PROCEEDINGS, 51st Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 9-11, 2026
SGP-TR-230

Mining Legacy Geothermal Databases with Unsupervised Learning for Dual-Resource Exploration

Lokesh Kumar SEKAR, Esuru Rita OKOROAFOR

Harold Vance Department of Petroleum Engineering, Texas A&M University, College Station, TX, USA

lokesh_kumar_sekar@tamu.edu

Keywords: Geologic Hydrogen, Geothermal, Dual Resource Exploration, Unsupervised Machine Learning

ABSTRACT

Legacy geothermal datasets provide a rich yet underutilized resource for advancing data-driven subsurface exploration. We applied unsupervised machine learning to a digitized U.S. Geological Survey database containing over 1,800 water and 300 gas chemistry samples (1930–2006) from geothermal and hot spring systems across the western United States. The objectives were twofold: (a) to delineate geothermal provinces and fluid sources; (b) to identify sites with geochemical conditions conducive to geological hydrogen generation through serpentinization of ultramafic rocks. A multi-stage analysis was performed using k-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Gaussian Mixture Models on standardized chemical, isotopic, and temperature data. Geochemical fingerprinting revealed coherent clusters representing magmatic–volcanic, sedimentary, and deep-circulation systems, providing analogs to productive geothermal fields and flagging unexplored zones. Cluster tagging based on high pH–Mg waters and reducing conditions identified candidate ultramafic settings favorable for H_2 formation and preservation. Temporal analyses across seven decades corrected historical biases and improved confidence in H_2 prospectivity screening. This framework illustrates how legacy datasets, when integrated with unsupervised learning, can delineate geothermal provinces, track geochemical evolution, and identify ultramafic terrains for exploration and engineering of hydrogen generation systems. The results highlight the untapped potential of historical data to accelerate sustainable subsurface energy discovery.

1. INTRODUCTION

Geothermal energy and geologic hydrogen are rapidly emerging as reliable, low-carbon baseload resources. Advances in subsurface engineering, particularly in Enhanced Geothermal Systems, together with underground thermal energy storage, are enabling continuous heat extraction from deep and low-permeability formations (Sekar and Okoroafor, 2024). In parallel, both naturally occurring and engineered geologic hydrogen systems are gaining momentum, leveraging fluid–rock reactions to generate and store hydrogen in the subsurface. Within this framework, water geochemistry becomes a powerful exploration tool, revealing thermal regimes, reaction pathways, and hydrogen-generation signatures that directly guide geothermal and geologic hydrogen prospecting. Legacy geothermal databases contain decades of hydrochemical and gas measurements that remain largely underexploited for modern subsurface energy exploration (Ahmmed and Vesselinov, 2022). Recent studies have demonstrated that shallow groundwater geochemistry, when analyzed with unsupervised machine learning, can reveal latent geothermal signatures even within sparse, heterogeneous legacy datasets. Ahmmed and Vesselinov (2022) showed that non-negative matrix factorization applied to regional geochemical data successfully delineates geothermal prospectivity without reliance on expert-defined thresholds or play fairway weighting. Recent work has also demonstrated that unsupervised machine learning can quantify the relative importance of geothermal exploration attributes, eliminating the need for subjective weighting in play fairway analysis. Mudunuru et al. (2023) showed that non-negative matrix factorization can extract hidden geothermal signatures and objectively rank thermal, structural, and geochemical controls on prospectivity in the Tularosa Basin.

Groundwater geochemistry also contains latent information that can distinguish low, medium, and high-temperature geothermal systems. Ahmmed et al. (2021) showed that unsupervised machine learning can extract temperature-dependent geochemical signatures across the Great Basin, revealing systematic transitions from major-ion-dominated low-temperature systems to trace-element-dominated high-temperature systems. Trace-element geochemistry has long been recognized as a powerful indicator of deep hydrothermal and geothermal processes. Moradpouri and Sabeti (2024) demonstrated that multivariate analysis of stream-sediment trace elements can identify geothermal pathfinders and subsurface mineralization, with arsenic-centered elemental associations reflecting deep fluid circulation and volcanic heat sources.

Previous studies have demonstrated that geothermal reservoir temperature can be reliably inferred from hydrogeochemical data using machine-learning methods. Tut Haklidir and Haklidir (2020) showed that deep neural networks accurately predict reservoir temperatures in Western Anatolia using multivariate chemical indicators, outperforming classical regression and geothermometry-based approaches. Building on these foundations, the present study leverages unsupervised learning of legacy hydrogeochemical datasets without pre-labeled classes to extract latent, attribute- and trace-element–informed geochemical signatures, enabling dual-resource screening of subsurface fluid regimes favorable for both geothermal energy extraction and geological hydrogen generation through ultramafic serpentinization.

In this study, we apply unsupervised machine learning to integrated water and gas chemistry datasets from geothermal systems across the western United States (Mariner et al., 1977; Fournier, 1980) to delineate geochemical provinces and fluid sources. By systematically grouping physical, chemical, and isotopic attributes, we extract coherent signatures of magmatic, sedimentary, meteoric, and ultramafic-

influenced systems. This data-driven framework further identifies geochemical conditions favorable for geological hydrogen generation, demonstrating the untapped value of historical datasets for dual-resource geothermal and hydrogen exploration.

2. DATA ANALYSIS

The data analysis section describes the data sourcing and compilation of legacy hydrogeochemical datasets, followed by data preprocessing and exploratory data analysis.

2.1 Data Overview

For this study, we compiled a legacy geothermal water-chemistry database comprising ~245 water samples and ~104 gas samples (1930s–2000s) from geothermal springs and wells across the western United States (primarily California, Oregon, and Washington). This dataset is an excerpt of a larger USGS collection (originally ~1,800 water and 300 gas analyses) digitized from historical reports (Mariner et al., 1977; Fournier, 1980). The data and their relevance to our study are shown in Appendix 1.

2.2 Data Preprocessing

The raw data required careful cleaning. We standardized units and handled missing or below-detection values. Non-detected entries marked as "---" were set to null. We merged multiple data files from different decades and sources, ensuring consistent column definitions. After cleaning, numeric features were log-transformed (for highly skewed concentration distributions) and then standardized (zero-mean, unit-variance) for clustering. Notably, older records (pre-1980) often lacked a full ionic analysis. For example, only ~54 of 148 samples in one 1990s sub-dataset had complete analyses of Ca, Mg, K, and SO_4. To avoid bias from incomplete vectors, we focused clustering on samples with a complete major-ion profile. This reduced the working set to ~75 water samples with all major parameters (Na, K, Ca, Mg, Cl, SO_4, HCO_3/alkalinity, SiO_2, pH, temperature) reported. Geochemical variables spanned several orders of magnitude, so $\log_{10}$-transformation was applied to concentrations prior to standardization to prevent extreme salinities (e.g., brine samples) from unduly dominating cluster-distance calculations. Spatially, the sample set spans volcanic arcs, sedimentary basins, and crystalline provinces of the U.S. West, providing a broad basis for delineating geothermal provinces, as stated in our objectives.

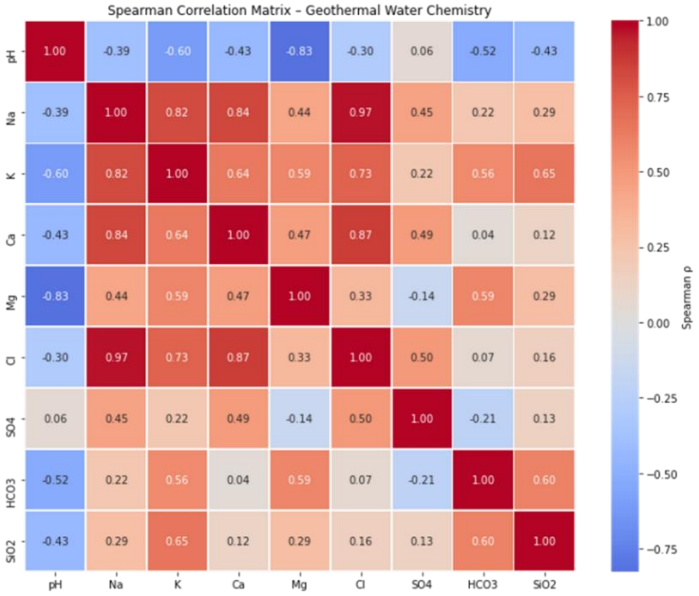

Figure 1: Spearman Correlation analysis.

2.3 Exploratory Data Analysis

Initial exploratory data analysis using Spearman's rank correlation (Figure 1) revealed distinct geochemical trends corresponding to water origins and thermal histories. Figure 1 illustrates a strong positive correlation between Na and Cl ($\rho \approx 0.97$), indicating that salinity in these geothermal fluids is primarily controlled by sodium chloride content. Chloride also correlates tightly with Ca ($\rho \approx 0.87$), suggesting that the most chloride-enriched waters often carry high calcium as well. Many of these are likely brines or magmatic waters where prolonged water–rock interaction or evaporation has enriched both Cl and Ca (Nicholson, 1993).

By contrast, bicarbonate (HCO_3) shows little correlation with Cl (ρ ~0.07), but correlates positively with Mg ($\rho \approx 0.59$). This HCO_3 + Mg coupling is characteristic of meteoric groundwater that has equilibrated with carbonate minerals or Mafic rocks, and has also been observed in serpentines. Also, as groundwater gains alkalinity (HCO_3) from CO_2 and mineral dissolution, it often picks up Mg from silicate weathering (Bruni et al., 2002). We also find that pH is inversely correlated with both HCO_3 and Mg ($\rho \approx$ –0.52 and –0.83, respectively), an indicator that the highest-pH waters are extremely low in dissolved inorganic carbon and divalent cations. This inverse relation foreshadows the presence of "hyperalkaline" waters (high pH, very low dissolved CO_2 and Mg) likely produced by water–ultramafic rock interaction (serpentinization) (Bruni et al., 2002).

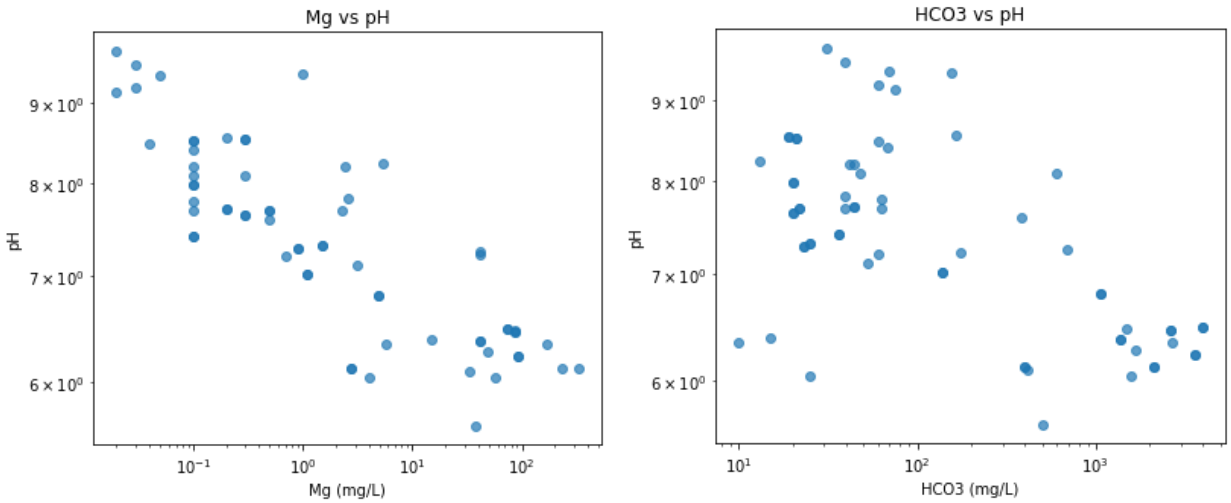

Figure 2: Correlation analysis.

In terms of general water chemistry ranges, the temperature of sampled waters ranges from near ambient (~2–10 °C) in cold springs to ~98 °C in the hottest springs (many boiling-point springs ~95 °C at elevation). Warmer springs clearly show higher dissolved solids: for example, chloride concentrations span <1 mg/L in cold dilute springs up to ~18,000 mg/L in the most saline hot spring (Mono Lake brine). A broad trend is that hotter waters are typically more saline, enriched in Na, Cl, SO_4, and silica relative to cooler waters. This aligns with geothermal paradigms: deep, high-enthalpy fluids leach salts and silica from rocks more effectively. Our data confirm that the thermal waters contain markedly higher concentrations of Na^+, Cl^-, SO_4^{2-}, and SiO_2 than the cooler springs. For instance, median SiO_2 in thermal waters (>50 °C) is ~90 mg/L, whereas many cooler springs have <20 mg/L SiO_2, which is consistent with silica geothermometry indicating higher reservoir temperatures in the former. The general water chemistry ranges are shown in Appendix 1.

The gas chemistry EDA, though limited by many nondetections, indicates two dominant gas types: N_2-dominated gas (often >90% N_2) with minor CO_2, thought to represent either entrained atmospheric gases or metamorphic N_2; and CO_2-rich gas in a subset of samples (e.g. some geothermal wells) where CO_2 can exceed 80–90% of the free gas (e.g. in magmatic areas). The median composition across all gas samples is ~93% N_2, ~0.6% CO_2, ~0.4% CH_4, and ~0.05% He, with O_2 generally low (<0.1%), indicating anoxic sources. Notably, only three samples had measurable H_2 gas: two at Mount Shasta (~1.17% H_2) and one at Mount St. Helens (8.6% H_2). The elevated H_2 (and He) at Mt. St. Helens reflects magmatic outgassing during the 1980 eruption (the sample was collected 11/1980). In contrast, no significant H_2 was detected in most non-volcanic springs, implying any hydrogen produced via water–rock interaction is either very low-level or consumed by subsurface reactions/microbes before reaching the surface. This emphasizes the use of indirect indicators (e.g., geochemistry) to infer subsurface H_2-generating potential, as elaborated below.

3. METHODOLOGY - UNSUPERVISED CLUSTERING OF GEOCHEMICAL SIGNATURES

The methodology section outlines the unsupervised clustering framework used in this study to identify geochemical regimes relevant to dual-resource exploration. We applied multiple unsupervised learning algorithms (K-means, DBSCAN, and Gaussian Mixture Models) to the standardized geochemical dataset to delineate natural clusters of geothermal fluids.

3.1 Cluster Identification and Geochemical Provinces

K-means (k=5) applied to the water chemistry data identified five distinct clusters of geothermal waters. The K-means clustering results and the respective parameter mean values are shown in Figure 3 and Table 1, respectively.

Table 1: The parameter Mean values for the K-means clusters.

cluster	pH	Na	K	Ca	Mg	Cl	SO4	HCO3	SiO2
0	6.35	1396.12	90.6	305.12	84.6	1762.7	118.7	2091.45	105.27
1	8.8	151.7	3.75	8.23	0.2	111.7	37.4	163.8	59.90
2	7.7	590.7	17.8	201.9	1.95	1078.45	221.25	46.5	85.7
3	6.91	29.4	3.8	25.0	14.78	3.6	28.6	201.78	79.67
4	7.30	5750.0	6.85	7025.0	10.2	21000.0	375.0	14.0	9.5

These clusters correspond to geochemically coherent groups that we interpret below. First, cluster 0 (Dilute Cold Waters - Meteoric Baseline) consists of low-temperature (often <15 °C) springs with very low mineralization. These have Cl < ~5 mg/L, HCO_3 ~20–50 mg/L, and total ionic content akin to shallow groundwater or stream water. pH is neutral ~6.5–7, and composition is typically Ca–Mg–HCO_3 dominated. We consider these non-thermal or marginally thermal springs, which are often included in the dataset as background samples adjacent to geothermal areas (e.g., cool springs or river water near hot springs). They serve as background hydrochemical references. In our clustering, this group anchors the low end of the thermal spectrum, ensuring that the algorithms can distinguish true geothermal waters from cold meteoric inputs.

Cluster 1 (Magmatic Chloride Waters) is characterized by high Cl (hundreds to ~1000 mg/L), high Na, significant Ca, and very low Mg. pH is neutral to slightly alkaline (~7.5–8), and temperatures are among the highest (60–95 °C). Examples include Springs at Long Valley (Casa Diablo, CA), and some Cascade volcanic springs fall into this cluster. These are classic volcanic or magmatic geothermal brines, often referred to as neutral pH chloride springs. The low Mg^{2+} (~0.1–0.3 mg/L) and high Ca^{2+} in these waters indicate extensive high-temperature water–rock reaction – Mg is scavenged by clay alteration at depth, while Ca can be leached from host rocks. Chloride serves as a conservative tracer of magmatic input or prolonged water circulation (Giggenbach, 1996). Similarly, all clusters are connected to their respective systems through the literature review, as shown below.

We interpret Cluster 2 as "deep-circulation meteoric" systems, often found in extensional terrains or volcanic peripheries. They serve as analogs to low-enthalpy geothermal fields where outflow mixing has occurred. For instance, springs in the Oregon Cascades foothills or Basin-and-Range margins fall here – they yield neutral Na–HCO_3–Cl waters of moderate temperature. These areas constitute a separate geothermal province distinct from the purely magmatic or purely sedimentary endmembers.

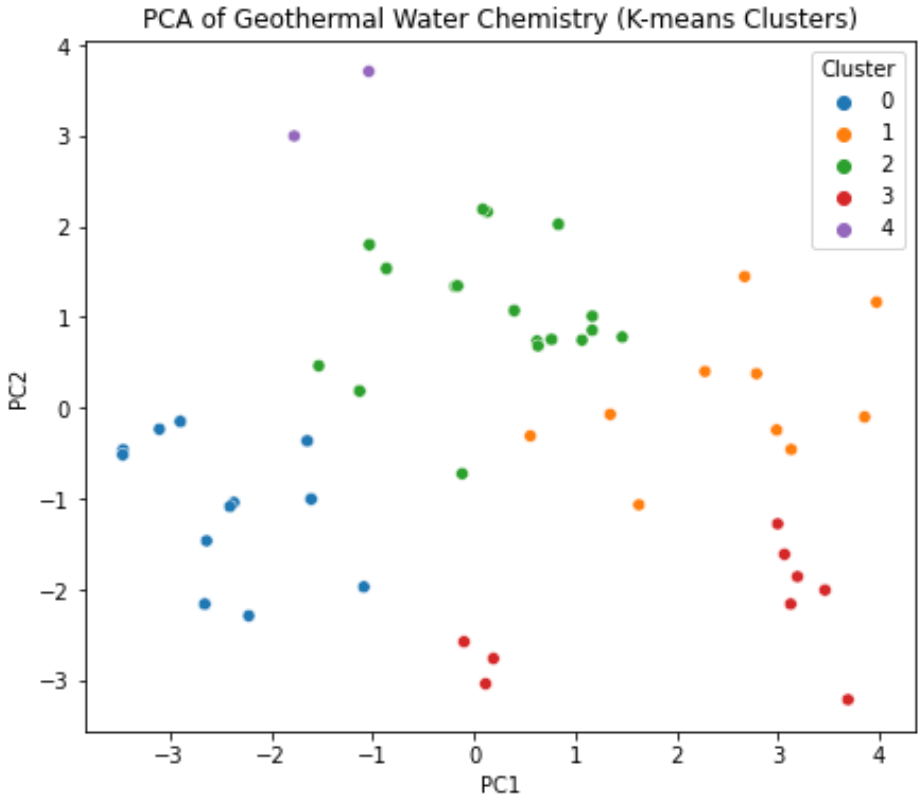

Figure 3: K-means clustering results.

The characteristics of cluster 3 (very high pH (9–10+), extremely low dissolved Mg (<0.1–0.5 mg/L), low total dissolved solids, and modest temperatures (~30–40 °C)} matches ultramafic rock (serpentinization) systems: when water reacts with peridotite, OH^- is generated, raising pH into the 9–12 range, while Mg^{2+} is stripped out to form brucite and serpentine. The resulting waters are Ca–OH type and very reducing, often containing hydrogen gas. Indeed, high-pH ultramafic waters are typically dilute and low in dissolved inorganic carbon, exactly what we observe. Our cluster analysis identified a handful of springs in this category, including Rock Creek Hot Springs (WA) (pH 9.7, Mg ~0 mg/L), Bonneville Spring (WA) (pH 9.5, Mg ~0 mg/L), and Lester/Scenic Hot Springs (WA) (pH ~9.1, Mg ~0 mg/L). These are all springs in or near the Cascade forearc region, where ultramafic mantle rocks are present in the subsurface. Interestingly, Casa Diablo (CA) also clustered close to this group because boiling in that system had stripped CO_2 and Mg, yielding a high-pH (9.2) fluid; however, Casa Diablo's geology is volcanic (rhyolitic), not ultramafic, so it's an exception in this cluster due to process (degassing) rather than rock type. Such conditions are known to actively produce H_2 in regions such as Oman and New Caledonia. Although H_2 was not directly measured in our water samples, the cluster's features align with those of known H_2-rich serpentinizing systems. This cluster is the key to our second objective: identifying hydrogen prospectivity.

Table 2: The Key Geochemical Signature for the clusters.

Cluster	Interpreted System Type	Key Geochemical Signature	References
Cluster 0	Dilute meteoric groundwater	Low TDS, low Cl–Na, Ca–Mg–HCO_3 dominated, low temperature	Freeze & Cherry, 1979; Drever, 1997
Cluster 1	Magmatic–volcanic geothermal fluids	High Cl–Na, elevated Ca, low Mg, neutral pH, high temperature	Giggenbach, 1996
Cluster 2	Deep-circulation meteoric geothermal systems	Moderate Cl–Na, high HCO_3, low Mg, moderate–high temperature	Fournier 1989
Cluster 3	Ultramafic influenced / serpentinization systems	High pH (>9), ultra-low Mg, low Cl, reducing conditions	Barnes & O'Neil 1969; Etiope & Lollar 2013
Cluster 4	Sedimentary brine / basinal fluids	Very high TDS, high Cl–Na, elevated HCO_3, slightly acidic to neutral pH	Hanor 1987; Kharaka & Hanor 2003

The cluster 4 chemistries are typical of sedimentary basin fluids or oilfield brines, where prolonged water–rock interaction and hydrocarbon-related CO_2 input yield chloride-rich, bicarbonate-rich waters. For example, our dataset includes Mono Lake, CA (a saline alkaline lake) which has ~18 g/L Cl and ~35 g/L HCO_3 serves as an extreme endmember.

3.2 Validation of clusters

The clusters above were cross-validated with DBSCAN and GMM. DBSCAN (density-based clustering) effectively removed outliers: it identified the acid-sulfate spring (pH 2, high SO_4) as a noise point and grouped the hyperalkaline springs (Cluster 3) into a small, tight cluster distinct from the main bulk. This ability to detect the small "ultramafic" cluster without forcing it in a k-means partition was valuable. Meanwhile, the Gaussian mixture model (GMM) provided fuzzy boundaries between clusters, reflecting mixing. For example, several cluster 1 vs. cluster 2 borderline samples showed 40–60% membership probability in each, consistent with mixed volcanic and meteoric fluid signatures. GMM also highlighted a continuum between Cluster 1 (magmatic Cl) and Cluster 4 (sedimentary brine) along the salinity axis, and between Cluster 1 and Cluster 2 along the alkalinity axis. These continua likely represent binary mixing trends (e.g., magmatic brine mixing with meteoric bicarbonate water), which we will quantify in a follow-up geochemical inversion.

Each clustering method provided a complementary view: K-means yielded an initial partition into a chosen number k of clusters, DBSCAN identified outlier or minority clusters (density-based approach, ideal for detecting unique small groups like hyperalkaline springs), and GMM (Gaussian mixtures) allowed probabilistic cluster memberships, highlighting transitional compositions (mixing) between endmembers.

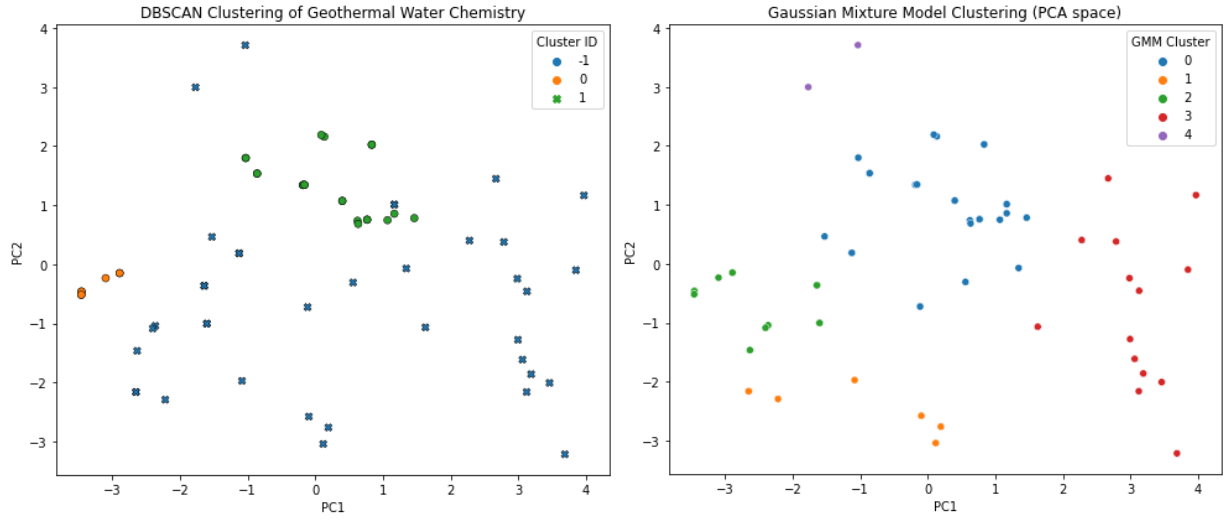

Figure 4: DBSCAN and GMM clustering results.

Further geographic mapping of cluster membership reveals a clear provincial organization across the western United States, with clusters aligning closely to regional geologic settings. Cluster 0 (mixed meteoric) samples commonly occur along the margins of volcanic centers and within extensional Basin-and-Range environments, such as Oregon. Cluster 1 (magmatic) samples are concentrated in young volcanic

provinces, including the Cascade volcanic arc in Oregon and Washington and the Long Valley Caldera in California. Cluster 3 (ultramafic) springs are preferentially located near ophiolitic belts and fore-arc regions west of the main Cascade volcanic front, particularly in Washington State, where high-pH springs occur in the Puget Sound Lowlands and Cascade foothills associated with mantle peridotite emplacement. In contrast, Cluster 4 (sedimentary brine) samples are relatively sparse and are primarily associated with closed basins and hydrocarbon-bearing regions, such as the Mono Basin and parts of Nevada.

3.3 Clustering Discussion and Conclusion

Overall, the unsupervised learning successfully delineated geothermal geochemical provinces. Each cluster serves as a geochemical "fingerprint" of a subsurface environment: volcanic-magmatic, meteoric-hydrothermal, sedimentary brine, ultramafic-serpentinizing, or non-thermal. These fingerprints can be used to map regions of similar subsurface conditions and infer the presence of heat sources or lithologies. In essence, legacy data, once clustered, provides analogues to known productive systems and flags unexplored zones with similar signatures. For example, a cluster 1 chemistry found in an area with no known geothermal development would immediately highlight the potential for a volcanic heat source there. Likewise, cluster 3 occurrences highlight ultramafic rocks at depth, which could be sources of hydrogen.

4. IDENTIFYING HYDROGEN GENERATION POTENTIAL

A central goal was to use the clusters identified above to identify sites with geochemical conditions conducive to geological H_2 generation (via water–rock reactions such as serpentinization). The cluster analysis pointed strongly to Cluster 3 (high-pH ultramafic waters) as the primary candidate. These springs have all the hallmarks of active H_2 production: extremely low redox potential (indicated by absence of sulfate and dissolved O_2, plus field observations of precipitated native metals or sulfides in some cases), abundant OH^- (high pH) to drive water dissociation, and likely presence of ultramafic minerals. Such conditions are known to generate hydrogen gas:

- Serpentinization of olivine produces H_2 and hydrogen-derived alkalinity, yielding Ca–OH waters with pH 10–12.
- Our cluster 3 springs, while not quite pH 11+, are ~pH 9–10 at surface; their reservoir pH could be higher, but CO_2 uptake or cooling may lower pH by the time they emerge.
- They are also likely H_2-saturated at depth. Although not directly measured here, analogous systems (e.g., Oman ophiolite, California Coast Range ophiolites) have free H_2 gas in the spring discharge. Past studies note that ultramafic spring waters with pH >11 often contain dissolved H_2 at 1–10 mM.

In the absence of direct H_2 measurements in our water dataset, we used proxy indicators to tag H_2 prospects:

- **High pH and ultra-low Mg:** We filtered for pH ≥9 and Mg <0.5 mg/L, as this combination is uniquely indicative of serpentinization-driven water chemistry. This highlighted a handful of springs (Rock Creek, Bonneville, Sulphur Springs WA, Scenic, Lester, Bagby OR, etc.). These precisely match Cluster 3 (and a couple from Cluster 1, like Casa Diablo, which have a different origin).
- **Reducing conditions:** Many of these springs also have high field pH and high measured H_2S or CH_4 in water analyses (when available). For example, Sulphur Hot Springs (WA) has high sulfide odor (H_2S was detected qualitatively). Reducing, oxygen-poor waters are needed to preserve H_2 (otherwise it oxidizes). The absence of SO_4^{2-} in these waters (SO_4 often <0.1 mg/L) confirms strongly reducing conditions. This environment is favorable for H_2 accumulation.

Prospective H_2 sites: Based on the above criteria, we compiled a short-list of springs most likely to be generating and preserving hydrogen. These include:

- **Rock Creek Hot Springs (WA)** – pH 9.7, Mg ~0; a candidate ultramafic spring likely sourcing H_2 at depth.
- **Bonneville (WA)** – pH 9.5, very low dissolved solids; located in an area of serpentinized ultramafics along the Columbia Gorge. This site has been flagged for highly reducing water chemistry, consistent with H_2 presence.
- **Sulphur Hot Springs (WA)** – pH 9.3, H_2S-rich, low Mg; despite the name, likely influenced by mafic rocks, possibly an ultramafic slice. Its reducing, alkaline nature suggests H_2 could accumulate.
- **Scenic Hot Springs and Lester Hot Springs (WA)** – pH ~9.1, very low Mg, in the Cascade Mountains. These are likely meteoric waters that interact with peridotite bodies at depth. Their helium isotope ratios (if measured) could be telling, but even from chemistry, they appear H_2-favorable.
- **Bagby Hot Springs (OR)** – pH 9.4, Mg ~0.5 mg/L. Although Bagby is in a volcanic area of Oregon, its unusually high pH hints at subsurface reactions that consumed CO_2 (possibly interaction with basal crustal rocks). It could be a mixed case but warrant consideration.

It is also notable that none of these H_2-prospect springs are associated with high geothermal temperatures; most are <50 °C at the surface. This is because surface discharge temperatures may significantly underestimate formation conditions, as ascending fluids and gases undergo conductive and advective heat loss during upward migration; therefore, the depth of the source ultramafic rocks and the associated thermal gradient must be considered when interpreting surface temperature observations and evaluating H_2 generation and preservation potential. Rock Creek and Bonneville Springs rank highest for hydrogen prospectivity in the Pacific Northwest. They exhibit near-identical profiles to those of proven H_2-producing sites (dilute, Ca-OH-type water, pH ≈ 10).

5. CONCLUSIONS AND NEXT STEPS

Through unsupervised learning on this legacy dataset, we successfully delineated multiple geothermal fluid types corresponding to geological provinces and identified a subset of sites with strong potential for subsurface hydrogen generation. The data-driven clusters

align well with known conceptual models (volcanic vs. sedimentary vs. meteoric systems) and spotlight previously under-recognized ultramafic-influenced springs. This illustrates the untapped potential of historical data: when mined with modern techniques, even decades-old records can yield new targets for sustainable energy exploration.

REFERENCES

Sekar, L.K. and Okoroafor, E.R., 2024, June. Maximizing Geothermal Energy Recovery from Enhanced Geothermal Systems Through Huff-And-Puff: A Comprehensive Simulation Study. In ARMA US Rock Mechanics/Geomechanics Symposium (p. D022S021R001). ARMA.

Ahmmed, B. and Vesselinov, V.V., 2022. Machine learning and shallow groundwater chemistry to identify geothermal prospects in the Great Basin, USA. Renewable Energy, 197, pp. 1034-1048.

Mariner, R.H., Presser, T.S. and Evans, W.C., 1977. Hot springs of the central Sierra Nevada, California: U.S. Geological Survey Open-File Report 77-559, 37 p.

Fournier, R.O., Thompson, J.M. and Austin, C.F., 1980. Interpretation of chemical analyses of water collected from two geothermal wells at Coso, California. Journal of Geophysical Research: Solid Earth, 85. B5, pp.2405-2410.

Mudunuru, M.K., Ahmmed, B., Rau, E., Vesselinov, V.V. and Karra, S., 2023. Machine learning for geothermal resource exploration in the Tularosa Basin, New Mexico. Energies, 16(7), p.3098.

Ahmmed, B., Vesselinov, V.V., Mudunuru, M.K., Middleton, R.S. and Karra, S., 2021. Machine Learning on the Geochemical Characteristics of Low-, Medium-, and Hot-temperature Geothermal Resources in the Great Basin, USA (No. LA-UR-21-20071). Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).

Moradpouri, F. and Sabeti, H., 2024. Geochemical evaluation of a geothermal region for the trace elements related to the subsurface mineralization using machine learning methods. Jordan J Earth Environ Sci, 15(2), pp. 94 -106.

Tut Haklidir, F.S. and Haklidir, M., 2020. Prediction of reservoir temperatures using hydrogeochemical data, Western Anatolia geothermal systems (Turkey): a machine learning approach. Natural Resources Research, 29(4), pp. 2333-2346.

Nicholson, K. (1993). Geothermal Fluids: Chemistry and Exploration Techniques. Springer.

Appelo, C. A. J., and Postma, D. (2005). Geochemistry, Groundwater and Pollution (2nd ed.). Balkema.

Bruni, J., Canepa, M., Chiodini, G., Cioni, R., and Longinelli, A. (2002). Irreversible water–rock mass transfer accompanying the generation of hydrothermal fluids in the Larderello geothermal field, Italy. Applied Geochemistry, 17, 455–474.

Giggenbach, W.F., 1996, January. Are Tokaanu chloride waters the outflow from Ketetahi or Hipaua. In Proceedings of the 18th New Zealand Geothermal Workshop. University of Aukland (Vol. 18, pp. 175-182).

Fournier, R.O., 1989. Geochemistry and dynamics of the Yellowstone National Park hydrothermal system. Annual Review of Earth and Planetary Sciences, Vol. 17, p. 13, 17, p.13.

Cherry, J.A. and Freeze, R.A., 1979. Groundwater (Vol. 370). Englewood Cliffs, NJ: Prentice-Hall.

Drever, J.I. and Stillings, L.L., 1997. The role of organic acids in mineral weathering. Colloids and Surfaces A: physicochemical and engineering aspects, 120(1-3), pp.167-181.

Barnes, I. and O'NEIL, J.R., 1969. The relationship between fluids in some fresh alpine-type ultramafics and possible modern serpentinization, western United States. Geological Society of America Bulletin, 80(10), pp. 1947-1960.

Etiope, G. and Sherwood Lollar, B., 2013. Abiotic methane on Earth. Reviews of Geophysics, 51(2), pp. 276-299.

Hanor, J.S., 1987. Kilometre-scale thermohaline overturn of pore waters in the Louisiana Gulf Coast. Nature, 327(6122), pp. 501-503.

Kharaka, Y.K. and Hanor, J.S., 2003. Deep fluids in the continents: I. Sedimentary basins. Treatise on geochemistry, 5, p. 605.

APPENDIX

Table A1: Description of the available data and the general water chemistry ranges

	Count	mean	std	min	25%	50%	75%	max
pH	66.0	7.410455	1.013256	5.64	6.460	7.37	8.10	9.7
Na	66.0	823.663636	1124.317475	4.10	151.250	405.00	920.00	6400.0
K	66.0	33.321212	54.548047	0.10	6.100	9.80	31.00	230.0
Ca	66.0	383.353030	1202.581289	1.20	19.250	76.00	320.00	7450.0
Mg	66.0	26.249848	56.848062	0.02	0.125	1.10	36.75	325.0
Cl	66.0	1575.719697	3624.505732	0.20	122.500	722.50	1350.00	22000.0
SO_4	66.0	143.828788	135.689180	0.40	30.000	137.50	190.00	510.0
HCO_3	66.0	642.181818	1074.750589	10.00	25.000	62.00	667.50	3935.0
SiO_2	66.0	84.030303	42.361074	9.00	48.000	81.00	102.25	199.0

Table A2: Available data and its relevance to our study

Attribute Group	Included Attributes (Water + Gas)	Relevance to This Study
Sample Metadata & Spatiotemporal Context	Sample name, ID, Type, Source, Collection date & time, Region, State, County, Reported location, Latitude, Longitude, Easting, Northing, Elevation, Location resolution, Location error, Author & digitizer comments	Provides spatial and temporal framework for integrating legacy datasets, correcting historical bias, and enabling regional geothermal province mapping.
Sampling, Hydraulic, and Physical Conditions	Well depth, Collection depth, Discharge, Temperature, Specific conductance, Salinity, Total dissolved solids	Constrains circulation depth, system enthalpy, and fluid concentration state distinguishing geothermal reservoirs from shallow groundwater.
Acid–Base System and Carbon Speciation	pH (field and lab), Carbonate (CO_3), Bicarbonate (HCO_3), Alkalinity (CO_3 and HCO_3), Carbonic acid (H_2CO3), Dissolved inorganic/organic carbon, δ13C, δ13C DIC	Captures CO_2 buffering and water–rock interaction processes critical for identifying high-pH ultramafic systems and hydrogen generation conditions.
Major Ions and Bulk Water Chemistry	Na, K, Ca, Mg, Cl, SO_4, SiO_2, F, Br, NO_3, PO_4, NH_4, Reported cations and anions	Forms the primary geochemical fingerprint used in unsupervised clustering to delineate geothermal fluid types.
Trace Elements and Redox Indicators	Al, As, B, Ba, Cs, Cu, Fe, Hg, I, Li, Mn, Mo, Ni, Pb, Rb, Sr, U, V, Zn, Hydrogen sulfide (H_2S)	Acts as sensitive tracers of lithology, redox state, temperature, and ultramafic influence relevant to hydrogen-producing environments.
Gas Composition and Dissolved Gases	Total gas, N_2, CO_2, CH_4, H_2, O_2, Ar, H_2S, NH_3, C_2H_6, dissolved gas species	Defines gas endmembers and gas–water partitioning behavior used to identify magmatic, biogenic, and serpentinization-related hydrogen sources.
Isotopic Tracers and Fluid Origin Indicators	δ2H, δ18O H2O, δ18O SO4, δ15N, δ13C CO2, Tritium (3H), 14C, He, 3He/4He, corrected 3He/4He	Constrain fluid origin, residence time, mixing, and deep mantle or ultramafic contributions to geothermal and hydrogen systems.

Table A3: A brief Summary and comparison of the used algorithms in this study.

Algorithm	Brief Description	Strengths	Weaknesses
K-means	Partitions data into *k* clusters by minimizing within-cluster variance.	Simple, fast, scalable to large datasets, easy to interpret cluster centroids.	Requires predefined *k*; assumes spherical clusters; sensitive to outliers and scaling.
DBSCAN	Density-based clustering that groups points based on neighborhood density.	Identifies arbitrarily shaped clusters; detects outliers; no need to predefine number of clusters.	Sensitive to parameter selection (ε, min_samples); struggles with varying density and high dimensions.
Gaussian Mixture Models (GMM)	Probabilistic model assuming data is generated from a mixture of Gaussian distributions.	Captures overlapping clusters; provides soft cluster membership; flexible cluster shapes.	Assumes Gaussian distributions; sensitive to initialization; computationally expensive.

Table A4: A brief Summary and comparison of the used algorithms in this study (Cont.).

Algorithm	Computation	No. of Hyperparameters	Type of Applications
K-means	Very fast; scales efficiently with dataset size.	Few (number of clusters *k*, initialization).	Large-scale exploratory analysis and initial province delineation.
DBSCAN	Moderate computational cost; depends on neighborhood queries.	Few (ε, min_samples).	Outlier detection, identification of rare or extreme geochemical systems.
Gaussian Mixture Models	Computationally intensive due to likelihood optimization.	Moderate (number of components, covariance type).	Modeling mixed fluids, transitional geochemical signatures, and uncertainty quantification.