field (Section 2.3).
 The training set covers the three-dimensional conditional parameter space spanned by ~~source~~ effective release height $H \in$

[0, 50] m, reference wind speed $U_{\text{ref}} \in [1, 15] \text{ m s}^{-1}$, and turbulence intensity $I_{\text{ref}} \in [0.05, 0.3]$. 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. ~~This setting applies to many tracer-release activities under moderate to strong winds, but the framework has not yet been evaluated under stable or unstable stratification (Section ??). A total of~~

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.

We uniformly sample 1 000 parameter combinations ~~are uniformly sampled~~ from this space. ~~This sample size covers the three-dimensional parameter space while keeping the LSM generation cost manageable. Smaller training sets were not systematically evaluated.~~ For each combination, the LSM is integrated over a $520 \times 200 \times 100 \text{ m}^3$ computational domain ($x \in [-20, 500]$, $y \in [-100, 100]$, $z \in [0, 100] \text{ m}$), with time step $\Delta t = 0.5 \text{ s}$, 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 $(H, U_{\text{ref}}, I_{\text{ref}})$ to form the supervised training set for neural field compression.

2.3 Conditional neural field training

Stage 2 ~~of NeuPlume~~ compresses the discrete ~~concentration fields generated by the LSM~~ LSM concentration fields into a continuous, conditionable ~~, low-dimensional representation to reduce storage and interpolation cost in diffusion modeling and inversion representation.~~ For each ~~source-meteorological condition $\theta = (H, U_{\text{ref}}, I_{\text{ref}})$~~ $\theta = (H, U_{\text{ref}}, I_{\text{ref}})$, an implicit ~~neural function $f_{\psi}(\mathbf{x}; \theta)$ maps any spatial coordinate $\mathbf{x} \in \mathbb{R}^3$~~ function $f_{\psi}(\mathbf{x}; \theta)$ maps $\mathbf{x} \in \mathbb{R}^3$ to concentration. ~~This representation preserves spatial continuity and conditional dependence while compressing each field for the diffusion model. Conditional~~ We implement this conditional neural field with Hash-INR; conditional neural fields have also been used for ~~spatial-dimension reduction and continuous-field representation~~ continuous representations of turbulence data Guo et al. (2026). ~~We implement this compression with Hash-INR.~~

Hash-INR ~~contains a~~ combines a 16-level multiresolution hash encoder ~~and with~~ a condition-concatenated MLP. ~~Spatial coordinates pass through a 16-level hash encoder. Each level has a learnable hash table with multilayer perceptron. Each hash level contains~~ two features per entry, ~~and resolutions span geometrically from 16^3 to 512^3 with resolutions spanning 16^3 to 512^3 voxels. Trilinear interpolation within levels and concatenation across levels produce a 32-dimensional spatial feature vector. Before concatenation, the conditional parameters θ are standardized to zero mean and unit variance with training-set statistics, then joined with the spatial and concatenation yield 32 spatial features. The standardized condition vector is concatenated with these features and a sample-specific 192-dimensional latent code $\mathbf{z}$. This 227-dimensional vector enters a~~ A four-hidden-layer MLP network with 256 ReLU units per layer ~~and ReLU activations, to predict~~ predicts $\log(1 + C)$ ~~at each query point. We also evaluated a SIREN architecture with periodic activations, but it was unstable in the current setting; Appendix ?? gives the comparison.~~

Hash-INR is ~~jointly-optimized-over-all- N~~ optimized jointly over all training conditions for 300 epochs. Network weights and ~~per-sample~~ latent codes use Adam ~~with~~ learning rates of 10^{-3} and 5×10^{-4} , respectively. The objective combines log-space mean squared error, a linear-space penalty ~~for high-concentration regions~~above the 90th concentration percentile, and ℓ_2 regularization of ~~the~~ latent codes. ~~The penalty is applied to values above the 90th percentile of the concentration distribution.~~ ~~For stability, the gradient norm is~~Gradients are clipped to 1, and the learning rate is halved after a validation plateau. ~~Key hyperparameters are summarized in~~

The complete training and inversion settings are listed in Appendix A (Table A1).

~~Key hyperparameters of the NeuPlume framework. Component Parameter Value Hash-INR Levels 16 Features/level 2 Layers 4 Hidden units 256 Latent dimension 192 Epochs 300 Diffusion Blocks 6 Heads 8 Hidden dim. 512 Steps 1000 Epochs 100 Inversion Coarse points 88 Fine points 75 DPS 1000 Scale 2.5 τ 1.0~~

After Hash-INR ~~converges, each training condition θ_i~~ training, each condition has a latent code $\mathbf{z}_i \in \mathbb{R}^{192}$ that, with the ~~shared network weights, that~~ encodes its concentration field. The set $\{\mathbf{z}_i\}_{i=1}^N$ is a low-dimensional representation of the ~~diffusion fields. We train a denoising diffusion probabilistic model (DDPM) to learn~~with the shared decoder weights. A DDPM learns $p(\mathbf{z} | \theta)$ and use it as the latent space prior for posterior sampling in Section 2.4~~over these codes~~. Inspired by the ~~CoNFILD latent space diffusion framework~~CoNFILD Du et al. (2024), the denoiser uses a ~~Transformer-based DiT architecture.~~ The 192 latent dimensions are sequence tokens, projected to a 512-dimensional hidden space and combined with learnable positional encodings. The denoiser has DiT architecture with six Transformer blocks, each with 8-head ~~eight-head~~ self-attention, cross-attention, and a feed-forward ~~network with expansion ratio~~expansion ratio of 4. The ~~diffusion timestep is embedded by a 256-dimensional sinusoidal encoding and a two-layer MLP, and θ by an independent MLP. Their sum is used as a single token context. The forward process uses a 1000-timestep~~192 latent dimensions are projected into a 512-dimensional hidden space. Diffusion timestep and condition embeddings form the context. Training uses a 1000-step linear noise schedule, with β increasing from 10^{-4} to 2×10^{-2} . The diffusion model is trained for $\beta \in [10^{-4}, 2 \times 10^{-2}]$, 100 epochs~~with~~, batch size 64, Adam ~~with~~ a 10^{-4} learning rate, and cosine annealing. ~~Latent codes and condition vectors are z-score normalized before training.~~

~~The diffusion model uses classifier-free guidance (CFG) to improve conditional consistency during inference. During training, Classifier-free guidance is trained by zeroing the condition vector is zeroed with probability p_{drop} , so a single denoiser learns both.~~ During generation, conditional and unconditional ~~noise prediction. During inference, the two predictions are com-~~ bined with guidance scale $s > 1$: as

$$\hat{\epsilon} = \epsilon_{\emptyset} + s(\epsilon_{\theta} - \epsilon_{\emptyset}), \quad (4)$$

where $\epsilon_{\emptyset}$ and ϵ_{θ} denote unconditional and conditional predictions, respectively. Together, the trained Hash-INR decoder and latent diffusion model form a conditional generative model: for any s is the guidance scale. The trained DDPM generates latent candidates at a specified θ , the reverse diffusion process draws latent samples from $p(\mathbf{z} | \theta)$, and Hash-INR decodes them into complete concentration fields ~~, which for the inversion procedure in Section 2.4 combines with observational constraints. The Hash-INR and diffusion architectures and training hyperparameters are listed in Table A1.~~

2.4 Point-source emission inversion

NeuPlume ~~performs inversion in the candidate~~ searches a finite source and meteorological parameter space grid. For each candidate θ , the conditional diffusion model generates ~~concentration-field-samples-a concentration field~~ in the Hash-INR latent space. The algorithm estimates the emission rate Q ~~is estimated analytically~~ from the linear ~~relationship between the observed concentration and the relation between the~~ predicted unit field. ~~The reconstruction loss is then compared across candidate parameters to construct a discrete approximate posterior over source parameters. The result is a weighted set of physically plausible solutions that explain the observations and the observations, evaluates the observation mismatch, and returns the lowest-loss candidate field and parameter vector.~~

The generative backbone ~~of NeuPlume follows the idea of latent space~~ follows latent-space diffusion for conditional neural fields, ~~running the reverse diffusion process in the latent space of INR weights to generate conditional concentration-field samples~~ Du et al. (2024); Liu et al. (2025a). Observational constraints are introduced through DPS, which incorporates the likelihood gradient of a known forward observation operator into the reverse diffusion process. This step balances the generative prior and measurement consistency Chung et al. (2023). ~~Du et al. (2024); Liu et al. (2025a).~~ Starting from Gaussian noise $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ~~z_T~~ , the reverse process uses the noise-prediction network $\epsilon_\phi(z_t, t, \theta)$ ~~to iteratively denoise the latent vector~~ $\epsilon_\phi(z_t, t, \theta)$. Classifier-free guidance combines conditional and unconditional predictions:

$$\hat{\epsilon}_t = \epsilon_\phi(z_t, t, \mathbf{0}) + s[\epsilon_\phi(z_t, t, \theta) - \epsilon_\phi(z_t, t, \mathbf{0})], \quad (5)$$

where s is the guidance scale. The standard DDPM relation gives the predicted clean latent vector $\hat{z}_0$ ~~is obtained through the standard DDPM relation~~ $\hat{z}_0 = (z_t - \sqrt{1 - \bar{\alpha}_t} \hat{\epsilon}_t) / \sqrt{\bar{\alpha}_t}$ and is decoded by the INR to produce, which the INR decodes into a concentration field, ~~which is then compared with the observations. When observation guidance is enabled, the observation-likelihood gradient is injected.~~

NeuPlume uses diffusion posterior sampling (DPS) Chung et al. (2023), which introduces the gradient of the observation-mismatch objective into the reverse process ~~to guide sampling. Following the DPS framework and score-based posterior sampling in related geophysical inverse problems Meng et al. (2024), the noise update is modified as:~~

$$\epsilon_t^* = \hat{\epsilon}_t - \zeta \nabla_{z_t} \mathcal{L}_{\text{obs}} \cdot \sqrt{1 - \bar{\alpha}_t} \sqrt{1 - \bar{\alpha}_t}, \quad (6)$$

where ζ controls ~~the strength of observation guidance, and $\mathcal{L}_{\text{obs}}$ is the observation loss defined below. Because each stochastic reverse trajectory produces a distinct ψ sample, running multiple trajectories from independent noise realizations yields an ensemble of concentration fields, providing candidate samples for subsequent observation guidance. DPS supplies observation-guided candidate fields for~~ grid-node loss evaluation and approximate posterior construction. This ensemble represents possible plumes ~~compatible with the observations under the given candidate parameters.~~ evaluation, and the minimum-loss rule in Eq. (2) produces a point estimate.

The observation operator maps the generated concentration field to measurement space. Given a candidate parameter vector $\theta = (H, U_{\text{ref}}, I_{\text{ref}})$ and emission rate Q , the CNF-reconstructed field is For a generated unit field $c(\theta)$ sampled at M observation locations ~~and scaled by Q , giving~~ $\hat{y} = Q \cdot \mathcal{H}[c(\theta)]$, where $\mathcal{H}$ denotes spatial interpolation at

the sensor locations. The observation loss is defined in are $\hat{y} = Q\mathcal{H}[c(\theta)]$. Candidate quality is measured with a smoothed L_1 (Huber) form: observation-mismatch objective:

$$\mathcal{L}_{\text{obs}}(\theta, Q) = \frac{1}{M} \sum_{j=1}^M h\left(\hat{y}_j - y_j^{\text{obs}}\right), \quad h(e) = \begin{cases} \frac{1}{2}e^2 & |e| \leq 1, \\ |e| - \frac{1}{2} & \text{otherwise.} \end{cases} \quad (7)$$

This loss behaves as mean-squared error (MSE) for small residuals and transitions to L_1 loss for large errors, reducing sensitivity to outlying observations. Because the concentration field of an inert tracer in steady transport The Huber form limits the influence of large residuals. The field of a passive tracer scales linearly with source strength, for each candidate θ , the emission rate so Q is analytically estimated as initialized by the least-squares scaling factor between predicted and observed concentrations, rather than being included in the grid search. The remaining parameters are estimated with a coarse-to-fine grid search strategy. In the coarse-search stage, the algorithm between the generated unit field and the observations and may be refined during the reverse process.

The coarse stage evaluates a broad grid covering physically reasonable ranges of H , U_{ref} , and I_{ref} . For each grid node, one or more DPS reverse trajectories are executed, and the lowest reconstruction loss is recorded. The fine-search stage refines the search around the maximum a posteriori (MAP) estimate from the coarse-search stage, using a denser grid and additional DPS samples at each node, thereby improving the precision and reliability of subsequent uncertainty estimation. 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.

NeuPlume supports two complementary inversion modes. In known- U mode, meteorological measurements fix the known wind speed mode, the reference wind speed is provided by meteorological measurements and kept fixed during inversion. The search space is reduced to (H, I_{ref}) . Because a major source of degeneracy is removed, the grid search can converge efficiently. This mode is appropriate when reliable wind-speed observations are available at or near the measurement points. In and the blind search covers (H, I_{ref}) , In blind mode, the conditional parameters $(H, U_{\text{ref}}, I_{\text{ref}})$ are jointly searched without external wind-speed data, while Q is still analytically recovered at each candidate node. The blind mode is more broadly applicable, but it faces the inherent search covers $(H, U_{\text{ref}}, I_{\text{ref}})$. Blind inversion exposes the H - U - Q degeneracy: different combinations of source height, wind speed, and emission rate can produce similar downwind concentration patterns. Coarse-to-fine grid search partially mitigates this degeneracy by systematically exploring the parameter space. Compared with the known-wind-speed case, the residual ambiguity appears as a wider approximate posterior distribution. Both modes share the same generative backbone (DDPM + CFG), and only the dimension and range of the parameter grid differ.

2.4.1 Uncertainty quantification

because different parameter combinations can yield similar downwind concentrations. The multi-sample grid search provides the discrete loss set needed to construct an approximate posterior over source parameters. Limited measurements may simultaneously support multiple source-term and environmental-parameter combinations, so the posterior distribution itself is part of the

inversion result rather than only an error bar after point estimation. At each grid node θ_k , the best reconstruction loss ℓ_k from N independent DPS trajectories serves as a surrogate for the negative log-likelihood. A set of Boltzmann-type weights is then computed:-

$$w(\theta_k) \propto \exp\left(-\frac{\ell_k - \ell_{\min}}{\tau}\right),$$

where $\ell_{\min} = \min_k \ell_k$ improves numerical stability, and τ is the energy-scale parameter controlling the sharpness of the resulting distribution. The weighted mean and standard deviation of each marginal parameter distribution are obtained directly from these weights. The shape of the joint distribution is then used to diagnose parameter correlations, multimodality, and identifiability differences caused by the observation geometry. This construction is a discretized Gibbs distribution on the evaluated grid, not a strict Bayesian posterior. We accordingly term it an *approximate posterior* observation-mismatch surface and the distance between competing low-loss candidates describe this ambiguity.

The default energy scale $\tau = 1.0$ was selected to balance posterior sharpness and coverage reliability. To verify the calibration of the approximate posterior, we adopted a holdout protocol: 30 synthetic test conditions that were independently drawn from the training distribution and were not used in any parameter tuning were inverted, and the empirical coverage of nominal credible intervals was evaluated. A well-calibrated approximate posterior should produce coverage close to the nominal level. Detailed calibration results are reported in Section ??.

3 Experimental setup

The experimental setup includes synthetic cases, real UAV observations, method configurations, and evaluation metrics. These experiments test inversion accuracy against known truth, posterior calibration across conditions, and emission rate consistency in field observations. For synthetic validation, we constructed six test cases whose source and meteorological parameters cover limiting conditions in the Lagrangian stochastic model (LSM) training distribution ($H \in [0, 50]$ m, $U_{\text{ref}} \in [1, 15]$ m s⁻¹, $I_{\text{ref}} \in [0.05, 0.30]$). Under neutral atmospheric stability, the LSM simulates particle dispersion in a $520 \times 200 \times 100$ m³ domain and discretizes the

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 $H = 10\text{--}40$ m, $U_{\text{ref}} = 3\text{--}12$ m s⁻¹, $I_{\text{ref}} = 0.08\text{--}0.27$, and five emission rates from 0.75 to 1.25 kg s⁻¹. Each time-averaged concentration field onto 10^5 grid points. Table 1 lists the true parameters of each case. Case field contains 10^5 evaluation points in the $520 \times 200 \times 100$ m³ domain, and 1 represents typical mid-range conditions ($H = 26.8$ m, $U = 7.7$ m s⁻¹, $I = 0.132$). The other cases represent high turbulence intensity (Case 2, $I = 0.251$, and Case 6, $I = 0.294$), a low source with weak wind (Case 3, $H = 9.4$ m, $U = 4.0$ m s⁻¹), a high source under strong wind (Case 4, $H = 47.9$ m, $U = 14.1$ m s⁻¹), and extremely low wind speed (Case 5, $U = 1.3$ m s⁻¹). % of these points (1 000 observations) are selected by a case-specific random volumetric mask.

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 kg s⁻¹ (Table B2). The emission rate is fixed at $Q = 1.0$ kg s⁻¹ in all cases, so performance differences mainly arise from differences in meteorological parameters and source-height conditions. For each case, the LSM generates the full concentration field, and a subset of grid points is randomly sampled as virtual observations to mimic sparse, irregular data in UAV measurements; the corresponding full-field morphologies are shown in Figure ???. This three-dimensional random sampling protocol tests NeuPlume's ability to recover source terms from volumetric concentration data. It differs from cross-sectional flux flight geometries, so the synthetic baseline results should be interpreted in light of the observation geometry. All synthetic inversions use a fixed random seed (seed = 42) for reproducibility. Cross-condition robustness is verified with an independent 30-case uncertainty quantification calibration experiment (Section ??). 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.

The real-data experiment uses a methane measurement campaign at 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 Pniówek coal mine (shaft V) in the Upper Silesian Coal Basin, Poland. We selected four AirCore-UAV flights above shaft V (in Poland Andersen et al. (2023)). 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 10⁵ evaluation points. The complete, partial, and off-center protocols transform trajectories from flight #11, or #12, #14, and #15, each covering different meteorological conditions and plume-interception quality. All four flights were conducted on 21 May 2018 under moderate wind conditions (ground-station $U = 6-7$ m s⁻¹). This wind-speed range falls within the near-neutral training range assumed by the LSM library, but atmospheric stability on the measurement day was not directly measured. Therefore, a mismatch between the real boundary-layer state and the neutral training library cannot be excluded; this limitation is discussed in Section ???. In preprocessing, we reversed the raw analyzer concentration sequence and time-aligned it with the GPS trajectory, assigning 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.

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 $w(x) = 0.6 \exp(-x/80)$ 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.

The field experiment applies NeuPlume to the methane measurements from these four flights. Ground-station wind speed was 6–7 m s⁻¹. Preprocessing reverses the analyzer sequence, aligns it with GPS time, and assigns each methane observation to a spatial location along the flight path. External comparisons include the corresponding flight location. From the same study,

descriptive external references are the same-shaft hourly emission inventory ($13.3 \pm 4.1 \text{ kt yr}^{-1}$) inventory ($13.3 \pm 4.1 \text{ kt yr}^{-1}$) and the published inverse campaign-level integrated Gaussian (IG) plume-fitting result ($9.5 \pm 3.5 \text{ kt yr}^{-1}$, $9.5 \pm 3.5 \text{ kt yr}^{-1}$) $n = 13$) and mass-balance (MB) integration result ($9.8 \pm 5.0 \text{ kt yr}^{-1}$, $9.8 \pm 5.0 \text{ kt yr}^{-1}$, $n = 14$) from the same campaign Andersen et al. (2023) estimates.

NeuPlume is compared with two baseline methods GP and MB baselines using the same observation data observations and meteorological inputs. The GP model assumes an analytical steady-state plume solution and inverts Q , H , and U by minimizing the GP minimizes residual sum of squares between predicted and observed concentrations. GP inversion uses differential evolution for global search, followed by gradient-based local refinement (with differential evolution followed by L-BFGS-B) initialized at the global optimum. The MB method interpolates observed concentrations onto a two-dimensional cross-section perpendicular to the mean wind direction and estimates the emission rate by integrating the resulting concentration flux. Q is obtained directly from the integrated flux without an explicit plume model. GP inversion is run in two modes. In blind-estimation mode, all parameters are estimated only from observations. In known- U mode, wind speed is specified from meteorological station data to isolate the effect of wind speed uncertainty on emission estimation. The MB method is run only in known- U mode. 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 because it requires wind speed to calculate advective flux through the cross-section. NeuPlume also supports both modes, and its known- U variant modes; the latter fixes wind speed during diffusion posterior sampling. Thus, candidate generation and grid search.

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 comparison covers both fully data-driven inversion and inversion constrained by a given wind speed; 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. Stress-test degradation is defined as the tested value minus the matched baseline for parameter errors and normalized RMSE, and as the baseline minus the tested value for spatial correlation; positive values therefore denote worse performance for every displayed metric. For field cases, ground-truth parameter errors are unavailable. We report the minimum-loss NeuPlume estimates and their descriptive differences from the inventory center and range and from the published IG and MB references.

Evaluation metrics are defined by data source. For synthetic cases, parameter-estimation error is measured as the relative error (in-

4 Results

4.1 Synthetic data validation

Table 1 summarizes inversion results for the six synthetic test cases. All cases assume neutral atmospheric stability and span a range of source and meteorological conditions: wind speed $U = 1.3\text{--}14.1 \text{ m s}^{-1}$. Across the 30 synthetic validation cases, the mean absolute relative errors were 2.39% for effective release height $H = 9.4\text{--}47.9 \text{ m}$, and turbulence intensity $I = 0.08\text{--}0.29$.

The design includes a near-calm low-source release (Case 3), a high-source high-wind-speed release (Case 4) H , 23.13% for wind speed U , 5.74% for turbulence intensity I , and strong-turbulence conditions (Cases 2–6), testing how the learned field prior constrains parameter posterior distributions under different morphologies. NeuPlume’s coarse-to-fine (c2f) inversion jointly estimates all source parameters and achieves a six-case mean relative error of 10.0–0.76 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.

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 6), testing how the learned field prior constrains parameter posterior distributions under different morphologies. NeuPlume’s coarse-to-fine (c2f) inversion jointly estimates all source parameters and achieves a six-case mean relative error of 10.0–0.76 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 ?? shows the Lagrangian stochastic model ground-truth concentration fields for the six cases as maximum-intensity projections (MIPs). The fields show asymmetric, non-Gaussian near-source dispersion, revealing spatial structures that are poorly captured by a Gaussian plume assumption. Figure B1 reports the grid-search loss surfaces from diffusion-posterior sampling. In each case, the loss in the (U, H) plane has a single, clear minimum aligned with the true parameter location. This unimodal-loss structure indicates that, within the investigated range, the posterior mode provides a stable point estimate for the source parameters. Together with the error statistics in Table 1, these results show that NeuPlume constrains source parameters by preserving non-Gaussian plume structures in the learned field prior.

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.

4.1.1 Parameter accuracy Six-case benchmark results

Coarse-to-fine search improves the estimation accuracy of Q and H . With coarse grid search alone, the mean relative error of The six representative cases support the GP and MB comparisons, Gaussian-plume control, and coarse-to-fine analysis (Table 1). Across these cases, NeuPlume’s mean absolute relative errors are 12.5% for Q is 25.0–

This gain is most pronounced in cases with strong turbulence or low release height. For the strong-turbulence cases (Cases 2 and 6, $I > 0.25$) and the low-release-height case (Case 3, $H = 9.4$ m), broader or near-ground plumes produce flatter coarse-resolution loss surfaces, reducing coarse search’s ability to distinguish the optimum. In these cases, coarse-only Q errors all exceed 34–

The parameters also differ in identifiability. 2.5% for H is the most stable parameter across the, 25.6% for U , and 7.3% for I .

The fine stage selects a lower minimum loss in every case. Across the six cases, with a relative-error range of 0.3–4.8, mean minimum loss decreases from 0.172 to 0.142; the case-wise decrease ranges from 0.001 to 0.080. Appendix Fig. B1 reports

Lagrangian-stochastic-model ground-truth concentration fields for six synthetic test cases (top row: Cases 1–3; bottom row: Cases 4–6), shown as z -axis maximum-intensity-projections (MIPs). The cases cover non-Gaussian plume morphologies.

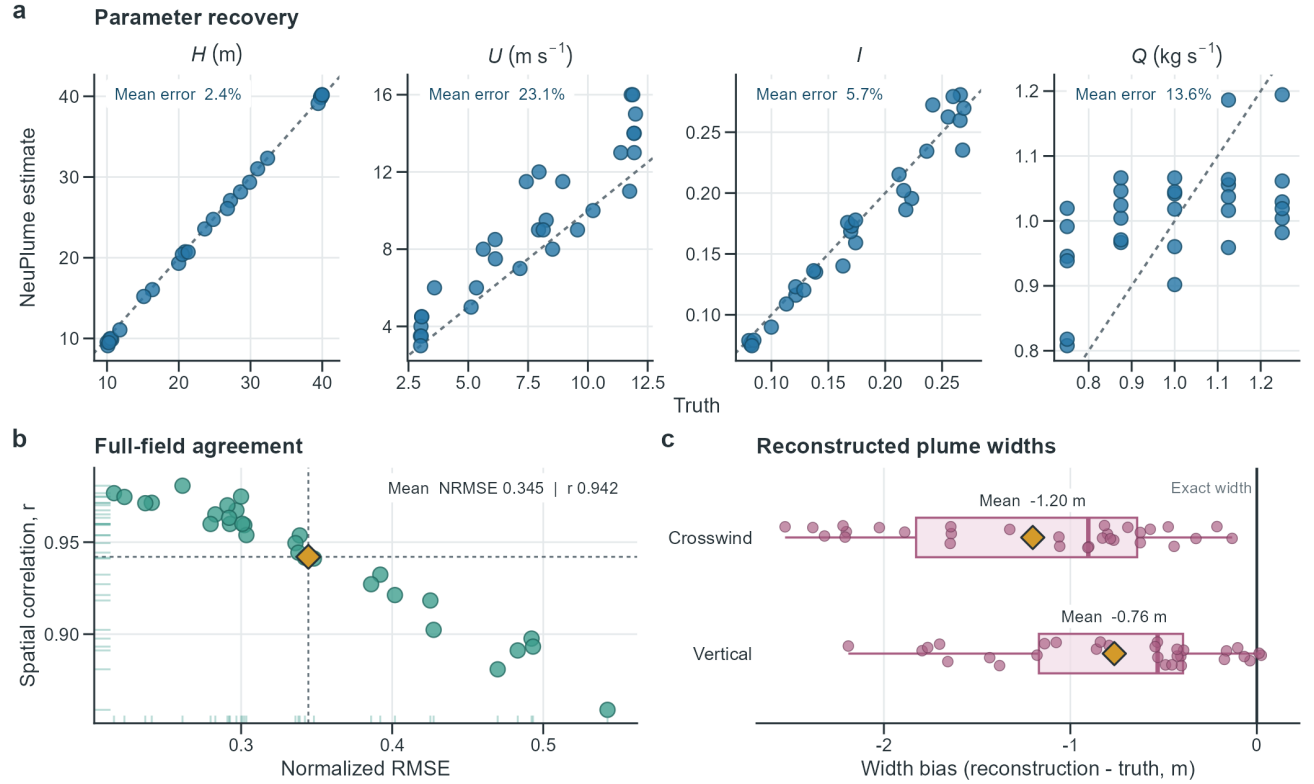


Figure 2. Synthetic validation across 30 held-out cases. (a) Each point is one synthetic case. Estimated values are plotted against the true H , U , I , and Q ; dashed diagonal lines denote exact recovery and inset labels give mean absolute relative errors. (b) Joint full-field normalized RMSE and spatial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

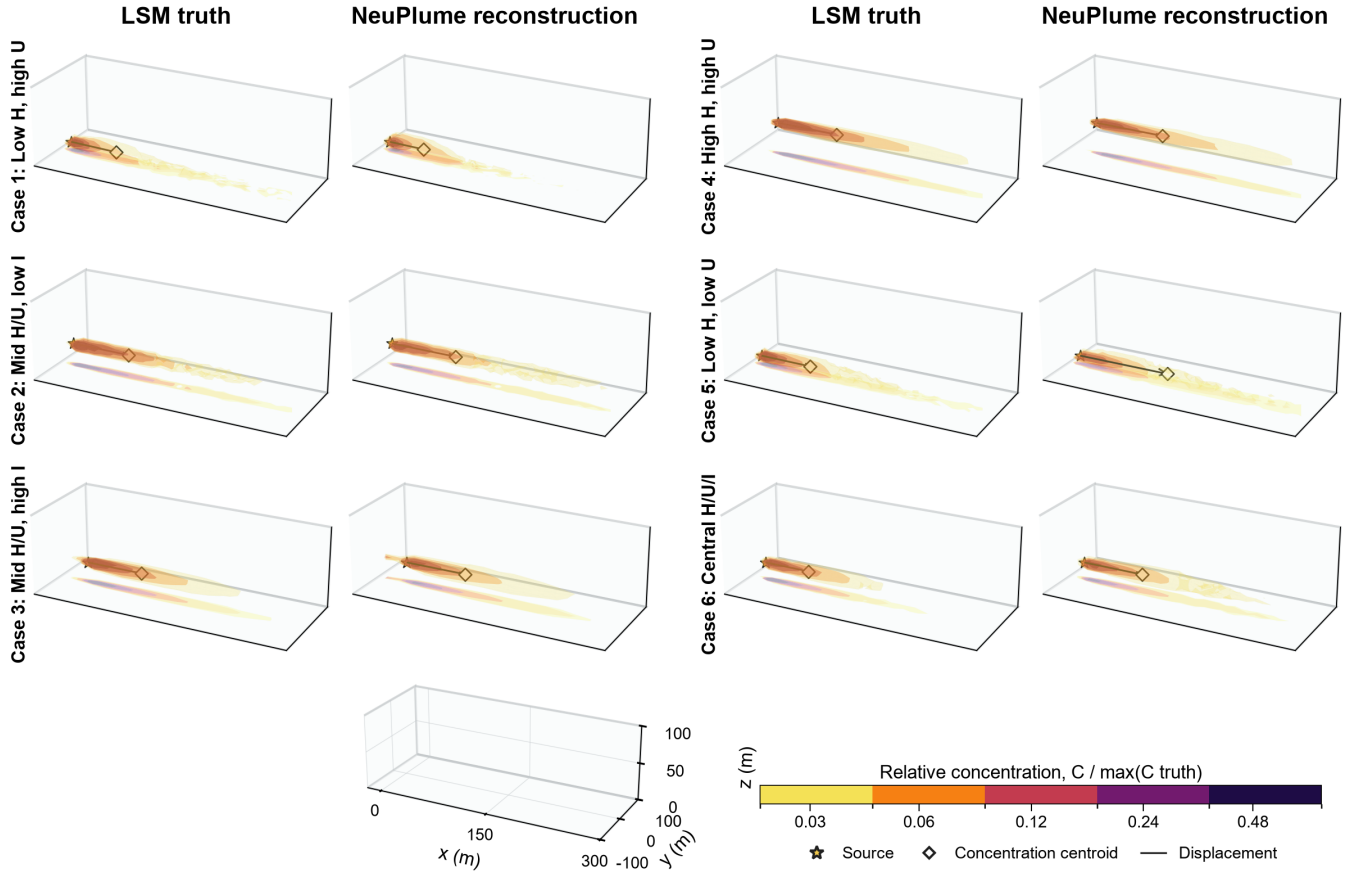


Figure 3. Coarse-to-fine grid search loss surfaces—Three-dimensional plume comparisons for the six synthetic test representative cases in-. The first two columns compare the (H, U) -parameter-plane LSM truth and NeuPlume reconstruction for Cases 1–3: Case 1 (low H , high U), Case 2 (mid H/U , low I), and Case 3 (mid H/U , high I). The cost functions from diffusion posterior sampling form a clear minimum aligned with last two columns show the true value same comparison for Cases 4–6: Case 4 (white diamond high H , high U), indicating a stable unimodal objective in Case 5 (low H , low U), and Case 6 (central $H/U/I$). Nested isosurfaces show the investigated range 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.

465

the case-wise changes and the coarse and local fine parameter spaces for the largest- and smallest-improvement cases. Table 1 reports the corresponding parameter accuracy.

Table 1. Inversion results for the six synthetic test representative cases. Values are absolute relative errors (%) against true values. Dashes mark parameters not reported by NeuPlume, GP, and MB use the baseline same 1 000 random-volume observations.

NeuPlume c2f GP blind GP known U MB known U Case Q H I Q U Q H Q 1—Typical 14.2 1.1 1.0 1836 —383 —525 2—High I 10.9 1.3 4.0 71 —232 —494 3—Low H , low U 17.3 4.8 6.8 6707 —441 —1656 4—High H , high U 2.1 0.3 7.5 444 —946 —1998 5—Low U 7.3 0.5 5.2 937 —12 —387 6—Very high I 8.2 0.6 12.2 4896 —58 —408 **Mean 10.0 1.4 6.1 2482 —345 —912**

Case	NeuPlume				GP blind		GP known U		MB known U
	Q	H	U	I	Q	U	Q	H	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

4.1.2 Baseline comparison

470

A baseline comparison evaluates the contribution of non-Gaussian field representation to estimating source parameters. We apply NeuPlume and two classical point source methods to the same synthetic observations: NeuPlume, GP regression Shah et al. (2019), and the MB method. This protocol uses random volumetric point observations rather than plume crossing transects, which is unfavorable to GP and MB because both baselines are designed around analytical plume shapes or cross-sectional flux integration. The comparison therefore tests how each method handles observation geometry mismatch, not its performance under ideal sampling conditions. Table 1 shows that, when all source parameters are unknown in blind inversion, GP has a mean relative Q error of 2482% applied to the same observations from the six representative cases. Under this random volumetric geometry, mean Q error of 2482-

475

Even when the true wind speed is provided (is 1 134.5% for blind GP, 137.6% for known- U mode), GP still produces a mean relative Q error of 345-

For MB, Q errors range from 387 to 1998% in GP, 329.3% for known- U mode, with a mean of 912%. Under the observation protocol formed by 1 000 random sampling points within the computational domain, random point observations do not provide the cross-section information required by the MB method. In contrast, NeuPlume uses the learned concentration-field structure directly, keeping Q errors one to two orders of magnitude below the baseline methods MB, and 12.5% for NeuPlume; the corresponding medians are 140.0%, 115.1%, 321.3%, and 12.4% (Table 1). The comparison tests how methods designed for analytical plume fitting or cross-sectional flux integration transfer to sparse volumetric observations.

4.1.3 Gaussian plume control experiment

The Gaussian-plume control experiment separates model-assumption mismatch from implementation error. We use Gaussian-plume control uses the same source parameters as the six synthetic cases, generate synthetic data representative cases and generates observations with an analytical Gaussian-plume model, and repeat the baseline comparison (Supplementary Table S2). On data satisfying the Gaussian assumption, Gaussian-plume regression in Gaussian-plume model (Table C1). In known- U mode, GP recovers all six cases, with Q and H errors close to zero ($< 10^{-6}$).

By contrast, GP in blind-inversion mode still overestimates below 0.001%. In blind mode, its mean Q on Gaussian data, with a mean error of 718.

In the fully blind-inversion configuration, NeuPlume also treats wind speed and U as unknown and jointly estimates errors are both 126.3%, consistent with the Q , U ambiguity when both parameters are inferred from the Gaussian-plume equation. The known- U result confirms that the GP implementation recovers parameters when its field assumption matches the data generator.

4.1.4 Observation geometry and transport-model mismatch

The geometry and transport stress tests use the six representative conditions within the trained parameter support. Figure F1 shows the case-wise transport-mismatch changes, Fig. G1 shows the transformed UAV sampling geometries, and Fig. G2 shows their case-wise degradation relative to random-volume sampling.

With 1 000 random-volume observations per case, the mean errors in H , U , and I . Compared with known- U , errors in, 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 remain similar, but 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 error

increases, with a six-case mean of 13.2 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.

4.2 Real-data case study 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 (Fig. 4). 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.

4.3 Real-data case study

The real-data experiment tests whether NeuPlume gives physically reasonable emission rate estimates from field UAV observations and how observation geometry and wind-speed input affect inversion. We analyze four Pniówek V flights introduced in the experimental setup: #11 ($N_{\text{obs}} = 362$), #12 ($N_{\text{obs}} = 214$), #14 ($N_{\text{obs}} = 108$), and #15 ($N_{\text{obs}} = 130$). Each flight's observations are inverted in two settings. Known- U We evaluate the Pniówek V flights in known- U mode fixes wind speed to the co-located meteorological station, whereas blind inversion estimates U . As the main external comparator, we use the and blind modes. The same-shaft hourly emission inventory (13.3 ± 4.1 kt yr $^{-1}$). The integrated Gaussian result reported by Andersen et al. (2023) ($Q_{\text{IG}} = 9.5 \pm 3.5$ kt yr $^{-1}$, $N = 13$ flights) and the mass-balance result ($Q_{\text{MB}} = 9.8 \pm 5.0$ kt yr $^{-1}$, $N = 14$ flights) are supplementary method references from the same campaign. Known- U is used for the main inventory comparison, and blind inversion diagnoses wind-input accuracy and parameter identifiability. 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 Andersen et al. (2023). Table 2 summarizes the results; Figure 5 shows the observation geometry and posterior distributions and Fig. 5 report NeuPlume's selected minimum-loss estimates.

Flights Known- U estimates for flights #11 and #12 show that, with sufficient coverage, NeuPlume can infer emission rate estimates from real UAV data that are compatible with the hourly inventory. In known- U mode, #14 are 2.14 and 2.55 kt yr $^{-1}$ above the inventory center and fall within its reported 9.2–17.4 kt yr $^{-1}$ range. Flights #11 gives $\hat{Q} = 8.33 \pm 4.07$ kt yr $^{-1}$ and #12 gives $\hat{Q} = 7.27 \pm 3.42$ kt yr $^{-1}$. Both posterior intervals overlap the inventory range, although their centers are low. Measurement points cover downwind distances of about 252–297 m (15 are 7.36 and 8.03 kt yr $^{-1}$ below the center and lie 3.26 and 3.93 kt yr $^{-1}$ below the lower end of that range.

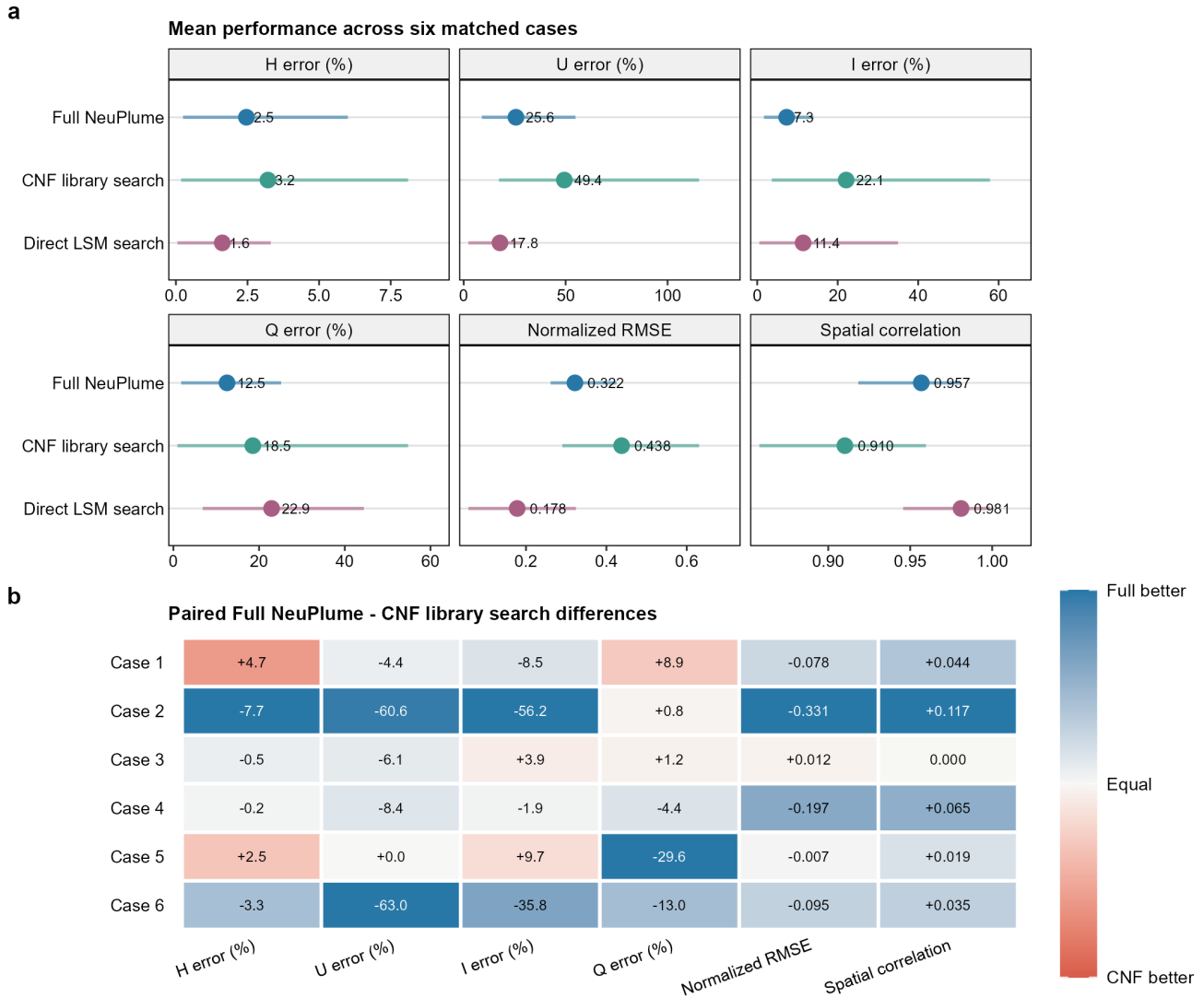


Figure 4. Matched accuracy comparison for the six representative cases. (a) Mean absolute relative errors for effective release height H , wind speed U , turbulence intensity I , and emission rate Q , together with mean full-field normalized RMSE and spatial correlation; horizontal intervals span the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

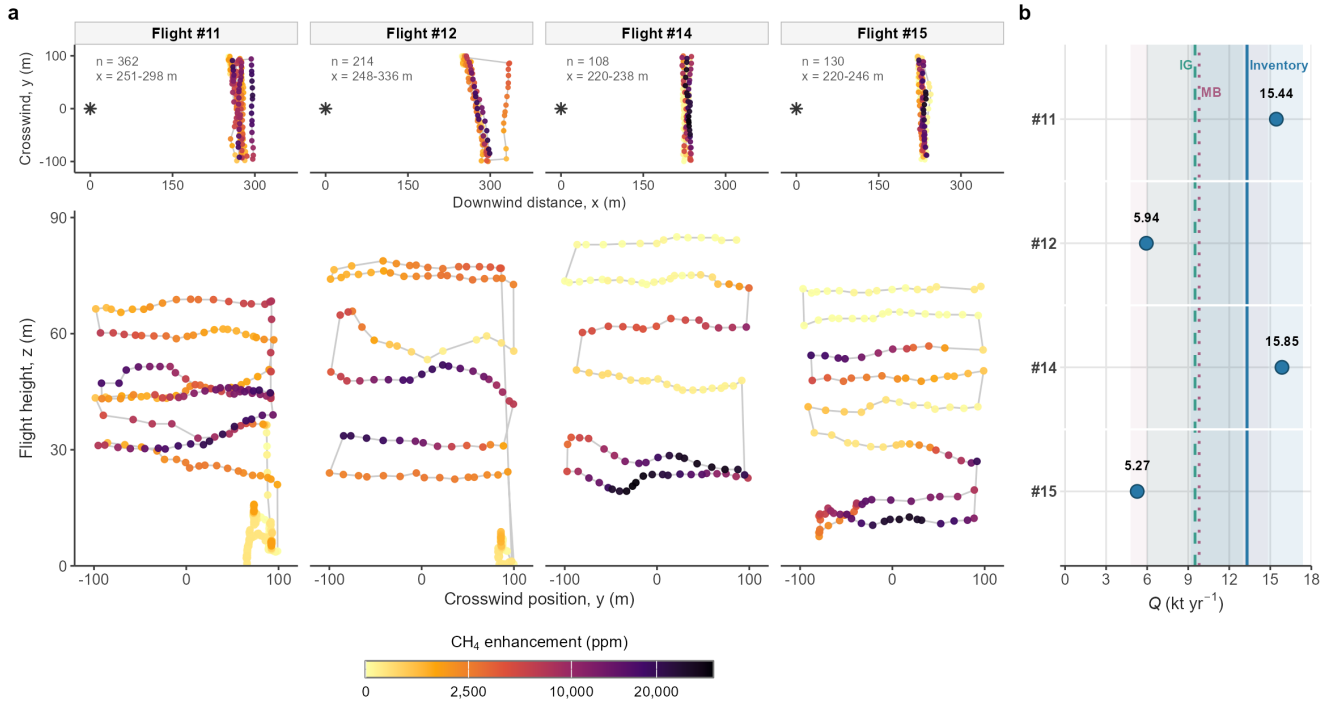


Figure 5. Real-data inversion results for the Pniówek V coal-mine-ventilation shaft. (a) xy -plane and (b) xz -plane views show UAV observations sampling geometry for flights #11 each flight. The upper strips locate the observations in the downwind-crosswind plane; stars mark the shaft. The lower panels show the crosswind-height sampling curtains, #12, #14, with grey lines tracing the flight paths and #15; color denotes measured denoting methane enhancement. Labels report the number of observations and sampled downwind range. (eb) Known-NeuPlume known- U posterior emission-rate minimum-loss Q estimates (Q_{points}) and credible intervals. The main-external comparator is the same-shaft hourly-inventory ($13.3 \pm 4.1 \text{ kt yr}^{-1}$); integrated-Gaussian (and published IG) and mass-balance (MB) estimates are supplementary-references values retain their reported external ranges.

Blind inversion selects lower wind speeds for flights #11) and 248–335 m (, #12), with 362 and 214 observations, respectively. This spatial coverage constrains the likelihood and separates the emission-rate signal from background noise. Blind inversion gives lower wind speeds ($\hat{U} \approx 2.6 \text{ m s}^{-1}$) and lower emission rates, indicating that $Q-U$ coupling still shifts the posterior when wind-speed is weakly-constrained. Thus, known- U results are the main estimates for inventory comparison.

545 Flight, and #14 reflects the identifiability limit caused by limited near-source sampling. Known- U gives $\hat{Q} = 11.74 \pm 3.75 \text{ kt yr}^{-1}$, with a posterior interval overlapping the hourly inventory. However, blind inversion yields a very broad emission-rate posterior ($\hat{Q} = 8.01 \pm 23.81 \text{ kt yr}^{-1}$), and the maximum-a-posteriori source height is only $H = 2.24 \text{ m}$. This flight has the fewest in-domain observations ($N_{\text{obs}} = 108$), concentrated at short downwind distances (221–238 m). Under this geometry, near-source concentration gradients allow multiple parameter combinations to produce similar concentrations at the observation locations.

550 The main limitation is therefore parameter identifiability under limited near-source coverage.

Table 2. Real-data inversion results—Minimum-loss NeuPlume estimates for the Pniówek V ventilation shaft flights. The main external comparator is the same-shaft hourly inventory $13.3 \pm 4.1 \text{ kt yr}^{-1}$. As supplementary references, Andersen et al. (2023) reported $\text{IG} = 9.5 \pm 3.5 \text{ kt yr}^{-1}$ and $\text{MB} = 9.8 \pm 5.0 \text{ kt yr}^{-1}$. Known- U is used for fixes wind speed to the main emission-rate comparison. Blind meteorological-station value; blind inversion jointly estimates selects Q and U to diagnose wind-speed representativeness and parameter degeneracy jointly.

Flight	N_{obs}	$U_{\text{meteo}} \text{ (m s}^{-1}\text{)}$	Known- U : $\hat{Q}$ Known- U : $\bar{Q} \pm \sigma_Q \text{ (kt yr}^{-1}\text{)}$ (kt yr ⁻¹)	Blind: $\hat{Q}$ Blind: $\bar{Q} \pm \sigma_Q \text{ (kt yr}^{-1}\text{)}$ (kt yr ⁻¹)	Blind: $\hat{U}$ Blind: $\bar{U} \pm \sigma_U \text{ (m s}^{-1}\text{)}$ (m s ⁻¹)
#11	362	7.0	8.33 ± 4.07 15.44	3.59 ± 0.82 5.17	2.57 ± 0.69 3.0
#12	214	7.0	7.27 ± 3.42 5.94	3.68 ± 1.02 3.49	2.55 ± 0.70 2.5
#14	108	6.0	11.74 ± 3.75 15.85	8.01 ± 23.81 2.84	1.81 ± 0.70 1.3
#15	130	6.0	4.10 ± 1.16 5.27	13.97 ± 5.87 13.26	13.00 ± 0.71 14.0

and correspondingly lower Q values. Flight #15 shows that wind-input accuracy strongly affects inversion. Known-selects $U = 14.0 \text{ m s}^{-1}$, more than twice the station value, and $Q = 13.26 \text{ kt yr}^{-1}$. Across all four flights, the selected U gives $\hat{Q} = 4.10 \pm 1.16 \text{ kt yr}^{-1}$, below the hourly inventory range. Blind inversion returns $\hat{Q} = 13.97 \pm 5.87 \text{ kt yr}^{-1}$ and estimates $\hat{U} = 13.0 \text{ m s}^{-1}$, above the meteorological-station value $U = 6.0 \text{ m s}^{-1}$. This agrees with the and Q - U scaling in steady transport: if station wind speed does not represent the effective transport speed at the UAV sampling height, fixed-wind inversion compensates through a lower Q .

The baseline comparison further shows that the physical latent space prior provides effective regularization when field observations are sparse. In blind inversion mode, the Gaussian plume baseline gives $\hat{Q}$ values ranging from 25 to 1792 kt yr^{-1} , indicating that an unconstrained analytical plume surface can produce physically implausible concentration fields under incomplete data. Even with fixed wind speed, Gaussian plume estimates reach $\hat{Q} \in [31, 49] \text{ kt yr}^{-1}$. MB deviates in the opposite direction, with near-zero estimates ($\hat{Q} \approx 0.009\text{--}0.055 \text{ kt yr}^{-1}$), because sparse and irregular UAV sampling does not satisfy its dense cross-section requirement. In contrast, NeuPlume move in the same direction relative to the known- U posterior means are $\hat{Q} \in [4.1, 11.7] \text{ kt yr}^{-1}$ across the four flights, and three posterior intervals overlap the hourly inventory (full comparison in Supplementary Table S3)solution.

4.4 Uncertainty quantification

Known- U inversion selects effective heights of 52.2, 43.8, 3.3, and 9.2 m for flights #11, #12, #14, and #15, respectively. The corresponding blind-mode heights are 61.4, 47.9, 2.2, and 10.6 m.

After evaluating emission rate point estimates, uncertainty quantification tests whether NeuPlume’s approximate posterior intervals have reasonable empirical coverage. In point-source inversion with limited measurements, credible intervals bound the interpretation of point estimates under underdetermined observations. We evaluate calibration on 30 holdout samples independently drawn from the training distribution. Each sample uses the default temperature $\tau = 1.0$ and runs the full diffusion posterior sampling and importance sampling workflow (Section 2.4). For each sample, the approximate posterior distributions

of (H, U, I) are summarized by symmetric credible intervals at 50%, 68%, 90%, and 95% nominal levels, with coverage defined as the fraction of samples whose true parameter lies inside the interval.

575 The overall coverage indicates that NeuPlume’s posterior intervals are empirically calibrated on the current holdout set, with conservative behavior at high nominal levels. Figure ??a shows that, after pooling parameters, empirical coverage is 73.3% for the 68% interval, 97.8% for the 90% interval, and 100% for the 95% interval. These values are slightly above the nominal levels but remain close to the ideal calibration line. Parameter-level coverage identifies the source of this pattern: at the 68% nominal level, wind speed
 580 The GP and MB baselines produce substantially larger or smaller point estimates under these sparse trajectories (Table D1). GP blind estimates span 25–1792 kt yr⁻¹ and GP known- U (70.0%) and turbulence intensity I (66.7%) are close to nominal coverage, whereas release height H (83.3%) is overcovered (Figure ??b).

Coverage for wind speed estimates span 31–49 kt yr⁻¹, while MB estimates span 0.009–0.055 kt yr⁻¹. NeuPlume known- U deviates further from nominal levels at narrower intervals, suggesting that the H – U – Q degeneracy mainly affects the marginal posterior of wind speed. At the 50% nominal level, empirical coverage for U decreases to 26.7%. When the nominal level
 585 increases to 90% and 95%, U coverage returns to 100%. This undercoverage at narrow intervals is consistent with transport physics: wind speed changes can be partly compensated by effective release height and source strength adjustments, making the marginal posterior of U more sensitive to interval width. Therefore, applications should not interpret the U point estimate alone; they should report posterior intervals and, where possible, cross-check them against independent wind speed observationsestimates
span 5.27–15.85 kt yr⁻¹.

590 Uncertainty quantification calibration on 30 independent holdout cases ($\tau = 1.0$). (a) Overall empirical coverage after pooling parameters versus the nominal credible interval level. The dashed diagonal denotes ideal calibration. (b) Per-parameter coverage at the 68% nominal level: U and I are close to nominal coverage (70.0% and 66.7%), while H is overcovered (83.3%).

5 Discussion

The temperature parameter τ in the diffusion posterior sampling likelihood (Eq. ??) controls the sharpness of the approximate
 595 posterior: smaller τ concentrates samples in high-likelihood regions, whereas larger τ flattens the distribution. A fixed default temperature is useful only if coverage and effective sample size are insensitive to reasonable perturbations. We scan $\tau \in \{0.01, 0.05, 0.1, 0.5,$
 over 18 parameter-case pairs (six synthetic cases $\times$ three parameters) and compute coverage at each nominal level (Figure ??a).

The temperature-sensitivity results support using $\tau = 1.0$ as the default. For $\tau \in [0.1, 10]$, both 90% and 95% coverage
 600 remain stable at 94.4%, showing that high-confidence intervals are insensitive to temperature changes across two orders of magnitude. Only at $\tau = 0.01$ do 90% and 95% coverage fall to 66.7% and 72.2%, respectively, indicating that excessive sharpening overconcentrates the posterior in high-likelihood regions. The 50% and 68% intervals are more sensitive to posterior shape and therefore vary more with τ . However, for $\tau \geq 0.05$, no collapse similar to $\tau = 0.01$ occurs. At $\tau = 1.0$, the mean effective sample size across the six cases is 70.7 out of 75 diffusion samples, indicating that the importance weights are not

605 dominated by a few samples. With only 18 parameter-case pairs, these results support $\tau = 1.0$ as a stable default within the tested range but do not guarantee universal calibration.

Approximate posterior distribution analysis. (a) Coverage sensitivity to Boltzmann temperature τ over 18 parameter-case pairs. For $\tau \in [0.1, 10]$, 90% and 95% coverage are stable at 94.4%. (b, c) Joint (H, U) posterior distributions for Case 1 (compact, low epistemic uncertainty) and Case 4 (diffuse, high epistemic uncertainty with negative $H-U$ correlation), showing
610 different observation constraints.

Beyond coverage, the geometry of the posterior diagnoses parameter identifiability. Figure ??b and c show the joint (H, U) posteriors for two synthetic cases. Importance weights are obtained by applying the Boltzmann transform to the diffusion posterior sampling loss (Eq. ??), and the complete marginal and joint distributions for Case 1 are shown in Supplementary Fig. S1. In Case 1, a typical mid-range scenario, the joint posterior forms a compact unimodal cluster near the truth with weak
615 $H-U$ correlation (Figure ??b).

By contrast, the high-release-height 30-case validation shows that NeuPlume preserves plume location and large-scale spatial organization more reliably than cross-sectional spread. Strong positional agreement together with negative mean width bias identifies cross-sectional spread as the systematic field-reconstruction error. In the six representative cases, high-wind-speed conditions of Case 4 ($H = 47.9$ m, $U = 14.1$ m s⁻¹) produce an elongated joint posterior with negative $H-U$ correlation
620 (Figure ??c), reflecting physical degeneracy: a higher release height can partly offset lower wind speed in the downwind concentration pattern. Traditional point estimates cannot express this uncertainty structure. Because observations in Case 4 cover only 1% of the downwind cross-section, the data cannot independently constrain H and U , and the emission estimate must be interpreted together with the posterior shape. The contrast between Cases 1 and 4 illustrates how posterior geometry diagnoses the strength of observational constraints. NeuPlume also yields lower mean emission-rate error than GP and MB
625 under random volumetric sampling. This comparison establishes NeuPlume's advantage for emission-rate estimation under the tested sparse three-dimensional observation geometry.

5.1 Ablation study Measurement geometry, wind representativeness, and identifiability

5.1.1 Coarse-to-fine parameter search

The coarse-to-fine search described in Section 2.4 balances computational cost and parameter accuracy by first locating
630 low-loss regions on a coarse grid and then performing a fine grid search around the winning cell. To quantify the contribution of the refinement stage, we compare the full two-stage search with a coarse-only baseline across the six synthetic cases. Under coarse-only search, the mean emission rate error is $\bar{\epsilon}_Q = 25.0\%$, and The field inversions expose strong coupling between wind-speed representativeness and the mean release-height error is $\bar{\epsilon}_H = 8.6\%$. After adding fine search, these errors decrease to $\bar{\epsilon}_Q = 10.0\%$ and $\bar{\epsilon}_H = 1.4\%$, corresponding to mean improvements of 15.0 and 7.2 percentage points.

635 Most of the refinement gain comes from improved release-height estimation. In the high- H region, the spacing of the coarse grid can reach $\Delta H = 10$ m, which is coarse 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 sensitivity of the plume to release height. Fine search restores subgrid accuracy by locally subdividing the winning cell. Case-wise results also show that refinement is not monotonically beneficial for every metric: the Q error in Case 1 increases slightly from 13.7% to 14.2%, and that in Case 4 increases from 1.7% to 2.1% (Supplementary Table S1). However, the H error in Case 4 simultaneously decreases from 4.3% to 0.3%, and the six-case mean still shows a clear benefit. Overall, the coarse-to-fine decomposition preserves the global search capability of the coarse scan while avoiding high-resolution enumeration over the entire parameter domain. known- U solutions. This behavior is consistent with wind sensitivities reported in cross-sectional-flux and controlled-release studies Meier et al. (2026); Eastwood et al. (2025); Conrad et al. (2023)

5.1.1 Observation-coverage sensitivity

The observation coverage r_{obs} is defined as the proportion of computational grid cells that contain at least one measurement point, and it determines how much information the likelihood term can use to constrain source parameters. We scan r_{obs} from 0.5% to 10% across the six synthetic cases while keeping all other settings unchanged (Figure ??). When coverage decreases from 1% to 0.5%, the mean emission rate error increases from 26.4% to 42.1%, a degradation of 15.7 percentage points. This jump indicates that, under extremely low coverage, the observational likelihood surface is insufficient to reliably distinguish candidate parameter combinations, making grid search more likely to select a suboptimal solution.

Once coverage exceeds 1%, the mean emission rate error enters a plateau and mainly remains within 25%–29%. Under the current three-dimensional random sampling protocol, the marginal benefit of adding observation points decreases rapidly. Inversion quality is limited by observation geometry and parameter degeneracy, so the number of measurement points is not the only limiting factor. Approximately 1% coverage can serve as a practical minimum target under the current setting, below which inversion quality degrades sharply.

Sensitivity of emission rate error to spatial observation coverage r_{obs} . When coverage is below 1%, the mean Q error increases markedly (42.1% at 0.5% and 26.4% at 1%). Above 1% coverage, the curve enters a plateau, indicating limited marginal benefit from further sampling densification.

5.1.1 DPS observation-likelihood guidance

The observation-likelihood guidance ablation evaluates the marginal contribution of the gradient term in the reverse diffusion trajectory. Both configurations use the same DPS reverse diffusion process to generate conditional concentration-field samples; they differ only in whether the observation-likelihood gradient is injected into the reverse trajectory. Under the current coarse-grid setting, the guided configuration has a mean Q error of 29.7%, whereas the unguided configuration has a mean error of 25.4%. The difference is approximately 4.4 percentage points, with the unguided configuration giving the lower mean error.

In the current implementation, the main constraint comes from explicit enumeration and loss comparison on the candidate parameter grid. The observation-likelihood gradient does not provide additional error reduction under this discrete search configuration. Subsequent improvements could prioritize continuous parameter optimization, allowing candidate parameters to update along the observation-likelihood gradient rather than only selecting among finite grid nodes. 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 (Figs. G1–G2). 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. UAV studies likewise identify plume coverage, platform configuration, and wind sampling as major determinants of emission accuracy Liu et al. (2024); Mohammadloo et al. (2025); Vollrath et al. (2026); Scheutz et

6 Discussion

5.1 Physical scope and boundary effects

5.2 Wind-speed representativeness and observation-geometry error

The real-data experiments indicate that the main errors in NeuPlume arise from the joint constraint that wind-speed representativeness and observation geometry impose on the source parameters. As shown in Section 4.3, Flight 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 #14 overlaps the hourly inventory in known- U mode but yields an unstable blind-inversion posterior because of its short downwind distance and incomplete plume coverage. Flight 11 and #15 illustrates the complementary problem: blind mode adjusts the posterior wind speed to $\hat{U} \approx 13 \text{ m s}^{-1}$, more than twice the ground-station value, shifting the emission rate posterior along the Q – U direction. When the station wind speed does not represent the effective transport speed at the UAV sampling height, the posterior moves along this coupling direction and the resulting emission estimate can deviate substantially from the inventory range 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 effect of wind-speed representativeness error can be explained by the analytical scaling of steady-state passive-tracer transport. Because the conditional neural field (CNF) is trained at a unit emission rate ($Q = 1$), the inversion first estimates $k = Q/U_{\text{eff}}$ and then recovers the emission rate through $Q = k \cdot U_{\text{eff}}$. The emission rate is therefore linearly coupled to effective wind speed. This coupling follows from the linear response of the steady concentration field to source strength and its inverse scaling with advective speed; it does not depend on a particular Gaussian plume approximation. Previous studies show the same sensitivity. Meier et al. (2026) showed that wind speed contributed up to 99.4

Transport-model error further changes the interpretive range of this coupling. Model error can not only widen the posterior distribution but also move the posterior location nonuniformly in parameter space De Meutter et al. (2021). When the wind-profile assumptions of the Lagrangian stochastic model (LSM) are violated, the blind-mode posterior can still have systematic displacement.

Other deep-learning methods face the same wind-speed sensitivity. Zhao et al. (2026) combined a U-Net encoder with inverse Gaussian fitting and reduced single-source mean absolute percentage error (MAPE) from 38% to 28.5% on controlled-release data. That framework differs from NeuPlume in its target: it estimates only Q and does not jointly invert Q , 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 , and U . It returns point estimates absorbs part of the missing vertical transport and functions as an effective height rather than a full posterior, and it relies on dense cross-sectional concentration data rather than limited point measurements. Accordingly, its large-eddy simulation (LES)-validated setup covers a wider range of atmospheric stability (Pasquill A–D) and has been validated in 11 controlled-release experiments, providing a reference for validation scenarios not yet covered in this study. physical outlet elevation.

Observation geometry is the second major error source and explains why inventory-compatible Flight #14 still gives an unstable blind-inversion posterior. This flight samples only the near-source region, with downwind distances of 221–238 m, and the limited downwind range makes the inversion depend strongly on steep concentration gradients. Under these conditions, small spatial offsets can translate into large intensity differences. The DPS likelihood surface therefore becomes flat, and multiple (Q, H, U) combinations can produce similar concentrations at the observation locations. This degeneracy is consistent with the sensitivity to observation geometry, platform configuration, and wind-field sampling documented in UAV and controlled-release studies Liu et al. (2024); Vollrath et al. (2026); Scheutz et al. (2025). Among them, Mohammadloo et al. (2025) found in 1,001 UAV flights that 55 Open-boundary diagnostics in Table E1 show that most particle exits occur through the downwind boundary and quantify the smaller side and top losses. Measurements and forward domains for weak-wind or high-turbulence regimes require sufficient crosswind and vertical extent to capture the broader plume. The current LSM also excludes flow regimes in which material exits and later re-enters through recirculation or building wakes.

5.2 Steady-state and atmospheric-stability assumptions

The neutral atmospheric-stability assumption defines the current scope of the framework. All LSM training simulations use a neutral wind profile ($\alpha = 1/7$), so the current results mainly apply to near-field point-source dispersion under moderate-to-strong wind conditions. Stability changes near-surface wind shear and vertical mixing: stable conditions generally correspond to a larger wind-shear exponent and weaker vertical mixing, whereas unstable conditions generally correspond to a smaller wind-shear exponent and stronger vertical exchange Newman and Klein (2014). In a Lagrangian stochastic model, non-neutral boundary layers also change the The forward-library design defines a direct extension path for additional regimes. Stable and unstable boundary layers require stability-conditioned simulations because they alter wind shear, vertical exchange, turbulent-velocity distribution, drift term, and implementation of the distributions, and the LSM drift needed to satisfy the well-mixed condition Cassiani et al. (2014). Because the CNF has not been trained under these states, its reconstruction fidelity under non-neutral conditions has not been verified. Incorporating stability as a conditional parameter in the training library is a necessary step for extending the framework to a wider range of atmospheric states.

735 The steady-state emission assumption also limits the application boundary of NeuPlume. The current framework applies to continuous point sources such as ventilation shafts, pipeline leaks, and industrial stacks, but it does not directly apply to intermittent or time-varying releases. Bayesian time-varying [Newman and Klein \(2014\)](#); [Cassiani et al. \(2014\)](#). Time-varying releases require a temporally resolved library and field representation, for which sequential Bayesian source-term estimation methods, such as the Bayesian source-term estimation of [Xu et al. \(2025\)](#), can handle temporal variability by assimilating sequential observations. Such methods require dense temporal coverage and explicit turbulence parameterization. A later extension of NeuPlume could use a time-conditioned diffusion model so that the latent space prior represents time-varying concentration fields, but this direction still requires separate implementation and validation. [methods provide an established reference Xu et al. \(2025\)](#).

5.2 Computational cost [Method positioning and scalability validation](#)

745 NeuPlume's scalability comes from separating the physical forward model from the probabilistic inversion engine. The current implementation uses the LSM to build a parameter-to-concentration-field mapping for a single point source, neutral stability, and flat terrain. The key requirement is not a particular analytical formula or simulator, but a forward model that can generate parameterized concentration-field samples for the target scenario. Those samples can then define the field prior and, when limited measurements are available, support posterior inference of the control parameters. Multi-source emissions, complex terrain, building effects, or higher-fidelity LES/CFD models can all be added as physical front ends along this route. The corresponding posterior object can also be extended from single-source (H, U, I, Q) to source locations and emission strengths for multiple sources, terrain or boundary layer parameters, and the full concentration field.

This extension direction has a clear computational cost. Whenever the scene geometry changes, for example from flat to complex terrain or from a single source to a multi-source configuration, the training data must be regenerated and the CNF must be retrained, because the learned concentration manifold is specific to the training parameter space. In the current implementation, the LSM ensemble and CNF training require about 50 CPU-h and 116 GPU-h, respectively. Diffusion-model training requires about 4 GPU-h, and one coarse-to-fine inversion requires about 3 GPU-h; the computational-cost breakdown is given in Appendix Table ???. These orders of magnitude make NeuPlume most appropriate for deployments in which an offline simulation library can be reused, rather than for cases that require rebuilding the full training pipeline for each new geometry.

Increasing parameter dimension further limits the current coarse-to-fine grid search strategy. If multi-source locations or time-varying emission profiles (>6 degrees of freedom) are jointly estimated, discrete grid enumeration will face combinatorial growth. The current ablation results show that, under the discrete coarse grid, the DPS observation-likelihood gradient provides no measurable benefit and even slightly increases the mean error (Section ??). This result indicates that the effective information in the present configuration mainly comes from grid enumeration. Gradient-driven continuous parameter optimization still needs separate implementation and validation. In atmospheric inversion, 4D-Var and adjoint models have been used for emission rate correction, chemical state estimation, and high-resolution CO_2 flux inversion [Elbern et al. \(2007\)](#); [Maksyutov et al. \(2021\)](#). A more scalable approach would update source terms and meteorological parameters on a continuous parameter manifold with

the observation-likelihood gradient, rather than continuing to enlarge the discrete grid. A complementary route is amortized inference: Liu et al. (2025b) demonstrated a two-stage conditional diffusion framework for partial-differential-equation (PDE) constrained inverse problems, with inference 60 times faster than DPS while maintaining an empirical coverage rate of 94.35%

5.3 Method positioning and future validation

NeuPlume is positioned as a complement to existing source-term estimation methods, not as a replacement for all traditional methods. In far-field settings with a homogeneous wind field where the Gaussian plume assumption holds, integrated Gaussian and MB approaches remain simpler and faster and have been widely used Shah et al. (2019); Krings et al. (2013); Conley et al. (2017). NeuPlume’s comparative advantage appears in three scenarios: targets near-field sampling dominated by settings where a simulation-trained representation can describe non-Gaussian structure, cases that require joint estimation of multiple source and environmental parameters, and settings that need posterior uncertainty quantification for regulatory or decision support. In these settings, the posterior distribution can state which explanations are supported by the observations and which cannot be excluded, which is closer to practical decision needs than a single optimum. This methodological complementarity is consistent with Aghdasi et al. (2025), who found that variational assimilation can strongly constrain point-source emissions under in situ observations but has diminishing benefits for diffuse area sources. A recent structure and several source and environmental parameters must be searched jointly. GP and MB remain efficient choices when their plume and sampling assumptions hold Shah et al. (2019); Krings et al. (2013); Conley et al. (2017). 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 evaluation with fence-line sensors likewise shows that Gaussian plume inversion and backward Lagrangian stochastic methods are sensitive to wind-sector width, averaging window, and multi-source interference Mbua et al. (2025). Controlled-release assessments further show that source type, observation geometry, and platform configuration change the error structure of quantification methods Liu et al. (2024). Different source types and observation geometries call for different inversion strategies. Because the current real dataset has no certified true experiment with a known reference emission rate for individual flights, cross-method comparison cannot replace absolute-accuracy validation. Validating NeuPlume against controlled-release experiments, the established benchmark for method comparison Morales et al. (2022); Vollrath et al. (2026), is the next step. Combining UAV, ground-based, and satellite observations will further establish the framework’s operational each flight would provide the decisive test of NeuPlume’s absolute emission-rate accuracy under field conditions Morales et al. (2022); Vollrath et al. (2026). Such validation should span wind regimes and observation geometries while retaining the minimum-loss selection rule used here.

6 Conclusions

This paper presents NeuPlume, a probabilistic inversion framework for NeuPlume estimates point-source emissions under limited observations. It uses Lagrangian stochastic model (LSM) forward simulations to construct parameterized concentration fields for the target scenario, learns a field prior with a conditional neural field (CNF) and a latent space diffusion model, and

uses diffusion-posterior sampling (DPS) to infer source parameters, environmental parameters, and the complete concentration field under observational constraints. Synthetic experiments, real UAV observations, and uncertainty assessment support three conclusions.

First, across six synthetic cases under neutral atmospheric stability that cover different source heights, wind speeds, and turbulence intensities, NeuPlume achieves mean errors of 10.0%, 1.4%, and 6.1% for emission rate Q , effective release height H , and turbulence intensity I , respectively. Under the same volumetric point observation setting, the mean Q errors of blind GP inversion and MB are 2482% and 912%. This gap indicates that the non-Gaussian concentration representation learned by the CNF provides stronger physical constraints for near-field point-source inversion than simplified dispersion assumptions.

Second, in the real UAV data from the Pniówek coal-mine ventilation shaft, NeuPlume gives posterior mean emission rates of 4.1–11.7 kt yr⁻¹ in known- U mode. Three of the four flights are compatible with the hourly inventory. The remaining low estimate is attributable to unrepresentative wind speed. These results support proof-of-concept transfer from synthetic training data to real UAV observations, but do not validate absolute accuracy.

Third, the posterior intervals achieve empirical coverage close to nominal levels on the holdout set. Among parameters and complete three-dimensional concentration fields from sparse observations by searching a generative representation of LSM simulations. Across 30 holdout samples, the empirical coverage of the 68% credible interval is 73.3%, and the coverages of the 90% and 95% intervals are 97.8% and 100%, respectively. These results show that, beyond returning point estimates, NeuPlume provides posterior uncertainty estimates that reflect parameter degeneracy and observational constraint strength.

The ablation experiments further define the key design conditions of the current implementation. The coarse-to-fine parameter search reduces the mean Q error from 25.0% to 10.0%. When observation coverage falls below approximately 1%, inversion quality clearly decreases. These results indicate that current performance depends on local parameter refinement and sufficient spatial observational constraints, held-out synthetic cases, fields retained plume-scale spatial organization (mean spatial correlation, 0.942), and emission-rate error averaged 13.6%; reconstructed plumes remained systematically narrow. In six matched cases, diffusion improved parameter and field metrics over direct CNF library search, and NeuPlume reduced emission-rate error relative to GP and MB under random-volume sampling.

These results also define the scope of applicability and directions for extending NeuPlume. Because the physical front end and the probabilistic inverter are modular, multiple point sources, complex terrain, or higher-fidelity simulators such as LES/CFD can enter the same posterior inference workflow by redefining the forward mapping from parameters to Stress tests identified observation geometry and forward-model fidelity as distinct controls. UAV-like trajectories weakened field reconstruction at the same observation budget. Under transport mismatch, shear shifted parameters while preserving mean field similarity, whereas effective plume rise degraded parameter and field recovery. The Pniówek inversions exposed a positive wind-speed-emission-rate coupling: higher selected wind speeds required larger emission-rate estimates to reproduce the observed concentration enhancements, whereas lower selected wind speeds produced lower estimates. NeuPlume thus provides an extensible framework for jointly estimating source parameters and concentration fields. The empirical results remain limited to the LSM training library used here; therefore, substantially different source configurations, terrain, or atmospheric stability conditions require reconstructing the training data and retraining. The current discrete grid search is

suitable for the low-dimensional parameter space considered here, but not for direct extension to higher-dimensional settings such as multiple sources or time-varying emissions. framework was validated here for passive, low-height, neutral releases over flat terrain; additional transport regimes can be incorporated through regime-specific simulation ensembles and neural-model retraining.

840 Future work should prioritize continuous parameter optimization or amortized inference strategies to replace high-dimensional grid enumeration. The framework should also be validated in controlled release experiments with known emission rates to establish its operational accuracy and calibration under field conditions.

Appendix A: ~~Comparison between Hash-INR Training and SIREN architectures~~inversion hyperparameters

845 During development of the conditional neural field component, we evaluated two candidate architectures: SIREN, which uses periodic sinusoidal activation functions, and the Hash-INR described in Section 2.3. Both architectures were trained on the same data with the same optimizer settings and loss function. For SIREN, we further compared two spatial sampling strategies: uniform random sampling and adaptive sampling that oversamples high-concentration regions, to determine whether training instability was driven by the sampling scheme. Table A1 lists the neural representation, diffusion training, candidate-generation, coarse-to-fine search, and emission-rate refinement settings used in the reported experiments. The real-data experiments use the same DPS settings, search I_{ref} on a discrete grid, and set $n_{\text{fine}} = 3$.

850 Figure ?? shows the training loss curves for all three configurations.

Both SIREN variants showed initial loss reductions during the first 7 training epochs, reaching minimum values of 0.30 (uniform sampling) and 0.34 (adaptive sampling), respectively. After epoch 7, the losses of both configurations began to increase and continued to deteriorate during subsequent training. The loss of the uniform-sampling SIREN rose from 0.30 at epoch 7 to 1.94 at epoch 67. The adaptive-sampling variant showed the same qualitative failure pattern, reaching a loss of 0.62 at epoch 16 before training was terminated. Neither sampling strategy eliminated the instability, indicating that, under this training setup, the convergence problem of SIREN was not mainly caused by the sampling distribution.

Table A1. Training ~~loss curves for the three CNF architectures. Both SIREN variants (uniform sampling and adaptive sampling) became unstable after training epoch 7, whereas Hash-INR converged stably to a final loss~~ inversion hyperparameters of 0.07, below the best loss achieved by the SIREN variants NeuPlume framework.

Component	Parameter	Value
<i>Hash-INR</i>		
	Hash levels	16
	Features per level	2
	Spatial resolutions	$16^3\text{--}512^3$
	Hidden layers	4
	Hidden units	256
	Latent dimension	192
	Training epochs	300
	Network learning rate	10^{-3}
	Latent-code learning rate	5×10^{-4}
<i>Conditional diffusion model</i>		
	Transformer blocks	6
	Attention heads	8
	Hidden dimension	512
	Diffusion steps	1000
	Training epochs	100
	Batch size	64
	Learning rate	10^{-4}
	Noise schedule	Linear, $\beta \in [10^{-4}, 2 \times 10^{-2}]$
<i>Coarse grid search</i>		
	H grid	2, 5, 10, 15, 20, 30, 40, 50 m
	U_{ref} grid	0.8–15.0 ms ^{−1} (11 levels)
	I_{ref} grid	0.05–0.30 (coarse-node values)

Appendix B: Detailed synthetic validation

860 By contrast, Hash-INR trained under all 1000 conditions converged stably over 100 training epochs, with the loss decreasing from 427.09 at epoch 1 to 0.07 at epoch 100. Its final loss was lower than the best value reached by either SIREN variant and also markedly lower than the final loss of the uniform-sampling SIREN. Based on this convergence stability and reconstruction accuracy, NeuPlume uses Hash-INR as the conditional neural field component.

Appendix C: Inversion hyperparameters

865 Table ?? lists hyperparameters for the three-stage DPS inversion workflow in Section 2.4. The Table B1 gives the complete 30-case parameter and field summary. Table B2 lists the physical conditions of the six preselected representative cases used for method comparisons and controlled stress tests. Figure B1 reports the coarse-to-fine grid search defines discrete parameter spaces for Stages 1 and 2; DPS sampling and emission rate optimization control the Stage 3 reverse diffusion process and learnable scaling factor. Boltzmann temperature controls pseudo-posterior sharpness from the grid-search loss surface. The synthetic experiments (Table 1) use the fixed settings below. The real-data experiments (Table 2) use the same DPS and 870 optimization settings, but search I_{ref} on a discrete grid and set $n_{\text{fine}} = 3$. observation-mismatch results.

Table B1. Hyperparameters of Performance across the three-stage DPS inversion workflow30-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.

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 B2. Physical conditions for the six representative cases.

Parameter <u>Case</u>	Value <u>Parameter-space description</u>	
Samples-per-point (n_{coarse}) height <u>Case 1</u>	Coarse-grid-DPS runs Stage 2: fine-grid search ΔH-offsets <u>Low H, high U</u>	
ΔI_{ref}-offsets <u>Case 2</u>	$\{-0.10, -0.05, 0, +0.05, +0.10\}$ <u>Mid H/U, low I</u>	Fine-grid
Noise-schedule <u>Case 3</u>	Linear, $\beta \in [10^{-4}, 2 \times 10^{-2}]$ <u>Mid H/U, high I</u>	
Parameter-learning-rate <u>Case 4</u>	0.01 <u>High H, high U</u>	
Stage-3: emission-rate-optimization k-learning-rate <u>Case 5</u>	5×10^{-4} <u>Low H, low U</u>	
k-mixing-weight <u>Case 6</u>	0.7 <u>Central $H/U/I$</u>	$Q = 0.7k$

Appendix C: Computational-cost

Table ?? reports the order of computational cost for the current implementation. The offline training cost mainly comes from LSM data generation and CNF training. Diffusion model training and single-case inversion are less expensive, indicating that the computational burden is concentrated mainly in the construction of the reusable forward model library.

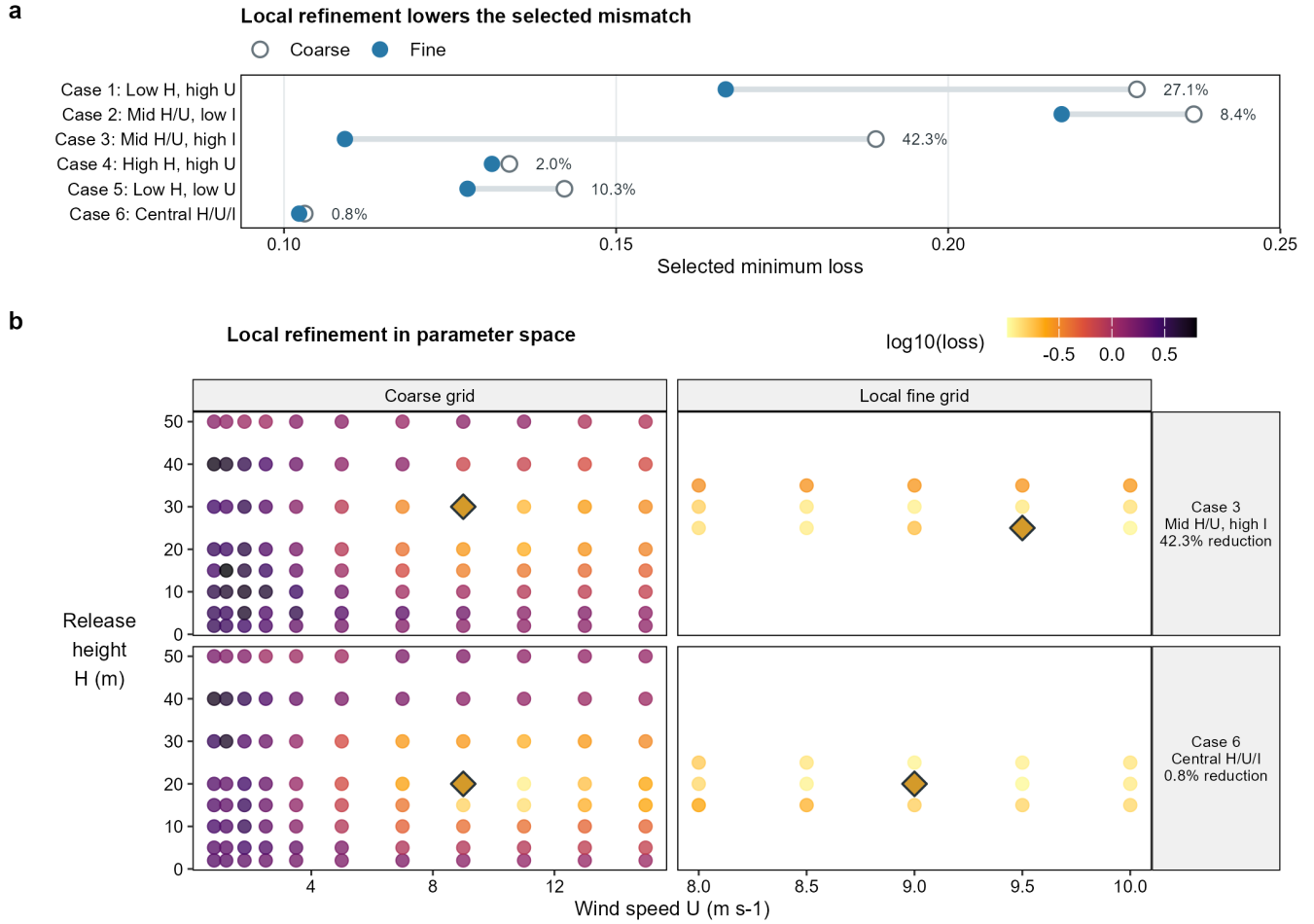


Figure B1. Coarse-to-fine optimization of the observation-mismatch objective. (a) Selected minimum losses for all six representative cases, ordered by relative reduction. The fine search lowers the selected loss in every case, with reductions from 0.8% to 42.3%. Descriptive labels summarize each case's location in the H – U – I parameter domain. (b) Coarse and local fine H – U search spaces for Case 3 (mid H/U , high I) and Case 6 (central $H/U/I$), which have the largest and smallest reductions, respectively. Coarse points show node losses; fine-grid color shows the minimum loss across candidate I values at each H – U location. Yellow diamonds mark the selected node at each stage. Each case has 88 coarse and 75 fine nodes.

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.

Table C1. 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.

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

Appendix D: [Field-data baseline estimates](#)

[Table D1 reports the flight-level GP and MB estimates used to bound the field-data comparison in Sect. 4.3.](#)

Table D1. ~~Computational-cost-breakdown (hardware: NVIDIA RTX 4080 SUPER, 16 GB VRAM)~~[GP and MB estimates for the four Pniówek V flights. The same-campaign published IG estimate is \$9.5 \pm 3.5\$ kt yr⁻¹.](#)

Flight	GP blind	GP known U	MB known U
	$\hat{Q}$ (kt yr ⁻¹)	$\hat{Q}$ (kt yr ⁻¹)	$\hat{Q}$ (kt yr ⁻¹)
#11	831	46	0.009
#12	31	41	0.013
#14	1792	49	0.055
#15	25	32	0.040

The LSM open-boundary accounting provides the physical-domain diagnostic discussed in Sect. 5.1.

Table E1. Boundary-loss diagnostics for the six representative cases using the LSM configuration specified in Sect. 2.2 and a unit emission rate.

<u>Stage</u>	<u>Metric</u>	<u>Task</u>	<u>Mean</u>	<u>Time</u>	<u>Minimum</u>
<i>Offline training (one-time)</i>	Particle exit fraction		0.920		0.913
1	Retained particle fraction	LSM data-generation	0.080		~50 CPU-h
2	Side/top share of exits	Hash implicit neural representation (CNF) training	0.032		~116 GPU-h
3	Downwind share of exits	Diffusion model training	0.968		~4 GPU-h
<i>Online deployment (per case)</i>	4 Side/top share of emitted particles	Inversion (coarse-to-fine search)	0.029		~3 GPU-h

Appendix F: Transport-model mismatch

Figure F1 reports the case-wise changes produced by strong shear and effective plume rise relative to the matched simple-transport baseline.



Figure 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.

Figure G1 defines the three transformed UAV sampling protocols, and Fig. G2 reports their case-wise changes relative to matched random-volume sampling.

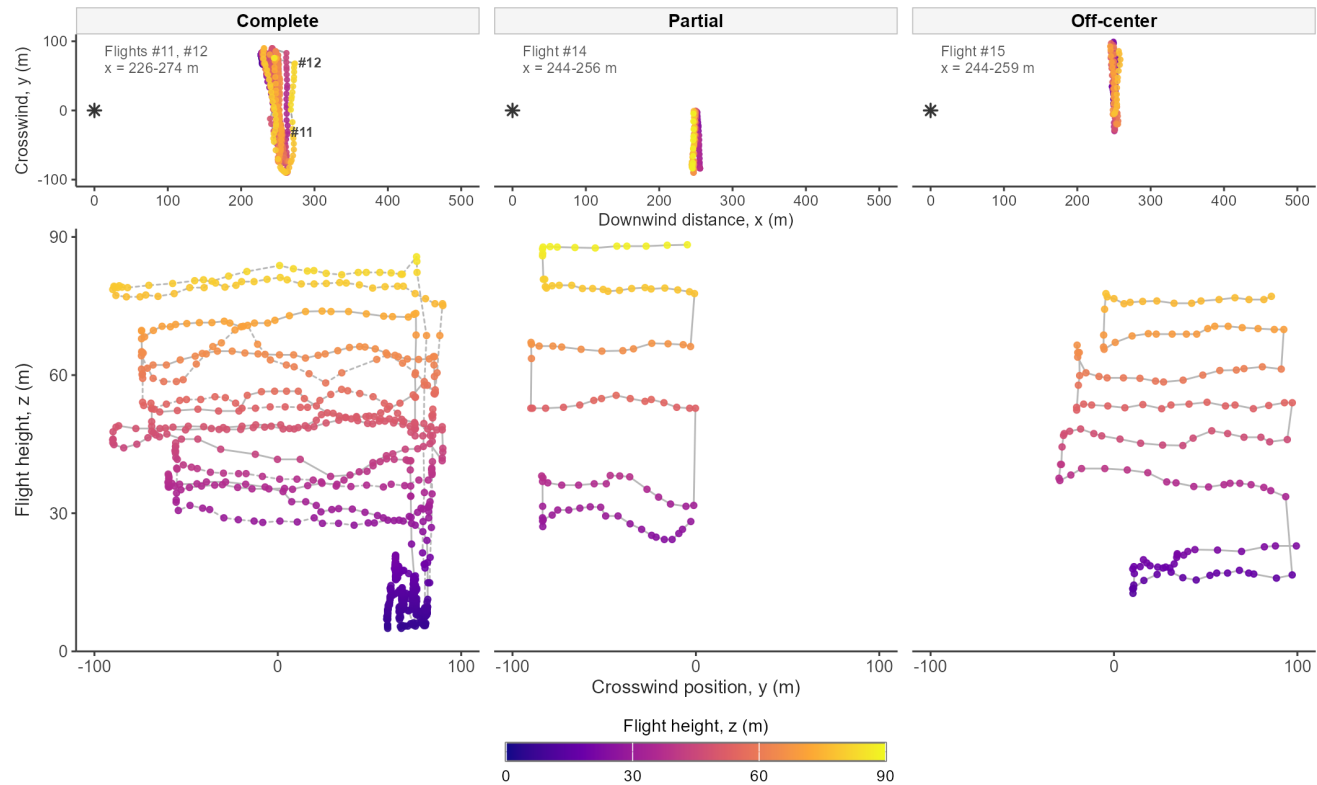
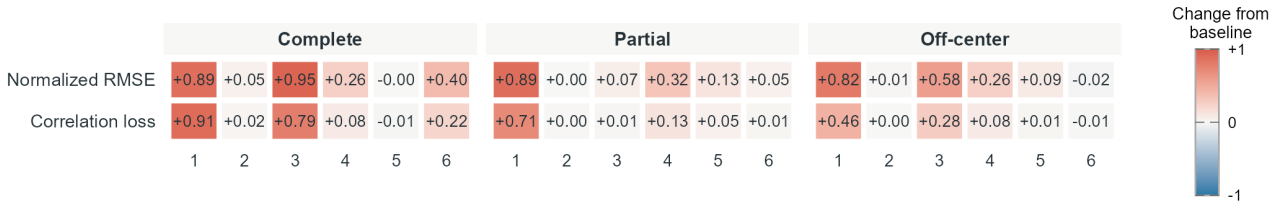


Figure G1. Complete, partial, and off-center transformed UAV sampling trajectories used in the controlled observation-geometry stress test. The upper strips locate the trajectories in the downwind–crosswind plane; stars mark the source and annotations identify the source flights and sampled downwind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

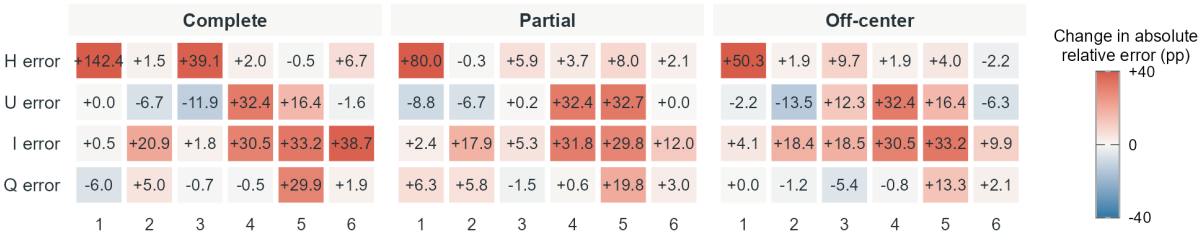


Figure G2. Case-wise degradation under three transformed UAV sampling protocols across six representative conditions. Rows show changes in normalized RMSE, correlation loss, and the absolute relative errors of H , U , I , and Q ; columns identify Cases 1–6 for each protocol. Changes are relative to each 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.

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

890 *Author contributions.* Lei Wang: conceptualization, methodology, software, validation, formal analysis, investigation, visualization, writing–original draft, and writing–review and editing. Xin Ma: conceptualization, supervision, project administration, funding acquisition, resources, and writing–review and editing.

Competing interests. The authors declare that they have no competing interests.

895 *Financial support.* This work was supported by the National Natural Science Foundation of China (Grant U25D9004 and 42171464), the Fundamental Research Funds for the Central Universities (ZNJC202415 and 2042026kf0075), the National Key R&D Program of China (2024YFC3015600), and the Science and Technology Program of Hubei Provincial (2025BEB017).

Acknowledgements. The numerical calculations reported in this paper were performed on the supercomputing system at the Supercomputing Center of Wuhan University. ~~Figure 1 was prepared as an AI-assisted schematic; an initial draft was generated using the gemini-3-pro-image-preview model, and the authors subsequently reviewed and manually revised the layout, icons, labels, and scientific content. The authors take full responsibility for the final figure.~~

- Aghdasi, S., Rayner, P., Deutscher, N., and Silver, J.: Exploring uncertainty reduction in high-resolution methane emissions in Gippsland through in-situ data: A Bayesian inverse modeling and variational assimilation method, *ATMOSPHERIC RESEARCH*, <https://doi.org/10.1016/j.atmosres.2025.107911>, 2025.
- Andersen, T., Zhao, Z., de Vries, M., Necki, J., Swolkien, J., Menoud, M., Röckmann, T., Roiger, A., Fix, A., Peters, W., and Chen, H.:
905 Local-to-regional methane emissions from the Upper Silesian Coal Basin (USCB) quantified using UAV-based atmospheric measurements, *ATMOSPHERIC CHEMISTRY AND PHYSICS*, <https://doi.org/10.5194/acp-23-5191-2023>, 2023.
- Cassiani, M., Stohl, A., and Brioude, J.: Lagrangian Stochastic Modelling of Dispersion in the Convective Boundary Layer with Skewed Turbulence Conditions and a Vertical Density Gradient: Formulation and Implementation in the FLEXPART Model, *Boundary-Layer Meteorology*, <https://doi.org/10.1007/s10546-014-9976-5>, 2014.
- 910 Chung, H., Kim, J., McCann, M. T., Klasky, M. L., and Ye, J. C.: Diffusion Posterior Sampling for General Noisy Inverse Problems, in: International Conference on Learning Representations, <https://doi.org/10.48550/arXiv.2209.14687>, 2023.
- Conley, S., Faloon, I., Mehrotra, S., Suard, M., Lenschow, D. H., Sweeney, C., Herndon, S., Schwietzke, S., Pétron, G., Pifer, J., Kort, E. A., and Schnell, R.: Application of Gauss's theorem to quantify localized surface emissions from airborne measurements of wind and trace gases, *Atmospheric Measurement Techniques*, <https://doi.org/10.5194/amt-10-3345-2017>, 2017.
- 915 Conrad, B. M., Tyner, D. R., and Johnson, M. R.: Robust probabilities of detection and quantification uncertainty for aerial methane detection: Examples for three airborne technologies, *Remote Sensing of Environment*, <https://doi.org/10.1016/j.rse.2023.113499>, 2023.
- De Meutter, P., Hoffman, I., and Ungar, K.: On the model uncertainties in Bayesian source reconstruction using an ensemble of weather predictions, the emission inverse modelling system FREAR v1.0, and the Lagrangian transport and dispersion model Flexpart v9.0.2, *GEOSCIENTIFIC MODEL DEVELOPMENT*, <https://doi.org/10.5194/gmd-14-1237-2021>, 2021.
- 920 Du, P., Parikh, M., Fan, X., Liu, X., and Wang, J.: Conditional neural field latent diffusion model for generating spatiotemporal turbulence, *NATURE COMMUNICATIONS*, <https://doi.org/10.1038/s41467-024-54712-1>, 2024.
- Eastwood, M., Thompson, D. R., Green, R. O., Fahlen, J., Adams, T., Brandt, A. R., Brodrick, P., Chlus, A., Kort, E. A., Reuland, F., and Thorpe, A. K.: Direct measurement of plume velocity to characterize point source emissions, *Proceedings of the National Academy of Sciences of the United States of America*, <https://doi.org/10.1073/pnas.2507350122>, 2025.
- 925 Elbern, H., Strunk, A., Schmidt, H., and Talagrand, O.: Emission rate and chemical state estimation by 4-dimensional variational inversion, *Atmospheric Chemistry and Physics*, <https://doi.org/10.5194/acp-7-3749-2007>, 2007.
- Guo, J., Du, P., Fan, X., Li, Y., and Wang, J.: Conditional neural field for spatial dimension reduction of turbulence data: A comparison study, *PHYSICS OF FLUIDS*, <https://doi.org/10.1063/5.0310238>, 2026.
- He, Z., Gao, L., Liang, M., and Zeng, Z.-C.: A survey of methane point source emissions from coal mines in Shanxi province of China using
930 AHSI on board Gaofen-5B, *Atmospheric Measurement Techniques*, <https://doi.org/10.5194/amt-17-2937-2024>, 2024.
- Jia, M., Fish, R., Daniels, W., Sprinkle, B., and Hammerling, D.: A fast and lightweight implementation of the Gaussian puff model for near-field atmospheric transport of trace gasses, *SCIENTIFIC REPORTS*, <https://doi.org/10.1038/s41598-025-99491-x>, 2025.
- Katata, G., Chino, M., Kobayashi, T., Terada, H., Ota, M., Nagai, H., Kajino, M., Draxler, R., Hort, M., Malo, A., Torii, T., and Sanada, Y.: Detailed source term estimation of the atmospheric release for the Fukushima Daiichi Nuclear Power Station accident by coupling
935 simulations of an atmospheric dispersion model with an improved deposition scheme and oceanic dispersion model, *ATMOSPHERIC CHEMISTRY AND PHYSICS*, <https://doi.org/10.5194/acp-15-1029-2015>, 2015.

- Krings, T., Gerilowski, K., Buchwitz, M., Hartmann, J., Sachs, T., Erzinger, J., Burrows, J. P., and Bovensmann, H.: Quantification of methane emission rates from coal mine ventilation shafts using airborne remote sensing data, *Atmospheric Measurement Techniques*, <https://doi.org/10.5194/amt-6-151-2013>, 2013.
- 940 Li, S. and Du, K.: Comparisons of forward-in-time and backward-in-time Lagrangian stochastic dispersion models for micro-scale atmospheric dispersion, *JOURNAL OF THE AIR & WASTE MANAGEMENT ASSOCIATION*, <https://doi.org/10.1080/10962247.2020.1728424>, 2020.
- Lin, J. C., Gerbig, C., Wofsy, S. C., Andrews, A. E., Daube, B. C., Davis, K. J., and Grainger, C. A.: A near-field tool for simulating the upstream influence of atmospheric observations: The Stochastic Time-Inverted Lagrangian Transport (STILT) model, *Journal of Geophysical Research: Atmospheres*, <https://doi.org/10.1029/2002JD003161>, 2003.
- 945 Liu, X., Parikh, M., Fan, X., Du, P., Wang, Q., Chen, Y., and Wang, J.: CoNFILd-inlet: Synthetic turbulence inflow using generative latent diffusion models with neural fields, *PHYSICAL REVIEW FLUIDS*, <https://doi.org/10.1103/PhysRevFluids.10.054901>, 2025a.
- Liu, Y., Paris, J.-D., Broquet, G., Roy, V. B., Fernandez, T. M., Andersen, R., Berlanga, A. R., Christensen, E., Courtois, Y., Dominok, S., Dussenne, C., Eckert, T., Finlayson, A., de la Fuente, A. F., Gunn, C., Hashmonay, R., Hayashi, J. G., Helmore, J., Honsel, S., Innocenti, F., Irjala, M., Log, T., López, C., Martínez, F. C., Martinez, J., Massardier, A., Nygaard, H. G., Reboredo, P. A., Rousset, É., Scherello, A., Ulbricht, M., Weidmann, D., Williams, O., Yarrow, N., Zaréa, M., Ziegler, R., Sciare, J., Vrekoussis, M., and Bousquet, P.: Assessment of current methane emission quantification techniques for natural gas midstream applications, *Atmospheric Measurement Techniques*, <https://doi.org/10.5194/amt-17-1633-2024>, 2024.
- 950 Liu, Y., Zhang, Z., Yu, D., Liu, P., Yan, X., and Zhang, K.: Conditional diffusion inversion for Darcy flows in heterogeneous porous media with random parameters, *PHYSICS OF FLUIDS*, <https://doi.org/10.1063/5.0281578>, 2025b.
- Maksyutov, S., Oda, T., Saito, M., Janardanan, R., Belikov, D., Kaiser, J. W., Zhuravlev, R., and Ganshin, A.: Technical note: A high-resolution inverse modelling technique for estimating surface CO₂ fluxes based on the NIES-TM-FLEXPART coupled transport model and its adjoint, *Atmospheric Chemistry and Physics*, <https://doi.org/10.5194/acp-21-1245-2021>, 2021.
- Mbua, M., Riddick, S., Kiplimo, E., Shonkwiler, K., Hodshire, A., and Zimmerle, D.: Evaluating the feasibility of using downwind methods to quantify point source oil and gas emissions using continuously monitoring fence-line sensors, *ATMOSPHERIC MEASUREMENT TECHNIQUES*, <https://doi.org/10.5194/amt-18-5687-2025>, 2025.
- 960 Meier, S., Vögtli, M., Hueni, A., McManemin, A., Brandt, A. R., Juéry, C., Blandin, V., Brunner, D., and Kuhlmann, G.: Quantifying CH₄ point source emissions with airborne remote sensing: first results from AVIRIS-4, *Atmospheric Measurement Techniques*, 19, 333–358, <https://doi.org/10.5194/amt-19-333-2026>, 2026.
- 965 Meng, C., Gao, J., Tian, Y., Chen, H., Zhang, W., and Luo, R.: Stochastic Solutions for Simultaneous Seismic Data Denoising and Reconstruction via Score-Based Generative Models, *IEEE Transactions on Geoscience and Remote Sensing*, <https://doi.org/10.1109/TGRS.2024.3421597>, 2024.
- Mohammadloo, T. H., Jones, M. C., van de Kerkhof, B., Dawson, K., Smith, B. J., Conley, S. A., Corbett, A., and Ijzermans, R.: Quantitative estimate of several sources of uncertainty in drone-based methane emission measurements, *Atmospheric Measurement Techniques*, <https://doi.org/10.5194/amt-18-1301-2025>, 2025.
- 970 Morales, R., Ravelid, J., Vinkovic, K., Korbe, P., Tuzson, B., Emmenegger, L., Chen, H., Schmidt, M., Humbel, S., and Brunner, D.: Controlled-release experiment to investigate uncertainties in UAV-based emission quantification for methane point sources, *ATMOSPHERIC MEASUREMENT TECHNIQUES*, <https://doi.org/10.5194/amt-15-2177-2022>, 2022.

- Newman, J. F. and Klein, P.: The Impacts of Atmospheric Stability on the Accuracy of Wind Speed Extrapolation Methods, *Resources*, <https://doi.org/10.3390/resources3010081>, 2014.
- Scheutz, C., Knudsen, J., Vechi, N., and Knudsen, J.: Validation and demonstration of a drone-based method for quantifying fugitive methane emissions, *JOURNAL OF ENVIRONMENTAL MANAGEMENT*, <https://doi.org/10.1016/j.jenvman.2024.123467>, 2025.
- Shah, A., Allen, G., Pitt, J., Ricketts, H., Williams, P., Helmore, J., Finlayson, A., Robinson, R., Kabbabe, K., Hollingsworth, P., Rees-White, T., Beaven, R., Scheutz, C., and Bourn, M.: A Near-Field Gaussian Plume Inversion Flux Quantification Method, Applied to Unmanned Aerial Vehicle Sampling, *ATMOSPHERE*, <https://doi.org/10.3390/atmos10070396>, 2019.
- Shaw, J., Shah, A., Yong, H., and Allen, G.: Methods for quantifying methane emissions using unmanned aerial vehicles: a review, *PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A-MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES*, <https://doi.org/10.1098/rsta.2020.0450>, 2021.
- Stohl, A., Forster, C., Frank, A., Seibert, P., and Wotawa, G.: Technical note: The Lagrangian particle dispersion model FLEXPART version 6.2, *Atmospheric Chemistry and Physics*, <https://doi.org/10.5194/acp-5-2461-2005>, 2005.
- Thomson, D. J.: Criteria for the Selection of Stochastic Models of Particle Trajectories in Turbulent Flows, *Journal of Fluid Mechanics*, <https://doi.org/10.1017/S0022112087001940>, 1987.
- Varon, D., Jacob, D., McKeever, J., Jervis, D., Durak, B., Xia, Y., and Huang, Y.: Quantifying methane point sources from fine-scale satellite observations of atmospheric methane plumes, *ATMOSPHERIC MEASUREMENT TECHNIQUES*, <https://doi.org/10.5194/amt-11-5673-2018>, 2018.
- Vollrath, C., Barchyn, T., Munn, A., Wearmouth, C., and Hugenholtz, C.: A human-portable mass flux method for methane emissions quantification: controlled release testing performance evaluation, *ATMOSPHERIC MEASUREMENT TECHNIQUES*, <https://doi.org/10.5194/amt-19-583-2026>, 2026.
- Whites, E., Rigby, M., Lunt, M., Smallman, T., Comyn-Platt, E., Manning, A., Ganesan, A., O'Doherty, S., Stavert, A., Stanley, K., Williams, M., Levy, P., Ramonet, M., Forster, G., Manning, A., and Palmer, P.: Quantifying the UK's carbon dioxide flux: an atmospheric inverse modelling approach using a regional measurement network, *ATMOSPHERIC CHEMISTRY AND PHYSICS*, <https://doi.org/10.5194/acp-19-4345-2019>, 2019.
- Xie, Z., Tang, J., Li, R., Tian, G., and Ma, R.: A Review of Source-Term Estimation for Continuous Methane Monitoring: From Data Acquisition to Modeling and Estimation, *ACS OMEGA*, <https://doi.org/10.1021/acsomega.5c13215>, 2026.
- Xu, Y., Dong, X., and Fang, S.: Efficient Bayesian source reconstruction and uncertainty quantification of atmospheric radionuclide releases by replacing release rate sampling with Maximum-A-Posteriori estimation of time-varying release rates, *JOURNAL OF HAZARDOUS MATERIALS*, <https://doi.org/10.1016/j.jhazmat.2025.138171>, 2025.
- Yuvaraj, R., Lauvaux, T., Abdallah, C., Ciais, P., Guery, J. A., Bonne, J., Groshenry, A., Hoang, N. M., and Joly, L.: High-Resolution Modeling of Methane Plumes: Validation and Sensitivity Experiments to Explore Emission Quantification Approaches., *Environmental science & technology*, <https://doi.org/10.1021/acs.est.5c15941>, 2026.
- Zhao, Z., Sun, S., Zhou, B., Sun, J., Ciais, P., Houweling, S., Krol, M., and Chen, H.: Advancing Point Source Methane Emissions Quantification With UAV Sampling and Deep Learning: A Large-Eddy Simulation Study, *JOURNAL OF GEOPHYSICAL RESEARCH-ATMOSPHERES*, <https://doi.org/10.1029/2025JD045952>, 2026.