Supplementary Materials

InverPINN: Diagnosing Source-Strength Bias in Physics-Informed Neural Networks for Sparse Pollution Source Inversion

Khadis Aidyn

Scope and companion data

This supplement accompanies the final controlled synthetic study. J1 recovered **23/30** scenarios, below the preregistered requirement of **27/30**. Final confirmation **FAILED**. The classical comparator recovered **30/30**. No failed case or outlier has been removed.

S1. Full distributions and paired effects

The 90-row table retains every final scenario for every method. Distribution summaries report mean, median, sample SD, quartiles, 90th percentile and worst value. Missing classical continuous-PDE diagnostics remain missing; a finite-difference field must not be assigned an autograd residual of zero.

The paired table reports J1 minus each control. Negative error differences favor J1. The report-only bootstrap uses 10,000 paired resamples with seed 2009202818 and the same resampled indices for all metrics. These effects and intervals are independently reproduced from raw rows; the original summary is preserved byte-for-byte. This is analysis reproduction, not a new hypothesis test, estimator or selection decision. Ties use the frozen tolerance.

Companion CSV and units

The separately supplied **InverPINN_Final_Benchmark_Results.csv** is a byte-identical copy of the canonical 90-row results table. Each row is one method/scenario pair. All 30 scenario IDs appear once for each of B_revised, J1 and classical. The primary optimization seed is 42.

In the CSV, relative_strength_error, signed_relative_Q_error, relative_l2 and reference_error_estimate are fractions, not percentages. Multiply by 100 to express them as percent. Locations and localization errors use the synthetic unit-square domain, not geographic coordinates or kilometres. Q is peak Gaussian source intensity. Concentration RMSE/MAE and PDE residuals use the synthetic units defined in the manuscript. runtime_seconds is in seconds.

Blank classical PDE/BC/IC diagnostic fields mean not evaluated, not zero. failure, training/evaluation status, recovery_success, numerical-reference flags and the observation/reference hashes are retained. Joint recovery requires localization error at most 0.08 **and** relative Q error at most 0.20.

S1.1. Primary error distributions

Every row below summarizes 30 scenarios. SD is the sample standard deviation. Q25, Q75 and Q90 are the 25th, 75th and 90th percentiles. Worst is the maximum error. Display rounding does not change the companion CSV.

Table S1a. Localization error (synthetic domain units)

Method	Mean	Median	SD	Q25	Q75	Q90	Worst
B_revised	0.019853	0.009186	0.024431	0.000985	0.034711	0.056342	0.082475
J1	0.017605	0.005574	0.023457	0.001410	0.025877	0.048607	0.084081
classical	0.001438	0.001111	0.000788	0.001001	0.001542	0.002687	0.003642

Table S1b. Relative Q error (%)

Method	Mean	Median	SD	Q25	Q75	Q90	Worst
B_revised	16.193	9.319	18.908	0.807	25.073	41.669	64.029
J1	13.101	5.975	16.902	0.811	17.507	34.783	62.183
classical	1.464	1.218	0.964	0.902	1.611	3.345	3.537

Table S1c. Concentration relative L2 error (%)

Method	Mean	Median	SD	Q25	Q75	Q90	Worst
B_revised	19.458	11.048	20.803	2.075	36.587	50.675	69.362
J1	16.660	7.423	19.263	2.333	27.219	47.535	69.724
classical	1.905	1.507	1.101	1.083	2.070	3.686	4.474

For Tables S1b and S1c, the SD is in percentage points. Percent conversion is applied only to these displayed tables. Full-precision values are available in the linked distribution summaries.

S1.2. Concentration, physics and runtime distributions

Table S2a. Concentration RMSE (synthetic concentration units)

Method	Mean	Median	SD	Q25	Q75	Q90	Worst
B_revised	0.009842	0.006259	0.010837	0.001084	0.018352	0.024412	0.040785
J1	0.008398	0.003882	0.009985	0.001115	0.015564	0.021172	0.040998
classical	0.000965	0.000815	0.000599	0.000495	0.001160	0.001992	0.002413

Table S2b. Concentration MAE (synthetic concentration units)

Method	Mean	Median	SD	Q25	Q75	Q90	Worst
B_revised	0.003068	0.001976	0.003112	0.000561	0.005692	0.007424	0.011518
J1	0.002654	0.001370	0.002867	0.000511	0.004845	0.006322	0.011560
classical	0.000307	0.000257	0.000151	0.000196	0.000354	0.000519	0.000703

Table S2c. Independent PDE residual RMSE (concentration/time)

Method	Mean	Median	SD	Q25	Q75	Q90	Worst
B_revised	0.006721	0.005562	0.003650	0.004458	0.006676	0.011845	0.017761
J1	0.006197	0.005361	0.003287	0.004244	0.006692	0.010084	0.016032
classical	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Table S2d. Per-case fitting runtime (seconds)

Method	Mean	Median	SD	Q25	Q75	Q90	Worst
B_revised	91.31	91.41	1.46	90.44	92.18	93.06	94.09
J1	93.41	93.45	3.98	91.78	95.62	97.34	102.45
classical	64.92	64.57	6.06	61.14	68.15	74.47	76.24

Classical continuous-PDE diagnostics were not evaluated. Its reusable response-bank setup is excluded from the per-case fitting times above and must be counted separately in any end-to-end runtime comparison. These tables do not establish hardware-independent speed.

S1.3. Paired effects on identical observations

Differences are J1 minus control. Negative error differences favor J1. Intervals are the existing report-only 95% paired bootstrap intervals for the mean difference, using 10,000 paired resamples and seed 2009202818. They are not new model-selection tests. Q and L2 differences below remain **fractions**, not percentage points.

Table S3a. J1 compared with B_revised

Metric	Mean delta	95% interval	J1 better	Control better	Tied
Localization	-0.00225	[-0.00398, -0.00077]	18	12	0
Relative Q	-0.03093	[-0.05093, -0.01302]	21	9	0
Relative L2	-0.02798	[-0.04651, -0.01199]	21	9	0
RMSE	-0.00144	[-0.00240, -0.00063]	21	9	0
MAE	-0.00041	[-0.00070, -0.00018]	22	8	0
PDE RMSE	-0.00052	[-0.00124, 0.00002]	20	10	0

Table S3b. J1 compared with classical

Metric	Mean delta	95% interval	J1 better	Control better	Tied
Localization	0.01617	[0.00862, 0.02495]	6	24	0
Relative Q	0.11636	[0.06222, 0.17873]	11	19	0
Relative L2	0.14755	[0.08558, 0.21751]	2	28	0
RMSE	0.00743	[0.00429, 0.01117]	2	28	0
MAE	0.00235	[0.00143, 0.00342]	0	30	0

The frozen metric tie tolerances are absolute 1e-8 and relative 1e-6. Classical PDE comparisons are absent because no comparable continuous-PDE diagnostic was evaluated. Full-precision intervals, paired medians and fitting-time comparisons remain available in the paired table.

S1.4. Supplementary physical and signal diagnostics

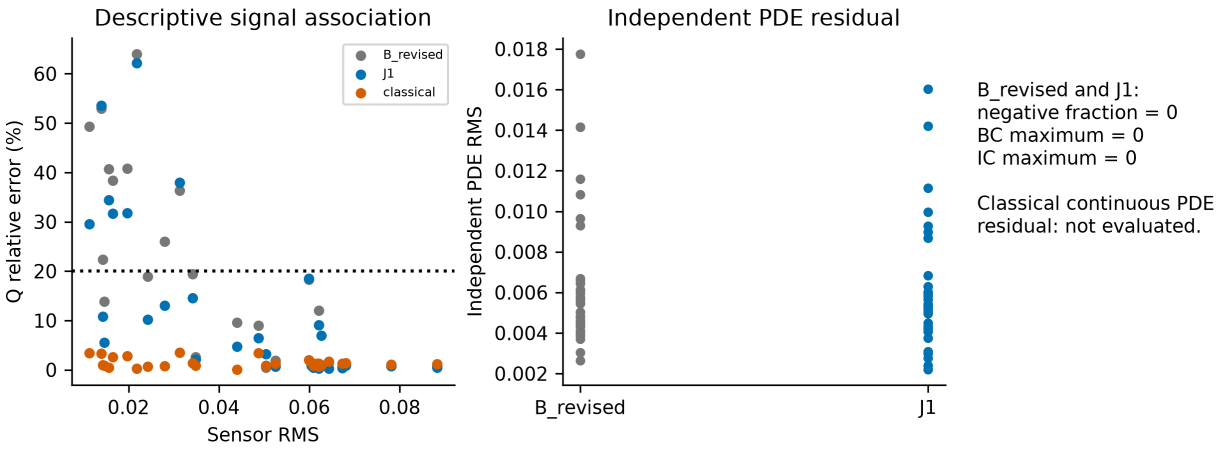


Figure S1. All final cases are retained in the sensor-signal scatter. The relation is descriptive, not a randomized causal experiment. Independent neural PDE diagnostics are shown separately from hard-constraint values. Exact BC/IC and nonnegativity do not certify accurate source recovery.

The figure reproduces the original scientific plot without changing its plotted values. The sensor-signal grouping is descriptive. Neither this figure nor the reported PDE residuals establish robustness to noise, unknown meteorology, multiple sources or real-world pollution observations.

Interpretation of physical diagnostics

The constrained neural fields retain exact nonnegativity, zero initial concentration and zero prescribed boundary concentration. These properties hold by construction and do not guarantee correct source recovery. Independent neural PDE residuals are evaluated separately from fitting losses. No autograd residual is fabricated for the classical finite-difference solution.

S2. Historical evidence is not another final test set

Experiment	Bundled evidence	Scientific role
Numerical validation	Convergence CSV, summary	Numerical verification and estimated accuracy
Original B0	Initial summary	Development and then-consumed held-out evidence
Physical correction	Independent comparison summary	B0 versus B_revised; not pooled with final cases
Observability	Diagnostic summary, conditional Q, resolution study	Oracle and classical diagnostics, not final validation
Formulation intervention	Candidate table, selection record	24-case development only; no B_revised_2 promotion
Final confirmation	Raw J1, B_revised, classical, failed decision	30 new cases; fixed recipes and failed final gate

Original detailed reports remain available: numerical validation, observability diagnosis, source/field derivation, development intervention, and final confirmation. These are historical records, not instructions to resume model development. Large historical diagnostic plots remain in their original local result directories; no unexecuted sweep is represented as evidence.

S3. Field display selection

Before inspecting the display images, reporting selected the median localization rank among J1 joint recoveries (ascending localization, then ID; zero-based index floor(n/2)), giving **final_blind_v1_003**. Failure selection is maximum localization, then ascending ID, giving **final_blind_v1_030**. All four saved 161² preview frames are copied unchanged; main figures show the final time 0.8. This is an explicit reporting selection, not an originally preregistered inferential endpoint. Full 641² archived metrics, not downsampled display values, determine rankings and captions.

S4. Numerical and statistical cautions

Final cases 016,017,027 have estimated reference errors 1.063%,1.184%,1.126%. All are retained in scores. The worst three localization cases (030,008,024) met the numerical target. Rejected hypotheses and failed recoveries are evidence, not missing data. The current report draws no noise, changed-wind, sensor-count, multiple-source or real-data performance conclusions from the existence of older software modules.

S5. Bibliographic verification

References 2-5 were checked against their authors' arXiv records. Publisher-deposited title/author/date/abstract metadata for Golub-Pereyra were verified through Crossref's DOI record; the publisher page was access-blocked. Almaty reference 1 title/authors/pages/DOI were checked against its publisher-deposited record; the full publisher page was access-blocked. It is used only for general motivating context, not quantitative pollution claims, licensing assumptions, or emitter attribution. No measurements were downloaded or scraped.

Evidence and reproducibility

Linked repository evidence in this PDF is pinned to the reporting commit. The reproducibility guide explains analysis reproduction and the limits of the compact evidence bundle. The bundle does not contain the complete training/reference archive and is not a deposited complete experiment archive. No public data DOI or real Almaty validation is claimed.