https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Open Access Earth System Science Data Discussions

A national inventory of well lithology and hydraulics for the conterminous United States

Vahid Rafiei[1]

[1]Independent Researcher, ORCID: 0009-0009-8309-1895

Correspondence: Vahid Rafiei (vahidr32@gmail.com)

Abstract. Continental and coupled groundwater models require a three-dimensional picture of the subsurface — the vertical sequence of geologic materials and their hydraulic properties with depth — yet no harmonized national compilation of this information has existed for the United States, where it is generated as a regulatory byproduct in millions of water-well driller's logs held by individual state agencies in heterogeneous, partly-digitized forms. We present an open, harmonized inventory of well lithology and hydraulics for the 48 conterminous states, compiled entirely from primary state-agency sources through public endpoints with no privileged access. The inventory brings together ∼**28.5 million depth-resolved lithology intervals for** ∼**5.2 million wells across 34 states**, harmonized to a single canonical schema (well identifier, interval top and bottom depth, verbatim material description, a derived normalized material class, and per-row source provenance); structured hydraulics — static water level, total depth, yield, drawdown, specific capacity, screened interval, and, where reported, transmissivity — for ∼**6.9 million wells** in 41 states; and coordinates for ∼**7.8 million wells** spanning all 48 conterminous states, with the agency's own precision qualifier attached. Every value is reported as published by the source agency, with no interpolation, gap-filling, or modeled substitution: release quality assurance nulls only documented agency filler codes, and flags — never alters — implausible or internally inconsistent values. Every record is traceable to its agency source. We document per-state coverage and the exact access pattern for every source, so the compilation is fully reproducible and extensible; a companion availability census classifies each of the ∼13.3 million registered wells by how — or whether — its lithology is published. Where a state exposes lithology only as scanned driller's-log images, we recover it with a reproducible, cost-bounded vision-language procedure validated against a state agency's hand-transcriptions (91.5% material-class accuracy, Cohen's $\kappa = 0.90$; 96.0% coarse/fine texture, $\kappa = 0.92$). The inventory is released openly on Zenodo and is browsable through a live web platform. It supplies the depth-resolved subsurface-structure and hydraulics layer that national well-location databases omit, ready for use in groundwater modeling, aquifer characterization, and data-driven hydrogeology.

1 Introduction

Continental and coupled groundwater models require the subsurface's vertical structure — the sequence of geologic materials and their hydraulic properties with depth. In the United States that information is generated as a regulatory byproduct: when a well is drilled, the driller records a log of the materials encountered. Well over ten million such logs exist in state custody (our census indexes ∼13.3 million digitally registered wells alone, before paper archives), but they are held by individual state

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

agencies in heterogeneous, often non-machine-readable forms. Interoperability standards such as GroundWaterML2 exist to bridge these schema differences (Brodaric et al., 2018), but a standard cannot substitute for the data itself: no harmonized national compilation of their lithology exists. Every regional model therefore re-digitizes its basin's driller's logs by hand. Our primary contribution is that compilation itself: an open, national, depth-resolved inventory of well lithology and hydraulics for the conterminous United States, harmonized from primary state-agency sources to one canonical schema, reported strictly as published (no imputation), and released openly with full per-source provenance so it is reproducible and extensible. Two further contributions support and extend it: an availability census that quantifies how — or whether — each state publishes its lithology, mapping where the national record is machine-readable, scanned-only, or absent; and a reproducible, cost-bounded procedure to recover lithology from the scanned-only remainder, reported with per-state coverage, accuracy, and cost. The inventory is the contribution; the census situates it and the recovery procedure shows how the remaining gap can be closed.

This inventory is complementary to the United States Groundwater Well Database (USGWD; Lin et al., 2024), which harmonizes 14.26M well records to a data standard of location, purpose, status, well depth, screen depth/length, and well capacity — but, by design, carries no lithology and no water levels. USGWD is the national well-location layer; this inventory is the subsurface-structure-and-hydraulics layer that USGWD omits. The two compose. Likewise, national hydrogeologic frameworks such as the USGS Secondary Hydrogeologic Regions partition the conterminous US into terrane units at the land surface (Belitz et al., 2019); our depth-resolved lithology populates the third dimension those map-based frameworks leave empty.

1.1 Who needs this, and how the gap is bridged today

The demand is documented on three independent lines. The continental-modeling community states the lack explicitly: "no consistent 3D lithologic, hydrogeologic or hydrofacies data set exists" at continental scale (de Graaf et al., 2020); hydrogeologic heterogeneity "requires subsurface borehole . . . data that are often scarce" (Condon et al., 2021); and even USGWD — the harmonized national well-*location* database — excludes lithology "due to [its] limited availability" (Lin et al., 2024), naming the precise gap this inventory fills. Agencies act on the need at regional scale: the USGS mined ∼458,000 driller's logs for a single flow model (Arihood, 2009) and standardized ∼14M records across the 24 glaciated states (Bayless et al., 2017). Most telling is duplicated effort: because no harmonized layer exists, every team that needs subsurface lithology re-compiles raw driller's logs from scratch — thousands to hundreds of thousands of logs per study, each in a bespoke workflow (representative studies are tabulated in supplement Table S1) (Faunt et al., 2010; Arihood, 2009; Vahdat-Aboueshagh et al., 2021; Riddle et al., 2025; Herzog et al., 2003; Bayless et al., 2017). That recurring, state-by-state re-compilation is the cost of the missing layer.

In the absence of harmonized borehole data, CONUS-scale platforms either collapse the subsurface conceptually (the NOAA National Water Model and USGS National Hydrologic Model resolve a ∼2 m soil column above lumped calibrated reservoirs) or parameterize it from surface geologic maps: ParFlow-CONUS assigns conductivity to 5–10 broad deep units from the GLHYMPS/GLiM map lineage (Maxwell et al., 2015; Yang et al., 2023; Gleeson et al., 2014; Hartmann and Moosdorf, 2012), and its developers describe the result as "limited not by computational expense but by data availability" and "better viewed as a shallow aquifer storage model" (O'Neill et al., 2021); map generalization alone has a measured cost on simulated fluxes

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Open Access Earth System Science Data Discussions

(roughly −59% to +38%; Moosdorf et al., 2010), and the same few map-derived inputs recur field-wide (Fan et al., 2013; de Graaf et al., 2017; Pelletier et al., 2016; Hengl et al., 2017; Zamrsky et al., 2025). Depth-resolved alternatives do not close the gap: USGS 3-D geologic framework models are regional (38 models, ∼49% of CONUS; U.S. Geological Survey, 2023), the seamless national geologic map is a surface polygon product (Horton et al., 2017), the one bulk driller's-record effort is confined to the glaciated states and outputs derived property grids rather than lithology (Bayless et al., 2017), and research boreholes are exquisite but number in the single digits per campaign (Twining et al., 2017; Uhlemann et al., 2019). Continental subsurface data thus splits between coarse-everywhere maps and exquisite-nowhere research holes (Figure 1); the scalable middle — millions of driller's logs — is what this inventory harmonizes. A further, less-remarked barrier is access position: the comprehensive prior compilations were assembled *by or with* the custodian agencies (inter-agency data requests), whereas for anyone outside those walls the same records are scattered across dozens of independent portals with no common index. We therefore built the inventory deliberately through the public pathway — every record reachable through a public endpoint with no privileged access — which is both why per-state counts can trail an agency's internal totals and why a reproducible, outside-the-firewall assembly is itself part of the contribution. Where records stayed locked, the reasons were statutory rather than physical (California's logs were confidential until ∼2017; §4.1).

2 Methods

2.1 Harvest — and the heterogeneity it exposes

State sources are ArcGIS REST (most), bulk file, or by-request. A registry-driven harvester paginates each REST layer to FlatGeobuf. We encountered, and the supplement catalogues (Table S2), about ten distinct access modes. The central methodological finding is that the access mode is not predictable from the portal type: an HTML viewer resolved to a structured server-side lithology table for some states (OK, IA) and to scanned PDFs for others (UT, ID); a structured-looking column was 100% empty (KS); and a state booked as carrying one interval per well (IL) in fact exposed the full multi-interval log through a hidden, born-digital text-layer PDF endpoint. Every source must be dereferenced and probed before it is classified.

2.2 Harmonization to a canonical interval schema

Each state's native lithology is mapped to one canonical interval schema — exactly the columns shipped in the release: `state | well_id | seq | top_ft | bottom_ft | description | normalized | source_id | source_type | qa_flag`. Depths are retained in feet (1 ft = 0.3048 m), the unit of the source driller's logs. Adapters handle separate interval tables, wide inline columns (un-pivoted to long), and multi-file bulk sources. Verbatim driller text is always retained in `description`; `normalized` is an explicitly *derived* convenience column — a rule-based keyword map onto the same 25-class controlled vocabulary used by the vision-language extraction schema (§2.4), with `other` where no rule matches. The downstream goal is parity with the per-well hydraulic attributes that MODFLOW-style groundwater-model

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Open Access Earth System Science Data Discussions

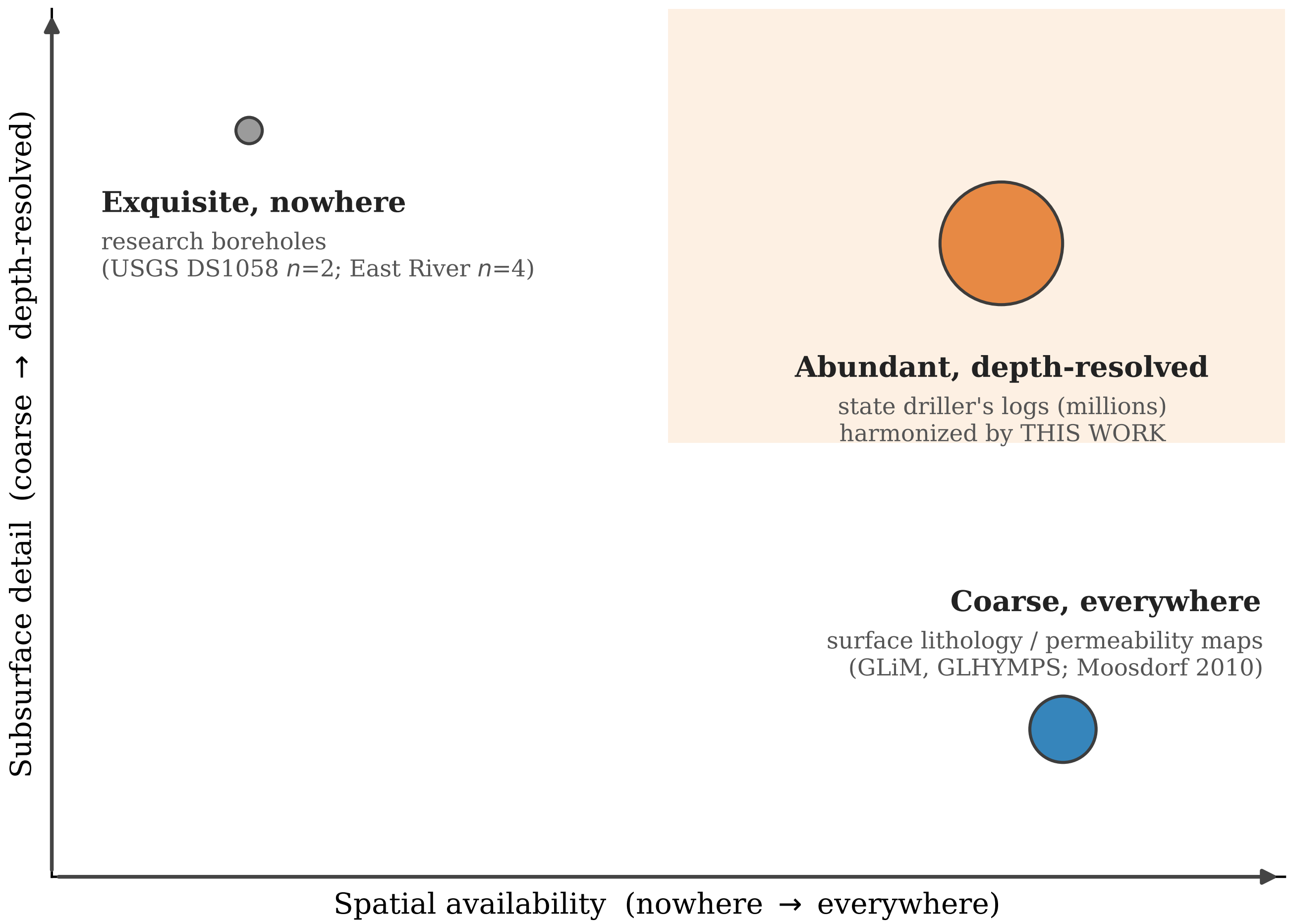

Figure 1. How the CONUS subsurface is represented today, on two axes the existing sources trade off against each other: spatial availability (horizontal) and depth detail (vertical). Surface lithology and permeability maps are coarse but everywhere (lower right); research boreholes are exquisite but almost nowhere (upper left, n of order single digits). Only state driller's logs — millions of them, and depth-resolved — occupy the abundant-and-detailed quadrant (upper right); they are the sole scalable source of three-dimensional subsurface information, and the contribution of this work is to harmonize them. Marker size is schematic, scaled to relative abundance.

construction consumes (horizontal and vertical hydraulic conductivity, aquifer thickness, and static water level); §4.3 reports what raw ingredients exist for that derivation.

2.3 Quality assurance and provenance — flag, never impute

The release policy is an honest harvest: every value is shipped as published by the source agency, and no value is interpolated, gap-filled, or modeled. Release QA distinguishes two cases. *Documented encoding artifacts are nulled*: per-source filler codes

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Open Access Earth System Science Data Discussions

that an agency uses to mean "not reported" (e.g. Michigan Wellogic's 9999-family static-water-level fillers — every occurrence of 9999, 9999.9, 999.99 and 999.9 ft in the compiled static-water-level column is a Michigan row — and its 999-hour pump-test filler) and non-finite numeric parse artifacts. *Everything else is kept and flagged*: a `qa_flag` column marks physically implausible magnitudes (e.g. depths beyond 15,000 ft, static water levels beyond 5,000 ft) and internal inconsistencies (static water level below total depth, reversed screen intervals), with generous bounds so genuine extremes — Wyoming basin water levels beyond 4,000 ft, flowing-artesian negative water levels — are never touched. The complete rule set is a machine-readable file (`qa_rules.json`) archived with the method code. Provenance is per-row: `source_id` keys every record to the per-state source ledger (supplement), and lithology `source_type` separates primary driller-log records from geologist-coded stratigraphy databases, vision-language-recovered samples (§2.4), and hydrogeologic-framework unit contacts interpreted at real boreholes. Interpreted framework *grid* products are excluded from the well tables entirely: the Lake Michigan under-lake framework ships as a separate table of model cells (§**??**), so `state` is a true USPS jurisdiction everywhere and no interpreted grid cell can be mistaken for a well.

2.4 PDF triage and extraction — a full-population, cost-definitive design

For the PDF-only states we do not run a blind full mine, nor rely only on sampling. The classification step—does a log have a text layer?—is free (`pdftotext`) and needs no vision model, so it can be run over the whole state. For each well's log we stream-and-discard: fetch, run `pdftotext`; if a text layer is present, extract the lithology now via a free state rule-parser (where the layout is a clean table) or a cheap text-only large language model; if the page is a raster scan, the well is recorded to an exact "needs-vision" worklist and deferred. No PDF is retained. This yields, per state, the harvested digital intervals, an exact count of vision-required logs, and therefore a definitive cost (digital cost incurred + vision count × measured per-log vision cost) rather than an extrapolated estimate. Raster logs are read with a vision-language model (Gemini 2.5, native PDF input, forced JSON schema), an approach grounded in a method base that matured only recently: multimodal models now read text-rich and handwritten documents (Gemini Team, Google, 2024; OpenAI, 2024; Bai et al., 2025; Chen et al., 2024; Liu et al., 2024; Humphries et al., 2024; Crosilla et al., 2025), and the "documents→structured database" paradigm is established in adjacent domains (Dagdelen et al., 2024). Within hydrogeology, machine learning has been used to predict lithofacies from borehole geophysics (Hammond and Allen, 2024), and natural-language processing is only now being introduced to groundwater archives (Christenson and McCoy, 2026); within geoscience more broadly, prior LLM work classifies already-transcribed borehole text (Ghorbanfekr et al., 2025). None recovers structured lithology from scanned driller's logs at scale. The shift that makes this tractable is recent: vision-language models from 2023 onward collapse OCR, handwriting recognition, layout parsing, and semantic typing into a single prompt-and-validate step, changing the extraction economics from per-log hand-transcription to metered inference. The closest scanned-record prior art quantifies the pre-VLM barrier and clarifies our methodological departure. Ma et al. (2024) extract well location and depth from 160 historical oil-and-gas records via an OCR-to-text conversion feeding a text LLM (Llama 2): accuracy holds at ∼100% on clean digital PDFs but falls to ∼70% on image-based scans, precisely where OCR breaks down. Ghorbanfekr et al. (2025)'s domain-adapted model (GEOBERTje) classifies already-transcribed borehole descriptions into lithology classes, outperforming a GPT-4 prompt by 11%, but never

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

touches a scan. We differ on all three axes: a vision-language model reads the scanned form directly (no OCR stage to fail), the target is the harder material-by-depth lithology log rather than point attributes, and it runs over a multi-state corpus rather than hundreds of documents.

3 Results

3.1 Availability census

Of ∼13.3M wells across 48 CONUS states, after systematic probing lithology is machine-readable (free, no vision model) for 21 states (∼8.4M wells), available only as scanned PDF logs for 11 states (∼3.1M), and without public digital lithology for 16 (∼1.8M) (Table 1, Figure 2). The central finding is that the machine-readable category is ∼2.6× what a surface audit suggests: an initial classification placed only 8 states (∼3.3M wells) in category A, but systematic probing revealed structured sources — server-rendered HTML lithology tables (Tennessee, New Mexico, Vermont, Oregon), bulk strata files and archives (Texas, Kansas, North Dakota), born-digital report PDFs (Ohio, Massachusetts), an overlooked bulk GeoPackage (Minnesota), and a public lithology table behind a map viewer (Indiana) — for thirteen states first taken to be scanned-only or data-less, moving ∼4.1M wells from the vision-model column and ∼1.0M from the no-data column into the free column. Classification is evidence-based and bidirectional, and can take several passes: Oklahoma reclassified PDF→digital once its viewer was found to render a server-side HTML lithology table (243,801 wells); Kansas moved digital→PDF when its `GEOLOGICAL_FORMATION` column proved 100% empty, then PDF→digital again when the state survey's bulk archive was found to carry staff-transcribed lithology for the full registry; North Dakota's scanned driller's logs proved to be a red herring — the same agency publishes tabular lithology in a separate bulk export. Classification is by each state's *primary well-record corpus*; small structured research subsets (Florida's GEODES geologist logs, Colorado's Denver Basin formation picks, Maryland's MCPAIS framework boreholes — all harvested, Table 2) do not reclassify a registry whose driller's logs remain scanned or unpublished. Category C is "no harmonized digital lithology," not "no data": a state-by-state regulatory audit (§4.2) found that records in these states are public and almost always merely undigitized (paper/county-health custody) or held behind an unwired portal, rather than legally sealed today. Outright sealing of the report itself is now rare and was largely historical: California's reports were statutorily confidential until their ∼2017 release (§4.1), and the main current restriction we found is Delaware's withholding of public-supply well locations.

3.2 Harmonized digital inventory

The harmonized release spans 34 states (Table 2): 28,518,082 depth-resolved lithology intervals across 5,165,761 wells, with Michigan and Texas dominant; structured hydraulics for 6,895,791 wells in 41 states; and coordinates for 7,796,626 wells across all 48 conterminous states. Of the lithology wells, 98.1% carry a coordinate row; the 95,873 without one are wells whose agency publishes no usable coordinate (Massachusetts) or whose lithology postdates the frozen location snapshot (Texas) — documented rather than filled (§2.3). A companion table ships the USGS Lake Michigan Basin under-lake hydrogeologic

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Open Access Earth System Science Data Discussions

Table 1. Availability census of depth-resolved lithology, 48 CONUS states (per-state counts and classification evidence in the archived census workbook). Well counts are each state's public well-record universe as accessed by us; for states with no public well database the USGS NWIS groundwater-site count is used.

Category	States	Wells	Access
A — machine-readable lithology (free)	21	∼8.4M	REST / inline / bulk / structured-HTML / text PDF
B — scanned-PDF lithology (vision model)	11	∼3.1M	raster scan (this method)
C — no public digital lithology / no DB	16	∼1.8M	none / by-request / access-blocked

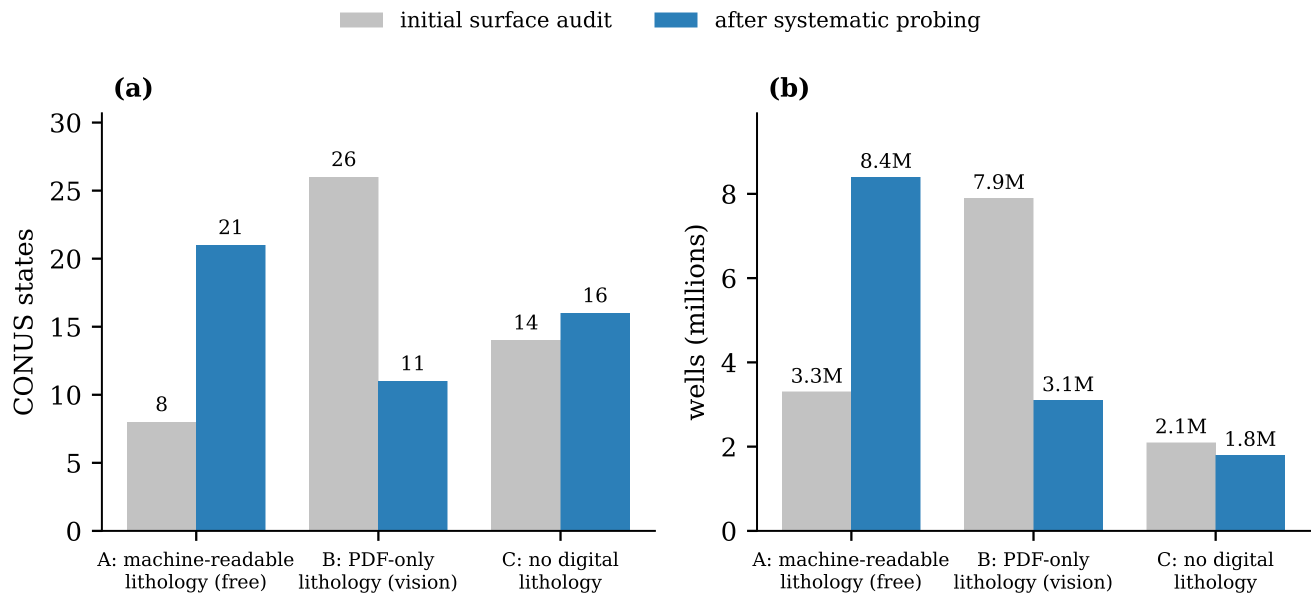

Figure 2. Lithology availability across 48 CONUS states by (a) number of states and (b) number of wells, shown before and after systematic per-state probing (§2.4). The initial surface audit (grey) is what a casual inspection of state portals suggests; systematic probing (blue) is the classification carried forward everywhere else in this paper (Table 1). Probing reclassifies thirteen states and ∼5.2M wells into the free, machine-readable category A (∼4.1M from the vision-model category B, ∼1.0M from the no-data category C) — the central availability finding. Totals are conserved: 48 states and ∼13.3M wells in both columns.

framework (260,637 interpreted unit intervals on 30,290 model grid cells) separately from the well tables, because its records are model cells, not wells.

Coverage is far from uniform across states and data elements (Figure 3): lithology dominates the structured states, while the hydraulic and derivation-readiness fields (static water level, depth, yield, surface elevation, completion date) are present in widely varying fractions, and the scanned-only states remain to be extracted.

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

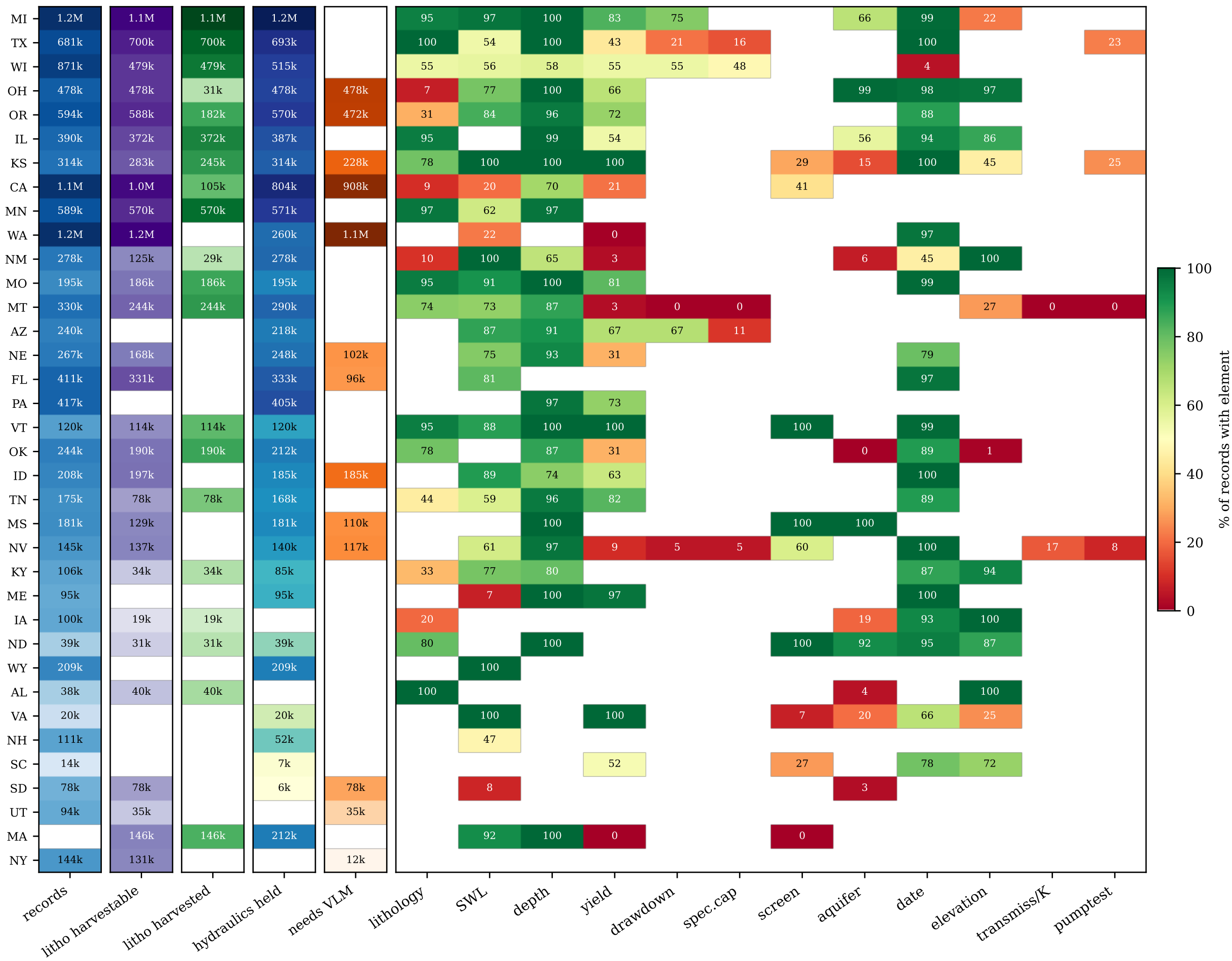

Figure 3. Per-state coverage of the harmonized inventory across the conterminous United States, ranked by estimated count of fully-attributed records. The five count columns are: *records* (wells held); *litho. harvestable* (wells whose lithology is obtainable); *litho. harvested* (wells with lithology now in the inventory); *hydraulics held* (wells with any structured hydraulic value); and *needs VLM* (scanned lithology still to extract). The heatmap reports, for each element, the percentage of a state's records that actually hold that element — harvested lithology plus the structured hydraulics and derivation-readiness fields (static water level, depth, yield, drawdown, specific capacity, screen, aquifer, completion date, surface elevation, measured transmissivity/conductivity, and pump-test duration). Blank cells are not-yet-harvested or not held, not a prediction; Michigan is drawn from the full Wellogic source.

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Table 2. Harmonized lithology inventory by state (valid depth-resolved intervals, 34 states; generated from the release build summary). Sources are primary state-agency databases unless marked: † regional USGS/state hydrostratigraphic framework (interpreted unit contacts at real boreholes, shipped with `source_type='framework_interpreted'`); ‡ validated vision-language sample recovered from that state's scanned driller's-log images (`source_type='vlm_extracted'`; the bulk scanned remainder is characterized in §3.5, not shipped). Wells are state-scoped distinct counts. Alabama's structured source publishes intervals almost entirely without usable depths, so only 3 wells survive the valid-interval criterion (§4.3). The Lake Michigan under-lake framework (260,637 intervals / 30,290 grid cells) ships as a separate framework table, not counted here.

St	Source	Intervals	Wells	St	Source	Intervals	Wells
MI	Wellogic (EGLE)	5,470,675	1,105,336	KY	KGS	167,906	34,374
TX	TWDB SDR	4,621,818	618,797	OH	ODNR	162,430	31,427
MN	CWI	2,804,726	568,935	IA	Iowa GS GeoSam	114,542	19,286
IN	IDNR	2,085,744	404,049	NE	NE DNR‡	78,091	2,894
IL	ISGS ILWATER	2,012,210	329,753	NM	NM OSE	52,487	29,129
WI	WI DNR	1,736,410	478,814	ID	IDWR‡	32,223	3,347
KS	KGS WWC5	1,591,408	244,610	WA	WA Ecology‡	29,336	4,009
MT	MBMG GWIC	1,533,645	243,818	SD	SD DANR‡	28,537	302
CA	DWR OSWCR	1,330,410	102,129	NV	NDWR‡	23,317	3,283
OK	OWRB	1,094,459	185,932	NY	NYSDEC‡	11,761	2,836
MO	MO DNR/MGS	755,037	184,363	MS	MS MDEQ‡	9,727	1,648
MA	MassDEP	742,326	146,383	UT	UT DWRi‡	8,062	770
OR	OWRD	724,650	182,336	CO	CO DWR†	5,508	3,525
VT	VT DEC	376,300	114,194	MD	MGS MCPAIS†	3,076	858
FL	FGS GEODES + VLM‡	349,227	9,861	NC	USGS PP1773†	1,415	182
TN	TDEC	306,914	77,691	SC	USGS PP1773†	1,126	113
ND	ND DWR	252,575	30,774	AL	GSA	4	3
				34 states	**total**	**28,518,082**	**5,165,761**

3.3 Dataset characteristics

Figure 4 summarizes what the released lithology actually looks like, computed directly from the frozen release. Intervals are decameter-scale: median thickness is 15 ft ($\approx$5 m), with a long tail to hundreds of feet (the comb structure at small values reflects drillers' whole-foot reporting). The record is shallow-biased but reaches aquifer depth: 90% of interval tops lie above $\approx$250 ft, and the 99.9th percentile of interval bottoms is $\approx$3,600 ft — i.e. the inventory resolves precisely the depth range that shallow aquifer systems and the upper units of regional groundwater models occupy, complementing the $\sim$100–400 m single-unit deep zones of continental platforms (§1.1). The derived material vocabulary is dominated by the unconsolidated glacial/alluvial classes — clay (9.0M intervals), sand (4.8M), gravel and sand-and-gravel (3.1M combined) — followed by the common sedimentary rocks (shale 2.1M, sandstone 1.7M, limestone 1.4M); the rule-based normalization leaves 7.0% of intervals as "other" (regional vocabulary, refinable post hoc without touching the verbatim descriptions). A typical well carries

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

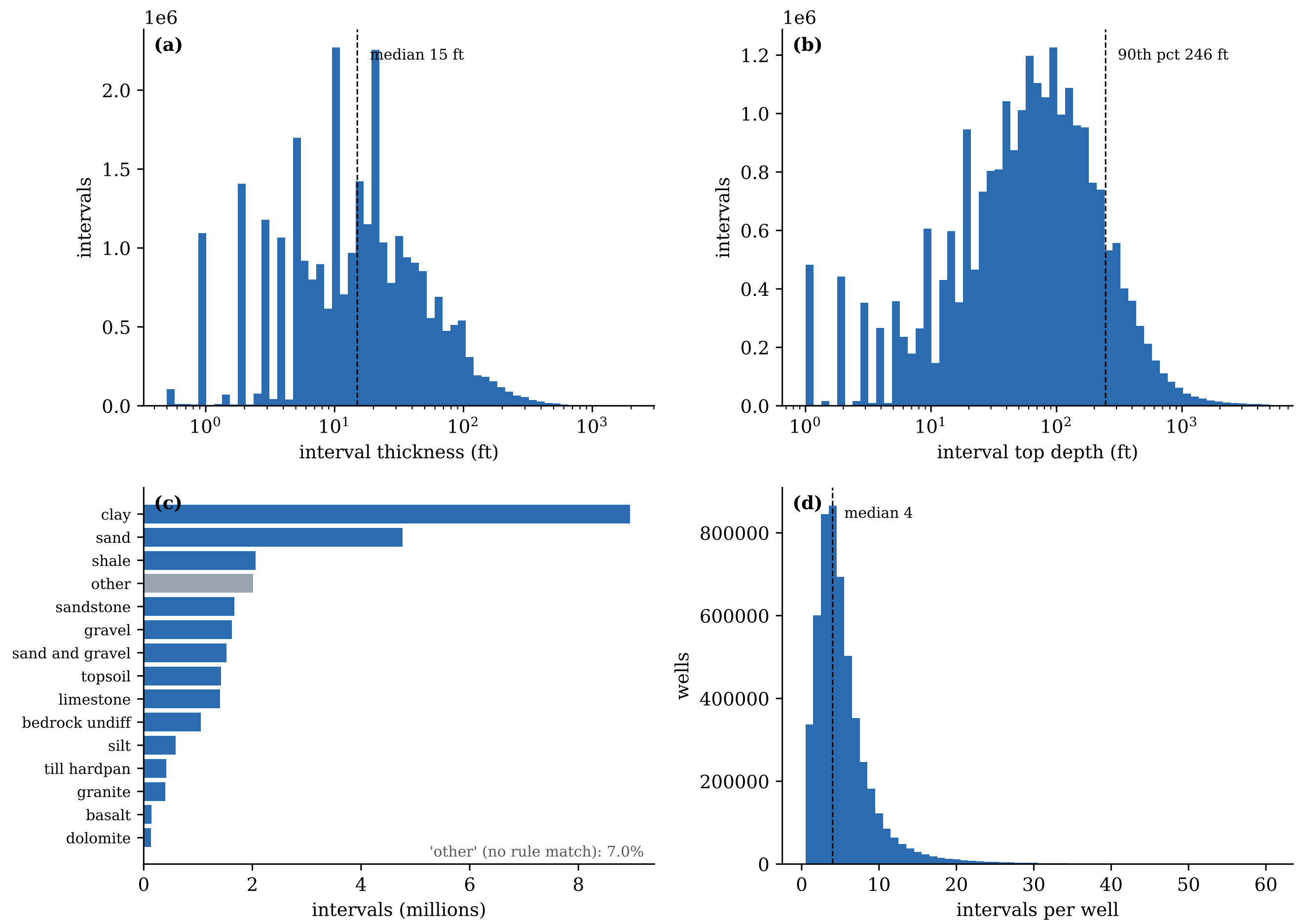

Figure 4. Characteristics of the released lithology table, computed from the frozen release: (a) interval-thickness distribution (log scale; dashed line = median; the comb at integer feet reflects drillers' whole-foot reporting); (b) interval top-depth distribution (log scale; dashed line = 90th percentile); (c) the fifteen most frequent derived `normalized` material classes (grey = "other", i.e. no keyword rule matched); (d) intervals per well (dashed line = median).

175 a median of 4 intervals (mean ≈5.5), consistent with the driller's-log convention of logging major material changes rather than laboratory-grade stratigraphy.

3.4 Extraction triage: the LLM/VLM split and its cost

An initial per-state calibration (n=100; Table 3, Figure 5) establishes the routing split the rest of the evaluation builds on. A retrievable log is present for 87–100% of wells everywhere; what varies sharply is the vision-required fraction — the share of logs that are raster scans needing a vision model rather than born-digital text — which spans 0–100% and is anti-

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Table 3. Initial per-state PDF triage calibration (n=100, Gemini 2.5 flash-lite, sorted by vision share): the born-digital (text-LLM) versus raster (vision-model) routing split. The $/1k-logs column is illustrative only (June-2026 list-price conversion of the blended token cost); Table 4 gives the definitive large-sample yield and token-metered cost.

State	has-PDF %	text-LLM %	raster-VLM %	intervals/log	$/1k
SD	100	0	100	10.9	0.37
ID	96	4	96	10.6	0.32
NV	93	5	95	7.2	0.25
OR	100	15	85	4.1	0.19
KS	90	29	71	9.8	0.30
FL	89	67	33	15.0	0.47
WI	98	78	22	3.2	0.20
IL	87	100	0	5.0	0.19

correlated with e-permitting era: recent e-permitting states (IL, WI, FL) are mostly born-digital (the free/cheap text route), while legacy paper-archive states (OR, NV, SD, ID) are almost all scans. Extraction is faithful across all states — driller shorthand transcribed verbatim, legibility flagged, and empty arrays returned (rather than fabricated intervals) on blank logs. The definitive large-sample yield and the token-metered cost follow in Section 3.5, and per-well validation against agency transcriptions in Section 3.6.

3.4.1 Definitive cost via full triage.

Because classification is free, the full-population triage (§2.4) converts the sampled estimate into a definitive per-state figure while harvesting the digital share for free. Illinois is the first completed case: triaging all 390,443 Illinois wells confirmed 0% vision-required (born-digital text layers), harvesting 2,198,466 lithology intervals from 382,245 wells (≈5.8 intervals/well) at zero vision-model cost. A national projection over the ∼3.1M scanned-PDF (category B) wells routed to the vision model — the raster remainder once probing has moved the born-digital share into the free category A — is on the order of 3 billion tokens[1] (≈$1.5k); the binding constraints are wall-clock and normalization-vocabulary QA, not inference cost or storage (stream-and-discard avoids a ∼500 GB archive).

3.5 Large-n yield and hydraulic-field coverage

We evaluated eleven scanned-PDF states (1,000–1,500 logs each; Table 4, Figure 6). A single "success rate" conflates three distinct outcomes, so we report them separately. Extraction yield — the fraction of documents the model actually read that yielded a log — is 94.6% pooled (11,806 documents) and 99–100% for most states: once a scan is legible, the model recovers a

[1]Token counts are the reproducible unit of cost; the dollar equivalents quoted in this paper derive from the Gemini 2.5 flash-lite list price as of June 2026 and will drift with provider pricing.

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

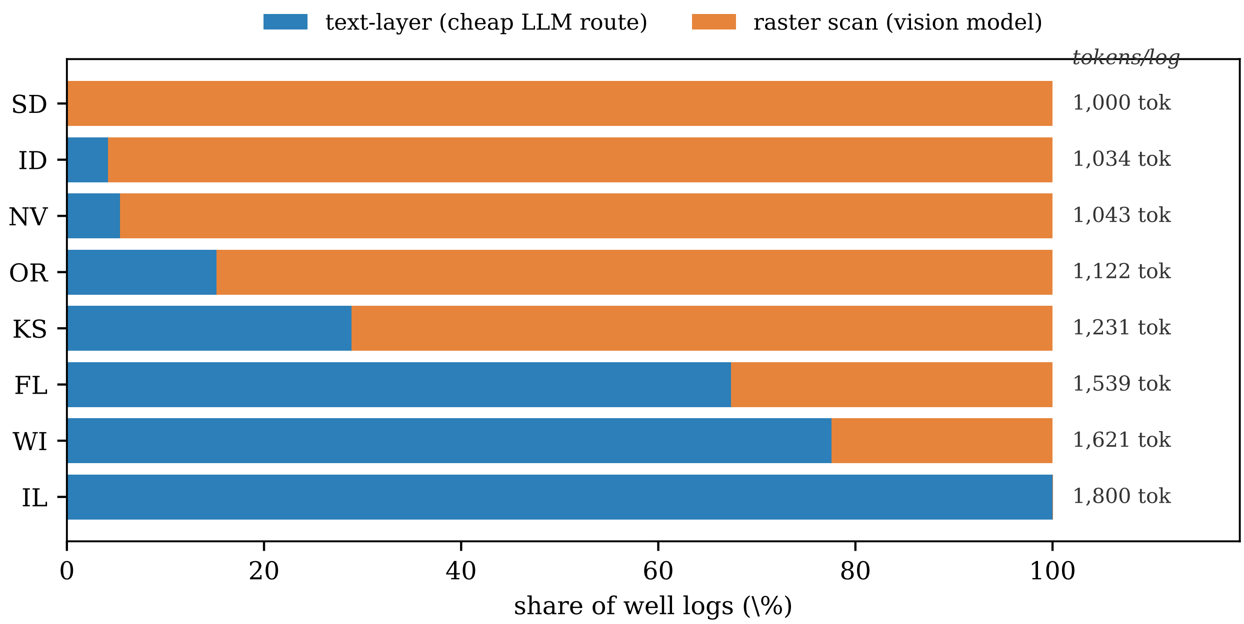

Figure 5. Per-state PDF triage (n=100): the share of well logs that are born-digital (cheap LLM route) versus raster scans (vision model), with the blended token cost per log (text-layer ≈1,800 tokens/log; vision-model raster ≈1,000). The vision-required fraction spans 0–100%, anti-correlated with e-permitting adoption.

log essentially always, and it returns an empty array rather than fabricating intervals when none is present. The recovery gaps lie on two other axes. Document availability varies widely: 28% of Mississippi and 63% of Utah wells simply have no scan on file ("no-PDF"). And one state is a genuine extraction outlier — Nebraska (65.5%) — because its modern registrations increasingly omit the lithology log (the temporal heterogeneity of Section 3.6); New York's apparent shortfall is instead transient server flakiness (∼24% retryable fetch failures), not extraction. For the spatially/temporally heterogeneous states (MS, NE, NV) the reported figure is from the dual-stratified run. We report cost in tokens, the reproducible unit, since per-token prices drift and vary by provider: the VLM (raster) route uses ≈1,000 tokens per log (≈600 input — which includes the rendered-page image tokens — plus ≈410 output), while the born-digital text route is higher at ≈1,800 tokens/log because the extracted text exceeds a single page image in token count. Across all evaluation logs the spend was ≈13 million tokens (≈$3).

Critically, recording every hydraulic field—returning null when absent—turns the evaluation into a coverage map of what the documents contain (Figure 6). Static water level is present on roughly half of documents (23–61%) and yield on 15–43%, giving a usable hydraulic backbone. But the pump-test fields are sparse—drawdown and pump rate are mostly below 20%—and specific capacity is essentially never written on the forms (0.1–0.5% across all 6,000 documents). This confirms, at scale, that the specific-capacity→transmissivity→conductivity pathway — the standard route for estimating aquifer hydraulics from well records (Rotzoll and El-Kadi, 2008) — is not broadly recoverable from driller's logs; derived conductivity is feasible only

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Open Access Earth System Science Data Discussions

Table 4. VLM evaluation, eleven scanned-PDF states, reported on three separate axes. Extraction yield = documents-with-a-log / documents-read (Wilson 95% CI); no-PDF = wells with no scan on file; design = age-stratified or dual spatial×temporal. n_{read} is documents the model read.

State	design	n_{read}	extraction yield % (95% CI)	no-PDF %	int/log
WA	both	1,497	100.0 (99.7–100)	0.1	5.8
SD	age	970	99.9 (99.4–100)	0.1	7.2
ID	age	928	99.9 (99.4–100)	5.0	9.8
NV	both	1,421	99.7 (99.3–99.9)	4.8	9.6
KS	age	872	99.7 (99.0–99.9)	9.1	7.5
UT	both	559	99.6 (98.7–99.9)	62.7	10.1
OR	both	1,491	99.2 (98.6–99.5)	0.3	6.8
MS	both	1,055	100.0 (99.6–100)	28.5	9.0
NY	age	722	94.7 (92.9–96.1)	3.4	9.9
FL	age	861	90.2 (88.1–92.1)	10.3	7.1
NE	both	1,430	65.5 (63.0–67.9)	3.6	8.9
pooled	—	**11,806**	**94.6**	—	—

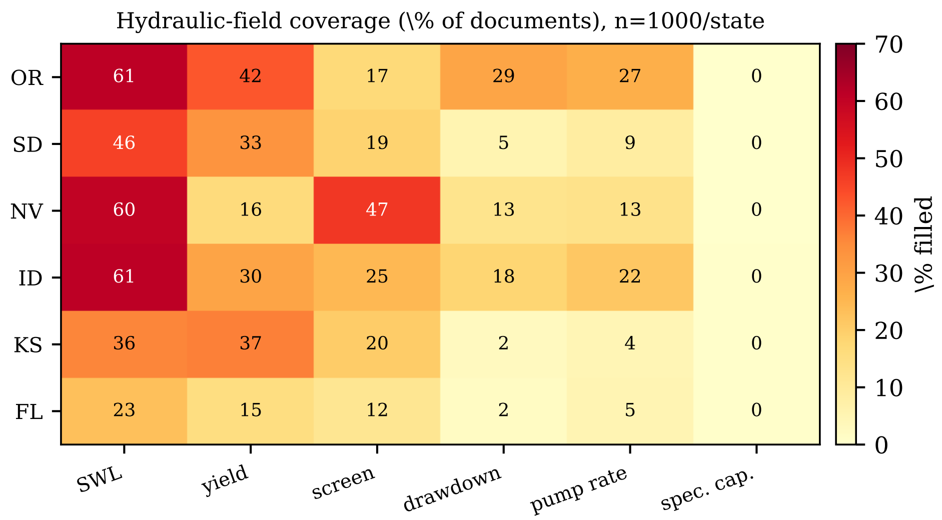

Figure 6. Hydraulic-field coverage: % of documents on which each field is present (n=1,000 per state). Static water level and yield form a backbone; specific capacity is essentially never reported.

where yield and drawdown co-occur (here, OR and ID). The zeros are themselves a result: each field was searched in every document, and its absence is now quantified rather than assumed.

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Table 5. VLM extraction accuracy versus DWR hand-transcribed geologic logs, California (n=1,499 wells; 516,317 ft of section compared; ground truth = DWR FreeForm transcription).

Metric	Value
Interval count — Pearson r / MAE	0.94 / 0.91
Depth-boundary F_1 ($\pm$5 ft)	0.97
recall / precision	0.98 / 0.97
Logged-depth extent — r	0.95
Material class accuracy (12-class)	91.5%
Cohen's κ	0.90
Coarse/fine accuracy	96.0%
Cohen's κ	0.92

3.6 Extraction accuracy: validation against agency transcriptions

Yield (Section 3.5) measures whether a scanned log is readable, not whether the model reads it correctly. California permits a direct accuracy test: 86,653 wells carry both a DWR hand-transcribed FreeForm geologic log and the original scanned report. We sampled 1,499 such wells, passed each scan through the same flash-lite pipeline, and scored its output against the agency transcription as ground truth (Table 5). Agreement is high on every axis that matters: interval count Pearson $r = 0.94$ (means 12.9 vs. 13.1; MAE 0.91 intervals), depth-boundary $F_1 = 0.97$ ($\pm$5 ft; recall 0.98, precision 0.97), and 12-class material accuracy 91.5% (Cohen's $\kappa = 0.90$) over 516,317 compared feet. On the coarse/fine binary that texture and conductivity models actually consume, agreement is 96.0% ($\kappa = 0.92$). Per-class recall is highest for rock/bedrock (0.96), shale (0.95), sand (0.92) and gravel (0.90), and lowest for the rare silt (0.79) and boulders (0.77). Two residual error sources are reported rather than hidden: (i) $\sim$6% of wells split their log across multiple scanned-page files, and our single-page comparison understates recall for those (a fixable stitching gap that produces the long tail in depth-extent error); and (ii) part of the material "disagreement" is wording granularity — terse human entries ("Soil & Rock", color-only lines) versus the fuller scan — rather than model error, making 91.5% a conservative floor. Against the agency's own labels, the model matches coarse/fine texture essentially perfectly and material class to $\kappa \approx 0.90$, validating the recovered layer for hydrogeologic use.

3.6.1 Sampling representativeness.

A 1,000-log estimate is only as good as its sample, so we re-evaluated six states (n=1,500) under a combined spatial×temporal stratification (a $\sim$1° latitude/longitude grid crossed with completion decade) and tested whether yield is homogeneous across strata. For a proportion, precision depends on the sample size, not the population: at n=1,000 the worst-case 95% margin is $\pm$3.1% and the finite-population correction is $\approx$1.000, so sampling 1,000 of $\sim 10^6$ scans is statistically tight provided the sample is unbiased. Three states (WA, OR, UT) are design-robust — per-stratum yield spreads $\leq$9 percentage points with all strata mutually inside Wilson 95% intervals. Two are not: Mississippi varies by geography (north-west counties $\sim$50% versus

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Open Access Earth System Science Data Discussions

~86–89% in the south, tracking per-county scan coverage), and Nebraska and Nevada vary by age — Nebraska's modern registrations increasingly omit the lithology log (76% recovery in the 1960s falling to 23% in the 2020s). These heterogeneities, invisible to an age-only or spatially-clustered sample, both justify the dual stratification and are themselves findings about how within-state completeness varies. The stratified runs also separate extraction yield from fetch reliability: success as a fraction of documents the model actually read is the design-robust quantity, whereas success as a fraction of attempts also absorbs transient server failures (Washington's apparent 77% rises to ~99.9% of fetched documents once download flakiness is excluded).

4 Discussion

The contribution is the harmonized framework plus a reproducible, cost-bounded method — not a finished national mine. We need not download and process every log to deliver value: we provide the method, the per-state cost/accuracy/coverage estimates, and the harmonized data we can already release. The same harmonized lithology, paired with the pump-test ingredients that often accompany it, is the input continental models currently approximate with GLHYMPS-class generalized geology (§1.1), offering a path to replace coarse deep-conductivity units with observation-based stratigraphy where log density allows.

4.1 California: a case study in locked-but-recoverable data

California concentrates both themes of this paper — the data exists but is locked, and vision-language methods newly unlock it — and we treat it as a worked example. The California Department of Water Resources hosts on the order of ~1 million scanned Well Completion Reports (our enumeration of the CKAN "WCR PDF Links" resource returns 1,012,784; DWR itself typically states " ~800,000–1,000,000 ", so we treat the figure as approximate), each carrying the driller's lithology log. Yet only ~104,000 unique wells' lithology has been transcribed into DWR's structured geologic-log tables (across overlapping format categories, not additive: FreeForm ≈87,633, Generalized ≈22,260, QuickPick ≈6,800, USCS ≈3,071), leaving roughly 90% of California's per-well lithology locked inside un-transcribed scans.

Three compounding reasons kept it locked. (i) Legal confidentiality: the logs were sealed under California Water Code § 13752 (State of California, 2024) and became broadly public only after SGMA (2014) and AB 1755 (2016) (State of California, 2016), so the public bulk dataset is only ~8 years old (§1.2). (ii) Manual-transcription cost: even the USGS GAMA program has hand-transcribed only ~19,247 reports verbatim (<3% of the corpus), basin-by-basin (Haugen et al., 2023), the barrier being handwriting and heterogeneous historical forms that classic OCR cannot read. (iii) Aggregate-not-per-well, low ROI: every major use collapsed lithology into smoothed percent-coarse or texture fields rather than per-well logs. The USGS Central Valley Hydrologic Model obtained >150,000 scanned logs from DWR but used only ~8,500, as binary coarse/fine percent-coarse on a ~1-mile grid (Faunt, 2009; Faunt et al., 2010), following the foundational drillers'-log texture method of Laudon and Belitz (1991); CVHM2 continued this lineage (Marcelli et al., 2022); the attributed-WCR releases carry ~260–290k construction attributes but not lithology (Borkovich et al., 2023); and Stanford's airborne-electromagnetic work used ~11,645 drillers'-log sediment profiles only to calibrate airborne EM (Knight et al., 2018; Kang et al., 2025). To our knowledge, no published work performs statewide, per-well lithology extraction from the scanned California WCR corpus, and none

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Open Access Earth System Science Data Discussions

applies vision-language models to it — the per-well, full-corpus extraction our triage pipeline (§2.4) is designed to make routine. As a direct test we streamed 1,000 randomly sampled un-transcribed California WCRs through the pipeline: every report fetched as a scanned (raster) PDF, 97.2% yielded at least one lithology interval (mean 7.85 intervals per log), at ≈1,000 tokens per log. At that yield the entire ∼908,000-well un-transcribed remainder is recoverable for ≈0.9 billion tokens (≈$260) — turning a 90%-locked statewide corpus into a structured per-well lithology layer at metered-inference cost.

4.2 Why the data is scattered — a governance outcome, not absence

The fragmentation has a specific, documentable cause, and it is not that the data is missing or withheld. Surface-water data is federally unified because one agency both collects and houses it (USGS NWIS, via the Cooperative Water Program); groundwater never acquired such a home. Water wells are regulated by state law — there is no federal well-construction or well-log-reporting standard — so record creation follows fifty schemas, with custody often further divided within a state by agency mission. Federal unification attempts were structurally limited — the National Ground-Water Monitoring Network is a voluntary aggregator, the Safe Drinking Water Act reaches only public water systems, and the National Geothermal Data System (2009–2014) federated only the already-digitized geological-survey holdings before losing funding (Allison et al., 2014) — and none targeted, reached, or outlived the shallow domestic water-well driller's logs that carry the lithology.

Privacy is real but narrow today: modern state public-records law is public-by-default, and where records remain restricted it is the owner's identity or a well's exact location (e.g. Delaware's public-supply wells), not the lithology — although several states' reports were fully confidential until recently (California's until ∼2017; §4.1), and domestic-well permitting exemptions mean some logs are never collected at all (Maxwell et al., 2019). A state-by-state audit of the apparent "no-data" states found most to be a harvesting miss rather than an absence: of eleven states initially flagged, seven expose completion reports through a named public portal, and the genuine gaps (Georgia, Maryland, Rhode Island, West Virginia) are paper/county-health records obtainable by request. "No digital lithology" therefore means *not yet harmonized or digitized*, essentially never absent or legally sealed.

4.3 Known gaps and shortcomings — a living inventory, not a closed product

We catalogue current holes so users can judge fitness and later versions can close them. *Lithology coverage is partial*: the release (Table 2) is the harvested-to-date subset of the wells the census finds carry machine-readable lithology; the remainder is recoverable by the same pipeline, the scanned-only states are characterized by validated sampling (yield, accuracy, cost) rather than fully harvested — their rows ship flagged `source_type='vlm_extracted'` so no user mistakes a validated sample for a full-state harvest — and a few states publish no public log at all. *Locations are not universal*: a small share of lithology wells (reported per state in the release summary) has no coordinate row, because the agency itself publishes none (Massachusetts: ∼48,000 wells carry null/zero coordinates at the source) or because the lithology postdates the frozen location snapshot (Texas); we never invent a coordinate for them. Published coordinates also vary in precision: where the agency itself marks a coordinate as PLSS-derived or not field-verified (California's "Derived from TRS" reports; Minnesota CWI's *unlocated wells* layer), the release carries `loc_precision='approximate'` — section-center class, hundreds of meters.

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Hydraulic ingredients are present but uneven: static water level and total depth form a broad backbone, but the pump-test fields needed to derive conductivity are scarce — specific capacity is recoverable mainly where yield and drawdown co-occur (e.g. Texas, Arizona, Wisconsin). We separately audited the derivation-readiness inputs that govern whether transmissivity can be reconstructed downstream. Surveyed land-surface elevation — reported as-is, never substituted with a modeled DEM fill — is bulk-available for eleven states, often complete (Alabama, Iowa, New Mexico at 100%; Ohio 97%, Kentucky 94%, Illinois 86%) but absent from most records elsewhere because elevation, unlike the driller-observed depths and water levels, is not measured at the wellhead. A well completion or construction date fares far better: twenty-two states report it, most at 90–100%, giving the inventory a temporal axis (record vintage) for age-stratified analysis. Pump-test duration — the input that lets specific capacity be promoted to transmissivity via the Theis/Logan relations — is harvestable for Texas ($\sim$156,000 wells), Kansas, and Nevada. *Transmissivity and conductivity need care*: the release's only bulk T/K columns are Michigan's statewide values, which are the agency's own *model estimates* derived from lithology and specific capacity, not aquifer-test measurements — they ship marked `hyd_prop_provenance='agency_estimated'` and should be treated as a screening layer, not observations. Measured aquifer-test transmissivity in public bulk form is essentially confined to Nevada's driller-report portal ($\sim$25,000 wells with reported transmissivity, screen, and test duration), which publishes structured hydraulics but keeps lithology as scanned images and is not yet integrated here — the strongest evidence that a national *measured*-conductivity layer must be derived from the yield/drawdown/duration ingredients rather than read off the records. Normalization leaves a minority of intervals as "other" (regional vocabulary; the derived `normalized` column is refinable post hoc without touching the verbatim text); and the scanned extraction, although validated against agency transcriptions for $\sim$1,500 wells (Section 3.6), is not yet run over every state's full corpus. Some states resist the driller's-log model entirely: Colorado's public subsurface data is geophysical-log and aquifer-framework based, so its release rows are Denver Basin formation picks (`source_type='framework_interpreted'`), not driller strata. Finally, a source detail worth naming: Alabama's structured interval table publishes almost all intervals without usable depths, so only 3 Alabama wells survive the valid-interval criterion even though the coverage census credits the state with $\sim$38,000 lithology-bearing records — the census counts what agencies hold, Table 2 counts what passes release validation. None of these is fatal; each is a documented, improvable hole.

4.4 Transferability beyond the United States

The conditions that make U.S. well lithology hard to assemble — records created under dozens of independent jurisdictions, partially digitized, sometimes behind document portals, never unified — describe the global state of water-well data. The assembly pattern demonstrated here depends only on a public list of wells carrying a document link plus, for scanned archives, a vision-language model: no privileged agency access, per-form templates, or country-specific training. The same triage (born-digital text versus raster scan) and token-metered extraction therefore apply wherever national well records remain image archives, at a marginal cost of roughly a thousand tokens of inference per well.

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Open Access Earth System Science Data Discussions

5 Conclusions

We present an open, harmonized inventory of well lithology and hydraulics for the conterminous United States, compiled entirely from primary state-agency sources through public endpoints: depth-resolved lithology for ∼5.2 million wells (28.5 million intervals, 34 states), structured hydraulics for ∼6.9 million wells (41 states), and coordinates for ∼7.8 million wells (48 states), every value as published by the source agency with per-row provenance, a documented flag-never-impute quality layer, and a separate interpreted-framework table. A companion 48-state availability census shows that machine-readable lithology exists at the source for far more of the national record than a surface audit suggests (21 states, ∼8.4 million wells), that the scanned-only remainder (∼3.1 million wells) is recoverable at validated accuracy ($\kappa = 0.90$ against agency hand-transcriptions) for on the order of 3 billion tokens of vision-language inference, and that the residual no-data category is a governance outcome — undigitized or unwired public records — rather than an absence of data. The inventory supplies the depth-resolved subsurface-structure and hydraulics layer that national well-location databases omit; the frozen, citable release and the archived method code make it reproducible and extensible as further states are unlocked.

6 Data availability

A frozen snapshot of the harmonized inventory is archived in a public repository (Zenodo) with a citable DOI (https://doi.org/10.5281/zenodo.21196958; Rafiei, 2026), as three Apache Parquet well tables joined on `(state, well_id)` plus one framework table: `gw_lithology.parquet` (28,518,082 depth-resolved intervals for 5,165,761 wells in 34 states, with verbatim descriptions, a derived normalized material class, and per-row `source_id`/`source_type` provenance), `gw_hydraulics.parquet` (structured hydraulics for 6,895,791 wells in 41 states, with numeric transmissivity/conductivity carrying an estimated-versus-reported provenance marker), `gw_well_locations.parquet` (coordinates for 7,796,626 wells in all 48 states, EPSG:4326, with the agency's own precision qualifier), and `gw_framework_lake_michigan.parquet` (the USGS Lake Michigan Basin under-lake hydrogeologic framework, kept separate because its records are model grid cells, not wells), with a data dictionary and per-source provenance. Quality assurance is documented and machine-readable (`qa_rules.json`; §2.3): only agency filler codes and non-finite parse artifacts are nulled, implausible values are flagged in `qa_flag` and kept as published, and nothing is imputed. Because the upstream public sources mutate and disappear, this frozen version is the authoritative record. The release is distributed under CC-BY-4.0; the underlying records are public US state well-completion data, and the release carries no well-owner names, addresses, or other personal information — only the agencies' public well identifiers, coordinates, lithology, and hydraulics as published. The method code — the vision-language extraction prompt and output schema, the validation harness, the harmonization routines, and the QA rule set — is archived alongside; the live per-state access resolvers are documented in the provenance supplement but are deliberately not distributed as a turnkey pipeline, because state endpoints change. The full inventory is additionally browsable through a live web platform (https://www.swatgenx.com/groundwater-inventory), whose per-state counts, coverage maps, and per-well lithology logs are served from the same harmonized database and update as new sources are integrated; the archived DOI, not the website, is the citable source.

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Author contributions. VR: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Visualization, Writing — original draft, Writing — review & editing.

Competing interests. The author develops and maintains SWATGenX (https://swatgenx.com), which hosts the optional interactive viewer referenced above; this presents no financial conflict with the findings reported here. The author otherwise declares no competing interests.

Disclaimer. During the preparation of this work the author used a large language model (Anthropic Claude) to assist with drafting and editing text and code. The author reviewed and edited all content and takes full responsibility for it. Separately, generative vision-language models are employed as a core *research method* for lithology extraction; that use is described in full in the Methods and is distinct from this manuscript-preparation declaration.

Acknowledgements. We thank the many state agency staff whose public well-record systems make this inventory possible.

Financial support. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Allison, M. L. et al.: Sustaining the National Geothermal Data System, Tech. rep., Arizona Geological Survey / U.S. Department of Energy, oSTI 1130323; DOE Geothermal Technologies Office, 2014.

Arihood, L. D.: Processing, analysis, and general evaluation of well-driller logs for estimating hydrogeologic parameters of the glacial sediments in a ground-water flow model of the Lake Michigan Basin, Scientific Investigations Report 2008-5184, U.S. Geological Survey, https://doi.org/10.3133/sir20085184, 2009.

Bai, S. et al.: Qwen2.5-VL Technical Report, arXiv:2502.13923, 2025.

Bayless, E. R., Arihood, L. D., Reeves, H. W., Sperl, B. J. S., Qi, S. L., Stipe, V. E., and Bunch, A. R.: Maps and grids of hydrogeologic information created from standardized water-well drillers' records of the glaciated United States, Scientific Investigations Report 2015-5105, U.S. Geological Survey, https://doi.org/10.3133/sir20155105, 2017.

Belitz, K., Watson, E., Johnson, T. D., and Sharpe, J. B.: Secondary Hydrogeologic Regions of the Conterminous United States, Groundwater, 57, 367–377, https://doi.org/10.1111/gwat.12806, 2019.

Borkovich, J. G. et al.: Attributed California water-supply well completion report data, U.S. Geological Survey data release (v6.0, 2025), https://doi.org/10.5066/P93ICKAF, 2023.

Brodaric, B., Boisvert, E., Chery, L., Dahlhaus, P., Grellet, S., Kmoch, A., Létourneau, F., Lucido, J., Simons, B., and Wagner, B.: Enabling global exchange of groundwater data: GroundWaterML2 (GWML2), Hydrogeology Journal, 26, 733–741, https://doi.org/10.1007/s10040-018-1747-9, 2018.

Chen, Z. et al.: InternVL: Scaling up Vision Foundation Models, arXiv:2312.14238, 2024.

Christenson, C. and McCoy, K.: Natural Language Processing for Groundwater Insights, Groundwater, https://doi.org/10.1111/gwat.70067, online first, 2026.

Condon, L. E., Kollet, S., Bierkens, M. F. P., Fogg, G. E., Maxwell, R. M., Hill, M. C., et al.: Global groundwater modeling and monitoring: Opportunities and challenges, Water Resources Research, 57, e2020WR029 500, https://doi.org/10.1029/2020WR029500, 2021.

Crosilla, L., Klic, L., and Colavizza, G.: Benchmarking Large Language Models for Handwritten Text Recognition, arXiv:2503.15195, 2025.

Dagdelen, J. et al.: Structured information extraction from scientific text with large language models, Nature Communications, 15, 1418, https://doi.org/10.1038/s41467-024-45563-x, 2024.

de Graaf, I. E. M., van Beek, R. L. P. H., Gleeson, T., Moosdorf, N., Schmitz, O., Sutanudjaja, E. H., and Bierkens, M. F. P.: A global-scale two-layer transient groundwater model, Advances in Water Resources, 102, 53–67, https://doi.org/10.1016/j.advwatres.2017.01.011, 2017.

de Graaf, I. E. M., Condon, L. E., and Maxwell, R. M.: Hyper-resolution continental-scale 3-D aquifer parameterization for groundwater modeling, Water Resources Research, 56, e2019WR026 004, https://doi.org/10.1029/2019WR026004, 2020.

Fan, Y., Li, H., and Miguez-Macho, G.: Global patterns of groundwater table depth, Science, 339, 940–943, https://doi.org/10.1126/science.1229881, 2013.

Faunt, C. C.: Groundwater Availability of the Central Valley Aquifer, California, Professional Paper 1766, U.S. Geological Survey, https://doi.org/10.3133/pp1766, https://pubs.usgs.gov/pp/1766/, 2009.

Faunt, C. C., Belitz, K., and Hanson, R. T.: Development of a three-dimensional model of sedimentary texture in valley-fill deposits of Central Valley, California, Hydrogeology Journal, 18, 625–649, https://doi.org/10.1007/s10040-009-0539-7, 2010.

Gemini Team, Google: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, arXiv:2403.05530, 2024.

https://doi.org/10.5194/essd-2026-527
Preprint. Discussion started: 22 July 2026

Ghorbanfekr, H., Kerstens, P. J., and Dirix, K.: Classification of geological borehole descriptions using a domain adapted large language model (GEOBERTje), Applied Computing and Geosciences, 25, 100 229, https://doi.org/10.1016/j.acags.2025.100229, 2025.

Gleeson, T., Moosdorf, N., Hartmann, J., and van Beek, L. P. H.: A glimpse beneath earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity, Geophysical Research Letters, 41, 3891–3898, https://doi.org/10.1002/2014GL059856, 2014.

Hammond, Z. and Allen, D. M.: Evaluating the feasibility of using artificial neural networks to predict lithofacies in complex glacial deposits, Hydrogeology Journal, 32, 509–526, https://doi.org/10.1007/s10040-023-02726-2, 2024.

Hartmann, J. and Moosdorf, N.: The new global lithological map database GLiM: A representation of rock properties at the Earth surface, Geochemistry, Geophysics, Geosystems, 13, Q12 004, https://doi.org/10.1029/2012GC004370, 2012.

Haugen, E. A., Bennett, G. L., Arroyo-Lopez, J. A., et al.: Compilation of lithologic data from selected well completion reports submitted to the California Department of Water Resources, U.S. Geological Survey data release (v3.0, 2025), https://doi.org/10.5066/P9M85U0T, 2023.

Hengl, T. et al.: SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE, 12, e0169 748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Herzog, B. L., Larson, D. R., Abert, C. C., Wilson, S. D., and Roadcap, G. S.: Hydrostratigraphic modeling of a complex glacial-drift aquifer system for importation into MODFLOW, Ground Water, 41, 57–65, https://doi.org/10.1111/j.1745-6584.2003.tb02568.x, 2003.

Horton, J. D., San Juan, C. A., and Stoeser, D. B.: The State Geologic Map Compilation (SGMC) geodatabase of the conterminous United States, Data Series 1052, U.S. Geological Survey, https://doi.org/10.3133/ds1052, 2017.

Humphries, M. et al.: Unlocking the Archives: Large Language Models Achieve State-of-the-Art Performance on the Transcription of Handwritten Historical Documents, arXiv:2411.03340, 2024.

Kang, S., Knight, R., et al.: Integrating airborne electromagnetic data and driller's logs to characterize the Central Valley aquifer system, Earth and Space Science, https://doi.org/10.1029/2024EA003958, 2025.

Knight, R., Smith, R., Asch, T., Abraham, J., Cannia, J., Viezzoli, A., and Fogg, G.: Mapping aquifer systems with airborne electromagnetics in the Central Valley of California, Groundwater, 56, 893–908, https://doi.org/10.1111/gwat.12656, 2018.

Laudon, J. and Belitz, K.: Texture and depositional history of Late Pleistocene–Holocene alluvium in the central part of the western San Joaquin Valley, California, Environmental & Engineering Geoscience, bulletin of the Association of Engineering Geologists 28(1); page-level details unverified, 1991.

Lin, C.-Y., Miller, A., Waqar, M., and Marston, L. T.: A database of groundwater wells in the United States, Scientific Data, 11, 335, https://doi.org/10.1038/s41597-024-03186-3, 2024.

Liu, Y. et al.: OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models, Science China Information Sciences, https://doi.org/10.1007/s11432-024-4235-6, arXiv:2305.07895, 2024.

Ma, Z., Santos, J. E., Lackey, G., Viswanathan, H. S., and O'Malley, D.: Information extraction from historical well records using a large language model, Scientific Reports, 14, 31 702, https://doi.org/10.1038/s41598-024-81846-5, 2024.

Marcelli, M. F., Shepherd, M. M., and Faunt, C. C.: CVHM2 well-log lithology database and texture model for the Central Valley, California, U.S. Geological Survey data release, https://doi.org/10.5066/P9IZRO3V, 2022.

Maxwell, K. et al.: State-level policies concerning private wells in the United States, International Journal of Environmental Research and Public Health, 16, 2539, pMC6656387, 2019.

Maxwell, R. M., Condon, L. E., and Kollet, S. J.: A high-resolution simulation of groundwater and surface water over most of the continental US with the integrated hydrologic model ParFlow v3, Geoscientific Model Development, 8, 923–937, https://doi.org/10.5194/gmd-8-923-2015, 2015.

Moosdorf, N., Hartmann, J., and Dürr, H. H.: Lithological composition of the North American continent and implications of lithological map resolution for dissolved silica flux modeling, Geochemistry, Geophysics, Geosystems, 11, Q11 003, https://doi.org/10.1029/2010GC003259, 2010.

O'Neill, M. M. F., Tijerina, D. T., Condon, L. E., and Maxwell, R. M.: Assessment of the ParFlow–CLM CONUS 1.0 integrated hydrologic model, Geoscientific Model Development, 14, 7223–7254, https://doi.org/10.5194/gmd-14-7223-2021, 2021.

OpenAI: GPT-4o System Card, arXiv:2410.21276, 2024.

Pelletier, J. D. et al.: A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses, Journal of Advances in Modeling Earth Systems, 8, 41–65, https://doi.org/10.1002/2015MS000526, 2016.

Rafiei, V.: A national inventory of well lithology and hydraulics for the conterminous United States, https://doi.org/10.5281/zenodo.21196958, data set, Zenodo, 2026.

Riddle, M., Arihood, L. D., Naylor, S., and Lampe, D. C.: A volumetric hydrogeologic framework and texture model for glacial sediments in parts of Michigan, Indiana, and Ohio, Scientific Investigations Report 2025-5008, U.S. Geological Survey, https://doi.org/10.3133/sir20255008, 2025.

Rotzoll, K. and El-Kadi, A. I.: Estimating hydraulic conductivity from specific capacity for Hawaii aquifers, USA, Hydrogeology Journal, 16, 969–979, https://doi.org/10.1007/s10040-007-0271-0, 2008.

State of California: Open and Transparent Water Data Act (AB 1755), https://water.ca.gov/ab1755, chaptered 2016; Stats. 2016, ch. 506, 2016.

State of California: California Water Code Section 13752 — confidentiality of water-well completion reports, California statute, releasable only to authorized agencies or with property-owner consent, 2024.

Twining, B. V., Hodges, M. K. V., Schusler, K., and Mudge, C.: Drilling, Construction, Geophysical Log Data, and Lithologic Log for Boreholes USGS 142 and USGS 142A, Idaho National Laboratory, Idaho, Data Series 1058, U.S. Geological Survey, https://doi.org/10.3133/ds1058, 2017.

Uhlemann, S., Carr, B., Dafflon, B., and Williams, K.: Geophysical borehole logging data of wells ER-GLS1, ER-GUM1, ER-PLM7, and ER-PLM8 at the East River Watershed, Colorado, ESS-DIVE, https://doi.org/10.15485/1650355, 2019.

U.S. Geological Survey: An inventory of three-dimensional geologic models—U.S. Geological Survey, 2004–22, Data Report 1183, U.S. Geological Survey, 2023.

Vahdat-Aboueshagh, H., Tsai, F. T.-C., Bhatta, D., and Paudel, K.: Irrigation-intensive groundwater modeling of complex aquifer systems through integration of big geological data, Frontiers in Water, 3, 623 476, https://doi.org/10.3389/frwa.2021.623476, 2021.

Yang, C., Tijerina-Kreuzer, D. T., Tran, H. V., Condon, L. E., and Maxwell, R. M.: A high-resolution, 3D groundwater-surface water simulation of the contiguous US: ParFlow CONUS 2.0, Journal of Hydrology, 626, 130 294, https://doi.org/10.1016/j.jhydrol.2023.130294, 2023.

Zamrsky, D., Ruzzante, S., Compare, K., Kretschmer, D., Zipper, S., Befus, K. M., Reinecke, R., Pasner, Y., Gleeson, T., Jordan, K., Cuthbert, M., Castronova, A. M., Wagener, T., and Bierkens, M. F. P.: Current trends and biases in groundwater modelling using the community-driven groundwater model portal (GroMoPo), Hydrogeology Journal, 33, 355–366, https://doi.org/10.1007/s10040-025-02882-7, 2025.