Los Alamos
NATIONAL LABORATORY
EST. 1943

LA-UR-21-24325
Approved for public release; distribution is unlimited.

Title: Cloud Fusion of Big Data and Multi-Physics Models using Machine Learning for Discovery, Exploration, and Development of Hidden Geothermal Resources

Author(s): Vesselinov, Velimir Valentinov; O'Malley, Daniel; Frash, Luke Philip; Ahmmed, Bulbul; Rupe, Adam Thomas; Karra, Satish; Middleton, Richard Stephen (Richard); Alexandrov, Boian; Mudunuru, Maruti K; Mims, Mark; Jameson, Kevin; Sun, Alex; Scanlon, Bridget; Banerji, Amy; Tartakovsky, Daniel M.; Horne, Roland N.; Zhou, Zitong; Maguire, Conor; Skillman, Sam; Scharer, John

Intended for: Report

Issued: 2021-05-05

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

GeoThermalCloud:

Cloud Fusion of Big Data and Multi-Physics Models using Machine Learning for Discovery, Exploration, and Development of Hidden Geothermal Resources

FOA Number: 1956-1516
FOA Title: Machine Learning for Geothermal Energy
Award Number: DE-EE3.1.8.1
Recipient: LANL

PI: Velimir Vesselinov (“monty”)
Institutions: LANL, Stanford, Google Inc, Descartes Labs Inc, and the University of Texas-Austin (Bureau of Economic Geology)
Participants: Daniel O’Malley (LANL), Luke Frash (LANL), Bulbul Ahmmed (LANL), Adam Rupe (LANL), Satish Karra (LANL), Richard Middleton (LANL), Boian Alexandrov (LANL), Maruti Mudunuru (PNNL), Mike Mims (Google), Kevin Jameson (Google), Alex Sun (UT Austin BEG), Bridget Scanlon (UT Austin BEG), Amy Banerji (UT Austin BEG); Daniel Tartakovsky (Stanford), Roland N. Horne (Stanford); Zitong Zhou (Stanford), Conor Maguire (Descartes Lab), Sam Skillman (Descartes Lab), John Scharer (Descartes Lab)

1 Executive Summary

The project is motivated by the challenges, risks, and costs associated with geothermal exploration and production. Many processes and parameters impacting geothermal conditions are poorly understood. Diverse datasets are available to help characterize subsurface geothermal conditions (public and proprietary; satellite, airborne surveys, vegetation/water sampling, geological, geophysical, etc.). Yet, it is not clear how to properly leverage these datasets for geothermal exploration due to an incomplete understanding of how physical processes impacting subsurface geothermal conditions are represented in these observations. Recent advancements in machine learning (ML) provide great promise to resolve these issues.

The tremendous challenges and risks of geothermal exploration and production bring the demand for novel ML methods and tools that can (1) analyze large field datasets, (2) assimilate model simulations (large inputs and outputs), (3) process sparse datasets, (4) perform transfer learning (between sites with different exploratory levels), (5) extract hidden geothermal signatures in the field and simulation data, (6) label geothermal resources and processes, (7) identify high-value data acquisition targets, and (8) guide geothermal exploration and production by selecting optimal exploration, production, and drilling strategies. Our goals and work under Phases 1 and 2 (as proposed) of this project address all these needs.

Under Phase 1, we have developed and applied our novel LANL-developed machine learning (ML) methodology to discover and extract new (unknown/hidden) geothermal signatures present in existing site, synthetic, and regional datasets. Our ML analyses also identified high-value data acquisition strategies that can reduce geothermal exploration/production costs and risks. Our ML methods also categorized geothermal data, which is applied to generate geothermal data labels (e.g., geothermal resource types). The end product of our effort is the development of a flexible open-source cloud-based ML framework for geothermal exploration, called ***GeoThermalCloud***, which can fuse existing geothermal datasets and multi-physics models. ***GeoThermalCloud*** will also allow for the treatment of both public and proprietary datasets. This is an essential feature considering the high sensitivities associated with the use of proprietary data. Processing simultaneously public and proprietary datasets will significantly increase the quality and applicability of the obtained ML results. Moreover, ***GeoThermalCloud*** framework includes a series of advanced pre-processing, post-processing, and visualization tools which tremendously simplify its application for real-world problems. These tools make the ML results understandable and visible even for non-experts. ML and subject-matter expertise are not a critical requirement to use our ***GeoThermalCloud*** framework.

ML methods embedded in the ***GeoThermalCloud*** have been extensively tested and validated. In this report, as well as in a series of presentations and submitted research papers, we have demonstrated that ***GeoThermalCloud*** can be applied to discover hidden geothermal signatures for a series of diverse datasets and regions. The analyzed ML problems are (results for all sites are available in our project presentation):

- **Southwest New Mexico (SWNM)**: Identified low- and medium-temperature hydrothermal systems; found dominant attributes and spatial distribution of extracted hidden hydrothermal signatures; demonstrated blind predictions of the regional physiographic provinces[1–3].
- **Great Basin**: Extracted hidden geothermal signatures associated with low-, medium-, high-temperature hydrothermal systems, their dominant characterization attributes, and spatial distribution within the study area[4,5]. The analyses are based on the public data available at the Nevada Bureau of Mines and Geology website [http://www.nbmg.unr.edu/Geothermal/GeochemDatabase.html].
- **Brady site, Nevada**: Successfully defined relations between well types (production, injection, non-production) and attributes characterizing site conditions (faulting, geology, in situ state of stress)[6].
- **West Texas**: Subdivided the region into three areas; the western portion has higher geothermal potential at a lower depth than the middle and eastern portions.

- **Tularosa Basin, New Mexico**: Analyzed 21 Play Fairway Analysis (PFA) attributes at 120 locations[7]; data comes from past PFA work in this region[8]. ML analyses identified geothermal signatures associated with low-, medium-, and high-temperature hydrothermal systems. Dominant attributes and spatial distribution of the geothermal signatures are also defined.
- **Tohatchi Springs, New Mexico**: Explored 19 geothermal attributes at 43 locations[9]. Successfully defined geothermal signatures associated with low- and medium-temperature hydrothermal systems, their dominant attributes, and spatial distribution.
- **Hawaii**: Analyzed four islands' data separately and jointly; ML identified low-, medium-, and high-temperature hydrothermal systems and their dominant characterization attributes[10].
- **Utah FORGE**: Performed prospectivity analysis to identify future drilling locations using geological, geochemical, and geophysical attributes[11]. Maps of temperature at depth, and heat flow constructed based on the available data. Processed data includes satellite (InSAR), geophysical (gravity, seismic), geochemical, and geothermal attributes. Prospectivity maps generated and drilling locations proposed for future geothermal field exploration.
- **EGS Collab:** Field experiment data processed to extract dominant temporal patterns observed in 49 data streams; erroneous measurement attributes and periods automatically identified; interrelated data streams automatically identified.

In addition, two synthetic datasets were generated to (1) validate our ML methods, (2) demonstrate how our ML tools can fuse model outputs in the geothermal analyses, (3) identify high-value data acquisition and reservoir production strategies, and (4) estimate exploration/production costs and risks.

The first synthetic dataset was generated utilizing ***GeoDT***, a novel LANL-developed multi-physics code for prediction of the performance of geothermal energy systems. ***GeoDT*** evaluates how geothermal site data conditions impact design decisions related to construction of enhanced geothermal systems (EGS). ***GeoDT*** is applied to evaluate the combined effect of >90 input parameters on thermal power and electrical power output based on >2000 random realizations; the analyses are representative of the Utah FORGE site conditions. The model inputs and outputs are analyzed using our ***GeoThermalCloud*** ML tools. They were able to identify key controlling attributes, separate the relative impact of different physical processes on production, and associate these impacts to ***GeoDT*** model inputs. Our study focused on the influence of stress state and natural fractures on geothermal well drilling and well production. ML analyses identified well spacing and well orientation as critical parameters for energy production.

The second synthetic dataset represented the SWNM geothermal conditions and was generated using the LANL-developed simulator PFLOTRAN. PFLOTRAN simulations were executed to create a large synthetic dataset exploring existing uncertainty in the geothermal conditions. The dataset is designed to explore how geothermal reservoir attributes such as temperature and heat flow inform observables near the ground surface, such as tracer element concentrations. Using our ML tools, this dataset is analyzed to understand interdependencies between model input and outputs. We also developed an ML model capable of performing fast prediction and uncertainty quantification of underground geothermal conditions.

These two synthetic-data analysis efforts demonstrate how our ML methods can be applied in tandem with forward-looking multi-physics models to address questions related to value-added data acquisition strategies and to reduce exploration and development costs. These synthetic analyses also validate our novel ML methodology.

For the Brady and Utah FORGE sites, our ML characterization particularly focused on impact in-situ stress on geothermal production.

All the data and codes required to reproduce the ML results for SWNM, Brady, and Great Basin studies presented in this report are available at the GDR (https://gdr.openei.org/submissions/1297) and at the ***GeoThermalCloud*** GitHub repository (https://github.com/SmartTensors/GeoThermalCloud.jl). The data and the codes for other problem analyses performed under this project will be available shortly on the website as well.

ML tools in ***GeoThermalCloud*** are based on a series of novel LANL-developed patented ML tools called ***SmartTensors*** (https://github.com/SmartTensors). ***SmartTensors*** have already been applied to solve a wide range of real-world problems (from COVID-19[12] to wildfires; http://tensors.lanl.gov). ***SmartTensors*** has been recently nominated for an R&D 100 award.

GeoThermalCloud and ***SmartTensors*** are written in Julia: a novel, fast programming language specifically designed for technical computing and machine learning (https://julialang.org). Julia is two orders of magnitude faster than Python, R, and MATLAB. It provides computational speed equivalent to C and FORTRAN. Moreover, it is a scripting language that is as easy to use as Python, R, and MATLAB. Julia can be applied to develop Jupyter and Pluto notebooks. (Pluto is a novel reactive Julia-centric notebook framework; https://github.com/fonsp/Pluto.jl). Julia is also scalable and can be executed without any code changes on handhelds, laptops, supercomputers, and cloud computing platforms. Julia programming language is actively used in numerous fields, including financing, oil/gas production, climate modeling, and pharmaceutical research. We have active collaborations with developers of the Julia language. We will involve Julia Computing as a collaborator in the proposed Phase 2 of our project.

GeoThermalCloud, as well as our ***SmartTensors*** ML tools in general, are designed to be computationally efficient and scalable. Our ML analyses can be executed to diverse computational platforms: from handhelds and laptops to supercomputers and cloud frameworks. The ML tools are designed to automatically and efficiently utilize available hardware accelerators such as GPUs (Graphics Processing Units) and TPUs (Tensor Processing Units) and diverse computing platforms involving a large number of compute nodes. ***SmartTensors*** framework has already been demonstrated to process tens of terabytes (TBs) of data using DOE supercomputers.

Our project is a collaborative effort that includes researchers from LANL, Stanford, Google Inc, Descartes Labs Inc, and the University of Texas-Austin (Bureau of Economic Geology).

Our Phase 1 completion report summarizes our efforts under this project. All the project goals, milestones, deliverables, and go/no-go decision points outlined in the original Phase 1 proposal have been successfully achieved. No changes have been made in the originally proposed work, milestones, tasks, and go/no-go decisions.

2 GeoThermalCloud

2.1 Goals

The main project goal is to apply machine learning (ML) to discover and extract new (unknown/hidden) geothermal signatures present in existing synthetic, site, and regional datasets. Our ML analyses also aim to identify high-value data acquisition strategies that can reduce geothermal exploration/production costs and risks. Our ML methods also provide categorization of geothermal data, which is applied to generate geothermal data labels. In summary, our project goals are summarized as follows:

1. Apply ML to discover and extract new (unknown/hidden) geothermal signatures in existing large datasets
2. Categorize geothermal data and generate labels
3. Identify high-value data acquisition strategies
4. Develop a general open-source cloud-based ML framework for geothermal exploration
5. Fuse big data and multi-physics models
6. Test & validate that ML methods can discover hidden geothermal signatures

Within Phase 1, we have achieved all these goals. In addition, we have identified the following go/no-go decision criteria: (1) extraction of new geothermal signatures, (2) generation of labeled datasets, (3) construction of subsurface heat and geothermal prospectivity maps, and (4) identification of high-value data acquisition targets via ML analysis. As demonstrated by our analyses presented in Section 3, we have achieved these go/no-go decision points as well.

2.2 Design

The end product of our project is to develop a flexible open-source cloud-based ML framework for geothermal exploration, called ***GeoThermalCloud***, which can fuse existing geothermal datasets and multi-physics models. ***GeoThermalCloud*** will allow for the treatment of both public and proprietary datasets. This is an essential feature considering the high sensitivities associated with the use of proprietary data. Interpreting simultaneously public and proprietary datasets will increase the quality and applicability of the obtained ML results tremendously. ***GeoThermalCloud*** framework also includes a series of advanced pre-processing, post-processing, and visualization tools which tremendously simplify its application for real-world problems. These tools make the ML results understandable and visible even for non-experts. ML and subject-matter expertise are not a critical requirement to use our ML framework.

ML methods embedded in the ***GeoThermalCloud*** have been extensively tested and validated. This report demonstrates how ***GeoThermalCloud*** can be applied to discover hidden geothermal signatures for a series of diverse datasets and regions within the U.S. ***GeoThermalCloud*** is open-source and available at GDR (https://gdr.openei.org/submissions/1297) and GitHub (https://github.com/SmartTensors/GeoThermalCloud.jl). The GitHub repository is up-to-date because it is used by our team to facilitate our collaborative coding effort; the GDR repository is regularly updated as well. The data and scripts, including Jupyter and Pluto notebooks, required to reproduce the ML results discussed here are also available at the GitHub and GDR repositories. Below, we present the ML methodology applied and the obtained results.

2.3 Methodology

GeoThermalCloud utilizes our novel, open-source, LANL-developed, patented ML methods and computational tools. All these methods are distributed as ***SmartTensors*** (http://tensors.lanl.gov, https://github.com/SmartTensors). ***SmartTensors*** is a toolbox for unsupervised and physics-informed

machine learning based on matrix/tensor factorization constrained by penalties enforcing robustness and interpretability (e.g., nonnegativity; physics and mathematical constraints; etc.). ***SmartTensors*** has already been successfully applied to analyze diverse datasets related to a wide range of problems: from COVID-19[12] to wildfires and text mining. Also, the ***SmartTensors*** framework can handle big data and has already been demonstrated to perform tens of TBs of data using DOE supercomputers[13].

The two most commonly used ML algorithms in SmartTensors are NMF*k* and NTF*k*. They perform nonnegative matrix/tensor factorization coupled with customized *k*-means clustering[14–16]. NMF*k* and NTF*k* are capable of identifying (i) the optimal number of hidden signatures in data, (ii) the dominant set of attributes in data that correspond to identified hidden signatures, and (iii) locations associated with each hidden signature.

All datasets are formed by directly observable quantities, while the underlying processes or data signatures usually remain unseen, hidden, or latent[17]. These hidden signatures (or features/signals) can be either impossible to measure directly or are simply unknown. For example, let us assume that a series of microphones are placed in a noisy ballroom[18] where many people are talking. The collected data records the mixtures of voices, sounds, and noises. The latent signatures are the individual voices that cannot be recorded separately but can be extracted from the collected data. Extracting latent signatures reduces the dimensionality of the data and defines low-dimensional subspaces[19,20] that represent the entire dataset. After the extraction, the obtained information is post-processed by subject-matter experts to identify the physical meaning (e.g., broken glass) or the origin (e.g., recognize voices of individuals) of the extracted signatures.

Similarly, our unsupervised ML techniques have been applied here to extract latent signatures and hidden (mixed) physical processes embedded in large, diverse geothermal datasets. Hidden (latent) signatures provide a low-dimensional and compressed representation of the processed dataset. They can also be viewed as basis vectors providing optimal data projection. In the case of geothermal applications, these signatures typically represent information about a series of physical processes which occur in observable and/or simulated datasets. These signatures can be multi-dimensional capturing processes occurring in spatiotemporal space and captured by a set of diverse attributes. Geothermal attributes we have processed in this report include temperature, gradients, geothermometers, conductivity, permeability, fluxes, fracture densities, in-situ stresses, etc.

To discover hidden signatures and their optimal number in large geothermal datasets, NMF*k* and NTF*k* are at the forefront among various unsupervised ML methods such as nonnegative matrix factorization (NMF), principal component analysis (PCA), independent component analysis (ICA), regular and high-order singular value decomposition (SVD/HOSVD), nonnegative tensor factorization (NTF), and Gaussian process/mixture modeling. In contrast, with traditional NMF[21], NMF*k* allows for automatic identification of the optimal number of signatures (features) present in the data[15]. The nonnegativity constraint makes the decomposed matrices easier to interpret than PCA, SVD, and ICA because the extracted signatures are additive. Moreover, NMF*k* and NTF*k* can handle huge (TBs), real, categorical, and missing data. Dealing with missing data is challenging or impossible for other supervised and unsupervised ML methods. Even more importantly, the missing data (some or all of it) can be reconstructed from available data using the obtained NMF*k* and NTF*k* results. Our ML methods also provide estimates of uncertainties associated with the estimated missing data. All of these features of our ML methods make them very well suitable for geothermal ML analyses.

ML methods, in general, can be subdivided into supervised, physics-informed, and unsupervised. The supervised methods require attributes and corresponding labels of the analyzed data[22]. The labeling should be done by subject-matter experts who can identify, for example, locations with high-, intermediate-, and low-temperature geothermal prospectivity or specific geologic features such as fault offsets. The supervised methods are then applied to learn geothermal prospectivity 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. In essence, the supervised methods cannot discover something that is not already known and provided as labels in the training dataset. The supervised methods are highly efficient to

process large datasets and find out how the processed data can be categorized. For example, they can be trained to recognize images of cats and dogs; however, they will not recognize horses if they have not been trained already to identify them. Furthermore, the supervised ML methods are sensitive to noise in the analyzed data or the so-called "adversarial examples"[23] where small, undetectable for the human eye, can fool the detection capabilities of these methods. All of these features make the application of supervised ML methods challenging for real-world science applications[24]. Commonly used supervised methods include deep neural networks[25], convolutional neural networks[26], recurrent neural networks[27], and random forest[28].

Physics-informed ML (PIML) methods also learn from data as the supervised methods, but they also include preconceived science knowledge through equations and models representing physics laws, constitutive relationships, and processes. Physics information can be (1) directly embedded in the ML framework[29] or (2) added as penalties in the ML loss minimization process[30]. However, the physics-informed neural networks (PINN) are problem specific and not general as the traditional supervised deep neural networks. Therefore, the construction PINN needs subject-matter expertise related to the analyzed problem. However, PIML analyses have better efficiency, accuracy, and robustness compared to the traditional ML analyses. The PINN development requires computationally efficient and general differentiable programming tools currently available only in the Julia programming language. Under the proposed Phase 2 of our project, the ***GeoThermalCloud*** ML framework will be expanded to incorporate PIML methods.

In contrast, the unsupervised ML techniques extract information from existing datasets without any prior labeling or subject-matter preprocessing. The unsupervised ML is applicable to discover unknown features and unmix mixed signals present in the processed data. Similarly, unsupervised ML techniques have been applied here to extract latent features and hidden (mixed) physical processes embedded in large, diverse geothermal datasets. Our novel nonnegative matrix/tensor factorization methods coupled with customized *k*-means clustering[14–16] (NMF*k* and NTF*k)* are unsupervised. Other commonly used methods include SVD[31], PCA[32], ICA[33], *k*-means clustering[34], Gaussian process/mixture modeling[35], nonnegative matrix/tensor factorization (NMF/NTF)[21]. Moreover, the labeling process required by the supervised methods discussed above can also be automated by applying unsupervised ML[22].

2.4 Nonnegative matrix/tensor factorization

Detailed descriptions of our novel NMF*k* and NTF*k* algorithms are provided in our papers[15,16]. Here, we give just a high-level description of the NMF*k* and NTF*k* methodology to support the geothermal analyses' discussion in this report. NMF*k* performs matrix factorization of a data matrix, $X_{m\times n}$, where the *m* rows represent here measurement locations, and the *n* columns are the values of the geothermal attributes. The goal of NMFk is to find the optimal number of signatures *k* that describe the analyzed dataset. This is accomplished by matrix factorization, which can be represented as using:

$$X \cong W \times H \tag{1}$$

where $W_{m\times k}$ is an "attribute" matrix characterizing the significance of attributes and $H_{k\times n}$ a "location" matrix captures the importance of locations and their spatial association. It is important to note that all the elements of matrices *W* and *H* are unknown. The number of signatures *k* is also unknown. The matrix factorization in (1) provides an approximate representation of the data *X*. To solve for all the unknowns, NMF*k* performs a series of matrix factorization with random initial guesses for *W* and *H* elements and for a range of values of *k*; theoretically, *k* can range between 2 and min(*m*,*n*). For a given number of signatures k, Equation 1 is solved iteratively by minimizing the reconstruction error $O(k)$:

$$O(k) = ||X - W \times H||_F \tag{2}$$

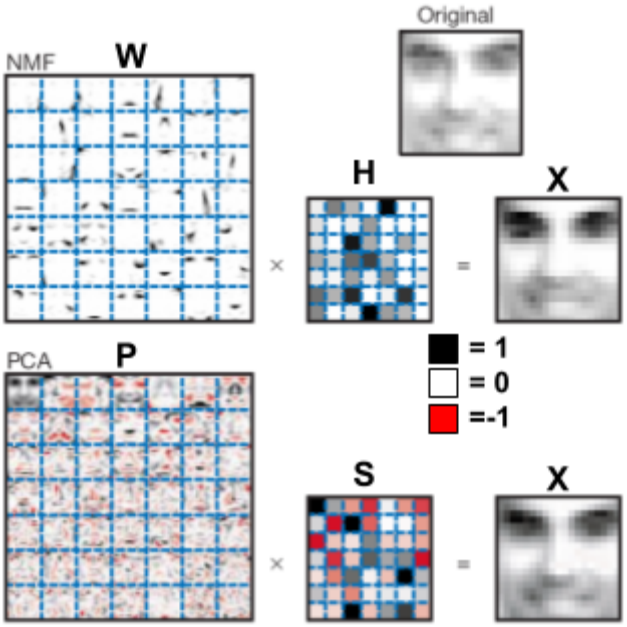


Fig 2.1.1: ML of faces using NMF and PCA. Nonnegativity constraint provides additive, sparse and interpretable results where facies features such as eyes and noses are well defined. Interpretation of PCA results is challenging (after Lee & Seung, 1999).

by constraining the W and H elements to be greater or equal to zero (nonnegative), and F defines the Frobenius norm[36]. Under the NMFk 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 H solutions are clustered into k clusters using a customized k-means clustering[15,16]. The average silhouette width $S(k)$ based on cosine norm is computed for all k clusters. This metric measures how well the random NMF solutions are clustered for a given value of [15,16]. 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 k increases. The average silhouette width behavior is more complicated; $S(k)$generally declines as k increases from 1 to -1. 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 reconstruction 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 here[15,37].

We listed the benefits of NMFk over similar unsupervised ML tools above. Here, we will provide a brief mathematical and graphical comparison between NMFk and PCA to show how NMFk differentiates itself from PCA. Another data analysis method similar to NMFk is PCA[32]. It factorizes the data matrix X into score (S) and principal component (P) matrices. The factorization can be represented as::

$$X \cong SP \tag{3}$$

S is a diagonal matrix. As in NMF, Equation 3 is solved iteratively by minimizing the following function:

$$L = \| |X - SP| \|_F \tag{4}$$

PCA searches for linear combinations in the data by projecting each data point onto an optimal set of principal components (PCs) to obtain a low-dimensional representation of data while preserving maximum data variation. PCs are ordered by the magnitude of data variance as captured by the S diagonal elements.

Even though NMF and PCA are mathematically similar, the ML results obtained by both methods are very different. As already demonstrated, the NMF and PCA can reconstruct human faces very well (Fig.2.1.1). Both methods extracted 49 basic facial features (can also be called dictionaries, basis vectors,

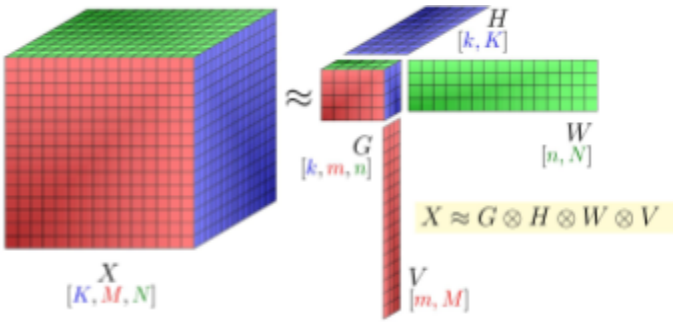


Fig.2.1.2: Example factorization of a data tensor X into a smaller core tensor G and three matrix factors for each dimension. Matrix factors represent signatures in different dimensions. G defines how these signatures are mixed.

or eigenvectors; the matrices on the left) needed to reconstruct the entire training set of ~1,000 faces (i.e., both methods performed data dimensionality reduction from ~1,000 to 49). However, the nonnegativity constraints provide additive, sparse, and interpretable results where facial features such as eyes and noses are well defined (Fig.2.1.1; W matrix; top left). NMF face reconstruction is obtained by adding a series of dominant features shown as black squares in the H matrix. In contrast, the interpretation of PCA results is challenging. The first face (upper left corner of matrix P) is the average face of the training set, and the reconstruction of face X is obtained by adding and subtracting a series of facies features (in P) based on the weights (in S; red defines negative values or feature subtraction; black represents positive values or feature addition).

In addition to matrices, our ML methods can process multi-dimensional datasets, i.e., tensors. Most of the geothermal data are multi-dimensional. The data indices can be space coordinates and time for each spatiotemporal location, and there might be numerous observables coming from different data sources (streams). Similarly, geothermal model outputs are multi-dimensional. There are a limited number of ML methods that can process multi-dimensional datasets, and our novel nonnegative tensor factorization method coupled with k-means clustering (NTFk) is at the forefront. The factorization process converts a data tensor (labeled as ***X*** in Fig.2.1.2) into a smaller core tensor (labeled as ***G*** in Fig.2.1.2) and three matrix factors for each dimension (labeled as ***H***, ***W***, and ***V*** in Fig.2.1.2). Matrix factors represent signatures in different dimensions. In the example presented in Fig.2.1.2, ***H***, ***W*** and ***V*** contain 5, 4, and 3 signatures, respectively. ***G*** defines how these signatures occurring in different dimensions are mixed to reproduce the original data tensor. The tensor factorization is again achieved through solving a minimization problem[15,37]. The estimation of the optimal number of signatures in each dimension is performed using customized *k*-means clustering[15,37].

2.5 Tests and Examples

All our codes and scripts are subject to unit testing and validation. The testing and validation of our codes are performed automatically using existing tools developed in Julia and automated GitHub workflows. The documentation of the functions in our codes is also automatically generated. All the information about the tests and examples (including Jupyter and Pluto notebooks) demonstrating our ML methods are available (GitHub: https://github.com/SmartTensors; GDR: https://gdr.openei.org/submissions/1297).

3 GeoThermalCloud Analyses

Here, we demonstrate the application of ***GeoThermalCloud*** to a series of problems involving synthetic, site, and regional datasets. This report focuses on (1) Southwest New Mexico (SWNM), (2) Great Basin, and (3) Brady site analyses. This work is discussed in this section. However, ML analysis also performed under this project include interpretations of the following datasets:

- **West Texas**: Subdivided the region into three areas; the western portion has higher geothermal potential at a lower depth than the middle and eastern portions.
- **Tularosa Basin, New Mexico**: Analyzed 21 Play Fairway Analysis (PFA) attributes at 120 locations. Data comes from past PFA work in this region[8]. ML analyses identified geothermal signatures

associated with low-, medium-, and high-temperature hydrothermal systems. Dominant attributes and spatial distribution of the geothermal signatures are also defined[7].

- **Tohatchi Springs**: ML analyses applied on 19 attributes at 42 locations in the study area. Successfully defined geothermal signatures associated with low- and medium-temperature hydrothermal systems, their dominant attributes, and spatial distribution[9].
- **Hawaii**: Analyzed four islands' data separately and identified low-, medium-, and high-temperature hydrothermal systems and their dominant characterization attributes[10].
- **Utah FORGE**: Prospectivity analysis has been performed to identify future drilling locations using geological, geochemical, and geophysical attributes[38]. Maps of temperature at depth and heat flow constructed based on the available data. Processed data includes satellite (InSAR), geophysical (gravity, seismic), geochemical, and geothermal attributes. Prospectivity maps generated and drilling locations proposed for future geothermal field exploration.
- **EGS Collab:** Field experiment data processed to extract dominant temporal patterns observed in 49 data streams; erroneous measurement attributes and periods automatically identified; interrelated data streams automatically identified.

These analyses and obtained results are available in our project presentation. They will also be discussed in upcoming conference presentations.

This report section also presents ML analyses of synthetic datasets relevant to (1) UtahForge and (2) SWNM sites. The first dataset was generated utilizing ***GeoDT,*** *a multi-physics* code for rapid prediction of the performance of geothermal energy systems, dependent on design decisions and site data uncertainty. The ML tools from our ***GeoThermalCloud*** were applied to identify key controlling attributes, separate the relative impact of different physical processes on production, and associate these simulated impacts to the model inputs. For the SWNM study, we executed PFLOTRAN simulations to generate a large synthetic dataset designed to predict the geothermal reservoir attributes such as temperature and heat flow based on observables near the ground surface, such as tracer element concentrations. This dataset is currently applied to perform ML analysis to develop an ML model, which will perform fast prediction and uncertainty quantification of underground geothermal conditions.

3.1 Southwest New Mexico

A geothermal dataset of Southwest New Mexico (SWNM) is processed using ***GeoThermalCloud***. SWNM is broadly divided into four physiographic provinces: the Colorado Plateau, the Mogollon-Datil Volcanic Field (MDVF), the Basin and Range, and the Rio Grande Rift[39,40,97]. Each physiographic province is associated with different types of unique hydrothermal systems with temperatures ranging from low (<90℃) to medium (90-150℃)[1,2,40,41]. Some of the SWNM systems are already utilized for commercial and recreational purposes. At 23 locations, energy-extraction facilities are providing both electricity and direct-use heating. For example, the Basin and Range province has one geothermal power plant (Lightning dock) of gross ~14 MWe power, five greenhouse farms[42], and numerous medium temperature wells and springs. There are 14 spas and recreational facilities utilizing the SWNM geothermal resources[42]. Recent Play Fairway Analysis (PFA) study of SWNM revealed more potential geothermal resources[8,41,43]. The dataset analyzed in this report comes from this past PFA work. It includes two geochemical, two geophysical, five geological, four hydrogeological, and four geothermal attributes (total 18) at 44 locations in SWNM (Fig.3.1.1). One of the attributes is the reservoir temperature estimates based on *silica geothermometry*[39,41]. *Boron*[39,41] and *lithium*[39,41] are tracer elements in thermal water. *Drainage density*[39,41], *spring density*[44], *hydraulic gradient*[39,41], and *precipitation*[39,41] are hydrogeological attributes. *Gravity anomaly*[39,41], *magnetic intensity*[39,41], and *seismicity*[39,41] are geophysical attributes. *Silica geothermometer*[29] temperature and *heat flow*[39,41] are geothermal attributes. *Silica geothermometer* indicates potential reservoir temperature, and *heat flow* is a proxy of the thermal gradient of the reservoir. Geological attributes include *crustal thickness*[45], *depth to the basement*[46], *fault intersection density*[40], *quaternary fault density*[40], *state map fault density*[47], *volcanic dike density*[44], and *volcanic vent density*[48]. The data are preprocessed before the ML analyses. The *boron* and *lithium* concentration values are

log-transformed. The attribute values at each location are shown in Table 3.3.1. Next, all attributes are rescaled within the range of 0.0 to 1.0 using unit range transformation. To apply NMF*k*, we create an 18×44 matrix ($X_{m\times n}$) where the m = 18 rows represent attributes, and the n = 44 columns are the values of the measured locations.

It is important to note that on purpose, the analyses did not include labeling of the hydrothermal systems based on their type and their association with physiographic provinces. Our ML analyses were also performed without providing information about the coordinates of the analyzed locations. In this way, we tested if the algorithm can blindly group the locations based on their type and province association based only on the observed geothermal attributes.

The NMFk analysis revealed 5 hidden geothermal signatures. It also estimated the elements of ***W*** (attributes) and ***H*** (locations) matrices (Fig. 3.1.2 and 3.1.3). Based on the ***H*** matrix estimates, Fig.3.1.4 shows the predominant association of the 44 locations with the extracted 5 geothermal signatures (labeled ***A***, ***B***, ***C***, ***D,*** and ***E***). The associations are estimated automatically by our ML tool; however, these associations are not one-to-one for all the locations; some of the locations are associated with more than one signature (Fig.3.1.2). Nevertheless, our ML analysis identified the association of the 44 locations with the 4 physiographic provinces blindly (Fig.3.1.4). The reproduction of the provinces is not perfect and captures the complexity of the geothermal conditions in the study area. ***Signatures A*** and ***E*** cover MDVF. ***Signature A*** encompasses the area below the Jemez lineament (we call it the southern MDVF). ***Signature E*** covers the Jemez lineament and its contiguous north area (we call it the northern MDVF). ***Signatures B***, ***C***, and ***D*** capture the Basin and Range, the Colorado Plateau, and the Rio Grande Rift provinces, respectively.

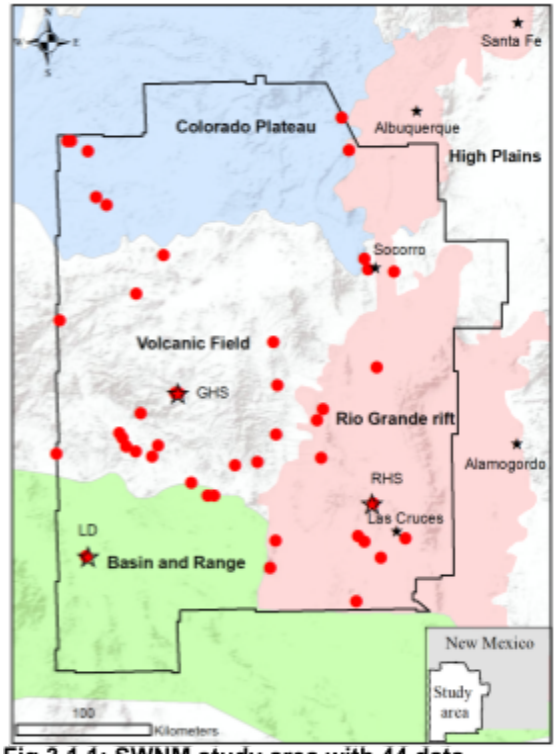


Fig.3.1.1: SWNM study area with 44 data locations. open stars are mid-temperature hydrothermal systems. GHS, LD, and RHS stand for Gila hot springs, lightning dock geothermal plant, and Radium hot springs, respectively.

Based on the ***W*** matrix estimates (Fig.3.1.3), our ML algorithm categorizes signatures associated with low- or medium-temperature hydrothermal systems based on the contribution of *silica geothermometer* in each signature: low and high silica values define low- and medium-temperature systems, respectively. Below, we describe how geology, hydrogeology, and geothermal attributes relate to each signature and how the extracted geothermal signatures define the hydrothermal systems within SWNM based on the ***H*** matrix estimates in Fig.3.1.3.

Signature A is representative of low-temperature hydrothermal systems because of the low contribution of the *silica geothermometer*. This signature's dominant attributes are Li^+ *concentration, drainage density*, and *magnetic intensity* (Table 3.1.2); *gravity anomaly* and *volcanic dike density* are also essential attributes (Fig.3.1.3). *Magnetic intensity, volcanic dike density*, and *gravity anomalies* indicate the manifestation of plutonic mafic rocks due to Tertiary volcanic events[49]. The locations associated with Signature A are in the southern MDVF. This portion of the MDVF has a history of active volcanism in the past[50–53] that might have further enhanced *Volcanic dike density* and secondary mineralization. The resultant secondary mineralization is expected to elevate *gravity anomaly* and *magnetic intensity* in this region. The dominant attributes except *drainage density* are indicators that the hydrothermal systems in this region are prospective geothermal resources. Yet, NMF*k* did not diagnose the Signature A locations as medium-temperature hydrothermal systems because of the low *silica geothermometer* contribution to

this geothermal signature. A possible explanation is a lack of high-temperature gradient in this area due to increased infiltration caused by high *drainage density*. It is also important to note that elevated Li^+ *concentration* here is not correlated with high B^+ *concentrations*. This suggests different subsurface mechanisms are causing elevated Li^+ / B^+ *concentrations* in geothermal systems.

Signature B represents medium-temperature hydrothermal systems because of the high contribution of the *silica geothermometer*. The other dominant attributes of this signature are *gravity anomaly* and *depth to the basement* (Table 3.1.2). Also substantial are B^+ and Li^+ *concentrations*, *magnetic density*, *quaternary fault density*, and *heat flow* (Fig.3.1.3). *Heat flow* and *depth to the basement* are unique dominant attributes of this signature. *Heat flow* is also an indicator of a temperature gradient, while *depth to the basement* is an indicator of a high reservoir depth and a large distance of the heat source distance from the ground surface. 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 went through frequent Tertiary and some Quaternary volcanic events[49]. Therefore, it is not surprising that *magnetic intensity*, *gravity anomaly*, and *volcanic dike density* are dominant attributes. This area also has a low crustal thickness[49,54,55], which indicates that this area is also closer to the mantle heat source. *Depth to the basement* is the deepest in the study area that may assist in entrapping the heat originating from the mantle. The high-temperature gradient, deep basement, and lower-crustal thickness may cause the medium-temperature hydrothermal systems in this region. Further field explorations and data collection activities are required to define better the locations associated with this geothermal signature and associated hydrothermal resources.

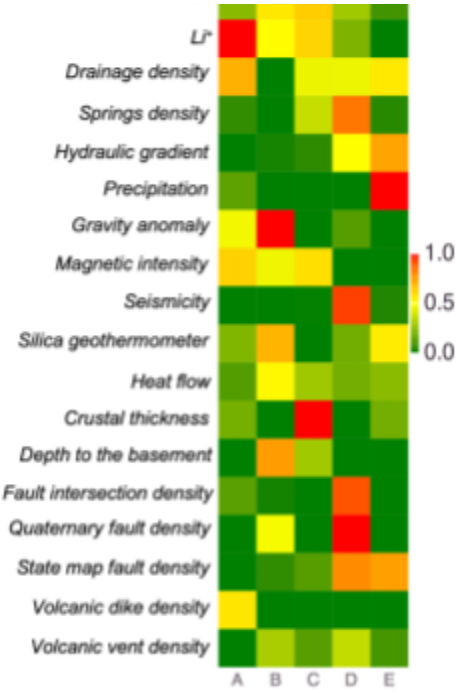


Fig.3.1.3: Estimated attribute matrix W. High (red) and low (green) matrix values define high and low significance, respectively, of attributes (rows) to inform extracted geothermal signatures (columns). Based on *silica geothermometer* values, Signatures B and E represent mid-temperature systems, and Signatures A, C and D represent low-temperature systems.

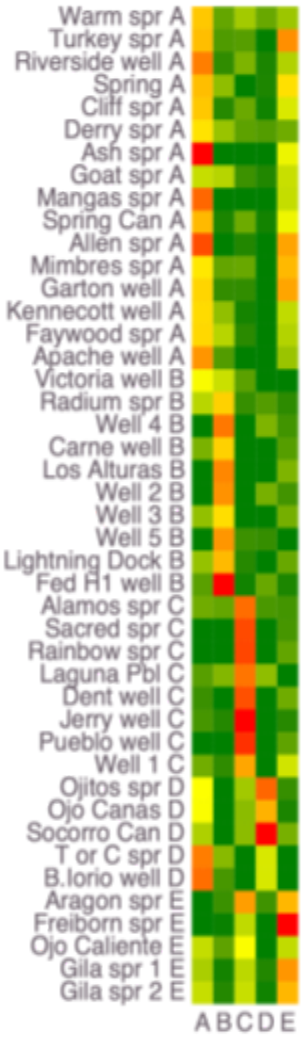


Fig.3.1.2: Estimated attribute matrix H. High (red) and low (green) matrix values define high and low significance, respectively, of locations (rows) to inform extracted geothermal signatures (columns).

Signature C represents low-temperature hydrothermal systems because of the low contribution of the *silica geothermometer*. The dominant attributes of this signature are *crustal thickness*, *magnetic intensity*, and B^+ and Li^+ *concentrations* (Table 3.1.2); *drainage density* is essential as well (Fig.3.1.3). B^+ and Li^+ here are collocated (in contrast with ***Signature A*** above) and may be released from the subsurface due to the nearby heat source, while *magnetic intensity* may indicate secondary mineralization due to Tertiary volcanic events, producing plutonic mafic rocks[56–58]. These three attributes suggest potential heat sources at depth. *Drainage density* and *crustal thickness* are a unique combination of attributes for this signature, indicating that it

might represent lateral hydraulics. The locations associated with Signature C are within the Colorado Plateau. However, the high significance of B^+ and Li^+ *concentrations*, *magnetic intensity*, *drainage density* in this signature are good indicators of geothermal resources. NMF*k* did not designate this signature as medium-temperature hydrothermal systems due to the low *silica geothermometer* impact on this geothermal signature. The large crustal thickness may preclude heat flow in this region that might be a potential explanation for the designation as low-temperature hydrothermal systems.

Signature D represents low-temperature hydrothermal systems because of the low contribution of the *silica geothermometer*. The dominant attributes of this signature are *quaternary fault intersection density, fault intersection density*, *seismicity*, *state map fault density*, and *spring density* (Table 3.1.2); *drainage density* and *hydraulic gradient* are also critical (Fig.3.1.3). *Quaternary fault intersection density, fault intersection density*, *seismicity*, and *state map fault density* suggest that this signature represents tectonic features. The locations associated with this Signature D are in the Rio Grande Rift and the Jemez lineament, which went through both extension and subduction tectonic events[49,55]. Active tectonic events increase *fault intersection density*, which imporves *drainage density*, *fault intersection density*, *spring density*, and *hydraulic gradient*. Also, *seismicity* indicates the presence of active faults. This signature did not get high contribution from attributes, which are good indicators of medium-temperature hydrothermal systems such as B^+, Li^+, *gravity anomaly*, *magnetic intensity*, *heat flow*, and *silica geothermometer*. The dominant attributes indicate that the locations associated with this signature have good groundwater flow characteristics. These locations are prospective hydrothermal systems for further exploration.

Table 3.1.1. SWNM geothermal dataset. Values along each column are color coded: minimum (green), intermediate (yellow) and maximum (red) values.

Location	Boron	Gravity	Magnet	Dikes	Drain	Fault	Qfault	Seism	NMFit	Springs	Vents	Lithium	Precip	Silica	Δh	Qheat	Crust	Bsmt
Alamos Spring	-0.2	-203.3	136.2	0.431	7.4	0.000	0.00	0.004	16.2	0.010	0.003	-3.1	264.8	16.5	5.6	4.6	38.7	1439
Allen Springs	-3.2	-189.3	184.6	3.625	17.3	0.000	0.01	0.002	15.6	0.003	0.001	-4.0	514.5	24.0	13.9	4.4	32.5	51
Apache Tejo Warm Springs well	-1.8	-181.2	15.0	3.807	17.3	0.001	0.03	0.001	0.7	0.003	0.000	-8.6	326.3	52.0	4.7	4.6	30.7	24
Aragon Springs	1.5	-229.1	-317.7	0.010	19.0	0.000	0.00	0.000	41.1	0.005	0.003	-7.5	387.0	56.5	4.0	4.5	38.8	1486
Ash Spring	-2.7	-193.2	66.6	4.914	17.0	0.000	0.00	0.002	9.3	0.003	0.000	-5.0	492.0	29.3	4.1	4.4	32.2	-92
B. Iorio 1 well	-2.1	-196.5	-48.2	1.936	18.8	0.057	21.02	0.000	9.1	0.003	0.003	-2.6	260.4	59.4	0.9	4.0	30.9	-188
Cliff Warm Spring	-2.5	-199.1	-47.1	1.290	22.8	0.001	2.58	0.002	11.0	0.002	0.001	-6.9	364.2	64.2	1.8	4.2	33.1	-191
Dent windmill well	-2.1	-230.8	89.3	0.000	13.4	0.000	0.00	0.000	0.0	0.005	0.000	-7.3	341.7	19.7	2.4	4.7	43.5	865
Derry Warm Springs	-1.5	-161.6	197.0	0.659	18.3	0.007	9.16	0.000	15.9	0.002	0.000	-7.5	276.1	37.4	3.0	4.6	30.0	-120
Faywood Hot Springs	-2.6	-172.1	-49.8	0.939	16.6	0.002	2.81	0.000	1.9	0.003	0.000	-4.8	346.4	67.2	4.2	5.5	30.0	619
Federal H 1 well	-0.4	-132.0	35.0	0.000	5.8	0.004	20.31	0.001	7.2	0.000	0.015	-5.0	253.8	78.7	2.7	4.9	27.3	2906
Freiborn Canyon Spring	-2.5	-225.0	-242.0	0.401	13.1	0.000	0.00	0.001	19.8	0.001	0.004	-12.6	538.6	49.8	13.0	4.6	38.4	1138
Garton well	-3.2	-196.8	35.6	0.150	18.0	0.000	0.00	0.000	28.9	0.002	0.001	-5.0	489.9	70.0	4.3	3.9	30.9	-266
Gila Hot Springs 1	-1.9	-221.6	-149.3	0.127	24.2	0.000	0.00	0.001	25.5	0.003	0.003	-7.8	422.6	69.9	6.6	4.4	34.0	413
Gila Hot Springs 2	-1.8	-222.9	-138.8	0.112	24.7	0.000	0.00	0.001	23.7	0.003	0.003	-6.7	425.9	70.8	3.2	4.6	33.9	519
Goat Camp Spring	-2.1	-159.2	-29.7	0.751	10.0	0.001	2.22	0.007	10.6	0.002	0.001	-8.0	344.0	68.9	5.8	4.4	32.4	19
Jerry well	-0.8	-219.6	172.4	0.111	15.5	0.000	0.00	0.000	6.3	0.004	0.005	-7.9	243.9	13.4	1.0	4.4	42.3	1190
Kennecott Warm Springs well	-2.4	-178.3	-69.9	1.422	17.8	0.002	1.76	0.000	1.1	0.003	0.000	-6.9	355.0	66.1	4.3	5.0	30.0	409
Laguna Pueblo	0.4	-204.2	62.5	0.406	8.6	0.004	4.58	0.006	14.6	0.018	0.005	-3.3	259.7	42.9	2.6	4.4	37.2	1506
Lightning Dock	-1.0	-168.0	-168.1	0.086	4.6	0.008	8.40	0.002	4.3	0.000	0.000	-3.9	291.5	107.3	0.8	5.0	29.8	1800
Los Alturas Estates	-1.5	-141.4	-127.5	0.004	7.6	0.003	0.05	0.002	6.6	0.001	0.000	-12.7	265.3	71.9	2.2	6.3	27.4	4321
Mangas Springs	-2.6	-201.0	-227.1	3.503	20.2	0.000	0.91	0.002	11.5	0.002	0.000	-4.5	393.5	53.6	0.3	4.2	32.4	-178
Mimbres Hot Springs	-2.3	-200.6	43.4	0.670	15.4	0.002	1.13	0.000	19.0	0.004	0.000	-3.8	445.9	68.3	9.1	4.9	31.0	50
Ojitos Springs	-1.6	-202.1	-7.5	1.342	19.6	0.044	19.74	0.037	31.0	0.020	0.005	-4.5	257.5	57.6	7.2	4.5	33.0	-255
Ojo Caliente	-2.6	-226.5	-168.4	0.000	20.5	0.000	0.00	0.000	8.3	0.004	0.000	-2.9	333.6	48.4	3.5	5.5	33.8	2415
Ojo De las Canas	-1.7	-188.5	-85.8	0.839	22.3	0.036	12.55	0.036	28.0	0.013	0.003	-6.0	270.5	14.2	4.0	4.5	31.8	101
Pueblo windmill well	-1.2	-228.8	315.9	0.029	15.2	0.000	0.00	0.000	6.1	0.004	0.003	-12.0	265.8	18.3	2.9	4.3	42.5	1027
Radium Hot Springs	-0.8	-151.4	-7.8	0.010	8.8	0.013	11.40	0.003	10.6	0.001	0.000	-5.3	264.2	63.6	0.3	5.4	28.2	1191
Rainbow Spring	-1.7	-227.1	-48.5	0.000	11.0	0.000	0.00	0.001	0.0	0.006	0.000	-7.0	307.8	21.7	3.3	4.7	43.9	755
Riverside Store well	-1.3	-196.1	-102.9	1.562	22.6	0.000	2.50	0.002	11.7	0.002	0.001	-2.4	356.1	60.8	0.9	4.3	32.9	-165
Sacred Spring	-1.8	-228.4	-80.4	0.000	10.9	0.000	0.00	0.001	0.0	0.006	0.000	-7.0	298.4	21.2	1.3	4.6	43.9	742
Socorro Canyon	-1.8	-204.7	-136.5	1.203	21.1	0.051	28.88	0.034	33.8	0.020	0.005	-6.7	284.1	44.6	11.1	5.0	32.6	-229
Spring	-4.1	-183.5	334.5	0.218	20.1	0.011	1.81	0.000	20.1	0.001	0.006	-6.8	361.9	117.2	5.1	3.8	31.5	-104
Spring Canyon Warm Spring	-2.1	-194.2	117.3	2.293	21.9	0.000	1.50	0.002	12.7	0.002	0.000	-8.3	361.7	51.6	5.8	4.2	32.6	-57
Truth or Consequences spring	-1.1	-168.2	-54.3	2.175	18.4	0.064	20.51	0.000	10.3	0.003	0.002	-3.3	265.9	55.3	0.6	4.3	31.0	304
Turkey Creek Spring	-3.2	-196.4	54.8	0.984	19.2	0.001	3.69	0.002	28.1	0.002	0.002	-3.7	493.4	81.3	5.8	4.4	33.6	56
Victoria Land and Cattle Co. well	-1.8	-165.9	-65.4	0.478	6.4	0.003	0.06	0.001	0.9	0.001	0.000	-2.9	253.0	43.0	1.9	4.1	30.7	2014
Warm Springs	-2.1	-193.3	113.5	0.220	19.0	0.029	2.63	0.000	16.5	0.004	0.003	-2.5	314.6	56.0	5.4	4.3	32.7	1252
Well 1	-1.4	-230.7	-31.3	1.190	15.7	0.000	0.75	0.001	22.1	0.004	0.002	-6.6	345.4	49.0	1.7	4.4	40.0	1961
Well 2	-1.2	-162.5	0.8	0.000	4.5	0.008	24.24	0.003	11.8	0.000	0.006	-10.1	279.5	70.5	1.7	4.8	27.8	2993
Well 3	-2.5	-140.0	31.7	0.839	2.1	0.001	2.11	0.001	5.0	0.001	0.000	-7.3	369.0	51.0	4.1	4.3	28.0	3073
Well 4	-1.3	-161.7	-56.1	0.000	3.4	0.008	28.49	0.003	10.6	0.000	0.006	-10.0	274.3	94.0	1.9	4.7	27.7	3373
Well 5	-1.9	-167.2	-29.9	0.000	2.5	0.008	15.48	0.002	3.1	0.000	0.005	-6.8	243.8	47.0	0.3	4.0	27.4	5460
Well south of Carne	-2.4	-156.7	-129.6	0.457	4.3	0.000	2.11	0.002	6.0	0.001	0.000	-6.8	269.7	87.1	1.4	4.5	28.4	2761

Signature E represents medium-temperature hydrothermal systems because of the high contribution of the *silica geothermometer*. The remaining dominant attributes of this signature are *precipitation*, *hydraulic gradient,* and *state map fault density* (Table 3.1.2). *State map fault density* and *hydraulic gradient* represent deep flow circulation, meaning that this signature may capture vertical groundwater flow characteristics. The locations associated with this signature are in or around the Jemez lineament within the northern part of the MDVF. The Jemez lineament went through rigorous Tertiary and Quaternary volcanism events[40,50–53]. Also, this region has an intensive tectonic history when compared to the surrounding areas. Both high volcanism and tectonic events make this area more amenable to have geothermal resources. Besides, this area has high precipitation, which increases deep flow circulation. A fluid circulation from the depth where hot water exists is a sound indicator of good hydrothermal systems. The combination of high-reservoir temperature (*silica geothermometer*) and deep fluid circulation characteristics make the northern MDVF favorable to medium-temperature hydrothermal systems. Further field explorations and data collection activities are required to define better the locations associated with this geothermal signature and associated hydrothermal as geothermal resources.

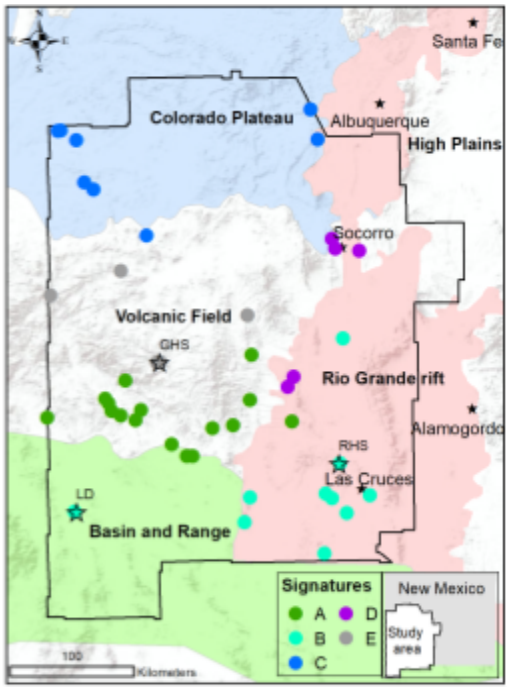


Fig.3.1.4: Sampling locations associated with the extracted 5 geothermal signatures.

Table 3.1.2: Geothermal hidden signatures and their association with a geothermal resource type, dominant data attributes, physical significance, and physiographic provinces. Dominant attributes are color coded as high and mid important based on the estimated attribute matrix weights (Fig,3.1.2).

Sig	Hydro-thermal system type	Dominant attributes	Physical signifi-cance	Province
A	Low temperature	*Li^+ concentration* *Drainage density* *Magnetic intensity*	Shallow heat flow	Southern MDVF
B	Medium temperature	*Gravity anomaly* *Depth to the basement* *Silica geothermometer*	Deep heat flow	Rio Grande rift
C	Low temperature	*Crustal thickness* *Magnetic intensity* *B^+ / Li^+ concentrations*	Thick crust	Colorado Plateau
D	Low temperature	*Quaternary fault density* *Fault intersection density* *Seismicity* *State map fault density* *Spring density*	Tectonics	Central Rio Grande rift
E	Medium temperature	*Precipitation* *Hydraulic gradient* *State map fault density*	Vertical hydraulics	Northern MDVF

Based on ML results, among the 18 analyzed attributes, the 12 dominant attributes related to medium-temperature hydrothermal systems are B^+ and Li^+ *concentrations*, *silica geothermometer*, *heat flow*, *gravity anomaly*, *magnetic intensity*, *quaternary fault density*, *state map fault density*, *depth to the basement*, *drainage density*, *precipitation*, and *hydraulic gradient*. All of these attributes are related to geothermal signatures B and E (Table 3). The remaining six attributes are *volcanic dike density*, *volcanic vent density*, *fault intersection density*, *springs density*, *crustal thickness*, and *seismicity*. These attributes are dominant in Signatures A, C, and D (Table 3.1.2), representing low-temperature hydrothermal systems.

Our ML analyses designated two unique geothermal signatures, B and E, associated with medium-temperature hydrothermal systems. A biplot is generated to show the correlations between these two signatures (Fig.3.1.5). Here, the signatures are viewed as basis vectors similarly to how eigenvectors under PCA analyses are defined[32]. A biplot is an exploratory scatterplot showing the mutual relation between two signatures based on how attributes and locations associated with these signatures are weighted. In a biplot, an attribute is well correlated if its significance is high for both signatures. In contrast, some attributes might be important for one signature but not for another signature. A biplot is also a good indicator if an attribute is not critical for both signatures; these attributes will be located close to the plot origin. In a biplot, well-correlated attributes lie on the diagonal between the two axes and away from the origin, while uncorrelated attributes lie close to the axes. The biplot in Fig. 3.1.5 demonstrates the correlation of attributes between the geothermal Signatures B and E. The only correlated attribute is the *silica geothermometer*, which classifies these signatures as medium-temperature hydrothermal systems. This lack of correlation among the other geothermal attributes reveals that they are uniquely associated with these two signatures. Because the geology of each province is unique, their controls on hydrothermal systems also vary. Signature B falls mainly in the southern Rio Grande Rift zone. The hydrothermal system of this area is primarily defined by *gravity anomaly*, *depth to the basement,* B^+ and Li^+ *concentrations, and heat flow*. On the other hand, Signature E falls in the northern MDVF, which is in the north portion of the study area. Hydrothermal systems in this area are defined by *precipitation*, *hydraulic gradient, state map fault density*, *state map fault density*, and *drainage density*.

Signatures B and E are associated with the northern MDVF and Rio Grande Rift provinces, respectively. Both areas went through Tertiary and Quaternary volcanisms, but the northern MDVF went through more frequent volcanic events than the Rio Grande Rift zone. Also, the northern MDVF is tectonically more active than the Rio Grande Rift zone. However, a tectonic extension feature is present in the western and easter portions of the Rio Grande Rift zone, but it is absent in the northern MDVF. Moreover, the Rio Grande Rift zone has a lower crustal thickness than the northern MDVF. All these observations demonstrate the unique geological and hydrological characteristics of the two regions. Therefore, they represent unique hydrothermal systems, and these differences were successfully captured and extracted by our NMF*k* analyses. This demonstrates that the ML algorithm can learn the geothermal characteristics of the analyzed dataset just based on provided data describing the geothermal conditions in the SWNM region.

Fig.3.1.5: Biplot of correlations between medium-temperature resource attributes as defined by extracted geothermal signatures B and E. Red dots show correlated attributes between Signatures A and B. Green color dots show poor correlation; blue dots have no correlation but are dominant in Signature B, and orange dots also show no correlation but are dominant in Signature E.

In conclusion, our ML analyses characterized a geothermal dataset of SWNM to (1) identify hidden geothermal signatures, (2) estimate the optimal number of signatures, (3) find dominant attributes associated with each signature, (4) map spatial association of the signatures. The locations represent hydrothermal systems within the four physiographic provinces present within the SWNM study domain. On purpose, the analyses did not include the labeling of the hydrothermal systems based on their type and their association with physiographic provinces. Our ML analyses were also performed without providing information about the coordinates of the analyzed locations. In this way, we tested if the algorithm can blindly group the locations

based on their type and province association based only on observed geothermal attributes.

Out of five extracted geothermal signatures, two (B and E) are associated with medium-temperature hydrothermal systems. The zones associated with these signatures (northern MDVF and the Rio Grande Rift) required further exploration to designate them as geothermal resources. The PFA work[41] generated a preliminary geothermal prospectivity map. In Phase 2 of our project (if funded), we will combine these PFA results and the knowledge accumulated in this study to make an ML-enhanced geothermal prospectivity map of the SWNM region. This map will assist in discovering hidden resources and their accurate locations for geothermal heat extraction. All the data and codes, including Jupyter and Pluto notebooks, required to reproduce these results are available at the ***GeoThermalCloud*** GitHub and GDR repositories (https://github.com/SmartTensors/GeoThermalCloud.jl/tree/master/SWNM).

3.2 Great Basin

Under this project, we analyzed geochemistry data of the Great Basin (Fig. 3.2.1). The Great Basin covers Nevada, much of its neighboring states Oregon, Utah, California, Idaho, and Wyoming. It has multiple geothermal reservoirs ranging from low- to high-temperature resources, and a vast area is yet to be explored to discover hidden geothermal resources. Plenty of data have been collected over several decade to characterize the regional geothermal resources. Here, we process public data available at the Nevada Bureau of Mines and Geology website [http://www.nbmg.unr.edu/Geothermal/GeochemDatabase.html]. The size of the data for this study is 14341 x 18; at 14341 locations, 17 geochemical attributes (water cations/anions) and groundwater temperature are observed[59,60]. The 18 attributes are pH, total dissolved solids (TDS), Al^{3+}, B^{+}, Ba^{2+}, Be^{2+}, Br^{-}, Ca^{2+}, Cl^{-}, HCO_3^{-}, K^{+}, Li^{+}, Mg^{2+}, Na^{+}, $\partial^{18}O$, groundwater temperature, quartz geothermometer, and chalcedony geothermometer. pH represents alkalinity of water, TDS is the total amount of major and tracer cations/anions, Ca^{2+}, K^{2+}, Mg^{2+}, Na^{+} are major cations, HCO_3 and Cl are major anions, Al^{3+}, B^{+}, Ba^{2+}, Be^{2+}, Br^{-}, are Li^{+} trace elements, and $\partial^{18}O$ is an oxygen isotope. Major anions/cations define the ionic type of water. The $\partial^{18}O$ describes the origin (e.g., meteoric, magmatic, connate) of the water. Groundwater temperature indicates the water temperature at a shallow depth rather than at the actual geothermal reservoir depth. Quartz and chalcedony geothermometers indicate potential reservoir temperature. Table 3.2.1 lists the minimum, maximum, mean, and missing values/sparsity in the data. The minimum and maximum values demonstrate that the dataset attributes vary over a wide range. The missing data column in the table indicates that the dataset is heavily sparsed. Here, we applied the ***GeoThermalCloud*** ML methods to analyze this sparse geothermal/geochemical data and better understand/predict the spatial distribution of the available geothermal resources.

The dataset described above was used to perform NMFk analyses. Before the ML runs, the dataset was log-transformed and normalized between 0 to 1. ML analysis was performed for k= 2, 3, …, 15 number of signatures. The ML algorithm selected the k=3 solution to represent the optimal number of hidden geothermal signatures for the Great Basin dataset. The k > 3 solutions overfitted the problem. Fig. 3.2.2(a) demonstrates the attribute matrix of the optimal NMFk solution; the attribute matrix depicts the importance of attributes to represent extracted signatures. Next, we define types of hydrothermal systems based on the contribution of groundwater temperature in the extracted 3 signatures. Based on this assumption, ***Signatures A, B, and C*** define low-, high-, and medium-temperature hydrothermal systems,

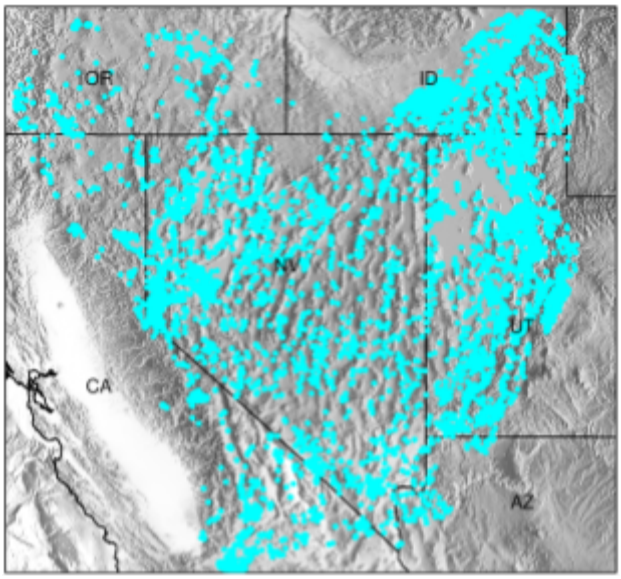


Fig 3.2.1: Geothermal data locations in the Great Basin.

respectively. ***Signature A*** represents low-temperature hydrothermal systems because of the low contribution of groundwater temperature in this signature. The dominant attributes of this signature are TDS, Br^{+}, B^{+}, and $\partial^{18}O$. ***Signature B*** represents high-temperature hydrothermal systems due to the high contribution of temperature in this signature. The dominant attributes of the signature are pH, Al^{3+}, Be^{2+}, as well as quartz and chalcedony geothermometers. ***Signature C*** defines medium-temperature hydrothermal systems because of the medium contribution of temperature. The dominant attributes of the signature are Mg^{2+} and Ca^{2+}.

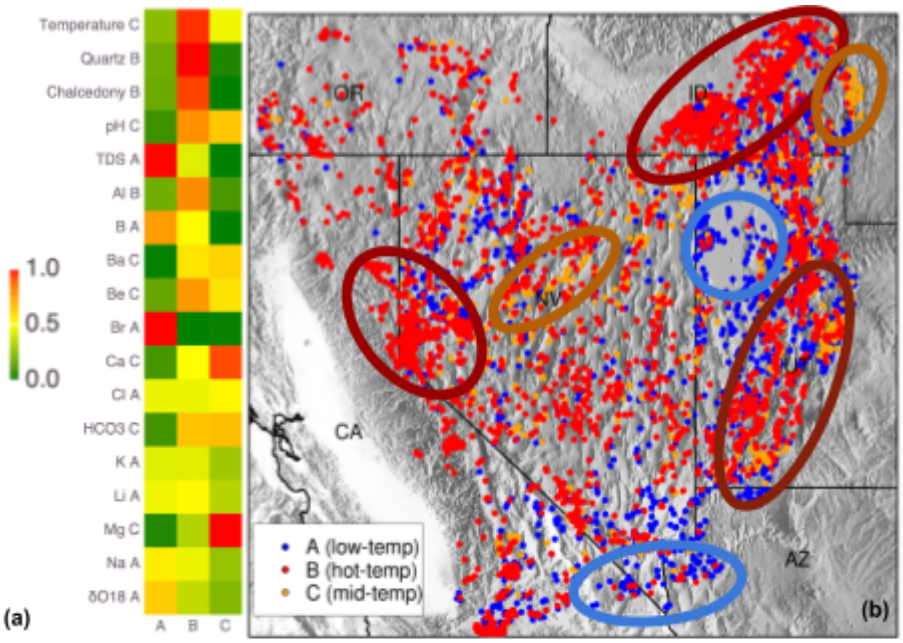


Fig 3.2.2: Optimal hidden geothermal signatures (a) and their spatial distribution (b); ellipses mark areas with high-density of signature locations of similar type.

The spatial distribution of each signature is shown in Fig.3.2.2(b), where blue, red, and orange colors represent low-, high-, and medium-temperature hydrothermal systems. The distribution of Signatures B and C suggests that the significant portions of the Great Basin region have prospective geothermal resources. Areas with a high density of B and C locations are labeled with ellipses in the figure. Some of these locations also align with existing geothermal resources and sites such as Dixie Valley and Brady geothermal areas in Nevada. Maps on the upper row of Fig.3.2.3 further demonstrate the spatial distribution of the extracted geothermal signatures.

Table 3.2.1: Great Basin dataset attributes / summary statistics.

Attribute	Minimum	Mean	Maximum	Missing (%)
Groundwater temperature (℃)	0.1	23.7	275	2.6
Quartz geothermometer (℃)	-50.8	81.0	273	39.1
Chalcedony geothermometer (℃)	-81.6	50.3	271	39.1
pH	1	7.5	11.7	35.0
TDS (PPM)	0	5770	329000	87.8
Al^{3+} (PPM)	0	7.3	6400	90.5
B^{+} (PPM)	0	3.1	590	61.7
Ba^{2+} (PPM)	0	0.1	27.4	82.4
Be^{2+} (PPM)	0	0	0.7	88.5
Br^{-} (PPM)	0	2.0	84	86.4
Ca^{2+} (PPM)	0	97.0	2570	33.6
Cl^{-} (PPM)	0	2870	240000	29.2
HCO_3^{-} (PPM)	0	278	37000	76.1
K^{+} (PPM)	0	101	13000	40.8
Li^{+} (PPM)	0	4.95	970	80.3
Mg^{2+} (PPM)	0	86.8	8500	34.8
Na^{+} (PPM)	0	1960	160000	38.2
$\delta^{18}O$ (‰)	-19.2	-14.6	7.8	89.7

The maps in Fig.3.2.2(b) and Fig.3.2.3(upper row) present the same ML results; however, they are shown in two different ways. Fig.3.2.2(b) shows how the locations are labeled as A, B, or C depending on the weights in the location matrix estimated by the ML analyses. The signature associations are determined automatically by our ML algorithm. Fig.3.2.3(upper row) shows the spatial interpolations of the weights in the location matrix for the 3

signatures separately. The spatial mapping is obtained by binning the locations shown in Fig.3.2.1 into a regular grid (with dimensions 108 x 128 in horizontal and vertical directions, respectively). The weights are also normalized between 0 and 1. The color scales in Fig.3.2.3(upper row) emphasize the range from 0 to 0.5; nevertheless, the actual values vary from 0 to 1. As a result, the maps show the probability of occurrence of the extracted geothermal signatures within the study domain. Therefore, the maps represent prospectivities associated with the extracted hidden geothermal signatures related to different geothermal reservoir types.

Using our ML tool, we can perform analyses on sparse datasets and make predictions for missing values. For example, B^+, $\partial^{18}O$, Br^+, and TDS are dominant attributes of ***Signature A***, and all of them are sparse (bottom row of Fig. 3.2.3). Yet, our ML methodology estimates a continuous spatial distribution for ***Signature A*** (top row of Fig. 3.2.3). Similarly, the dominant attributes of ***Signature B*** and ***C*** are also sparse (Fig. 3.2.3). Still, the ML algorithm reconstructs a continuous signature distribution over the study domain. This is possible because NMFk and NTFk can learn from only a partially represented object. This capability is generally absent in many traditional machine learning techniques, such as PCA, deep neural networks (convolutional or recurrent), etc.

As discussed above, all attributes in the Great Basin dataset have some level of sparsity (Table 3.2.1). For example, $\partial^{18}O$ has 90% sparsity (Table 3.2.1, Fig. 3.2.4(left)). After learning the mapping function among all attributes and generating the signature mappings (Fig. 3.2.3), our ML algorithm can estimate a continuous distribution of all the attributes, including $\partial^{18}O$ (Fig. 3.2.4). In this process, our ML method is superior to alternative statistical approaches such as kriging and co-kriging (i.e., Gaussian process modeling) for interpolation. The kriging-based methods require additional information to account for interrelationships among analyzed attributes (e.g., variograms and co-variograms). Our ML approach identifies the interrelationships among the attributes automatically based on the provided data. Both NMFk and NTFk can be applied to find mapping functions among all attributes, both in attribute and spatial domain. As a result, we constructed a continuous distribution of all attributes in the dataset. This continuous distribution of data can be further utilized for identifying geothermal resources either in the whole Great Basin or part of the Great Basin.

In addition to making predictions about the attribute values at the locations where data are missing, our ML methodology estimates uncertainties in these predictions. For example, the developed ML model is also applied to predict the temperature based on all other attributes. To test ML predictive capabilities, the data are split into training and prediction sets (Table 3.2.2). Furthermore, different levels of artificial noise were added to the training dataset (Table 3.2.2) to evaluate the ML sensitivity to measurement errors. The accuracy of the blind temperature predictions was assessed using a coefficient of determination (R^2) between actual and estimated values for a series of test problems (Table 3.2.2). The results listed in Table 3.2.2 demonstrate that accurate prediction ($R^2 > 0.9$) can be obtained even if we use only 50% of the data with <10% measurement errors. The above results also validate the applicability of our ML methods to predict geothermal conditions based on limited data.

Table 3.2.2: Accuracy of the blind temperature predictions evaluated by R^2 between true and estimated values for a series of test problems with (1) different percent of measurements applied for training, (2) different levels of measurement error added to the training dataset.

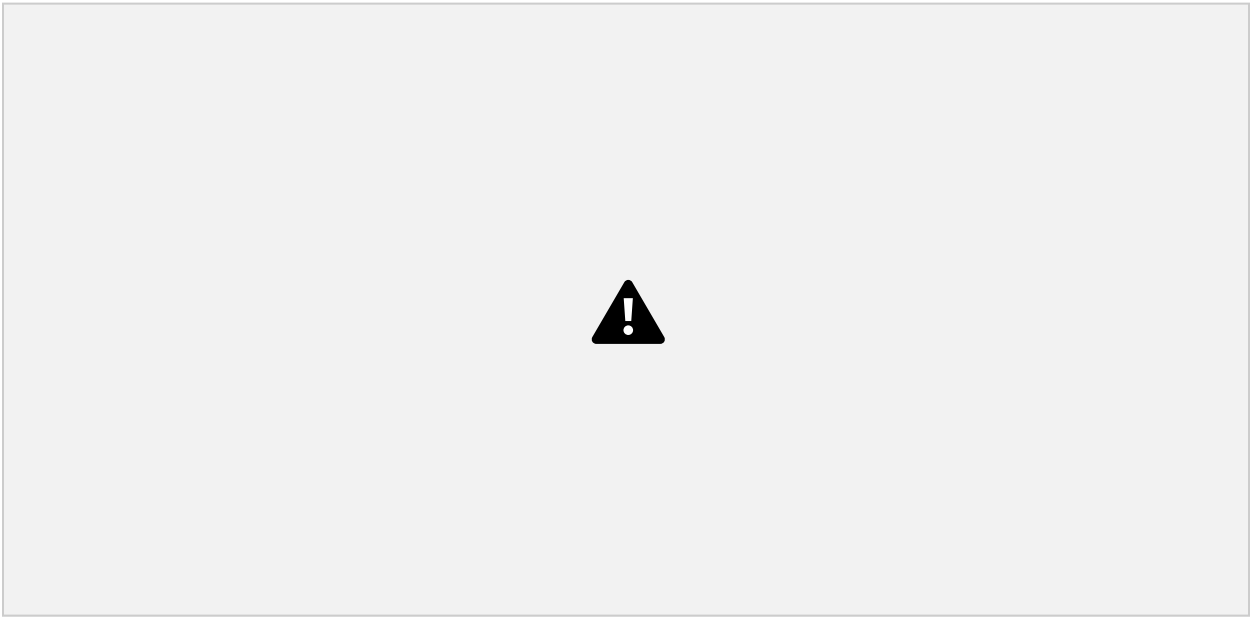
Training dataset	Measurement Error [%]			
	100%	50%	20%	10%
90%	0.675	0.823	0.939	0.976
80%	0.616	0.769	0.919	0.951
50%	0.574	0.749	0.870	0.917
20%	0.565	0.714	0.838	0.887
10%	0.441	0.623	0.755	0.876

In conclusion, the ML analyses identified hidden geothermal signatures associated with low-, medium-, high-temperature hydrothermal systems, their dominant characterization attributes, and spatial distribution within the study area. Also, we generated continuous maps of low-, medium-, and high-temperature hydrothermal systems that will assist in developing geothermal resources in

the Great Basin. Furthermore, we constructed continuous distribution from the sparse distribution of attributes that will help analyze other geological/geophysical/geothermal attributes with geochemical attributes.

It is important to note that there are already new geothermal and geochemical datasets that have been collected in the Great Basin area[59,60], which are not reflected here because they are not currently publicly available yet. Once these datasets become accessible, we can rerun our analyses to assimilate the new data. Although the ML processing of the Great Basin is computationally intensive, it does not take more than a week using available cutting-edge computing resources at LANL and on Google Cloud. Furthermore, the GeoDAWN dataset is currently collected in this region. As already demonstrated, our ML tools are computationally efficient and robust to process large datasets. Therefore, the assimilation of the GeoDAWN dataset together with other existing Great Basin datasets will not be methodologically or computationally a challenge for us; we are planning to conduct this work under Phase 2 of this project if we get selected to be funded. The new work will also compare how different data with different pedigree and quality will impact the extracted geothermal signatures and ultimately geothermal prospectivity of the Great Basin region.

All the data and codes, including Jupyter and Pluto notebooks, required to reproduce these results are available at the ***GeoThermalCloud*** GitHub and GDR repositories (https://github.com/SmartTensors/GeoThermalCloud.jl/tree/master/GreatBasin).



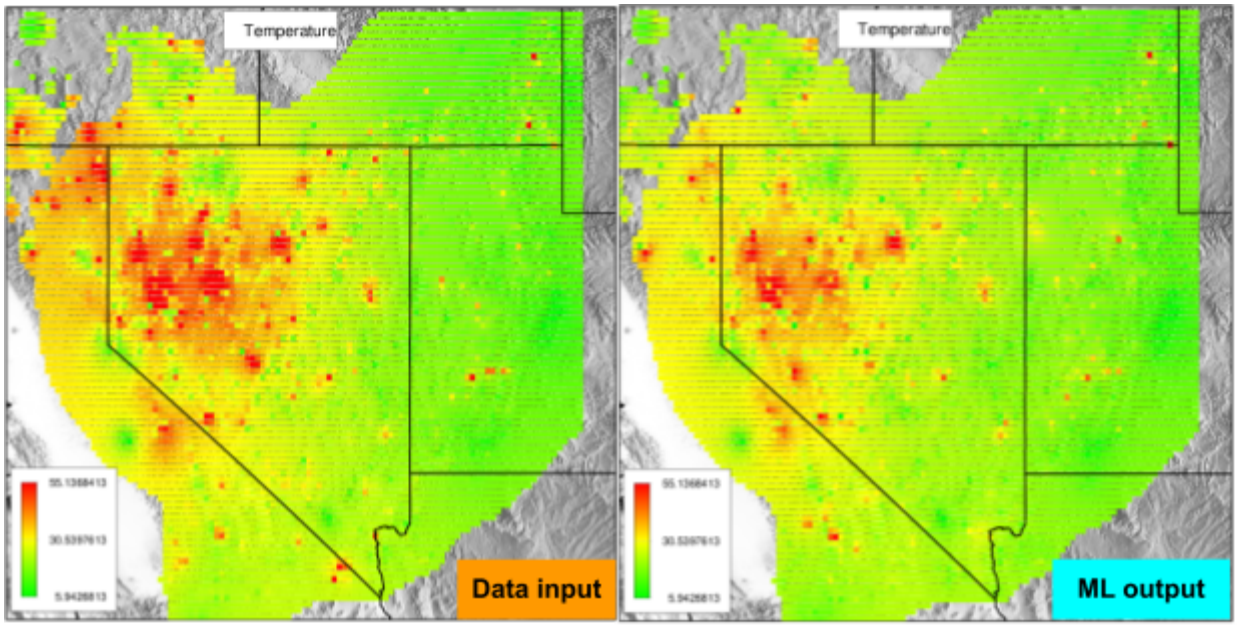


Fig 3.2.4: Actual data (left) and corresponding ML prediction (right) of oxygen isotope spatial distribution.

3.3 Brady site, Nevada

Brady geothermal field is located in northwestern Nevada, USA. It is located in the Basin and Range physiographic province. It has seen geothermal electricity production since 1992 and research or exploration since at least 1959. The existing hydrothermal system supplies hot fluid to two power stations and a direct-use vegetable drying facility. Electricity production capacity at Brady is 26.1 MWe, and ~7 MWe is supplied to the drying facility. Temperatures of produced fluid have been ~130-185°C, though temperatures as high as 219°C have been measured as well. These relatively high temperatures at relatively shallow levels (300-600 depth for some production wells) occur due to either convective upwelling driven by temperature control differences in fluid density or hydraulic head-driven circulation through the hot rock. In either case, relatively high heat flow at the site is associated with crustal thinning providing the heat. Here, our ***GeoThermalCloud*** framework is applied to identify the key geologic factors contributing to the geothermal production in the Brady site. Transmissive fluid flow pathways are relatively rare in the subsurface but are critical components of hydrothermal systems like Brady and many other similar fluid flow systems in fractured rocks. Our ML methods are applied to a library of fourteen 3D geologic characteristics hypothesized to control hydrothermal circulation in the Brady geothermal field.

At the Brady site, as well as at many other geothermal sites, crustal permeability is a key parameter in geothermal models for the exploration and development of these geothermal resources. Permeability is, however, highly variable and heterogeneous in space[61–65], and this substantially complicates the characterization of subsurface hydrothermal processes. Accordingly, it is common in developed geothermal systems to produce fluid flow from a few relatively small (sub-meter- to meter-long) intervals of a borehole that may be 100s or 1000s of meters in total length (based on Nevada Division of Minerals, publicly available data). This compartmentalization of hydrothermal fluid flow means that the volume of rock that transmits fluids at rates suitable for power production or direct use is much smaller than the volume of rock that does not transmit fluid (or transmits at sub-commercial flow rates). This presents a significant challenge to efficient exploration, development, and management of these renewable energy resources. Compartmentalization of the fluid flow system may be associated with a variety of geologic characteristics. For instance, spatial changes in fracture permeability throughout a

fault network and/or permeability variation in the stratigraphic succession may control compartmentalization.

The stratigraphic section at Brady consists of metamorphic basement rocks overlain by Oligocene to late Miocene volcanic rocks and late Miocene to Holocene sedimentary rocks. The Brady fault zone is a west-dipping, north-northeast-striking system of normal faults that cuts this stratigraphic section (Fig.3.4.1). The step-over[66,66–68] is an area where parallel but non-collinear strands of the Brady fault zone come together (e.g., Peacock and Sanderson, 1991, 1994; Fossen and Rotevatn, 2016). The southern segment of the Brady fault zone steps to the left to meet the northern segment (Figure 1[69]). It has been suggested[70] that the occurrence of the hydrothermal system within the step-over is related to focused stress and strain that periodically occur at the step-over during fault slip, resulting in progressive generation and maintenance of a dense fracture network over geologic time. Advection of heat to shallow levels by hydrothermal circulation within this fracture network is evident from a ~3 km-wide × 6 km-long (across strike × along strike) temperature anomaly centered on the step-over (Fig.3.4.1). Geothermal production wells at Brady are situated within the step-over. Fluids are produced from two levels; ~300-600 m and ~1750 m depth (based on Nevada Division of Minerals, publicly available data).

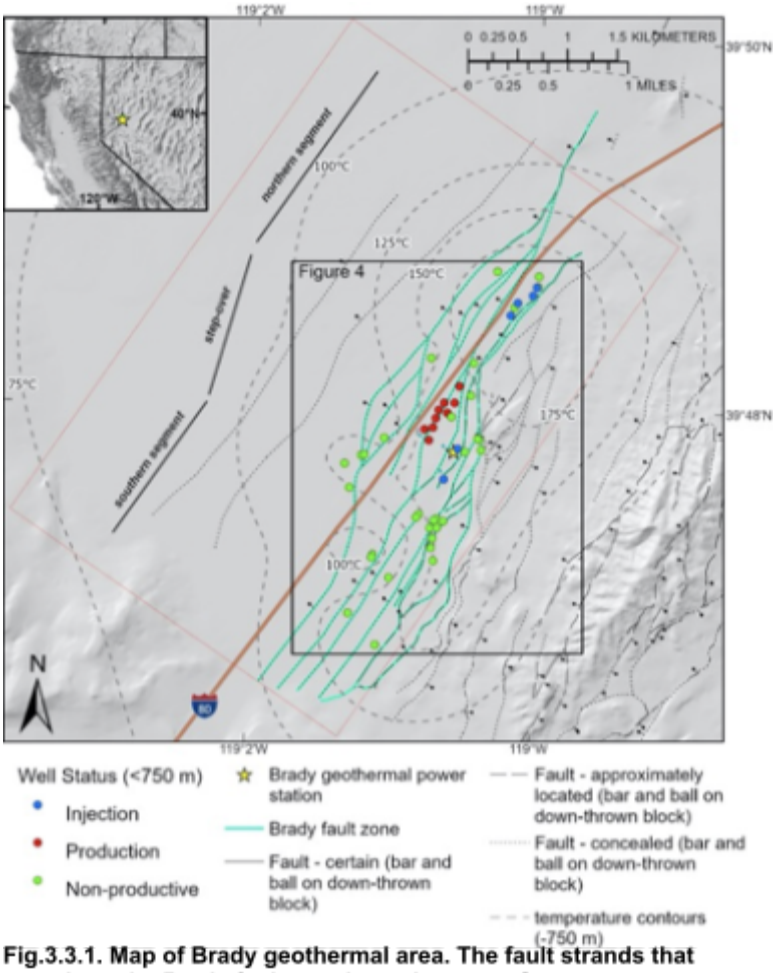


Fig.3.3.1. Map of Brady geothermal area. The fault strands that constitute the Brady fault are shown in green. Contours represent modeled temperature at 750 m depth. Wells are colored by their usage (production, injection, and non-productive) for depths shallower than 750 m. The general geometry of the step-over is shown to the left of the fault system.

Our ML methods are applied to reveal the geologic attributes that influence this compartmentalization of fluid flow in hydrothermal systems. The ML analyses evaluate three-dimensional (3D) geologic characteristics estimated along the production, injection, and non-productive wells at the Brady site. The dataset includes (1) ***Fault***, (2) ***Fault network***, (3) ***Stress and strain***, (4) ***Stratigraphic***, and (5) ***Temperature*** attributes. How these attributes are computed is discussed in greater detail in our research paper (Siler et al., 2021).

Fault attributes are computed for each of the thirty-two faults defined by the 3D geologic map[67,68,71,72]. The ***faulting*** attribute has a value of '1' where a well is located within a fault zone and '0' for well intervals not situated within a fault zone. The ***fault curvature*** attribute is the along-strike and down-dip curvature calculated along each fault. The ***fault dilation tendency*** and ***fault slip tendency*** values are computed using existing methods[73,74] and a local stress model calculated at Brady[75]. The ***fault dilation tendency*** and ***fault slip tendency*** attributes are the ratios of the resolved normal stresses and the normal to shear stress ratio on faults, respectively. Fault segments that are either highly dilatant (high ***fault dilation tendency***) or stress loaded for slip (high ***fault slip tendency***) are likely to host fluid flow[76].

For all wells, the ***inverse distance from faults*** attribute is computed as the difference between the distance to the nearest 3D mapped fault plane and the maximum distance to a fault in the dataset.

Fault network attributes capture areas in the subsurface with especially dense faulting and fracturing are expected to have relatively high permeability and thus host hydrothermal circulation. The ***fault density*** and ***fault intersection density*** are calculated in 3D space[71]. Fault intersections represent structural discontinuities, where stresses become concentrated and accentuated fracturing is expected[77,78]. Similarly, areas with many closely spaced faults are also likely to have a relatively high density of fractures, i.e., high permeability.

Stress and strain attributes characterize the step-over in the Brady fault, which is an important attribute controlling the presence of hydrothermal circulation at Brady[66,66–71]. Stress and strain become concentrated at the step-over when slip occurs on the Brady fault, and the location of the stress and strain perturbation is largely concomitant with the production well field and the local temperature anomaly[70]. The 2D modeled ***dilation***, ***normal stress,*** and ***coulomb shear stress*** as a result of 1-meter normal slip on the Brady fault are calculated at 250-meter-depth intervals from the surface to 750 m depth[70]. Siler et al. (2018)[70] suggest that the stress and strain perturbations that occur with fault slip result in a zone of accentuated secondary faulting and fracturing that is an essential attribute in localizing hydrothermal circulation in the step-over. To account for the uncertainties associated with the state of stress at the site, in a set of additional ML analyses, the ***dilation***, ***normal stress,*** and ***coulomb shear stress*** are computed for a series of stress ratios: 1:1, 1:2, 1:3, 1:4, 2:1, 3:1, and 4:1.

Stratigraphic attributes characterize changes in the permeability caused by stratigraphy. The distance from the nearest stratigraphic contact (***inverse distance from contacts***) is calculated as the difference between the distance to the closest stratigraphic contact along each well and the maximum distance to a contact in the dataset. In this case, high values of ***inverse distance from contacts*** would be expected to correlate with hydrothermal fluid flow. Alternatively, relatively thick geologic units, i.e., relatively large, intact volumes of rock distal to stratigraphic contacts, may focus strain on a relatively small number of dominant, high-aperture fractures. Areas with high values for the thickness of each stratigraphic unit (***stratigraphic unit thickness***) from the 3D geologic map could be favorable for localizing hydrothermal circulation in this case. The ~300-600-m-depth production reservoir at Brady occurs in Miocene mafic to intermediate volcanic rocks. It is possible that these specific stratigraphic units have high matrix porosity and permeability and/or are particularly favorable for developing highly transmissive fracture systems when faulted. The ***good lithology*** attribute is '1' for well intervals with these stratigraphic units and '0' for intervals in other units.

Temperature attribute: Advection is a much more efficient means of heat transport than conduction. Higher temperatures, therefore, are expected within or near transmissive fluid flow conduits. Equilibrated temperature logs from thirty-nine deep (as deep as ~2 km) geothermal wells and seventy-nine shallow (~150 m) temperature gradient wells[79] were utilized to build a 3D temperature model. The ***modeled temperature*** is projected to each of the forty-seven wells at the Brady site.

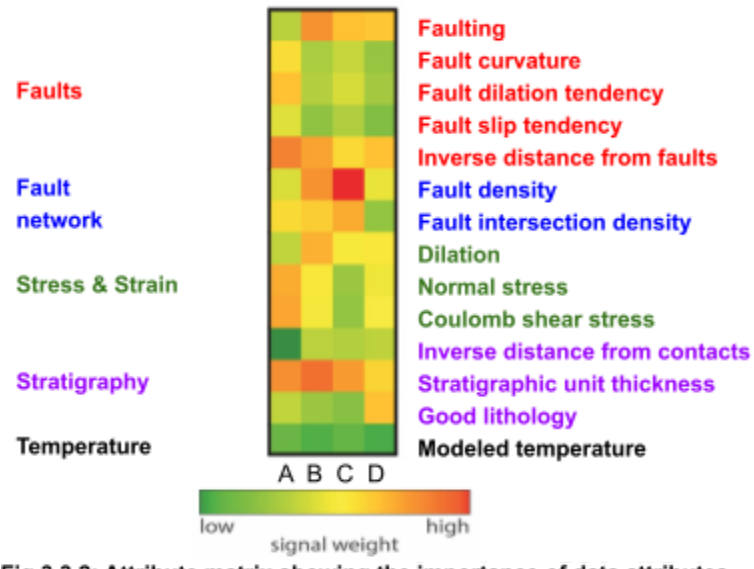


Fig.3.3.2: Attribute matrix showing the importance of data attributes (rows) to represent extracted geothermal signatures (columns). Red colors indicate attributes with high importance with a particular signature, green colors show that an attribute has a low weight with a specific signature.

Based on all the data discussed above, a 3D data tensor has been

generated. The first tensor dimension represents the 47 wells. The second dimension is indexing the 14 attributes (discussed above). For each well and attribute, we have data for 750 vertical depths 1 m apart, and this defines the third tensor dimension. The processed Brady dataset is available in our ***GeoThermalCloud*** repositories on GitHub and GDR. The 3D data tensor is analyzed using both matrix and tensor factorization methods. Both methods produced similar results. The ML algorithm extracted 4 hidden geothermal signatures labeled A, B, C, and D[6].

How the hidden geothermal signatures are related to analyzed geothermal attributes is summarized in Fig.3.4.2. The figure plots the attribute matrix, which defines the weights of each attribute pertaining to the 4 geothermal signatures. In Fig.3.4.2, red colors indicate attributes with high importance with that particular signature, green colors show that an attribute has a low weight with that specific signature.

The associations of the four signatures with the site wells are also identified. Fig.3.4.3 shows the Brady site and the geothermal signatures to which each well falls into.

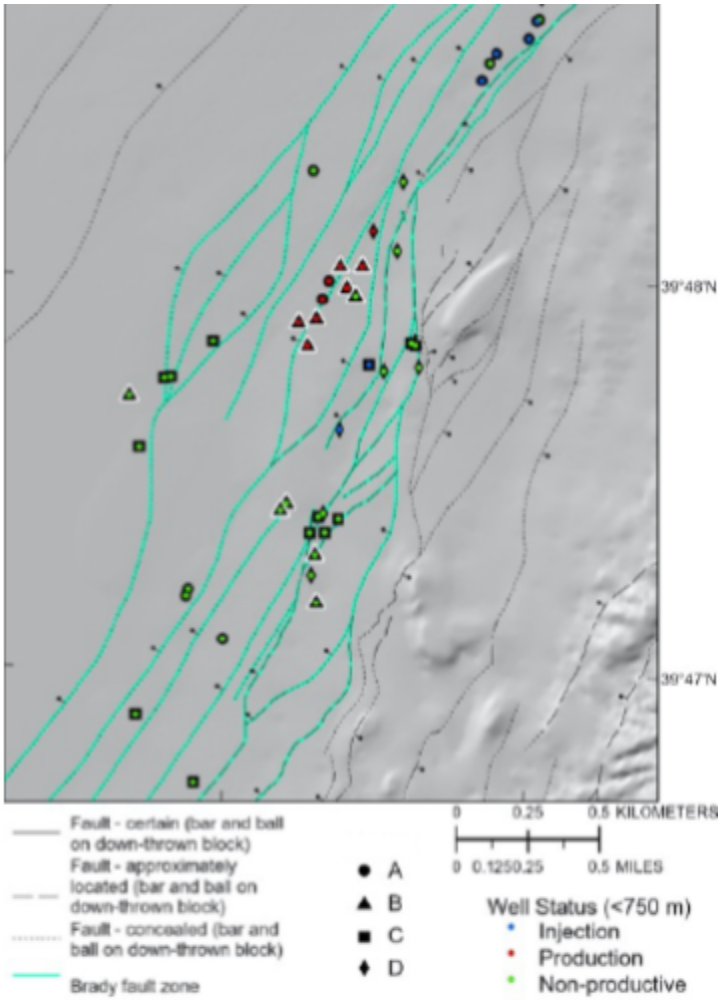


Fig.3.3.3. Map of the Brady well field and fault system. Wells are shown by their use (production, injection, or non-productive) and their cluster (A, B, C, or D). Cluster B (triangles) contains six of the nine production wells.

Figure 3.4.4 shows a biplot of signatures B and A. These two signatures most effectively separate the production wells from the injection and non-productive wells. The production wells have relatively high B values and relatively low A values. The injection wells have low B values. The attributes with relatively high B values and relatively low A values are those that most effectively separate the production wells from the other wells. These results show that six of the nine wells that have been used for geothermal production at Brady from the shallow (~300-600 m depth) reservoir (June 1992-August 2019) are associated with signature B. NMFk results indicate that the ***fault density*** and ***fault intersection density*** attributes are more dominant in signature B (and therefore predominate along the production wells) relative to injection wells and non-productive wells.

Fault attributes: The ***faulting*** is the predominant fault attribute associated with B (Fig.3.4.2). This is evident on Fig.3.4.4 in which ***faulting*** plots in the upper left quadrant, with relatively high B and relatively low A; a similar pattern is observed at the majority of the production wells. Though ***fault intersection density*** also has high B values, it has somewhat higher values in A, C, and D (Fig.3.4.2), so it is less distinctly related to B relative to ***faulting***. In Fig.3.4.4, this is evident from ***inverse distance from faults*** plotting in the upper-right quadrant, farther to the right relative to the production wells. These results suggest that the presence or absence of distinct, macro-scale fault zones is strongly related to production wells, more so than to the other fault attributes such as ***inverse distance from faults***, ***fault curvature***, ***slip tendency***, or ***dilation tendency***.

Fault network attributes: The ***fault density*** is the predominant fault network attribute related to signature B and the production wells (Figure Fig.3.4.2). This is evident on Fig.3.4.4, in which ***fault density*** plots in the upper left quadrant, with relatively high B and relatively low A, similar to the majority of the production wells. Though the ***inverse distance from faults*** has a high B value, it also has relatively high A and C values (Fig.3.4.2) and plots to the right of the production wells relative to ***fault density***. This indicates that ***fault density*** is more strongly related to production wells than ***fault intersection density***.

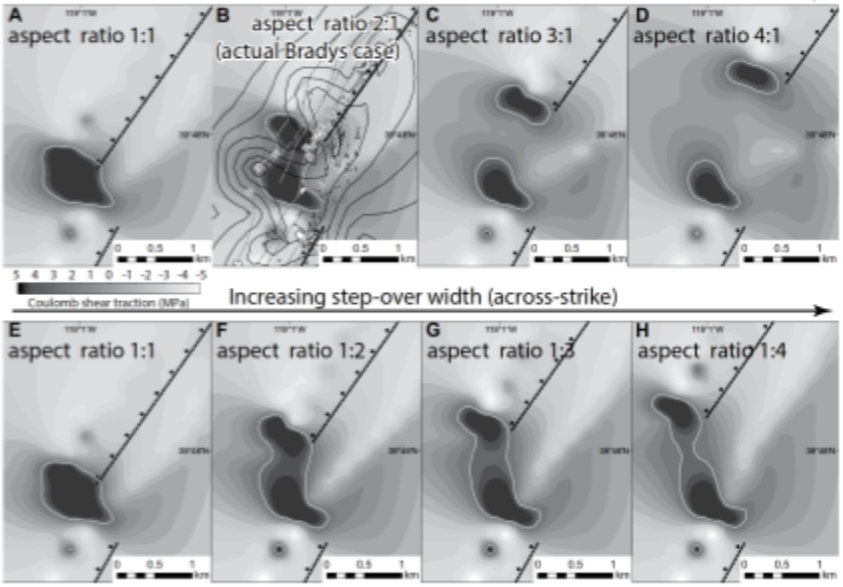


Fig.3.3.5: Coulomb shear traction estimated at 1000 m depth. Dark colors show high Coulomb shear traction on optimally oriented normal faults caused by slip. An aspect ratio of 2:1 most probably represents the state of stress at the Brady case (after Siler et al., 2018)

Stress and strain attributes: ***Dilation*** occurring due to modeled fault slip is the predominant stress/strain attribute related to B and the production wells. This is evident on Fig.3.4.4, in which ***dilation*** plots in the upper left quadrant, with relatively high B and relatively low A values, similar to the majority of the production wells. These results indicate that ***dilation*** is strongly related to production wells relative to ***normal*** or ***coulomb***, the other stress/strain network attributes examined herein.

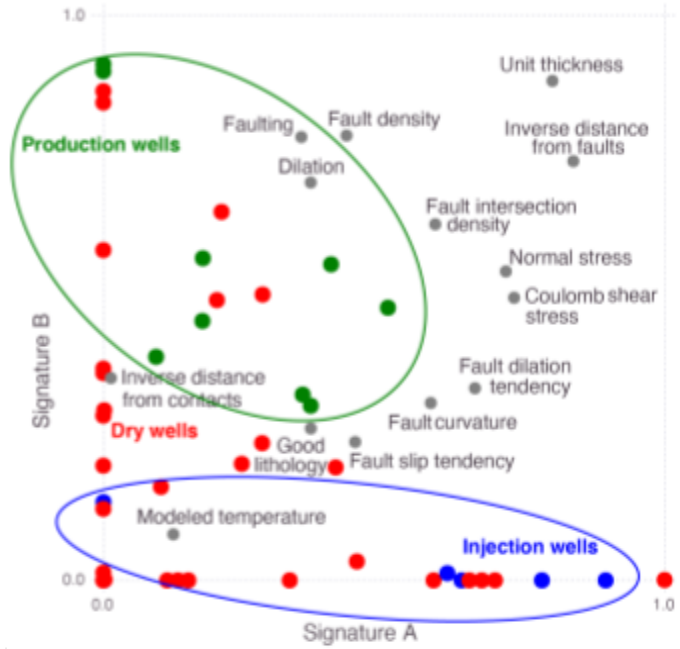


Fig.3.3.4. Biplot of Signatures A and B. Production wells (green) and the *faulting*, *dilation*, and *fault density* attributes have high B values and relatively low A values indicating that these attributes control the separation of the production wells from the rest of the data set.

Stratigraphic attributes: ***Stratigraphic unit thickness*** attribute has high B values relative to A values (Fig.3.4.2). Relative to the nearness to geologic contacts (***inverse distance from contacts***), and the specific geologic units associated with geothermal production (***good lithology***), ***unit thickness*** is more strongly related to B and the production wells. However, A values for ***unit thickness*** are high relative to the production wells (***unit thickness*** plots in the upper right on Fig.3.4.4), so ***unit thickness*** appears to be less strongly related to the production wells relative to ***dilation***, ***fault density***, and ***faulting***.

The ***modeled temperature*** has relatively low values for all extracted hidden signatures. This suggests that the ***modeled temperature*** is not significantly higher or lower along with any subset of wells relative to the others.

Geologic controls: The ML results suggest that two dominant characteristics of the geologic structure control hydrothermal processes at <750 m depth at Brady are: (1) the distinct, macro-scale faults and (2) the step-over in the

Brady fault system. The macro-scale faults are those that are mappable in 3D at the local scale using geologic and geophysical evidence (Siler et al., 2021). Interestingly, this relatively simple fault attribute, the binary occurrence or non-occurrence of a 30-m-wide fault zone (Figure 2), is more closely related to the production wells than the static stress state of the faults (***dilation*** or ***slip tendency***), the geometry of the faults (***fault curvature***), or the nearness to the mapped fault planes (***inverse distance from faults***). These are important conclusions that can be made without the use of our advanced ML methods.

Brady fault zone: The step-over in the Brady fault zone (Fig.3.3.1) is the other dominant geologic attribute controlling hydrothermal processes at Brady. The geometry and location of the step-over control the spatial density of fault planes (***fault density***) since faults are most dense in the step-over (Fig.3.3.1). The step-over also controls the dilatational strain that occurs as a result of modeled fault slip on the Brady fault zone (***dilation***)[70]. The ML results suggest that the spatial density of fault planes and the modeled dilation are the most effective indicators of step-over's control on hydrothermal processes, relative to the other stress and strain attributes (***normal*** and ***coulomb***) and the spatial density of fault intersections (***inverse distance from faults***). Thicker geologic units (***unit thickness***) may also influence hydrothermal processes. This may indicate that faults cutting through thicker geologic units preferentially transmit the high flow rates necessary for geothermal production relative to faults cutting thinner units. However, based on the ML analyses, this control appears to be secondary to the macro-scale faults and the step-over. The ***modeled temperature*** attribute is not strongly related to production wells relative to the other wells. It is likely that our extrapolation of the existing temperature data does not sufficiently resolve advective relative to conductive heat transport, and thus ***modeled temperature*** is relatively ineffective for determining discrete fluid flow pathways.

State of stress: To better estimate the capability of our ML algorithms to evaluate the impact of the state of stress for geothermal production at the Brady site, we performed additional ML analyses. To account for the uncertainties associated with the state of stress at the site, in a set of additional ML analyses, the ***dilation***, ***normal stress,*** and ***coulomb shear stress*** are computed for a series of stress ratios: 1:1, 1:2, 1:3, 1:4, 2:1, 3:1, and 4:1. The impact of the alternative ratios is visualized in Fig.3.3.5. All these 7 stress cases are used together with other geologic attributes to create 7 alternative datasets. Each dataset is represented by 3D data tensors with dimensions equivalent to the datasets analyzed above. The only differences between the 7 tensors are in the 3 stress attributes: ***dilation***, ***normal stress,*** and ***coulomb shear***. The goal of the ML analyses is to select which of these 7 stress ratios are the most representative for the site. This is evaluated based on the quality of the reconstruction of the original tensors, which is achieved by the ML algorithm. The better the reconstruction, the higher the ML estimated consistency between the 3 stress attributes (***dilation***, ***normal stress,*** and ***coulomb shear***) and the remaining 11 geothermal attributes. Estimates of the location matrix under the 7 different stress ratios are shown in Fig.3.3.6. The production wells, injection, and non-productive wells have similar associations

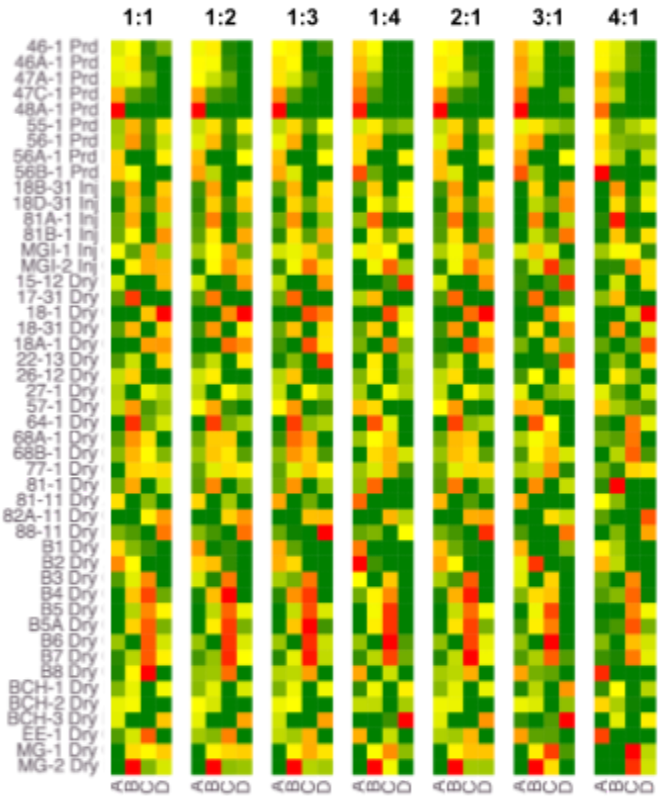


Fig.3.3.6: Estimates for the location matrix under 7 different stress ratios. The production wells (first 9 rows), injection (nest 6 rows), and non-productive (dry; remaining rows) wells have similar associations with the extracted signatures for the 7 different stress scenarios.

with the extracted signatures for the 7 different stress scenarios. However, the overall reconstruction error is the lowest for the 2:1 ratio case (the seven reconstruction errors are: 1:1 - 12772, 1:2 - 12773, 1:3 - 12691, 1:4 - 12685, 2:1 - 12666, 3:1 - 12928, and 4:1 - 12788; the lowest estimate is for 2:1). In fact, this is the most probable stress ratio at the site (2:1) based on previous studies.

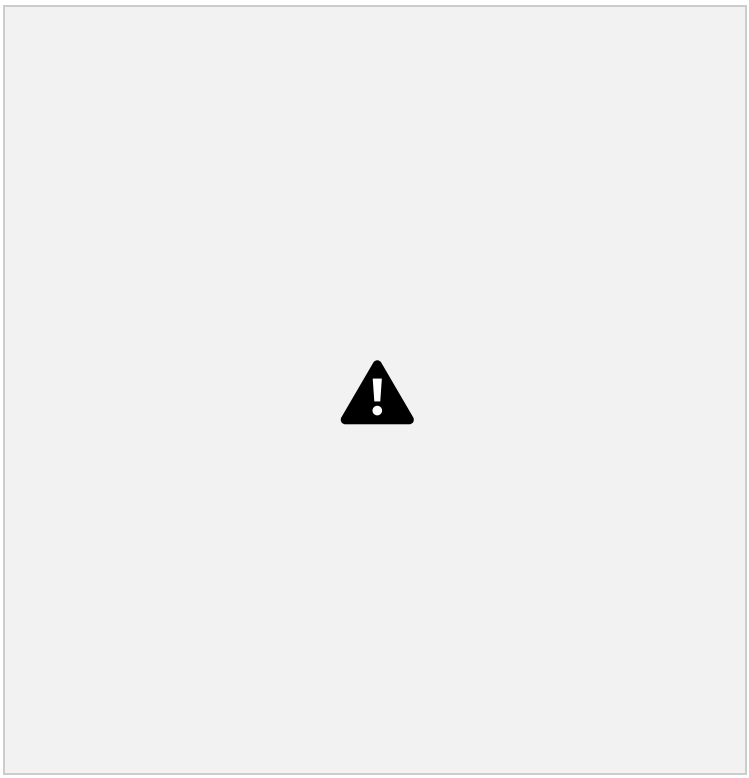
The ML analyses conducted on a 3D geological dataset from Brady geothermal field elucidate the geologic characteristics that control hydrothermal circulation in the shallow (~300-600 m depth) geothermal reservoir. The ML results show that known, macro-scale faults, i.e., those that have been mapped in 3D based on geological and geophysical evidence, are strongly associated with the production wells at the Brady site. Geologic attributes that occur most prominently within the Brady step-over, such as high spatial densities of faults, and dilatation brought on by modeled fault slip, are also critical geologic attributes associated predominantly with production wells relative to other wells. These results suggest that the shallow hydrothermal reservoir at Brady is hosted by relatively prominent faults. Locations where such faults lie within the subsurface projection of the step-over (i.e., the volume of rock with relatively high fault and fracture density and where fractures tend to dilate due to periodic fault slip) are exceptionally well suited for geothermal production. In concert and not either independently, these two attributes control the presence of the Brady hydrothermal system that has been developed for electricity production and direct uses. The NMFk methodology successfully differentiates production wells amongst a large number of non-productive wells using just these geologic data. This suggests that these geologic attributes may be effective as training data for using ML techniques to identify areas within unexplored subsurface volumes that have the geologic characteristics that constitute productive geothermal wells.

All the data and codes, including Jupyter and Pluto notebooks, required to reproduce these results are available at the ***GeoThermalCloud*** GitHub and GDR repositories (https://github.com/SmartTensors/GeoThermalCloud.jl/tree/master/Brady).

3.4 GeoDT synthetic dataset

Our rapid multi-physics GeoDT model[80] was used to generate a library of over 2000 geothermal production scenarios based on the UtahFORGE site's parameters. This GeoDT modeling approach enables valuation that considers the interplay between general site parameters (e.g., depth and thermal gradient), in-situ stress attributes (e.g., stress anisotropy), rock mechanical attributes (e.g., elastic moduli), natural fracture strength and permeability characteristics (e.g., hydraulic aperture and friction angle), natural fracture intensity (e.g., number, orientation, and spacing for fractures), fracture complexity (e.g., roughness), and site design decisions (e.g., well spacing and well orientation). GeoDT also predicts maximum induced seismic magnitudes using a built-in length, displacement, aperture, and stress scaling relationship that is based on existing power-law scaling relationships[81]. The site-specific parameters from the UtahFORGE site used for the model are given in Table 3.8.1. Each modeled scenario included stochastically generated natural fractures. An example system is visualized in Fig. 3.8.1. To solve this system, GeoDT completes the following computational sequence:

1. Natural fracture placement, well placement, and calculation of fracture activation pressure (Pc) based on the far-field stress state, mechanical properties, and orientations.
2. Hydraulic stimulation by simultaneous injection into all of the intervals. This ignores sequencing and staging but accelerates the solver.
3. Long-term flow calculation with consideration of continued stimulation, far-field leakoff, and 3D connectivity issues through the well and fracture network.
4. Long-term transient heat-extraction and production simulation where heat from the rock transfers to the injected fluid.
5. Electrical power output calculation via the Single-flash Rankine steam cycle. More advanced and higher efficiency cycles are not evaluated at this time, so the estimates will be lower than what is achievable by the best available technologies.



Combining these mechanisms, this tool models the whole geothermal development cycle from initial well design to the end of production. Since GeoDT includes geomechanical coupling between fracture properties and stress, results allow us to probe the influence of stress data on geothermal production potential at a given site. GeoDT also enables us to investigate links to seismicity and the benefits or consequences of key design attributes such as well spacing, orientation, and diameter.

Results from GeoDT (Fig. 3.8.2) predict the time series of geothermal power production for each of over 2000 subsurface scenarios. For each scenario, GeoDT estimates when thermal breakthrough (i.e., produced fluid cooling) begins, and the thermal and electrical power outputs over time. Initial inspection of the results from our GeoDT analysis shows an apparent link between the well spacing and the electrical power output of the system after 20 years of production. There also appears to be a strong link between the number of injection intervals and power output. However, these links are only a small portion of what can be identified using ML methods developed by our team.

Applying ***GeoThermalCloud*** ML methods reveals four constitutive multi-attribute input signatures that control the time series of the produced fluid enthalpy (i.e., geothermal fluid energy) and the related electrical power potential. The structure of these signatures is shown in Fig.3.8.3. The roles of all four signatures are stronger and more varied for enthalpy output (i.e., thermal power output) than for electrical power output, where two of the signatures are almost flat. Each signature is constructed from multiple input attributes and captures the impact of model inputs onto the model outputs. The complete composition of each controlling signature is shown in Fig.3.8.4. Based on this result, the dominant attribute in each signature is identified by the largest numbers (marked with red boxes in Fig. 3.8.4) in each signature. We can use these dominant attributes to categorize the signatures into combining (1) well spacing & other attributes, (2) stress & other attributes, (3) system (i.e., site conditions) & other attributes, and (4) well dip (i.e., orientation) & other attributes. The signature that includes well spacing is the only input that links to strongly increased power production over time. Increased in situ stress causes decreased production over time. Here it is important to note that increased stress will cause fracture closure after stimulation which will likely reduce production, but this stress increase will also provide for more shear stress. Shear stress is a prerequisite for shear stimulation of fractures to increase reservoir performance, but it is also a driver for induced seismicity. Additional work is needed to parse out the meaning of these signatures and implications for site-specific geothermal energy production. GeoThermalCloud coupled with GeoDT provides a good platform for this future work owing to its ability to rapidly model the effect of complex interactions and design decisions on production for an extensive range of site conditions.

GeoThermalCloud ML methods also allow investigation of the effects of the input attributes (Fig.3.8.4) on other outputs such as maximum induced seismic magnitude, far-field leakoff, and the number of fractures that interlink the injection and production wells (Fig.3.8.5). Interestingly, there appears to be a link between the well-spacing dominated signature and the maximum induced seismic event magnitude. It is not yet clear what underlying mechanism drives this connection. Less surprisingly,

the system attributes (e.g., natural fractures, well length, well diameter, and rock properties) have a strong influence on the amount of fluid loss (i.e., boundary outflow rate) from the system. Stress effects on the GeoDT results are clearly evident, but a clear causal pattern is not immediately apparent. Instead, stress appears to associate with mixed effects, some positive and others negative. Another surprising result is the importance of well dip and azimuth (i.e., well orientation). The cause of this importance is suspected to be linked to the natural fracture orientations, especially Joint Set 3, which is northeast striking and southeast dipping, making it a prime target for shear slip. At the UtahFORGE site, the nearby Opal Mound Fault is also northeast striking and southeast dipping. The planned well orientation at UtahFORGE is nearly perpendicular (i.e., face on) to this fault. Note that the presented results are preliminary, and the GeoDT model was only just completed in 2021. More validation of GeoDT is needed to gain confidence in these model predictions and their importance to guide field exploration and drilling decision making. Further investigation of the identified signatures is required to more clearly understand the links and implications. This additional work is planned for Phase 2 of our project.

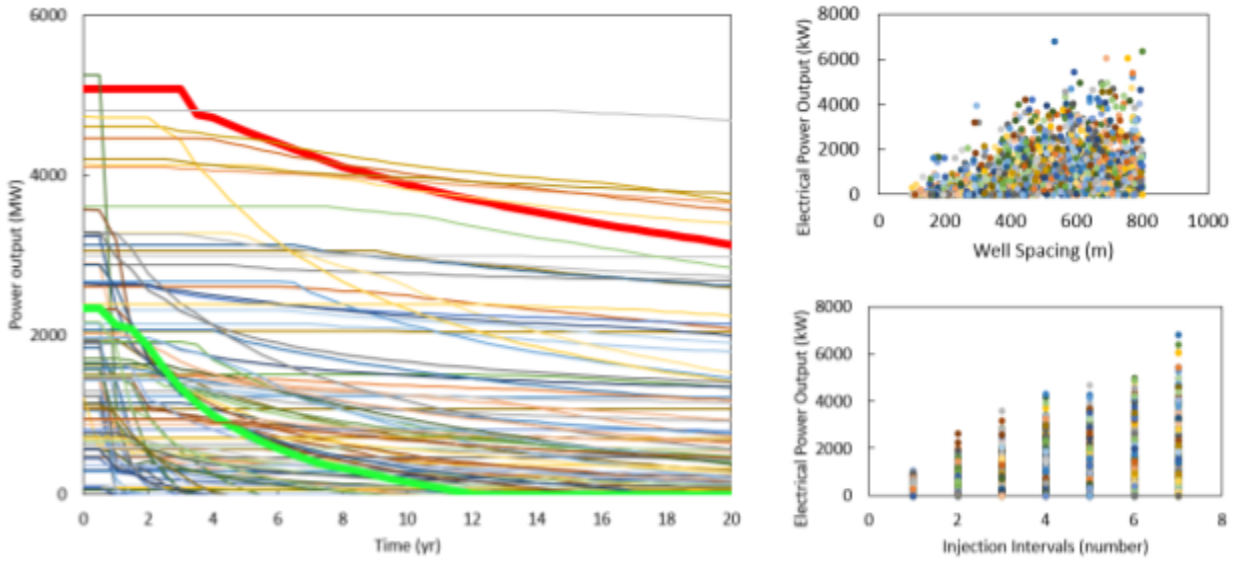


Fig. 3.8.2. Compiled results from more than 2000 geothermal power production simulations based on the parameters described in Table 3.8.1. In the time series plot, a high-performing case is highlighted in red, and a poor performer is highlighted in green. There is also a clear link between the well spacing and power output in addition to the number of injection intervals (i.e., isolated zones) and power output (plots on the right).

Table 3.8.1: Partial list of properties for GeoDT modeling scenarios.

Attribute	Minimum	Mean	Maximum
Domain size (m)	3000	3000	3000
Depth (m)	4000	6000	8000
Geothermal gradient (C/km)	40	45	50
***Well orientation & dip (deg AzN & deg down)**	89 & -15	104 & 0	119 & 15
Well length & spacing (m & m)	1500 & 100	1500 & 450	1500 & 800
Well diameter (m)	0.15	0.28	0.41
Number of injection zones (#)	1	5	8
***Joint set 1 strike & dip (deg AzN & deg down)**	-15 & -7	96 & 80	+15 & +7
***Joint set 2 strike & dip (deg AzN & deg down)**	-15 & -7	185 & 48	+15 & +7
***Joint set 3 strike & dip (deg AzN & deg down)**	-15 & -7	35 & 64	+15 & +7
Minimum horizontal stress anisotropy (%)	30%	60%	90%
Intermediate horizontal stress anisotropy (%)	60%	90%	120%
Rock density (kg/m3)	2700	2700	2700
Thermal conductivity (W/m-K)	2.1	2.45	2.8
Specific heat capacity (kJ/m3-K)	1900	2050	2200
Rock Young's modulus & Poisson's ratio (GPa & ratio)	30 & 0.15	60 & 0.25	90 & 0.35
Production lifespan (yr)	20	20	20
Injection temperature (C)	95	95	95
Production pressure drawdown (MPa)	2.0	2.0	2.0
Stimulation rate & volume (m3/s & m3)	0.08 & 100000	0.08 & 100000	0.08 & 100000
Long term injection rate per zone (m3/s)	0.01	0.01	0.01
Fracture frictional angle & cohesion (deg & MPa)	25 & 5	35 & 10	45 & 15
Fracture roughness factor (no unit)	0.7	0.85	1.0
Gutenberg-Richter b-value (no unit)	1.0	1.0	1.0

*data based on UtahFORGE

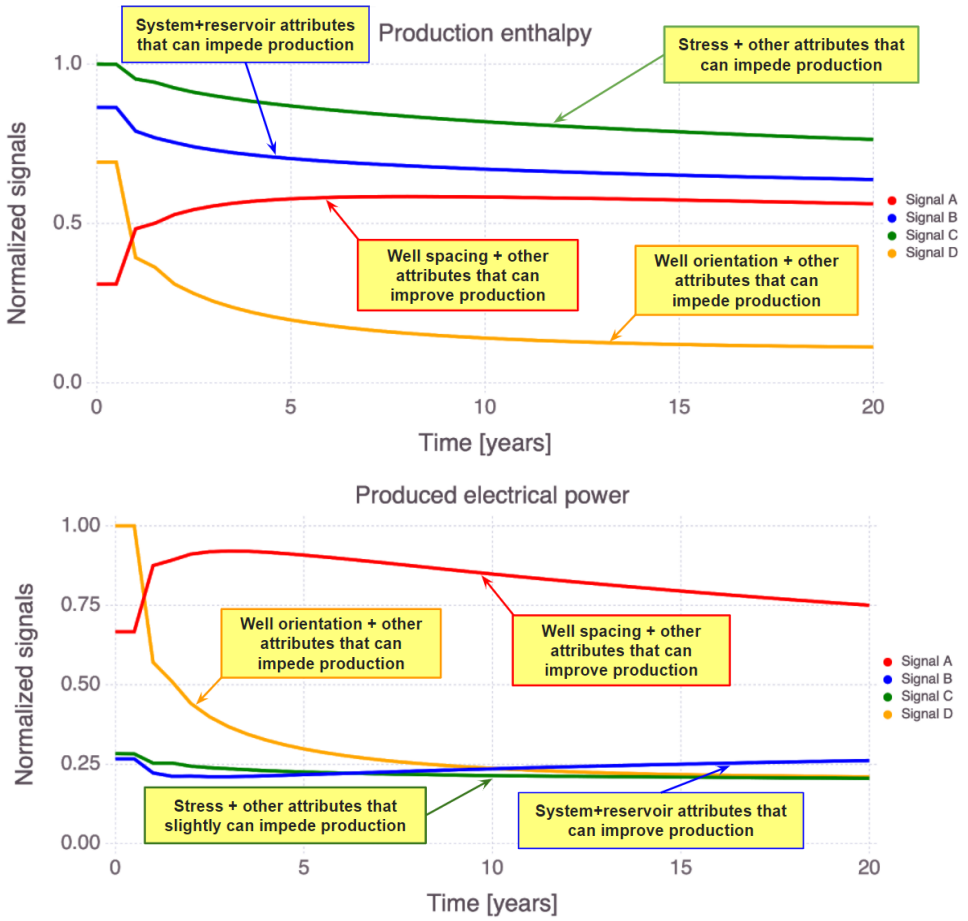
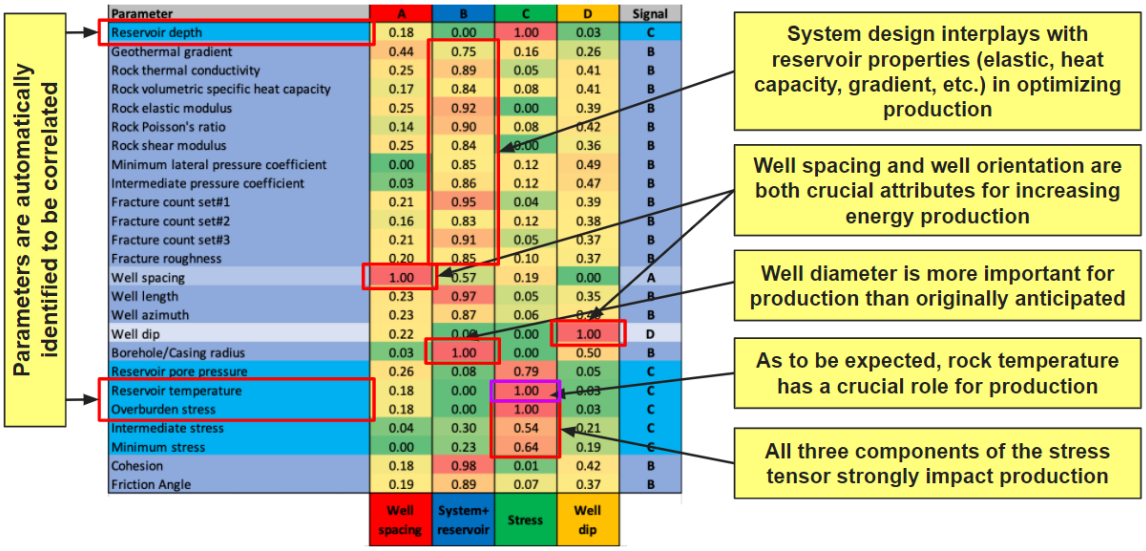


Fig. 3.8.3. ML identifies the signature structure of the enthalpy and power production time series predicted by GeoDT. The primary physical components of each mixed signature are provided to aid interpretation. Only one of the signatures (red) shows inputs that associate with increased production over time.



Parameter	A	B	C	D	Signal
Reservoir depth	0.18	0.00	1.00	0.03	C
Geothermal gradient	0.44	0.75	0.16	0.26	B
Rock thermal conductivity	0.25	0.89	0.05	0.41	B
Rock volumetric specific heat capacity	0.17	0.84	0.08	0.41	B
Rock elastic modulus	0.25	0.92	0.00	0.39	B
Rock Poisson's ratio	0.14	0.90	0.08	0.42	B
Rock shear modulus	0.25	0.84	0.00	0.36	B
Minimum lateral pressure coefficient	0.00	0.85	0.12	0.49	B
Intermediate pressure coefficient	0.03	0.86	0.12	0.47	B
Fracture count set#1	0.21	0.95	0.04	0.39	B
Fracture count set#2	0.16	0.83	0.12	0.38	B
Fracture count set#3	0.21	0.91	0.05	0.37	B
Fracture roughness	0.20	0.85	0.10	0.37	B
Well spacing	1.00	0.57	0.19	0.00	A
Well length	0.23	0.97	0.05	0.35	B
Well azimuth	0.23	0.87	0.06	0.34	B
Well dip	0.22	0.00	0.00	1.00	D
Borehole/Casing radius	0.03	1.00	0.00	0.50	B
Reservoir pore pressure	0.26	0.08	0.79	0.05	C
Reservoir temperature	0.18	0.00	1.00	0.03	C
Overburden stress	0.18	0.00	1.00	0.03	C
Intermediate stress	0.04	0.30	0.54	0.21	C
Minimum stress	0.00	0.23	0.64	0.19	C
Cohesion	0.18	0.98	0.01	0.42	B
Friction Angle	0.19	0.89	0.07	0.37	B

Fig. 3.8.4. Combined inputs of the four ML identified signatures that control geothermal power production. Callouts are included to highlight the primary physical components of each signature. We categorize each signature by its most dominant component. Red colors indicate parameters with high importance with that particular signature, green colors show that a parameter has a low weight with that specific signature.

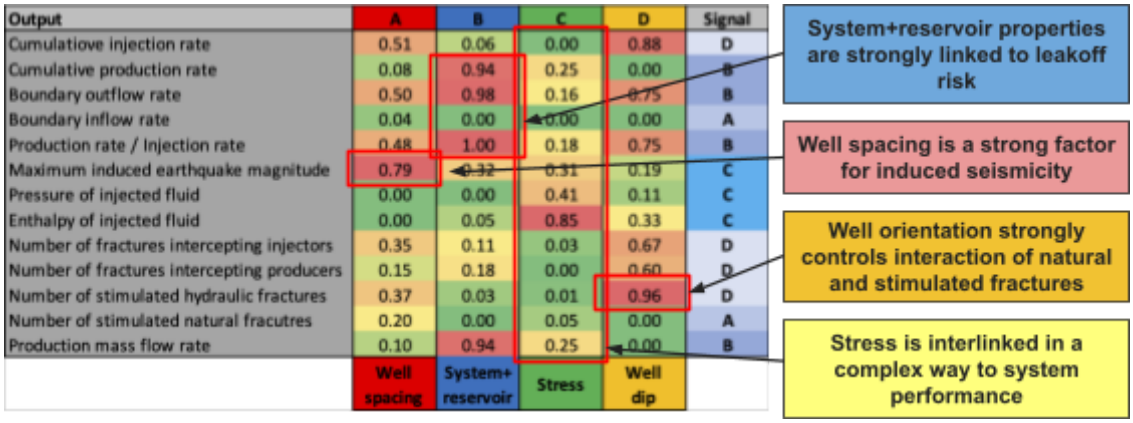


Fig. 3.8.5. ML predicted effects of the identified multi-attribute signatures on various outputs from GeoDT. Callouts are included to highlight significant links that aid understanding of subsurface geothermal processes coupled with system design. Red colors indicate outputs with high importance with that particular signature, green colors show that a specific output has a low weight with that specific signature.

3.5 Southwest New Mexico synthetic analyses

In section 3.1, we showed that the northern MDVF and southern Rio Grande provinces have mid-temperature geothermal resources. These areas require further investigation to better characterize geothermal conditions and pinpoint the prospective geothermal resource locations. To do this, numerical-model analyses of the subsurface geothermal process are critical. However, in the SWNM region, there are significant uncertainties associated with subsurface geothermal conditions, including rock permeabilities, heat flux, and thermal gradient. Only a few studies[40,41,82] discussed the permeability uncertainty ranges. Literature values for heat flux and thermal gradients are also scarce, and existing estimates are also uncertain.

To address these issues, we built a 3-D thermo-hydro-chemical (THC) model using the LANL simulator PFLOTRAN (Fig. 3.5.1)[83]. The model simulates heat and mass transport and predicts the spatiotemporal distribution of temperature, B+, and Li+ concentrations in the subsurface. To capture stratigraphic uncertainties, the model has nine layers with uncertain permeabilities (Table 3.5.1). Also uncertain are the heat influx at the bottom and the thermal gradient. These simulations are slow and computationally expensive. Using such models, making model diagnostic analyses such as calibration against data, uncertainty quantification, sensitivity evaluation, and decision-support analyses are challenging because they require numerous model runs.

To reduce the computational time and to be able to perform the above-mentioned analyses, we have developed a neural network-based surrogate model as a faster alternative to the THC model. The ML model is using a multi-layer perceptron (MLP) based neural network[84]. The MLP network has three layers because which have been demonstrated to be sufficient to capture the underlying data and physics complexity. The optimal MLP hyperparameters. (e.g., activation function, loss function, number of layers, learning rates) are estimated through hyperparameter tuning[85]. To train, test, validate the ML

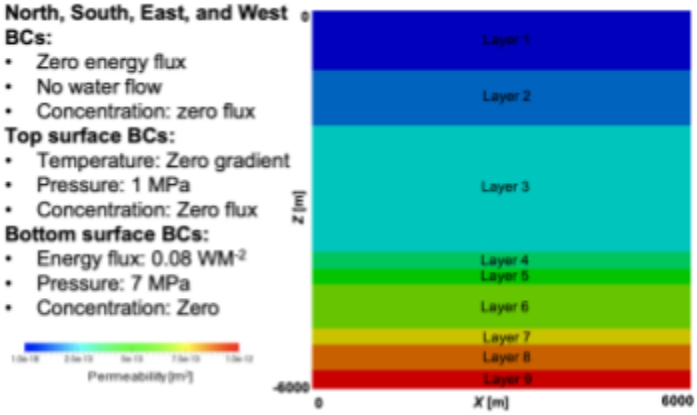


Fig. 3.5.1: Schematic of the THC model domain, layers, boundary and initial conditions for a SWNM site.

model, we generated 10,000 input parameter realizations using Markov Chain Monte Carlo with the delayed rejection adaptive Metropolis sampling technique[86], and executed our THC model to predict corresponding temperature, Li+, and B+ concentrations. The realizations are generated by varying the 11 model inputs in Table 3.5.1. More uncertain parameters will be included during the next project phase.

This ML work is ongoing, and the results will be ready soon. The developed ML surrogate model predicts temperature, Li+, and B+ concentrations much faster than numerical simulations (in order of milliseconds). After the ML model is finalized, we will be used for (1) calibration against observed SWNM data, and (2) quantification of the SWNM uncertainties in permeabilities, heat flux, and thermal gradients. Data-acquisition optimizations, sensitivity (local and global) evaluations, and decision-support investigations related to SWNM geothermal exploration and production will be performed as well.

Table 3.5.1: Model inputs and their ranges (Pepin, et al., 2015).

Attribute	Minimum (m^2)	Maximum (m^2)
Permeability L1	1E-13	1E-11
Permeability L2	1E-17	1E-15
Permeability L3	1E-18	1E-16
Permeability L4	1E-18	1E-16
Permeability L5	1E-16	1E-14
Permeability L6	1E-12	1E-10
Permeability L7	1E-13	1E-11
Permeability L8	1E-14	1E-12
Permeability L9	1E-16	1E-14
Heat flux (Wm-2)	0.06	0.09
Thermal gradient	0.023	0.041

4 Conclusions

Geothermal exploration and production challenges come from the fact that various processes and parameters impacting geothermal conditions are poorly understood. Diverse datasets are available that can be applied to characterize subsurface geothermal conditions (public and proprietary; satellite, airborne surveys, vegetation/water sampling, geological, geophysical, etc.). However, how exactly all these datasets can be applied for geothermal exploration is poorly understood due to an imperfect understanding of how physical processes impacting subsurface geothermal conditions are represented in these observations. As demonstrated in our report, our novel ***GeoThermalCloud*** technology provides a solution to address these challenges.

Our project has successfully developed, tested, demonstrated, and validated a novel framework for geothermal data processing and machine learning (ML) called ***GeoThermalCloud***. To address the challenges and risks associated with geothermal exploration and production, ***GeoThermalCloud*** brings novel computational methods and tools that can (1) analyze large field datasets, (2) assimilate model simulations (large inputs and outputs), (3) process sparse datasets, (4) perform transfer learning (between sites with different exploratory levels), (5) extract hidden geothermal signatures in the field and simulation data, (6) label geothermal resources and processes, (7) identify high-value data acquisition targets, and (8) guide geothermal exploration and production by selecting optimal exploration, production, and drilling strategies. It is important to note that the capability to process sparse datasets is critical for geothermal exploration; this feature is generally absent in traditional machine learning methods. This report demonstrates the applicability of our novel ML methods within ***GeoThermalCloud*** to process and analyze geothermal information. All the project goals, milestones, deliverables, and go/no-go decision points outlined in the original Phase 1 proposal have been successfully achieved.

GeoThermalCloud is a flexible open-source cloud-based ML framework for geothermal exploration. ***GeoThermalCloud*** will also allow for the treatment of both public and proprietary datasets (a proposed task under Phase 2 of our project). ***GeoThermalCloud*** framework also includes a series of advanced pre-processing, post-processing, and visualization tools which tremendously simplify its application for real-world problems. These tools make the ML results understandable and visible even for non-experts. ML and subject-matter expertise are not a critical requirement to use our ML framework.

ML tools in ***GeoThermalCloud*** are based on a series of novel LANL-developed patented ML tools called ***SmartTensors*** (https://github.com/SmartTensors). ***SmartTensors*** have already been applied to solve a wide range of real-world problems (from COVID-19[12] to wildfires; http://tenosrs.lanl.gov). ***SmartTensors*** has been recently nominated for an R&D 100 award. ***SmartTensors*** is written in Julia programming language: a novel, fast (two orders of magnitude faster than Python, R, and MATLAB; https://julialang.org) language specifically designed for technical computing and machine learning. Julia programming language is actively used in numerous fields, including oil/gas exploration, financing, climate modeling and pharmaceutical research. We have active collaborations with developers of the Julia language. We will involve Julia Computing as a collaborator in Phase 2 of our project.

GeoThermalCloud, as well as our ***SmartTensors*** ML tools in general, are designed to be computationally efficient and scalable. Our ML analyses can be executed to diverse computational platforms: from handhelds and laptops to supercomputers and cloud frameworks. The ML tools are designed to automatically and efficiently utilize available hardware accelerators such as GPUs (Graphics Processing Units) and TPUs (Tensor Processing Units) and diverse computing platforms involving a large number of compute nodes. ***SmartTensors*** framework has already been demonstrated to process tens of TBs of data using DOE supercomputers[13].

ML methods embedded in the ***GeoThermalCloud*** have already been extensively tested and validated. In this report, as well as in a series of presentations and submitted research papers, we have demonstrated that ***GeoThermalCloud*** can be applied to discover hidden geothermal signatures for a series of diverse datasets and regions. Our work so far demonstrated the applicability of ***GeoThermalCloud*** to process diverse multi-source, multi-scale, and multi-physics geothermal datasets for the following sites: ***Southwest New Mexico***, ***Brady, Nevada***, ***Great Basin***, ***West Texas***, ***Tularosa Basin, New Mexico***, ***Tohatchi Springs, New Mexico***, ***Hawaii***, ***Utah FORGE***, and ***EGS Collab***. In addition, two synthetic datasets were generated using ***GeoDT*** and PFLOTRAN. They are applied to (1) validate our ML methods, (2) demonstrate how our ML tools can fuse model outputs in the geothermal analyses, (3) identify high-value data acquisition strategies, and (4) estimate exploration/production costs and risks. These two synthetic-data analysis efforts demonstrate how our ML methods can be applied in tandem with forward-looking multi-physics models to address questions related to value-added data acquisition strategies and to reduce exploratory and development costs. For the Brady and Utah FORGE sites, our ML characterization particularly focused on impact in-situ stress on geothermal production. For Utah FORGE, prospectivity maps are generated and drilling locations are proposed for future geothermal field exploration.

All the data and codes required to reproduce the ML results for SWNM, Brady, and Great Basin studies presented in this report are available at the GDR (https://gdr.openei.org/submissions/1297) and at the ***GeoThermalCloud*** GitHub repository (https://github.com/SmartTensors/GeoThermalCloud.jl). The data and the codes for other problem analyses performed under this project will be available shortly on the website as well.

5 Future Work

A series of tasks will be executed under proposed Phase 2 of our project that will further enhance the ***GeoThermalCloud*** capabilities:

1. ***GeoThermalCloud*** will allow for the treatment of both public and proprietary datasets. This is an essential feature considering the high sensitivities associated with the use of proprietary data in general. Processing simultaneously public and proprietary datasets will tremendously increase the quality and applicability of the obtained ML results. The work will be executed by Julia Computing; a new collaborator in Phase 2 of our project. We have already active collaborations with Julia Computing under several other projects. Julia Computing has already past experience in developing computing systems, which deal with both public and proprietary datasets.

2. ***GeoThermalCloud*** will be deployed at JuliaHub (https://juliahub.com). JuliaHub is maintained by Julia Computing. JuliaHub will allow for simplified cloud deployment and executions of various types of geothermal ML analyses available with ***GeoThermalCloud***.
3. Our ***GeoThermalCloud*** framework already includes a series of advanced pre-processing, post-processing, and visualization tools. These tools will be expanded to include GIS (Geographic Information System) based visualization capabilities. ***GeoThermalCloud*** already includes the open-source framework GMT (Generic Mapping Tools; https://www.soest.hawaii.edu/gmt). Additional open-source GIS tools will be added in the future (e.g., QGIS, https://qgis.org).
4. ML methods and tools in our ***GeoThermalCloud*** framework will be further enhanced and validated to extract geothermal information. Particular key areas for improvements include:
    a. Physics-Informed ML (PIML): Development of PIML methods and tools specific for geothermal exploration[29,30]. PIML methods have already been included in ***SmartTensors***. However, they need to be modified to address the challenges and needs related to geothermal resource exploration and production.
    b. Koopman theory: Develop efficient and optimized data-driven reduced order models[87,88] using recent advances in nonlinear dynamics[89,90] and Koopman theory[91–93] to model geothermal resources based on limited observational data and hidden features[94] discovered by our unsupervised ML methods.
    c. ML using limited and missing data: Here, we have demonstrated that our ML methods can successfully process small datasets with data gaps. The capabilities of our ML methods to process limited and missing data will be further enhanced incorporating PIML and Koopman theory. In addition, we will also explore novel data augmentation techniques such as few-shot learning[95] and meta-learning[96].
    d. Transfer learning: Our ML methods have already been successfully applied for transfer learning[2,4]. However, they will be further enhanced by incorporating PIML methods and by advancing our unsupervised ML methods based on matrix/tensor factorization.
5. We will further validate our novel LANL-developed simulator ***GeoDT***. This validation process enables us to apply ***GeoDT*** to study a wider range of geothermal production scenarios to identify crucial data needs for geothermal energy extraction. Further investigation of the ML extracted geothermal signatures is also required to more clearly understand the implications of physical site attributes for geothermal exploration and production.
6. Looker tool (https://looker.com) for improved understanding of data analytics and ML results has been recently added in the Google Cloud Platform toolbox. We will be working with the developers of the Looker tool at Google to incorporate it into our ***GeoThermalCloud*** framework.
7. Add additional datasets for ***GeoThermalCloud*** analysis. This will include assimilation of satellite data, GeoDAWN (Geoscience Data Acquisition for Western Nevada) dataset, physics simulation outputs, and open-source geothermal datasets. We will also process jointly datasets that are available in overlapping (e.g., Brady, Utah FORGE, GeoDAWN and Great Basin) and adjacent (e.g., SWNM, Tularosa, Tohatchi) regions.
8. Our already accomplished work for the Brady and Utah FORGE sites focused on ML characterization of impact in-situ stress on geothermal production. These analyses will be further expanded under Phase 2 of our project.
9. The test and example problems in ***GeoThermalCloud*** can already be executed as Jupyter and Pluto notebooks. The utilization of notebooks will be expanded to incorporate GIS and interactive visualization tools. Interactive plots (using Plotly.js https://plotly.com/javascript) are already implemented. Other interactive tools such as Dash (https://plotly.com/dash) and D3 (https://d3js.org) will be added as well. This work will further facilitate the interpretation of ML results.
10. The data and codes required to reproduce all ML analyses will be available at the GDR (https://gdr.openei.org/submissions/1297) and at the ***GeoThermalCloud*** GitHub repository (https://github.com/SmartTensors/GeoThermalCloud.jl). The data and codes will be also available for execution at JuliaHub (https://juliahub.com).

6 REFERENCES

1. Vesselinov, V. V., Mudunuru, M. K., Ahmmed, B., S, K., & Middleton R.S. Discovering signatures of hidden geothermal resources based on unsupervised learning. *45th Annu. Stanf. Geotherm. Workshop* (2020).
2. Vesselinov, V. V., Ahmmed, B., Mudunuru, M. K., Karra, S. & Middleton, R. S. Hidden geothermal signatures of the southwest New Mexico. in *Proceedings of the World Geothermal Congress, Reykjavik, Iceland* (2021).
3. Vesselinov, V. V. *et al.* Discovering Hidden Geothermal Signatures using Unsupervised Machine Learning. *Geothermics* (in review).
4. Ahmmed, B., Vesselinov, V. V., Mudunuru, M. K., Middleton, R. S. & Karra, S. Geochemical characteristics of low-, medium-, and hot-temperature geothermal resources of the Great Basin, USA. in *World Geothermal Congress, Reykjavik, Iceland* (2021).
5. Ahmmed, B. Machine learning to characterize regional geothermal reservoirs in the western USA. (2020).
6. Siler, D. L., Pepin, J. D., Vesselinov, V. V., Mudunuru, M. K. & Ahmmed, B. Machine learning to identify geologic factors associated with production in geothermal fields: A case-study using 3D geologic data, Brady geothermal field, Nevada. *Geotherm. Energy* (in review).
7. Vesselinov, V. V., Ahmmed, B. & M.K, Mudunuru. *Unsupervised machine learning to discover attributes that characterize low, moderate, and high-temperature geothermal resources*. (2020).
8. Bennett, C. R. & Nash, G. D. *The Convergence of Heat, Groundwater & Fracture Permeability: Innovative Play Fairway Modelling Applied to the Tularosa Basin*. (2017).
9. Ahmmed, B. & Vesselinov, V. V. *Geothermal resource analysis at Tohatchi hot springs, New Mexico*. (2020).
10. Ahmmed, B., Lautze, N., Vesselinov, V. V., Dores, D. & Mudunuru, M. K. *Unsupervised machine learning to extract dominant geothermal attributes in Hawaii Island Play Fairway data*. (2020).

11. Ahmmed, B., Vesselinov, V. V. & M.K, Mudunuru. *Non-negative matrix factorization to discover dominant attributes in Utah FORGE Data*. (Geothermal Resources Council, 2020).

12. Vesselinov, V. V., Middleton, R. S. & Talsma, C. J. *COVID-19: Spatiotemporal social data analytics and machine learning for pandemic exploration and forecasting*. 9 (2021).

13. Chennupati, G., Vangara, R., Skau, E., Djidjev, H. & Alexandrov, B. Distributed non-negative matrix factorization with determination of the number of latent features. *J. Supercomput.* **76**, 7458–7488 (2020).

14. Alexandrov, B. S. & Vesselinov, V. V. Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization. *Water Resour. Res.* **50**, 7332–7347 (2014).

15. Vesselinov, V. V., Mudunuru, M. K., Karra, S., O'Malley, D. & Alexandrov, B. S. Unsupervised machine learning based on non-negative tensor factorization for analyzing reactive mixing. *J. Comput. Phys.* **395**, 85–104 (2019).

16. Iliev, F. L., Stanev, V. G., Vesselinov, V. V. & Alexandrov, S., B. Nonnegative matrix factorization for identification of unknown number of sources emitting delayed signals. *PloS One* **13**, e0193974 (2018).

17. Everett, B. *An introduction to latent variable models*. (Springer Science & Business Media, 2013).

18. Haykin, S. & Chen, Z. The cocktail party problem. *Neural Comput.* **17**, 1875–1902 (2005).

19. Parsons, L., Haque, E. & Liu, H. Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explor. Newsl.* **6**, 90–105 (2004).

20. Constantine, P. G. *Active subspaces: Emerging ideas for dimension reduction in parameter studies*. (SIAM, 2015).

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

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

23. Huang, S., Papernot, N., Goodfellow, I., Duan, Y. & Abbeel, P. Adversarial attacks on neural network policies. *ArXiv Prepr. ArXiv170202284* (2017).

24. Goodfellow, I. J. *et al.* Challenges in representation learning: A report on three machine learning contests. in *International conference on neural information processing* (eds. Lee, M., Hirose, A., Hou, Z. & Kil, R. M. ") 117–124 (Springer, 2013).

25. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? *ArXiv Prepr. ArXiv14111792* (2014).

26. Gu, J. *et al.* Recent advances in convolutional neural networks. *Pattern Recognit.* **77**, 354–377 (2018).

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

28. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

29. Chen, R. T. Q., Rubanova, Y., Bettencourt, J. & Duvenaud, D. Neural ordinary differential equations. *ArXiv Prepr. ArXiv180607366* (2018).

30. Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).

31. Klema, V. & Laub, A. The singular value decomposition: Its computation and some applications. *IEEE Trans. Autom. Control* **25**, 164–176 (1980).

32. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **2**, 37–52 (1987).

33. Comon, P. Independent component analysis, a new concept? *Signal Process.* **36**, 287–314 (1994).

34. Hartigan, J. A. & Wong, M. A. A K-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **28**, 100–108 (1979).

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

36. Böttcher, A. & Wenzel, D. The Frobenius norm and the commutator. *Linear Algebra Its Appl.* **429**, 1864–1885 (2008).

37. Vesselinov, V. V., Alexandrov, B. S. & O'Malley, D. Contaminant source identification using semi-supervised machine learning. *J. Contam. Hydrol.* **212**, 134–142 (2018).

38. Ahmmed, B. & Vesselinov, V. V. Prospectivity Analyses of the Utah FORGE Site using Unsupervised Machine Learning. in *Geothermal Rising, San Diego, CA* (2021).

39. Bielicki, J. *et al.* Hydrogeologic windows and estimating the prospectivity of geothermal resources. in *Proceedings, 41st Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California, February* 22–24 (2016).

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

41. Bielicki, J. *et al. Hydrogeolgic windows: Regional signature detection for blind and traditional geothermal play fairways, Final Reports, Los Alamos National Laboratory*. (2015).

42. Kelley, S. Geothermal Energy, Lite Geology. *N. M. Bur. Geol. Miner. Resour. Div. N. M. Tech* (2010).

43. Levitte, D. & Gambill, D. T. *Geothermal potential of west-central New Mexico from geochemical and thermal gradient data*. (1980).

44. USGS. National Water Information System. (2018).

45. Keller, G. R. *et al.* A comparative study of the Rio Grande and Kenya rifts. *Tectonophysics* **197**, 355–371 (1991).

46. Person, M. *et al.* Assessment of the sustainability of geothermal development within the Truth or Consequences hot-springs district. *New Mexico: New Mexico Bureau of Geology and Mineral Resources Open-File Report* **551**, (2013).

47. USGS. Energy and Environment in the Rocky Mountain Area. (2018).

48. UNM. *Digital Geologic Map of New Mexico – Volcanic Vents*. (Earth Data Center, University of New Mexico, 2018).

49. Nakai, J. S., Sheehan, A. F. & Bilek, S. L. Seismicity of the rocky mountains and Rio Grande Rift from the EarthScope Transportable Array and CREST temporary seismic networks, 2008–2010. *J. Geophys. Res. Solid Earth* **122**, 2173–2192 (2017).

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

51. McIntosh, W. C., Geissman, J. W., Chapin, C. E., Kunk, M. J. & Henry, C. D. Calibration of the latest Eocene-Oligocene geomagnetic polarity time scale using $^{40\$}Ar/^{39\$}Ar$ dated ignimbrites. *Geology* **20**, 459–463 (1992).

52. Chapin, C. E., Wilks, M., McIntosh, W. C. & Cather, S. M. Space-time patterns of Late Cretaceous to present magmatism in New Mexico—Comparison with Andean volcanism and potential for future volcanism. *N. M. Bur. Geol. Miner. Resour. Bull.* **160**, 13–40 (2004).

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

54. Elston, W. E., Rhodes, R. C., Coney, P. J. & Deal, E. G. *Progress Report on the Mogollon Plateau Volcanic Field Southwestern New Mexico, no. 3—Surface Expression of a Pluton*. *Cenozoic volcanism in southwestern New Mexico: New Mexico Geological Society Special Publication* vol. 5 3–28 (1976).

55. Sanford, A. R., Lin, K., Tsai, I. & Jaksha, L. H. Earthquake catalogs for New Mexico and bordering areas: 1869–1998. *N. M. Bur. Geol. Miner. Resour. Circ.* **210**, 1–9 (2002).

56. Hunt, C. B. Cenozoic Geology of the Colorado Plateau. *US Geol. Surv. Prof. Pap.* (1956).

57. Thompson, G. A. & Zoback, M. L. Regional geophysics of the Colorado Plateau. *Tectonophysics* **61**, 149–181 (1979).

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

59. Goff, F., Bergfeld, D. h & Janik, C. J. *Geochemical data on waters, gases, scales, and rocks from the Dixie Valley region, Nevada (1996-1999)*. (2002).

60. Zehner, R. E., Coolbaugh, M. F. & Shevenell, L. Regional groundwater geochemical trends in the

Great Basin: implications for geothermal exploration. *Geotherm. Resour. Counc. Trans.* **30**, 117–124 (2006).

61. Caine, J. S. & Forster, C. B. Fault zone architecture and fluid flow: Insights from field data and numerical modeling. *Geophys. Monogr.-Am. Geophys. Union* **113**, 101–128 (1999).

62. Caine, J. S., Evans, J. P. & Forster, C. B. Fault zone architecture and permeability structure. *Geology* **24**, 1025–1028 (1996).

63. Fairley, J. P. & Hinds, J. J. Rapid transport pathways for geothermal fluids in an active Great Basin fault zone. *Geology* **32**, 825–828 (2004).

64. Fairley, J., Heffner, J. & Hinds, J. Geostatistical evaluation of permeability in an active fault zone. *Geophys. Res. Lett.* **30**, (2003).

65. Sanderson, D. J. & Zhang, X. Stress-controlled localization of deformation and fluid flow in fractured rocks. *Geol. Soc. Lond. Spec. Publ.* **231**, 299–314 (2004).

66. Faulds, J., Moeck, I., Drakos, P. & Zemach, E. Structural assessment and 3D geological modeling of the Brady's geothermal area, Churchill county (Nevada, USA): A preliminary report. in *Proceedings, thirty-fifth workshop on geothermal reservoir engineering, Stanford: Stanford University* (2010).

67. Siler, D. L. & Faulds, J. E. *Three-dimensional geothermal fairway mapping: examples from the western Great Basin, USA*. (2013).

68. Siler, D. L., Hinz, N. H., Faulds, J. E. & Queen, J. 3D analysis of geothermal fluid flow favorability: Brady's, Nevada, USA. in *Proceedings, forty-first workshop on geothermal reservoir engineering. Stanford: Stanford University* (2016).

69. Faulds, J. E. *et al.* Preliminary geologic map of the Bradys geothermal area, Churchill County, Nevada. *Nev. Bur. Mines Geol. Open-File Rep.* 17–4 (2017).

70. Siler, D. L., Hinz, N. H. & Faulds, J. E. Stress concentrations at structural discontinuities in active fault zones in the western United States: Implications for permeability and fluid flow in geothermal fields. *Bulletin* **130**, 1273–1288 (2018).

71. Siler, D. L., Faulds, J. E., Hinz, N. H. & Queen, J. H. *Three-dimensional geologic map of the Brady geothermal area, Nevada*. (2021).

72. Witter, J. B., Siler, D. L., Faulds, J. E. & Hinz, N. H. 3D geophysical inversion modeling of gravity data to test the 3D geologic model of the Bradys geothermal area, Nevada, USA. *Geotherm. Energy* **4**, 14 (2016).

73. Morris, A., Ferrill, D. A. & Henderson, D. B. Slip-tendency analysis and fault reactivation. *Geology* **24**, 275–278 (1996).

74. Ferrill, D. A. *et al.* Stressed rock strains groundwater at Yucca Mountain, Nevada. *GSA Today* **9**, 1–8 (1999).

75. Jolie, E., Moeck, I. & Faulds, J. E. Quantitative structural–geological exploration of fault-controlled geothermal systems—a case study from the Basin-and-Range Province, Nevada (USA). *Geothermics* **54**, 54–67 (2015).

76. Siler, D. L., Pepin, J. D., Vesselinov, V. V., Mudunuru, M. K. & Ahmmed, B. Machine learning to identify geologic factors associated with production in geothermal fields: A case study using 3D geologic data, Brady geothermal field, Nevada. *Geotherm. Energy* (2021).

77. Peacock, D. C. P. & Sanderson, D. J. Displacements, segment linkage and relay ramps in normal fault zones. *J. Struct. Geol.* **13**, 721–733 (1991).

78. Fossen, H. & Rotevatn, A. Fault linkage and relay structures in extensional settings—A review. *Earth-Sci. Rev.* **154**, 14–28 (2016).

79. Shevenell, L., Oppliger, G., Coolbaugh, M. & Faulds, J. Bradys (Nevada) InSAR anomaly evaluated with historical well temperature and pressure data. *Geotherm. Resour. Counc. Trans.* **36**, 1383–1390 (2012).

80. Frash, L. P. Geothermal Design Tool (GeoDT). in *Proceedings of the 46th Workshop on Geothermal Reservoir Engineering Stanford University, Stanford, February 15-17* (2021).

81. Frash, L. P., Welch, N. J., Meng, M., Li, W. & Carey, J. W. A Scaling Relationship for Fracture Permeability after Slip. in *Proceedings of the 55th US Rock Mechanics/Geomechanics Symposium,*

Houston, TX, June (2021).

82. Pepin, J. *et al.* Deep fluid circulation within crystalline basement rocks and the role of hydrologic windows in the formation of the Truth or Consequences, New Mexico low-temperature geothermal system. *Geofluids* **15**, 139–160 (2015).

83. Mudunuru, M. K. *et al.* Site-scale and regional-scale modeling for geothermal resource analysis and exploration. *45th Annu. Stanf. Geotherm. Workshop* (2020).

84. Gardner, M. W. & Dorling, S. R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos. Environ.* **32**, 2627–2636 (1998).

85. Bardenet, R., Brendel, M., Kégl, B. & Sebag, M. Collaborative hyperparameter tuning. in *International Conference on Machine Learning* 199–207 (PMLR, 2013).

86. Zhou, Z. & Tartakovsky, D. M. Markov chain Monte Carlo with neural network surrogates: application to contaminant source identification. *Stoch. Environ. Res. Risk Assess.* **35**, 639–651 (2021).

87. Chorin, A. J. & Lu, F. Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics. *Proc. Natl. Acad. Sci.* **112**, 9804–9809 (2015).

88. Gilani, F., Giannakis, D. & Harlim, J. Kernel-based Prediction of Non-Markovian Time Series. *Phys. Nonlinear Phenom.* **418**, 132829 (2021).

89. Zhao, Z. & Giannakis, D. Analog forecasting with dynamics-adapted kernels. *Nonlinearity* **29**, 2888 (2016).

90. Alexander, R. & Giannakis, D. Operator-theoretic framework for forecasting nonlinear time series with kernel analog techniques. *Phys. Nonlinear Phenom.* **409**, 132520 (2020).

91. Budišić, M., Mohr, R. & Mezić, I. Applied koopmanism. *Chaos Interdiscip. J. Nonlinear Sci.* **22**, 047510 (2012).

92. Mezić, I. Analysis of fluid flows via spectral properties of the Koopman operator. *Annu. Rev. Fluid Mech.* **45**, 357–378 (2013).

93. Klus, S., Koltai, P. & Schütte, C. On the numerical approximation of the Perron-Frobenius and Koopman operator. *J. Comput. Dyn.* **3**, 51–79 (2016).

94. Comeau, D., Giannakis, D., Zhao, Z. & Majda, A. J. Predicting regional and pan-Arctic sea ice anomalies with kernel analog forecasting. *Clim. Dyn.* **52**, 5507–5525 (2019).

95. Generalizing from a Few Examples: A Survey on Few-shot Learning: ACM Computing Surveys: Vol 53, No 3. https://dl.acm.org/doi/abs/10.1145/3386252.

96. Peng, H. A Comprehensive Overview and Survey of Recent Advances in Meta-Learning. *ArXiv200411149 Cs Stat* (2020).

97. Bielicki, J. *et al. Hydrogeologic Windows: Regional Signature Detection for Blind and Traditional Geothermal Play Fairways, Final Report, Los Alamos National Laboratory*. (2015).