Making the most from abrupt and sparse data: Data driven interpolation from sporadic sensors of contaminant concentration with extreme spatial gradients in saturated porous media

Alon Feldman,[†] Shai Kendler,[‡] Barak Fishbain,[*,‡] and Yaniv Edery[‡]

[†]*Faculty of Mathematics, The Technion - Israel Institute of Technology*

[‡]*Faculty of Civil and Environmental Engineering, The Technion - Israel Institute of Technology*

E-mail: fishbain@technion.ac.il

Phone: +972 (0)4 8293177

Abstract

Soil and aquifer contamination pose a persistent environmental challenge; however, contamination monitoring is severely limited by the high cost, destructive nature, and logistical constraints associated with sampling. Constructing dense contamination maps from sparse measurements is, therefore, a critical yet unresolved task. This study presents a data-driven interpolation method that successfully adapts a artificial intelligence, encoder-decoder framework from air pollution to the domain of soil contamination. This effective implementation underscores the framework's adaptability and potential as a unifying approach to interpolation. Interpolation is cast here as a signal reconstruction problem, where sparse soil measurements act as the encoded input and a trained decoder reconstructs the entire contamination field. A synthetic

dataset was generated using a contaminant transport simulator in a porous medium, utilizing a particle tracker that follows the simulated flow and stochastic diffusion. In this way, the decoder statistically emulates a physical process, learning the dependencies between sparse sampling points and the underlying concentration field. Results demonstrate that the framework successfully reproduces high-resolution contamination maps and outperforms classical interpolation techniques, such as Kriging and Inverse Distance Weighting. This preliminary study adopts a favorable scenario—with high sensor density and known conductivity—to demonstrate the approach's feasibility and robustness, supporting its future application in more realistic and challenging settings.

Introduction

Monitoring the transport of contaminants in soils and aquifers presents a persistent challenge in hydrology and environmental science. Accurate characterization of chemical migration through porous media—including soils, aquifers, and fractured rock formations—is essential for predicting reactive transport processes and guiding remediation efforts.[1–6] Subsurface heterogeneity plays a central role in shaping contaminant behavior, giving rise to preferential flow pathways that result in non-Fickian dispersion, diffusion, and heavy-tailed breakthrough curves.[7–12]

These preferential pathways channel contaminants through narrow, tortuous routes, producing highly heterogeneous spatial concentration distributions. Such distributions often include under-sampled regions, complicating efforts to monitor and assess contamination accurately.[13–15] Sampling networks, typically constrained by economic, logistical, and accessibility limitations, may fail to capture the full extent of contaminant spread.[16–18] Estimating contaminant distributions from discrete sensor networks is therefore inherently challenged by subsurface complexity and sensor placement limitations. This highlights the need for more effective and adaptive monitoring strategies.[19,20]

Traditional spatial interpolation methods often struggle to resolve the intricate features

of contaminant plumes in heterogeneous media. Recent advances have introduced data assimilation techniques that integrate observational data with physical transport models, significantly improving plume estimation accuracy.[21,22] In parallel, physics-informed machine learning approaches have emerged, embedding governing physical equations within data-driven frameworks to enhance predictive fidelity and enable robust reconstruction of contaminant distributions from sparse datasets.[23,24] Machine learning also facilitates automated spatial interpolation and optimal sensor placement, offering strategies to balance resource constraints with information gain.[25–27] These approaches address the limitations of sparse sensor networks in complex subsurface environments, reducing uncertainty in contamination mapping and improving monitoring efficiency.

Despite these technological advancements, significant challenges remain. Subsurface heterogeneity, dynamic flow paths, and limited sensor density continue to hinder comprehensive contaminant assessment. To overcome these barriers, adaptive monitoring protocols that integrate iterative interpolation, sensor layout optimization, and high-resolution data acquisition are essential. Such strategies can support more informed environmental management and decision-making.[28,29]

To this end, this paper presents the application of the Ridiculously simple interpolation method (RSIM) as an interpolation mechanism for saturated porous media. RSIM was originally developed for air pollution interpolation by Feldman et. al.[30] RSIM is a Machine Learning (ML) approach that uses simulated data, so the machine follows the physical nature of the given environment and capture statistical dependencies between limited sensor data and the whole contamination field. Unlike air pollution, which typically exhibits smooth and continuous dispersion in space and time, soil contamination often displays irregular and highly heterogeneous spreading patterns, driven by porous-media heterogeneity, nonuniform pollutant infiltration, and slow transport within the pore structure. The successful implementation of RSIM in saturated porous media highlights the method's ability to reconstruct fields without prior assumptions. Neither mathematical properties of the field—such as

boundary conditions, gradient bounds, or function type—nor physical information such as the conductivity field are required as inputs to the model. Instead, the statistical inference approach enables the model to mimic the physical behavior generated by the physics-based simulator. These characteristics emphasize the need for monitoring and modeling strategies tailored to the unique complexities of subsurface environments.

Methods

Problem Formulation

The interpolation of soil contamination is formulated as a signal reconstruction problem, using the Ridiculously Simple Interpolation Method (RSIM)[30]. To this end, let Ω denote the domain of interest, representing the soil or groundwater system under investigation, and let $\mathcal{F}$ be the set of scalar fields that describe pollutant concentrations over Ω:

$$\forall f \in \mathcal{F}; f : \Omega \to \mathbb{R}$$

A set of sparse measurement locations $\Omega_s \subset \Omega$ is considered, representing borehole samples, soil probes, or water-table extraction points, with the number of sampling points $s = |\Omega_s|$. The encoding function is then defined as:

$$\forall \omega \in \Omega_s \; ; \; z_\omega = e(f(\omega)), \tag{1}$$

where z_ω are the measured concentrations. To mimic the uncertainty inherent in soil sampling, similarly to Feldman et. al[30], measurement noise was introduced through a Gaussian perturbation, reflecting both laboratory errors and small-scale heterogeneity:

$$z_\omega = f(\omega) \cdot (1 + \eta_\omega), \quad \eta_\omega \sim \mathcal{N}(0, \epsilon^2). \tag{2}$$

Then, the reconstruction function (decoder) that aim at estimating the entire soil contamination field is given by:

$$d : \mathbb{R}^s \to \mathcal{F} \; ; \quad \hat{f} = d(\vec{z}), \tag{3}$$

where $\vec{z}$ denotes the vector of sensor readings. The optimal decoder d^* minimizes the expected loss across the probability distribution of pollutant configurations:

$$d^* = \arg\min_{d \in D} \; \mathbb{E}_{f \sim P_{\mathcal{F}}} \left[l(f, \hat{f}) \right] \tag{4}$$

with the loss defined as the mean squared error. The distribution $P_{\mathcal{F}}$ was approximated using a large set of simulated contamination scenarios generated with a soil transport simulator, incorporating infiltration, adsorption, groundwater flow, and boundary conditions.

System Overview

The interpolation task was implemented as an encoder–decoder pipeline, following the formulation presented in Section 2.1. The sparse soil measurements constitute the encoded representation, whereas the decoder reconstructs the full concentration field. The framework itself is model-agnostic: any predictive model capable of learning a mapping from sparse inputs to dense outputs may be used as the decoder.

In the present study, a linear regression decoder was selected due to its transparency, robustness, and computational efficiency. However, the training procedure—learning statistical dependencies between sparse samples and the underlying contamination field—remains identical for any decoder class. In this sense, the framework naturally accommodates more expressive models such as artificial neural networks, which were part of the original conceptualization of the method. Future work may therefore employ deep learning–based decoders to enhance performance while preserving the same statistical structure.

Importantly, the hydraulic-conductivity fields used in this work are *not* provided as input to the interpolation model and do not play any role during training or inference. Their

sole purpose is to generate physically plausible contaminant plumes under heterogeneous subsurface conditions. Thus, any criticism regarding the synthetic nature of the conductivity fields pertains only to the data-generation process, not to the interpolation model itself, which operates exclusively on the contaminant concentration fields.

Synthetic Case Studies

Synthetic soil contamination scenarios were generated to reflect typical pollution in saturated porous media. Each simulation generated dense pollution fields over a discrete soil grid and time interval. In total, m realizations (in our case, $m = 2,170$) of pollution maps were generated, forming the database:

$$\{(\vec{z}_j, f_j)\}_{j=1}^{m}, \tag{5}$$

where $\vec{z}_j$ denotes the sensor vector for the j-th simulation, and f_j denotes the corresponding ground truth dense contamination map. The optimization objective is then:

$$d^* = \arg\min_{d \in D} \frac{1}{m} \sum_{j=1}^{m} \|f_j - d(\vec{z}_j)\|^2. \tag{6}$$

Flow and transport simulation

We construct the flow and transport field layout using a set of 2D numerical simulations, where a second-order stationary random domain of hydraulic conductivities is distributed according to a lognormal distribution with a mean of $ln(\hat{K}) \approx 0$ and a heterogeneity level characteristic of soil and rock formation marked by a variance of $\sigma^2 = 3$, established by a sequential Gaussian simulator (GCOSIM3D).[31] This conductivity domain is composed of a mesh of 120×300 conductivity bins (each of size $\Delta = 0.2[m]$), forming a domain of $24 \times 60m^2$. Each domain is produced by a statistically homogeneous and isotropic Gaussian distribution in $ln(K)$, with a dimensionless correlation length $l_c/L = 0.016$, where L is the domain length along the main flow direction and $l_c = 1[m]$ corresponds to a dimensionless value of

$\Delta/l_c = 0.2$, which provides an accurate description of the velocity distribution generated by the $ln(K)$ domain and advective transport.[32,33] The statistical significance of this model was verified in previous studies where the relation between the heterogeneity and anomalous nature of the transport,[34] the preferential flow characteristics,[13] and the Shannon entropy,[35] as well as the correlation length effect on transport[36] have been investigated.

Two deterministic pressure drops were used in the simulation, translated to a total head drop of $H = 10$ and $H = 100$. Both drops are imposed from the inlet (top) to the outlet (bottom), and a finite element numerical model with a Galerkin weighting function was used to calculate the local head drop in 2D for each bin.[37] The local head provided the streamlines and local velocities, as calculated using the local conductivity and porosity ($\theta = 0.3$). At $t = 0$, three particle pulse types were introduced at the inlet: 1. a Dirac delta function comprising 10^5 particles. 2. a uniform pulse spanning four time extents ($\Delta T = 1, 5, 10, 50[sec]$). 3. a Gaussian pulse spanning four standard deviations ($\sigma_T = 1, 5, 10, 50[sec]$). All pulses type advance according to the local advection and diffusion term, following the Langevin equation:

$$\lambda_n = v[x_n(t)]\delta t + \lambda_{\Lambda_n} \tag{7}$$

where λ_n is the displacement (bold marks a vector), $\mathbf{x}_n(t)$ is the known location of the n particle at time t, v the fluid velocity at that location (v is the modulus of $|v|$), $\delta t = \delta s/v$ the temporal displacement magnitude calculated from the fixed displacement size δ_s, and the diffusive displacement for particle n calculated from the diffusion of ions in water ($\Lambda_m = 10^{-9}[\frac{m^2}{sec}]$).[38] The local fluid velocity field is obtained as $v = q(x)/\theta$, where $q(x)$ is the local Darcy flux:

$$\nabla \cdot q(x) = 0; \quad q(x) = -K(x) \cdot \nabla h(x) \tag{8}$$

The displacement size $\delta_s$is selected to be an order of magnitude less than Δ to interpolate the velocity within each bin correctly. The Particle Track (PT) simulation employed reflective boundary conditions, similar to a Dirichlet boundary condition, over the y-axis boundaries,

and constant pressure boundary conditions at the inlet and outlet, similar to the Neumann boundary condition, over the x-axis. This displacement size robustness, as well as the statistical convergence and numerical stability, and particle volume, was reported in previous studies for similar conditions.[13,34,35] The diffusive displacement, d_{D_n}, is randomly generated for the x and y dimensions, whose entries are mutually independent and sampled from a Normal distribution between 0 and 1 ($\mathbf{N}[0,1]$) multiplied by the square root of the diffusion coefficient and the temporal displacement magnitude, as depicted in the following equation:

$$\lambda_{\Lambda_n} = \mathbf{N}[0,1]\sqrt{(2\Lambda_m \delta t)} \tag{9}$$

The same GCOSIM3D model proved valuable in reproducing and analyzing field data,[39,40] uncertainty,[41–44] and upscaling,[45] while the PT methodology proved to be very robust and appropriate in this modeling configuration.[46,47]

Figure 1 shows eight examples of contamination patterns generated by the porous-medium flow simulator. For visual clarity, values are presented in this paper on a $\log_{10}$ scale with colors ranging from dark blue for near zero values (less than 10 parts per bin), through light-blue, green, yellow, orange, and red for higher concentrations. Each image represents a different synthetic scenario with varying shapes and intensities of the contaminant plume. It is important to clarify that all scenarios are varying in pollution release speed and head difference. The montage illustrates the diversity of spatial patterns used for training and testing the reconstruction model. In all concentration maps presented in this paper, the porous medium is identical, the head gradient is imposed from top to bottom, and accordingly, the flow direction is oriented downward. An intriguing feature can be observed in the case of soil contamination: rather than dispersing uniformly throughout the domain, the contaminant streams tend to merge and coalesce as they migrate. Such behavior is non-trivial in the context of environmental pollution processes more broadly.[48]

Each image represents a different synthetic scenario with varying plume shapes and

intensities, driven by differences in pollution-release rate and hydraulic head. Because these scenarios span different concentration ranges, each map is independently normalized for visualization. Consequently, no unified colorbar is shown: applying a single scale would misleadingly imply comparability of absolute concentrations across simulations, whereas the purpose of the montage is purely to illustrate morphological diversity.

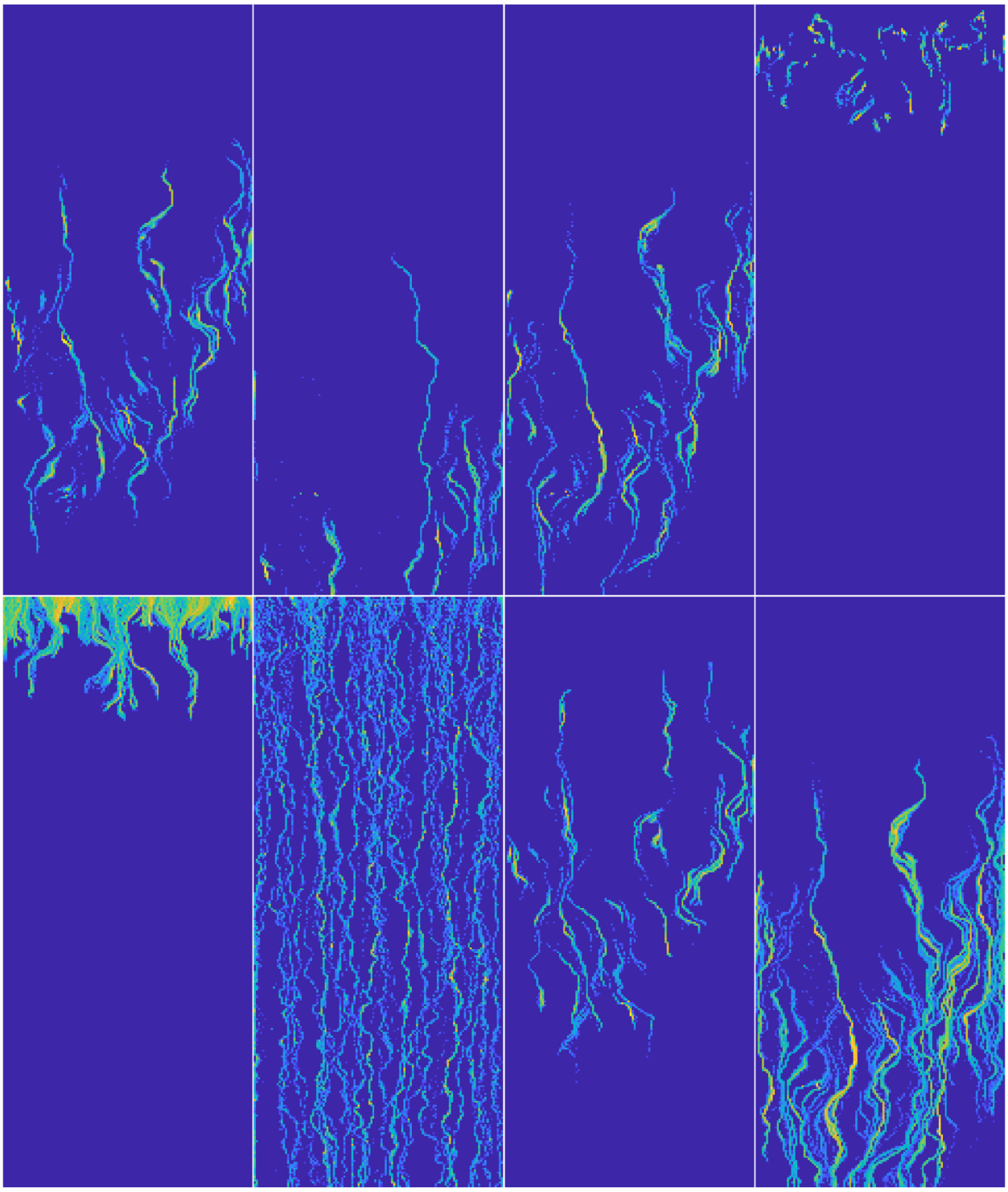


Figure 1: Eight synthetic contamination scenes are arranged in a 2x4 grid. The examples show different plume shapes and intensities generated by the porous-medium simulator.

Model Specification

A linear regression model was selected for the decoder, following its simplicity, explainability, and performance. For each spatial location $\omega \in \Omega$:

$$\hat{f}(\omega) = \langle W_\omega, \vec{z} \rangle + b_\omega, \tag{10}$$

where W_ω is a weight vector estimated from the training data, and b_ω a bias term. This formulation resembles both Inverse Distance Weighting and Kriging approaches, but in the present framework, the weights are directly learned from simulated data rather than imposed through deterministic assumptions. While a linear model is used here, the framework is equally compatible with non-linear models, including deep neural networks as potential decoders.

Sensor Placement Strategy

Sensor locations were selected according to a local Shannon-entropy metric derived from Mano et al.[49]. For each spatial location $\omega \in \Omega$, an empirical distribution of contaminant concentrations was obtained from all simulated realizations, and the corresponding local Shannon entropy was computed. This entropy quantifies the variability of the concentration field at each location; regions that exhibit greater variability across the ensemble are expected to yield higher information gain when sampled.

Maximizing local Shannon entropy effectively aligns with maximizing information gain. With these properties of the transport, and given the practical minimum-distance (boxing-out) constraint imposed on borehole placement, high-entropy selection provides a principled and tractable strategy for placing informative sampling points.

Sensors were therefore placed sequentially at locations with the highest entropy values, subject to minimum horizontal and vertical spacing constraints that reflect realistic drilling limitations. This approach ensures that sampling focuses on regions where contaminant

concentrations are most variable across realizations, while avoiding redundant measurements in low-entropy zones that contribute little to reconstruction accuracy.

The resulting sensor configuration is shown in Figure 2. The background depicts the entropy map derived from the simulated ensemble of the transport, where brighter regions indicate higher variability. The white circles mark the selected sensor positions. As shown, the chosen locations align with high-entropy regions while satisfying the required spacing of 20 grid cells in both horizontal and vertical directions. The entropy field highlights patterns of contaminant migration under the imposed flow field, as the plumes move from top to bottom and tend to converge along preferential pathways, creating localized regions of elevated variability that guide the resulting sampling layout.

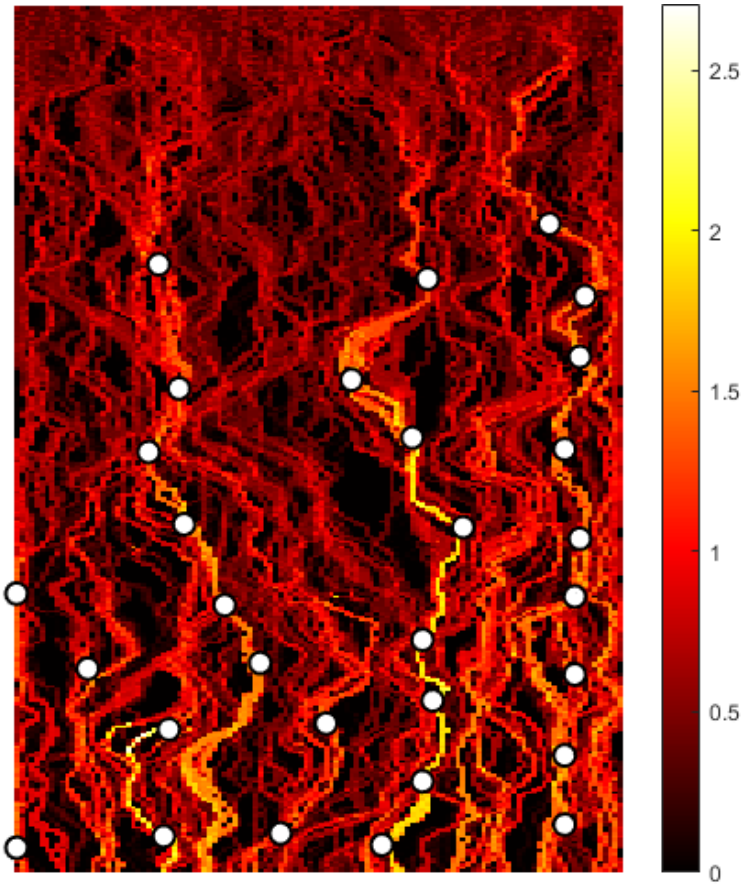


Figure 2: Entropy map of the simulated domain with high entropy locations for sensor placement.

Evaluation Metrics

Model performance was assessed using both qualitative and quantitative measures. Visual inspection of reconstructed contamination maps provided insights into the preservation of sharp discontinuities induced by soil heterogeneity. Quantitative benchmarks included the Mean Root Mean Square Error (MRMSE) and the Mean Correlation Coefficient (MCorr) between predicted and true concentration fields:

$$\text{MRMSE} = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{|\Omega|}\sum_{\omega\in\Omega}(\hat{f}_j(\omega) - f_j(\omega))^2}, \tag{11}$$

$$\text{MCorr} = \frac{1}{m}\sum_{j=1}^{m}\text{corr}(\hat{f}_j, f_j). \tag{12}$$

These metrics allowed for direct comparison with alternative interpolation methods.

Classical Interpolation Methods

For benchmarking, several standard interpolation techniques were implemented. These methods represent widely used baselines for reconstructing spatial fields from sparse samples:

- **Linear Interpolation** — piecewise interpolation assuming locally linear variation.
- **Natural (Sibson) interpolation** - Sibson interpolation computes the value at a query point as a weighted average of surrounding data points, where the weights are derived from Voronoi tessellation geometry.
- **Cubic Interpolation** — smooth interpolation using cubic polynomials with continuous derivatives.
- **Nearest Neighbor** — assigns each unsampled location the value of the closest measurement.

- **Inverse Distance Weighting (IDW)** — weighted average with weights proportional to inverse distance.
- **Ordinary Kriging** — geostatistical interpolation based on a fitted variogram and second-order stationarity.
- **Universal Kriging** — extends Ordinary Kriging by modeling deterministic spatial trends.

A comprehensive overview of these interpolation families can be found in Cressie [50]. These classical methods were used solely as benchmark baselines to evaluate the performance of the proposed data-driven interpolation framework.

Computation Setup

In practice, all computations were conducted using synthetic data generated from the porous medium simulator, with sensor placement determined by maximum local entropy under minimum-distance constraints. Unless otherwise stated, the reconstruction model was trained with 30 sensors, a multiplicative Gaussian noise factor of 10%, and a minimum horizontal and vertical spacing of 20 grid cells between sensors. The simulation grid size was fixed at 300×120 cells. The data were split into training, validation, and test sets in proportions of 0.7/0.1/0.2, respectively. Training was performed with a mini-batch size of 128, a maximum of 300 epochs, and early stopping after 50 validation checks, using an initial learning rate of 10^{-1} with scheduled reductions during training.

Results

Reconstruction from sparse measurements

Figure 3 presents eight representative reconstruction cases from the test set, arranged in a 3×2 grid. Each triplet illustrates the input sensor vector (left), the predicted contamination

field (middle), and the corresponding ground truth field (right). Across all cases, the model accurately reconstructs dominant plume structures and preserves sharp gradients caused by soil heterogeneity, even with only sparse sensor data. The prediction panels closely resemble the ground truth fields, with high spatial correlation and low reconstruction error. The comparison further highlights that the model successfully interpolates fine-scale patterns that classical smooth interpolation techniques would otherwise miss.

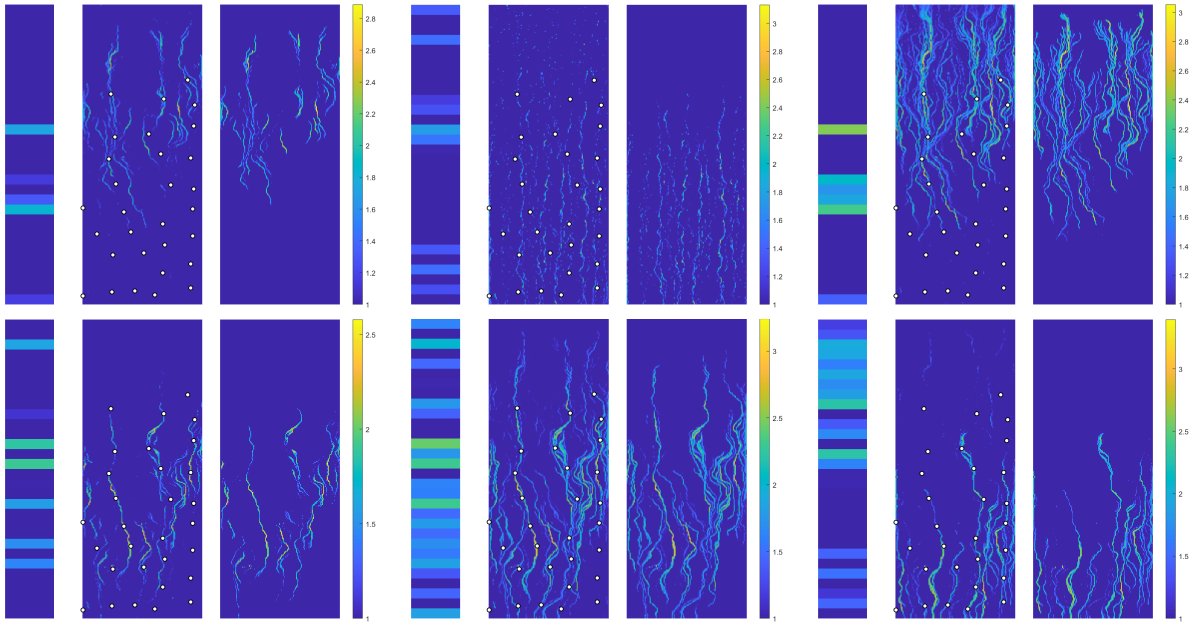


Figure 3: Representative reconstruction results from the test set. Each triplet shows, from left to right: the vector of input sensor readings, the reconstructed contamination field by the proposed encoder–decoder model (including sensor positions), and the corresponding ground truth contamination field.

Comparison to Classical Interpolation

To highlight the advantages of the proposed framework, the reconstructions were compared to standard interpolation techniques: Linear, Natural , Nearest Neighbor, cubic, Inverse Distance Weighting (IDW), Universal Kriging, and Ordinary Kriging methods. Figure 4 Demonstrates a sample from contamination event in the porous media in it's peak. The wave of contamination has activated all sensors in the field of interests and a rich reconstruction is shown in a logarithmic scale. As shown in Figure 4, these methods were unable

to reproduce the steep gradients present in the contamination field, instead producing overly smoothed maps with loss of fine-scale structures. By contrast, the proposed encoder–decoder framework effectively mimicked the statistical patterns embedded in the training simulations, resulting in reconstructions that align more closely with the ground truth.

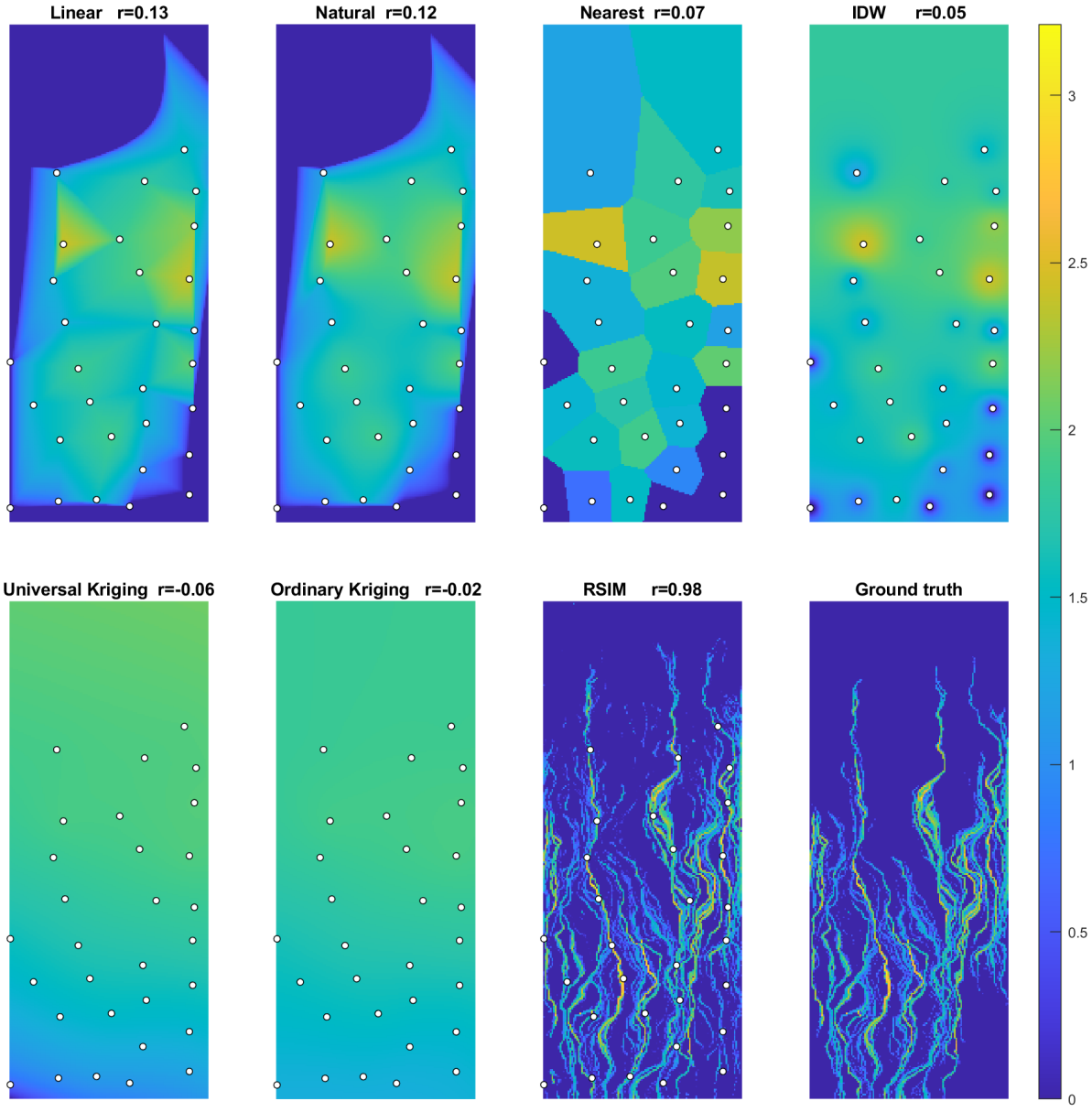


Figure 4: Description of the figure goes here.

Table 1 summarizes the quantitative evaluation of the different methods using MCorr and MRMSE. The proposed model, trained with 30 sensors and 10% noise, achieved the highest

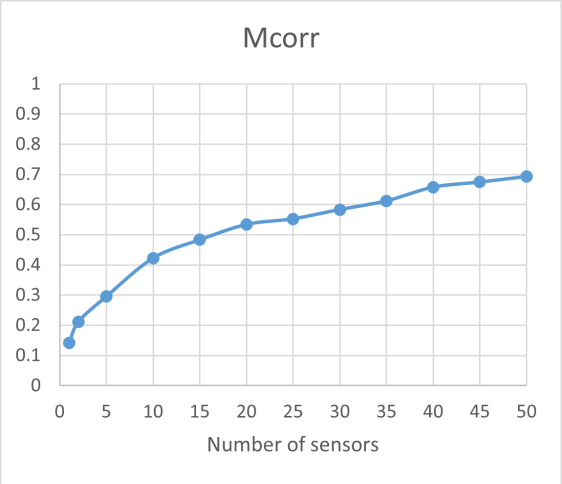
correlation and lowest error, consistent with its ability to exploit the statistical variability of the simulated data. This highlights the advantage of a data-driven approach that learns from representative simulations rather than relying on deterministic smoothness assumptions.

Table 1: Comparison of interpolation methods on the test set.

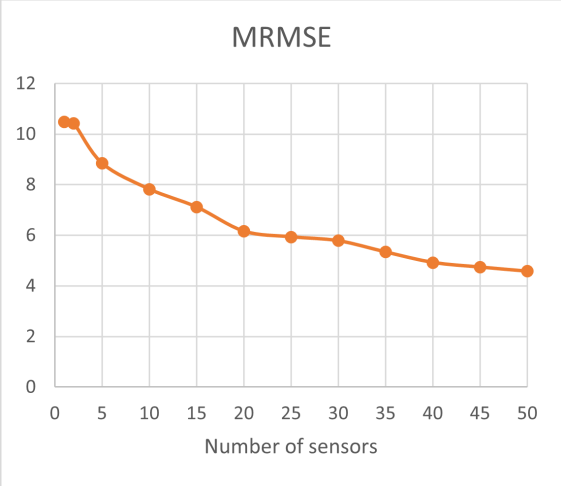
Method	**MRMSE**	**MCorr**
Linear	19.779	0.10569
Natural	19.764	0.10649
Nearest	16.686	0.10414
IDW	12.979	0.1196
Universal Kriging	14.428	0.093224
Ordinary Kriging	13.19	0.088727
RSIM	**5.39**	**0.72228**

Sensitivity Analysis

To further assess performance, a sensitivity analysis was conducted by varying the number of sensors used as input. In this analysis, the noise factor was set to **0%** to isolate the effect of sensor number. Figure 5 depicts MCorr and MRMSE as a function of the number of sensors, ranging from **1 to 50**. As expected, increasing the number of sensors improved reconstruction quality, but with diminishing returns beyond a certain threshold. This behavior is consistent with the intuition that once the principal spatial patterns are captured, using the Shannon-entropy method, additional sensors provide limited additional information.



(a)



(b)

Figure 5: Description of the two panels: (a) explanation for the first, (b) explanation for the second.

Discussion

This study demonstrates that an interpolation framework originally developed for air pollution can be effectively adapted to the very different context of contamination in saturated porous media. Although air and porous media differ markedly in their transport physics, sensing constraints, and spatial heterogeneity, the encoder-decoder model successfully reconstructed high-resolution concentration fields from sparse measurements. This outcome is not trivial: unlike air quality monitoring, where sensor density is increasingly feasible, subsurface characterization relies on sparse, expensive, and invasive sampling. The fact that the statistical model still performs well underscores that its predictive ability stems from learning structured relationships in the data rather than from any domain-specific physical continuity.

The implications extend beyond the specific case examined. Conceptualizing interpolation as a statistical reconstruction problem allows the method to remain agnostic to the physical medium, provided that the training data capture the relevant variability.

Several limitations should be acknowledged. The synthetic scenarios assume access to

a relatively large number of soil sensors, a condition that may be impractical in real field deployments. Soil sampling is constrained by cost, logistics, and land-use restrictions, and sensor locations cannot always be optimized for ideal information gain. These limitations emphasize the need for strategies that combine prior knowledge of average soil properties with a minimal set of initial measurements.

Future work should therefore test the framework on real contaminated sites and investigate adaptive sampling strategies that iteratively refine subsurface characterization. Additional extensions may include coupling the interpolation with physical drivers of variability or integrating it with sensor-placement methods grounded in information theory. Across these directions, the broader message remains: statistical reconstruction offers a promising path toward generalizable, cross-domain environmental inference.

Conclusion

This work shows that a data-driven encoder–decoder framework, originally developed for air pollution, can be extended to contamination in porous media and outperform standard interpolation methods under sparse and noisy sampling. The approach thus provides a flexible tool for reconstructing dense pollution fields in complex, data-limited settings. Future research should validate the method using real subsurface contamination data and develop adaptive sensor-placement strategies to improve practicality and scalability.

Acknowledgments

During the preparation of this work, the authors used Chat-GPT to support language refinement and improve clarity in the manuscript. All intellectual content, scientific reasoning, and final editing were carried out and approved by the authors. Y.E. acknowledge the support of the Israel Science Foundation (grant no. 3774/24).

Author CREDIT statement

Alon Feldman: Conceptualization, Software, Formal analysis, Investigation, Visualization, Writing – original draft.

Shai Kendler: Supervision, Writing – review & editing.

Yaniv Edery: Conceptualization, Software, Data acquisition, Writing – review & editing.

Barak Fishbain: Conceptualization, Methodology, Supervision, Writing – review & editing.

References

(1) Bear, J. *Dynamics of fluids in porous media*; Courier Corporation, 2013.

(2) Edery, Y.; Geiger, S.; Berkowitz, B. Structural controls on anomalous transport in fractured porous rock. *Water Resources Research* **2016**, *52*, 5634–5643.

(3) Haggerty, R.; Fleming, S. W.; Meigs, L. C.; McKenna, S. A. Tracer tests in a fractured dolomite: 2. Analysis of mass transfer in single-well injection-withdrawal tests. *Water Resources Research* **2001**, *37*, 1129–1142.

(4) Edery, Y.; Porta, G. M.; Guadagnini, A.; Scher, H.; Berkowitz, B. Characterization of bimolecular reactive transport in heterogeneous porous media. *Transport in Porous Media* **2016**, *115*, 291–310.

(5) Edery, Y.; Stolar, M.; Porta, G.; Guadagnini, A. Feedback mechanisms between precipitation and dissolution reactions across randomly heterogeneous conductivity fields. *Hydrology and Earth System Sciences* **2021**, *25*, 5905–5915, Publisher: Copernicus GmbH.

(6) Wu, Y.; Ajo-Franklin, J. B.; Spycher, N.; Hubbard, S. S.; Zhang, G.; Williams, K. H.; Taylor, J.; Fujita, Y.; Smith, R. Geophysical monitoring and reactive transport mod-

eling of ureolytically-driven calcium carbonate precipitation. *Geochemical transactions* **2011**, *12*, 7.

(7) Berkowitz, B.; Cortis, A.; Dentz, M.; Scher, H. Modeling non-Fickian transport in geological formations as a continuous time random walk. *Reviews of Geophysics* **2006**, *44*.

(8) Berkowitz, Y.; Edery, Y.; Scher, H.; Berkowitz, B. Fickian and non-Fickian diffusion with bimolecular reactions. *Phys. Rev. E* **2013**, *87*, 032812.

(9) Cirpka, O. A.; Kitanidis, P. K. Characterization of mixing and dilution in heterogeneous aquifers by means of local temporal moments. *Water Resources Research* **2000**, *36*, 1221–1236.

(10) Cushman, J. H.; Ginn, T. Nonlocal dispersion in media with continuously evolving scales of heterogeneity. *Transport in Porous Media* **1993**, *13*, 123–138.

(11) Dullien, F. A. L. *Porous Media: Fluid Transport and Pore Structure*; Academic Press, 2012; Google-Books-ID: pPJCOKfZGPkC.

(12) Le Borgne, T.; Dentz, M.; Carrera, J. Lagrangian statistical model for transport in highly heterogeneous velocity fields. *Phys. Rev. Lett.* **2008**, *101*.

(13) Dagan, A.; Edery, Y. Bifurcating paths: The relation between preferential pathways, channel splitting, under sampled regions, and tortuosity on the Darcy scale. *Advances in Water Resources* **2024**, *184*, 104622.

(14) Comolli, A.; Dentz, M. Anomalous dispersion in correlated porous media: a coupled continuous time random walk approach. *The European Physical Journal B* **2017**, *90*, 166.

(15) Webb, E. K.; Anderson, M. P. Simulation of Preferential Flow in Three-Dimensional, Heterogeneous Conductivity Fields with Realistic Internal Architec-

ture. *Water Resources Research* **1996**, *32*, 533–545, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/95WR03399.

(16) Zhan, Z.; Wei, Y.; Yeh, T. C. J.; Chen, Y.; Chen, Y.; Li, Y.; Zhang, J.; Wen, Y.; Li, H. Small Data Insights for Groundwater Management. *Environmental Science & Technology* **2025**, *59*, 3339–3343.

(17) Man, J.; Chen, Y.; Fan, H.; Chen, Q.; Yao, Y. Optimizing Soil Sampling with Information Entropy at Heavy-Metal Sites. *ACS ES&T Engineering* **2023**, *3*, 1350–1358.

(18) Man, J.; Zeng, L.; Luo, J.; Gao, W.; Yao, Y. Application of the Deep Learning Algorithm to Identify the Spatial Distribution of Heavy Metals at Contaminated Sites. *ACS ES&T Engineering* **2021**, *2*, 158–168.

(19) Kim, S. G.; Kim, G.-B. Are Groundwater Monitoring Networks Economical? Cost-Benefit Analysis on the Long-Term Groundwater Supply Project of South Korea. *Water* **2019**, *11*, 753, Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

(20) Aral, M. M.; Guan, J.; Maslia, M. L. Optimal Design of Sensor Placement in Water Distribution Networks. *Journal of Water Resources Planning and Management* **2010**, *136*, 5–18, Publisher: American Society of Civil Engineers.

(21) Honeycutt, J.; Johnson, H.; Kelly, S. Using data assimilation to better predict contaminant transport in fluids. *SIAM Undergr. Res. Online* **2018**, *11*, 1–16.

(22) Camporese, M.; Girotto, M. Recent advances and opportunities in data assimilation for physics-based hydrological modeling. *Frontiers in Water* **2022**, *4*, 948832.

(23) Cao, W.; Song, J.; Zhang, W. Solving high-dimensional parametric engineering problems for inviscid flow around airfoils based on physics-informed neural networks. *Journal of Computational Physics* **2024**, *516*, 113285.

(24) Huang, X.; Tang, J.; Shen, Y.; Zhao, Y.; Hao, S. A physically informed neural network approach for modeling wave transformation in vegetated waters. *Engineering Applications of Artificial Intelligence* **2025**, *159*, 111803.

(25) Yan, B.; Harp, D. R.; Chen, B.; Pawar, R. A physics-constrained deep learning model for simulating multiphase flow in 3D heterogeneous porous media. *Fuel* **2022**, *313*, 122693.

(26) Ohmer, M.; Liesch, T.; Wunsch, A. Spatiotemporal optimization of groundwater monitoring networks using data-driven sparse sensing methods. *Hydrology and Earth System Sciences* **2022**, *26*, 4033–4053.

(27) Zhao, W.; Ma, J.; Liu, Q.; Dou, L.; Qu, Y.; Shi, H.; Sun, Y.; Chen, H.; Tian, Y.; Wu, F. Accurate Prediction of Soil Heavy Metal Pollution Using an Improved Machine Learning Method: A Case Study in the Pearl River Delta, China. *Environmental Science & Technology* **2023**, *57*, 17751–17761.

(28) Rizzo, C. B.; de Barros, F. P. Minimum hydraulic resistance and least resistance path in heterogeneous porous media. *Water Resources Research* **2017**, *53*, 8596–8613.

(29) Tyukhova, A. R.; Kinzelbach, W.; Willmann, M. Delineation of connectivity structures in 2-d heterogeneous hydraulic conductivity fields. *Water Resources Research* **2015**, *51*, 5846–5854.

(30) Feldman, A.; Kendler, S.; Pisoni, E.; Fishbain, B. Ridiculously Simple Data-Driven Air Pollution Interpolation Method. *Available at SSRN* **2025**,

(31) Gómez-Hernández, J. J.; Journel, A. G. In *Geostatistics Tróia '92: Volume 1*; Soares, A., Ed.; Springer Netherlands: Dordrecht, 1993; pp 85–94.

(32) Ababou, R.; McLaughlin, D.; Gelhar, L. W.; Tompson, A. F. B. Numerical simulation

of three-dimensional saturated flow in randomly heterogeneous porous media. *Transport in Porous Media* **1989**, *4*, 549–565.

(33) Riva, M.; Guadagnini, A.; Neuman, S. P.; Janetti, E. B.; Malama, B. Inverse analysis of stochastic moment equations for transient flow in randomly heterogeneous media. *Advances in Water Resources* **2009**, *32*, 1495–1507.

(34) Edery, Y.; Guadagnini, A.; Scher, H.; Berkowitz, B. Origins of anomalous transport in heterogeneous media: Structural and dynamic controls. *Water Resources Research* **2014**, *50*, 1490–1505, ISBN: 0043-1397 Publisher: Wiley Online Library.

(35) Zehe, E.; Loritz, R.; Edery, Y.; Berkowitz, B. Preferential pathways for fluid and solutes in heterogeneous groundwater systems: self-organization, entropy, work. *Hydrology and earth system sciences* **2021**, *25*, 5337–5353, Publisher: Copernicus GmbH.

(36) Edery, Y. The Effect of varying correlation lengths on Anomalous Transport. *Transport in Porous Media* **2021**, *137*, 345–364.

(37) Guadagnini, A.; Neuman, S. P. Nonlocal and localized analyses of conditional mean steady state flow in bounded, randomly nonuniform domains: 1. Theory and computational approach. *Water Resources Research* **1999**, *35*, 2999–3018.

(38) Domenico, P.; Schwartz, F. 1990, Physical and chemical hydrogeology. Publisher: New York, John Wiley and Sons.

(39) Eze, P. N.; Madani, N.; Adoko, A. C. Multivariate mapping of heavy metals spatial contamination in a Cu–Ni exploration field (Botswana) using turning bands co-simulation algorithm. *Natural Resources Research* **2019**, *28*, 109–124.

(40) Obi, I. S.; Onuoha, K. M.; Obilaja, O. T.; Dim, C. Understanding reservoir heterogeneity using variography and data analysis: an example from coastal swamp deposits, Niger Delta Basin (Nigeria). *Geologos* **2020**, *26*.

(41) Ciriello, V.; Guadagnini, A.; Di Federico, V.; Edery, Y.; Berkowitz, B. Comparative analysis of formulations for conservative transport in porous media through sensitivity-based parameter calibration. *Water Resour. Res.* **2013**, *in press.*

(42) Ciriello, V.; Edery, Y.; Guadagnini, A.; Berkowitz, B. Multimodel framework for characterization of transport in porous media. *Water Resources Research* **2015**, *51*, 3384–3402.

(43) Franssen, H. H.; Stauffer, F.; Kinzelbach, W. *geoENV IV—Geostatistics for Environmental Applications*; Springer, 2004; pp 223–234.

(44) Riva, M.; De Simoni, M.; Willmann, M. *Geostatistics for Environmental Applications*; Springer, 2005; pp 273–284.

(45) Li, L.; Zhou, H.; Gómez-Hernández, J. J. Transport upscaling using multi-rate mass transfer in three-dimensional highly heterogeneous porous media. *Advances in Water Resources* **2011**, *34*, 478–489.

(46) Salamon, P.; Fernàndez-Garcia, D.; Gómez-Hernández, J. Modeling mass transfer processes using random walk particle tracking. *Water Resources Research* **2006**, *42*.

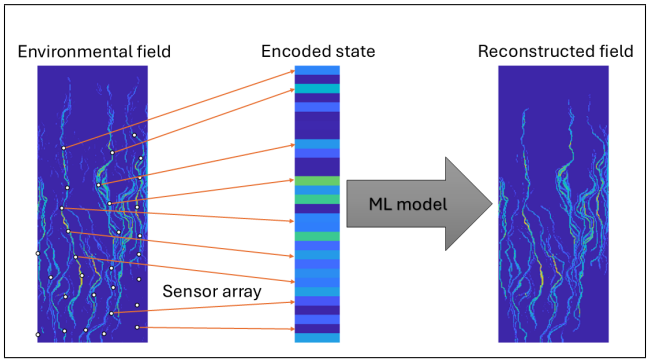
(47) Salamon, P.; Fernàndez-Garcia, D.; Gómez-Hernández, J. J. A review and numerical assessment of the random walk particle tracking method. *Journal of contaminant hydrology* **2006**, *87*, 277–305.

(48) Sausa, A. R. P.; Li, S.; Kaye, N. B.; Flynn, M. R. The coalescence of adjacent turbulent plumes in a stratified and unstratified environment. *Environmental Fluid Mechanics* **2024**, *24*, 923–951.

(49) Mano, Z.; Kendler, S.; Fishbain, B. Information Theory Solution Approach to the Air Pollution Sensor Location–Allocation Problem. *22*, 3808.

(50) Cressie, N. *Statistics for Spatial Data*; Wiley, 1993.

TOC Graphic


Environmental field
Encoded state
Reconstructed field
ML model
Sensor array