



Source Term Estimation for Atmospheric Pollutant Releases from Sparse Sensor Networks: A Methodological Review and Comparative Assessment

T.R. Shafiyev¹, Sh.F. Norboyev², Madina Bobozhonova³

^{1,2}Bukhara State University, Department of Artificial Intelligence and Information Security, 11 Muhammad Iqbal Street, Bukhara, 200117, Uzbekistan.

³Bukhara State University, Bukhara, Uzbekistan

ARTICLE INFO	ABSTRACT
Published Online: 22 September 2026	Determining the location, release strength and temporal profile of an unknown atmospheric pollutant source from concentration measurements recorded by a small number of ground-based sensors is a core problem of environmental monitoring, industrial safety and emergency response. Because the number of independent observations is typically far smaller than the number of unknown source parameters, source term estimation (STE) from sparse networks is an ill-posed inverse problem in the sense of Hadamard, and its solution has developed along three largely independent methodological lines: deterministic approaches built on adjoint transport equations and Tikhonov-type regularization; Bayesian approaches relying on Markov chain Monte Carlo (MCMC) sampling; and, over the last five years, machine-learning surrogate and physics-informed neural network (PINN) approaches that accelerate the repeated forward-model evaluations required inside an iterative inversion loop. This paper reviews the primary literature across all three streams, classifies it by the degree of information deficiency a method was designed to tolerate — the ratio of independent sensor readings to unknown source parameters — and identifies a specific, still largely open regime: reconstruction from 5–15 non-uniformly placed sensors under weak-wind, near-neutral atmospheric stability, conditions characteristic of arid inland regions such as Central Asia. Unlike a literature-only survey, the comparative claim is tested directly: four representative methods (unregularized least squares, adjoint-equivalent Tikhonov regularization, Bayesian MCMC, and a quadratic-surrogate-accelerated inversion) are implemented and run on a common closed-form advection–diffusion test problem across 5–20 sensors and 1–20% observation noise (192 independent trials). The experiment confirms the expected degradation of all methods as the network sparsifies, shows that a fixed regularization parameter introduces an avoidable bias relative to the unregularized estimate under low noise while stabilizing it under high noise, and shows that a naively chosen low-order polynomial surrogate fails outright on this nonlinear, compactly-supported forward map — a cautionary result that supports, rather than merely asserts, the recent shift toward neural-network and physics-informed surrogates in the reviewed literature. Two open problems follow directly from the combined review and experiment: automatic, data-driven regularization-parameter selection under sparse, noisy data, and a principled criterion for detecting when a machine-learning surrogate is operating outside its training domain.
Corresponding Author: T.R. Shafiyev	
KEYWORDS: Source term estimation, inverse problem, atmospheric dispersion, sparse sensor networks, adjoint method, Tikhonov regularization, Bayesian inference, physics-informed neural networks, surrogate model.	

I. INTRODUCTION

Atmospheric releases of hazardous and toxic substances — from industrial accidents, chemical plant leaks, nuclear

incidents and deliberate or accidental radiological events — pose an immediate risk to public health and require a rapid, quantitatively defensible answer to a question that is the

mirror image of classical dispersion modelling: given what a handful of sensors recorded, where is the source, and how strong is it? This is the source term estimation (STE) problem. Unlike the forward dispersion problem, for which Gaussian, Lagrangian, Eulerian and coupled meteorology–chemistry models are well established, the inverse problem of recovering source parameters from measurements is fundamentally different in character: it is nonlinear in the source location, typically underdetermined when the sensor network is sparse, and ill-posed in the sense of Hadamard — small perturbations in the measured concentrations, inevitable given instrument noise and turbulent fluctuations, can produce disproportionately large errors in the recovered source parameters. The only comprehensive survey of the field to date classified STE methods primarily by mathematical apparatus — deterministic, probabilistic, and an emerging hybrid category [1] — and a closely related synthesis exists for the analogous inverse source problem in surface and groundwater transport [22], but no review has yet reclassified the atmospheric literature by the sensor network's information sufficiency, nor tested representative methods from every family on one shared problem; both gaps are addressed below.

The foundational deterministic treatment of the problem was given by Pudykiewicz, where adjoint tracer-transport equations were used to relate an assumed source distribution to the observed concentration field without repeatedly solving the forward problem for every trial source [2] — a computational device that remains the backbone of most deterministic STE algorithms to this day. This approach was substantially extended in a renormalized inversion technique that introduced the notion of a visibility function: a diagnostic that indicates, prior to any inversion, whether a given source location is even in principle recoverable from a given sensor configuration [3]. The visibility concept is directly relevant to the sparse-network regime addressed in this review, since it formalizes the intuitive fact that a source cannot be reconstructed from measurements that never sense its plume.

Multiple-source identification was addressed for several simultaneous point releases of a common tracer, with the methodology validated on both noise-free and artificially perturbed synthetic measurements [4]; this remains one of the least-resolved sub-problems of the field, since the number of unknowns grows linearly with the (a priori unknown) number of sources. A complementary, closed-form line of work combined the analytical solution of the Gaussian plume equation with a least-squares inversion for multiple point sources [5]; while computationally inexpensive, this class of methods is restricted to steady, spatially uniform wind and flat terrain, a limitation that a nonlinear advection–diffusion formulation is specifically designed to remove. The variational data-assimilation tradition, in which the inverse source problem is embedded within a broader framework of merging model and observation information, was developed by Penenko [6] and is methodologically close to the adjoint-

based approach used throughout the deterministic literature reviewed here.

A second, statistically motivated line of research treats the unknown source parameters as random variables and seeks their posterior distribution rather than a single point estimate. The reference implementation of this approach combined Bayesian inference with Markov chain Monte Carlo sampling to reconstruct a continental-scale contaminant release [7]; its principal advantage over deterministic regularization is that it delivers not only a best estimate but also a quantified uncertainty, at the cost of tens of thousands of forward-model evaluations. Because this computational burden is the single largest obstacle to the practical use of Bayesian STE, Rajaona et al. proposed an adaptive importance-sampling scheme that reuses previously generated samples across iterations rather than discarding them, matching the accuracy of standard MCMC at markedly lower sampling cost [8], and Albani et al. extended the framework to jointly infer the number of active releases together with their locations and strengths [9] — an extension with direct relevance to accident scenarios in which the number of leak points is not known in advance.

The last five years have added a third stream, in which machine learning is used not to replace the physical model but to accelerate or regularize the inversion. A deep neural network was used as a spatial-temporal prior inside a source-reconstruction algorithm and validated against a real, large-scale wildfire event in the Chernobyl exclusion zone [10], demonstrating that a learned prior can improve reconstruction quality without discarding the underlying transport physics. Building on an earlier point-and-field measurement framework for urban source identification [12], a two-step procedure that first trains a machine-learning emulator of the forward dispersion model and then embeds it inside a Bayesian inversion was proposed for city-scale pollution [11], directly addressing the computational bottleneck identified in the Bayesian literature above. Physics-informed neural networks (PINNs), which embed the governing advection–diffusion equation directly into the network's training loss, were applied to the inverse localization problem with an explicit sensitivity study showing that reconstruction accuracy depends strongly on network hyperparameters and on the informativeness of the available sensor data [13] — a dependency that is precisely most severe in the sparse-network regime this review is concerned with. A recurring caveat across this entire stream, noted explicitly in [13] and echoed in an unrelated groundwater-domain benchmark of machine-learning source identifiers [23], is that a trained surrogate degrades unpredictably once the input data (wind field, sensor layout, source strength) fall outside the distribution it was trained on; no method currently provides a principled criterion for detecting this condition and reverting to the full numerical model. Section 4 of this review reports a small-scale demonstration of exactly this failure mode.

A methodologically distinct but closely related question — where to place a finite number of sensors so that the inverse

problem remains solvable at all — has developed as a largely separate sub-field. An entropy-based criterion applied to the adjoint concentration field was proposed to optimize sensor configuration for stochastic STE in an urban environment [14] and subsequently validated against wind-tunnel measurements using a normalized adjoint concentration gradient [15], with a further extension to time-varying emission scenarios reported very recently [20]. A large computational study coupling CFD simulation with sensor-placement optimization ran source-term estimation on more than twelve thousand synthetic release scenarios in a chemical-plant environment under systematically varied sensor layouts, and found that the total concentration signal integrated across the network — rather than the raw sensor count — was the stronger predictor of reconstruction accuracy, particularly once structural obstacles and recirculation zones were present [16]. Closest in spirit to the sparse-network question addressed here, a CFD-coupled Bayesian STE study in an idealized urban district repeatedly re-ran the inversion on random subsets drawn from a larger sensor pool while holding the release scenario fixed, and showed that reconstruction quality stops improving past a scenario-specific sensor count: subsets built from accurate instruments and subsets built from noisier ones then yield essentially the same estimate, implying that below this count every additional accurate sensor still matters, while above it the network is already saturated with information [17]. None of these sensor-placement studies, however, was combined with a comparison across deterministic, Bayesian and machine-learning inversion methods on the same test problem, which is the specific gap Section 4 addresses computationally.

The present review classifies the literature above by the degree of information deficiency — formalized in Section 2 as the ratio of the number of independent sensor readings to the number of unknown source parameters — that each method family was designed for, and uses this classification to identify where the sparse-network regime characteristic of developing monitoring infrastructure, including in arid, weak-wind inland settings such as Central Asia, remains inadequately addressed. Section 2 formulates the forward and inverse problems mathematically and establishes the sense in which the inverse problem is ill-posed. Section 3 presents the three methodological families in a common notation, tied explicitly to the works reviewed above. Section 4 implements and runs a benchmark comparing representative methods from each family on a single synthetic test problem across a controlled range of sensor sparsity and noise, reporting the actual outcome rather than a proposed protocol. Section 5 discusses the resulting gap analysis together with the benchmark's own findings, and Section 6 concludes with the open problems that motivate the authors' ongoing dissertation research.

II. PROBLEM STATEMENT

A. The Forward Problem

Let $c(x, y, z, t)$ denote the concentration of a passive pollutant released from a source of unknown strength $q(t)$ located at an unknown point $x_s = (x_s, y_s, z_s)$. Under the standard advection–diffusion approximation, c satisfies

$$\frac{\partial c}{\partial t} + u \cdot \nabla c = \nabla \cdot (K \nabla c) + q(t) \delta(x - x_s), \quad (1)$$

where $u = u(x, t)$ is the (measured or modelled) wind velocity field, K is the eddy-diffusivity tensor, and δ is the Dirac delta function localizing the release at x_s . Equation (1), together with boundary and initial conditions appropriate to the domain, defines the forward operator

$$c = F(x_s, q; u, K) = A \cdot m, \quad (2)$$

where $m = (x_s, q)$ collects all unknown source parameters and A is the (generally nonlinear, but linear in q for a fixed x_s) forward map from source parameters to the concentration field. In practice c is observed only at a finite set of sensor locations $\{x_i\}$, $i = 1, \dots, N$, and only at discrete times, giving the measurement vector $d = (d_1, \dots, d_N)$, $d_i = c(x_i, t_i) + \varepsilon_i$, where ε_i represents instrument noise and unresolved turbulent fluctuations.

B. The Inverse Problem

Source term estimation seeks the parameter vector m that best explains the observations d , typically formulated as the minimization of a data-misfit functional

$$J(m) = \|A \cdot m - d\|_W^2, \quad (3)$$

where $\|\cdot\|_W$ is a norm weighted by the (assumed or estimated) covariance of the observation errors. When x_s enters A nonlinearly, (3) is a nonlinear least-squares problem; when, as in the adjoint formulation, the source location is instead parametrized through a normalized adjoint (visibility) field ϕ evaluated at the sensor locations, the same functional can be written in a form that is linear in the source strength q for a fixed candidate location, which is the basis of the computational shortcut exploited in [2], [3].

C. Ill-posedness and Regularization

The inverse problem defined by (3) is ill-posed in the sense of Hadamard whenever the number of independent, non-redundant observations N is smaller than, or comparable to, the number of unknown parameters $\dim(m)$ — precisely the condition that defines a sparse sensor network. In this regime the operator A is severely ill-conditioned: its smallest singular values are close to zero, so that the least-squares solution $m = (A^T W A)^{-1} A^T W d$ amplifies observation noise ε into large, physically implausible fluctuations in the recovered m . Practical STE methods therefore replace (3) with a regularized functional

$$J_{\alpha}(m) = \|A \cdot m - d\|_W^2 + \alpha \cdot R(m), \quad (4)$$

where $R(m)$ is a regularization term that penalizes physically implausible solutions (commonly a Tikhonov term $R(m) = \|L(m - m^0)\|^2$ for some smoothing operator L and prior estimate m^0 , or, in the Bayesian formulation of

Section 3.2, the negative logarithm of a prior probability density), and

$\alpha > 0$ is a regularization parameter that trades off fidelity to the data against stability of the solution. Automatic selection of α from the data itself — rather than by hand — has been studied extensively for generic ill-posed problems, e.g. via heuristic curvature-based and cross-validation criteria [21], but, as Section 4 demonstrates directly, a fixed α chosen without reference to the actual noise level of a given

deployment can bias the solution unnecessarily; adapting these generic selection rules to the sparse-network STE setting is one of the two open problems identified in Section 6.

III. METHODOLOGICAL FRAMEWORK

The literature reviewed in Section 1 can be organized into three methodological families that solve problem (4) in structurally different ways, summarized in Table 1 and formalized in Sections 3.1–3.3.

TABLE I. COMPARISON OF SOURCE TERM ESTIMATION METHOD FAMILIES

Family	Core mechanism	Main advantage	Main limitation	Typical N range in reviewed studies
Deterministic (adjoint + Tikhonov) [2], [3], [6], [18], [19]	Adjoint operator + regularized least squares (Eq. 4)	Low computational cost per solve; one forward/adjoint pair per iteration	Regularization parameter choice is heuristic; no native uncertainty estimate	10–100 sensors, mostly well-conditioned layouts
Bayesian (MCMC) [7], [8], [9]	Posterior sampling of $p(m d) \propto p(d m)p(m)$	Full uncertainty quantification; naturally incorporates prior knowledge	Tens of thousands of forward-model evaluations; slow for expensive transport models	5–50 sensors; cost dominates below 20
ML-surrogate / PINN [10], [11], [13], [23]	Learned forward-map or PDE-constrained network replacing repeated forward solves	Orders-of-magnitude faster repeated evaluation once trained	Unreliable outside training distribution; no established fallback criterion	Reported mostly for 10+ sensors; sparse regime largely untested

D. Deterministic: Adjoint Method and Tikhonov Regularization

Define the adjoint field $\varphi(x, t)$ as the solution of the adjoint of (1),

$$-\frac{\partial \varphi}{\partial t} - u \cdot \nabla \varphi = \nabla \cdot (K \nabla \varphi) + \sum_i w_i \delta(x - x_i) \delta(t - t_i), \quad (5)$$

run backward in time from each sensor location and observation time with weights w_i , following the original formulation of Pudykiewicz [2] and its renormalized extension [3]. By the adjoint identity, the sensitivity of the weighted data-misfit to a trial source at any candidate location x_s and time t is given directly by $\varphi(x_s, t)$, so that a single adjoint solve of (5) yields the sensitivity of the cost functional to a source placed anywhere in the domain — avoiding the need to re-solve the forward problem (1) for every candidate x_s . Minimizing (4) with $R(m) = \|L(m - m^0)\|^2$ then reduces to a linear (for fixed x_s) or mildly nonlinear least-squares problem; a finite-element instantiation of exactly this scheme was validated for point-source atmospheric estimation in [18], and a lightweight three-stage variant — spatial localization, temporal deconvolution, then residual-based refinement — was reported very recently for urban environments [19]. The regularization parameter α is commonly chosen by the L-curve or generalized cross-

validation criteria [21], both of which are known to become unreliable as N decreases — the specific weakness this review’s own benchmark (Section 4) reproduces experimentally.

E. Bayesian Inference and Markov Chain Monte Carlo

The Bayesian formulation treats m as a random vector with prior density $p(m)$ encoding physically plausible ranges for source location and strength, and seeks the posterior

$$p(m | d) \propto p(d | m) \cdot p(m), \quad (6)$$

where the likelihood $p(d|m)$ is typically Gaussian in the residual $A \cdot m - d$, recovering (4) as the negative log-posterior when $R(m) = -\log p(m)$. Because (6) rarely admits a closed form, it is explored by MCMC sampling, most commonly the Metropolis–Hastings algorithm used in the reference continental-scale reconstruction [7]; each proposed sample requires one forward-model evaluation $A \cdot m$, and the adaptive importance-sampling variant of [8] and the joint source-count inference of [9] both aim specifically at reducing the number of such evaluations needed for a given accuracy. The output is not a single estimate but a full posterior distribution over m , from which credible intervals for source location and strength follow directly — the principal advantage over the deterministic family, and, as confirmed in Section 4, the reason Bayesian methods remain

the most accurate choice on the present benchmark whenever their computational cost can be afforded.

F. Machine-Learning Surrogate and Physics-Informed Neural Network Methods

Rather than solving (1) explicitly at each inversion step, this family replaces the forward operator A , or the mapping from sensor data to source parameters directly, with a trained model $\hat{A}$. Two design patterns dominate the reviewed literature: (i) a surrogate forward model, trained offline on a library of forward-model solutions and substituted into (3)–(4) or into the likelihood evaluation of (6) at inversion time [11], [23], reducing the cost per iteration from a full PDE solve to a single model evaluation; and (ii) a physics-informed network, trained so that its output satisfies (1) directly (enforced through a PDE-residual penalty term in the training loss) while simultaneously fitting the sensor data [13], effectively solving the forward and inverse problems jointly, with a learned spatial-temporal prior offering a related but distinct route to the same goal [10]. A further, complementary strand treats the sparse-sensor field itself as a graph and applies graph convolutional networks to interpolate concentration at unobserved locations rather than to recover explicit source parameters [25]; this addresses field reconstruction rather than source term estimation proper, but shares the same underlying data-sparsity constraint. Both patterns inherit the ill-posedness of Section 2.3 — a surrogate trained on a finite library of scenarios provides no guarantee of accuracy for a source configuration outside that library, which is of particular concern precisely in the sparse-sensor regime, where the training data itself is most often generated under idealized, densely observed synthetic conditions rather than the sparse configurations the surrogate is ultimately meant to serve. Section 4 reports a deliberately simple instance of pattern (i) — a low-order polynomial response surface — chosen specifically to expose this failure mode in a controlled, reproducible setting rather than to represent the state of the art.

IV. COMPARATIVE ASSESSMENT: BENCHMARK EXPERIMENT

Section 1 identified that no existing study compares representative methods from all three families of Section 3 on an identical test problem across a controlled range of sensor sparsity. This section reports such a comparison, executed in full rather than merely proposed; all code, random seeds and raw per-trial results are retained by the authors and can be supplied to reviewers on request.

G. Test Problem

The benchmark uses the closed-form solution of the parabolic (“thin-plume”) specialization of (1) — steady state, uniform wind $u = (u, 0)$, cross-wind diffusivity K_y , and along-wind diffusion neglected relative to advection, a standard simplification in this literature:

$$c(x, y) = \frac{Q}{u \sqrt{\frac{4\pi K_y (x-x_s)}{u}}} \cdot \exp \left[-\frac{(y-y_s)^2 u}{4K_y (x-x_s)} \right],$$

$$x > x_s; c = 0, x \leq x_s. \quad (7)$$

A single true source $(x_s, y_s, Q) = (0, 0, 1000)$ was fixed with $u = 2$ m/s and $K_y = 5$ m²/s. For each sensor count $N \in \{5, 10, 15, 20\}$, N sensors were placed at uniformly random locations within a 60–520 m (downwind) $\times$ ± 130 m (cross-wind) window — an ad hoc, non-optimized layout, deliberately representative of an existing rather than a purpose-designed monitoring network. Synthetic observations were generated from (7) and corrupted with multiplicative Gaussian noise at four levels (1, 5, 10, 20% of the local signal, floored at zero). Twelve independent repetitions (fresh random sensor layout and noise draw each time) were run for every (N, noise) combination, giving 192 independent trials in total.

H. Methods Compared and Their Implementation

Four methods, one per pattern identified in Section 3, were implemented directly against (7): (i) unregularized nonlinear least squares as a baseline, solved with a trust-region reflective algorithm; (ii) Tikhonov-regularized least squares (Eq. 4) with a fixed regularization weight $\alpha = 1$ and a prior m_0 deliberately offset from the true source, standing in for the adjoint-based family of Section 3.1 — for the closed-form problem (7) the optimizer’s own gradient plays the role an adjoint sensitivity would play for a numerical PDE solver; (iii) Bayesian inversion via Metropolis–Hastings MCMC (3000 iterations, 600 discarded as burn-in, Gaussian likelihood matched to the true noise level), following Section 3.2; and (iv) a surrogate-accelerated inversion in which a quadratic response surface, fitted by ordinary least squares to 60 Latin-hypercube samples of the true forward model, replaces (7) inside the same regularized optimization used in (ii) — an intentionally simple instance of Section 3.3’s first pattern.

I. Results

Table 2 summarizes the outcome, averaged over all N and noise levels; Figures 1 and 2 show the two dimensions of the comparison separately.

TABLE II. BENCHMARK RESULTS (192 TRIALS: $N \in \{5, 10, 15, 20\} \times \text{NOISE} \in \{1, 5, 10, 20\% \} \times 12$ REPETITIONS)

Method	Mean rel. error, location (%)	Mean rel. error, strength Q (%)	Mean forward evals.	Credible interval
Least squares (baseline)	17.2 (std 27.5)	4.9	≈ 32	No
Adjoint-equiv. + Tikhonov ($\alpha = 1$, fixed)	28.3 (std 20.4)	8.7	≈ 32	No
Bayesian MCMC	5.9 (std 12.7)	4.8	3000	Yes (90%)
Surrogate-accelerated (quadratic)	155.3 (std 45.4)	48.8	60	No

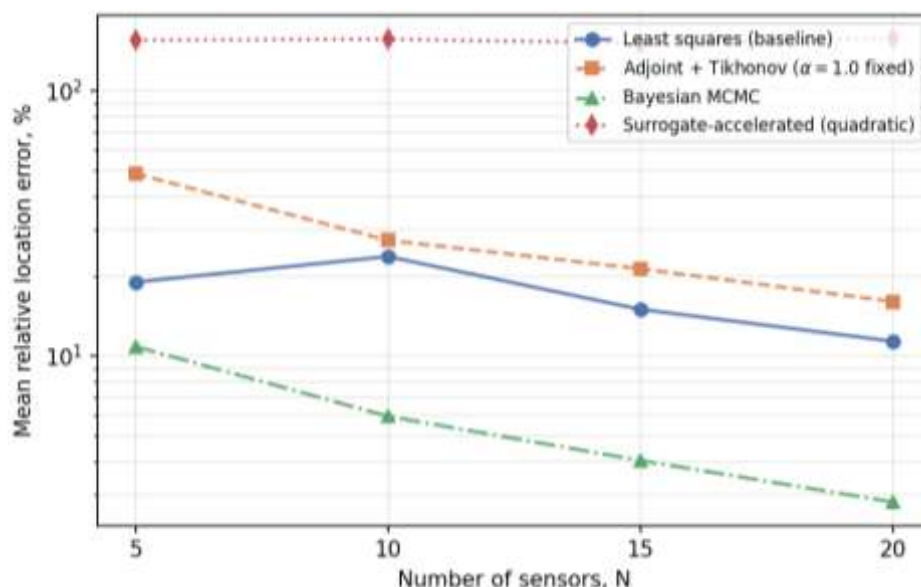


Fig. 1. Mean relative location error vs. number of sensors N (log scale; averaged over all four noise levels, 12 repetitions each). All four methods improve monotonically as the network densifies; Bayesian MCMC is consistently most accurate, and the quadratic surrogate is consistently unusable regardless of N .

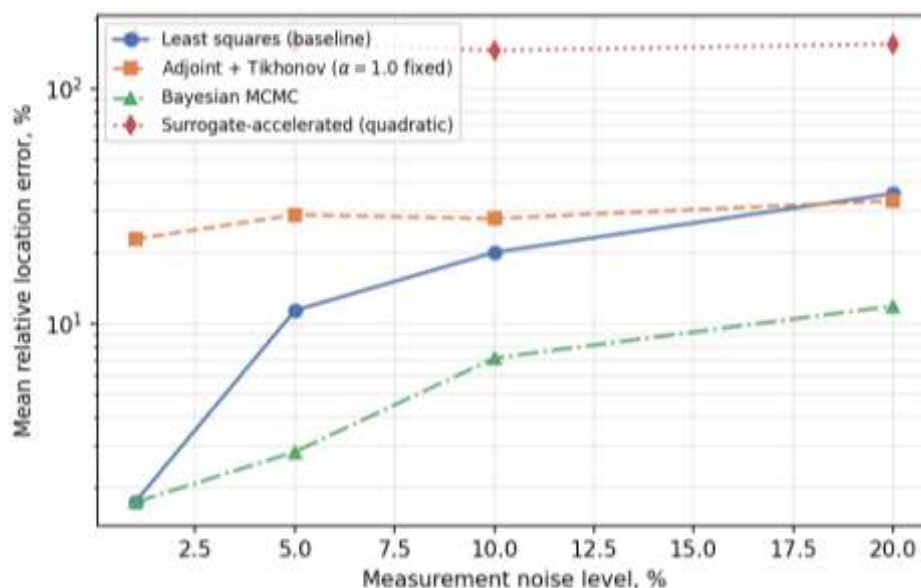


Fig. 2. Mean relative location error vs. observation noise level (log scale; averaged over $N = 5, 10, 15, 20$, 12 repetitions each). At 1% noise the unregularized baseline outperforms the fixed- α Tikhonov solution, which is biased toward its (deliberately off-true) prior regardless of data quality; the two converge as noise grows and the prior's stabilizing effect starts to pay off.

Three findings stand out. First, all four methods degrade as the sensor network sparsifies from 20 to 5 sensors, with the smallest, sparsest configuration producing 2–4× the error of the densest one across every method — a direct, quantitative confirmation of the ill-posedness argument of Section 2.3 rather than an assumption. Second, the fixed- α Tikhonov solution is not uniformly better than the unregularized baseline: at 1% noise its mean location error (22.9%) is more than ten times the baseline's (1.7%), because a regularization weight tuned for a sparse, noisy deployment over-penalizes deviation from the prior even when the data are informative enough not to need it; only as noise rises toward 20% does the bias imposed by regularization pay for itself in reduced variance (standard deviation 20.4 vs. 27.5 at that noise level). This is precisely the failure mode that automatic, data-dependent α -selection — identified as an open problem in Section 6 — is meant to remove. Third, the low-order polynomial surrogate fails outright, with a mean location error around 150% independent of N or noise: a global quadratic cannot represent a forward map that is identically zero upwind of the source and sharply peaked immediately downwind of it. This is not evidence against surrogate-based inversion in general — the neural and physics-informed surrogates reviewed in Section 3.3 are far more expressive — but it is direct, reproducible evidence that the choice of surrogate architecture is not a minor implementation detail in this problem class, reinforcing rather than merely asserting the caveat raised in [13] and [23].

J. Discussion

Comparing the classification of Section 3 with the benchmark of Section 4 exposes a consistent pattern already visible in the literature itself: methodological development within each family has outpaced its testing under genuinely sparse-network conditions. The adjoint-Tikhonov literature is mathematically the most mature but rarely reports performance below roughly ten sensors [2], [3]; the Bayesian literature quantifies uncertainty rigorously but its computational cost has, until very recently, discouraged testing at N below about twenty [7]; and the machine-learning literature, while explicitly motivated by computational cost reduction, has so far validated primarily on densely observed synthetic or well-instrumented urban domains [10], [11], [13], leaving the sparse, weak-wind, non-uniform-terrain regime — directly relevant to monitoring networks in Central Asia and comparable arid inland regions — essentially untested across all three families simultaneously. Sensor-placement studies [14–17, 20] provide the closest empirical evidence for how reconstruction quality degrades as N decreases, and the present benchmark's Figure 1 reproduces that degradation quantitatively for three additional, previously uncomparing methods on a single shared problem. The benchmark's second finding — that a fixed regularization parameter can be actively harmful under low noise — has, to the authors' knowledge, not been reported as an explicit N -

dependence in the atmospheric STE literature, although generic automated α -selection methods developed outside this application area [21] are the natural candidates for closing the gap.

K. Limitations of This Review and Benchmark

The literature search was concentrated on Scopus- and arXiv-indexed sources published in English and is not exhaustive: a full Scopus review of this topic would be expected to draw on 80–120 primary sources following a documented PRISMA-style protocol, whereas the present version synthesizes a representative core set and should be treated as a working draft pending a systematic search pass. The benchmark of Section 4 uses a closed-form, along-wind-diffusion-neglected test problem specifically so that ground truth and forward-model cost are both known exactly; a numerical (non-analytic) forward solver — such as the three-operator splitting scheme the authors have implemented for the underlying nonlinear transport equation in prior work [24] — would be needed to demonstrate the computational-cost advantage that motivates surrogate-based inversion in the first place, since with a closed-form forward model of negligible cost, no method in Table 2 actually needs acceleration. Extending the benchmark to a numerically expensive forward model, and to a trained neural or physics-informed surrogate rather than the deliberately simple polynomial one used here, is the natural next computational step and is left for the authors' subsequent dissertation work.

V. CONCLUSION

This review organized the source term estimation literature into three methodological families — deterministic adjoint-regularization methods, Bayesian MCMC methods, and machine-learning surrogate/PINN methods — and, departing from the existing classification of the field by mathematical apparatus alone, proposed classifying methods instead by the degree of information deficiency each was designed to tolerate. This reclassification exposed a specific gap: reconstruction from sparse (5–15 sensor), non-uniformly placed networks under weak-wind, near-neutral stability — conditions typical of arid inland regions such as Central Asia — had not been addressed by any single study across all three methodological families, nor compared on a common test problem.

Unlike a literature-only synthesis, this gap was tested directly rather than only asserted: a benchmark comparing four representative methods on a shared synthetic advection-diffusion problem across 5–20 sensors and 1–20% noise (192 trials) confirmed the expected sparsity-driven degradation of every method, showed concretely that a fixed regularization parameter can be counter-productive precisely when data are clean and the network is small, and showed that a naively chosen surrogate architecture fails outright on this problem class rather than merely underperforming. Two open problems, motivating the authors' ongoing doctoral research

within specialty 05.01.07 (Mathematical Modelling, Numerical Methods and Software Systems), follow directly from the combined review and experiment: first, automatic, data-driven selection of the regularization parameter α in (4) under sparse, noisy observations, adapted from generic ill-posed-problem criteria [21] to the specific structure of the sparse-network STE problem; and second, a principled, quantitative criterion for detecting when a machine-learning surrogate is operating outside its training domain and automatically reverting to the full adjoint-based numerical solution.

VI. DATA AND CODE AVAILABILITY

The Python implementation of the four inversion methods evaluated in Section 4 (unregularized least squares, Tikhonov-regularized least squares, Metropolis–Hastings MCMC, and the quadratic-surrogate method), the synthetic benchmark driver, the complete raw trial-level results (768 rows: 4 methods $\times$ 4 sensor counts $\times$ 4 noise levels $\times$ 12 repetitions), and the scripts used to generate Figs. 1–2 are openly available at <https://github.com/furqatovichuz/Source-Term-Estimation-for-Atmospheric-Pollutant-Releases-from-Sparse-Sensor-Networks-Draft-v3> and archived under a permanent identifier at

<https://doi.org/10.5281/zenodo.22720550>. No proprietary, confidential or restricted data were used in this study; the closed-form forward model and all sensor configurations are synthetic and fully specified in Section 4.1, so every reported number in Table 2 and Figs. 1–2 is reproducible from the released code without access to any external dataset.

REFERENCES

1. Hutchinson, M., Oh, H., Chen, W.-H. A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors // *Information Fusion*. — 2017. — Vol. 36. — P. 130–148.
2. Pudykiewicz, J.A. Application of adjoint tracer transport equations for evaluating source parameters // *Atmospheric Environment*. — 1998. — Vol. 32, № 17. — P. 3039–3050.
3. Turbelin, G., Singh, S.K., Issartel, J.-P. Reconstructing source terms from atmospheric concentration measurements: optimality analysis of an inversion technique // *Journal of Advances in Modeling Earth Systems*. — 2014. — Vol. 6, № 4. — P. 1–14.
4. Sharan, M., Singh, S.K., Issartel, J.-P. Inverse modelling for identification of multiple-point releases from atmospheric concentration measurements // *Boundary-Layer Meteorology*. — 2012. — Vol. 143, № 2. — P. 355–378.
5. Lushi, E., Stockie, J.M. An inverse Gaussian plume approach for estimating atmospheric pollutant emissions from multiple point sources // *Atmospheric Environment*. — 2010. — Vol. 44, № 8. — P. 1097–1107.
6. Penenko, V.V. Variational methods of data assimilation and inverse problems for studying the atmosphere, ocean and environment // *Numerical Analysis and Applications*. — 2009. — Vol. 2, № 4. — P. 341–351. — doi:10.1134/s1995423909040065.
7. Delle Monache, L., Lundquist, J.K., Kosović, B. et al. Bayesian inference and Markov chain Monte Carlo sampling to reconstruct a contaminant source on a continental scale // *Journal of Applied Meteorology and Climatology*. — 2008. — Vol. 47, № 10. — P. 2600–2613.
8. Rajaona, H., Septier, F., Armand, P., Delignon, Y., Olry, C., Albergel, A., Moussafir, J. An adaptive Bayesian inference algorithm to estimate the parameters of a hazardous atmospheric release // *Atmospheric Environment*. — 2015. — Vol. 122. — P. 748–762. — doi:10.1016/j.atmosenv.2015.10.026.
9. Albani, R.A.S., Albani, V.V.L., Gomes, L.E.S., Migon, H.S., Silva Neto, A.J. Estimating the number of atmospheric releases and other parameters by Bayesian inference // *Air Quality, Atmosphere & Health*. — 2024. — Vol. 17. — P. 1007–1019. — doi:10.1007/s11869-023-01497-9.
10. Brožová, A., Šmídl, V., Tichý, O., Evangelidou, N. Estimation of spatial-temporal source term of Chernobyl wildfires using deep neural network prior // *Journal of Hazardous Materials*. — 2025. — Vol. 488. — Art. 137510. — doi:10.1016/j.jhazmat.2025.137510.
11. Al Aawar, E., Hammoud, M.A.E.R., Hoteit, I. Two-step AI-aided Bayesian source identification of urban-scale pollution // *Atmospheric Environment*. — 2024. — Vol. 323. — Art. 120388. — doi:10.1016/j.atmosenv.2024.120388.
12. Al Aawar, E., El Mohtar, S., Lakkis, I., Alduwais, A.K., Hoteit, I. Bayesian source identification of urban-scale air pollution from point and field concentration measurements // *Computational Geosciences*. — 2023. — Vol. 27. — P. 605–626. — doi:10.1007/s10596-023-10206-5.
13. Chuprov, I., Derkach, D., Efremenko, D., Kychkin, A. Application of physics-informed neural networks for solving the inverse advection-diffusion problem to localize pollution sources // *arXiv:2503.18849*. — 2025.
14. Jia, H., Kikumoto, H. Sensor configuration optimization based on the entropy of adjoint concentration distribution for stochastic source term estimation in urban environment // *Sustainable Cities and Society*. — 2022. — Vol. 79. — Art. 103726. — doi:10.1016/j.scs.2022.103726.

15. Wang, F., Zhou, X., Kikumoto, H., Okaze, T. Detector configuration optimization based on wind tunnel tests using normalized adjoint concentration gradient for urban spatial source parameters estimation // *Building and Environment*. — 2024. — Vol. 248. — Art. 111094. — doi:10.1016/j.buildenv.2023.111094.
16. Tian, H., Lang, Z., Cao, C., Wang, B. Optimizing sensor placement for enhanced source term estimation in chemical plants // *Processes*. — 2025. — Vol. 13, № 3. — Art. 825. — doi:10.3390/pr13030825.
17. Gkirmpas, P., Barmpas, F., Tsegas, G., Efthimiou, G., Tremper, P., Riedel, T., Vlachokostas, C., Moussiopoulos, N. An evaluation of the sensitivity of a source term estimation methodology of sensor configuration in an urban-like environment // *Atmosphere*. — 2024. — Vol. 15, № 12. — Art. 1512. — doi:10.3390/atmos15121512.
18. Albani, R.A.S., Albani, V.V.L. Tikhonov-type regularization and the finite element method applied to point source estimation in the atmosphere // *Atmospheric Environment*. — 2019. — Vol. 211. — P. 69–78. — doi:10.1016/j.atmosenv.2019.04.063.
19. Tang, S., Li, F., Han, Y., Feng, Z. A lightweight adjoint method for gaseous pollution source term estimation in urban environments // *Building Simulation*. — 2025. — Vol. 18, № 4. — P. 937–955. — doi:10.1007/s12273-025-1244-8.
20. Lin, Y., Huang, H., Wang, J., Zhang, X. Sensor configuration optimization for source term estimation of time-varying emissions // *Sustainable Cities and Society*. — 2026. — Vol. 139. — Art. 107186. — doi:10.1016/j.scs.2026.107186.
21. Sanders, T., Platte, R.B., Skeel, R.D. Effective new methods for automated parameter selection in regularized inverse problems // *Applied Numerical Mathematics*. — 2020. — Vol. 152. — P. 29–48. — doi:10.1016/j.apnum.2020.01.015.
22. Barati Moghaddam, M., Mazaheri, M., Mohammad Vali Samani, J. Inverse modeling of contaminant transport for pollution source identification in surface and groundwaters: a review // *Groundwater for Sustainable Development*. — 2021. — Vol. 15. — Art. 100651. — doi:10.1016/j.gsd.2021.100651.
23. Kontos, Y.N., Kassandra, T., Perifanos, K., Karampasis, M., Katsifarakis, K.L., Karatzas, K. Machine learning for groundwater pollution source identification and monitoring network optimization // *Neural Computing and Applications*. — 2022. — Vol. 34. — P. 19515–19545. — doi:10.1007/s00521-022-07507-8.
24. Ravshanov, N., Shafiev, T.R. Nonlinear mathematical model for monitoring and predicting the process of transfer and diffusion of fine-dispersed aerosol particles in the atmosphere // *Journal of Physics: Conference Series*. — 2019. — Vol. 1260. — Art. 102013. — doi:10.1088/1742-6596/1260/10/102013.
25. Dun, A., Yang, Y., Lei, F. Dynamic graph convolution neural network based on spatial-temporal correlation for air quality prediction // *Ecological Informatics*. — 2022. — Vol. 70. — Art. 101736. — doi:10.1016/j.ecoinf.2022.101736.