



Regional Groundwater Database for Arequipa and Surroundings in Southern Peru from Unstructured Sources Using an Integrated Optical Character Recognition and Large Language Model Workflow

Héctor L. Venegas-Quiñones¹, Madeleine Guillen², Pablo A. Garcia-Chevesich¹³, Edgard González²,
 Javier Ticona², José Díaz², Julia Zea⁴, Francisco Alejo⁴, and John E. McCray¹⁵

¹ Department of Civil and Environmental Engineering, Colorado School of Mines, Golden, CO 80401, USA

² Facultad de Geología, Geofísica y Minas, Universidad Nacional de San Agustín de Arequipa, Arequipa, Peru

³ Intergovernmental Hydrological Programme, UNESCO, Montevideo, Uruguay

⁴ Departamento Académico de Química, Universidad Nacional de San Agustín de Arequipa, Arequipa, Peru

⁵ Hydrologic Science and Engineering Program, Colorado School of Mines, Golden, CO 80401, USA

Correspondence to: Madeleine Guillen (mguillen@unsa.edu.pe)

Abstract

Groundwater management in the arid Andes is constrained less by an absence of measurements than by their inaccessibility: hydraulic parameters measured over six decades remain in unpublished theses and consultancy reports that are not machine-readable and are indexed nowhere. We present a regional groundwater database for Arequipa, southern Peru, containing 4775 georeferenced records recovered from 3675 source documents together with the national monitoring portal, spanning 1966 to 2025. Records carry depth to water, well geometry, saturated thickness, hydraulic conductivity (K), transmissivity (T), storage coefficient (S), lithology and test type, each linked to the page of the document it came from. Extraction combined optical character recognition with a large language model, followed by a three-phase human-in-the-loop protocol in which every recovered value from the principal thesis corpus was checked against its source page; that audit corrected 4.1 % of records, predominantly unit misassignments. Because verifying transcription establishes fidelity but not validity, we additionally evaluate the data against evidence independent of the extraction workflow. Reported transmissivity agrees with the product of reported conductivity and saturated thickness within a factor of two for 94 % of records carrying all three quantities, with a median ratio of 1.00. Transmissivities re-derived independently from primary pumping-test reports released by the national water authority are located in the database, and the extreme values in two basins reproduce figures published in the corresponding official bulletins exactly. Conductivities from saturated-zone pumping tests fall inside published material envelopes in 96 % of cases. The evaluation also exposes the dominant hazard for reuse: the conductivity column mixes seven measurement supports, and pooling them without filtering biases a regional estimate downward by a factor of about 48. The dataset is therefore distributed with explicit unit and support vocabularies and with guidance to filter on test type before any aggregate statistic is computed. Data are archived on HydroShare under CC BY 4.0 (Venegas-Quiñones et al., 2026).

1 Introduction

Quantitative groundwater management requires the parameters that govern aquifer response, namely hydraulic conductivity, transmissivity and storage. Across much of the arid Andes these parameters are



38 nearly absent from the public record, a situation usually described as a data gap. For southern Peru that
39 description is misleading in an important respect. Measurements exist, and have existed for decades, but
40 they are held in university theses, consultancy reports and government inventories that were never
41 digitised, were never assigned persistent identifiers, and are not discoverable by any search a
42 hydrogeologist outside the region could run. The barrier is one of access and structure rather than of
43 measurement.

44 The distinction matters because the two problems have different remedies. A genuine absence of data can
45 only be addressed by new field campaigns, which are slow and costly. An accessibility problem can be
46 addressed by systematic recovery, provided the recovered values can be shown to be trustworthy.
47 Recovery at the necessary scale, thousands of Spanish-language documents, much of it scanned
48 typescript, has until recently been impractical by hand. Optical character recognition combined with large
49 language models makes it tractable (Luscombe et al., 2020; Daqqah, 2024), and comparable recovery
50 efforts have produced national-scale hydrological datasets elsewhere (Venegas-Quñones et al., 2024;
51 Hannaford et al., 2023). That capability introduces its own difficulty: models of this class generate fluent
52 output whether or not it is correct (Zhao et al., 2023; Jančařík and Dušek, 2024), so a database assembled
53 this way carries a burden of proof that a hand-compiled one does not.

54 We treat that burden as the central methodological contribution of this paper. Our position is that
55 verifying transcription, confirming that a number in the database matches the number on the source page,
56 is necessary but not sufficient. A value can be transcribed perfectly and still mislead, because the source
57 itself was in error, because its units were misdeclared, or because it measures a different physical quantity
58 than the column implies. Establishing trustworthiness therefore requires corroboration against evidence
59 independent of the extraction process. We provide four such lines of evidence: internal physical
60 consistency between quantities related by a known equation, agreement between values reported for the
61 same point by different documents, comparison against primary measurement reports and official
62 publications that we re-analysed ourselves, and plausibility against published hydraulic-property
63 envelopes for the relevant materials.

64 The paper is organised accordingly. Section 2 describes the study area, the corpus and the recovery
65 workflow. Section 3 reports the data evaluation and is the core of the paper. Section 4 states the
66 uncertainty budget, separating what further work could reduce from what is irreducible. Section 5
67 describes the data product. Section 6 positions it among comparable compilations. Section 7 gives reuse
68 guidance and states explicitly the interpretations the data will not support. The file inventory and the
69 column dictionary are deliberately not reproduced here: they belong to the dataset and are maintained in
70 the repository README, which is versioned with the data.

71 **2 Study area, corpus and data recovery**

72 **2.1 Study area**

73 The Arequipa region occupies the Pacific flank of the southern Peruvian Andes, rising from sea level to
74 volcanic summits above 6000 m within about 150 km. The regional geology comprises the Precambrian
75 metamorphic basement of the Arequipa Massif (Shackleton et al., 1979), Mesozoic and Cenozoic



76 sedimentary sequences of the forearc basins (Decou et al., 2011), and extensive Neogene to Quaternary
77 volcanic cover associated with subduction of the Nazca plate (Jordán et al., 1983). Precipitation is
78 negligible on the coastal plain, of the order of 200 mm per year in the intermediate zone and up to about
79 850 mm per year on the high plateau, so recharge is generated almost entirely at altitude and moves
80 westward through fractured volcanic sequences before discharging into alluvial valley fill; interbasin flow
81 at high elevation has been demonstrated isotopically in comparable settings of the Western Cordillera
82 (Alvarez-Campos et al., 2022). Groundwater supports irrigated agriculture in the coastal valleys,
83 including the Majes-Siguas scheme (Zapana, 2018; Latorre Farfan et al., 2017), mining in the highland
84 headwaters, and about one million urban inhabitants. The database covers 14.88° S to 16.50° S and
85 71.30° W to 73.30° W.

86 2.2 Source corpus

87 Two structurally different bodies of information exist for this territory and both are represented. The first
88 is the national monitoring network operated by the water authority, which is systematic, spatially broad
89 and almost entirely restricted to static water levels and basic water-quality indicators. The second is an
90 academic and consultancy literature, dominated by university theses and environmental impact
91 assessments, which is spatially clustered but is the only source of measured hydraulic parameters. Table 1
92 summarises the corpus. This asymmetry is not incidental: it is the reason the dataset exists and it
93 conditions how the data should be used.

94 Table 1. Composition of the source corpus by contributing institution. Counts for the university repositories are
95 approximate because a small number of documents are held in more than one repository.

Institution	Type	Documents processed	Records extracted	Time span
UNSA	Theses	~3024	~3044	1966–2024
UCSM	Theses	~400	~500	1980–2023
UCSP	Theses	~250	~300	1990–2022
ANA / SNIRH	Government database	1	931	2010–2025
Total	All	3675	4775	1966–2025

96

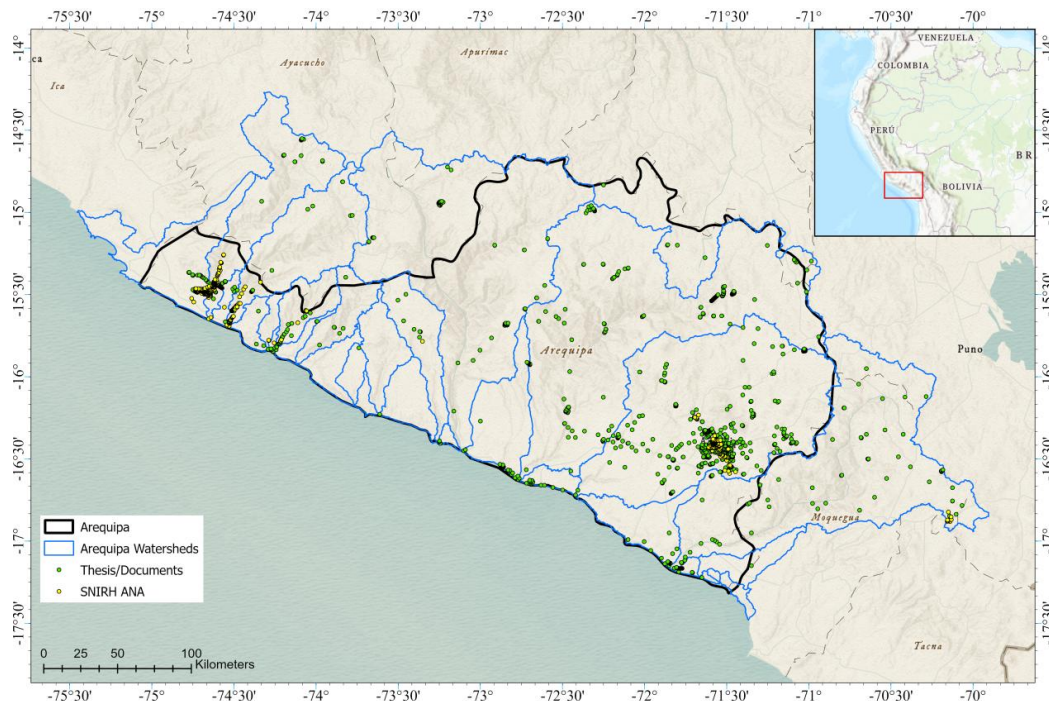


Figure 1. Spatial distribution of groundwater measurement points derived from the study's literature review, encompassing academic theses and technical reports (1969–2024). The national monitoring network follows the coastal highway and the principal irrigation valleys, whereas the recovered documentary records form clusters around mining operations, water-conflict zones and the university city of Arequipa, and supply the only measured hydraulic parameters in the region. Basemaps: Esri Modern Antique (main map) and Esri World Topographic Map (inset), both overlaid on Esri World Hillshade. Basemap sources: Esri, TomTom, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community. Hillshade sources: Esri, Vantor, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap, and the GIS user community | Powered by Esri. Department boundary: simplemaps.com (Pareto Software, LLC), CC BY 4.0. Watershed boundaries: hydrographic units of the Autoridad Nacional del Agua (ANA, 2012).

2.3 Semi-automated extraction from unstructured sources

Candidate documents were identified by scripted harvesting of Peruvian institutional repositories and by direct request to the holding institutions. Scanned documents, which comprise most of the pre-2010 material, were converted with the Tesseract engine, an established open-source system whose behaviour on degraded documents is documented (Luscombe et al., 2020). Recognition was applied page by page and the page index was retained, because the page number is what makes later verification and third-party audit possible.

Structured extraction from the recognised text used Gemini 2.5 Pro through its application programming interface. The model was prompted per document with a fixed schema and instructed to return the source page for every value and to return a null rather than an inferred value when a field was absent. That instruction is the single most important configuration choice, because the failure mode of concern is plausible invention rather than omission. A language model was preferred over regular-expression parsing for three properties that matter with this corpus: tolerance of character-level recognition noise, retention



of associations between a value and a column header separated by intervening layout, which follows from the attention mechanism (Vaswani et al., 2017), and disambiguation of abbreviations that recur with different meanings across documents. We emphasise, because it bears on how this dataset should be judged, that the model was used to locate and transcribe values that already existed in the sources. It was not used to estimate, interpolate or infer any hydraulic parameter, and no value in the released dataset is model-generated.

2.4 Standardisation of units, coordinates and vocabularies

Source documents report hydraulic conductivity in at least fifteen distinct unit strings and transmissivity in nine, including legacy and non-metric forms. Rather than silently converting to a single unit, which would destroy the evidence needed to detect a misdeclaration, we retain the value exactly as printed together with the unit string as printed, and supply a controlled conversion table with the dataset. The consequences of this choice for reuse are quantified in Sects. 3.1 and 7. Coordinates were reported in PSAD56 UTM zones 18S and 19S and in WGS 84; all were transformed to WGS 84 using the official NTv2 grid of Peru's Instituto Geográfico Nacional, with typical shifts of 200 to 300 m in this region, and the original coordinate string and datum are retained for every record.

2.5 Quality assurance: three-phase human-in-the-loop protocol

All 3044 records recovered from the principal thesis corpus were checked against their source page by four trained hydrogeologists using the retained page index, a complete audit of that corpus rather than a sample. Errors were found and corrected in 4.1 % of records (121 of 3044) and no record required rejection. The error spectrum is informative about the failure modes of this class of workflow: about 60 % were unit misassignments, in which the model normalised a legacy or non-metric unit to an SI form the source did not state; about 30 % were coordinate confusions in documents tabulating two datums simultaneously, producing offsets of 200 to 400 m; and about 10 % were over-simplifications of compound lithological descriptions. Two further screens were applied: an automated flag for values beyond three interquartile ranges within a basin returned 145 records that review confirmed as genuine geological extremes rather than errors, and a spatial check of every point against hydrographic boundaries and imagery identified eight additional coordinate errors. We report this audit as evidence of transcription fidelity only. What it cannot establish is whether the transcribed values are themselves reliable, which is the subject of the next section.

3 Data evaluation against independent evidence

Four evaluations are reported. Each uses information that was not available to the extraction workflow, so none can be satisfied by an internally consistent but incorrect extraction. Table 2 summarises the outcomes and Fig. 2 shows the two quantitative tests. All are reproducible from the released evaluation code.

3.1 Unit declarations and the risk they carry

Of the 224 conductivity values, only 11 are declared in metres per day. The remainder are declared in metres per second (40), centimetres per hour in three spellings (77), centimetres per second (41), Spanish



or English per-day variants (31), Lugeon units (16) and a small number of other forms including inches per hour. For transmissivity, 36 of 58 values are declared in units of square metres per second rather than per day. One transmissivity record is declared in litres per second, which is a discharge unit and not a transmissivity; it is flagged in the released data. These declarations are correct in that they faithfully reproduce the sources, and the audit of Sect. 2.5 confirmed them. They nonetheless constitute the largest single hazard in the dataset, because a user who computes a statistic on the raw numeric column without reading the unit column will combine quantities differing by up to eight orders of magnitude.

3.2 Internal physical consistency

Transmissivity, hydraulic conductivity and saturated thickness are related by $T = Kb$. Where a source reports all three, that relation provides a test independent of the extraction process, because a transcription or unit error in any one of the three will break it. Thirty-four records carry all three quantities. After conversion to common units the median ratio of reported T to the product Kb is 1.00, 32 of 34 records (94 %) agree within a factor of two, and all 34 agree within a factor of ten (Fig. 2a). The median absolute value of $\log_{10}(T/Kb)$ is 0.01. This is a demanding test and the dataset passes it: an extraction that had invented values, or that had mixed units within a record, could not produce this result. The two records outside a factor of two are retained and flagged rather than removed, because departures of this size occur legitimately where the reported thickness is the screened interval rather than the full saturated thickness.

3.3 Corroboration between independent documents

Where the same physical point is described by two documents produced independently, the reported values constitute a replicate. Matching records on coordinates rounded to about 11 m, nine points carry a depth to water reported by two or more independent source documents, and all nine agree within a factor of two, with a median ratio between maximum and minimum reported value of 1.00. No point carries a hydraulic conductivity or transmissivity from two independent documents. That absence is itself a result worth stating plainly: the sources are almost entirely non-overlapping, each aquifer test having been performed once and reported once. It follows that redundancy cannot be relied upon for quality control in this dataset, which is precisely why the external corroboration of the next section was necessary rather than optional.

3.4 Corroboration against primary reports and official publications

The national water authority released to us, under Peru's transparency law, the primary pumping-test reports for five tests in the Chili valley as AQTESOLV analyses (ANA, 2018). We re-derived the fitted transmissivity for each test independently from those reports, obtaining ten values, one from the drawdown and one from the recovery analysis of each test. Six of the ten are present in the database within 2 % of the independently re-derived figure. The pattern of the four absences is systematic rather than random: the recovery analysis is present for all five tests, whereas the drawdown analysis is present for only one. The compilation therefore captured one of the two analyses reported for each test, a completeness limitation that we quantify here and that users of the transmissivity column should know about; the missing drawdown analyses have since been added to the working version of the data.



195 A second external comparison uses values published in the official INGEMMET hydrogeological
 196 bulletins, compiled independently of this work. In the Tambo basin the bulletin reports a single seven-day
 197 test at Cocachacra with a conductivity of about 125 m d^{-1} and a transmissivity of about $9900 \text{ m}^2 \text{ d}^{-1}$
 198 (Carpio and Peña, 2020); the maxima in the database for that basin are 125.0 m d^{-1} and $9900 \text{ m}^2 \text{ d}^{-1}$,
 199 reproducing the published figures exactly. In the Quilca-Vitor-Chili basin the bulletin reports a 2002
 200 campaign with conductivities between 4.7 and 138.5 m d^{-1} and transmissivities between 216 and 7260 m^2
 201 d^{-1} (Carpio et al., 2022); 51 of the 96 database conductivities and 13 of the 22 transmissivities for that
 202 basin fall inside those ranges, the remainder lying outside because the database also contains tests of other
 203 types and from other campaigns, as Sect. 3.5 makes clear. In the Caravelí basin both database
 204 transmissivities fall inside the published range of 27 to $144 \text{ m}^2 \text{ d}^{-1}$ (Machaca et al., 2022).

205 **3.5 Plausibility against published envelopes, and the measurement-support problem**

206 Restricting attention to records whose lithological description names exactly one material family, so that
 207 the expected range is unambiguous, 19 of 21 sand records (90 %) and all 19 volcanic-rock records fall
 208 inside the conductivity envelopes published for those materials (Freeze and Cherry, 1979; Domenico and
 209 Schwartz, 1998). Ten records described as gravel fall entirely outside the gravel envelope, with a median
 210 conductivity of 0.99 m d^{-1} against an expected 86 m d^{-1} or more. Rather than treat this as an error we
 211 traced it, and the explanation is diagnostic for the dataset as a whole: all ten are cylinder-infiltrometer or
 212 Lefranc tests and their lithology strings are agricultural soil-texture descriptions such as sandy loam with
 213 gravel. They are measurements of near-surface infiltration in the unsaturated zone, not of saturated
 214 aquifer conductivity, and they are correctly transcribed measurements of a different physical quantity.

215 This is a general property of the conductivity column and the most consequential finding of the
 216 evaluation. The 224 conductivity values comprise seven measurement supports: 83 cylinder-infiltrometer
 217 tests, 55 saturated-zone pumping or recovery tests, 25 packer tests in rock, 16 laboratory permeameter
 218 determinations, 13 Lefranc tests, 9 auger-hole tests and 23 unspecified. Restricted to the 55 pumping
 219 tests, the median conductivity is 23.0 m d^{-1} and 96 % of values fall inside the sand-to-gravel envelope.
 220 The remaining 169 records have a median of 0.48 m d^{-1} . Pooling the column without filtering therefore
 221 depresses a regional conductivity estimate by a factor of about 48 (Fig. 2b). The dataset is not wrong; the
 222 column is heterogeneous by construction, because the underlying literature is. The remedy is filtering, not
 223 correction, and Sect. 7 states how.

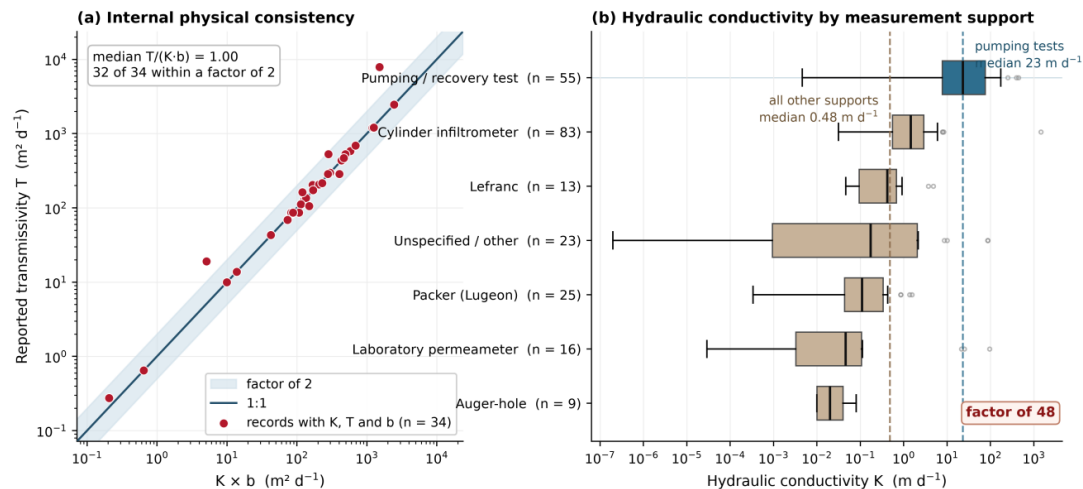


Figure 2. Evaluation of the recovered hydraulic parameters against independent evidence. (a) Internal physical consistency: reported transmissivity against the product of reported conductivity and saturated thickness for the 34 records carrying all three quantities, with the 1:1 line and a factor-of-two band. (b) Hydraulic conductivity by measurement support, plotted on a logarithmic axis. Saturated-zone pumping and recovery tests (blue) are the subset appropriate for aquifer characterization; the remaining supports measure near-surface infiltration, rock-mass permeability or sample-scale permeability, and their inclusion in an aggregate statistic biases it downward by about a factor of 48.

Table 2. Summary of the data evaluation. Each test draws on evidence independent of the extraction workflow.

Evaluation	Records tested	Outcome	What it establishes
Transcription audit against source page	3044 (complete audit of the thesis corpus)	4.1 % corrected; no record rejected	Fidelity of transcription, not validity of values
Internal consistency, T against Kb	34	median ratio 1.00; 94 % within a factor of 2	Mutual coherence of K, T and b; excludes invented values
Replication between independent documents	9 points (depth to water)	9 of 9 within a factor of 2	Agreement where redundancy exists; shows redundancy is rare
Re-derivation from primary pumping-test reports	10 transmissivity values	6 of 10 located within 2 %; absences systematic	Traceability to primary evidence and a completeness gap
Comparison with official published bulletins	3 basins	extremes reproduce published values exactly	Consistency with independent official compilations
Plausibility against material envelopes	55 pumping tests	96 % inside published envelopes	Physical reasonableness of the intended subset

4 Uncertainty budget

We distinguish uncertainty that further work could reduce from uncertainty inherent to the measurements and irreducible by any treatment of the compilation. Table 3 states both, with magnitudes where we can quantify them. The distinction is practical rather than philosophical: it tells a prospective user which limitations they might overcome with effort and which they must design around.

Table 3. Uncertainty budget, separating reducible (epistemic) from irreducible (aleatory) components.

Component	Type	Magnitude or extent	Reducible?
Transcription error remaining after the complete audit	Epistemic	below the 4.1 % pre-audit rate; residual not directly measurable	Yes, by independent re-audit
Unit misdeclaration in the source document	Epistemic	60 % of the 121 corrected records	Yes, by consulting the source



Component	Type	Magnitude or extent	Reducible?
Coordinate uncertainty, documentary stream	Epistemic	200–400 m where datums were conflated; larger where inferred from maps	Partly, by field checking
Measurement support mismatch in the K column	Epistemic	factor of about 48 between pumping tests and other supports	Yes, by filtering on test type
Completeness of extraction per source	Epistemic	one of two analyses captured in the ANA pumping-test check	Yes, by targeted re-extraction
Aquifer heterogeneity within a basin	Aleatory	K spans four to five orders of magnitude regionally	No
Interpretation of the original pumping test	Aleatory	drawdown and recovery analyses of one test differ by up to an order of magnitude	No
Temporal non-stationarity of water levels	Aleatory	records span 1966–2025 under changing abstraction	No

239

240 5 Data description

241 The dataset comprises 4775 georeferenced records with 56 fields covering identification and provenance,
 242 location and datum, well geometry, water level, hydraulic parameters with their declared units, test type,
 243 lithology, aquifer condition and water quality. Depth to water is available for the largest number of
 244 records, followed by saturated thickness, hydraulic conductivity, transmissivity and, most sparsely,
 245 storage coefficient. Temporal coverage spans 1966 to 2025, with the greatest density after 2000. The
 246 spatial distribution (Fig. 1) reveals both data-rich zones in the coastal irrigation valleys and knowledge
 247 gaps in high-elevation areas vulnerable to climate-induced hydrological shifts.

248 The repository is deliberately organised so that the final curated dataset is separable from the evidence it
 249 was derived from. Part A contains the analysis-ready table in CSV and spreadsheet form, the unit-
 250 conversion table, the test-type vocabulary and the README. Part B contains the complete archive of
 251 source documents, named so that the source field of any record resolves directly to a file, together with
 252 the traceability guide and the source catalogue. A user who wants only the analysis-ready table need not
 253 download the document archive. The complete file inventory and the column-level dictionary, including
 254 fill rates for every field, are held in the README, which is generated from the released file so that the
 255 documentation cannot drift from the data. They are not reproduced in this paper because the README,
 256 not the paper, is versioned with the dataset.

257 Every record carries the source document name and the page number from which the value was taken. To
 258 verify any value a user locates the document in the Part B archive and opens the cited page. This
 259 traceability is what allows the evaluation of Sect. 3 to be repeated or extended by a third party.

260 6 Relation to comparable datasets

261 This dataset occupies a position not filled by existing compilations, and stating that position precisely is
 262 the fairest way to convey its value. Global compilations of subsurface permeability provide continuous
 263 coverage but are derived from lithological mapping rather than measurement (Gleeson et al., 2011), so
 264 they cannot supply a measured conductivity for a specific well. National monitoring portals, including the
 265 Peruvian SNIRH from which part of this compilation is drawn, provide systematic water levels but, as
 266 Sect. 2.2 shows, essentially no hydraulic parameters; the interoperability infrastructure developed for such
 267 networks addresses distribution rather than the absence of parameters (Blodgett et al., 2015; Brodaric et



al., 2015). National groundwater-level datasets assembled from official records, such as the Chilean compilation produced by part of this author team (Venegas-Quiñones et al., 2024), and national projection datasets such as the enhanced Future Flows and Groundwater product (Hannaford et al., 2023), share the recovery and curation philosophy but address water levels or simulated series rather than measured aquifer properties. Regional hydrogeological studies in Peru provide authoritative interpretation but rest on a small number of in-situ tests and assign permeability largely by lithological analogy (Velasco and Capilla, 2019; Carpio et al., 2022).

What is distinctive here is therefore neither the region nor the variables, but the provenance: a compilation of individually traceable, in-situ measured hydraulic parameters recovered from literature that was previously unusable, each value linked to a page of a named document and released alongside the documents themselves. To our knowledge it is the first such compilation for the Peruvian Andes. The evaluation of Sect. 3 is offered as a template that other recovery efforts can apply, since the four tests require no information beyond what any such compilation already contains.

7 Data usage, reuse guidance and known pitfalls

The dataset supports several uses directly. It provides prior distributions and calibration targets for numerical groundwater models in the region; it supplies the parameter fields needed for aquifer vulnerability mapping when combined with depth to water; it enables sustainable-yield assessment where transmissivity and storage are available; and it establishes pre-development baselines, since a substantial share of records predate 1990 and therefore precede the expansion of large-scale irrigation and mining. These applications are directly relevant to projected water stress in the region (Chen et al., 2025; Quiroz et al., 2023). Three cautions apply, and we state them as instructions rather than caveats.

First, never compute a statistic on a numeric column without reading its unit column. Only 11 of the 224 conductivity values are declared in metres per day and 36 of 58 transmissivities are declared in per-second units. A conversion table is distributed with the data, and the released files include a derived column in harmonised units for convenience, but the source value and source unit remain the authoritative record.

Second, filter on measurement support before aggregating hydraulic conductivity. For aquifer characterization use records whose test type is a pumping or recovery test; infiltrometer, auger-hole and laboratory determinations measure different quantities at different scales and belong to soil physics rather than aquifer hydraulics. Ignoring this depresses a regional conductivity estimate by roughly a factor of 48, larger than any other error in the dataset.

Third, do not read spatial density as hydrogeological importance, and do not read the absence of a repeat measurement as stability. The distribution of records follows historical research and licensing priorities, so a basin with many records may hold no hydraulic parameters at all, while a basin with few records may simply never have been studied. Only a small fraction of wells were revisited, and almost all of those lie in one basin, so trend analysis is defensible there and is not supported elsewhere. Storage coefficients deserve particular reserve: there are 27 in the entire compilation and most derive from single-well tests, which do not determine storage reliably.



306 **8 Limitations and recommendations**

307 The database is limited by the quality and completeness of the original source materials. Many entries
 308 contain missing or estimated rather than measured values, particularly in older documents where certain
 309 measurements were not standard practice. Temporal inconsistency complicates the establishment of
 310 baselines and the detection of long-term trends, because measurement protocols evolved over the six
 311 decades covered. The spatial distribution reflects historical research priorities, with clustering in areas of
 312 mining interest, agricultural development and urban centres, while remote high-altitude regions remain
 313 under-represented. Reliance on published reports and theses means that some proprietary data from
 314 private companies are missing. All sources are in Spanish, and language-specific recognition artefacts,
 315 such as misread accented characters in older typescript, may introduce subtle errors not fully captured by
 316 the verification protocol.

317 For researchers developing similar databases elsewhere we recommend establishing data standards and
 318 validation protocols at the outset; building a crosswalk between coordinate systems and between the
 319 terminologies used across historical documents; retaining provenance to the page for every value, because
 320 that is what makes both verification and corroboration possible; and engaging local institutions early,
 321 since university thesis repositories proved to be valuable and frequently overlooked sources. Above all we
 322 recommend planning for corroboration and not only for verification. Our experience is that a complete
 323 manual audit, which is expensive, establishes only that transcription is faithful; the four independent tests
 324 in Sect. 3 cost far less and are what actually established that the values can be trusted. Future compilations
 325 built with machine assistance should budget for both.

326 **9 Data availability**

327 The database is archived on HydroShare under a Creative Commons Attribution 4.0 licence at
 328 <https://www.hydroshare.org/resource/1dfb97e6d48d7b307591868ca70c4/> (Venegas-Quiñones et al.,
 329 2026). The repository contains the curated dataset, the complete collection of source documents, the
 330 controlled vocabularies and unit-conversion table, and a README holding the full file inventory and the
 331 column-level dictionary. The Python scripts developed for harvesting Peruvian institutional repositories
 332 and for connecting to the extraction interface are available at
 333 <https://github.com/venegasquinones/Scrapers-de-Repositorios-Institucionales-de-Per->. The evaluation
 334 code that reproduces every number in Sect. 3 is released with the same repository.

335 **10 Conclusions**

336 We have shown that six decades of hydraulic measurements for a data-scarce Andean region can be
 337 recovered from unstructured sources, and, more importantly, that the recovered values can be shown to be
 338 trustworthy by evidence independent of the recovery process. The internal relation between
 339 transmissivity, conductivity and thickness is satisfied within a factor of two in 94 % of testable records;
 340 transmissivities re-derived from primary pumping-test reports are found in the database; the extreme
 341 values in two basins reproduce official published figures exactly; and conductivities from the appropriate
 342 subset of tests fall inside published material envelopes in 96 % of cases.



343 The evaluation also determined what the data cannot do. Redundancy between sources is almost absent,
344 so quality control cannot rely on replication. Extraction completeness is imperfect in a way that is
345 systematic rather than random. And the conductivity column mixes seven measurement supports whose
346 pooled median differs from the aquifer-relevant subset by a factor of about 48, the dominant hazard for
347 anyone reusing the data.

348 We suggest this pattern generalises. Where a compilation is assembled with machine assistance, verifying
349 transcription against the source establishes fidelity only. Trustworthiness requires corroboration against
350 independent evidence, and the four tests used here, internal physical consistency, replication between
351 sources, re-derivation from primary reports, and plausibility against published envelopes, need no
352 information that such a compilation does not already contain. We therefore offer them as a minimum
353 standard for machine-assisted data recovery in the geosciences.

354 **Author Contributions**

355 HLV-Q and PAG-C wrote the manuscript's first draft. JEM reviewed the first draft. MG, EG, JT, JD, JZ,
356 and FA contributed additional data by formally requesting information from agencies and reviewed the
357 first draft. MG, EG, JT, and JD performed the manual verification of LLM-extracted records (Phase 1
358 quality assurance). HLV-Q developed the Python codes and machine learning algorithms to extract the
359 information, build the database, and carry out the data evaluation of Sect. 3. FA funded the project. PAG-
360 C and JEM supervised the team.

361 **Competing Interests**

362 The authors declare that they have no conflict of interest.

363 **Disclaimer**

364 Even though all information in this database was carefully checked through a three-phase quality
365 assurance protocol and independently evaluated as described in Sect. 3, errors may still exist. The authors
366 advise users to consult the original source document, linked via the source document and page reference
367 fields, if verification is needed.

368 **Acknowledgements**

369 The authors thank the support of the Center for Mining Sustainability (Grant #470266), a joint venture
370 between Colorado School of Mines and Universidad Nacional de San Agustín. We thank the Autoridad
371 Nacional del Agua for releasing the primary pumping-test reports used in the corroboration of Sect. 3.4.
372 During the preparation of this manuscript, artificial intelligence tools were used to assist with the
373 structural organisation of the text and the synthesis of complex information; the role of automated tools in
374 constructing the dataset itself is described in full in Sect. 2.3, and all reported values originate in the
375 source documents rather than from any model.



References

- Alvarez-Campos, O., Olson, E. J., Welp, L. R., Frisbee, M. D., Zuñiga Medina, S. A., Díaz Rodríguez, J., Roque Quispe, W. R., Salazar Mamani, C. I., Arenas Carrión, M. R., Jara, J. M., Ccancapa-Cartagena, A., and Jafvert, C. T.: Evidence for high-elevation salar recharge and interbasin groundwater flow in the Western Cordillera of the Peruvian Andes, *Hydrol. Earth Syst. Sci.*, 26, 483–503, <https://doi.org/10.5194/hess-26-483-2022>, 2022.
- ANA: Delimitación y codificación de unidades hidrográficas del Perú: Memoria descriptiva, Autoridad Nacional del Agua, Dirección de Conservación y Planeamiento de Recursos Hídricos, Lima, Peru, <https://hdl.handle.net/20.500.12543/597>, 2012.
- ANA: Estudio hidrogeológico del acuífero del río Chili (Arequipa), Autoridad Nacional del Agua, Dirección de Calidad y Evaluación de Recursos Hídricos, Lima, Peru, obtained by public-information request under Law 27806 (Carta N° 0233-2026-ANA-DCERH), 2018.
- Blodgett, D., Lucido, J., and Kreft, J.: Progress on water data integration and distribution: a summary of select US Geological Survey data systems, *J. Hydroinform.*, 18, 226–237, <https://doi.org/10.2166/hydro.2015.067>, 2015.
- Brodaric, B., Booth, N., Boisvert, E., and Lucido, J.: Groundwater data network interoperability, *J. Hydroinform.*, 18, 210–225, <https://doi.org/10.2166/hydro.2015.242>, 2015.
- Campos, O. A., Olson, E., Welp, L. R., and Frisbee, M. D.: Groundwater chemistry within Arequipa, Peru in the Characato, Chiguata, and Lagunas Salinas study areas, *HydroShare*, 2022.
- Carpio, J. and Peña, F.: Hidrogeología de la cuenca del río Tambo (1318), regiones Arequipa, Moquegua y Puno, *Boletín Serie H: Hidrogeología N° 7*, Instituto Geológico, Minero y Metalúrgico, Lima, Peru, 288 pp., 2020.
- Carpio, J., Quispe, B., Peña, F., and Sulca, P.: Hidrogeología de la cuenca del río Quilca-Vitor-Chili (132), *Boletín Serie H: Hidrogeología N° 15*, Instituto Geológico, Minero y Metalúrgico, Lima, Peru, 381 pp., 2022.
- Chen, M., Huang, Y., Hu, X.-M., Xue, M., Hong, Y., Novoa, H. M., Martin, E. R., McPherson, R. A., Zhu, S., and Peraz, A. V.: Arequipa's Water in the Short Future: a Hydrologic Outlook in an Arid Peruvian Andes Region Utilizing Hyperresolution RCM and CREST-VEC Model Simulations Under SSP5-8.5, *Am. J. Water Resour.*, 13, 51–62, 2025.
- Cheneaux, A., Damborenea, C., Moreno-S, C., Cruz-R, C., and Coayla-Peñaloza, P.: Aquatic macroinvertebrate community and water quality in the Tambo River, Arequipa, Perú, *Rev. Biol. Trop.*, 72, <https://doi.org/10.15517/rev.biol.trop.v72i1.56670>, 2024.
- Contos, J. and Tripcevic, N.: Correct placement of the most distant source of the Amazon River in the Mantaro River drainage, *Area*, 46, <https://doi.org/10.1111/area.12069>, 2014.
- Daqqah, B. H.: Leveraging Large Language Models (LLMs) for Automated Extraction and Processing of Complex Ordering Forms, MS thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2024.
- Decou, A., Eynatten, H., Mamani, M., Sempere, T., and Wörner, G.: Cenozoic forearc basin sediments in Southern Peru (15–18° S): Stratigraphic and heavy mineral constraints for Eocene to Miocene evolution of the Central Andes, *Sediment. Geol.*, 237, 55–72, <https://doi.org/10.1016/j.sedgeo.2011.02.004>, 2011.
- Domenico, P. A. and Schwartz, F. W.: Physical and Chemical Hydrogeology, 2nd edn., John Wiley and Sons, New York, USA, 506 pp., 1998.
- Erwin, A., Silva, C. A., and Ma, Z.: Self-organization for community resilience in an invisible agricultural community, *Front. Sustain. Food Syst.*, 7, <https://doi.org/10.3389/fsufs.2023.1160109>, 2023.
- Freeze, R. A. and Cherry, J. A.: Groundwater, Prentice-Hall, Englewood Cliffs, NJ, USA, 604 pp., 1979.
- Gleeson, T., Smith, L., Moosdorf, N., Hartmann, J., Dürr, H. H., Manning, A. H., van Beek, L. P. H., and Jellinek, A. M.: Mapping permeability over the surface of the Earth, *Geophys. Res. Lett.*, 38, L02401, <https://doi.org/10.1029/2010GL045565>, 2011.
- Gramlich, D. and Walker, T.: Water risk modeling: A framework for finance, *J. Environ. Manage.*, 342, 117991, <https://doi.org/10.1016/j.jenvman.2023.117991>, 2023.
- Hannaford, J., Mackay, J. D., Ascott, M., Bell, V. A., Chitson, T., Cole, S., Counsell, C., Durant, M., Jackson, C. R., Kay, A. L., Lane, R. A., Mansour, M., Moore, R., Parry, S., Rudd, A. C., Simpson, M., Facer-Childs, K., Turner, S., Wallbank, J. R., and Wilcox, A.: The enhanced future Flows and Groundwater dataset: development and



- 424 evaluation of nationally consistent hydrological projections based on UKCP18, *Earth Syst. Sci. Data*, 15, 2391–
425 2415, <https://doi.org/10.5194/essd-15-2391-2023>, 2023.
- 426 INRENA: Inventario de fuentes de aguas subterráneas en el valle de Yauca: Informe final, Instituto Nacional de Recursos
427 Naturales, Lima, Peru, <https://hdl.handle.net/20.500.12543/3318>, 2002.
- 428 Jančařík, A. and Dušek, O.: The problem of AI hallucination and how to solve it, in: European Conference on e-Learning,
429 2024.
- 430 Jordán, T. E., Isacks, B. L., Allmendinger, R. W., Brewer, J. A., Ramos, V. A., and Ando, C. J.: Andean tectonics related
431 to geometry of subducted Nazca plate, *Geol. Soc. Am. Bull.*, 94, 341–361, 1983.
- 432 Latorre Farfan, J. P., Kristensen, I., and Jacobsen, S.-E.: Is quinoa cultivation on the coastal desert of Peru sustainable? A
433 case study from Majes, Arequipa, 2017.
- 434 Luscombe, A., Dick, K., Duncan, J., and Walby, K.: Access to Information and Optical Character Recognition (OCR): A
435 Step-by-Step Guide to Tesseract, Part One of the CAIJ Computer Literacy Series, 2020.
- 436 Machaca, D., Peña, F., and Vasquez, J.: Hidrogeología de la cuenca del río Pescadores-Caravelí, región Arequipa, *Boletín*
437 *Serie H: Hidrogeología* N° 14, Instituto Geológico, Minero y Metalúrgico, Lima, Peru, 212 pp., 2022.
- 438 Quiroz, J. A., Garcia-Chevesich, P. A., Martínez, G., Martínez, K., Tejada-Purizaca, T., Murray, K. E., and McCray, J. E.:
439 Water Resources Evaluation and Sustainability Considering Climate Change and Future Anthropic Demands in the
440 Arequipa Region of Southern Peru, *Sustainability*, 15, 16270, <https://doi.org/10.3390/su152316270>, 2023.
- 441 Rymes, M. N., Rasmussen, R., Garcia-Chevesich, P. A., Burgert, K., Long, J., McBride, E., Martínez, G., Martínez, K.,
442 Tejada-Purizaca, T. R., Murray, K. E., Vanzin, G., Sharp, J. O., and McCray, J. E.: Surface water quality database
443 for five watersheds in the Arequipa region of Southern Peru, *Data in Brief*, 56, 110770,
444 <https://doi.org/10.1016/j.dib.2024.110770>, 2024.
- 445 Shackleton, R., Ries, A., Coward, M., and Cobbold, P. R.: Structure, metamorphism and geochronology of the Arequipa
446 Massif of coastal Peru, *J. Geol. Soc.*, 136, 195–214, 1979.
- 447 Vargas Vilchez, L.: Geología del cuadrángulo de Arequipa – *Boletín A 24, INGEMMET*, Lima, Peru,
448 <https://hdl.handle.net/20.500.12544/142>, 1970.
- 449 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is
450 all you need, in: *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- 451 Velasco, A. and Capilla, J. E.: Hydrogeological characterization and assessment of anthropic impacts in the Lower Piura
452 Sub-basin Aquifer in Peru, *Hydrogeol. J.*, 27, 2755–2773, <https://doi.org/10.1007/s10040-019-02027-7>, 2019.
- 453 Venegas-Quñones, H. L., Valdés-Pineda, R., García-Chevesich, P., Valdés, J. B., Gupta, H. V., Whitaker, M. P. L., and
454 Ferré, T. P. A.: Development of Groundwater Levels Dataset for Chile since 1970, *Sci. Data*, 11, 170,
455 <https://doi.org/10.1038/s41597-023-02895-5>, 2024.
- 456 Venegas-Quñones, H. L., Garcia-Chevesich, P. A., Guillen, M., Alejo, F., and McCray, J. E.: Addressing the Problem of
457 Hard-to-Reach Unpublished Data from Theses in University Repositories, *Groundwater*, 63, 922–925,
458 <https://doi.org/10.1111/gwat.70007>, 2025.
- 459 Venegas-Quñones, H. L., Guillen, M., Garcia-Chevesich, P. A., Uhle, B., González, E., Ticona, J., Díaz, J., Zea, J., Alejo,
460 F., and McCray, J.: Regional Groundwater Database for Arequipa and Surroundings, *HydroShare*,
461 <http://www.hydroshare.org/resource/1dfbde97e6d48d7b307591868ca70c4>, 2026.
- 462 Vera Delgado, J. and Vincent, L.: Community Irrigation Supplies and Regional Water Transfers in the Colca Valley, Peru,
463 *Mt. Res. Dev.*, 33, 195–206, 2013.
- 464 Zapana, L.: Responses to the water crisis in agricultural and urban areas: Case study “Irrigation Project Majes Siguan I”
465 Arequipa, Peru, 2018.
- 466 Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., and Wen, J. R.: A survey of large language models, *arXiv*
467 preprint, *arXiv:2303.18223*, 2023.