Discovering Hidden Geothermal Signatures using Non-negative Matrix Factorization with Customized *k*-means Clustering

V. V. Vesselinov[1], B. Ahmmed[2], *, M. K. Mudunuru[3], J. D. Pepin[4], E. R. Burns[5], D. L. Siler[6], S. Karra[7], and R. S. Middleton[8]

[1]EnviTrace, LLC., Santa Fe, NM 87501

[2]Computational Earth Science Group, Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545

[3]Watershed & Ecosystem Science, Pacific Northwest National Laboratory, Richland, WA 99352

[4]U.S. Geological Survey, Albuquerque, NM 87113

[5]U.S. Geological Survey, Portland, OR 97201

[6]U.S. Geological Survey, Moffett Field, CA 94035

[7]Environmental Molecular Science Laboratory, Pacific Northwest National Laboratory, Richland, WA 99352

[8]Carbon Solutions LLC, Okemos, MI 48864

[*]**Corresponding author:** Bulbul Ahmmed, Staff Scientist, Computational Earth Science Group (EES-16), Earth and Environmental Sciences Division (EES), Los Alamos National Laboratory (LANL), Los Alamos, NM - 87545. Email: ahmmedb@lanl.gov

Abstract: Discovery of hidden geothermal resources is challenging. It requires the mining of large datasets with diverse data attributes representing subsurface hydrogeological and geothermal conditions. The commonly used play fairway analysis approach typically incorporates subject-matter expertise to analyze regional data to estimate geothermal characteristics and favorability. We demonstrate an alternative approach based on machine learning (ML) to process a geothermal dataset from southwest New Mexico (SWNM). The study region includes low- and medium-temperature hydrothermal systems. Several of these systems are not well characterized because of insufficient existing data and limited past explorative work. This study discovers hidden patterns and relations in the SWNM geothermal dataset to improve our understanding of the regional hydrothermal conditions and energy-production favorability. This understanding is obtained by applying an unsupervised ML algorithm based on non-negative matrix factorization coupled with customized *k*-means clustering (NMF*k*). NMF*k* can automatically identify (1) hidden signatures characterizing analyzed datasets, (2) the optimal number of these signatures, (3) the dominant data attributes associated with each signature, and (4) the spatial distribution of the extracted signatures.

Here, NMF*k* is applied to analyze 18 geological, geophysical, hydrogeological, and geothermal attributes at 44 locations in SWNM. Using NMF*k*, we find data patterns and identify the spatial associations of hydrothermal signatures within two physiographic provinces (Colorado Plateau and Basin and Range) and two sub-regions of these provinces (the Mogollon-Datil volcanic field and the Rio Grande rift) in SWNM. The ML algorithm extracted five hydrothermal signatures in the SWNM datasets that differentiate between low (<90°C) and medium (90-150°C)-temperature hydrothermal systems. The algorithm also suggests that the Rio Grande rift and northern Mogollon-Datil volcanic field are the most favorable regions for future geothermal resource discovery. NMF*k* also identified critical attributes to identify medium-temperature hydrothermal systems in the study area. The resulting NMF*k* model can be applied to predict geothermal conditions and their uncertainties at new SWNM locations based on limited data from unexplored regions. The code to execute the performed analyses as well as the corresponding data can be found at https://github.com/SmartTensors/GeoThermalCloud.jl.

Keywords: Geothermal energy, unsupervised machine learning, non-negative matrix factorization, custom *k*-means clustering, feature extraction, hidden signatures, hidden geothermal resources.

1. Introduction

Identifying hidden geothermal resources is challenging because they typically lie hundreds of meters to several kilometers below the ground surface without hydrothermal exposure at the ground surface. They may have deep water tables or be sealed by overlying impermeable rock that precludes the flow of hot water and heat towards the ground surface (Anderson, 2013; Brott et al., 1981; Dobson, 2016; Porro et al., 2012; Smith, 2004; Williams et al., 2009). A commonly used

approach for the geothermal exploration of hidden resources is based on play fairway analysis (PFA) (Faulds et al., 2015, 2018, 2019; Siler et al., 2017, 2019; Lautze et al., 2017, 2020; McClain et al., 2015; Shervais et al., 2015a-b, 2017).

PFA evaluates geothermal favorability by assimilating various geological, geophysical, geochemical, and geothermal attributes locally and at regional scales. The latter attributes provide direct geothermal evidence based on measurements, such as temperatures at various depths, heat flow, and thermal gradients. However, direct measurements of geothermal attributes are often challenging and expensive to acquire. PFA typically incorporates subject-matter expertise to process and analyze the available data and make conclusions about geothermal favorability. Ultimately, this can produce bias in the interpretations and limits the amount of data that can be efficiently mined. The general challenges of PFA applications relate to (1) defining relations between analyzed geothermal data attributes; (2) identifying critical easy-to-measure attributes that can be applied to estimate geothermal reservoir properties and favorability at new locations; and (3) removing bias introduced by experts in analysis.

To address these challenges, we conduct an alternative to the PFA approach that is based on machine learning (ML) to process existing regional data and to find the hidden data relations without interpretive biases and, more importantly, without requiring direct measurements of geothermal attributes (e.g., thermal gradient, heat flow). To achieve this, we detail an ML methodology that can efficiently analyze all available local or regional data to learn hidden relations between the attributes of known geothermal reservoir properties at sites that are more economic and accessible (e.g., drainage density, shallow groundwater geochemistry) and sites with

unknown properties. Also, the method can (1) provide a better understanding and robust prediction of geothermal favorability, (2) discover hidden geothermal resources without the direct human intervention, (3) identify the optimal number of hidden signatures characterizing the data, (4) isolate dominant sets of attributes in data that correspond to identified hidden signatures, and (5) pinpoint locations associated with each hidden signature. This alternative PFA approach is demonstrated here using a geothermal dataset of southwest New Mexico (SWNM). Through ML, we discover hidden geothermal signatures, their dominant attributes, and the spatial association of each hidden signature. Moreover, we delineate the spatial distribution of low (<90°C) and medium (90-150°C) temperature hydrothermal systems. Finally, we make a comparison of outputs between a few similar studies in the same study area (Bielicki et al., 2015; 2016; Pepin, 2019).

2. Background

2.1 Machine Learning

ML methods, in general, can be subdivided into supervised and unsupervised methods. Supervised methods require attributes and corresponding labels of the analyzed data (Johnson et al., 2021; Muller and Guido, 2016; Rouet-Leduc et al., 2020). The labeling needs to be done by subject-matter experts who can identify, for example, locations with high (>150°C), medium, and low temperature geothermal favorability or specific geologic features such as fault offsets. The labeling process can also be automated by unsupervised ML (Muller et al., 2016). The supervised methods can then be applied to learn geothermal favorability based on the available data. However, the successful training of supervised methods requires large, continuous (without data gaps), non-noisy (with small measurement errors) training datasets that are typically not available for geothermal exploration. Commonly used supervised methods include deep neural networks

(Yoshinki et al., 2014), convolutional neural networks (Gu et al., 2018), recurrent neural networks (Medsker & Jain, 1999), and random forest (Breiman, 2001).

In contrast, unsupervised ML techniques extract information from existing datasets without any prior labeling or subject-matter preprocessing. These methods find patterns with common features and the underlying signatures that lead to these patterns. The extracted information is then post-processed by subject-matter experts to identify the physical meaning of the patterns, features, and the underlying signatures. Commonly used unsupervised methods include singular value decomposition (SVD) (Klema & Laub, 1980), principal component analysis (PCA) (Wold et al., 1987), independent component analysis (ICA) (Comon, 1994), *k*-means clustering (Hartigan & Wong, 1979), Gaussian mixture models (Friedman & Tibshirani, 2001), non-negative matrix/tensor factorization (NMF/NTF) (Lee & Sung, 1999), and non-negative matrix/tensor factorization with customized *k*-means clustering (NMF*k*/NTF*k*; https://smarttensors.github.io; Alexandrov & Vesselinov, 2014; Vesselinov et al., 2018).

One or more unsupervised ML methods can be applied for finding hidden patterns in a geologic/geothermal dataset. For example, Watson et al. (2020) utilized *k*-means clustering on infrasound signals to characterize volcanic eruption activity. Anzieta et al. (2019) used *k*-means clustering, correntropy, and dynamic time warping to understand the precursor of the 2015 Cotopaxi volcano eruption. Alexandrov & Vesselinov (2014) and Vesselinov et al. (2018, 2019) applied NMF*k* for blind source separation and extraction of physics insights about complex geologic systems. Unsupervised ML also has been used to characterize hydrothermal systems. For example, Pepin (2019) applied PCA on a similar SWNM dataset to identify geothermal favorability. Siler et al. (2021) and Siler & Pepin (2021) used NMF*k* and PCA to identify geologic factors that control flow in the Brady, Nevada, geothermal site and found similar results by both

methods. Ahmmed et al. (2021 a-b; 2020 a-b) identified hidden geothermal signatures at the Utah FORGE site, the Great Basin, and Hawaii Islands, and Vesselinov et al. (2020, 2021) successfully identified hidden geothermal signatures in eight datasets of U.S. geothermal reservoirs. As recent work suggests, the application of NMF*k* to diverse multi-source, multi-scale, and multi-physics geothermal datasets may lead to discovering unknown geothermal signatures. These discovered signatures can be applied to improve the detection of hidden geothermal resources and identify the potential for geothermal play development.

Here, we applied NMF*k* to analyze an existing SWNM geothermal dataset. To discover hidden signatures, along with their optimal number in large geothermal datasets, NMF*k* is at the forefront among various unsupervised ML methods such as NMF, PCA, ICA, SVD and its variants, *k*-means clustering, and Gaussian mixture models. In contrast with traditional NMF (Lee & Seung, 1999), NMF*k* allows for automatic identification of the optimal number of signatures (features) present in the data (Vesselinov et al., 2018). Because the data attributes analyzed here are transformed to be non-negative, NMF*k* preserves non-negativity when extracting hidden signatures. The non-negativity constraint makes the decomposed matrices easier to interpret than PCA, SVD, and ICA because the extracted signatures are additive (Lee & Seung, 1999). Moreover, NMF*k* can handle real, categorical, and missing data (challenging or impossible with other supervised and unsupervised ML methods) (Vesselinov et al. 2018, 2019). Even more importantly, the missing data (some or all) can be reconstructed from available data based on the estimated matrix factorization. Note, NMF*k* is part of SmartTensors, which are a part of the GeoThermalCloud framework for geothermal exploration and can be found at https://github.com/SmartTensors/GeoThermalCloud.jl.

2.2 NMF*k*

NMF*k* performs matrix factorization of a data matrix, $X_{m\times n}$, where the m rows are the values of the geothermal attributes, and the n columns represent measurement locations. The goal of NMF*k* is to find the optimal number of signatures k that describe the analyzed dataset. The signatures are extracted by matrix factorization (Lee & Seung, 1999), which can be represented as:

$$X_{m\times n} \cong W_{m\times k} \times H_{k\times n} \quad (1)$$

where $W_{m\times k}$ is an "attribute" matrix characterizing the significance of the attributes and $H_{k\times n}$ a "location" matrix captures the importance of the locations. Note that all the elements of matrices W and H are unknown. The number of signatures, k, is also unknown. The matrix factorization in Equation1 provides an approximate representation of the data X. To solve for all the unknowns, NMF*k* performs a series of matrix factorizations with random initial guesses for W and H elements and for a range of values of k; theoretically, k can range between 1 and min(m,n). For a given number of signatures k, Equation 1 is solved iteratively to minimize the reconstruction error $O(k)$:

$$O(k) = ||X - W \times H||_F \quad (2)$$

by constraining the W and H elements to be greater than or equal to zero and F defines the Frobenius matrix norm (Böttcher & Wenzel, 2008). Under the NMF*k* algorithm, NMF is executed numerous times (typically 1,000), which generates a series of solutions for W and H matrices for a given k value. The resulting multiple solutions of H are clustered into k clusters using a customized k-means clustering. The average Silhouette width $S(k)$ (Rousseeuw, 1987) is computed for all k clusters based on the cosine norm. This metric (Vesselinov et al. 2018) measures how well the random NMF solutions are clustered for a given value of k. The values of $S(k)$ theoretically can vary from -1 to 1. These operations are repeated for a series of k values. The

optimal number of signatures, k, is estimated on how the reconstruction error, $O(k)$, and the average silhouette width, $S(k)$, vary with the increase of k. The reconstruction error decreases as the number of signatures increases. The average silhouette width behavior is more complicated; $S(k)$ generally declines from 1 to -1 as the number of signatures increases. However, $S(k)$ values frequently spike up for specific k values, indicating that these k values are potentially optimal. In an ideal case, a given k value is considered optimal when adding another signature does not significantly improve the estimate of X (i.e., lower $O(k)$) and does not lower $S(k)$. In practice, a solution with $S(k)$ greater than 0.5 and the lowest $O(k)$ value can be chosen as an optimal solution. The solutions with k values less than the optimal value and S(k) values > 0.5 are acceptable; they provide underfitting representations of the data matrix X. All the solutions with k values greater than the optimal value are not acceptable; they provide overfitting representations of the data matrix X. Implementation of the NMFk algorithm and details related to the selection of the optimal solution are further discussed in Alexandrov and Vesselinov (2014) and Vesselinov et al. (2018, 2019). The NMFk results are summarized using different analytical and visual methods detailed in the discussion section below.

2.3 Test dataset

This study analyzes a geothermal dataset from SWNM. SWNM is broadly divided into two physiographic provinces: the Colorado Plateau and the Basin and Range. The Mogollon-Datil volcanic field (volcanic field) is a sub-region of the Colorado Plateau and the Rio Grande rift is a sub-region of the Basin and Range (Bielicki et al. 2015, 2016; Pepin, 2019; Person et al. 2015). Each of the regions are associated with different types of hydrothermal systems with temperatures ranging from low to medium (Bielicki et al., 2015; Pepin, 2019; Vesselinov et al., 2020, 2021). Some of the SWNM systems are already utilized for commercial and recreational purposes. At 23

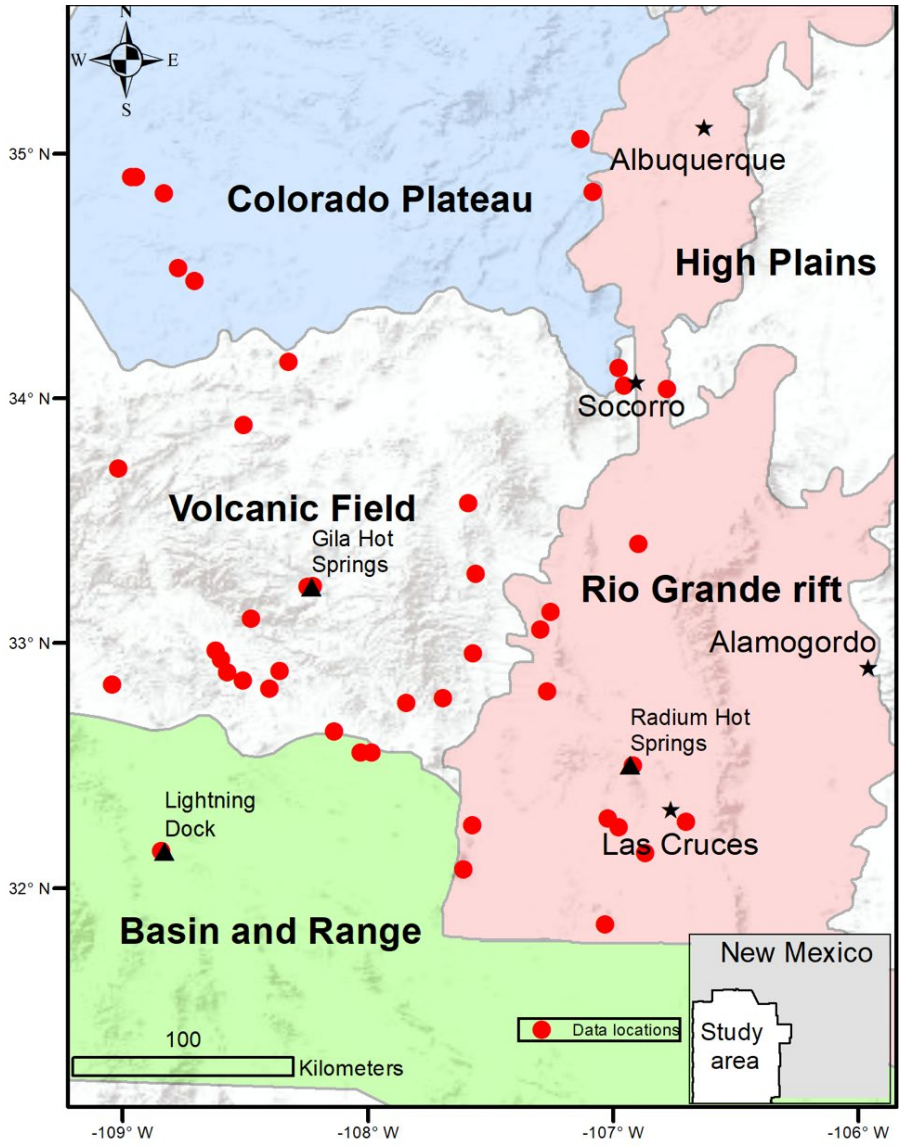
191 locations, energy-extraction facilities provide both electricity and direct-use heating (Kelley,
192 2010). A geothermal power plant is located in Lightning Dock within the Basin and Range
193 province (Figure 1), and it produces a gross of ~14 MWe power (Bonafin et al., 2019). One of the
194 largest greenhouses in the country, Masson Farms, is in Radium Hot Springs within the Rio Grande
195 rift (figure 1). SWNM has another medium-temperature geothermal system in the volcanic field,
196 Gila Hot Springs. There are 14 spas and recreational facilities utilizing the SWNM geothermal
197 resources (Kelley, 2010). A recent PFA Phase I study of SWNM revealed more potential
198 geothermal resources (Bennett & Nash, 2017; Bielicki et al., 2015; Levitte & Gambill, 1980).
199
200 The analyzed dataset includes two geochemical, three geophysical, seven geological, four
201 hydrogeological, and two geothermal attributes (total 18) at 44 locations in SWNM (Figure 1).
202 Each attribute may be related to geothermal processes, as discussed in detail in Table 1 (Bielicki
203 et al., 2015; Keller et al., 1991; Pepin, 2019; Person et al., 2013; UNM, 2018; USGS, 2018a,b).
204 The data are preprocessed prior to the ML analyses. During the preprocessing stage, the *boron* and
205 *lithium* concentration values are log-transformed to narrow down the distribution of values. Next,
206 all attributes are rescaled within the range of 0.0 to 1.0 using unit range transformation. The values
207 of each attribute at each location are shown in Table 2. To apply NMF*k*, we create an 18×44
208 matrix ($X_{m \times n}$) where the m = 18 rows are the values of the measured attributes, and the n = 44
209 columns represent each location.

210 ## 3. Results and Discussion

211 ### *3.1 Identification of the Optimal Number of Signatures*

212 For each k, NMF*k* analysis provides W and H matrices and several metrics regarding solution
213 accuracy including reconstruction errors and solution robustness. For this project, we run NMF*k*

for k = 2 to 15. Figure 2 summarizes the NMFk results related to reconstruction error and solution robustness for specified k values. The reconstruction error, $O(k)$, decreases as the number of signatures increases. However, the average silhouette width, $S(k)$ fluctuates over the number of signatures, as shown in Figure 2. Solutions with $S(k)$ less than zero are rejected (not shown in Figure 2). As discussed above, solutions with $S(k)$ greater than 0.5 can be acceptable. The solution with $S(k) > 0.5$ and the lowest $O(k)$ values is identified to be optimal. Based on these criteria, the solutions for k = 2, 3, 4, and 5 were accepted. The solutions for k > 5 are overfitting the analyzed dataset. The k = 5 solution is automatically identified by the NMFk algorithm as optimal. The solution with the optimal number of signatures is expected to provide the best physical interpretability of the analyzed data matrix. In the following section, we focus on the spatial association of the extracted signatures within the study area.



225

226 **Figure 1: The study area is in southwest New Mexico. The red dots show 44 geothermal data**

227 **locations in this study (after Pepin, 2019). Filled black color stars represent cities, while filled**

228 **black color triangles represent medium-temperature hydrothermal systems. Base map**

229 **source: ESRI, USGS, and NOAA.**

230

231 **Table 1. List of geothermal data attributes and their significance for geothermal resource**

232 **exploration, and units.**

Attribute number	Attribute (abbreviation in table 2)	Measurement type	Significance for geothermal resource exploration	Unit
1	*B^+ concentration (Boron)*	Geochemical	Potentially represents enhanced dissolution from high-temperature waters	mg/L
2	*Li^+ concentration (Lithium)*	Geochemical	Potentially represents deep heat source	mg/L
3	*Drainage density (Drain)*	Hydrogeological	Represents the structure of surface-water flow and permeability; may also represent groundwater recharge areas and the existence of geologic structures influencing the shape of the drainage network	count/area
4	*Springs density (Springs)*	Hydrogeological	Represents occurrence of conduits of groundwater from depth to the ground surface	count/area
5	*Hydraulic gradient (Hydr. Grad)*	Hydrogeological	Slope of the water table along the direction of decreasing head; may indicate the magnitude of groundwater flow and permeability	[-]
6	*Precipitation (Precip)*	Hydrogeological	The primary source of groundwater recharge	mm
7	*Gravity anomaly (Gravity)*	Geophysical	May represent secondary mineralization; also characterizes the geologic structure (Beihler, 1971)	mGal
8	*Magnetic intensity (Magnet)*	Geophysical	May represent secondary mineralization, and in some events, may characterize the geologic structure	nT
9	*Seismicity (Seism)*	Geophysical	Seismicity is related to recent fault and/or tectonic activities which may create and/or maintain secondary permeability fractures that support geothermal	count/area

			circulation. Seismicity might provide additional insight into the controls on geothermal resources relative to the fault density maps because it allows us to better understand the role that active seismicity (and perhaps permeability maintenance) plays in the location of geothermal resources.	
10	*Silica geothermometer (Silica)*	Geothermal	Indicator of the potential temperature of subsurface reservoirs	°C
11	*Heat flow (Qheat)*	Geothermal	Represents advective heat transport	mW/m^2
12	*Crustal thickness (Crust)*	Geological	Represents proximity of the deep heat source (Earth's mantle)	km
13	*Depth to the basement (Bsmt)*	Geological	Represents the thickness of the potential geothermal reservoir and the depth of groundwater circulation	m
14	*Fault intersection density (Fault)*	Geological	Represents connection of fault networks; the higher the density, the better for extracting hot water (Faulds et al. 2018)	count/area
15	*Quaternary fault density (Qfault)*	Geological	Act as conduits of (1) groundwater flow from depth to the ground surface and (2) groundwater recharge; may also indicate recent activity (Quaternary activity)	count/area
16	*State map fault density (NMFlt)*	Geological	Significance is the same as Qfault but does not limit the timespan for activity	count/area
17	*Volcanic dike density (Vents)*	Geological	Represents subsurface manifestation of volcanic events and can provide vertical permeability where fractured	count/area
18	*Volcanic vent density (Dikes)*	Geological	Indicate the occurrence of volcanic eruptions and may provide vertical permeability	count/area

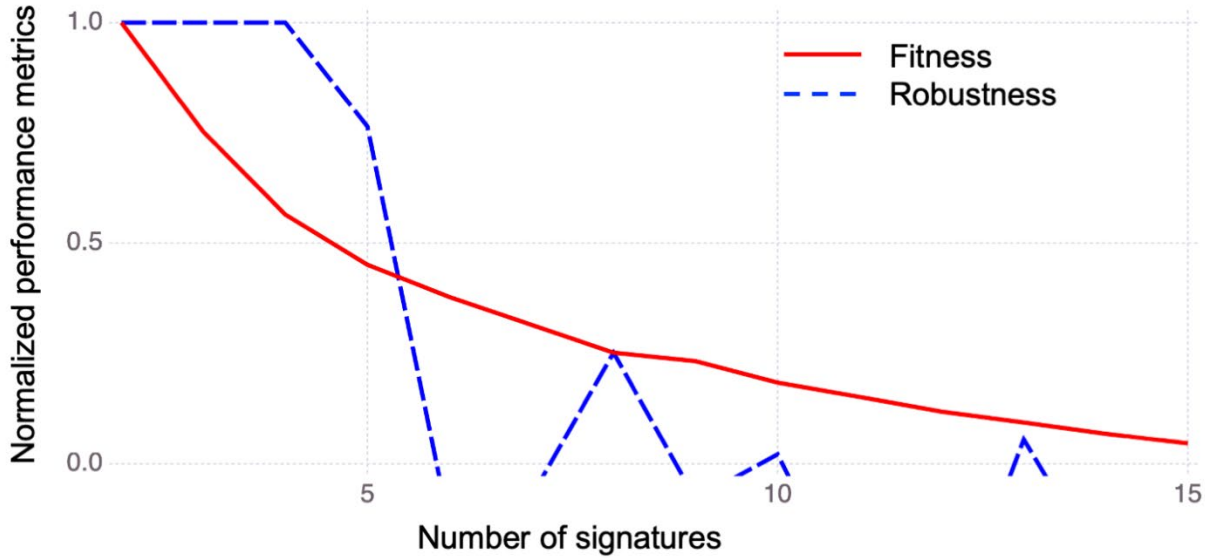
233

Table 2. Geothermal data applied for NMF*k* analysis. The table shows the data matrix (X) representing observations of 18 attributes (columns) over 44 locations in the study area. Each column is color-coded, where warm and cool colors represent high and low values, respectively. See Table 1 for attribute explanation.

Location	Boron	Lithium	Drain	Springs	Hydr. Grad	Precip	Gravity	Magnet	Seism	Silica	Qheat	Crust	Bsmt	Fault	Qfault	NMFlt	Vents	Dikes
Alamos Spring	-0.21	-3.1	7.4	0.010	5.6	264.8	-203.3	136.2	0.004	16.5	4.6	38.7	1439	0.000	0.00	16.2	0.003	0.431
Allen Springs	-3.2	-4.0	17.3	0.003	13.9	514.5	-189.3	184.6	0.002	24.0	4.4	32.5	51	0.000	0.01	15.6	0.001	3.625
Apache Tejo Warm Springs we	-1.8	-8.6	17.3	0.003	4.7	326.3	-181.2	15.0	0.001	52.0	4.6	30.7	24	0.001	0.03	0.7	0.000	3.807
Aragon Springs	1.5	-7.5	19.0	0.005	4.0	387.0	-229.1	-317.7	0.000	56.5	4.5	38.8	1486	0.000	0.00	41.1	0.003	0.010
Ash Spring	-2.7	-5.0	17.0	0.003	4.1	492.0	-193.2	66.6	0.002	29.3	4.4	32.2	-92	0.000	0.00	9.3	0.000	4.914
B. Iorio 1 well	-2.1	-2.6	18.8	0.003	0.9	260.4	-196.5	-48.2	0.000	59.4	4.0	30.9	-188	0.057	21.02	9.1	0.003	1.936
Cliff Warm Spring	-2.5	-6.9	22.8	0.002	1.8	364.2	-199.1	-47.1	0.002	64.2	4.2	33.1	-191	0.001	2.58	11.0	0.001	1.290
Dent windmill well	-2.1	-7.3	13.4	0.005	2.4	341.7	-230.8	89.3	0.000	19.7	4.7	43.5	865	0.000	0.00	0.0	0.000	0.000
Derry Warm Springs	-1.5	-7.5	18.3	0.002	3.0	276.1	-161.6	197.0	0.000	37.4	4.6	30.0	-120	0.007	9.16	15.9	0.000	0.659
Faywood Hot Springs	-2.6	-4.8	16.6	0.003	4.2	346.4	-172.1	-49.8	0.000	67.2	5.5	30.0	619	0.002	2.81	1.9	0.000	0.939
Federal H 1 well	-0.4	-5.0	5.8	0.000	2.7	253.8	-132.0	35.0	0.001	78.7	4.9	27.3	2906	0.004	20.31	7.2	0.015	0.000
Freiborn Canyon Spring	-2.5	-12.6	13.1	0.001	13.0	538.6	-225.0	-242.0	0.001	49.8	4.6	38.4	1138	0.000	0.00	19.8	0.004	0.401
Garton well	-3.2	-5.0	18.0	0.002	4.3	489.9	-196.8	35.6	0.000	70.0	3.9	30.9	-266	0.000	0.00	28.9	0.001	0.150
Gila Hot Springs 1	-1.9	-7.8	24.2	0.003	6.6	422.6	-221.6	-149.3	0.001	69.9	4.4	34.0	413	0.000	0.00	25.5	0.003	0.127
Gila Hot Springs 2	-1.8	-6.7	24.7	0.003	3.2	425.9	-222.9	-138.8	0.001	70.8	4.6	33.9	519	0.000	0.00	23.7	0.003	0.112
Goat Camp Spring	-2.1	-8.0	10.0	0.002	5.8	344.0	-159.2	-29.7	0.007	68.9	4.4	32.4	19	0.001	2.22	10.6	0.001	0.751
Jerry well	-0.8	-7.9	15.5	0.004	1.0	243.9	-219.6	172.4	0.000	13.4	4.4	42.3	1190	0.000	0.00	6.3	0.005	0.111
Kennecott Warm Springs well	-2.4	-6.9	17.8	0.003	4.3	355.0	-178.3	-69.9	0.000	66.1	5.0	30.0	409	0.002	1.76	1.1	0.000	1.422
Laguna Pueblo	0.4	-3.3	8.6	0.018	2.6	259.7	-204.2	62.5	0.006	42.9	4.4	37.2	1506	0.004	4.58	14.6	0.005	0.406
Lightning Dock	-1.0	-3.9	4.6	0.000	0.8	291.5	-168.0	-168.1	0.002	107.3	5.0	29.8	1800	0.008	8.40	4.3	0.000	0.086
Los Alturas Estates	-1.5	-12.7	7.6	0.001	2.2	265.3	-141.4	-127.5	0.002	71.9	6.3	27.4	4321	0.003	0.05	6.6	0.000	0.004
Mangas Springs	-2.6	-4.5	20.2	0.002	0.3	393.5	-201.0	-227.1	0.002	53.6	4.2	32.4	-178	0.000	0.91	11.5	0.000	3.503
Mimbres Hot Springs	-2.3	-3.8	15.4	0.004	9.1	445.9	-200.6	43.4	0.000	68.3	4.9	31.0	50	0.002	1.13	19.0	0.000	0.670
Ojitos Springs	-1.6	-4.5	19.6	0.020	7.2	257.5	-202.1	-7.5	0.037	57.6	4.5	33.0	-255	0.044	19.74	31.0	0.005	1.342
Ojo Caliente	-2.6	-2.9	20.5	0.004	3.5	333.6	-226.5	-168.4	0.000	48.4	5.5	33.8	2415	0.000	0.00	8.3	0.000	0.000
Ojo De las Canas	-1.7	-6.0	22.3	0.013	4.0	270.5	-188.5	-85.8	0.036	14.2	4.5	31.8	101	0.036	12.55	28.0	0.003	0.839
Pueblo windmill well	-1.2	-12.0	15.2	0.004	2.9	265.8	-228.8	315.9	0.000	18.3	4.3	42.5	1027	0.000	0.00	6.1	0.003	0.029
Radium Hot Springs	-0.8	-5.3	8.8	0.001	0.3	264.2	-151.4	-7.8	0.003	63.6	5.4	28.2	1191	0.013	11.40	10.6	0.000	0.010
Rainbow Spring	-1.7	-7.0	11.0	0.006	3.3	307.8	-227.1	-48.5	0.001	21.7	4.7	43.9	755	0.000	0.00	0.0	0.000	0.000
Riverside Store well	-1.3	-2.4	22.6	0.002	0.9	356.1	-196.1	-102.9	0.002	60.8	4.3	32.9	-165	0.000	2.50	11.7	0.001	1.562
Sacred Spring	-1.8	-7.0	10.9	0.006	1.3	298.4	-228.4	-80.4	0.001	21.2	4.6	43.9	742	0.000	0.00	0.0	0.000	0.000
Socorro Canyon	-1.8	-6.7	21.1	0.020	11.1	284.1	-204.7	-136.5	0.034	44.6	5.0	32.6	-229	0.051	28.88	33.8	0.005	1.203
Spring	-4.1	-6.8	20.1	0.001	5.1	361.9	-183.5	334.5	0.000	117.2	3.8	31.5	-104	0.011	1.81	20.1	0.006	0.218
Spring Canyon Warm Spring	-2.1	-8.3	21.9	0.002	5.8	361.7	-194.2	117.3	0.002	51.6	4.2	32.6	-57	0.000	1.50	12.7	0.000	2.293
Truth or Consequences spring	-1.1	-3.3	18.4	0.003	0.6	265.9	-168.2	-54.3	0.000	55.3	4.3	31.0	304	0.064	20.51	10.3	0.002	2.175
Turkey Creek Spring	-3.2	-3.7	19.2	0.002	5.8	493.4	-196.4	54.8	0.002	81.3	4.4	33.6	56	0.001	3.69	28.1	0.002	0.984
Victoria Land and Cattle Co. w	-1.8	-2.9	6.4	0.001	1.9	253.0	-165.9	-65.4	0.001	43.0	4.1	30.7	2014	0.003	0.06	0.9	0.000	0.478
Warm Springs	-2.1	-2.5	19.0	0.004	5.4	314.6	-193.3	113.5	0.000	56.0	4.3	32.7	1252	0.029	2.63	16.5	0.003	0.220
Well 1	-1.4	-6.6	15.7	0.004	1.7	345.4	-230.7	-31.3	0.001	49.0	4.4	40.0	1961	0.000	0.75	22.1	0.002	1.190
Well 2	-1.2	-10.1	4.5	0.000	1.7	279.5	-162.5	0.8	0.003	70.5	4.8	27.8	2993	0.008	24.24	11.8	0.006	0.000
Well 3	-2.5	-7.3	2.1	0.001	4.1	369.0	-140.0	31.7	0.001	51.0	4.3	28.0	3073	0.001	2.11	5.0	0.000	0.839
Well 4	-1.3	-10.0	3.4	0.000	1.9	274.3	-161.7	-56.1	0.003	94.0	4.7	27.7	3373	0.008	28.49	10.6	0.006	0.000
Well 5	-1.9	-6.8	2.5	0.000	0.3	243.8	-167.2	-29.9	0.002	47.0	4.0	27.4	5460	0.008	15.48	3.1	0.005	0.000
Well south of Carne	-2.4	-6.8	4.3	0.001	1.4	269.7	-156.7	-129.6	0.002	87.1	4.5	28.4	2761	0.000	2.11	6.0	0.000	0.457

It is important to note that typically there are general consistencies between NMF*k* solutions for different k values. For example, the solutions for k = 2, 3, 4, 5, and 8, where all these solutions have $S(k) > 0.25$, have interesting similarities which provide additional insights into how the ML algorithm works and how the hidden geothermal signatures are extracted. The relations between signatures for these 5 NMF*k* solutions are further discussed in Appendix A. The appendix provides

244 a brief explanation of the analysis and the results of the solutions for k = 2, 3, 4, 5, and 8. The
245 analyses suggest that the k = 5 solution is optimal for the studied problem and demonstrates that
246 all acceptable solutions (for k = 2, 3, 4, and 5) can be applied to describe the dataset.



247

Figure 2: NMF*k* results for normalized reconstruction error (fitness) $O(k)$ in red color and solution robustness (based on the average silhouette $S(k)$ width of the clusters) in blue color for different numbers of signatures k. $S(k)$ values less than zero are truncated because these solutions do not provide interpretative results.

248 Figure 3 shows the predominant association of the 44 measurement locations with the extracted
249 geothermal signatures for solutions with k = 2, 3, 4, 5, and 8. The NMFk solution for k = 2
250 separates the Colorado Plateau and the volcanic field (Signature A) from the Basin and Range and
251 the Rio Grande rift (Signature B) regions (Figure 3a). The k = 3 solution combines the locations
252 of the Colorado Plateau and the volcanic field in Signature A; however, Signature B mostly
253 represents the locations of the Basin and Range and Rio Grande rift. Signature C mainly covers

254 locations in the central and northern Rio Grande rift (Figure 3b). The locations of Signature A of
255 the k = 4 solution (Figure 3c) represent the southern volcanic field. The Signature B locations fall
256 in the Basin and Range and southern Rio Grande rift. The locations of Signature C cover part of
257 the northern volcanic field and the Colorado Plateau. The Signature D locations mostly encompass
258 the central and northern Rio Grande rift (Figure 3c).

259 The k = 5 solution (Figure 3d) regrouped the four signatures of the k = 4 solution into five
260 signatures. The locations of Signatures A and E mainly cover the volcanic field. The locations of
261 Signatures B, C, and D capture the remaining three areas: the Basin and Range and southern Rio
262 Grande rift, the Colorado Plateau, and the central and northern Rio Grande rift, respectively
263 (Figure 3d). Signature A encompasses the area mostly below the Gila Hot Spring (GHS) or the
264 southern volcanic field. In contrast, Signature E covers areas north of GHS or the northern volcanic
265 field.

266 In the k = 8 solution (Figure 3e), the locations of Signature B cover the Colorado Plateau province.
267 The locations of Signatures G and H encompass the a few watersheds within the Rio Grande rift
268 (Figure 3e). The locations of Signatures A, C, and D capture the spatial variability of geothermal
269 conditions within the volcanic field. The Signature E locations fall in Basin and Range and Rio
270 Grande rift. The Signature F locations fall in the southern Rio Grande rift.

271 The solution progression with increasing k demonstrates the power of our method to refine the
272 spatial characterization of the analyzed geothermal data. There are similarities between the
273 solutions for k = 2, 3, 4, 5, and 8. The solutions for k = 2, 3, and 4 provide a higher-level
274 generalization of the geothermal signatures, while the k = 8 solution offers a refined
275 characterization of the extracted geothermal signatures.

It should be noted that none of the extracted signatures of the k = 2, 3, 4, 5, and 8 solutions perfectly represent the four identified areas. It appears that some of the locations outside but in close vicinity of a given region have similar signatures. This observation signifies that the extracted signatures are less distinct in terms of their regional association as the number of signatures increases. This discrepancy could also be related to how the regions are drawn. They are presumably identified predominantly based on surface data, but the NMF*k* is used to analyze data that are affected by subsurface processes. Perhaps at the surface, a region may be characterized as the Basin and Range, but in the subsurface, there are volcanic features that make it 'act' like a volcanic field hydrogeologically. Also, the boundaries between these regions are certainly gradational. Perhaps the sites that are assigned to different regions in ML analyses with varying numbers of signatures are 'transitional' sites.

3.2 Interpretation of the NMFk solution with the optimal number of signatures

This subsection provides an interpretation of each signature in the optimal k = 5 solution. Figure 4 plots the attribute matrix for the k = 5 solution. High-value matrix entries (dark red) define high significance, while low-value matrix entries (light red) represent low significance. We define the low, medium, and high significance by the matrix entries of <0.11, 0.11-0.5, and >0.5, respectively. Data attributes dominant for a given signature (A, B, C, D, or E) have high values in each respective column. The dominant attributes for a given signature are automatically identified by the ML algorithm and listed in Table 3. The table also presents how the extracted hidden geothermal signatures are related to a hydrothermal system and region (Figures 5 and A-1). We further categorize each signature as being more favorable for low- or medium-temperature hydrothermal systems based on the contribution of *silica geothermometer* in each signature: low and medium silica values define low- and medium-temperature systems, respectively (Figure 4;

Table 3). In the following paragraphs, we describe in detail how geology, hydrogeology, and geothermal attributes relate to each other in each signature and how they define the hydrothermal systems within the study area. Some of this information is also summarized in Table 3.

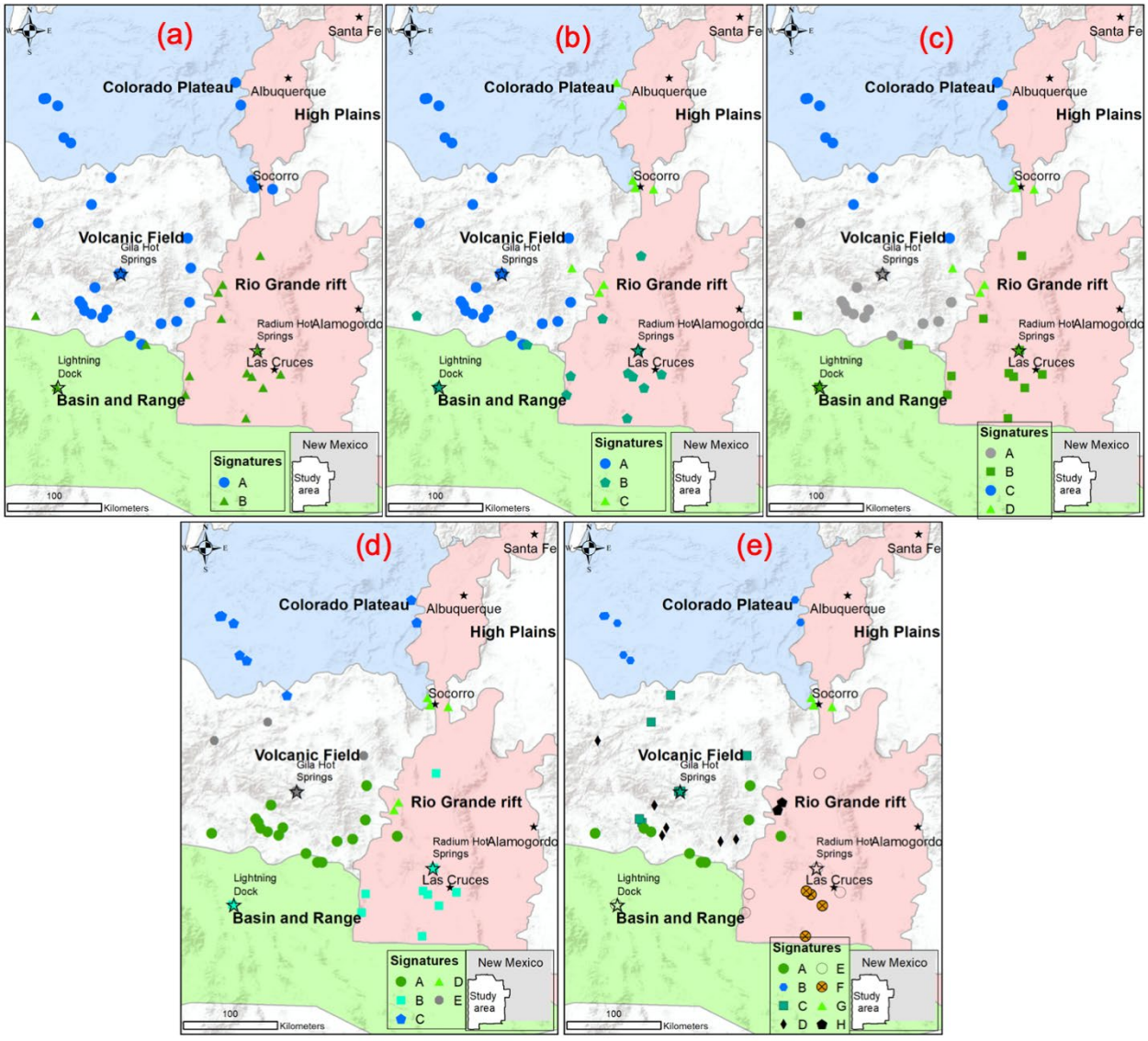
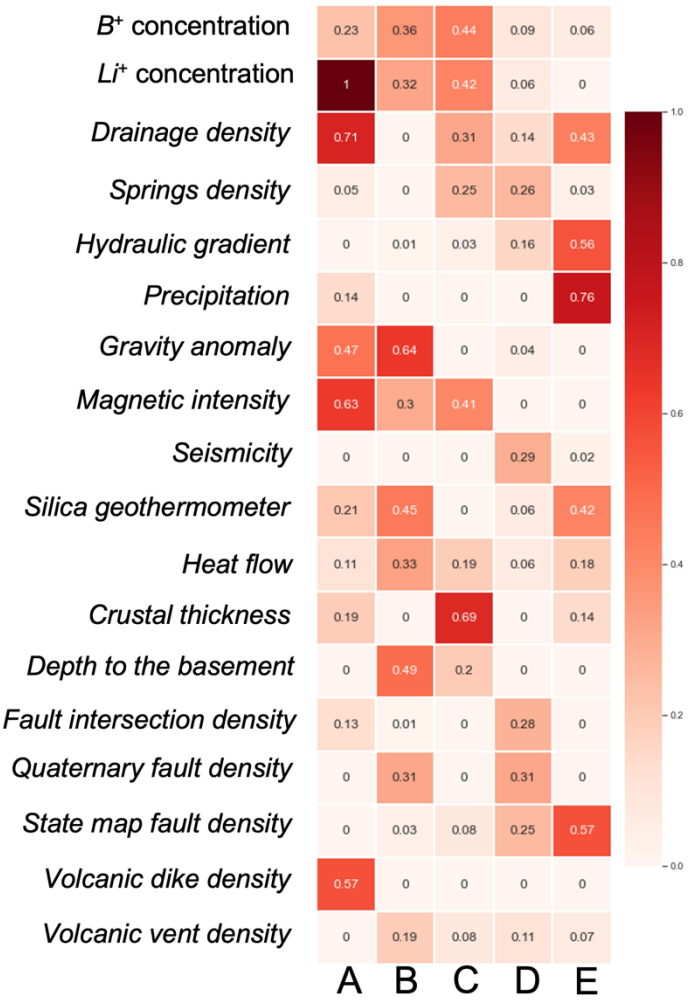


Figure 3: Spatial distribution of signatures associated with the NMF*k* solutions for the number of *k* equal to 2 (a), 3 (b), 4 (c), 5 (d), and 8 (e). Base map source: ESRI, USGS, and NOAA.



306

307 **Figure 4: NMF*k* attribute matrix for the *k* = 5 solution. High-value matrix entries (dark red)**
308 **define high significance, while low-value matrix entries (light red) represent low significance.**
309 **Medium and low values of *silica geothermometer* indicate medium- and low-temperature**
310 **geothermal resources, respectively. Data attributes dominant for a given signature (A, B, C,**
311 **D, or E) have high values along the respective columns. Base map source: ESRI, USGS, and**
312 **NOAA.**

313 **Table 3: NMF*k* extracted hidden geothermal signatures and their association with a**
314 **geothermal resource type, physical significance, dominant data attributes (Figure 4), and**
315 **regions (Figures 5 and A-1).**

Signature	Hydrothermal system type	Dominant attributes	Physical significance	Region
A	Low temperature	*Li^+ concentration* *Drainage density* *Gravity anomaly* *Magnetic intensity* *Volcanic dike density*	Shallow heat transport	Southern volcanic field
B	Medium temperature	*B^+ and Li^+ concentrations* *Gravity anomaly* *Magnetic intensity* *Silica geothermometer* *Heat flow* *Depth to the basement* *Quaternary fault density*	Deep heat transport	Rio Grande rift and Basin and Range
C	Low temperature	*B^+ and Li^+ concentrations* *Magnetic intensity* *Drainage density* *Crustal thickness*	Deep heat transport	Colorado Plateau
D	Low temperature	*Drainage density* *Spring density* *Hydraulic gradient* *Seismicity* *Fault intersection density* *Quaternary fault density* *State map fault density*	Elevated secondary permeability	Rio Grande rift
E	Medium temperature	*Drainage density* *Hydraulic gradient* *Precipitation* *Silica geothermometer* *State map fault density*	Deep groundwater circulation	Northern volcanic field

316

317 Signature A is potentially representative of low-temperature hydrothermal systems because of the
318 low contribution of the *silica geothermometer* attribute. This signature's dominant attributes are
319 *gravity anomaly*, *magnetic intensity*, *volcanic dike density*, *drainage density*, and *Li^+*
320 *concentration* (Figure 4; Table 3). *Volcanic dike density*, *gravity anomaly*, and *magnetic intensity*
321 indicate the manifestation of plutonic mafic rocks due to Tertiary volcanic events (Nakai et al.,

2017, Figure 5). The locations associated with Signature A are in the southern volcanic field. This portion of the volcanic field has a history of active Tertiary-Quaternary volcanism (Cather, 1990; Chapin et al. 2004; McIntosh et al. 1992; Ratté & Grotbo, 1979) that further enhanced *volcanic dike density* and secondary mineralization. The resultant secondary mineralization is expected to elevate *gravity anomaly* and *magnetic intensity* in this region (Beihler, 1971). However, geothermal resources are expected to be amagmatic in this area (Barroll & Reiter, 1990); therefore, these attributes are not representative of favorable hydrothermal systems in this area. *Drainage density* may represent low-permeability hard rock, which weathers in a brittle manner and with minimal infiltration as a result. Therefore, locations associated with this signature are less favorable for discovering hidden geothermal systems.

Signature B potentially represents medium-temperature hydrothermal systems because of the medium contribution of the *silica geothermometer* attribute. The other dominant attributes of this signature are B^+ and Li^+ *concentrations*, *gravity anomaly*, *magnetic intensity*, *Quaternary fault density*, *heat flow*, and *depth to the basement* (Figure 4; Table 3). *Heat flow* and *depth to the basement* are unique dominant attributes of this signature. *Heat flow* is also an indicator of a high-temperature gradient, while *depth to the basement* suggests a deep groundwater circulation. The locations associated with Signature B fall in the southern Rio Grande rift; there is also one location in the Basin and Range province, suggesting an extension of this rift signature within the Basin and Range. The area covered by Signature B spans frequent Tertiary and some Quaternary volcanic events (Nakai et al., 2017). Therefore, it is not surprising that *magnetic intensity* and *gravity anomaly* are dominant attributes. This area has thin crustal thickness (Elston et al. 1970; Nakai et al. 2017; Olson, 1979; Sanford et al., 2002), which indicates that the land surface in this area is also closer to the mantle heat source. *Depth to the basement* is the greatest depth in the study area

and may facilitate deep groundwater circulation. The high-temperature gradient, deep basement, and lower crustal thickness may be the potential cause of the medium-temperature hydrothermal systems in this region. Furthermore, locations of this signature fall in the region where two medium-temperature geothermal facilities (Lightning Dock and Radium Hot Springs) are already in use. Therefore, we postulate that the attributes associated with this signature are favorable for discovering hidden geothermal systems.

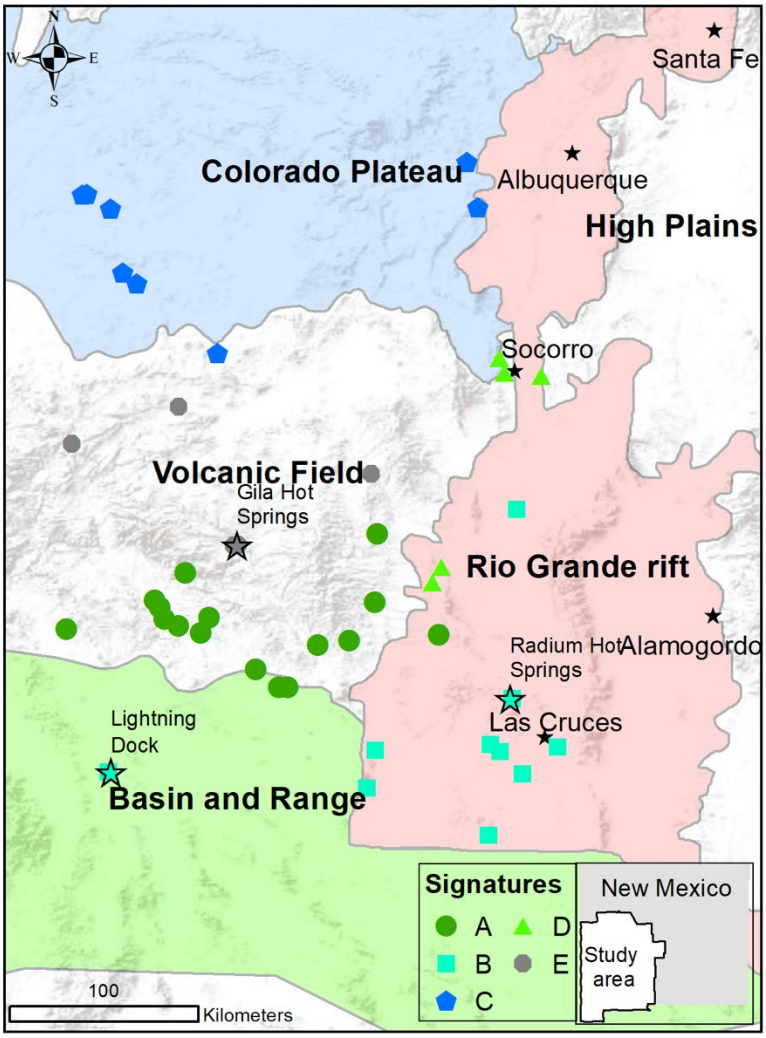


Figure 5: Locations associated predominantly with the five signatures A, B, C, D, and E. Filled stars represent cities while filled triangles represent moderate-temperature hydrothermal locations. Base map source: ESRI, USGS, and NOAA.

Signature C represents low-temperature hydrothermal systems because of the low contribution of the *silica geothermometer* attribute. The dominant attributes of this signature are B^+ and Li^+ *concentrations*, *magnetic intensity*, *drainage density*, and *crustal thickness* (Figure 4; Table 3). B^+ and Li^+ may go into solution in the groundwater because of the nearby heat source in the distant past (maybe in the Tertiary period or before), while *magnetic intensity* may indicate secondary mineralization due to Tertiary volcanic events, which may produce plutonic mafic rocks (Hunt, 1956; Lucchitta, 1979; Thompson & Zoback, 1979). These three attributes may represent volcanic activity in the past, but their effects have been dissipated (Barroll & Reiter, 1990); therefore, they are not indicators of favorable geothermal resources. *Drainage density* potentially represents low permeability in the surface that may act as a barrier to heat flow. On the other hand, *crustal thickness* is dominant and indicates that the locations associated with this signature have a large crustal thickness, which translates into a deep heat source and low regional heat flow. The combination of large crustal thickness and drainage density indicates that the locations associated with this signature are not favorable for discovering hidden geothermal systems.

Signature D represents low-temperature hydrothermal systems because of the low contribution of the *silica geothermometer* attribute. The dominant attributes of this signature are *drainage density*, *fault intersection density*, *seismicity*, *state map fault density*, *spring density*, *Quaternary fault density*, and *hydraulic gradient* (Figure 4; Table 3). *Fault intersection density*, *Quaternary fault density*, *seismicity*, and *state map fault density* suggest that this signature represents tectonic features with significant secondary permeability. The locations associated with this Signature D

375 went through extensional tectonic events (Nakai et al., 2017; Olson, 1979; Sanford, 2002).
376 Frequent tectonic events increase *fault intersection density*, which increases appreciable secondary
377 permeability. Also, *seismicity* indicates the presence of active faults along with Q*uaternary fault*
378 *density*. This signature did not get a high contribution from attributes that are indicators of medium-
379 temperature hydrothermal systems such as *heat flow* and *silica geothermometer*. The dominant
380 attributes indicate that the locations associated with this signature have enhanced/secondary
381 permeability, which is favorable for finding shallow hidden fault-controlled hydrothermal systems.

382 Signature E represents medium-temperature hydrothermal systems because of the medium
383 contribution of the *silica geothermometer* attribute. The remaining dominant attributes of this
384 signature are *drainage density*, *state map fault density*, *precipitation*, and *hydraulic gradient*
385 (Figure 4; Table 3). *State map fault density* suggests regional faults that may or may not be
386 permeable. High *precipitation* is an indicator of high recharge if there are any conduits (e.g., faults
387 and fractures) to the subsurface. A high *hydraulic gradient* potentially represents reduced
388 permeability. The high contribution of *silica geothermometer* may indicate deep groundwater
389 circulation. *Drainage density*, *state map fault density*, *precipitation*, and *hydraulic gradient* may
390 suggest a gravity-driven system with a recharge in higher elevations due to increased precipitation,
391 and groundwater discharge to lower elevations through conduits from hydrogeologic windows
392 (Witcher, 1988; Bielicki et al. 2016; 2015). A hydrogeologic window is a regional aquitard that is
393 thinned by erosions or breached by magmatic intrusions or faulting. For example, sub-vertical
394 dikes make a hydrogeologic window at Radium Hot Springs along the Rio Grande rift (Witcher,
395 1988; Bielicki et al. 2016; 2015). The dominant attributes suggest that the locations associated
396 with this signature are favorable for discovering hidden geothermal systems if a hydrogeologic
397 window is present in this area.

3.3 Characterization of Medium-temperature Hydrothermal Systems

This subsection explains the uniqueness of two medium-temperature hydrothermal systems designated by our NMF*k* analyses. NMF*k* extracts geothermal signatures and also estimates the significance of attributes and locations to define these signatures. A series of biplots are generated by pairing each of the extracted signatures to show the correlations between the signatures. In these analyses, the signatures are viewed as basis vectors, similarly to how eigenvectors are viewed in PCA analyses (Wold et al., 1987). A biplot is an exploratory scatterplot showing the mutual relation between two signatures based on how attributes (and/or locations) are associated with these signatures. In a biplot, attributes located along the axes are essential for one of the signatures but not for another signature. A biplot is also a good indicator if an attribute is not informative for both signatures; these attributes will be scattered close to the plot origin. In a biplot, well-correlated attributes are located close to each other and away from the plot origin, while uncorrelated attributes lie far apart from each other.

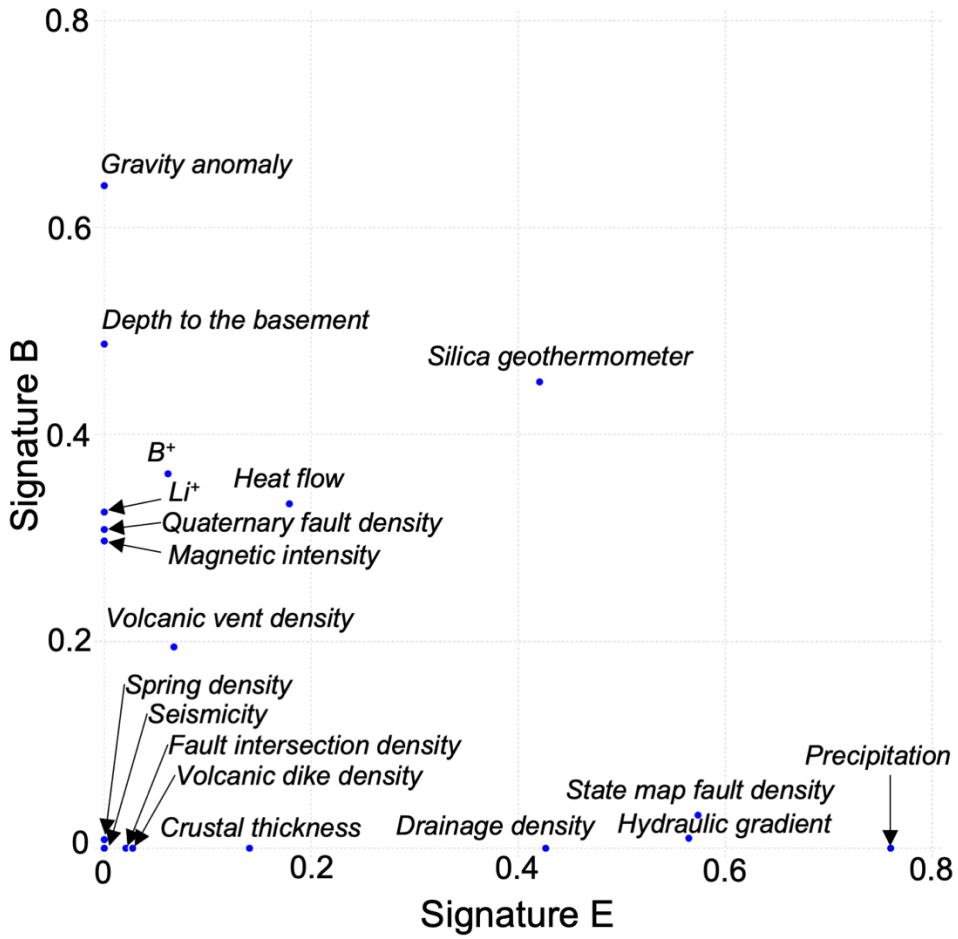
The only attribute important for both Signatures B and E is the *silica geothermometer*, which classifies these signatures as medium-temperature hydrothermal systems (figure 6). The other geothermal attributes are aligned along the axes, indicating that they are unique to Signatures B or E. Because the geology of each region is unique, the controls on hydrothermal systems also vary. Signature B falls mainly in the southern Rio Grande rift. The hydrothermal systems of this area are primarily defined by *gravity anomaly*, *depth to the basement,* B^+ and Li^+ *concentrations,* and *heat flow*. Conversely, Signature E falls in the northern volcanic field, hydrothermal systems in this area are defined by *precipitation*, *hydraulic gradient, state map fault density*, and *drainage density*.

As mentioned above, the geological, geophysical, hydrological, and geothermal characteristics of Signatures B (southern Rio Grande rift and Basin and Range) and E (the northern volcanic field) are distinct. Because these two signatures predominantly represent two different SWNM areas (the northern volcanic field and the Rio Grande rift zone), it is essential to understand what makes them distinct. Both areas went through Tertiary and Quaternary volcanic episodes, but the northern volcanic field experienced more frequent volcanic events than the Rio Grande rift zone (Cather, 1990; Chapin et al., 2004; McIntosh et al., 1992; Pepin, 2019; Ratte & Grotbo, 1979;). Also, the northern volcanic field was tectonically more active than the Rio Grande rift zone (Elston et al. 1970; Nakai et al. 2017; Olson, 1979; Sanford, 2002). However, a tectonic extensional feature that is present between the western and eastern portions of the Rio Grande rift zone but absent in the northern volcanic field (Nakai et al., 2017; Sanford, 2002) may cause the observed differentiation between the two hidden geothermal signatures. Moreover, the Rio Grande rift zone has a thinner crustal thickness than the northern volcanic field. All these observations demonstrate the unique geological and hydrological characteristics of the two areas. Therefore, the locations associated with Signature B and Signature E represent unique hydrothermal systems, and NMF*k* successfully captured these differences in the extracted geothermal signatures. It is important to note that all differences in conditions between these two areas were deduced blindly by the ML algorithm based solely on the data provided in Table 2.

3.4 Comparison with Relevant Studies

We found the potential medium-temperature hydrothermal systems primarily based on the contribution of the *silica geothermometer* attribute. *Silica geothermometer* provides a proxy of likely reservoir temperature. Although it is a questionable attribute because of its formulation and assumptions (Fournier, 1977), in the study area, it approximately matches the site groundwater

443 temperature. Our study area is large, but only 44 data locations are available. Despite the small
444 dataset, we found locations of potential low- and medium-temperature hydrothermal systems.
445 These locations are consistent with Bielicki et al. (2016; 2015) and Pepin (2019). Bielicki et al.
446 (2016; 2015) implemented the concept of the hydrogeologic windows while Pepin (2019) applied
447 PCA to a dataset for precisely the same locations and 18 similar and two additional (*depth to water*
448 and *ground surface elevation*) attributes. In both studies, critical attributes for favorable
449 hydrothermal systems are identical. Pepin (2019) found that Basin and Range and Rio Grande rift
450 region are favorable for discovering hidden geothermal systems. However, the current study found
451 the Basin and Range and Rio Grande rift regions and the northern volcanic field favorable for
452 finding hidden geothermal systems. This difference is not surprising, considering the substantial
453 difference in the applied data-analytics algorithms (PCA vs. NMF*k*). Moreover, Pepin (2019) used
454 two additional attributes *depth to water* and *ground surface elevation*. These attributes
455 substantially vary in space throughout the study area and could therefore affect the spatial
456 distribution of the clusters.



457

458 **Figure 6: Biplot showing the importance and correlations of attributes for medium-**
459 **temperature resources as defined by Signatures B and E (Table 3). Attributes far apart on**
460 **the biplot (e.g., along the opposite axes) are not correlated. Attributes near the origin are not**
461 **very important to characterize these signatures. Attributes close to each other and away**
462 **from the plot origin are correlated and essential for both signatures.**

463 **4. Conclusions**

464 Using an unsupervised ML tool (NMF*k*), this study characterized a geothermal dataset of 18
465 geothermal attributes measured at 44 locations in SWNM and (1) identified hidden geothermal
466 signatures, (2) estimated the optimal number of signatures, (3) found dominant attributes

associated with each signature, and (4) mapped spatial areas associated with the signatures. We identified potential regions suitable for further exploration to discover hidden geothermal resources based on the obtained results. By design, the analyses did not include the labeling of the hydrothermal systems based on their type and their association with a particular region. In this way, we confirmed that the algorithm could blindly group the locations based on their type and region association using only the provided data of observed geothermal attributes. Our analyses extracted a series of geothermal signatures and automatically found the optimal number of signatures (5) to characterize the data. The five signatures under this solution are labeled as A, B, C, D, and E. Based on our analyses; these signatures directly associate with either low- or medium-temperature geothermal favorability of the SWNM study area.

The extracted hidden geothermal signatures have been categorized as low- or medium-temperature hydrothermal systems based on the contribution of the *silica geothermometer* attribute on the corresponding signature. Signature A represents low-temperature hydrothermal systems, and it includes locations in the southern volcanic field. The dominant attributes are *gravity anomaly*, *magnetic intensity*, *volcanic dike density*, *drainage density*, and Li^+ *concentration*; these attributes are expected to characterize shallow heat transport. Signature B depicts medium-temperature hydrothermal systems, and it covers locations in the southern Rio Grande rift zone and the Basin and Range province. The dominant attributes are B^+ and Li^+ *concentrations*, *gravity anomaly*, *magnetic intensity*, *Quaternary fault density*, *silica geothermometer*, *heat flow*, and *depth to the basement;* these attributes potentially indicate deep heat transport. Signature C defines low-temperature systems and captures locations in the Colorado Plateau. The dominant attributes are B^+ and Li^+ *concentrations*, *magnetic intensity*, *drainage density*, and *crustal thickness;* these attributes likely demonstrate the existence of deep heat transport. Signature D represents low-

temperature hydrothermal systems, and it covers locations in the Rio Grande rift and volcanic field. The dominant attributes are *drainage density, spring density, hydraulic gradient, seismicity, fault intersection density, Quaternary fault density,* and *state map fault density;* these attributes predominantly capture the occurrence of elevated secondary permeability. Signature E is associated with medium-temperature hydrothermal systems, and it covers the northern volcanic field. The dominant attributes are *drainage density*, *state map fault density*, *precipitation*, *silica geothermometer*, and *hydraulic gradient;* these attributes likely portray deep groundwater circulation.

Out of five extracted geothermal signatures, only two signatures (B and E) are associated with medium-temperature features. The locations of these signatures have elevated regional heat flow, for example, Signature B is located mainly in the Rio Grande rift zone, including one location in the Basin and Range province where there is a high heat gradient. We should note that the only geothermal power plant in New Mexico is located in the Basin and Range province. This successful identification of correct hydrothermal system types without prior knowledge demonstrates the power of the proposed ML methodology based on NMF*k* for geothermal exploration. Signature E, another medium-temperature geothermal signature, is located in the northern volcanic field that has deep groundwater circulation.

The northern volcanic field and the Rio Grande rift zones require further exploration to find hidden geothermal systems. The PFA work by Bielicki et al. (2015) generated a preliminary geothermal favorability map. These PFA results and the knowledge accumulated in this study could be used to make an ML-enhanced geothermal favorability map of the SWNM region. Such maps could assist in discovering hidden resources and their accurate locations for geothermal heat extraction using well drilling.

To conclude, the extracted signatures by NMF*k* indicate dominant attributes to identify hydrothermal systems in each province. Moreover, the proposed NMF*k* analysis is widely applicable to extract signatures from large-scale geothermal data (including observations and simulation outputs). This broad applicability of our ML tools could aid researchers in the geothermal industry and institutions to discover, quantify, and assess hidden geothermal energy resources. Our algorithms are open source, and examples, test problems, notebooks, and documentation are available at https://smarttensors.github.io.

Acknowledgments and Disclaimer

This research is based upon work supported by the U.S. Department of Energy's (DOE) Office of Energy Efficiency and Renewable Energy (EERE) under the Geothermal Technology Office (GTO) Machine Learning (ML) for Geothermal Energy funding opportunity, Award Number DE-EE-3.1.8.1. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy (Contract No. 89233218CNA000001). Additional information regarding the datasets and codes can be obtained from Velimir V. Vesselinov (Monty) (vvv@lanl.gov) and Bulbul Ahmmed (ahmmedb@lanl.gov). Support for Jeff Pepin, Erick Burns, Drew Siler, and Jake DeAngelo was provided by the U.S. Geological Survey (USGS) Energy Resources Program. Support for Richard Middleton was provided by the DOE Small Business Innovation Research (SBIR) project *Development and Commercialization of SCO₂T to Maximize CO₂-based Subsurface Energy Potential* (DE-SC0021570).

warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe on privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Department of Energy. This article has been peer reviewed and approved for publication consistent with U.S. Geological Survey Fundamental Science Practices (https://pubs.usgs.gov/circ/1367/).

Appendix A: Discussion of the NMF*k* solutions for different number of signatures

NMF*k* analyses provided solutions for a different number of signatures. The optimal number of signatures is equal to 5, as discussed in Section 3.1. However, there is a general consistency between the extracted geothermal signatures. Here, we demonstrate these consistencies in the solutions for k = 2, 3, 4, 5, and 8; all these solutions have relatively high $S(k)$values (>0.25). The solutions for k = 2, 3, and 4 provide a higher-level generalization of the geothermal signatures (Figure 3), while the k = 8 solution allows us to refine further the characterization of the extracted geothermal signatures (Figure 3).

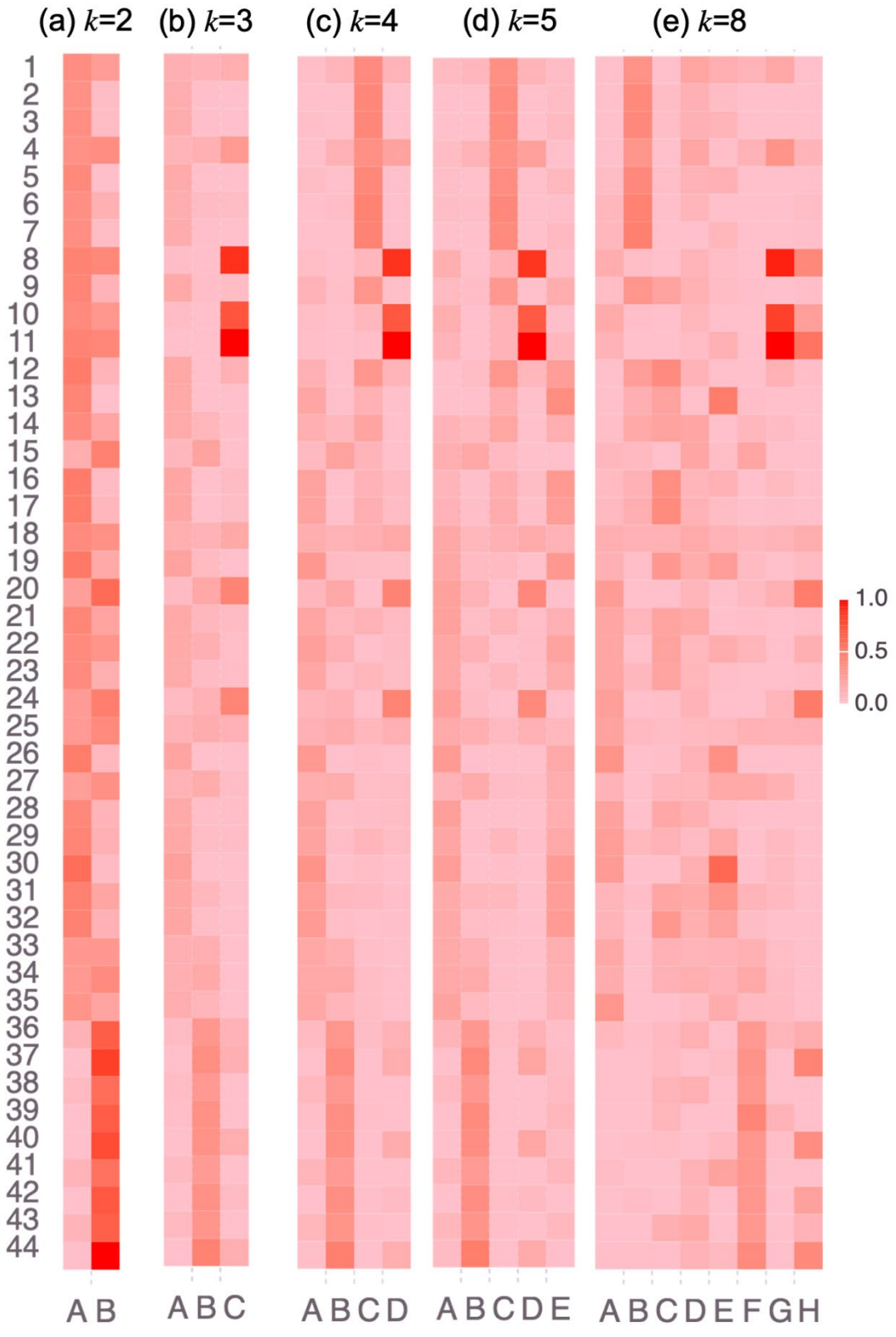
For these five solutions, the locations associated with each geothermal signature share a portion of one of the four regions or neighboring regions. If more than one geothermal signature is within a given region, they either characterize a spatial complexity or hydrothermal impacts from adjacent regions (Figure 3).

The k = 2 solution subdivides the region into two groups (Figure A-1a). Signatures A and B of the k = 3 solution (Figure A-1b) are split into Signatures A, B, and C of the k = 4 solution (Figure A-1c). Signature C for k = 3 (Figure A-1b) and Signature D for k = 4 (Figure A-1c) share similar properties. Signatures A, B, C, and D of both the k = 4 and 5 solutions (Figure A-1c, d) also possess similar properties. However, the k = 5 solution got an entirely new signature (Signature E) (Figure A-1d). The k = 8 solution (Figure A-1e) regrouped the k = 5 solution (Figure A-1d). Signature A of the k = 5 solution possesses similar properties to Signatures A and D of the k = 8 solution. Signature B of the k = 5 solution shares similar values to Signatures E and F of the k = 8 solution. Signature C of the k = 5 solution has similarities to both Signatures B and C of the k = 8 solution. Signature D for k = 5 and both Signatures G and H for k = 8 also have similar values. These associations among signatures for k = 2, 3, 4, 5, and 8 solutions are best visualized in Figures 3(a)-(e).

It is critical to mention that although the 44 locations in the W matrices are labeled (Figures 3 and A-1) to be associated predominantly with a given geothermal signature (i.e., a specific region; A, B, etc.), it does not mean the locations are related with only one signature. Instead, it means that those locations predominantly dominate commensurate signatures with contributions from other signatures too.

Figure A-2 shows the H matrices for signatures of the k = 2, 3, 4, 5, and 8, that show the progression of the extracted signatures related to the observed 18 geothermal attributes. This progression also represents the transformation of signatures as the number of signatures increases. For example, Signatures A, B, and C of the k = 3 solution (Figure A-2b) have similar properties to Signatures A, B, and both C and D of the k = 4 solution (Figure A-2c), respectively. Signatures of A, B, C, and D for k = 4 (Figure A-2c) possess similarities to signatures A and E, B, C, and D

for k = 5 (Figure A-2d), respectively. Signatures A, B, C, D, and E of the k = 5 solution share similar values with (1) A and E, F, B, (2) G and H, and (3) C and D of the k = 8 solution (Figure A-2e), respectively.



583

584 **Figure A-1: NMF*k* location (*W*) matrices for (a) *k* = 2, (b) *k* = 3, (c) *k* = 4, (d) *k* = 5, and (e) *k***

= 8. These matrix plots show the association of each location to the extracted geothermal signatures. High-value matrix entries (dark red) define high significance, while low-value matrix entries (light red) represent low significance.

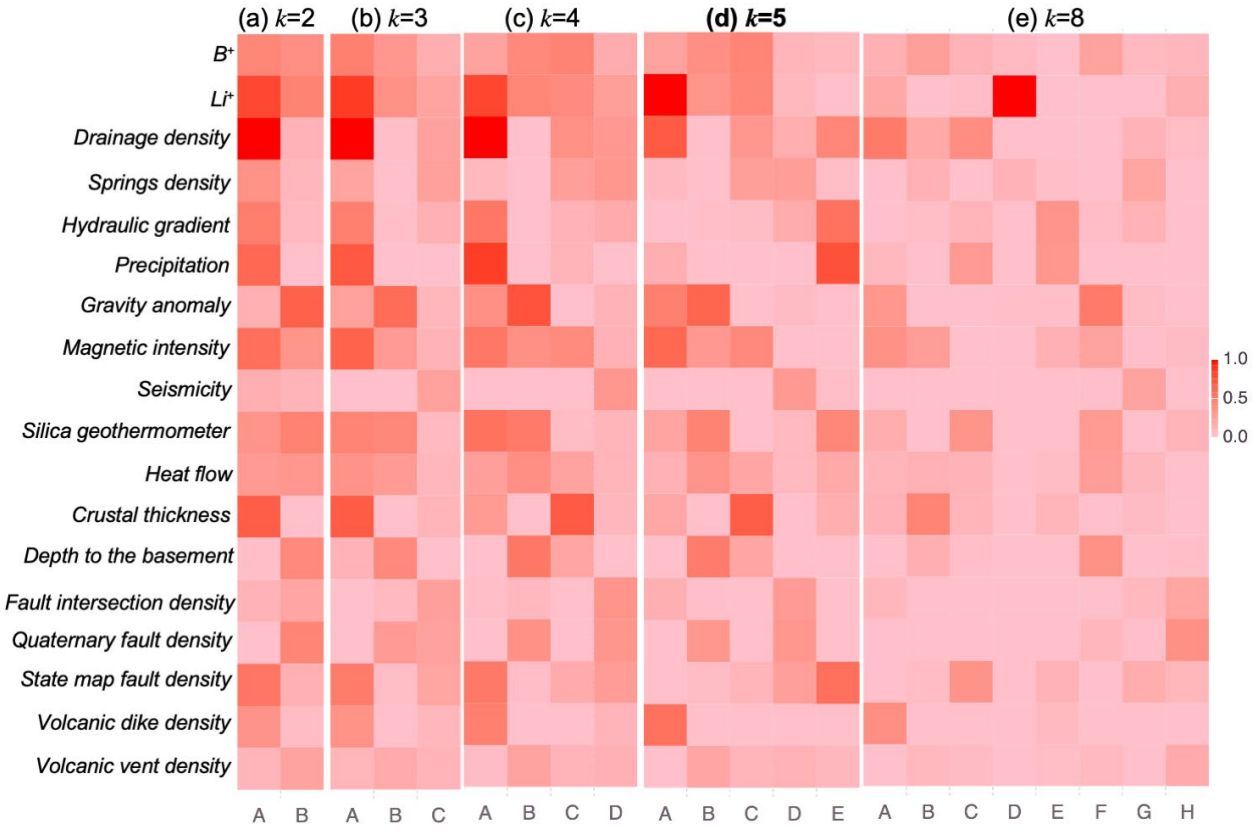


Figure A-2: NMF*k* attribute matrices for (a) *k* = 2, (b) *k* = 3, (c) *k* = 4, (d) *k* = 5, and (e) *k* = 8. These matrix plots show the contribution of each attribute on signatures. High-value matrix entries (dark red) define high significance, while low-value matrix entries (light red) represent low significance.

References

Ahmmed, B., and Vesselinov, V. (2021a). Prospectivity analyses of the Utah FORGE site using unsupervised machine learning. *Geothermal Rising, San Diego, CA*.

Ahmmed, B., Vesselinov, V., Mudunuru, M., Middleton, R., and Karra, S. (2021b). Geochemical characteristics of low-, medium-, and hot-temperature geothermal resources of the Great Basin, USA. In *World Geothermal Congress, Reykjavik, Iceland*.

Ahmmed, B., Lautze, N., Vesselinov, V., Dores, D., and Mudunuru, M. (2020a). Unsupervised machine learning to extract dominant geothermal attributes in Hawaii Island Play Fairway data. Geothermal Resources Council, Reno, NV, October 18–23.

Ahmmed, B., Vesselinov, V., and Mudunuru, M. (2020b). Machine learning to characterize regional geothermal reservoirs in the western USA. Fall Conference, Geological Society of America, Abstract T185-358249, October 26–29.

Alexandrov, B. and Vesselinov, V. V. (2014). Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization. *Water Resources Research*, 50(9):7332–7347.

Anderson, T. (2013). Geothermal potential of deep sedimentary basins in the United States. In *Unconventional Resources Technology Conference*, pages 1969–1978. Geothermal Resources Council Transaction.

Anzieta, J., Ortiz, H., Arias, G., and Ruiz, M. (2019). Finding possible precursors for the 2015 Cotopaxi Volcano eruption using unsupervised machine learning techniques. *International Journal of Geophysics*, 2019.

Barroll, M.W, and Reiter, M. (1990). Analysis of the Socorro hydrological system: Central New Mexico. *Journal of Geophysical Research*, 95, 21949-21963.

Bennett, C. and Nash, G. (2017). The convergence of heat, groundwater & fracture permeability: innovative play fairway modelling applied to the Tularosa Basin. Technical Report, Ruby Mountain Inc. and Energy & Geoscience Institute, Salt Lake City, UT.

Biehler, S. (1971). Gravity studies in the Imperial Valley. Cooperative geological-geophysical-geochemical investigations of geothermal resources in the Imperial Valley of California: Riverside, California, University of California–Riverside Education Research Service, 29-41.

Bielicki, J., Blackwell, D., Harp, D., Karra, S., Kelley, R., Kelley, S., Middleton, R., Person, M., Sutula, G., and Witcher, J. (2016). Hydrogeologic windows and estimating the prospectivity of geothermal resources. In *Proceedings, 41st Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California, February*, pages 22– 24.

Bielicki, J., Blackwell, D., Harp, D., Karra, S., Kelley, R., Kelly, S., Middleton, R., Pepin, J., Person, M., and Sutula, G. (2015). Hydrogeologic windows: regional signature detection for blind and traditional geothermal play fairways, Los Alamos National Laboratory. Technical Report, LA-UR-15-28360.

Bonafin, J., Goodman, G., and Dickey, H.K. (2019). The repowering of Lightning Dock Plant in New Mexico, Technical Report by Cyrq, USA.

Böttcher, A. and Wenzel, D. (2008). The Frobenius norm and the commutator. *Linear algebra and its applications*, 429(8-9):1864–1885.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Brott, C., Blackwell, D., and Ziagos, J. (1981). Thermal and tectonic implications of heat flow in the eastern Snake River Plain, Idaho. *Journal of Geophysical Research: Solid Earth*, 86(B12):11709–11734.

Cather, S. (1990). Stress and volcanism in the northern Mogollon-Datil volcanic field, New Mexico: Effects of the post-Laramide tectonic transition. *GSA Bulletin*, 102(11):1447– 1458. Publisher: GeoScienceWorld.

Chapin, C., Wilks, M., McIntosh, W., and Cather, S. (2004). Space-time patterns of Late Cretaceous to present magmatism in New Mexico—Comparison with Andean volcanism and potential for future volcanism. *New Mexico Bureau of Geology and Mineral Resources Bulletin*, 160:13–40.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314. Publisher: Elsevier.

Dobson, P. (2016). A review of exploration methods for discovering hidden geothermal systems. *Geothermal Resources Council Transactions*, pages 695–706.

Elston, W. E., Coney, P. J., & Rhodes, R. C. (1970). Progress report on the Mogollon Plateau volcanic province, southwestern New Mexico; no. 2. In *Tyrone-Big Hatchet Mountain-Florida Mountains Region: New Mexico Geological Society, 21st Annual Field Conference, Guidebook,* pages 75-86.

Faulds, J., Craig, J., Hinz, N., Coolbaugh, M., Glen, J., Earnery, T., Schermerhorn, W., Peacock, J., Deoreo, S., and Siler, D. (2018). Discovery of a blind geothermal system in southern Gabbs Valley, western Nevada, through application of the play fairway analysis at multiple scales. *GRC Transactions*, 42. Institution: Nevada Bureau of Mines and Geology, University of Nevada, Reno, Number: DOE-UNR-06731-08.

Faulds, J., Hinz, N., Coolbaugh, M., Ramelli, A., Glen, J., Ayling, B., Wannamaker, P., Deoreo, S., Siler, D., and Craig, J. (2019). Vectoring into potential blind geothermal systems in the Granite Springs Valley area, western Nevada: Application of the play fairway analysis at

multiple scales. *PROCEEDINGS, 44th Workshop on Geothermal Reservoir Engineering, Stanford University*. Institution: Nevada Bureau of Mines and Geology, University of Nevada, Reno, Number: DOE-UNR-06731-03.

Faulds, J., Hinz, N., Coolbaugh, M., Shevenell, L., Siler, D., dePolo, C., Hammond, W., Kreemer, C., Oppliger, G., and Wannamaker, P. (2015). Integrated geologic and geophysical approach for establishing geothermal play fairways and discovering blind geothermal systems in the Great Basin region, western USA: A progress report. *GRC Transactions*, 39 (DOE-UNR-06731-05). Nevada Bureau of Mines and Geology, University of Nevada, Reno.

Fournier, R. O. (1977). Chemical geothermometers and mixing models for geothermal systems. *Geothermics*, *5*(1-4), 41-50.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics, New York, New York.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., and Cai, J. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377.

Hartigan, J. and Wong, M. (1979). A *k*-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):100–108.

Hunt, C. (1956). Cenozoic geology of the Colorado Plateau. *US Geological Survey Professional Paper*.

Johnson, P., Rouet-Leduc, B., Pyrak-Nolte, L., Beroza, G., Marone, C., Hulbert, C., Howard, A., Singer, P., Gordeev, D., Karaflos, D., Levinson, C., Pfeiffer, P., Ming Puk, K., and Reade, W. (2021). Laboratory earthquake forecasting: A machine learning competition. *Proceedings of the National Academy of Sciences*, 118(5).

Keller, G. R., Khan, M. A., Morganc, P., Wendlandt, R. F., Baldridge, W. S., Olsen, K. H., Prodehl, C., and Braile, L. W. (1991). A comparative study of the Rio Grande and Kenya rifts. *Tectonophysics*, 197(2-4):355–371.

Kelley, S. (2010). Geothermal energy, Lite Geology. *New Mexico Bureau of Geology & Mineral Resources, a Division of New Mexico Tech*.

Klema, V. and Laub, A. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2):164–176.

Lautze, N., Ito, G., Thomas, D., Frazer, N., Martel, S. J., Hinz, N., Tachera, D., Hill, G., Pierce, H. A., Wannamaker, P. E., and Martin, T. (2020). Play fairway analysis of geothermal resources across the State of Hawai‘i: 4. Updates with new groundwater chemistry, subsurface stress analysis, and focused geophysical surveys. *Geothermics*, 86:101798.

Lautze, N., Thomas, D., Waller, D., Frazer, N., Hinz, N., and Apuzen-Ito, G. (2017). Play fairway analysis of geothermal resources across the state of Hawai‘i: 3. Use of development viability criterion to prioritize future exploration targets. *Geothermics*, 70:406–413.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Levitte, D. and Gambill, D. (1980). Geothermal potential of west-central New Mexico from geochemical and thermal gradient data. Technical Report, Los Alamos Scientific Lab., NM (USA), LA-8608-MS.

Lucchitta, I. (1979). Late Cenozoic uplift of the southwestern Colorado Plateau and adjacent lower Colorado River region. *Tectonophysics*, 61(1-3):63–95.

McClain, J. S., Dobson, P., Cantwell, C., Conrad, M., Ferguson, C., Fowler, A., Gasperikova, E., Glassley, W., Hawkes, S., Schiffman, P., Siler, D., Spycher, N., Ulrich, C., Zhang, Y., and

Zierenberg, R. (2015). Geothermal play fairway analysis of potential geothermal resources in NE California, NW Nevada, and southern Oregon: A transition between extension-hosted and volcanically-hosted geothermal fields. In *Geothermal Resources Council Annual Meeting, GRC 2015 - Geothermal: Always On*, pages 739–742. Geothermal Resources Council.

McIntosh, W., Geissman, J., Chapin, C., Kunk, M., and Henry, C. (1992). Calibration of the latest Eocene-Oligocene geomagnetic polarity time scale using 40Ar/39Ar dated ignimbrites. *Geology*, 20(5):459–463.

Medsker, L. and Jain, L. (1999). *Recurrent neural networks: design and applications*. CRC press.

Müller, A. C. and Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, Inc.

Nakai, J. S., Sheehan, A. F., and Bilek, S. L. (2017). Seismicity of the Rocky Mountains and Rio Grande rift from the EarthScope Transportable Array and CREST temporary seismic networks, 2008–2010. *Journal of Geophysical Research: Solid Earth*, 122(3):2173– 2192.

Olsen, K., Keller, G., and Stewart, J. (1979). Crustal structure along the Rio Grande rift from seismic refraction profiles. *Rio Grande rift: Tectonics and Magmatism*, 14:127–144.

Pepin, J. (2019). *New approaches and insights to geothermal resource exploration and characterization*. PhD thesis, New Mexico Institute of Mining and Technology.

Person, M., Kelley, S., Kelley, R., Karra, S., Harp, D., Witcher, J., Bielicki, J., Sutula, G., Middleton, R., Pepin, J. D., (2015). Hydrogeologic Windows: Detection of blind and traditional geothermal play fairways in southwestern New Mexico using conservative element concentrations and advective-diffusive solute transport. In Transactions - *Geothermal Resources Council*, Vol. 39.

Person, M., Phillips, F., Kelley, S., Timmons, S., Pepin, J., Blom, L., Haar, K., and Murphy,

M. (2013). Assessment of the sustainability of geothermal development within the Truth or Consequences hot-springs district. 551(65).

Porro, C., Esposito, A., Augustine, C., and Roberts, B. (2012). An estimate of the geothermal energy resource in the major sedimentary basins in the United States. *Geothermal Resources Council Transactions*, 36:1359–1369.

Ratté, J. and Grotbo, T. (1979). Chemical analyses and norms of 81 volcanic rocks from part of the Mogollon-Datil volcanic field, southwestern New Mexico. N. 79-1435, US Geological Survey.

Rouet-Leduc, B., Hulbert, C., McBrearty, I., and Johnson, P. (2020). Probing slow earthquakes with deep learning. *Geophysical Research Letters*, 47(4):e2019GL085870.

Sanford, A., Lin, K., Tsai, I., and Jaksha, L. (2002). Earthquake catalogs for New Mexico and bordering areas: 1869–1998. *New Mexico Bureau of Geology and Mineral Resources Circular*, 210:1–9.

Shervais, J. W., Glen, J. M., Liberty, L. M., Dobson, P., and Gasperikova, E. (2015a). Snake River Plain play fairway analysis – Phase 1 Report. *Transactions - Geothermal Resources Council*. Institution: Utah State Univ., Logan, UT (United States), Geothermal Resources Council.

Shervais, J. W., Glen, J. M., Liberty, L. M., Dobson, P., Gasperikova, E., Sonnenthal, E., … and Burns, E. (2015b). Snake River Plain play fairway analysis–Phase 1 Report. *Geothermal Resources Council Transactions*, *39*, 761-769.

Shervais, J. W., Glen, J. M., Nielson, D. L., Garg, S., Liberty, L. M., Siler, D., … and Snyder, N. (2017). Geothermal play fairway analysis of the Snake River Plain: Phase 2 (no. INL/con-17-42832). Idaho National Lab. (INL), Idaho Falls, ID (United States).

754 Siler, D.L. and Pepin, J.D., 2021. 3-D Geologic controls of hydrothermal fluid flow at Brady
755 geothermal field, Nevada, USA. Geothermics, 94, p.102112.
756 Siler, D., Pepin, J., Vesselinov, V., Mudunuru, M., and Ahmmed, B. (2021). Machine learning to
757 identify geologic factors associated with production in geothermal fields: A case study using
758 3D geologic data, Brady geothermal field, Nevada. *Geothermal Energy* 9(1), p. 1-17.
759 Siler, D. L., Faulds, J. E., Hinz, N. H., Dering, G. M., Edwards, J. H., and Mayhew, B. (2019).
760 Three-dimensional geologic mapping to assess geothermal potential: examples from Nevada
761 and Oregon. *Geothermal Energy*, 7(1):2.
762 Siler, D., Zhang, Y., Spycher, N., Dobson, P., McClain, J., Gasperikova, E., Zierenberg, R.,
763 Schiffman, P., Ferguson, C., Fowler, A., and Cantwell, C. (2017). Play-fairway analysis for
764 geothermal resources and exploration risk in the Modoc Plateau region. *Geothermics*, 69:15–
765 33.
766 Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of
767 cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.
768 Smith, R. (2004). Geologic setting of the Snake River Plain aquifer and vadose zone. *Vadose Zone*
769 *Journal*, 3(1):47–58.
770 Thompson, G. and Zoback, M. (1979). Regional geophysics of the Colorado Plateau.
771 *Tectonophysics*, 61(1-3):149–181.
772 UNM (2018). Digital geologic map of New Mexico – Volcanic vents.
773 https://catalog.data.gov/dataset/digital-geologic-map-of-new-mexicovolcanic-vents
774 USGS (2018a). Energy and environment in the Rocky Mountain area. https://my.usgs.gov/eerma/.
775 Accessed on 20th May 2022.

776 USGS (2018b). USGS water data for the Nation: U.S. Geological Survey National Water
777 Information System database, accessed May 20, 2022, at
778 http://dx.doi.org/10.5066/F7P55KJN.
779 Vesselinov, V., Ahmmed, B., Mudunuru, M., Karra, S., and Middleton, R. (2021). Hidden
780 geothermal signatures of southwest New Mexico. In *Proceedings of the World Geothermal*
781 *Congress, Reykjavik, Iceland*.
782 Vesselinov, V., Mudunuru, M., Ahmmed, B., S, K., and Middleton, R. (2020). Discovering
783 signatures of hidden geothermal resources based on unsupervised learning. In *Proceedings of*
784 *the 45th Annual Stanford Geothermal Workshop*.
785 Vesselinov, V. V., Mudunuru, M. K., Karra, S., O'Malley, D., & Alexandrov, B. S. (2019).
786 Unsupervised machine learning based on non-negative tensor factorization for analyzing
787 reactive-mixing. *Journal of Computational Physics*, *395*, 85-104.
788 Vesselinov, V. V., Alexandrov, B. S., and O'Malley, D. (2018). Contaminant source identification
789 using semi-supervised machine learning. *Journal of Contaminant Hydrology*, 212:134–142.
790 Watson, L., Johnson, J., Sciotto, M., and Cannata, A. (2020). Changes in crater geometry revealed
791 by inversion of harmonic infrasound observations: 24 December 2018 Eruption of Mount
792 Etna, Italy. *Geophysical Research Letters*, 47(19):e2020GL088077.
793 Williams, C., Reed, M., DeAngelo, J., and Galanis, S. (2009). Quantifying the undiscovered
794 geothermal resources of the United States. In *Geothermal Resources Council 2009 Annual*
795 *Meeting*, volume 33.
796 Witcher, J.C. (1988). Geothermal resources of southwestern New Mexico and southeastern
797 Arizona. *New Mexico Geological Society Guidebook*, 39, 191-198.

798 Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and*
799 *intelligent laboratory systems*, 2(1-3):37–52.
800 Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep
801 neural networks? *Advances in neural information processing systems*, *27*.