PROCEEDINGS, 47th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 7-9, 2022
SGP-TR-223

GeoThermalCloud: Machine Learning for Discovery, Exploration, and Development of Hidden Geothermal Resources

Velimir V. Vesselinov[1], Bulbul Ahmmed[1], Luke Frash[1] and Maruti K. Mudunuru[2]
[1]Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[2]Watershed & Ecosystem Science, Pacific Northwest National Laboratory, Richland, WA 99352, USA
Correspondence: vvv@lanl.gov, ahmmedb@lanl.gov

Keywords: GeothermalCloud, geothermal resources exploration, unsupervised machine learning, non-negative matrix factorization, feature extraction, hidden geothermal signatures.

Abstract

Discovery, exploration, and development of hidden geothermal resources have many risks and challenges because of the complex and uncertain subsurface conditions. To mitigate these risks, we have developed a tool called GeoThermalCloud, which utilizes unsupervised machine learning (ML) and physics-informed machine learning (PIML) methods to process the data and guide geothermal exploration and development efficiently. The unsupervised ML automates the data analyses and interpretations by extracting hidden signatures (features) characterizing geothermal resources/exploration/development. It also enables practitioners to identify observations that are important to represent the discovered hidden signatures. In addition to data, PIML adds physical constraints such as mass balance, constitutive relationships, and models, in the ML processes to characterize hidden geothermal resources better. GeoThermalCloud capabilities include (1) analyzing large field datasets, (2) assimilating model simulations (large inputs and outputs), (3) processing sparse datasets, (4) performing transfer learning (between sites with different exploratory levels), (5) extracting hidden geothermal signatures in the field and simulation data, (6) labeling geothermal resources and processes, (7) identifying high-value data acquisition targets, and (8) guiding geothermal exploration and production by selecting optimal exploration, production, and drilling strategies. The GeoThermalCloud is an open-source tool available at https://github.com/SmartTensors/GeoThermalCloud.jl (a part of our SmartTensors framework; http://tensors.lanl.gov, https://github.com/SmartTensors) We have used GeoThermalCloud on ten geothermal datasets, including a large and sparse dataset of the Great Basin, and all of them show promising results. Most of the data and analyses are available on GitHub as well. Obtained results can be reproduced and further expanded by adding additional data. Practitioners and researchers are welcome to utilize GeoThermalCloud to solve other geothermal problems.

1. Introduction

Geothermal community often utilizes a diverse set of attributes/parameters for geothermal resource exploration, geothermal field development, and geothermal power production rather than using only a set of attributes. For geothermal resource exploration, they may use surface exposures (e.g., springs) in combination with shallow water chemistry (e.g., anions, cations, tracer elements), geophysics attributes (e.g., gravity, magnetic, seismic), geologic attributes (e.g., fault, fault density, dike/dyke), geothermal attributes (e.g., thermal gradient, heat flow). There are not a set of attributes for geothermal exploration like in oil/gas field for various reasons. Often each geothermal field has unique geological characteristics that make the discovery of geothermal resources is challenging. Furthermore, processes and parameters impacting geothermal conditions are poorly understood. It is even more challenging to develop a geothermal field because it often requires too many well drillings, and the cost of well installation is very high. 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.

To facilitate geothermal exploration and production, we developed and applied our novel Los Alamos National Laboratory (LANL)-developed 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**. It is an open-source cloud-based ML framework for geothermal exploration, geothermal play development, and geothermal power production. It can fuse geothermal datasets and multi-physics codes. Datasets can range from small to big datasets; however, to our best knowledge, this is the best tool available in the market to deal with small datasets and data with missing values. Moreover, it can simultaneously handle both public and proprietary datasets keeping the sensitivities of private data hidden. This increases the quality and applicability of the obtained ML results. Additionally,

GeoThermalCloud has in-build preprocessing, postprocessing, and state-of-the-art visualization tools for non-experts. Therefore, both experts and non-experts can equally utilize this tool without going through steep learning curve.

GeoThermalCloud has been used to analyze 10 datasets including eight real/field and two synthetic datasets. Here, because of space constraints, we provide a glimpse of each dataset we analyzed and explain three datasets in brief. Also, we provide the capability of **GeoThermalCloud** and

2. GeothermalCloud Capability

GeoThermalCloud utilizes *SmartTensors*, which is an open-source, LANL-developed framework of patented ML methods and computational tools (http://tensors.lanl.gov, https://github.com/SmartTensors). *SmartTensors* is a toolbox for unsupervised and physics-informed ML based on matrix/tensor factorization constrained by penalties enforcing robustness and interpretability (e.g., nonnegativity; physics and mathematical constraints; etc.). It can also utilize hardware accelerators such as graphical and tensor processing (GPU and TPU) units to make computing faster. *SmartTensors* has already been successfully applied to analyze diverse datasets related to a wide range of problems, from COVID-19 (Vesselinov, Middleton, and Talsma 2021) to wildfires and text mining. The two most commonly used ML algorithms in *SmartTensors* are NMF*k* (https://github.com/SmartTensors/NMFk.jl) and NTF*k* (https://github.com/SmartTensors/NTFk.jl). They perform nonnegative matrix/tensor factorization coupled with customized *k*-means clustering (Alexandrov and Vesselinov 2014; Vesselinov et al. 2019; Iliev et al. 2018). 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. 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(Haykin and Chen 2005) 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(Parsons, Haque, and Liu 2004; Constantine 2015) 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. Detail descriptions of NMFk and NTFk are available at (Alexandrov and Vesselinov 2014; Vesselinov et al. 2019; Iliev et al. 2018). Another important tool, PIML, is also available in **GeoThermalCloud (**https://github.com/SmartTensors/GeoThermalCloud.jl**)**. Through PIML, users can utilize any physics code during the training phase of ML models.

3. Example Datasets

ML methods embedded in the **GeoThermalCloud** have been extensively tested and validated against various kinds of datasets (cite GTCloud report). Outputs of these applications have been published in a series of presentations, conference papers and peer-reviewed papers. The analyzed ML applications are:

1. **Southwest New Mexico (SWNM)**: Here, we analyzed 18 attributes at 44 locations and 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 (Vesselinov, Ahmmed, et al. 2021; Vesselinov et al. 2020; in review).
2. **Great Basin**: In this dataset, we analyzed 18 shallow water chemistry attributes at 14,342 locations. This work extracted hidden geothermal signatures associated with low-, medium-, high-temperature hydrothermal systems, their dominant characterization attributes, and spatial distribution within the study area (Ahmmed 2020; Ahmmed et al. 2021). The analyses are based on the public data available at the Nevada Bureau of Mines and Geology website.
3. **Brady site, Nevada**: We identified key geologic factors controlling geothermal production in the Brady geothermal field. Please see (Siler et al. 2021) for more details.
4. **Tularosa Basin, New Mexico**: Analyzed 21 Play Fairway Analysis (PFA) attributes at 120 locations (Vesselinov 2020); data comes from past PFA work in this region (Bennett and Nash 2017). ML analyses identified geothermal signatures associated with low-, medium-, and high-temperature hydrothermal systems. Dominant attributes and spatial distribution of the geothermal signatures were also defined.
5. **Tohatchi Springs, New Mexico**: Explored 19 geothermal attributes at 43 locations in Tohatchi Springs, New Mexico (Ahmmed, Vesselinov, and Middleton 2020). Successfully defined geothermal signatures associated with low- and medium-temperature hydrothermal systems. Also, we found their dominant attributes, and spatial distribution.
6. **Hawaii**: Analyzed four islands' data separately and jointly; ML identified low-, medium-, and high-temperature hydrothermal systems and their dominant characterization attributes (Ahmmed et al. 2020).
7. **Utah FORGE**: Performed prospectivity analysis to identify future drilling locations using geological, geochemical, and geophysical attributes (Ahmmed and Vesselinov 2021). Maps of temperature at depth, and heat flow are 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.
8. **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. This work has not been published yet.

9. **GeoDT synthetic dataset:** GeoDT, a novel LANL-developed multi-physics code for predicting the performance of geothermal energy systems. GeoDT evaluates how geothermal site data conditions impact design decisions related to the 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 (Vesselinov, Frash, et al. 2021). Our study focused on stress states and natural fractures on geothermal well drilling and well production. ML analyses identified well spacing and well orientation as critical parameters impacting energy production and induced seismicity.
10. **Thermo-hydro-chemical synthetic dataset**: Also, this tool was used to predict synthetic thermo-hydro-chemical states. The LANL simulator PFLOTRAN (Lichtner et al. 2015) was used to simulate a 3-D thermo-hydro-chemical (THC) model. The model simulates heat and mass transport and predicts the spatiotemporal distribution of temperature, B+, and Li+ concentrations in the subsurface. **GeoThermalCloud** was used to predict THC data faster than the PLFOTRAN simulation.

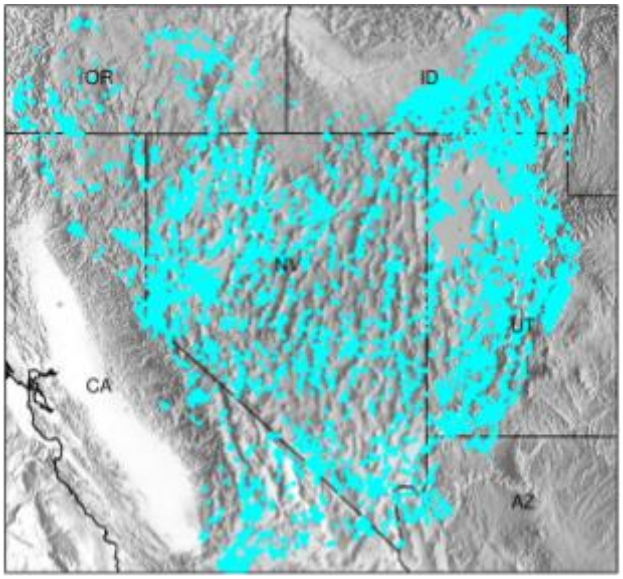

Figure 4.1.1: Geothermal data locations in the Great Basin.

4. RESULTS

This section provides a brief description of three analyses that we performed using **GeoThermalCloud**. The datasets are Great Basin, State of the stress of Brady geothermal site, and GeoDT synthetic dataset.

4.1 Great Basin

This case study showcases four important capabilities of **GeothermalCloud** that are (1) handling missing/sparse data, (2) characterizing geothermal resource types, (3) identifying critical attributes for different types of geothermal resources, and (4) reconstructing continuous data from sparse with quantified uncertainty.

The Great Basin covers Nevada, and 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 (Figure 4.1.1). Plenty of data have been collected over several decades 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 shallow water geochemical attributes (water cations/anions) and groundwater temperature are observed (Goff, Bergfeld, and Janik 2002; Zehner, Coolbaugh, and Shevenell 2006). 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 4.1.1 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

Table 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

missing data column in the table indicates that the dataset is heavily sparsed. Here, we applied the **GeoThermalCloud** ML methods to analyze these 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 ran, 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. Figure 4.1.2(a) demonstrates the attribute matrix of the optimal NMFk solution; the attribute matrix depicts the importance of attributes to represent extracted signatures. Next,

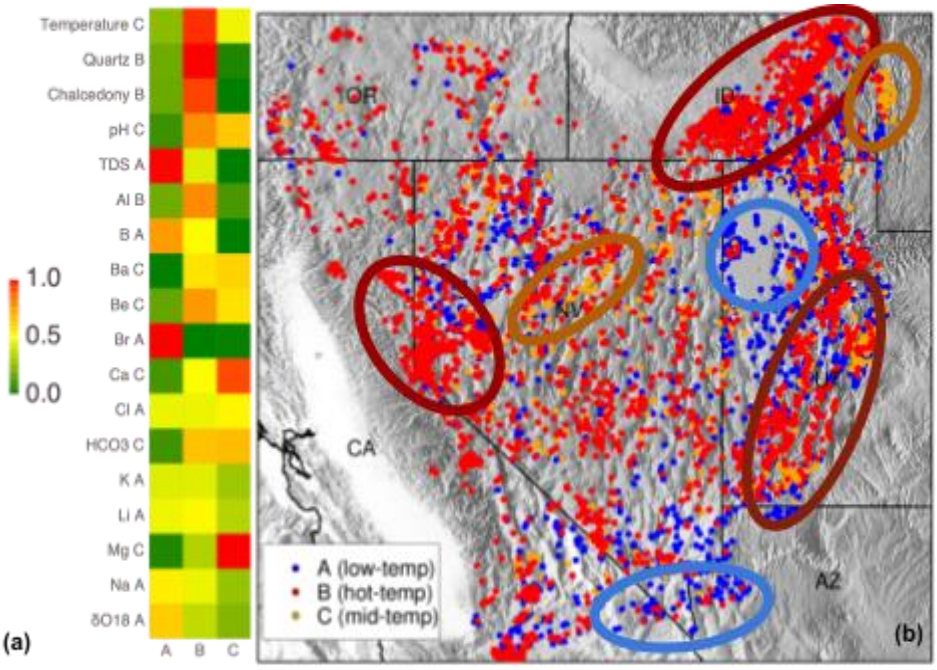

Figure 4.1.2: Optimal hidden geothermal signatures (a) and their spatial distribution (b); ellipses mark areas with high-density of signature locations of similar type.

we defined 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, 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+}.

The spatial distribution of each signature is shown in Figure 4.1.2(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 Figure 4.1.3 further demonstrate the spatial distribution of the extracted geothermal signatures.

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. Yet, our ML methodology estimates a continuous spatial

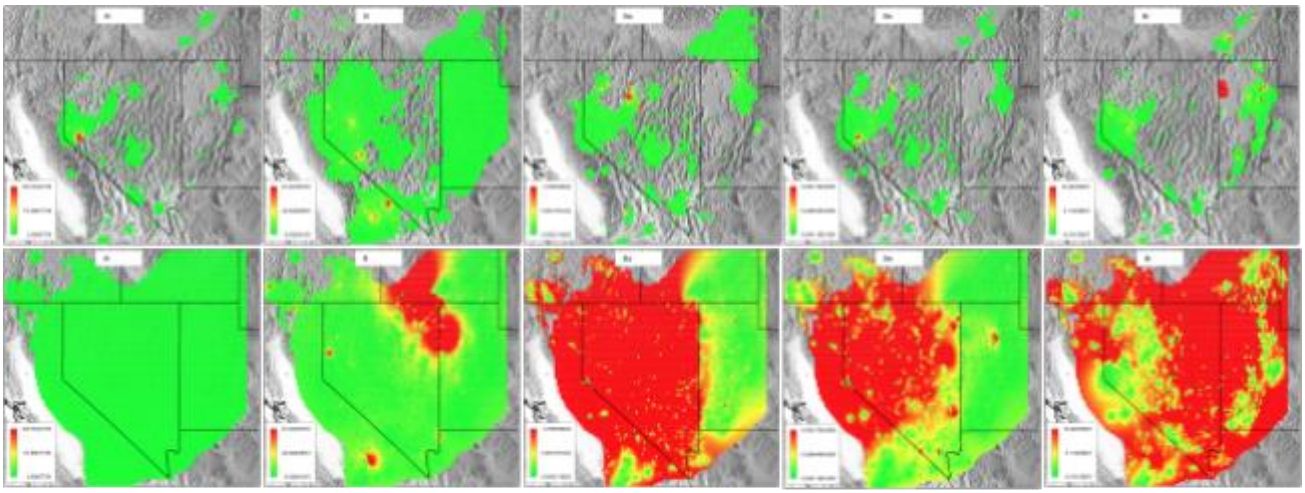

Figure 4.1.3: Maps of the spatial distribution of actual data (top) and corresponding ML reconstruction (bottom) of Al^{3+}, B^+, Ba^{2+}, Be^{2+}, and Br^-.

distribution for ***Signature A***. Similarly, the dominant attributes of ***Signature B*** and ***C*** are also sparse. 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). After learning the mapping function among all attributes and generating the signature mappings (Figure 4.1.3), our ML algorithm can estimate a continuous distribution of all the attributes, including $\partial^{18}O$ (Figure 4.1.3). 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 the 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.

Table 4.1.2: 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 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. 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.

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 attributes that will help analyze other geological/geophysical/geothermal attributes with geochemical attributes. 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).

4.2 State of the Stress of Brady Geothermal Site

This case study shows the capability of analyzing complex and big datasets; yet, finding a good mapping function in the dataset.

State of stress: State of stress in the Brady geothermal site is a critical factor determining geothermal production (Siler et al. 2021). 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 Figure 4.2.1. All these seven stress cases are used together with other geologic attributes to create seven alternative datasets. Each dataset is represented by 3D data tensors with dimensions equivalent to the analyzed datasets. The only differences between the seven tensors are in the three stress attributes: *dilation*, *normal stress*, and *coulomb shear*. The goal of the ML analyses is to select which of these seven 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 three stress attributes (*dilation*, *normal stress*, and *coulomb shear*) and the remaining 11 geothermal attributes. Estimates of the location matrix under the seven different stress ratios are shown in Figure 4.2.2. The production wells, injection, and non-productive wells have similar associations with the extracted signatures for the seven 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.

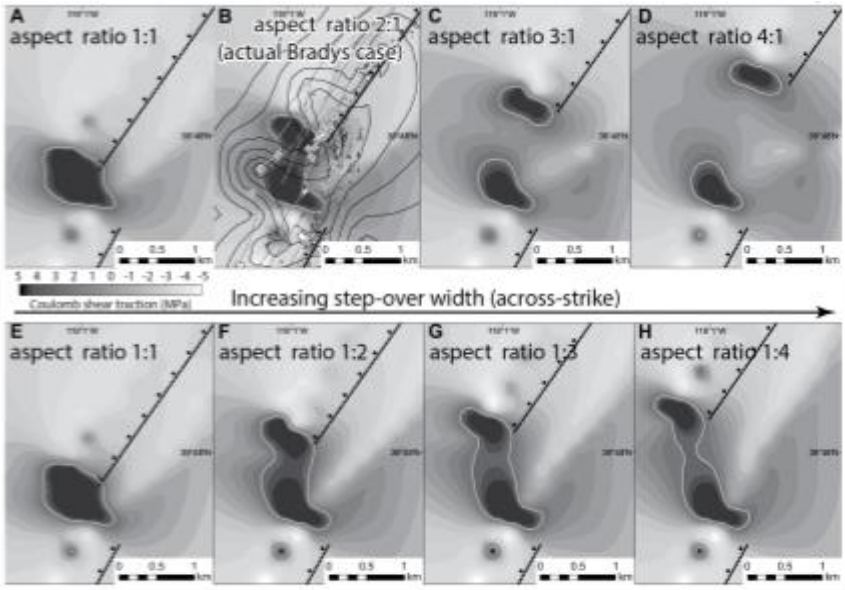

Figure 4.2.1: Coulomb shear traction estimated at 1000 m depth. Dark colors show high Coulomb shear traction on optimally oriented normal faults caused by slip. An aspect ratio of 2:1 most probably represents the state of stress at the Brady case (after Siler et al., 2018)

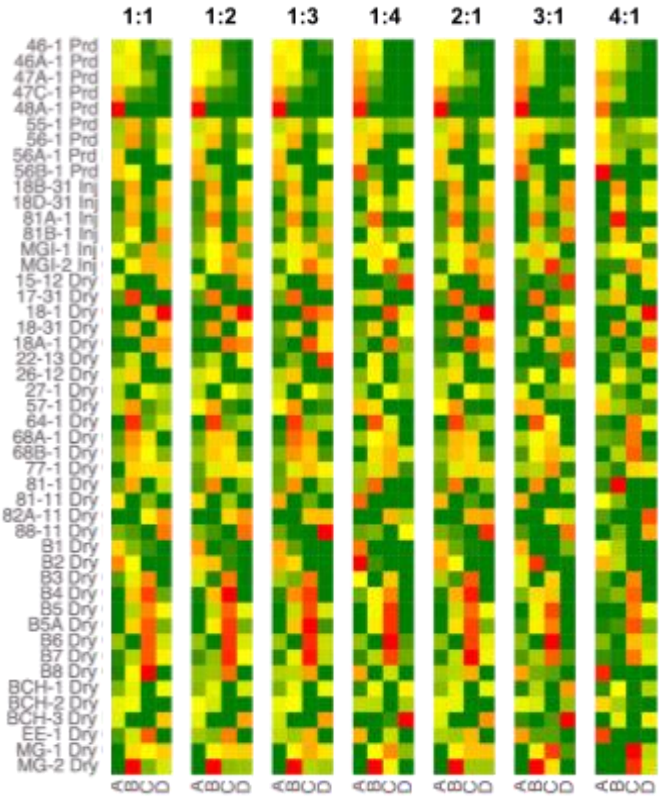

Figure 4.2.2: Estimates for the location matrix under 7 different stress ratios. The production wells (first 9 rows), injection (nest 6 rows), and non-productive (dry; remaining rows) wells have similar associations with the extracted signatures for the 7 different stress scenarios.

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 (Siler et al. 2021). 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).

4.3 GeoDT synthetic dataset

This case study shows the capability of finding relationships among numerous attributes in a big dataset. Also, it finds critical attributes defining geothermal power production in an enhanced geothermal system.

Our rapid multi-physics GeoDT model (Frash 2021) 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 (Frash et al., 2021). The site-specific parameters from the UtahFORGE site used for the model are given in (Vesselinov, Frash, et al. 2021). Each modeled scenario included stochastically generated natural fractures. An example system is visualized in Figure 4.3.1. To solve this system, GeoDT completes the following computational sequence:

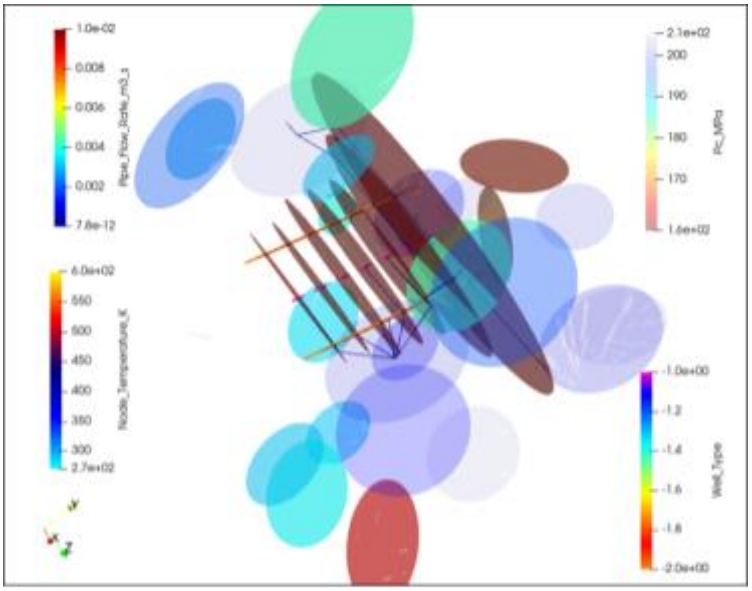

Figure 4.3.1: Example stochastically generated fracture and well scenario with injection into one well across seven isolated intervals and production from two bounding wells. The parallel hydraulic fractures propagated from each injection interval are shown in red, the color indicating that these fractures require relatively low pressure for activation (*Pc*). Note that most, but not all, of the scattered natural fractures require significantly higher pressures to activate.

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 (Figure 4.3.2) 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 Figure 4.3.3. 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 Figure 4.3.4. Based on this result, the dominant attribute in each signature is identified by the largest numbers (marked with red boxes in Figure 4.3.4) in each signature. We can use these dominant attributes to categorize the signatures into combining (1) well spacing and other attributes, (2) stress and other attributes, (3) system (i.e., site conditions) and other attributes, and (4) well dip (i.e., orientation) and 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 (Figure 4.3.4) on other outputs such as maximum induced seismic magnitude, far-field leakoff, and the number of fractures that interlink the injection and production wells (Figure 4.3.5). 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.

In conclusion, our ML analyses of GeoDT simulations focused on the influence of stress states and natural fractures on geothermal well drilling and well production. ML analyses identified well spacing and well orientation as critical parameters impacting energy production and induced seismicity. Our results also support the idea that "fracture caging" and "well caging" can limit induced seismic event magnitudes. "Caging" aims to optimize the drilling of injection and production wells so that they can contain the circulated fluids within a portion of the reservoir where fracture-dominated flow occurs.

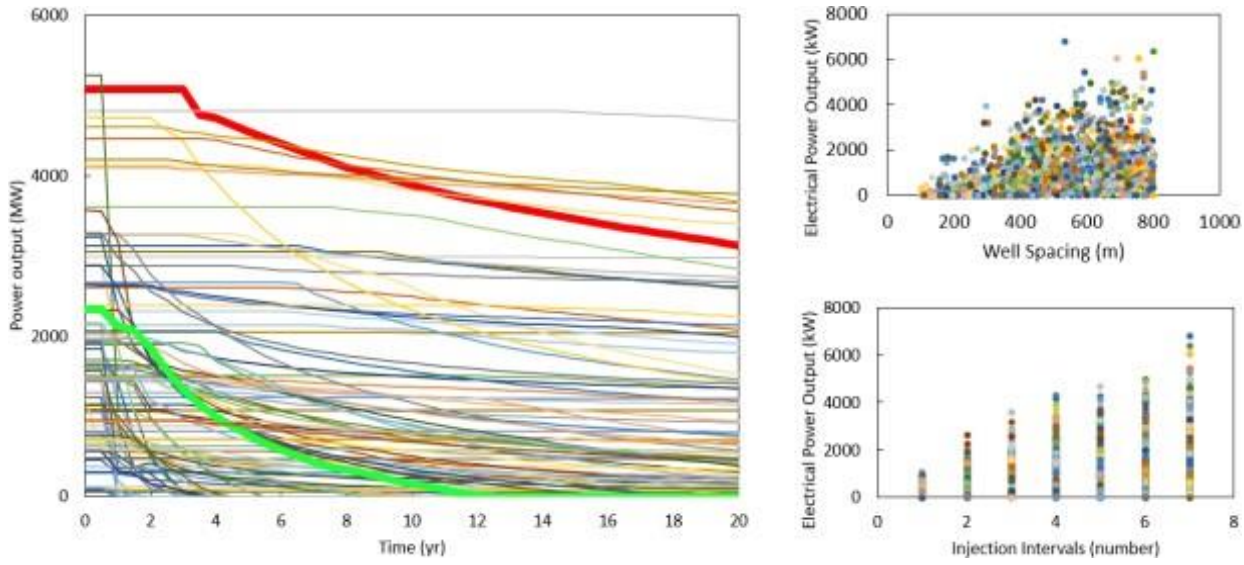

Figure 4.3.2: Compiled results from more than 2000 geothermal power production simulations based on the parameters described in Table 3.8.1. In the time series plot, a high-performing case is highlighted in red, and a poor performer is highlighted in green. There is also a clear link between the well spacing and power output in addition to the number of injection intervals (i.e., isolated zones) and power output (plots on the right).

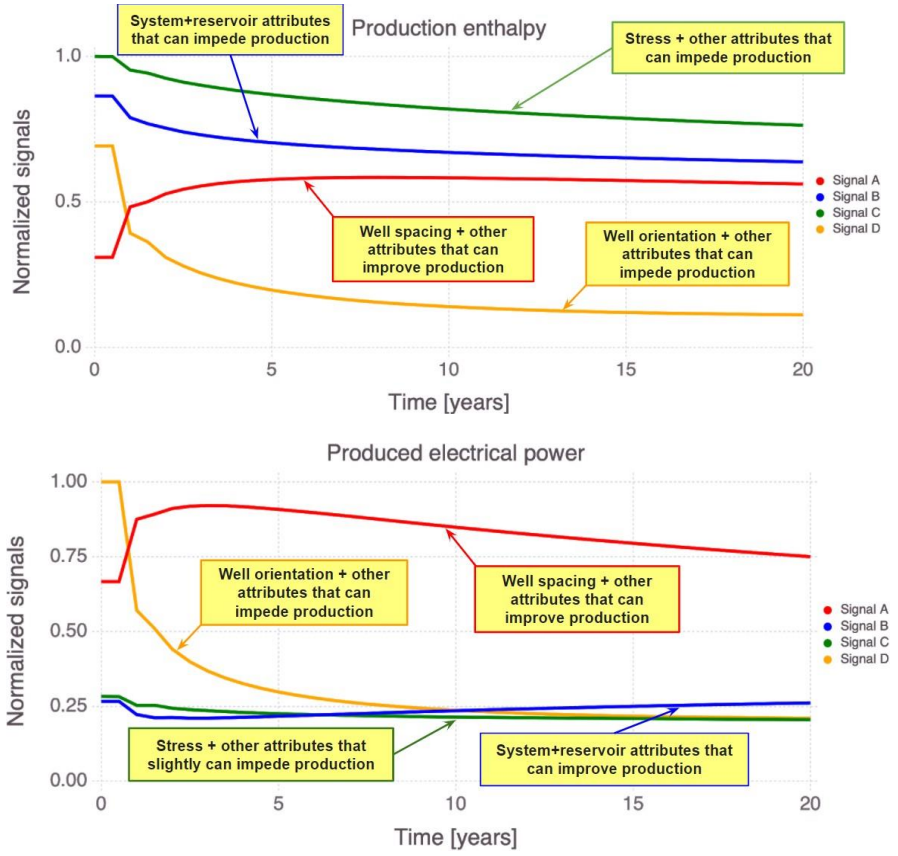

Figure 4.3.3: ML identifies the signature structure of the enthalpy and power production time series predicted by GeoDT. The primary physical components of each mixed signature are provided to aid interpretation. Only one of the signatures (red) shows inputs that are associated with increased production over time.

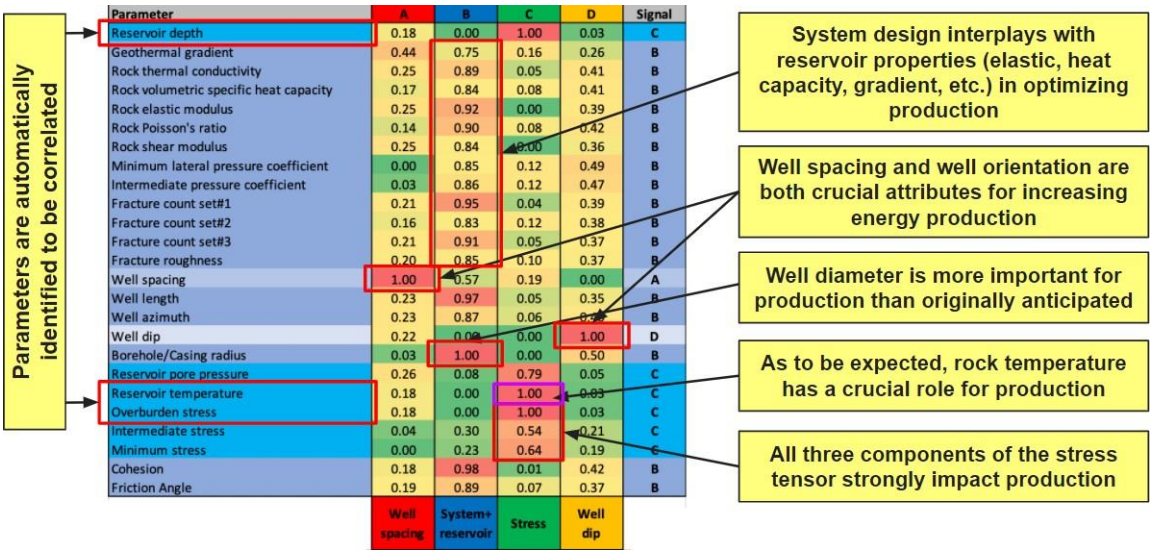

Fig. 4.3.4: Combined inputs of the four ML identified signatures that control geothermal power production. Callouts are included to highlight the primary physical components of each signature. We categorize each signature by its most dominant component. Red colors indicate parameters with high importance with that particular signature, green colors show that a parameter has a low weight with that specific signature.

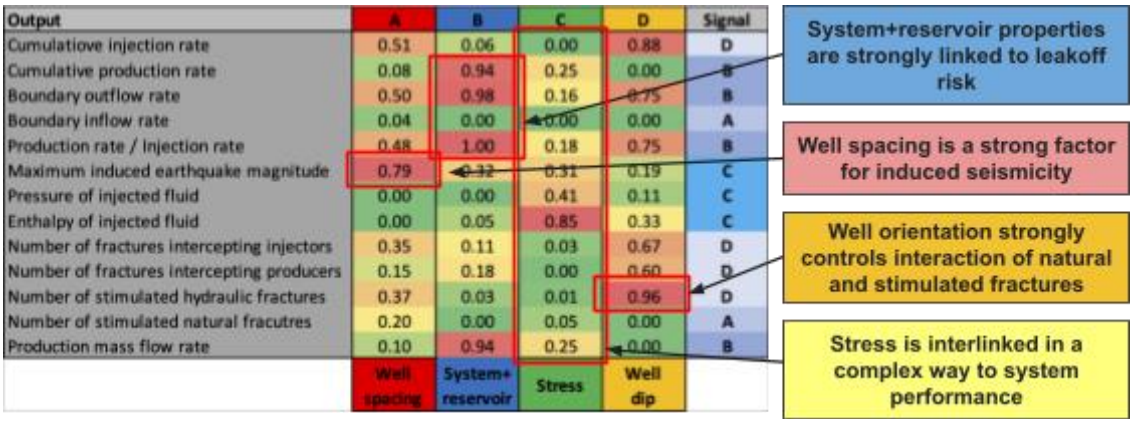

Fig. 4.3.5: ML predicted effects of the identified multi-attribute signatures on various outputs from GeoDT. Callouts are included to highlight significant links that aid understanding of subsurface geothermal processes coupled with system design. Red colors indicate outputs with high importance with that particular signature, green colors show that a specific output has a low weight with that specific signature.

5. Conclusions

GeoThermalCloud is a flexible open-source cloud-based ML framework for geothermal exploration. **GeoThermalCloud** can simultaneously handle both public and proprietary datasets. Also, **GeoThermalCloud** framework consists of a series of advanced pre-processing, post-processing, and visualization tools that tremendously simplify its application for real-world problems. These tools make the ML results understandable and visible even for non-experts; therefore, ML and subject-matter expertise are not critical requirements to use our ML framework.

GeoThermalCloud utilizes a series of novel LANL-developed patented ML tools called *SmartTensors* (https://github.com/SmartTensors). *SmartTensors* has already been applied to solve a wide range of real-world problems, from COVID-19(Vesselinov, Middleton, and Talsma 2021) to wildfires (http://tenosrs.lanl.gov), and it has won two 2021 R&D 100 awards, including a bronze award for market disruptor tools. *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, scientific, statistical, and machine learning computing.

GeoThermalCloud is designed to process and analyze diverse datasets including both small and large datasets. Also, it can handle sparse datasets with missing values. It does not only analyze but also finds actionable information for enabling decision makers to make sound decisions for geothermal exploration, development, and production. It finds such actionable information by finding mapping functions between all input parameters. We analyzed 10 diverse datasets and found critical information out of them that would not be possible by visual inspection or any other statistical tools. Overall, **GeoThermalCloud** 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.

ACKNOWLEDGEMENTS

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Geothermal Technology Office (GTO) Machine Learning (ML) for Geothermal Energy funding opportunity, Award Number DE-EE-3.1.8.1. The authors would like to thank Dr. Jeffrey Pepin, Dr. Drew Siler, and Dr. Erick Burns from U.S. Geological Survey (USGS) for providing us raw and processed data for ML analyses, and also for many useful and knowledgeable discussions.

REFERENCES

Ahmmed, B. 2020. "Machine Learning to Characterize Regional Geothermal Reservoirs in the Western USA." Online presented at the Geological Society of America, Online. https://www.osti.gov/biblio/1755886.

Ahmmed, B., N. Lautze, V.V. Vesselinov, D. Dores, and M.K. Mudunuru. 2020. *Unsupervised Machine Learning to Extract Dominant Geothermal Attributes in Hawaii Island Play Fairway Data*.

Ahmmed, B., and V.V. Vesselinov. 2021. "Prospectivity Analyses of the Utah FORGE Site Using Unsupervised Machine Learning." In *Geothermal Rising, San Diego, CA*.

Ahmmed, B., V.V. Vesselinov, and R.S. Middleton. 2020. "Geothermal Resource Analysis at Tohatchi Hot Springs, New Mexico." LA-UR-21-23827. Online: Los Alamos National Laboratory. https://www.osti.gov/biblio/1778753-geothermal-resource-analysis-tohatchi-hot-springs-new-mexico.

Ahmmed, B., V.V. Vesselinov, M.K. Mudunuru, R.S. Middleton, and S. Karra. 2021. "Geochemical Characteristics of Low-, Medium-, and Hot-Temperature Geothermal Resources of the Great Basin, USA." In *World Geothermal Congress, Reykjavik, Iceland*. https://www.researchgate.net/publication/348161100_Machine_Learning_on_the_Geochemical_Characteristics_of_Low-_Medium-_and_Hot-temperature_Geothermal_Resources_in_the_Great_Basin_USA.

Alexandrov, B.S., and V. V Vesselinov. 2014. "Blind Source Separation for Groundwater Pressure Analysis Based on Nonnegative Matrix Factorization." *Water Resources Research* 50 (9): 7332–47.

Bennett, C.R., and G.D. Nash. 2017. "The Convergence of Heat, Groundwater & Fracture Permeability: Innovative Play Fairway Modelling Applied to the Tularosa Basin." Ruby Mountain Inc. and Energy & Geoscience Institute, Salt Lake City, UT.

Constantine, P.G. 2015. *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*. SIAM.

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

Frash, L.P., N.J. Welch, M. Meng, W. Li, and J.W. Carey. 2021. "A Scaling Relationship for Fracture Permeability after Slip." In *Proceedings of the 55th US Rock Mechanics/Geomechanics Symposium, Houston, TX, June*.

Goff, F., D.h Bergfeld, and C.J. Janik. 2002. "Geochemical Data on Waters, Gases, Scales, and Rocks from the Dixie Valley Region, Nevada (1996-1999)." DOE–EEGTP (USDOE Office of Energy Efficiency and Renewable Energy Geothermal.

Haykin, S., and Z. Chen. 2005. "The Cocktail Party Problem." *Neural Computation* 17 (9): 1875–1902.

Iliev, F. L., V. G. Stanev, V. V. Vesselinov, and S. Alexandrov B. 2018. "Nonnegative Matrix Factorization for Identification of Unknown Number of Sources Emitting Delayed Signals." *PloS One* 13: e0193974.

Lichtner, P.C., G.E. Hammond, C. Lu, S. Karra, G. Bisht, B. Andre, R. Mills, and J. Kumar. 2015. "PFLOTRAN User Manual: A Massively Parallel Reactive Flow and Transport Model for Describing Surface and Subsurface Processes." LA-UR-15-20403. Los Alamos National Lab. (LANL), Los Alamos, NM (United States); Sandia National Lab. (SNL-NM), Albuquerque, NM (United States); Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States); Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States); OFM Research, Redmond, WA (United States). https://doi.org/10.2172/1168703.

Parsons, L., E. Haque, and H. Liu. 2004. "Subspace Clustering for High Dimensional Data: A Review." *Acm Sigkdd Explorations Newsletter* 6 (1): 90–105.

Siler, D.L., J.D. Pepin, V.V. Vesselinov, M.K. Mudunuru, and B. Ahmmed. 2021. "Machine Learning to Identify Geologic Factors Associated with Production in Geothermal Fields: A Case Study Using 3D Geologic Data, Brady Geothermal Field, Nevada." *Geothermal Energy*.

Vesselinov, V. V., B. Ahmmed, M. K. Mudunuru, J.D. Pepin, E. Burns, D.L. Siler, S. Karra, and R. Middleton. in review. "Discovering Hidden Geothermal Signatures Using Unsupervised Machine Learning." *Geothermics*.

Vesselinov, V. V, R.S. Middleton, and C.J. Talsma. 2021. "COVID-19: Spatiotemporal Social Data Analytics and Machine Learning for Pandemic Exploration and Forecasting." LA-UR-21-23230. Los Alamos, NM: Los Alamos National Laboratory.

Vesselinov, V. V., M. K. Mudunuru, S. Karra, D. O'Malley, and B. S. Alexandrov. 2019. "Unsupervised Machine Learning Based on Non-Negative Tensor Factorization for Analyzing Reactive Mixing." *Journal of Computational Physics* 395: 85–104.

Vesselinov, V.V. 2020. "Unsupervised Machine Learning to Discover Attributes That Characterize Low, Moderate, and High-Temperature Geothermal Resources." Presented at the Geothermal Resources Councils, Online.

Vesselinov, V.V., B. Ahmmed, M.K. Mudunuru, S. Karra, and R.S. Middleton. 2021. "Hidden Geothermal Signatures of the Southwest New Mexico." In *Proceedings of the World Geothermal Congress, Reykjavik, Iceland*. https://www.researchgate.net/publication/348161446_Discovering_the_Hidden_Geothermal_Signatures_of_Southwest_New_Mexico.

Vesselinov, V.V., L. Frash, B. Ahmmed, and M.K. Mudunuru. 2021. "Machine Learning to Characterize the State of Stress and Its Influence on Geothermal Production." *Geothermal Rising Conference, San Diego, CA*.

Vesselinov, V.V., M.K. Mudunuru, B. Ahmmed, Karra S, and Middleton R.S. 2020. "Discovering Signatures of Hidden Geothermal Resources Based on Unsupervised Learning." *45th Annual Stanford Geothermal Workshop*.

Zehner, R.E., M.F. Coolbaugh, and L. Shevenell. 2006. "Regional Groundwater Geochemical Trends in the Great Basin: Implications for Geothermal Exploration." *Geothermal Resources Council Transactions* 30: 117–24.