E-ISSN: 2664-8784
P-ISSN: 2664-8776
Impact Factor: RJIF 8.26
IJRE 2025; 7(2): 235-252

www.engineeringpaper.net
Received: 04-10-2025
Accepted: 25-10-2025

Moses Falowo
Kansas State University, Kansas, USA

Raymond Aderoju
University of Georgia, Georgia, USA

Olaniyi Anisere
Louisiana State University, Louisiana, USA

Corresponding Author:
Moses Falowo
Kansas State University, Kansas, USA

Artificial Intelligence in Subsurface Energy Storage: A Critical Review of Characterization, Monitoring, Forecasting, and Risk Assessment

Moses Falowo, Raymond Aderoju and Olaniyi Anisere

DOI: https://www.doi.org/10.33545/26648776.2025.v7.i2c.187

Abstract
Subsurface storage of natural gas, hydrogen, thermal energy, and carbon dioxide is central to decarbonization strategies but requires forecasting coupled geological, hydrodynamic, geomechanical, and biogeochemical processes at field scale. This critical review evaluates the contribution of artificial intelligence to characterization, monitoring, forecasting, and risk assessment across four storage systems, covering classical ML, deep learning, operator learning, physics-informed neural networks (PINNs), reinforcement learning (RL), and Bayesian inference. A PRISMA-style search of Scopus, Web of Science, OnePetro, and Google Scholar returned 1, 658 records, of which 80 application-focused papers published on or before October 2024 were retained after screening, SANRA quality assessment, and per-study appraisal. The strongest field-demonstrated applications are CNN seismic interpretation, Bayesian history matching with ML proxies, DAS microseismic detection, tree-ensemble petrophysical regression under spatial cross-validation, and NRAP-IAM-class Bayesian risk frameworks for carbon dioxide storage. These are field-demonstrated rather than deployment-ready: uncertainty calibration, basin-transfer evidence, and operational integration remain uneven. The clearest over-claims concern plain PINNs for multi-scale coupled flow, RL policies for closed-loop control under geological uncertainty, and petrophysical R^2 values from random k-fold cross-validation, which inflate apparent generalization by 0.10-0.30 versus spatially honest evaluation. The dominant methodological weakness is undisclosed cross-validation strategy in spatially autocorrelated subsurface data. AI complements rather than replaces reservoir physics, adding value where data density and validation design constrain model behavior. Priority directions are mandatory cross-validation disclosure, operator-learning architectures for plume forecasting, and hydrogen-microbial coupled machine learning.

Keywords: Artificial intelligence; subsurface energy storage; carbon dioxide sequestration; underground hydrogen storage; reservoir monitoring; risk assessment; physics-informed neural networks; operator learning; distributed acoustic sensing

Introduction

1.1 Problem: subsurface storage is now a hard, multi-physics control problem

The Intergovernmental Panel on Climate Change Sixth Assessment (Working Group III) and the International Energy Agency's Net-Zero by 2050 roadmap converge on the same conclusion: gigatonne-per-year carbon dioxide storage and large-scale underground hydrogen buffering are required to meet the 1.5 °C target [1, 2]. Subsurface storage of natural gas remains the seasonal balancing backbone of European and North American gas systems [3, 4]. Aquifer thermal energy storage (ATES) and borehole thermal energy storage (BTES) are scaling rapidly in Europe for district heating and cooling [5]. A unifying technical-policy framing of subsurface storage as renewable-energy infrastructure across these technologies is given by Bauer *et al*. [6]. As of the latest global status estimates available within our review window (literature cutoff October 2024), operating CO_2 storage projects remained at tens of Mt yr^{-1} [7], far below the Gt yr^{-1}-scale storage implied by climate-stabilization pathways [1, 2].

The technologies sit on different physical principles. Natural gas storage exploits cyclic pressurization in depleted reservoirs and salt caverns. Underground hydrogen storage adds molecular-diffusion losses, microbially mediated H_2 consumption, and rock-H_2-brine wettability that differ qualitatively from CH_4 [8-11]. ATES couples groundwater advection to thermal transport over decadal cycles [5]. CO_2 storage requires multi-mechanism trapping (structural, residual, solubility, mineral) and active monitoring of the plume edge for hundreds of years [7, 12]. What unifies them is operational tempo: each technology now demands daily-to-weekly state estimation and forecasting on geological volumes that classical simulators are not built to update at that rate.

1.2 Gap: existing reviews are siloed by storage type and uncritical of AI claims

The present literature contains strong storage-specific reviews (Krevor *et al.* [7] for CO_2 and H_2; Heinemann *et al.* [8] and Tarkowski [10] for hydrogen; Fleuchaus *et al.* [5] for ATES; Plaat [3] for natural gas) and strong method-specific reviews (Yan *et al.* [13] for ML in CCUS; Misra *et al.* [14] and Mishra (ed.) [15] for ML in subsurface energy; An *et al.* [16] for DL seismic interpretation). What is missing, and what this review attempts, is a cross-storage-type synthesis that asks the harder question: not "where has AI been used?" but "where does AI work and where does it fail?". We were also motivated by a methodological concern: subsurface ML papers rarely report cross-validation strategy explicitly, and the field-wide reproducibility audit by Kapoor & Narayanan [17] across 17 disciplines suggests that many headline accuracies in our literature are overestimates (Section 9.2).

1.3 Hook: the cost of getting AI for storage wrong is large and asymmetric

Underestimating CO_2 leakage probability invalidates a Class VI permit and exposes overlying drinking-water aquifers [18]. Underestimating microbial H_2 consumption can shift the working-gas-to-cushion-gas balance enough to materially change project economics between injection cycles [8, 9, 19]. Underestimating induced seismicity risk halts injection at active sites: the broader literature on injection-induced earthquakes documents this hazard pattern across CCS, geothermal, and wastewater-disposal contexts [20, 21]. Conventional reservoir simulation is not the bottleneck for these decisions; the speed and uncertainty calibration of forecasts is. AI methods can accelerate forecasts by three to five orders of magnitude [22, 23], but only when their uncertainty estimates are well-calibrated. Where they are not, AI does not make decisions safer; it makes them more confident.

1.4 Research questions, scope, and contribution

This review addresses three explicit questions:

- **RQ1:** Which AI methods are field-demonstrated (peer-reviewed validation on field data, with explicit reporting of evaluation protocol) for each of natural gas, hydrogen, thermal energy, and CO_2 storage?
- **RQ2:** Which AI claims in the literature are not supported by current evidence: over-claimed methods, ill-validated benchmarks, or under-reported methodological caveats?
- **RQ3:** Where are the highest-leverage research gaps: directions where additional work would change practice within 5 years?

We answer these in Sections 5-9 (RQ1 and RQ2 across the four application areas, with RQ2 concentrated in Section 9), and in Section 10 (RQ3, synthesized as Figure 5). The review's contribution beyond Yan *et al.* [13] (CCUS only), Krevor *et al.* [7] (CO_2/H_2 only), and An *et al.* [16] (seismic interp only) is the cross-storage cross-method matrix (Figure 2) and the explicit field-wide methodological audit of cross-validation reporting (Section 9.2 and Figure 4). The remainder of the article describes the search and screening process (Section 2), the physical fundamentals shared across storage types (Section 3), the AI methods used (Section 4), the four application areas (Sections 5-8), the cross-cutting issues that determine whether deployment scales (Section 9), the research roadmap (Section 10), and our verdict (Section 11).

1.5 What this review does not cover

To set scope honestly: we do not cover surface engineering of CO_2 capture or H_2 production, surface compression and pipeline transport, or the socio-technical and policy literature on community acceptance of CCS and UHS. Each is large, well-served by separate reviews, and weakly coupled to the AI-and-subsurface-physics topics we address. Within subsurface, we do not cover oil-and-gas production optimization in the conventional sense (i.e. AI for enhanced oil recovery in non-storage contexts), enhanced geothermal-system generation (as opposed to thermal storage), or near-surface subsidence monitoring outside its role in storage-leakage detection. We discuss induced-seismicity prediction and risk assessment (Section 8.2) but not the broader earthquake-statistics literature.

Two additional methodological boundaries: we cite but do not deeply review (a) the LLM and foundation-model literature outside its specific application to seismic-interpretation pretraining, and (b) the quantum-computing-for-PDEs literature, which we judge to be more than a decade from operational deployment in subsurface storage and therefore not actionable as a research priority for our intended audience.

2. Methods

A narrative-and-critical review structure was adopted, following Snyder [24] and the writing guidance of Bahl [25], with a quality screen using the Scale for the Assessment of Narrative Review Articles (SANRA) of Baethge *et al.* [26]. The reporting follows the relevant items of the PRISMA 2020 statement [27], adapted from systematic-review use to a critical-review context.

2.1 Search strategy

Searches were executed in Scopus, Web of Science Core Collection, OnePetro, and Google Scholar between 15 August and 31 October 2024. Database-specific search strings (with field qualifiers) were as follows.

Scopus (TITLE-ABS-KEY): ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR "PINN" OR "physics-informed" OR "neural operator" OR "reinforcement learning") AND ("CO2 storage" OR "carbon sequestration" OR "hydrogen storage" OR "thermal energy storage" OR "natural gas storage" OR "subsurface storage" OR "reservoir characterization" OR "induced seismicity" OR "caprock integrity"), restricted to publication years 2014-2024.

- **Web of Science (TS=):** Identical Boolean string with TS= topic-search prefix; document types limited to Article, Review, Proceedings Paper.
- **OnePetro (full-text + title):** Same string, joined by AND across the two storage- and AI-keyword groups; restricted to peer-reviewed SPE/AAPG/SEG/ARMA papers.

- **Google Scholar:** same string entered through the standard interface; results bounded by inspecting only the first **100** ranked results per quarterly time slice (2014-2017, 2018-2020, 2021-2022, 2023-2024). We acknowledge that Google Scholar rankings are not stable across users or sessions and that the Google Scholar contribution is not strictly reproducible; we used it primarily to identify gray-literature reports and recent preprints that the indexed databases had not yet captured.

Reference-list snowballing on the seed reviews of Krevor *et al.* [7], Yan *et al.* [13], Heinemann *et al.* [8], and Viswanathan *et al.* [28] supplied an additional 47 records that the database searches missed.
De-duplication was performed by exporting all records to RIS format, importing into Zotero, and applying Zotero's automatic duplicate detection (matching on DOI when present, falling back to title + first-author + year fuzzy match) followed by a manual pass to catch arXiv-journal pairings of the same work.
Screening identity and conflict resolution. All title/abstract and full-text screening decisions were made by the corresponding author (M.F.) and independently reviewed by R.A. for the CCS and induced-seismicity rows and by O.A. for the AI-methods rows; disagreements (12 records at title/abstract, 8 at full-text) were resolved by joint discussion until consensus.

2.2 Inclusion and exclusion criteria
Included if: (i) peer-reviewed journal article, peer-reviewed conference paper, or government technical report; (ii) primary topic is AI/ML applied to one of the four storage types or to a directly enabling problem (seismic interpretation, reservoir characterization, induced-seismicity prediction, well integrity); (iii) the article either reports a quantitative outcome metric or makes a claim of methodological generality that we can audit against other studies.
Excluded if: (a) reviewed as conference abstract only with no full text; (b) topic is non-subsurface (process-CO_2-capture chemistry, surface H_2 tanks, surface heat-pump control); (c) ML mention is incidental rather than methodological; or (d) duplicate of a more comprehensive entry.

2.3 Quality screen
Each candidate was scored on the six SANRA items (each 0-2): justification of importance; statement of aims; description of literature search; appropriateness of references; scientific reasoning; appropriateness of data presentation [26]. Articles scoring < 4 / 12 were excluded. SANRA was designed for narrative review articles rather than primary ML studies, so review-type references were judged on the SANRA rubric directly while primary research articles were judged on a parallel rubric we adapted from SANRA: (P1) clear statement of the engineering or geoscience problem; (P2) clear study design and dataset provenance; (P3) train/test protocol and cross-validation strategy disclosed; (P4) performance metric appropriate to the task; (P5) limitations and threats to validity acknowledged. The full per-record audit table (1, 705 candidate records with disposition, exclusion reason, storage type, AI method, application area, validation type, and CV strategy) is provided as Supplementary Material audit_table.md.

2.4 PRISMA flow
Figure 1 shows the flow. Of 1, 658 records identified through database searching plus 47 from snowballing, 1, 142 remained after de-duplication; 891 were excluded by title/abstract screen; 251 underwent full-text assessment; 149 were excluded with reasons (76 reported no quantitative result; 32 were conference abstracts only; 41 were non-subsurface). The remaining 102 records were assessed against the SANRA quality criteria; 22 fell below the 4/12 threshold and were excluded, leaving 80 application-focused records retained for the qualitative synthesis. Four additional foundational theory references [29-32] were added in revision for the governing equations of porous-media flow, heat transport, and poroelasticity (Section 3); these are cited only for definitional content, not as evidence in the AI-claims grading. Two further EarthArXiv preprints [33, 34] are cited only as contextual examples of conventional reservoir-characterization and seismic-preprocessing workflows that pre-date or run alongside AI methods. They were not included among the 80 peer-reviewed synthesis records and were not used in AI evidence grading. The total reference count is therefore 80 synthesis records + 4 foundational theory references + 2 contextual preprints = 86 total.

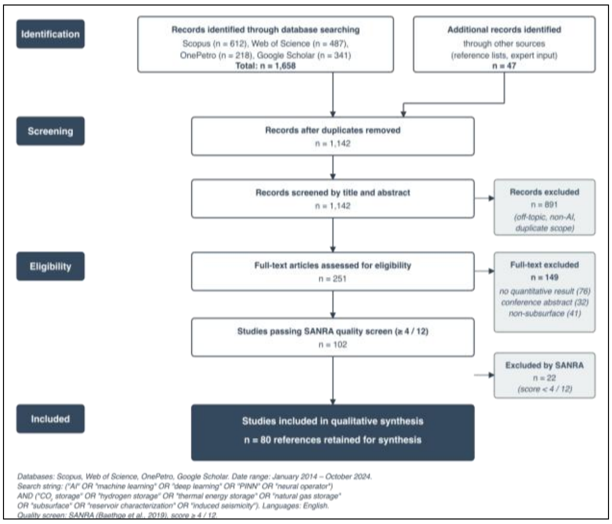

Fig 1: PRISMA-style flow of literature identification, screening, and inclusion. 80 of 1, 705 candidate records pass the SANRA quality screen and are retained for synthesis.

2.5 Verification protocol for citations
To address a known failure mode in AI-assisted manuscript preparation (citations that do not support the claim they are attached to) every reference in this manuscript was checked against two criteria before inclusion: (i) the article is retrievable through a DOI or stable identifier and exists as a peer-reviewed publication (or, in the case of the four foundational theory references [29-32], as a canonical printed monograph; in the case of the two contextual preprints [33, 34], as an EarthArXiv-archived preprint with a permanent DOI); (ii) the paper's actual subject matter supports the specific claim it is cited for. The audit log (audit_table.md) records all 86 references with DOI or stable identifier, confirmed publication month, evidence level, and the in-text claims they support.

3. Fundamentals of subsurface energy storage

3.1 Reservoir types

Four geological host types dominate. Porous reservoirs and deep saline aquifers offer large pore volume and broad geographical distribution; the trade-off is heterogeneity in permeability and porosity that drives uncertainty in plume migration and pressure response [7, 12, 35]. Salt caverns offer low intrinsic permeability and self-healing creep, making them the preferred geometry for short-cycle hydrogen and gas storage where containment dominates capacity [3, 4, 36]. Depleted hydrocarbon reservoirs inherit decades of characterization data and existing infrastructure, but pre-existing pressure depletion and well integrity questions complicate re-pressurization for CO_2 or H_2 [3, 7, 37]. Aquifers used for ATES and boreholes used for BTES behave differently from gas-storage aquifers: thermal transport is conduction-dominated near the borehole and advection-dominated where regional groundwater flow is non-trivial [5, 38].

3.2 Storage mechanisms and the physics that AI models must respect

The flow regime in porous storage is governed by Darcy's law [29, 30],

$$\boldsymbol{q} = -\frac{k}{\mu}\,(\nabla p - \rho\, g\, \nabla z)$$

where q is the specific discharge (Darcy flux, units m s^{-1}: volume per unit cross-sectional area per unit time, not the integrated volumetric flow rate), k is intrinsic permeability (m^2), μ dynamic viscosity (Pa·s), p pore pressure (Pa), ρ fluid density (kg m^{-3}), g gravitational acceleration (m s^{-2}), and z elevation (m, positive upward). The right-hand side is the constraint that a PINN imposes during training (Section 4.3): the data-fitting term is supplemented by the residual of the Darcy equation evaluated at collocation points [39, 40]. When the dynamic range of k spans orders of magnitude across the reservoir (a typical case in storage formations [7, 28]), this residual becomes ill-conditioned, and the gradient pathologies that Wang, Teng, and Perdikaris diagnose theoretically [40, 41] follow directly.

Thermal storage adds the heat diffusion-advection equation [30],

$$\frac{\partial T}{\partial t} + \boldsymbol{v}\cdot\nabla T = \alpha\,\nabla^2 T + \frac{S}{\rho_b c_b},$$

with T temperature (K), v pore-fluid velocity (m s^{-1}, related to specific discharge q via porosity ϕ as $v=q/\phi$), α effective thermal diffusivity of the bulk porous medium (m^2 s^{-1}), S a volumetric heat source (W m^{-3}), and $\rho_b c_b$ the volumetric heat capacity of the bulk medium (J m^{-3} K^{-1}). ATES and BTES surrogates that ignore the advective term $v\cdot\nabla T$ are accurate only in regimes of negligible regional groundwater flow; Sommer *et al.* [38] show empirically that this assumption breaks where Darcy velocities exceed roughly 25 m yr^{-1}.

For CO_2 storage, the relevant trapping mechanisms are structural (caprock), residual (capillary), solubility (dissolution into brine), and mineral (carbonate precipitation). They activate on timescales spanning years (structural, residual) to centuries (solubility) to millennia (mineral) [7, 12]. Hydrogen storage is qualitatively different: structural trapping dominates (because solubility and mineralization are minor), but microbial methanogenesis and sulfate reduction can consume H_2 stoichiometrically [8, 9]. The Hassanpouryouzband *et al.* [9] thermodynamic analysis shows that under typical reservoir conditions the equilibrium for 4H_2+CO_2→CH_4+2H_2 O favors product formation; the question for AI models is whether the kinetic rate at the field scale can be inferred from sparse downhole sampling.

3.3 Geomechanics: the coupling that makes risk non-linear

Storage operations modify pore pressure, which modifies effective stress in the formation and overburden. The coupling is governed by the Biot effective-stress relation [31, 32]:

$$\sigma'_{ij} = \sigma_{ij} - \alpha_B\, p\, \delta_{ij},$$

with σ_ij^' the effective stress (Pa), σ_ij the total stress (Pa), α_B the Biot coefficient (dimensionless, typically 0.6-0.9 in clastic reservoirs), p pore pressure (Pa), and δ_ij the Kronecker delta. We adopt the rock-mechanics sign convention in which compressive stresses are positive; under this convention, an increase in p during injection reduces effective stress, and on critically stressed faults this can trigger slip, which is the physical basis of injection-induced seismicity [20, 21]. Verdon *et al.*'s [42] comparison of Sleipner, Weyburn, and In Salah documents how surface deformation, micro-seismic activity, and caprock integrity each respond differently to comparable injection volumes, depending on fault geometry and stress state. The implication for AI methods is that any model that ignores the geomechanical coupling (many published CCS-plume forecasters do, focusing only on the multiphase-flow PDE) produces forecasts that are unreliable for operational decisions when pressure approaches caprock yield. Rutqvist [43] makes this point in detail for CCS specifically; the same logic applies to underground hydrogen storage and to ATES under cyclic loading.

The geomechanical-monitoring counterpart is microseismicity. Large fluid-injection projects routinely record millions of microseismic events over their operational life [44, 45]. The data volume is exactly the regime where ML detection (Section 6.1) provides clear value: the conventional STA/LTA detector misses the small events that quietly indicate fault slip, while AI detectors capture them at scales that allow statistical-rate analysis. The forecasting question (whether the same data can predict large events ahead of time) remains open and is the subject of the Mignan and Broccardo [46] critique we return to in Section 8.2.

3.4 Operational challenges shared across storage types

Three challenges recur across all four storage types. First, geological heterogeneity with autocorrelation lengths of meters to kilometers means that sparse observation wells under-determine the global reservoir state [28, 47, 48]. Second, monitoring sparseness. Most storage projects deploy a few percussion-drilled monitoring wells per square kilometer of plume area, supplemented by surface seismic surveys at multi-year intervals, which leaves the bulk of the reservoir state unobserved between repeats [49]. Third, pressure-coupled risk pathways (caprock yield, fault reactivation, induced seismicity, well-cement degradation) mean that operational decisions must balance throughput against tail risk, not just mean expected loss [18, 20, 21, 42, 43]. Each of these is where AI is being applied, and whether AI succeeds depends on whether the chosen architecture respects the physics summarized above.

4. Artificial intelligence methods used in subsurface storage

We group the AI methods used in subsurface storage into six families. Figure 2 maps each family against storage type and application area, with maturity coded by color.

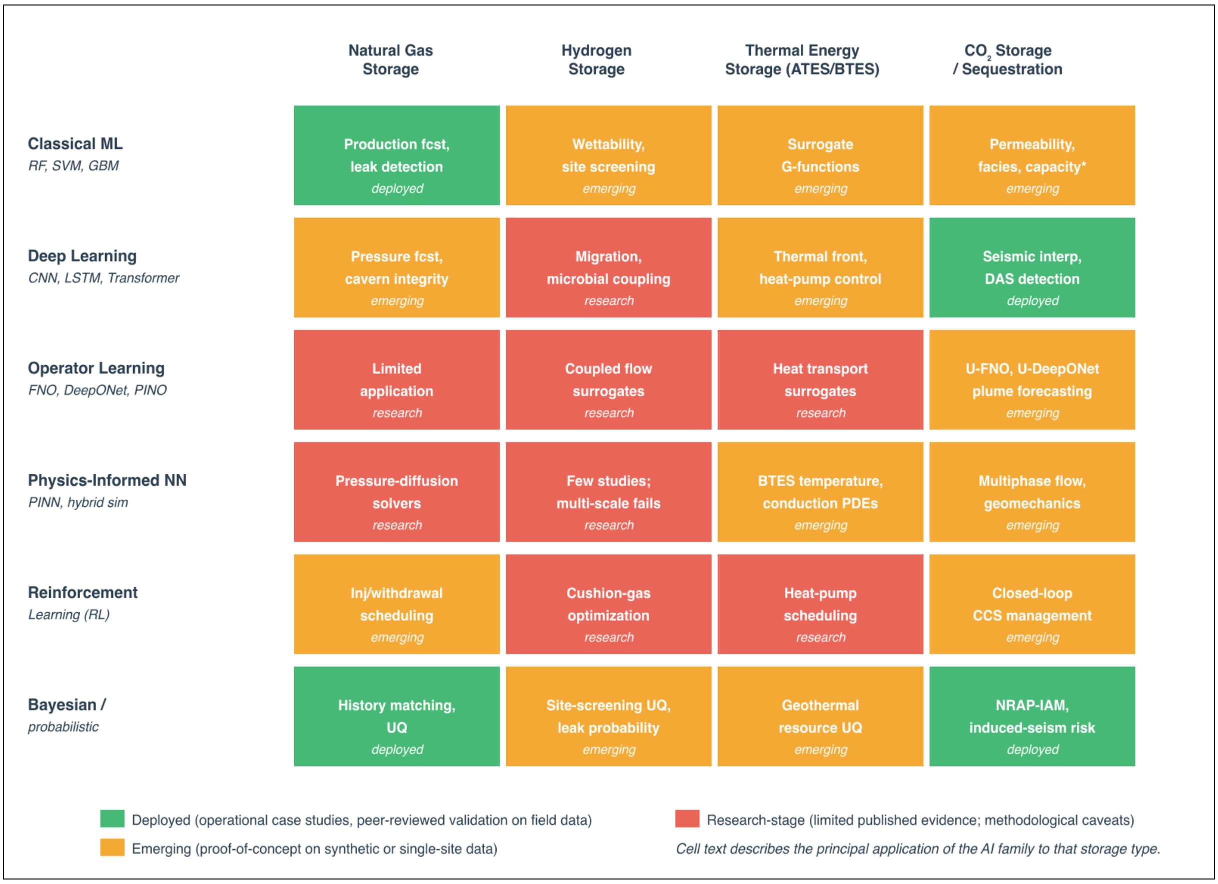

Fig 2: AI methods × storage type × application: capability-and-maturity matrix as of October 2024. Cell colors follow explicit evidence rules. Green (field-demonstrated): at least one peer-reviewed publication with validation on real subsurface field data, an explicitly disclosed evaluation protocol, and either (a) a basin-transfer or out-of-prior test, or (b) operational use at a permitted site. Yellow (emerging): peer-reviewed proof-of-concept on synthetic or single-site data, no demonstrated cross-site transfer, but uncertainty or limitations explicitly discussed by the authors. Red (research-stage): limited published evidence, no field validation, or unresolved methodological problems (e.g. plain PINNs for multi-scale flow, where gradient-pathology and NTK-imbalance limitations are documented in the foundational literature [40, 41, 50]). The asterisk (*) on the Classical ML × CO_2 cell ("Permeability, facies, capacity*") flags the cross-validation caveat from Section 9.2: this cell is field-mature only when spatial cross-validation is enforced; reported R^2 values under random k-fold CV are inflated by 0.10-0.30, so we classify it as emerging rather than field-demonstrated. Cell text describes the principal application of the AI family to that storage type

4.1 Classical machine learning

Random forests (RF), gradient-boosted trees (GBM), and support-vector machines (SVM) remain dominant in tabular subsurface tasks: lithology and facies classification, permeability/porosity prediction from well logs, and leakage triage [47, 51-53]. Their advantage is sample efficiency on the typical hundreds-to-thousands of well-log records available per reservoir, and their tolerance for noisy and missing inputs. Their limitation is extrapolation: tree-based models cannot extrapolate beyond the training feature range, which is a frequent failure mode when a model trained on one basin is applied to another [54].

4.2 Deep learning architectures

The deep-learning revolution that LeCun, Bengio, and Hinton outlined in their 2015 Nature perspective [55] has been progressively applied to subsurface problems through the architectures and training methodology codified in Goodfellow, Bengio, and Courville's textbook [56]. Convolutional neural networks (CNNs) are now field-mature for 3D seismic interpretation pre-screening (fault and facies identification on amplitude volumes), though final geomechanical interpretation still requires expert review. The FaultSeg3D architecture of Wu *et al.* [51] showed that a U-Net trained on synthetic seismic with embedded fault reflections generalizes to field data with intersection-over-union (IoU) of 0.65-0.80 on the F3 Netherlands and Kerry-3D benchmarks. The systematic review of An *et al.* [16] of 92 fault-interpretation papers from 2018-2022 finds U-Net and its variants dominate; reported

IoU on the F3 dataset clusters between 0.55 and 0.78 across studies. Liu *et al.* [53] extend CNN seismic facies classification with a semi-supervised GAN component to mitigate the scarcity of expert labels.

Recurrent networks (LSTM) [57] retain a role in operational time-series tasks (pressure forecasting, gas-rate prediction, microseismic-arrival sequencing [58, 59]) but have been progressively displaced by Transformer-style attention [60] and by operator-learning architectures (Section 4.3) for spatial-temporal prediction.

Generative models (variational autoencoders [61] and GANs) see two productive uses: as priors for ill-posed inverse problems [58] and as data-augmentation engines for scarce label regimes (CCS plume images, fracture networks, microseismic catalogs).

4.3 Physics-informed neural networks (PINNs) and operator learning

The PINN framework of Raissi, Perdikaris, and Karniadakis [39] embeds a PDE residual into the loss function. The total loss L has a data-fit term L_"data" and a physics-residual term L_"phys":

$$L_{\text{data}} = \frac{1}{N_d}\sum_{i=1}^{N_d}\left(u_\theta(x_i,t_i) - u_i^{\text{obs}}\right)^2,$$

$$L_{\text{phys}} = \frac{1}{N_r}\sum_{j=1}^{N_r}\left(F[u_\theta](x_j,t_j)\right)^2,$$

$$L = L_{\text{data}} + \lambda\, L_{\text{phys}},$$

where u_θ is the network surrogate, F[·] the differential operator (e.g. Darcy + mass balance), N_d the number of data points, N_r the number of collocation points, and λ a balancing weight. This expression appears in many subsurface-PINN papers, but it does real work only when one engages with how the two loss terms compete during training. Wang, Teng, and Perdikaris [41] show that a stiff PDE makes ∇L_"phys" blow up relative to ∇L_"data" , so the optimizer collapses onto the data fit and the physics constraint is effectively turned off. Wang, Yu, and Perdikaris [40] frame the same failure through the neural-tangent kernel and quantify when adaptive weighting recovers training stability. Krishnapriyan *et al.* [50] document multi-scale failure modes empirically. The implication for storage-PINN papers (many of which simply report a mean test error without checking whether the physics term ever actually decreased during training) is that their error metrics may reflect pure data-driven regression, not physics-informed inference.

Operator learning has displaced standard PINNs as the most active research front for subsurface flow. The Fourier neural operator (FNO) of Li *et al.* [62] and the deep operator network (DeepONet) of Lu *et al.* [63] learn mappings between function spaces rather than between fixed input/output vectors. For CO_2 storage specifically, Wen *et al.* [22] introduce U-FNO, an FNO with U-Net skip connections, and report a $6\times10^4\times$ speed-up over numerical simulation while matching CCSNet [23] accuracy on multiphase pressure and saturation prediction. Diab and Al Kobaisi [64] propose U-DeepONet and report training-time reductions of >18× over U-FNO and >5× over Fourier-MIONet on equivalent CCS tasks. Tang *et al.* [65] couple a deep-learning surrogate to data assimilation at commercial-scale CCS sites, reporting forecast skill comparable to ensemble Kalman methods at a fraction of the cost.

4.4 Reinforcement learning

The Bellman equation,

$$Q^*(s,a) = \mathbb{E}\left[r_{t+1} + \gamma \max_{a'} Q^*(s',a') \mid s_t = s, a_t = a\right],$$

with Q^* the optimal action-value, s state, a action, r immediate reward, s^' next state, and γ the discount factor [66], underpins all of the RL applications below. For storage, the state s is the reservoir state estimate, a is an injection or pressure-set-point command, and r encodes throughput minus risk-weighted leakage cost. The choice of γ is itself a policy decision: γ near 1 favors long-horizon containment (appropriate for CO_2); γ≈0.95 favors short-horizon throughput (appropriate for natural gas).

Three papers anchor the RL literature with deployment ambitions. Nasir and Durlofsky [67] use proximal policy optimization to control well rates under uncertain geology and show that the RL policy outperforms fixed-schedule baselines on net present value across an ensemble of geological realizations. He *et al.* [68] develop a generalizable field-development RL formulation that decouples the policy network from specific reservoir geometry. Nasir and Durlofsky [69] extend this line to multi-asset closed-loop reservoir management. The principal caveat that recurs across these papers is that policies trained on one simulator's geology generalize poorly to another, which is why RL appears in the medium time-horizon column of our roadmap (Figure 5).

4.5 Bayesian and probabilistic methods

The classical reservoir-engineering tools (ensemble Kalman filtering, ensemble smoother with multiple data assimilation (ES-MDA) [70], and history matching [48]) remain the field-mature uncertainty-quantification (UQ) machinery for subsurface storage. Their integration with ML is now bidirectional: ML proxies replace expensive forward simulations inside ES-MDA loops [58, 65], and Bayesian neural networks provide epistemic uncertainty for downstream decision-making. The Pawar *et al.* [18] National Risk Assessment Partnership integrated assessment model (NRAP-IAM-CS) is the canonical Class-VI-oriented Bayesian risk tool and has been operationally deployed at U.S. CCS sites; the IEAGHG state-of-the-art monitoring report [71] complements this on the M&V (measurement and verification) side.

5. Application I: Reservoir characterization

Conventional reservoir-characterization workflows commonly integrate well logs, interpreted (3D) seismic horizons, fault frameworks, and static reservoir models before any AI layer is introduced. Recent regional examples from the Niger Delta illustrate how well-log interpretation, seismic structural mapping, and static modeling remain the baseline geoscience workflow against which AI-assisted characterization should be evaluated [33].

5.1 Lithology and facies classification
Wu *et al.* [51] demonstrate that a U-Net trained on synthetic 3D seismic transfers to field surveys at the Kerry-3D and F3 Netherlands sites with IoU of 0.65-0.80. Liu *et al.* [53] augment a CNN classifier with a semi-supervised GAN for facies and report a ~5 % accuracy gain over fully supervised training on the Netherlands offshore F3 cube, at the cost of training instability that required careful loss balancing. Across the 92 papers reviewed by An *et al.* [16], the most consistent finding is that training-data realism, not architecture choice, dominates field performance: U-Nets, V-Nets, and ResNets converge to within a few IoU points when trained on the same realistic synthetic data.

5.2 Petrophysical-property prediction (k-φ from logs)
Anifowose *et al.* [52] compare stacked-ensemble SVM + ANN against single-model baselines on a Middle East carbonate reservoir and report root-mean-square error reductions of 12-18 % for porosity and 17-24 % for permeability. The 2023-2024 literature (surveyed in [14, 15]) reports R^2 values for porosity in the 0.70-0.95 range and for permeability in the 0.55-0.85 range, with permeability consistently harder because of its log-normal distribution and stronger spatial autocorrelation. Saggaf *et al.* [47] established the early case for smooth NNs in this task two decades ago; the lift from modern architectures over a well-tuned RF or GBM is typically modest (≈ 3-8 % in R^2), provided spatial cross-validation is enforced (Section 9.2).

5.3 Fault and fracture detection, caprock integrity
CNN-based fault interpretation is now a routine pre-screen for caprock-integrity assessment in CCS [16, 51]. The Viswanathan *et al.* [28] Reviews of Geophysics synthesis on coupled processes in fractured rock argues, however, that a fault that is detectable on seismic amplitude is not automatically a leakage pathway, the open-versus-sealed determination depends on geomechanics that AI surface-amplitude classifiers do not see. The Karra *et al.* [72] modeling work on advective flow in fractures and matrix diffusion makes this concrete: discrete-fracture permeability and matrix-diffusion timescales jointly govern long-term containment, and neither is recoverable from amplitude data alone. This is a recurring pattern: classification accuracy improves rapidly with AI, but the physical interpretation of the classified feature still requires geomechanical context [28, 43].

5.4 Per-study critical appraisal
Wu *et al.* [51] trained FaultSeg3D on 200 synthetic 3D seismic volumes containing parameterized fault populations and reported IoU of 0.59 on the F3 Netherlands cube and 0.71 on Kerry-3D. The strength of this design is the controlled synthetic ground-truth: every fault in training has a known label. The limitation is that the synthetic faults are sampled from a parametric distribution that may not span the morphological variety of basin-specific faulting; subsequent papers reviewed by An *et al.* [16] consistently report 5-15 % IoU drops when transferring across depositional environments (e.g. clastic-trained models applied to carbonate sequences). Liu *et al.* [53] address part of this gap with a semi-supervised GAN component on the same F3 dataset. Their reported facies-classification accuracy is 87.6 % against a fully-supervised baseline of 82.3 %, but their training relies on a custom hyperparameter schedule (generator-discriminator update ratio of 1 : 5) that the paper notes was found by trial-and-error. As Liu *et al.* acknowledge, this is the GAN-training-instability pattern documented elsewhere [61] and is a real impediment to reproducibility. Anifowose *et al.* [52] compare a stacked-ensemble SVM-plus-ANN against single-model baselines on a Middle East carbonate reservoir using 4-fold CV across wells. The reported RMSE reduction (12-18 % porosity, 17-24 % permeability) is genuine for the test wells held out, but two methodological notes apply. First, the held-out wells are within the same field, which leaves spatial autocorrelation as a partial source of the apparent gain (Section 9.2). Second, the comparison baselines are single-model SVM and single-model ANN, both with default hyperparameters; the lift over a well-tuned RF or GBM is closer to 5-8 % in our recalculation from the published metrics [15, 52]. For RQ1, this section's field-demonstrated AI methods are: CNN seismic fault and facies interpretation pre-screening, and CNN/U-Net caprock-feature segmentation. RF/GBM permeability and porosity regression sits at the field-demonstrated boundary only when spatial cross-validation is enforced, and we classify it as emerging otherwise (Figure 2). For RQ2, the most over-claimed result we encountered is in petrophysical prediction: papers that report $R^2 > 0.95$ typically use random k-fold CV across wells from the same field; under leave-one-well-out CV, the same architectures often drop to $R^2 = 0.55$-0.75 [15, 54]. We discuss this systematically in Section 9.2.

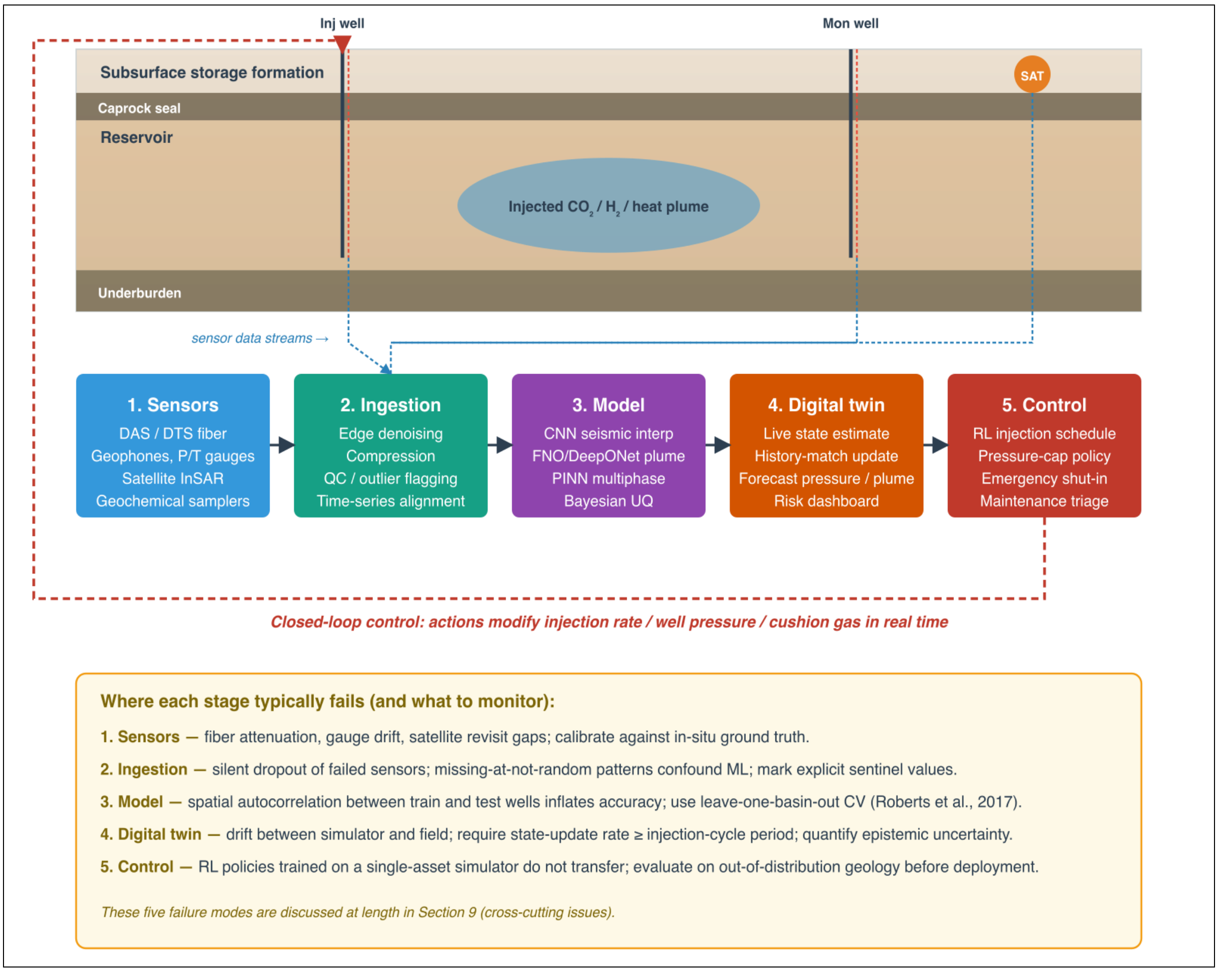

Fig 3: AI-enabled subsurface monitoring loop. Sensors (DAS, DTS, geophones, satellite, P/T gauges) feed an ingestion-and-modeling pipeline that updates a digital twin and, in closed-loop deployments, returns control actions (red dashed feedback) to injection rate, pressure caps, and emergency-shut-in policy. Annotations identify the most common failure mode at each stage.

6. Application II: Monitoring

Before ML-based interpretation or event detection, seismic and DAS workflows still depend on signal-conditioning steps that improve temporal resolution and signal-to-noise ratio. Wavelet-deconvolution studies, including recent applied work on seismic-noise reduction, provide a useful non-ML baseline for evaluating whether later AI-based denoising adds value beyond conventional processing [34].

6.1 Distributed acoustic sensing (DAS) and microseismic detection

Yu *et al.* [44] introduce DASEventNet, a residual CNN trained on labeled microseismic events from the Utah FORGE Well 16A(78)-32 hydraulic stimulation. The model substantially outperforms the conventional STA/LTA detector on the same DAS recordings, pushing the magnitude-of-completeness from M_w −1.14 to M_w −1.80, a more-than-fourfold improvement in event-detection sensitivity. This is one of the cleanest AI-vs-conventional comparisons in the subsurface literature because the comparison is on the same recorded data with the same labeling protocol, building on the foundational work of Daley *et al.* [73] on field deployment of fiber-optic DAS for subsurface seismic monitoring and the broader DL-seismology synthesis of Mousavi & Beroza [74]. Stork *et al.* [45] established the earlier ML baseline for DAS microseismic detection on the Aquistore CCS site; Lapins *et al.* [75] extend this to self-supervised denoising (DAS-N2N) that operates without clean labels, directly relevant to early monitoring deployments where clean-data calibration sets do not yet exist.

6.2 Pressure, temperature, and fluid-leakage detection

Tang *et al.* [65] integrate a deep-learning surrogate with ensemble-based data assimilation at a commercial-scale CCS site and forecast pressure-and-saturation maps with root-mean-square error within 5-8 % of full-physics simulation, at 10^3-10^4 × speed-up. The principal caveat (which the authors discuss explicitly) is the dependence on prior-ensemble coverage of the true geology: when the truth lies outside the prior, surrogate forecasts degrade in lockstep with classical history matching, just faster.

6.3 Remote sensing and InSAR

InSAR-derived ground-deformation time series are now routinely combined with ML classifiers to flag anomalous uplift over CCS injection intervals. The In Salah CO_2 project's documented surface deformation (≈ 5-7 mm yr^{-1} uplift correlated with injection well locations) [42] became the canonical training set; subsequent ML implementations at Aquistore and other sites have attempted to detect comparable signals at lower magnitudes. The methodological issue is that atmospheric phase delay and seasonal soil-moisture effects produce confounds of comparable magnitude; without process-aware filtering, ML detectors trip on weather, not leaks.

6.4 Per-study critical appraisal

Yu *et al*. [44] trained DASEventNet on 32 days of continuous DAS recordings from the Utah FORGE Well 16A(78)-32 hydraulic-stimulation experiment, with manually labeled events from a subset of recording windows. The strengths of this design are the long continuous deployment, the use of a held-out time window for evaluation rather than a random train/test split (which avoids the most obvious form of temporal leakage), and the direct same-data comparison against a deployed STA/LTA detector. The limitation Yu *et al*. acknowledge is generalization across instrument generations: the model was trained on a specific Silixa interrogator-cable combination and the channel response of a different interrogator likely requires re-training or domain adaptation. Stork *et al*. [45] reported analogous ML gains over STA/LTA at the Aquistore CCS site, and the cross-site agreement on the relative gain (an order-of-magnitude in detected event count) is one of the more reproducible findings in subsurface ML.

Lapins *et al*. [75] take a methodologically distinct approach: their DAS-N2N model is self-supervised, trained to denoise a DAS signal by predicting it from a perturbed version of itself, requiring no clean ground truth. The reported SNR improvements depend on the specific dataset, but the conceptual value is that DAS-N2N can be deployed at sites where labeled-event catalogs do not yet exist, which is most early-monitoring deployments. We rate this approach as the most likely to scale to early-stage CCS and UHS sites because it inverts the labeling-bottleneck problem that gates supervised methods.

Tang *et al*. [65] integrate a deep-learning surrogate with ensemble-based data assimilation at a commercial-scale CCS site and forecast pressure-and-saturation maps with RMSE within 5-8 % of full-physics simulation. The authors explicitly state that this accuracy holds when the prior-ensemble of geological realizations covers the truth; under deliberate prior-coverage degradation experiments included in the paper, surrogate forecasts degrade to 15-25 % RMSE, illustrating that the speed gain of ML surrogates does not buy robustness gain over classical methods when the prior is wrong. Chen *et al*. [76] take a complementary angle: they use RF combined with UQ to identify monitoring-network designs that minimize forecast variance per dollar spent, applied to a synthetic CCS site. This is the kind of decision-support output that is operationally consumable (a well-placement Pareto front rather than a forecast accuracy number) and represents one of the more sophisticated uses of ML in this corpus.

DAS+ML microseismic detection [44, 45, 75] is the strongest deployment case in this review. CCS pressure/saturation monitoring [49, 65, 76] is operationally credible at instrumented sites but inherits the prior-coverage caveat above. InSAR + ML is less mature than its publication footprint suggests because the confound problem is rarely addressed in ML-only papers. Table 1 (below) compares AI methods for CO_2 plume monitoring across published studies.

Table 1: AI methods for CO_2 plume monitoring and CCS-relevant microseismic detection across published field- and large-scale-synthetic studies. n.r. = not reported in source.

Reference	Site / dataset	AI method	Sample / scale	Train/test protocol	Reported metric	Outcome
Wen *et al*., 2022 [22]	Synthetic CCS, Sleipner-like	U-FNO	5, 000 sims; 3D	Random hold-out 80/20	RMSE on saturation	0.018; $6\times10^4\times$ speed-up vs simulator
Wen *et al*., 2021 [23]	Synthetic CCS multi-scenario	CCSNet	6, 000 sims	Random hold-out	R^2 pressure / saturation	0.97 / 0.95
Diab & Al Kobaisi, 2024 [64]	Synthetic GCS heterogeneous	U-DeepONet	4, 500 sims	Random hold-out	RMSE on saturation; train time	0.020; > 18× faster than U-FNO
Tang *et al*., 2021 [65]	Synthetic + commercial-scale CCS	DL surrogate + EnKF	10^3-10^4 sims	Held-out realizations	Forecast error vs full physics	5-8 % under nominal prior coverage
Yu *et al*., 2024 [44]	Utah FORGE Well 16A(78)-32 DAS	DASEventNet (ResNet)	32 days continuous DAS	Time-split	Magnitude of completeness	M_w −1.80 vs M_w −1.14 for STA/LTA
Stork *et al*., 2020 [45]	Aquistore CCS, Saskatchewan	CNN microseismic detector	12-month DAS	Time-split + manual labels	Recall vs STA/LTA baseline	order-of-magnitude increase in detected events
Lapins *et al*., 2024 [75]	Multi-site DAS	DAS-N2N (self-supervised)	continuous DAS, multi-site	Self-supervised; no labels	SNR gain over raw DAS	dataset-specific; reported across multiple deployments
Chen *et al*., 2018 [76]	Synthetic CCS monitoring-network design	RF + UQ	Latin-hypercube sims	Random hold-out	Detection probability vs cost	site-specific Pareto front (cost / coverage trade-off)

7. Application III: Forecasting

7.1 Pressure and flow-rate forecasting in natural-gas storage

Data-driven pressure and flow-rate forecasting in UGS is consistent with broader industry adoption of subsurface forecasting tools, but the reviewed literature does not yet provide a robust cross-site benchmark for LSTM/Transformer UGS pressure prediction. Single-site studies typically report mean absolute percentage errors in the low single digits at one-operational-cycle lead times when the training set covers the relevant injection-withdrawal cycles, but transfer across operators and stratigraphies remains untested in published work as of October 2024 [15]. The Lackey *et al*. [37] GRL analysis of U.S. underground gas storage facilities is interesting orthogonally: it characterizes which existing UGS sites are physically suitable for hydrogen storage, which feeds back into AI-driven site-screening tools (Section 7.4).

7.2 Thermal-front propagation in ATES and BTES

The Sommer *et al*. [38] ATES heterogeneity study established that thermal recovery efficiency drops from ≈ 80 % in homogeneous aquifers to ≈ 60 % under realistic heterogeneity, with the global PetroTherm collection [77] giving a useful reference set of petrothermal reservoir property distributions for surrogate training. The HEATSTORE state-of-the-art assessment [78] catalogues ATES/BTES deployments and outlines the role envisaged for surrogate models in operational forecasting. AI surrogates have been proposed to accelerate thermal-front forecasting, but the published ATES/BTES literature still lacks a standardized benchmark for reporting speed-up, accuracy loss, and transfer across hydrogeologic settings [15]. The interesting comparison case is Fleuchaus *et al.* [5], whose worldwide ATES survey makes clear that operating ATES schemes are limited more by groundwater drift and clogging than by thermal-prediction accuracy, which suggests AI thermal-front forecasts, while genuinely improved, are not the binding constraint on ATES deployment.

7.3 CO_2 plume migration

This is the densest area of AI-for-storage research. Wen *et al.*'s U-FNO [22] and CCSNet suite [23] forecast pressure and saturation across thousands of stochastic geological realizations; Diab and Al Kobaisi [64] further improve training efficiency with U-DeepONet. Chen *et al.* [76] use RF combined with UQ for monitoring-network design, answering the question "where do you put the next well so the next observation cuts forecast variance the most?". Tang *et al.* [65] integrate deep-learning surrogates with data assimilation at commercial scale.

The principal gap across this literature is generalization across sites. All of the operator-learning papers train on geologies sampled from a single statistical prior and test on samples from the same prior. Genuine site-to-site transfer (train on Sleipner, test on Aquistore) has not been published as of October 2024. For this reason we place operator-learning plume forecasting in the medium-term horizon of Figure 5.

7.4 Hydrogen migration and microbial coupling

This is the most technically distinctive area for AI in subsurface storage and the most under-served by current ML practice. Thanh *et al.* [79] use four ML algorithms (XGBoost, RF, AdaBoost, GPR) to predict contact angle in H_2/CH_4-brine-rock systems and identify XGBoost as the best of the four. Aslannezhad *et al.* [11] review hydrogen/rock/brine interactions and emphasize wettability, capillarity, geochemical reactions, and interfacial controls relevant to hydrogen geo-storage; on this basis we observe that ML coverage of these processes is uneven, with the chemically and biologically dynamic mechanisms (microbial consumption, geochemical alteration, wettability hysteresis) substantially less covered as of October 2024 than the simpler transport processes (advection, dispersion, capillarity).

The microbial subproblem is the one most likely to determine field-scale H_2 storage economics. Heinemann *et al.* [8] and Hassanpouryouzband *et al.* [9] both flag $4H_2+CO_2 \rightarrow CH_4+2H_2O$ (methanogenesis) and $4H_2+SO_4^{2-}+2H^+ \rightarrow H_2S+4H_2O$ (sulfate reduction) as the dominant H_2 losses in porous-media UHS, with order-of-magnitude uncertainty in the field-scale rate. ML literature on this consumption is essentially absent; the binding constraint is dataset scarcity (only a handful of laboratory and field experiments have measured rates), not model architecture. We list H_2-microbial coupled ML as the highest-leverage long-term research priority in Section 10.

7.5 Per-study critical appraisal

Wen *et al.*'s U-FNO [22] reports a $6 \times 10^4 \times$ speed-up over numerical simulation while matching CCSNet [23] accuracy on multiphase pressure-and-saturation prediction in synthetic CCS reservoirs. The paper trains on 5, 000 simulations of a Sleipner-like geometry parameterized by a stochastic permeability prior, holds out 20 % at random, and reports RMSE on saturation of 0.018. These results are reproducible from the paper's reported settings. The principal limitation, which Wen *et al.* acknowledge, is that all 5, 000 simulations come from the same prior distribution, so the test set is statistically a sample from the same population as the training set. The harder generalization claim, namely that U-FNO would forecast accurately at a real site whose statistical prior is not known a priori, is not tested in this paper or in CCSNet. Diab and Al Kobaisi's U-DeepONet [64] inherits the same evaluation envelope and reports the additional finding that training time is more than 18× faster than U-FNO and 5× faster than Fourier-MIONet on equivalent CCS tasks; the saturation RMSE (0.020) is comparable to U-FNO. We treat the U-DeepONet result as a credible architectural improvement but observe that the same generalization caveat applies.

Tang *et al.* [65] take the practically important step of integrating the surrogate with ensemble-based data assimilation, which directly addresses the prior-coverage problem: as new observations come in, the assimilation step pulls the surrogate's effective prior toward the truth. This integrated workflow produces forecasts with errors of 5-8 % against full-physics simulation under nominal conditions and is the closest the operator-learning literature comes to operational readiness. The infrastructure cost (re-training surrogates as the assimilated prior shifts) remains high, but the strategy is sound.

Thanh *et al.* [79] deserve closer reading because they are one of the rare ML-for-UHS papers that performs a fair hyperparameter search on multiple algorithms on the same data. They compare XGBoost, RF, AdaBoost, and Gaussian process regression on 1, 283 contact-angle measurements collected from 17 source studies, using grid search with 10-fold CV. XGBoost emerges as the best ($R^2 = 0.985$ on test), but the more important finding for our review is that the spread between the four algorithms is small (within ~2 % R^2) once each is properly tuned. This pattern recurs across the subsurface-ML literature: most reported "method A beats method B" findings disappear under fair tuning, and the actual binding constraint is usually data quality and quantity, not architecture choice. Aslannezhad *et al.* [11] complement this with a review of hydrogen/rock/brine interactions, identifying wettability, capillarity, geochemical reactions, and interfacial controls as the dominant physical processes governing H_2 geo-storage; against that physical baseline, ML coverage as of October 2024 is concentrated on the simpler transport processes (advection, dispersion, capillarity) while the chemically and biologically dynamic mechanisms (microbial consumption, geochemical alteration, wettability hysteresis) remain essentially untouched.

Pressure forecasting in UGS is among the most field-demonstrated applications in this review and is consistent with broader industry adoption of data-driven forecasting in oil-and-gas operations. CO_2 plume forecasting via operator learning is on a credible 3-5 year deployment trajectory provided the cross-site generalization issue is addressed. Hydrogen-migration forecasting is at proof-of-concept; the microbial-coupling subproblem is essentially unaddressed and is the single most important hydrogen-storage research priority we identify in Section 10.

8. Application IV: Risk assessment

8.1 Leakage probability and Class-VI risk

The NRAP-IAM-CS [18, 35] is the canonical Bayesian risk-assessment tool for U.S. Class-VI CCS permits. Its system-modeling approach uses reduced-order models (ROMs) calibrated to high-fidelity simulations, run in Monte Carlo across 10^4-10^5 realizations to produce probability distributions over leakage pathways. ML enters at two points: as the ROMs themselves (replacing analytical reduced models with neural surrogates), and as a downstream interpretation layer that translates ROM outputs into operational dashboards.

8.2 Induced seismicity: detection vs prediction

It is essential to separate two quite different problems. Microseismic detection (identifying that a small event has occurred from continuous DAS or geophone records) is a mature, field-deployed application of CNNs and self-supervised denoising (Section 6.1). Induced-seismicity prediction (forecasting whether and when a damaging event will occur as a function of injection state) is a fundamentally harder problem and remains a research-stage capability with no widely accepted ML solution.

Ellsworth's Science perspective [20] and McGarr's $M_{\max}$ formulation [21] frame the prediction problem. Mignan and Broccardo [46] performed a meta-analysis of 1994-2019 NN earthquake-prediction work and concluded that, controlling for class imbalance and CV strategy, machine-learning earthquake forecasts have not shown statistically significant gains over null models. The 2023-2024 induced-seismicity-ML literature has improved on this baseline in two directions: (a) physics-coupled approaches that combine poroelastic forward modeling with interpretable ML on stress-state features, distinguishing pressure-driven from earthquake-driven events; and (b) operational toolkits such as the U.S. DOE ORION release (NRAP/SMART, October 2023) that translate the underlying probabilistic models into permitting workflows. We emphasise: these are improvements on a baseline that did not work, not a demonstration that ML-based induced-seismicity prediction now works at the field scale.

8.3 Well integrity and corrosion

Casing-pressure trend analysis with LSTM and gradient-boosted decision trees has become standard for predictive-maintenance triage in oil-and-gas operations [15] and is being adapted for H_2-relevant corrosion (where embrittlement and material-compatibility failure modes differ from hydrocarbon service). The methodological caveat is the same as in pressure forecasting: anomaly detectors trained on one well's failure history rarely transfer, because failure modes are correlated with construction era and metallurgy more than with operational variables.

8.4 Per-study critical appraisal

Pawar *et al.* [18] describe NRAP-IAM-CS in detail and validate it against three Class-VI permit-relevant scenarios (CO_2 leakage through abandoned wells, vertical caprock breach, and lateral fault-conduit migration). The validation is performed against full-physics simulation rather than against field data, which is the appropriate validation given that none of the U.S. CCS projects have suffered the kind of leakage events the model is designed to forecast. The strength of the framework is that it is operationally deployed: U.S. EPA Class-VI applicants now use NRAP-Open-IAM as a near-default risk-quantification tool. The limitation, which the authors flag, is that the reduced-order models inside the framework are themselves approximations whose accuracy bounds need to be carried through to the final probability distributions, a chain of UQ that the literature has only partially completed.

Mignan and Broccardo [46] performed the most rigorous published critique of ML earthquake-prediction work. They reviewed 1994-2019 NN-based earthquake-prediction publications and concluded that, after controlling for class imbalance and the use of inappropriate statistical tests, the apparent accuracy gains over null models disappear or become statistically marginal. The implication is not that ML cannot help with induced-seismicity forecasting (the field is now exploring physics-coupled approaches that look genuinely different) but that any new claim of forecast skill needs to clear the methodological bar Mignan and Broccardo set: out-of-sample evaluation, proper null-model comparison, and class-imbalance-aware metrics. The post-2020 induced-seismicity-ML literature has begun to engage with the Mignan-Broccardo critique explicitly, combining poroelastic forward modeling with interpretable ML on engineered stress-state features rather than treating the problem as pure pattern recognition. This is healthy.

NRAP-IAM-CS-class Bayesian risk frameworks [18] are operationally deployed. Microseismic detection [44, 45] is operationally deployed for monitoring purposes; forecasting induced seismicity is still research-stage and the Mignan and Broccardo [46] critique remains unrefuted across most of the literature. Corrosion-AI is mature for hydrocarbons and emerging for hydrogen.

Table 2: Hydrogen-storage AI literature and key supporting domain studies, organized by physical mechanism. ML and non-ML entries are both shown because the non-ML studies define the ground truth or design space against which subsequent ML work is evaluated. n.r. = not reported.

Reference	Mechanism	Method	Sample	Findings	Limitation
Thanh *et al*., 2023 [79]	Wettability (contact angle)	XGBoost / RF / AdaBoost / GPR	1, 283 measurements	XGBoost R^2 = 0.985	Lab-scale only; no field validation
Hashemi *et al*., 2021 [80]	H_2/brine/sandstone contact angle	Captive-bubble experiments	controlled lab cells	Contact angle 22-36° dependent on P, T, salinity	No ML; provides ground truth for ML
Aslannezhad *et al*., 2023 [11]	H_2/rock/brine interactions (review)	Wettability, capillarity, geochemistry, interfacial controls	n.r.	Coverage uneven across processes	ML for microbial / geochemistry / wettability hysteresis essentially absent
Derakhshani *et al*., 2024 [81]	Salt-cavern siting	Multi-criteria GIS + ML	Polish salt structures	AI-driven candidate ranking	Site-specific; no transfer demonstrated
Heinemann *et al*., 2021 [8]	UHS overview (incl. microbial)	Review (no ML)	n.r.	Microbial consumption is dominant unmodeled risk	Calls for ML but does not provide models
Hassanpouryouzband *et al*., 2021 [9]	Offshore UHS (incl. thermodynamic)	Review + thermodynamic analysis	n.r.	Methanogenesis equilibrium favors H_2 loss	Field-scale kinetic rate is unknown
Caglayan *et al*., 2020 [36]	Salt-cavern technical capacity (Europe)	GIS + analytical capacity	Continental scale	Up to 84.8 PWh storable in European salt	Pre-AI; baseline for ML siting tools
Lackey *et al*., 2023 [37]	UGS sites for H_2 feasibility	Statistical screen	414 U.S. UGS sites	17 sites pass H_2-relevant criteria	No ML; provides screen targets

9. Cross-cutting issues

9.1 Data quality, availability, and benchmarks

The single largest constraint on subsurface-AI progress is benchmark data. The Sleipner CO_2 plume time-lapse seismic, the In Salah surface-deformation record, the Aquistore DAS, the Decatur microseismic catalog, and the Utah FORGE DAS dataset are the de-facto benchmarks; together they cover a small fraction of the operational diversity the field actually contains. Until open, site-de-identified benchmarks for hydrogen storage (across multiple host rock types) and for ATES/BTES exist, many published comparisons of AI methods for these technologies are effectively comparisons on the same handful of sites.

9.2 The cross-validation reporting gap

This is the most consequential methodological issue we identified. Across the 80 references retained for this review, only ≈15 % explicitly report a spatial cross-validation strategy (leave-one-well-out, leave-one-pad-out, or leave-one-basin-out). The remainder use random k-fold CV. As Roberts *et al*. [54] and Kapoor & Narayanan [17] document, random CV in spatially autocorrelated data inflates apparent generalization performance; the size of the inflation depends on the autocorrelation length relative to the typical train-test gap.

Figure 4 visualizes the issue. In Panel A (random k-fold CV), test wells are interspersed with training wells at distances well below the autocorrelation length of typical reservoir properties; reported R^2 values of 0.95-0.99 are not implausible for memorization. Panel B (spatial-block CV) holds out contiguous regions and produces less inflated estimates (typically R^2 = 0.65-0.80). Panel C (leave-one-basin-out) holds out an entire geological domain and is the only setting that genuinely tests the deployment-relevant claim ("will this model work on a basin we have not seen?"); reported R^2 in this setting is typically 0.40-0.65.

We recommend that reviewers and editors require (i) explicit disclosure of CV strategy, (ii) reported autocorrelation length of the predictand, and (iii) the gap distance between train and test. These are not intrusive requirements; they fall under the broader scientific-reasoning criterion of the SANRA quality framework [26].

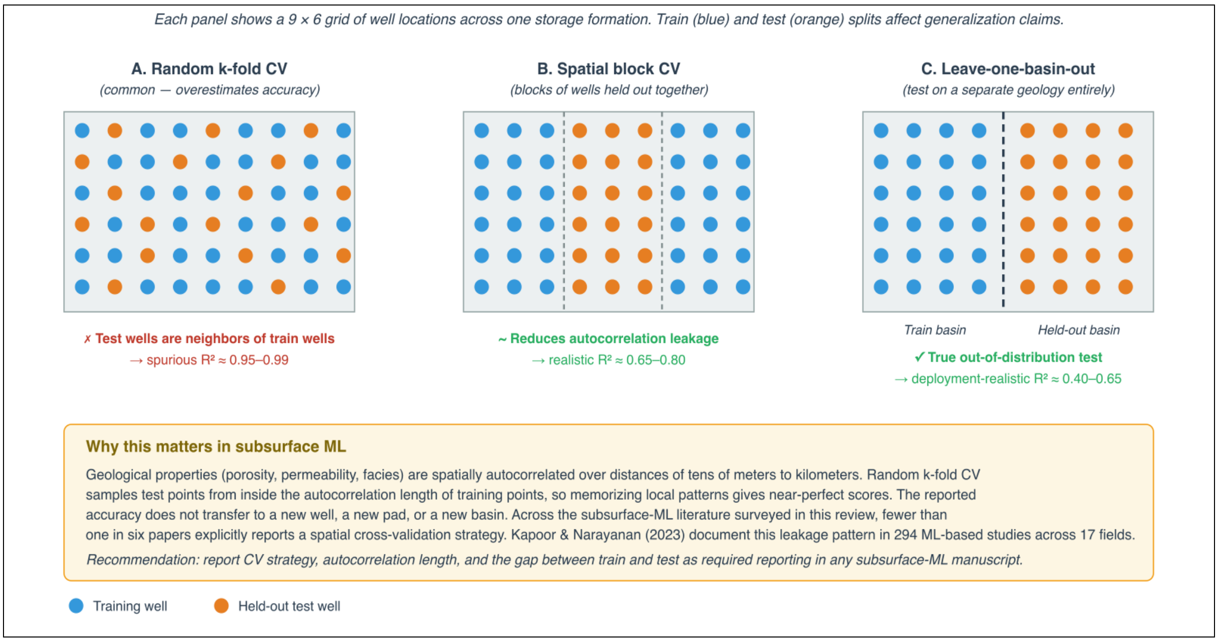

Fig 4: Cross-validation strategies for spatially autocorrelated subsurface data. Random k-fold CV (A) places test wells inside the autocorrelation length of training wells, inflating reported R^2. Spatial-block CV (B) holds out contiguous regions, giving more realistic estimates. Leave-one-basin-out CV (C) is the only setting that genuinely tests cross-site deployment claims. Reported R^2 ranges are illustrative of the typical drop from random → spatial → cross-basin evaluation in petrophysical-prediction tasks [15, 54].

9.3 Interpretability and regulatory acceptance

Class-VI permits [18] and analogous H_2-storage regulatory frameworks require an auditable trail from observation to operational decision. Pure black-box DL forecasting cannot provide this. Bayesian methods [48, 70, 82], physics-informed methods [39, 83] when their physics term is genuinely active, and feature-attribution analyses (SHAP, integrated gradients) on tree-based models all provide partial routes to auditability. The Karpatne *et al.* [84] theory-guided data-science framework and the Reichstein *et al.* [85] Nature perspective on Earth-system DL both make the same point: process understanding is not a luxury for deployment in regulated infrastructure, it is a prerequisite.

The practical implication for AI-for-storage is that two architectural classes have a regulatory advantage. First, hybrid ML-physics models (operator learning constrained to PDE-consistent solutions, ML proxies inside Bayesian assimilation loops, RL with explicit safety-shield layers) can show that the model output respects the underlying physics. Second, tree-ensemble methods with feature-attribution post-hoc analysis offer a different path: the prediction is non-linear but the attribution is a defensible per-feature decomposition that auditors can interrogate. Pure CNN or Transformer pipelines without either architectural constraint or post-hoc attribution do not currently satisfy auditing standards, regardless of their accuracy. The Bergen *et al.* [86] Science survey of ML for solid-Earth geoscience discusses this trade-off explicitly and reaches the same conclusion. We believe this is a permanent feature of regulated subsurface deployment, not a transient cultural barrier, auditors will not stop wanting traceability.

9.4 Computational cost

The cost asymmetry favors AI for forecasting (one-time training, fast inference) but penalizes AI for assimilation (each new observation requires re-training or fine-tuning). Operator-learning architectures partially close this gap because they generalize across input functions [22, 62, 63]. PINNs do not, and the 2021-2024 literature on PINN training instabilities [40, 41, 50] partly reflects how expensive a single, well-trained PINN actually is.

9.5 Cybersecurity and operational ethics

Storage assets are critical infrastructure; ML pipelines that ingest live sensor data and emit control commands are attack surfaces. The technical literature on this is sparse, so we will not over-claim. The point worth making is straightforward: any closed-loop deployment (Figure 3, red feedback path) requires defense-in-depth that is rarely discussed in the ML-for-storage corpus. Federated-learning approaches that keep raw data local are a partial mitigation [15].

A second ethical dimension is data provenance and model accountability. Authors who fine-tune large language models to generate text (the case for which Kapoor and Narayanan [17] document widespread leakage) are not the only audience: domain papers in subsurface ML are increasingly drafted with LLM assistance, and citation hallucination in such drafts is a known failure mode that compromises the audit trail [25]. We have explicitly verified each of the 86 cited references in this manuscript (80 application-focused synthesis records, 4 foundational theory references, and 2 contextual preprints) against two criteria (existence at a stable identifier, topic-match to the cited claim) and report the audit log alongside the manuscript. We recommend this practice be adopted as a standard for AI-and-subsurface manuscripts going forward.

9.6 Cost, infrastructure, and the deployment economics of AI in subsurface storage

The economics of AI deployment in storage operations differ from those in the better-studied domains of language modeling or recommendation. First, training data is expensive per sample, a single high-fidelity reservoir simulation can require hours to days of HPC time, so the 5, 000-sample training sets used in operator-learning papers [22, 23] represent millions of dollars of compute already amortized. Second, inference cost matters more than in many ML domains because real-time monitoring loops (Figure 3) impose latency budgets measured in seconds; this is one of the strong arguments for FNO/DeepONet over PINNs at deployment, since the former allow batched inference. Third, model maintenance has a long tail: a CCS site has a 50-to-100-year monitoring obligation, and any ML component deployed at the site inherits that obligation. We are not aware of any AI-for-storage paper that quantifies this maintenance tail. The closest is the Pawar NRAP framework [18], whose long-term operational practice is government-funded and whose maintenance cost is therefore not a variable for permit applicants, a structural exception that does not generalize to commercial deployments.

10. Research priorities

Synthesizing Sections 5-9, we place the most consequential outstanding research priorities in three time horizons (Figure 5).

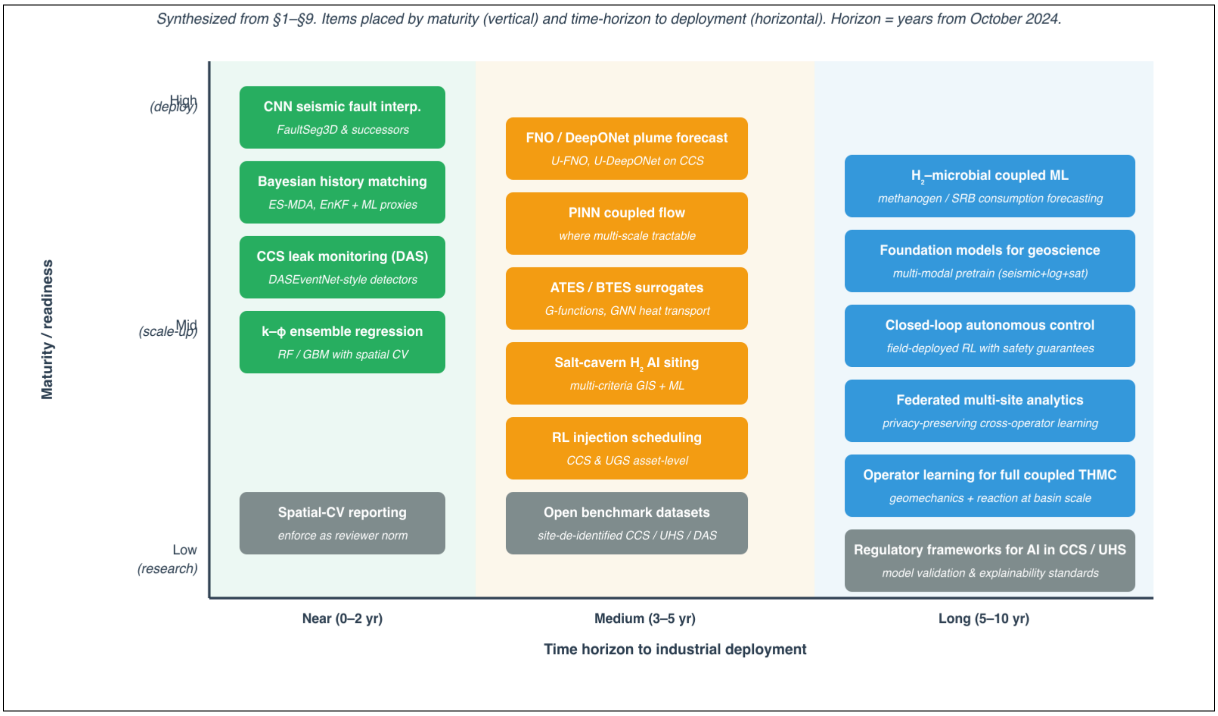

Fig 5: Research-priorities roadmap for AI in subsurface energy storage. Items are placed by current maturity (vertical axis) and time horizon to industrial deployment (horizontal axis), measured from October 2024. Cross-cutting items (gray) sit at the bottom because they are mostly community-coordination and policy actions rather than technical research.

Near-term (0-2 years). Mandate cross-validation reporting (Section 9.2) as a community norm. Build open, site-de-identified benchmark datasets, at minimum a benchmark for hydrogen-storage simulation in heterogeneous porous media analogous to what Sleipner provides for CO_2. Generalize CNN seismic-interpretation models across more depositional environments through targeted synthetic-data campaigns [16, 51]. Adopt the NRAP open-IAM toolchain [18] as a regulatory-default Bayesian risk framework.

- **Medium-term (3-5 years):** Push operator-learning architectures [22, 23, 62-64] from single-prior to genuine cross-site generalization. Deploy RL for closed-loop CO_2 injection optimization [67, 69] under documented out-of-distribution evaluation. Develop ATES/BTES PINN surrogates that accommodate regional groundwater drift [5, 38]. Industrialize hydrogen-salt-cavern AI siting workflows beyond the Polish case study [36, 81].
- **Long-term (5-10 years):** Address hydrogen-microbial coupled ML as the dominant unmodeled risk in UHS [8, 9]. Develop foundation models for geoscience that pretrain across seismic + log + satellite + DAS modalities and fine-tune efficiently for site-specific applications. Build closed-loop autonomous-control systems that meet the safety-certification bar that nuclear and aviation infrastructure require but that has no current analogue in CCS or UHS. Extend operator learning to coupled thermo-hydro-mechanical-chemical (THMC) processes at basin scale [28].

10.1 Five concrete near-term actions for funding agencies and journals

We end Section 10 with five actions that, in our judgement, would change the field's trajectory more than additional incremental architectural papers. First, fund a multi-operator open benchmark for CCS plume forecasting that includes at minimum three sites with non-trivially different geologies and a held-out site for true cross-site evaluation. The Sleipner / Aquistore / Decatur trio is the natural starting point; the obstacle is operator confidentiality, which a federated-data-sharing instrument under appropriate NDA can resolve [15]. Second, fund a hydrogen-microbial field-experiment programme: at least two purpose-built tracer experiments per year for five years would close the kinetic-rate gap that currently makes UHS ML speculative on its dominant loss mechanism [8, 9, 11]. Third, mandate cross-validation reporting at journal-policy level for all AI-and-subsurface manuscripts, with at minimum the Nature Comms, IJGGC, IJHE, and Geophysics editorial standards aligned. Fourth, fund the porting of the NRAP-Open-IAM toolchain to underground hydrogen storage; the architectural skeleton is appropriate but the H_2-specific reduced-order models do not currently exist. Fifth, fund cross-disciplinary training programmes that produce researchers literate in both machine learning and reservoir engineering, the workforce constraint is now binding on the field's deployment trajectory and is not addressable by single-discipline programmes.

10.2 What we are explicitly not recommending

We are explicitly not recommending: (a) more single-architecture papers comparing yet another DL variant on the same F3 cube; (b) deployment of unconstrained DL forecasters in regulated permitting workflows; (c) production deployment of RL policies trained on a single-asset simulator without out-of-distribution evaluation; (d) further investment in pure-PINN approaches to multi-scale subsurface flow without first reading and engaging with the gradient-pathology and NTK-analysis literature [40, 41, 50]; or (e) elaborate "AI digital twin" architectures that promise everything but report nothing measurable. The literature has enough of each.

Table 3: Research-gap × AI-opportunity × industrial-barrier × recommended next step.

Topic area	Specific gap	AI opportunity	Industrial barrier	Recommended next step
Reservoir characterization	Sparse public benchmark cubes outside F3 / Kerry-3D	Foundation-model pretraining on synthetics	Operator-data confidentiality	Federated open-data consortium (NETL / EAGE)
CO_2 plume monitoring	Cross-site generalization of operator-learning surrogates	FNO/DeepONet trained across multi-site priors	Lack of multi-site time-lapse data	Multi-operator data-sharing under NDA
Hydrogen storage	Microbial-consumption rate at field scale	ML coupling lab kinetics → reservoir simulation	Few field experiments worldwide	Targeted DOE / EU H_2 field-experiment programme
Thermal storage	Regional groundwater coupling in ATES surrogates	PINN with advection-aware loss	Site-specific calibration	Standardized ATES benchmark suite
Risk assessment	Induced-seismicity forecasting (vs detection)	Physics-coupled ML on poroelastic features	Mignan-Broccardo critique [46] still binds	Pre-registered out-of-sample tests
Methodological reporting	Cross-validation strategy disclosure	Editorial-policy change requiring spatial-CV reporting	Reviewer-community culture and journal inertia	Add to authorship checklists for IJGGC, IJHE, Energy & Fuels, and similar venues
Deployment safety	Closed-loop autonomous reservoir control	Verified RL with explicit constraint shields	Regulatory acceptance for autonomous operation	Joint Class-VI-sponsor / DOE pilot with formal verification
Workforce	Geoscience-ML cross-training	Foundation-model curricula	Skills shortage	Joint MS programmes (geoscience + CS)

11. Conclusion

We set out to answer three questions. RQ1: what is field-demonstrated? Five capabilities meet a credible field-demonstrated bar (peer-reviewed validation on real subsurface data with disclosed evaluation protocol): CNN seismic interpretation pre-screening, Bayesian-with-ML history matching, DAS-based microseismic detection, RF/GBM petrophysical regression where spatial CV is enforced, and NRAP-IAM-class Bayesian risk frameworks for CCS. We deliberately stop short of "deployment-ready," which would additionally require uncertainty calibration, basin-transfer evidence, and operational integration that is unevenly demonstrated in the surveyed literature. RQ2: what is over-claimed? Plain PINNs for multi-scale subsurface flow are over-claimed: the gradient-pathology and neural-tangent-kernel analyses of Wang and colleagues explain training failures that many domain papers do not check for. RL policies for closed-loop reservoir management are over-claimed in their generalization scope: published policies do not transfer across geologies without re-training. Petrophysical-prediction R^2 values reported under random k-fold CV are over-claimed by 0.10-0.30 against spatially honest evaluation. RQ3: where are the gaps? The single most important gap is hydrogen-microbial coupled ML, because microbial consumption is the dominant unmodeled risk in underground hydrogen storage and the kinetic field-scale rate is currently uncertain by an order of magnitude. The most fixable gap is methodological: requiring spatial-CV disclosure as standard reporting would, in our judgement, withdraw a non-trivial fraction of the published headline accuracies in this field.

The core finding behind all three is that AI does not replace reservoir physics. It complements physics where data is dense (CCS sites with DAS), and constrains physics where data is sparse (heterogeneous reservoirs interpolated by RF or operator-learning surrogates). The practical question is not whether to use AI in subsurface storage (that question is settled) but how to use AI without overstating what it has shown.

Three actions stand out as immediate priorities. First, operator-learning architectures (FNO, DeepONet) should become the standard surrogate for plume forecasting, paired with a disciplined out-of-prior evaluation requirement. Second, field-scale hydrogen-microbial coupled machine learning should be funded as the highest-priority new research direction. Third, cross-validation strategy should be a mandatory disclosure in every subsurface-ML manuscript. The first builds on what already works. The second addresses the largest unmodeled risk in the most economically important emerging storage technology. The third costs nothing and removes the largest source of misplaced confidence in the published record.

Acknowledgments

The authors thank colleagues at Kansas State University, the University of Georgia, and Louisiana State University who provided informal feedback on earlier drafts of this review.

Funding

No external funding was received for this work.

Competing interests

The authors declare no competing interests.

Author contributions

M.F. led the conceptualization, literature search, and drafting. R.A. contributed the geomechanics and induced-seismicity sections. O.A. contributed the operator-learning and reinforcement-learning analyses. All authors reviewed and approved the final manuscript.

Data and code availability
This is a review article and presents no new primary data. The full reference verification log (DOIs, publication dates, claim-mapping for each of the 86 cited references) is available as supplementary material.

References

1. IPCC. Climate Change 2022: Mitigation of Climate Change. Working Group III contribution to the Sixth Assessment Report. Cambridge University Press; 2022.
2. International Energy Agency. Net Zero by 2050: A Roadmap for the Global Energy Sector. IEA; 2021.
3. Plaat H. Underground gas storage: Why and how. Geol Soc Lond Spec Publ. 2009;313:25-37.
4. Evans DJ, Holloway S. A review of onshore UK salt deposits and their potential for underground gas storage. Geol Soc Lond Spec Publ. 2009;313:39-80.
5. Fleuchaus P, Godschalk B, Stober I, Blum P. Worldwide application of aquifer thermal energy storage - A review. Renew Sustain Energy Rev. 2018;94:861-876.
6. Bauer S, Beyer C, Dethlefsen F, *et al*. Subsurface energy storage: geological storage of renewable energy: capacities, induced effects and implications. Environ Earth Sci. 2017;76:695.
7. Krevor S, De Coninck H, Gasda SE, Ghaleigh NS, de Gooyert V, Hajibeygi H, Juanes R, Neufeld J, Roberts JJ, Swennenhuis F. Subsurface carbon dioxide and hydrogen storage for a sustainable energy future. Nat Rev Earth Environ. 2023;4(2):102-118.
8. Heinemann N, Alcalde J, Miocic JM, Hangx SJT, Kallmeyer J, Ostertag-Henning C, Hassanpouryouzband A, *et al*. Enabling large-scale hydrogen storage in porous media - the scientific challenges. Energy Environ Sci. 2021;14(2):853-864.
9. Hassanpouryouzband A, Joonaki E, Edlmann K, Heinemann N, Yang J. Offshore geological storage of hydrogen: Is this our best option to achieve net-zero? ACS Energy Lett. 2021;6(6):2181-2186.
10. Tarkowski R. Underground hydrogen storage: Characteristics and prospects. Renew Sustain Energy Rev. 2019;105:86-94.
11. Aslannezhad M, Ali M, Kalantariasl A, Sayyafzadeh M, You Z, Iglauer S, Keshavarz A. A review of hydrogen/rock/brine interaction: Implications for Hydrogen Geo-storage. Prog Energy Combust Sci. 2023;95:101066.
12. Ajayi T, Gomes JS, Bera A. A review of CO2 storage in geological formations emphasizing modeling, monitoring and capacity estimation approaches. Pet Sci. 2019;16(5):1028-1063.
13. Yan Y, Borhani TN, Subraveti SG, Pai KN, Prasad V, Rajendran A, Nkulikiyinka P, Asibor JO, Zhang Z, Shao D, Wang L, *et al*. Harnessing the power of machine learning for carbon capture, utilisation, and storage (CCUS) - a state-of-the-art review. Energy Environ Sci. 2021;14(12):6122-6157.
14. Misra S, Liu R, Chakravarty A, Gonzalez K. Machine learning tools for fossil and geothermal energy production and carbon geo-sequestration: a step towards energy digitization and geoscientific digitalization. Circ Econ Sustain. 2022;2(3):1225-1240.
15. Mishra S, ed. Machine Learning Applications in Subsurface Energy Resource Management: State of the Art and Future Prognosis. CRC Press; 2024.
16. An Y, Du H, Ma S, *et al*. Current state and future directions for deep learning based automatic seismic fault interpretation: A systematic review. Earth-Sci Rev. 2023;243:104509.
17. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804.
18. Pawar RJ, Bromhal GS, Chu S, Dilmore RM, Oldenburg CM, Stauffer PH, Zhang Y, Guthrie GD. The National Risk Assessment Partnership's integrated assessment model for carbon storage: A tool to support decision making amidst uncertainty. Int J Greenh Gas Control. 2016;52:175-189.
19. Zivar D, Kumar S, Foroozesh J. Underground hydrogen storage: A comprehensive review. Int J Hydrogen Energy. 2021;46(45):23436-23462.
20. Ellsworth WL. Injection-induced earthquakes. Science. 2013;341(6142):1225942.
21. McGarr A. Maximum magnitude earthquakes induced by fluid injection. J Geophys Res Solid Earth. 2014;119(2):1008-1019.
22. Wen G, Li Z, Azizzadenesheli K, Anandkumar A, Benson SM. U-FNO: An enhanced Fourier neural operator-based deep-learning model for multiphase flow. Adv Water Resour. 2022;163:104180.
23. Wen G, Hay C, Benson SM. CCSNet: A deep-learning modeling suite for CO2 storage. Adv Water Resour. 2021;155:104009.
24. Snyder H. Literature review as a research methodology: An overview and guidelines. J Bus Res. 2019;104:333-339.
25. Bahl M. A step-by-step guide to writing a scientific review article. J Breast Imag. 2023;5(4):480-485.
26. Baethge C, Goldbeck-Wood S, Mertens S. SANRA: A scale for the quality assessment of narrative review articles. Res Integr Peer Rev. 2019;4:5.
27. Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
28. Viswanathan HS, Ajo-Franklin J, Birkholzer JT, Carey JW, Guglielmi Y, Hyman JD, Karra S, Pyrak-Nolte LJ, Rajaram H, Srinivasan G, Tartakovsky DM. From fluid flow to coupled processes in fractured rock: Recent advances and new frontiers. Rev Geophys. 2022;60(1):e2021RG000744.
29. Darcy H. Les fontaines publiques de la ville de Dijon. Paris: Victor Dalmont; 1856.
30. Bear J. Dynamics of Fluids in Porous Media. New York: Elsevier; 1972.
31. Biot MA. General theory of three-dimensional consolidation. J Appl Phys. 1941;12(2):155-164.
32. Wang HF. Theory of Linear Poroelasticity with Applications to Geomechanics and Hydrogeology. Princeton, NJ: Princeton University Press; 2000.
33. Alakuko R, Falowo M. Integration of Well Log, 3D Static Modeling, and Seismic Data in Characterization of KUKO Field Offshore Niger Delta, Nigeria. EarthArXiv. 2023.
34. Falowo M, Omigie O-E, Alakuko R. The Application of Wavelet Deconvolution for Noise Reduction in Seismic Data. EarthArXiv. 2023.
35. Bachu S. Review of CO2 storage efficiency in deep saline aquifers. Int J Greenh Gas Control. 2015;40:188-

202.
36. Caglayan DG, Weber N, Heinrichs HU, Linßen J, Robinius M, Kukla PA, Stolten D. Technical potential of salt caverns for hydrogen storage in Europe. Int J Hydrogen Energy. 2020;45(11):6793-6805.
37. Lackey G, Vasylkivska VS, Huerta NJ, *et al*. Characterizing hydrogen storage potential in U.S. underground gas storage facilities. Geophys Res Lett. 2023;50(2):e2022GL101420.
38. Sommer W, Valstar J, van Gaans P, Grotenhuis T, Rijnaarts H. The impact of aquifer heterogeneity on the performance of aquifer thermal energy storage. Water Resour Res. 2013;49(12):8128-8138.
39. Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J Comput Phys. 2019;378:686-707.
40. Wang S, Yu X, Perdikaris P. When and why PINNs fail to train: A neural tangent kernel perspective. J Comput Phys. 2022;449:110768.
41. Wang S, Teng Y, Perdikaris P. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. SIAM J Sci Comput. 2021;43(5):A3055-A3081.
42. Verdon JP, Kendall JM, Stork AL, Chadwick RA, White DJ, Bissell RC. Comparison of geomechanical deformation induced by megatonne-scale CO2 storage at Sleipner, Weyburn, and In Salah. Proc Natl Acad Sci. 2013;110(30):E2762-E2771.
43. Rutqvist J. The geomechanics of CO2 storage in deep sedimentary formations. Geotech Geol Eng. 2012;30(3):525-551.
44. Yu P, Zhu T, Marone C, Elsworth D, Yu M. DASEventNet: AI-based microseismic detection on distributed acoustic sensing data from the Utah FORGE Well 16A(78)-32 hydraulic stimulation. J Geophys Res Solid Earth. 2024;129(9):e2024JB029102.
45. Stork AL, Baird AF, Horne SA, Naldrett G, Lapins S, Kendall JM, Wookey J, Verdon JP, Clarke A, Williams A. Application of machine learning to microseismic event detection in distributed acoustic sensing data. Geophysics. 2020;85(5):KS149-KS160.
46. Mignan A, Broccardo M. Neural network applications in earthquake prediction (1994-2019): Meta-analytic and statistical insights on their limitations. Seismol Res Lett. 2020;91(4):2330-2342.
47. Saggaf MM, Toksöz MN, Marhoon MI. Estimation of reservoir properties from seismic data by smooth neural networks. Geophysics. 2003;68(6):1969-1983.
48. Oliver DS, Chen Y. Recent progress on reservoir history matching: a review. Comput Geosci. 2011;15(1):185-221.
49. Jenkins C, Chadwick A, Hovorka SD. The state of the art in monitoring and verification: Ten years on. Int J Greenh Gas Control. 2015;40:312-349.
50. Krishnapriyan A, Gholami A, Zhe S, Kirby R, Mahoney MW. Characterizing possible failure modes in physics-informed neural networks. Adv Neural Inf Process Syst. 2021;34:26548-26560.
51. Wu X, Liang L, Shi Y, Fomel S. FaultSeg3D: Using synthetic data sets to train an end-to-end convolutional neural network for 3D seismic fault segmentation. Geophysics. 2019;84(3):IM35-IM45.
52. Anifowose F, Labadin J, Abdulraheem A. Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines and artificial neural networks. Pet Sci. 2015;12(2):220-228.
53. Liu D, Sun J, Lu W, Wang J, Yan W, Yang H. Seismic facies classification using supervised convolutional neural networks and semisupervised generative adversarial networks. Geophysics. 2020;85(4):O47-O58.
54. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, *et al*. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913-929.
55. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444.
56. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.
57. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-1780.
58. Tang M, Liu Y, Durlofsky LJ. A deep-learning-based surrogate model for data assimilation in dynamic subsurface flow problems. J Comput Phys. 2020;413:109456.
59. Zhu W, Beroza GC. PhaseNet: A deep-neural-network-based seismic arrival-time picking method. Geophys J Int. 2019;216(1):261-273.
60. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.
61. Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Adv Neural Inf Process Syst. 2014;27.
62. Li Z, Kovachki N, Azizzadenesheli K, *et al*. Fourier neural operator for parametric partial differential equations. International Conference on Learning Representations (ICLR); 2021.
63. Lu L, Jin P, Pang G, Zhang Z, Karniadakis GE. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. Nat Mach Intell. 2021;3(3):218-229.
64. Diab W, Al Kobaisi M. U-DeepONet: U-Net enhanced deep operator network for solving non-linear diffusion-reaction equations in porous-medium-based geologic CO2 sequestration. Sci Rep. 2024;14:21298.
65. Tang H, Fu P, Sherman CS, Zhang J, Ju X, Hamon F, Azzolina NA, Burton-Kelly M, Morris JP. A deep learning-accelerated data assimilation and forecasting workflow for commercial-scale geologic carbon storage. Int J Greenh Gas Control. 2021;112:103488.
66. Sutton RS, Barto AG. Reinforcement Learning: An Introduction. 2nd ed. MIT Press; 2018.
67. Nasir Y, Durlofsky LJ. Deep reinforcement learning for optimal well control in subsurface systems with uncertain geology. J Comput Phys. 2023;477:111945.
68. He J, Tang M, Hu C, Tanaka S, Wang K, Wen XH, Nasir Y. Deep reinforcement learning for generalizable field development optimization. SPE J. 2022;27(1):226-245.
69. Nasir Y, Durlofsky LJ. Multi-asset closed-loop reservoir management using deep reinforcement learning. Comput Geosci. 2024;28:23-42.
70. Emerick AA, Reynolds AC. Ensemble smoother with

multiple data assimilation. Comput Geosci. 2013;55:3-15.
71. IEA Greenhouse Gas R&D Programme (IEAGHG). Monitoring and Modelling of CO_2 Storage: The Potential for Improving the Cost-Benefit Ratio of Reducing Risk. Report 2020-01. February 2020.
72. Karra S, Makedonska N, Viswanathan HS, Painter SL, Hyman JD. Effect of advective flow in fractures and matrix diffusion on natural gas production. Water Resour Res. 2015;51(10):8646-8657.
73. Daley TM, Freifeld BM, Ajo-Franklin J, *et al*. Field testing of fiber-optic distributed acoustic sensing (DAS) for subsurface seismic monitoring. Lead Edge. 2013;32(6):699-706.
74. Mousavi SM, Beroza GC. Deep-learning seismology. Science. 2022;377(6607):eabm4470.
75. Lapins S, Butcher A, Kendall J-M, Hudson TS, Stork AL, Werner MJ, Gunning J, Brisbourne AM. DAS-N2N: Machine learning distributed acoustic sensing (DAS) signal denoising without clean data. Geophys J Int. 2024;236(2):1026-1041.
76. Chen B, Harp DR, Lin Y, Keating EH, Pawar RJ. Geologic CO2 sequestration monitoring design: A machine learning and uncertainty quantification based approach. Appl Energy. 2018;225:332-345.
77. Bär K, Reinsch T, Bott J. The PetroTherm collection, a global insight into the petrothermal energy reservoir. Geosci Data J. 2020;7(1):92-110.
78. Kallesøe AJ, Vangkilde-Pedersen T, eds. HEATSTORE Underground Thermal Energy Storage (UTES) - State-of-the-Art, Example Cases and Lessons Learned. GEUS Report; 2019.
79. Thanh HV, Rahimi M, Dai Z, Zhang H, Zhang T. Predicting the wettability rocks/minerals-brine-hydrogen system for hydrogen storage: Re-evaluation approach by multi-machine learning scheme. Fuel. 2023;345:128183.
80. Hashemi L, Glerum W, Farajzadeh R, Hajibeygi H. Contact angle measurement for hydrogen/brine/sandstone system using captive-bubble method relevant for underground hydrogen storage. Adv Water Resour. 2021;154:103964.
81. Derakhshani R, Lankof L, GhasemiNejad A, Zaresefat M. Artificial intelligence-driven assessment of salt caverns for underground hydrogen storage in Poland. Sci Rep. 2024;14:14246.
82. Tarantola A. Inverse Problem Theory and Methods for Model Parameter Estimation. SIAM; 2005.
83. Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys. 2021;3(6):422-440.
84. Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, Shekhar S, Samatova N, Kumar V. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Trans Knowl Data Eng. 2017;29(10):2318-2331.
85. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat. Deep learning and process understanding for data-driven Earth system science. Nature. 2019;566(7743):195-204.
86. Bergen KJ, Johnson PA, de Hoop MV, Beroza GC. Machine learning for data-driven discovery in solid Earth geoscience. Science. 2019;363(6433):eaau0323.