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

Machine Learning Estimates of Geothermal and Critical Mineral Prospectivity of the Great Basin

Velimir ("monty) Vesselinov, Tracy Kliphuis

EnviTrace LLC, Santa Fe, New Mexico, USA, https://envitrace.com

monty@envitrace.com, trais@envitrace.com

Keywords: artificial intelligence, machine learning, AI, ML, world models, data imputation, feature extraction, uncertainty analyses, risk assessment, hidden geothermal resources, critical mineral exploration, prospectivity.

ABSTRACT

Prospectivity mapping for geothermal and critical minerals is a complex undertaking reliant on integrating diverse datasets ranging from geological and geophysical surveys to geochemical analyses. Prospectivity mapping aims to identify areas likely to host commercially viable resources based on multiple criteria and risk factors. Subject-matter-driven methods, such as traditional Play Fairway Analysis (PFA) borrowed from the oil and gas industry, are commonly used. However, these approaches are often subjective and time-consuming. Also, the PFAs may not account for all the intricate details hidden in the analyzed datasets. This study explores the application and comparative performance of alternative artificial intelligence (AI) and machine learning (ML) techniques for automated and objective prospectivity assessment. In our work, we have investigated the efficacy of SmartTensors' Nonnegative Matrix Factorization (NMFk), Support Vector Machines (SVM), Graph Neural Networks (GNN), Gradient Boosting (XGBoost), Diffusion Models (DM), and Kriging Convolution Networks (KCN) in predicting geothermal potential. The analyses are based on Great Basin datasets collected under various past projects. These include the USGS/DOE's GeoDAWN and INGENIOUS projects. We merged surface, structural geology, gravity, magnetic, heat flow, and geochemical data attributes. The performance of each model is evaluated using a series of metrics, as well as through visual inspection of the resulting prospectivity maps. We also analyzed feature importance for each model to gain insights into the key factors influencing prospectivity. The tested ML methods and tools are deployed and demonstrated on EnviCloud (https://envitrace.com/#envicloud; https://envitrace.com/saas). EnviCloud is a proprietary, comprehensive, cloud-based platform designed to optimize the entire reservoir lifecycle, from initial exploration and site assessment to real-time well monitoring and production optimization. It is developed to support Software-as-a-Service (SaaS) licensing as well as project consulting work. By leveraging cloud/high-performance computing, cloud data management, AI/ML, data analytics, and GIS, EnviCloud streamlines resource utilization and enhances decision making. The platform offers key features such as AI-powered geologic mapping, reservoir simulations, near-real-time data analytics, and risk/decision analysis tools for project feasibility, all at a centralized cloud dashboard for multi-user collaboration. It also includes tools for tracking sustainability metrics and ensuring regulatory compliance, especially related to groundwater contamination and induced seismicity. EnviCloud reduces exploration time and maximizes the potential of geothermal and critical mineral resources. It targets a broad audience, including exploration companies, energy providers, investors, regulators, and research institutions, aiming to promote sustainable geoengineering. Here, we demonstrate EnviCloud's application in processing the Great Basin datasets. Our ML analyses successfully extracted key features relevant to evaluating geothermal and critical mineral prospectivity. This research contributes to a more robust and efficient methodology for exploration, potentially reducing exploration costs and accelerating the discovery of new geothermal and critical mineral resources.

1. INTRODUCTION

Driven by the escalating global need for detailed geothermal and critical mineral prospectivity analyses, the accurate, defensible, and robust high-resolution reconstruction and analyses of sparse geospatial datasets are of great importance. Geothermal and critical mineral prospectivity mapping, the identification of areas likely to host commercially viable resources, is a complex process reliant on integrating diverse datasets ranging from geological and geophysical surveys to geochemical analyses.

Subject-matter-driven methods, such as the traditional Play Fairway Analysis (PFA), are conventionally employed as an initial screening tool in resource exploration (Pauling *et al.*, 2023; Smith *et al.*, 2023). The PFA methodology, borrowed from the petroleum industry, typically involves a systematic and qualitative assessment of geological, geophysical, and geochemical data to identify and rank areas with high prospectivity. However, a significant drawback of these conventional approaches is their inherent subjectivity. The ranking and weighting of different geoscience factors often rely heavily on the expert knowledge, experience, and sometimes personal bias of the geoscientists involved. This may lead to variability in results across different teams or projects. Furthermore, this multi-disciplinary integration and interpretation process is time-consuming and expensive. Compiling, cross-referencing, and manually interpreting vast datasets to generate a robust PFA model can take months or years, which hinders the speed and efficiency of the overall exploration campaign, particularly in frontier or underexplored areas.

As an alternative to traditional, often subjective, and labor-intensive PFA methods, here we explore the application and comparative performance of several machine learning (ML) techniques for automated, objective, and high-throughput prospectivity assessment. This

study specifically evaluates the efficacy of semi-supervised and physics-informed ML models. The primary objective is to develop a robust, data-driven framework capable of identifying subtle, non-linear relationships between geological features and the occurrence of viable geothermal resources or economically significant critical mineral deposits, thereby significantly reducing exploration time, cost, and risk compared to conventional exploration campaigns. The comparative analysis will focus on metrics such as prediction accuracy, recall, precision, and model interpretability to determine the optimal ML architecture for various exploration scenarios.

A significant challenge in geothermal and critical mineral exploration lies in extracting and interpreting meaningful "hidden" (latent) features present within the available data. These concealed features, often overlooked or underestimated, can hold substantial potential in accurately assessing and evaluating prospectivity. Uncovering these "hidden" indicators requires advanced ML analysis techniques and a deep understanding of the geological, geophysical, and geochemical processes that govern geothermal systems. By effectively identifying and integrating these "hidden" features into the evaluation process, we can enhance the accuracy and reliability of resource assessments. This leads to more informed decision-making and, ultimately, a greater success rate in geothermal and critical mineral exploration and development projects.

Another significant challenge in geothermal and critical mineral exploration is the inherent disparity in the spatial scale and dimensionality of the supporting data. This diversity complicates the integration and analysis of the collected information. For instance, some crucial data points are representative of point measurements—zero-dimensional (0D) or highly localized, discrete values—such as downhole temperature readings, fluid chemistry analyses from a single well, or the specific concentration of a critical mineral within a hand-sample rock chip. These data are precise in their location but offer little direct information about the surrounding environment. Conversely, other essential datasets are representative of large-scale features—one, two, or even three-dimensional phenomena covering extensive areas often spanning several kilometers. Examples of these macro-scale features include mapped fault systems and fracture networks (often 2D or 3D surfaces), regional geological contacts and stratigraphy (3D volumes), airborne geophysical surveys (2D grids or 3D inversions), and seismic reflection profiles (2D cross-sections or 3D cubes). The heterogeneity in data scale creates major challenges for predictive modeling. A model must simultaneously honor the high-resolution constraints imposed by point measurements while also capturing the influence of vast, kilometer-scale geological structures that govern fluid flow, heat transport, and mineral deposition. Effective exploration and resource estimation, therefore, require advanced geostatistical and machine learning techniques capable of spatially correlating and fusing these disparate data types—from single, precise sensor readings to expansive, regional structural interpretations—into a coherent, predictive 3D subsurface model.

Furthermore, the sheer volume, velocity, and variety of data generated from modern exploration techniques—including high-resolution geophysical surveys, remote sensing, well logs, geochemical assays, and distributed sensor networks—often outpace the capacity of conventional geological and reservoir modeling software. This disparity creates a bottleneck where sophisticated analytical techniques, such as machine learning and high-fidelity simulations, cannot be effectively applied across the entire exploration target. Furthermore, the integration of these disparate data streams, which are often stored in heterogeneous formats and subject to different quality control standards, introduces significant uncertainty and complexity into the interpretation process, ultimately hindering the accurate delineation of viable geothermal reservoirs and high-concentration critical mineral deposits. Addressing this challenge necessitates the development of scalable, cloud-based data management and processing platforms capable of handling petabyte-scale datasets and facilitating rapid, multi-physics data fusion.

Our ML methods and tools are designed to be deployed within EnviCloud (https://envitrace.com/#envicloud; https://envitrace.com/saas) for efficient and near-real-time analyses. Our EnviCloud is a proprietary, comprehensive, cloud-based platform designed to optimize the entire geothermal energy lifecycle, from initial exploration and site assessment to real-time well monitoring and energy output optimization. It supports Software-as-a-Service licenses as well as project consulting work. By leveraging cloud/high-performance computing, AI/ML, data analytics, and GIS, EnviCloud streamlines geothermal resource utilization and enhances decision making. The platform offers key features such as AI-powered geologic mapping, near-real-time data analytics, reservoir simulations, and risk/decision analysis tools for project feasibility, all at a centralized cloud dashboard for multi-user collaboration. It also includes tools for tracking sustainability metrics and ensuring regulatory compliance, especially related to groundwater contamination and induced seismicity. EnviCloud reduces exploration time and maximizes the potential of geothermal resources. It targets a broad audience, including exploration companies, energy providers, investors, regulators, and research institutions, aiming to promote sustainable energy development.

In this context, the EnviCloud suite of Machine Learning (ML) algorithms represents a significant advancement, offering a robust framework specifically designed to address these complex challenges of spatial data analysis, exploration, imputation, modeling, and prediction. Our AI algorithms leverage advanced statistical learning techniques to infer missing information with greater accuracy and reliability than conventional approaches.

Our research builds on past work by us and others that emphasized the need for advanced imputation strategies to address the limitations of existing methods when applied to diverse and complex datasets, especially those related to geosciences (Jarrett *et al.*, 2022; V. Vesselinov *et al.*, 2022; V. V. Vesselinov *et al.*, 2022; Vesselinov, 2023). By comparing a range of state-of-the-art techniques, we aim to evaluate their accuracy and computational efficiency systematically.

2. METHODOLOGY

In this study, we evaluated and compared a series of state-of-the-art imputation methods—SmartTensors' NMFk (Nonnegative Matrix Factorization with k-means clustering), SVR (Support Vector Regression), XGBoost (eXtreme Gradient Boosting), Graph Neural Networks (GNN), and KCN (Convolution Network Kriging). Our analyses have so far demonstrated that the SmartTensors' NMFk has performed consistently the best for the analyzed datasets. Here, we are presenting only the results obtained using SmartTensors' NMFk.

2.1. Datasets

We focus on all the publicly available datasets related to geothermal conditions in the Great Basin area. This includes geophysical, geochemical, and hydrochemical measurements from sources such as the GeoDAWN (USGS, 2022) and INGENIOUS project (Ayling, 2022) and state geologic institutions. Information about a range of geochemical and hydrochemical elements (K, U, Th, Li, Mg, Na, Fe, Ba, B, Ca, As, HCO_3, SiO_2, F, SO_4, Cl, Tc) is analyzed. We also processed and gridded many additional features for this region that would benefit this analysis further. These include 500k-scale geologic maps, geologic structures, quaternary faults, well and spring chemistry, aquifers, and subsurface fluid flow. Furthermore, our analyses include the GeoDAWN data, which we have previously analyzed in great detail (Kliphuis *et al.*, 2025; Vesselinov *et al.*, 2025). The GeoDAWN dataset encompasses 149,030 line-kilometers of flight lines, covering an expansive area of 51,857 square kilometers. Data was gathered using airborne platforms flying at a spacing of 200 to 2000 meters apart, facilitating broad coverage while maintaining a reasonable resolution. This dataset is rich in detail, incorporating more than 20 distinct measured attributes, providing a multifaceted perspective on the surveyed region. This comprehensive dataset amounts to a substantial volume of data, exceeding 5 gigabytes. The collected information holds significant potential for advancing our understanding of subsurface geological structures and resources, particularly those relevant to geothermal energy exploration and development, as well as in-situ mining of rare-earth elements.

In total, we have processed 126 data attributes. This integrated dataset provides a comprehensive spatial representation of geological characteristics across the study area. **Figure 1** presents just 6 out of 126 data attributes.

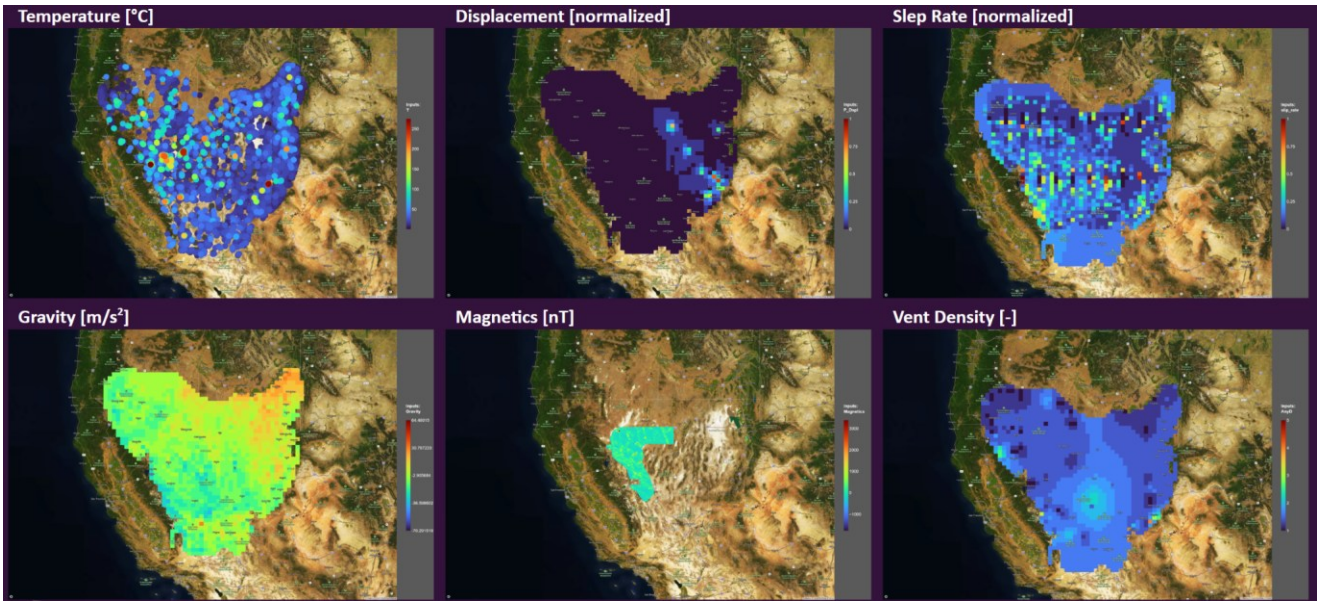

Figure 1: Various data attributes (just 6 out of 126) representing the Great Basin geologic conditions. The analysis assimilates all these datasets (126 in total).

It is important to note that the analyzed datasets, which form the basis of this investigation, are characterized by inherently different support scales. This disparity arises from the varied methodologies and instrumentation employed during the data acquisition process for each distinct dataset. Consequently, direct, point-to-point comparisons or simple aggregation of these data require careful consideration and appropriate normalization or upscaling/downscaling techniques to ensure that the heterogeneity in the underlying measurement volume or area (i.e., the support scale) does not unduly bias the subsequent statistical analyses or model parameterization. The implications of these differing support scales are particularly critical when integrating field measurements, laboratory analyses, and computational model outputs, as the scale of observation directly influences the observed variance and spatial correlation structures within the geothermal reservoir properties.

Vesselinov, Kliphuis.

2.2 Machine Learning Methods

2.2.1. SmartTensors' NMFk

SmartTensors' NMFk is at the forefront among various unsupervised ML methods, such as NMF, PCA, ICA, SVD, and its variants, k-means clustering, and Gaussian mixture models. In contrast with traditional NMF (Lee and Seung, 1999), NMFk allows for the automatic identification of the optimal number of signatures (features) present in the data (Vesselinov, Alexandrov and O'Malley, 2019, 2019). The non-negativity constraint makes the decomposed matrices easier to interpret than PCA, SVD, and ICA because the extracted signatures are additive (Lee and Seung, 1999). Moreover, our version of NMF (implemented in NMFk) can also handle categorical and missing data. Missing data is challenging or impossible to address with other supervised and unsupervised ML methods (Vesselinov, Alexandrov and O'Malley, 2018; Vesselinov *et al.*, 2019; Siler *et al.*, 2021). Even more importantly, the missing data can be reconstructed from available data based on the estimated matrix factorization. NMFk is part of our SmartTensors ML framework.

2.2.2 Support Vector Regression (SVR)

SVR is a kernel-based supervised learning algorithm that projects data into a higher-dimensional feature space to capture nonlinear relationships (Drucker *et al.*, 1996; Chang and Lin, 2013). By employing a suitable kernel (such as a Radial Basis Function), SVR can model complex interactions among the variables and the spatial structure of the missing data. One principal advantage of SVR lies in its robustness against overfitting, stemming from a well-defined regularization framework. Its capacity to accommodate moderate-sized datasets with meaningful complexity often leads to accurate imputations in scenarios with relatively smooth or structured patterns. However, tuning kernel parameters, regularization terms, and insensitivity thresholds can be computationally intensive and highly sensitive to data scale. Moreover, performance may degrade on massive datasets because of the time and memory required to compute and store kernel transformations.

2.2.3 XGBoost

XGBoost (eXtreme Gradient Boosting) is an ensemble method that builds decision trees in sequence, with each tree attempting to reduce the residual errors from the preceding iteration (Chen and Guestrin, 2016). Thanks to a combination of efficient tree construction, sparse data handling, and integrated regularization, XGBoost has gained attention for its strong predictive performance across diverse domains. The main strength of XGBoost is its flexibility in modeling complex interactions and scalability on large datasets when suitably optimized. Regularization techniques help control overfitting, and parallelization speeds up training. However, this flexibility entails an ample hyperparameter search space, including learning rates, tree depths, and the number of boosting rounds. Prolonged tuning procedures can significantly increase computational costs, potentially making XGBoost less appealing if time or computational resources are limited.

2.2.4 Graph Neural Networks (GNN)

Graph Neural Networks (GNNs) represent a class of deep learning models designed (Zhou *et al.*, 2021) to operate on data structured as graphs, such as networks of geological faults, interconnected wells, or spatially correlated sensor measurements. Unlike conventional machine learning models that assume data points are independent (like SVR or XGBoost), GNNs explicitly leverage the relationships (edges) between data points (nodes) to learn better feature representations. This is achieved by iteratively aggregating and transforming feature information from a node's local neighborhood. For spatial imputation, GNNs are highly effective because they inherently capture the non-Euclidean nature and spatial dependencies within geoscience data. A key advantage is their ability to model complex, non-linear dependencies and structural correlations that models might overlook, focused only on attribute values. Furthermore, they support inductive learning, meaning a trained model can generate predictions for entirely new, unseen graph structures (e.g., new areas or newly added wells) without re-training. However, GNNs can be computationally intensive, especially for extremely large or dense graphs, and their performance is highly sensitive to the quality and structure of the initial graph definition (how "neighborhoods" are defined).

2.2.5 Kriging Convolutional Networks (KCN)

Kriging Convolutional Networks (KCN) represent a hybrid spatial interpolation method that combines the statistical rigor of traditional Kriging with the flexibility of graph-based neural networks (Appleby, Liu and Liu, 2020). KCNs are designed to address the limitations of conventional kriging, which relies on strong Gaussian assumptions and can be computationally expensive on large datasets. Instead of treating all data as a single, static Gaussian process, KCNs construct small, dynamic K-nearest neighbor (KNN) graphs for each prediction point. KCN directly leverages known training labels from neighboring observations as input, whereas standard graph convolutional networks (GCNs) often only use features. GCNs are more flexible in modeling non-Gaussian, highly complex spatial data and do not require re-training for new test points (inductive learning). KCNs can emulate traditional kriging while being much more capable of capturing non-linear relationships.

3. RESULTS

3.1 Geothermal Prospectivity

Estimating the prospectivity of geothermal reservoirs is a multidimensional challenge rooted in the non-linear coupling of thermal-hydraulic-mechanical-chemical (THMC) processes within the subsurface. Geothermal "hidden" systems often lack surface manifestations

like hot springs, requiring the integration of disparate data—such as magnetotellurics (MT), microseismicity, and multicomponent geothermometry—to delineate permeable fluid pathways (Dobson, 2016; Vesselinov *et al.*, 2021; V. Vesselinov *et al.*, 2022; V. V. Vesselinov *et al.*, 2022). The primary complexity lies in the structural heterogeneity of fracture networks; faults act as either high-permeability conduits or impermeable barriers depending on their orientation relative to the regional stress field and the degree of hydrothermal mineralization (Young and Akar, 2015). Furthermore, the transition toward Enhanced Geothermal Systems (EGS) introduces epistemic uncertainties regarding induced seismicity and long-term reservoir sustainability, as fluid injection can trigger shear reactivation of pre-existing fractures (Huang *et al.*, 2021). These variables necessitate the use of stochastic inversion and recurrent neural networks to resolve "conceptual uncertainty" and reduce the high economic risk associated with exploratory drilling (Pollack, Horne and Mukerji, 2020; Norbeck and Latimer, 2023; El-Sadi *et al.*, 2024; Fercho *et al.*, 2025).

Figure 2 shows estimated geothermal prospectivity across the Great Basin region. The analysis uses our SmartTensors' NMFk algorithm, and it is built upon the assimilation of all the diverse and extensive datasets previously discussed above, ensuring a robust and well-informed foundation for the prospectivity mapping. The prospectivity here represents assimilation of the NMFk predicted "hidden" features representing the entire processed dataset.

It is important to note that based on our detailed geospatial and statistical analyses, a clear, direct, one-to-one correlation between the discrete groundwater temperature measurements (as visually represented in the left panel of **Figure 2**) and the spatially interpolated estimates of geothermal prospectivity (as mapped in the right panel of **Figure 2**) cannot be established for the entirety of the Great Basin study area. For example, in some of the areas where the ML model predicts high geothermal prospectivity, we either lack temperature measurements or the temperature measurements are low. Conversely, in areas with relatively high temperature measurements, the ML model predicts low geothermal prospectivity.

Although groundwater temperature is commonly used as a first-order geothermal indicator, our results demonstrate that elevated temperature does not necessarily correspond to high geothermal prospectivity. This apparent mismatch reflects the underlying physics of geothermal systems. Shallow temperature measurements may be suppressed by advective cooling due to regional groundwater flow, even where deep thermal anomalies are present. Conversely, elevated temperatures can occur in structurally isolated compartments where permeability is insufficient to sustain economically viable fluid circulation. Structural sealing, mineralized fault zones, and anisotropic fracture permeability further decouple surface thermal expression from deeper reservoir potential.

Additionally, hydrologic dilution and uneven sampling density introduce observational bias. Prospectivity therefore depends not solely on temperature magnitude, but on the coupled thermal–hydraulic–structural framework governing heat recharge, permeability architecture, and fluid transport. This mechanistic interpretation underscores the necessity of multi-attribute integration rather than reliance on temperature anomalies alone.

This lack of perfect congruence suggests that while groundwater temperatures are an important, easily accessible proxy indicator of underlying thermal conditions, they are not the sole or definitive factor controlling the overall geothermal potential of a given subsurface region. Factors such as subsurface hydrology, the presence and permeability of fault systems acting as conduits, the depth and nature of the thermal anomaly, and the influence of regional groundwater flow on shallow thermal gradients all contribute to the final assessment of prospectivity. Therefore, relying exclusively on shallow temperature measurements would lead to an incomplete or potentially misleading interpretation of the Great Basin's vast and heterogeneous geothermal resource. A multi-variate approach integrating geophysical data, structural mapping, and geochemistry is essential for accurate resource characterization.

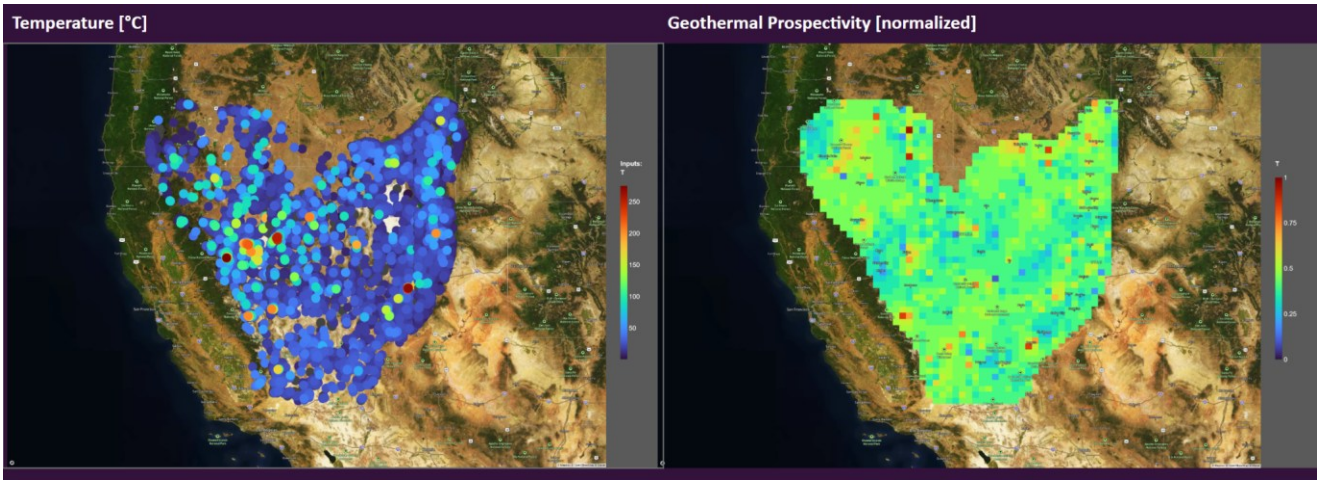

Figure 2: Temperature measurements (left) vs estimated geothermal prospectivity (right) in the Great Basin.

3.2 Critical Minerals and Rare-Earth Prospectivity

Estimating the prospectivity of geologic reservoirs for critical minerals and rare earth elements (REE) is an inherently complex endeavor due to the heterogeneous distribution and intricate geochemical behavior of these elements (Abah *et al.*, 2025). Unlike traditional base metals, REEs are often found in low concentrations, dispersed as "incompatible elements" within the lattice of accessory minerals, rather than forming large, pure mineral phases (Walsh and Spandler, 2022). This necessitates sophisticated multiphysics and multi-source data integration, combining airborne geophysics, remote sensing, and isotopic geochemistry to identify subtle structural controls and alteration halos.

Furthermore, the prospectivity modeling of unconventional reservoirs—such as ion-adsorption clays or coal-based formations—is frequently hindered by a lack of historical exploration data and the high degree of lithologic heterogeneity at depth (Creason *et al.*, 2023). As a result, the industry is increasingly shifting toward AI/ML-based predictive models and "mineral systems" frameworks to reconcile conflicting data streams and resolve the semantic complexities of deep-seated, concealed mineral targets.

Using our SmartTensors' NMFk algorithm, we have performed detailed analyses of critical minerals and rare earth elements (REE) prospectivity in the Great Basin region. **Figures 3 and 4** show estimated critical mineral and rare earth element prospectivity across the Great Basin region. **Figure 3** summarizes the Lithium results. **Figure 4** shows the Boron results. Results for the other critical minerals and rare earth elements in the region provide similar conclusions. As above, the analysis is built upon the assimilation of all the diverse and extensive datasets, ensuring a robust and well-informed foundation for the prospectivity mapping.

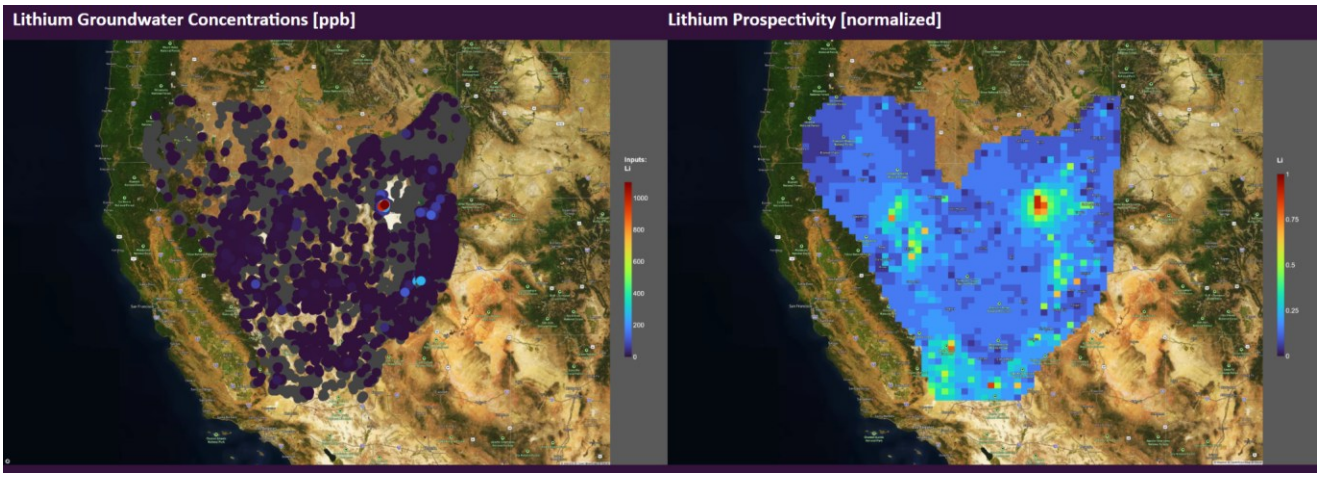

Figure 3: Lithium groundwater samples concentrations (left) vs estimated lithium prospectivity (right) in the Great Basin. Note that many of the processed groundwater samples lack lithium data; these sampling locations are shown as gray dots.

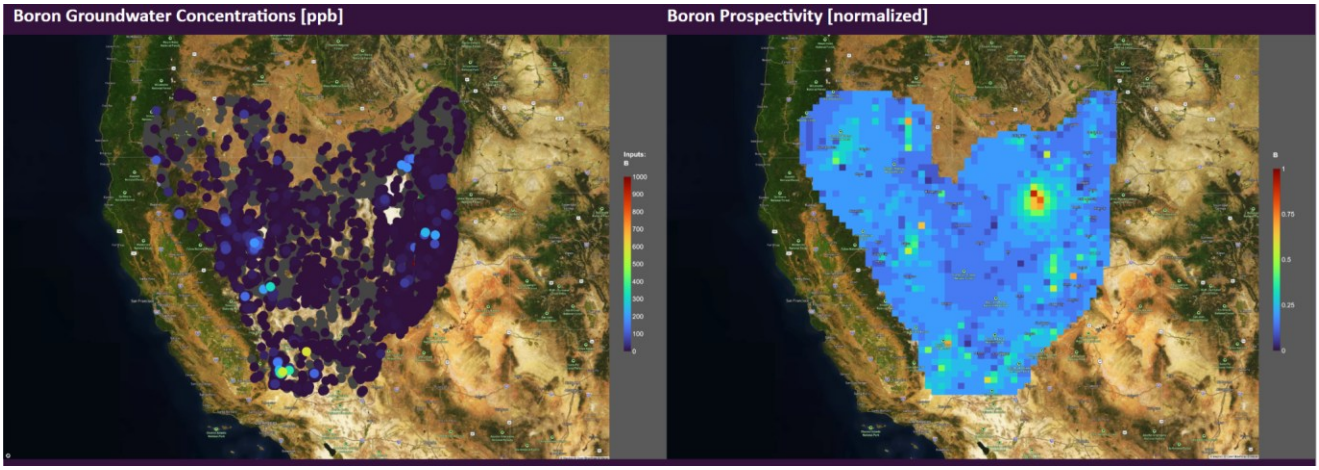

Figure 4: Boron groundwater samples concentrations (left) vs estimated boron prospectivity (right) in the Great Basin. Note that many of the processed groundwater samples lack boron data; these sampling locations are shown as gray dots.

A direct, one-to-one correlation between the discrete lithium and boron concentration measurements (**Figures 3 and 4, left panel**) and the spatially interpolated estimates of lithium prospectivity derived from the Machine Learning (ML) model (**Figures 3 and 4, right panel**) is not evident across the Great Basin study area. This is supported again by two key observations. First, certain areas predicted to have high lithium and boron prospectivity by the ML model either have no lithium or boron data or exhibit low lithium readings; note that

many of the processed groundwater samples lack data for lithium and boron; these sampling locations are shown as gray dots. Second, conversely, some areas with relatively high lithium and boron concentrations are predicted to have low lithium and boron prospectivity by the ML model.

The robust evaluation of these prospectivities is of paramount importance for several strategic and economic reasons. Firstly, securing a domestic supply of critical minerals and REE is essential for national security and the stability of high-tech industries, including defense, renewable energy technologies, and advanced electronics. Secondly, a precise understanding of resource locations significantly de-risks exploration and development efforts, reducing both the financial investment and the environmental footprint associated with resource extraction. This targeted approach ensures that resources are allocated efficiently, accelerating the time-to-market for these vital materials.

3.3 Feature Extraction

Our comprehensive analysis successfully employed a latent-variable modeling approach to identify and cluster 8 distinct geologic provinces, which we term "hidden" or "latent" signatures (i.e., features or signals), within the complex and multi-dimensional Great Basin dataset. Each of these 8 extracted signals represents a unique combination of geological, geophysical, and geochemical features, thereby reflecting the inherent diversity and complexity embedded within the raw data.

The spatial distribution and geographical extent of these 8 identified signals are visually documented in **Figure 5**. This figure provides a geographical context for the results, allowing for a clear understanding of where these unique geothermal characteristics manifest across the Great Basin. The 8 provinces are associated with the following 8 signals with their respective data attributes: A: C14 age and Basement Depth, B: Radon, C: TDS and Conductivity, D: Technicium-99, E: Li/Na, F: Carbonate hardness, G: Temperature, Boron, and Stepover Fault Systems, and H: Na/Ca/Mg (major cations). The association between the data signatures and data attributes is revealed in **Figure 6**.

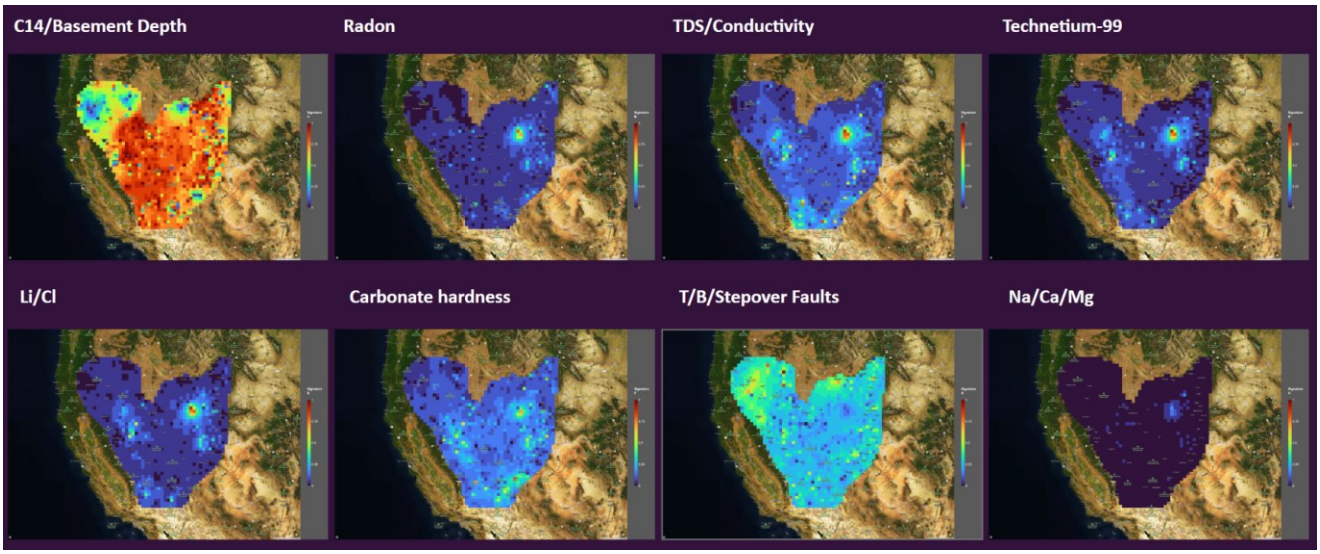

Figure 5: Map AI identified provinces (spatial domains representing hidden "latent" signatures) in the Great Basin dataset. The 8 provinces are associated with the following 8 signals with their respective data attributes: A: C14 age and Basement Depth, B: Radon, C: TDS and Conductivity, D: Technicium-99, E: Li/Na, F: Carbonate hardness, G: Temperature, Boron, and Stepover Fault Systems, and H: Na/Ca/Mg (major cations). See also the matrix in Figure 6.

Figure 6 presents a matrix plot (heatmap) capturing the association of the extracted "hidden" signatures in the Great Basin dataset and their association with the respective input data attributes. Note that only a subset of 30 attributes is plotted in **Figure 6**. There are 126 data attributes in total that have been processed. **Figure 6** shows only the 30 most important attributes associated with the 8 extracted hidden features (signatures).

Figure 6 is a color-coded heatmap of the association matrix capturing relationships between data attributes and data signatures. It shows close to 0 matrix elements as green squares. Green squares define low associations between signatures and attributes. The red matrix squares map matrix elements close to 1; they define strong associations between signatures and attributes.

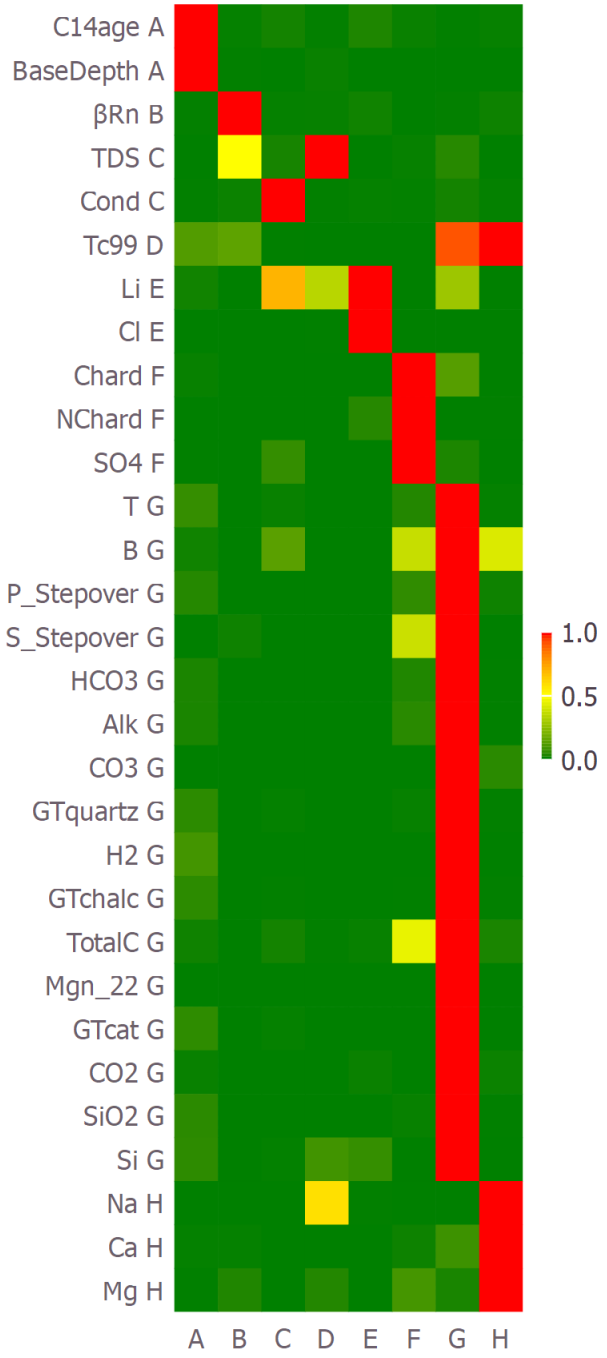

Figure 6: Matrix plot representing the association of the extracted "hidden" signatures in the Great Basin dataset and their association with the respective input data attributes. Note that only a subset of 30 attributes is plotted. There are 126 data attributes in total that have been processed. Here, we show only the 30 most important attributes associated with the 8 extracted hidden features (signatures). The extracted 8 hidden signatures are predominantly associated with the respective data attributes: A: C14 age and Basement Depth, B: Radon, C: TDS and Conductivity, D: Technicium-99, E: Li/Na, F: Carbonate hardness, G: Temperature, Boron, and Stepover Fault Systems, and H: Na/Ca/Mg (major cations).

Some of the data attributes are predominantly associated with a single signature. For example, the association between Mg and Signature H is represented by the red square in the lower right corner of the heatmap (**Figure 6**). Other data attributes are associated with multiple signatures (e.g., Lithium, Li). Similarly, some data signatures are associated with multiple data attributes (e.g., Signature G).

These results demonstrate the robustness of the NMFk clustering approach in delineating geochemical and geophysical provinces, revealing transparent and interpretable patterns in the Great Basin dataset.

3.4 Uncertainties

Robust geothermal and critical mineral prospectivity assessment requires not only predictive skill but also defensible characterization of uncertainty. Given the heterogeneity of the Great Basin datasets—including varying spatial supports, missing data fractions, and multi-scale geophysical signals—quantifying uncertainty is essential for reducing exploration risk and informing drilling decisions. Uncertainties in this study and our ML predictions arise from several primary sources:

1. **Data Sparsity and Missingness:** Many geochemical attributes (e.g., Li, B, Tc) are absent in a substantial fraction of sampling locations. Spatial gaps are non-uniform and often clustered.
2. **Support-Scale Disparities:** Integration of point measurements (e.g., well temperatures) with gridded airborne geophysics introduces scale-dependent variance and smoothing artifacts.
3. **Model Structural Uncertainty:** Different ML architectures (NMFk, SVR, XGBoost, GNN, KCN) embody different assumptions regarding linearity, spatial correlation, and feature interactions.
4. **Hyperparameter Sensitivity:** Model outputs depend on choices such as latent dimension (k), regularization strength, kernel parameters, and neighborhood definitions.
5. **Conceptual Geological Uncertainty:** Competing interpretations of structural permeability, fluid pathways, and heat transport mechanisms introduce epistemic uncertainty independent of data density.

Future work will explore the uncertainties in the data and ML predictions. The existing capabilities within our SmartTensors' NMFk methodology already provide a defensible approach to perform these analyses. In this work, we also rely on our existing MADS (Model Analyses and Decision Support) computational framework (Daniel O'Malley and Vesselinov, 2014; D. O'Malley and Vesselinov, 2014; O'Malley and Vesselinov, 2015; Vesselinov and O'Malley, 2016; Vesselinov, 2025).

4. CONCLUSIONS

Our findings thus far indicate that our SmartTensors' Non-Negative Matrix Factorization with k-means clustering (NMFk) methodology for AI/ML data analysis shows the most significant potential among the diverse range of ML approaches we have explored for evaluation of geothermal and critical mineral prospectivity. This conclusion is based on an evaluation of accuracy, computational efficiency, and scalability across our complex Great Basin datasets.

The results from the application of SmartTensors' NMFk strongly suggest that this advanced methodology is not only capable of but also highly effective at accurately and efficiently filling in the substantial number of missing data points (data gaps subject to data imputation) that are endemic within our large-scale geothermal datasets. This capability is crucial for improving the fundamental reliability and overall robustness of our Machine Learning (ML) model analyses in general. The reality of geothermal and critical mineral exploration data is that it is often incomplete, suffering from sensor failures, sporadic logging, or historical gaps; this incomplete nature can usually lead to significantly inaccurate predictions, biased model results, and ultimately, poor decision-making regarding reservoir management and drilling targets.

The NMFk approach is also uniquely suited to capture the latent (hidden), underlying physical relationships and structures inherent in the complex multivariate geologic data, allowing for physically-informed and statistically sound predictions. By continuing to refine and rigorously optimize our SmartTensors approach methodically, we can further develop significantly more accurate, predictive, and effective ML models for all phases of geologic exploration, resource assessment, and long-term resource management. Success in this area will directly translate into reduced exploration risk and more sustainable production.

SmartTensors algorithms are a sophisticated analytical tool designed with the innate capability to discern and extract pivotal ("hidden") spatial and attribute features embedded within complex datasets. These features encompass a broad spectrum of latent information, such as the identification of subtle yet significant patterns, long-term trends, or sudden anomalies in the geographical or spatial distribution of the data points. Furthermore, SmartTensors excel at recognizing and characterizing the key descriptive characteristics or intrinsic attributes intrinsically linked with specific data points. This powerful capability to unearth deep-seated information is not merely an academic exercise; it is the fundamental mechanism through which SmartTensors generate substantial, actionable insights, thereby significantly informing and optimizing complex decision-making processes across various fields, including, but not limited to, geothermal energy exploration, environmental monitoring, and resource management. The extracted intelligence transforms raw data into a strategic asset.

This prospectivity mapping is important as it directly informs and prioritizes future geothermal and critical mineral exploration and development efforts. By pinpointing areas of high prospectivity with data-driven confidence, the prospectivity maps allow developers and policymakers to allocate limited resources—such as drilling budgets and infrastructure planning—to the locations with the highest probability of success. This targeted approach minimizes financial risk, accelerates the transition to renewable energy sources, and

ultimately aids in realizing the vast, untapped geothermal and critical mineral potential of the Great Basin region, thereby contributing significantly to regional and national energy security and carbon reduction goals.

Future work will aim to:

- Incorporate more data representative of the subsurface conditions in the Great Basin region.
- Comparisons against similar AI/MLresults executed by others:
    - Stanford Thermal Earth Model (M. Aljubran and Horne, 2024; M. J. Aljubran and Horne, 2024).
    - Transparent Earth v2.0 developed by O'Malley et al. (2026) at the Los Alamos National Laboratory.
- Further improve the way we account for the spatial context of the data, including the support scale of the data and the scale (size) of the analyzed features.
- Capture vertical and temporal aspects of the analyzed datasets.
- Account for uncertainties in the data.
- Provide uncertainties related to model predictions.

5. ACKNOWLEDGMENT

EnviTrace work is partly funded by DOE SBIR Grant DE-SC0022697 titled "GeoTGo: Equitable and inclusive tool for community-based geothermal development" and DOE SBIR Grant DE-SC0023594 titled "GeoML: AI/ML for interpretation of geoscience data and prediction of geologic reservoir engineering activities".

References

Abah, M.A. *et al.* (2025) "Mineralogy and Geochemistry of Critical Minerals: Exploration, Extraction, and Sustainability," *Environmental Reports*, 7(2), pp. 246–251. Available at: https://doi.org/10.51470/ER.2025.7.2.246.

Aljubran, M. and Horne, R. (2024) "Stanford Thermal Earth Model for the Conterminous United States." DOE Geothermal Data Repository; Stanford University. Available at: https://doi.org/10.15121/2324793.

Aljubran, M.J. and Horne, R.N. (2024) "Thermal Earth model for the conterminous United States using an interpolative physics-informed graph neural network," *Geothermal Energy*, 12(1), p. 25. Available at: https://doi.org/10.1186/s40517-024-00304-7.

Appleby, G., Liu, L. and Liu, L.-P. (2020) "Kriging convolutional networks," *Proceedings of the AAAI conference on artificial intelligence*, pp. 3187–3194. Available at: https://aaai.org/ojs/index.php/AAAI/article/view/5716 (Accessed: February 20, 2026).

Ayling, B. (2022) *INGENIOUS - Great Basin Regional Dataset Compilation*. 1391. USDOE Geothermal Data Repository (United States); GBCGE, NBMG, UNR. Available at: https://doi.org/10.15121/1881483.

Chang, C. and Lin, C. (2013) "LIBSVM : A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology (TIST)* [Preprint]. Available at: https://doi.org/10.1145/1961189.1961199.

Chen, T. and Guestrin, C. (2016) "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD*. Available at: https://doi.org/10.1145/2939672.2939785 (Accessed: November 27, 2022).

Creason, C.G. *et al.* (2023) "A Geo-Data Science Method for Assessing Unconventional Rare-Earth Element Resources in Sedimentary Systems," *Natural Resources Research*, 32(3), pp. 855–878. Available at: https://doi.org/10.1007/s11053-023-10163-x.

Dobson, P.F. (2016) "A review of exploration methods for discovering hidden geothermal systems," *GRC Transactions*, 40. Available at: https://publications.mygeoenergynow.org/grc/1032385.pdf (Accessed: February 20, 2026).

Drucker, H. *et al.* (1996) "Support Vector Regression Machines," *Advances in Neural Information Processing Systems*. MIT Press. Available at: https://proceedings.neurips.cc/paper_files/paper/1996/hash/d38901788c533e8286cb6400b40b386d-Abstract.html (Accessed: January 29, 2025).

El-Sadi, K. *et al.* (2024) "Review of drilling performance in a horizontal EGS development," *Proceedings, 49th Stanford workshop on geothermal reservoir engineering*. Available at: https://pangea.stanford.edu/ERE/pdf/IGAstandard/SGW/2024/Elsadi.pdf (Accessed: February 20, 2026).

Fercho, S. *et al.* (2025) "Update on the geology, temperature, fracturing, and resource potential at the Cape Geothermal Project informed by data acquired from the drilling of additional horizontal EGS wells," *Proc. 50th Workshop on Geothermal Reservoir Engineering https://pangea. stanford. edu/ERE/pdf/IGAstandard/SGW/2025/Fercho. pdf (Stanford University, 2025)*. Available at: https://pangea.stanford.edu/ERE/pdf/IGAstandard/SGW/2025/Fercho.pdf (Accessed: February 20, 2026).

Huang, Y. *et al.* (2021) "Imaging Complex Subsurface Structures for Geothermal Exploration at Pirouette Mountain and Eleven-Mile Canyon in Nevada," *Frontiers in Earth Science*, 9, p. 782901. Available at: https://doi.org/10.3389/feart.2021.782901.

Jarrett, D. *et al.* (2022) "HyperImpute: Generalized Iterative Imputation with Automatic Model Selection," *Proceedings of the 39th International Conference on Machine Learning*. *International Conference on Machine Learning*, PMLR. Available at: https://proceedings.mlr.press/v162/jarrett22a.html (Accessed: January 29, 2025).

Kliphuis, T. *et al.* (2025) "GeoDAWN to GeoTGo: from Complex Data to Decisions Related to Geothermal Prospectivity," *Stanford Geothermal Workshop*. Available at: https://pangea.stanford.edu/ERE/db/GeoConf/papers/SGW/2025/Kliphuis.pdf.

Lee, D.D. and Seung, H.S. (1999) "Learning the parts of objects by non-negative matrix factorization.," *Nature*, 401(6755), pp. 788–91. Available at: https://doi.org/10.1038/44565.

Norbeck, J.H. and Latimer, T. (2023) "Commercial-scale demonstration of a first-of-a-kind enhanced geothermal system." Available at: https://eartharxiv.org/repository/view/5704/ (Accessed: February 20, 2026).

O'Malley, Daniel and Vesselinov, V.V. (2014) "A Combined Probabilistic/Nonprobabilistic Decision Analysis for Contaminant Remediation," *SIAM Journal on Uncertainty Quantification* [Preprint]. Available at: http://dx.doi.org/10.1137/140965132.

O'Malley, D. and Vesselinov, V.V. (2014) "Groundwater remediation using the information gap decision theory," *Water Resources Research* [Preprint]. Available at: http://doi.org/10.1002/2013WR014718.

O'Malley, D. and Vesselinov, V.V. (2015) "Bayesian-information-gap decision theory with an application to CO2 sequestration," *Water Resources Research* [Preprint]. Available at: https://doi.org/10.1002/2015WR017413.

Pauling, H. *et al.* (2023) *Geothermal Play Fairway Analysis Best Practices*. National Renewable Energy Laboratory (NREL), Golden, CO (United States). Available at: https://www.researchgate.net/profile/Hannah-Pauling-2/publication/374919835_Geothermal_Play_Fairway_Analysis_Best_Practices/links/6536947c24bbe32d9a6559e3/Geothermal-Play-Fairway-Analysis-Best-Practices.pdf (Accessed: February 20, 2026).

Pollack, A., Horne, R. and Mukerji, T. (2020) "What are the challenges in developing enhanced geothermal systems (EGS)? Observations from 64 EGS sites," *Proceedings of the World Geothermal Congress*. Available at: https://www.worldgeothermal.org/pdf/IGAstandard/WGC/2020/31027.pdf (Accessed: February 20, 2026).

Siler, D. *et al.* (2021) *Machine Learning to Identify Geologic Factors Associated with Production in Geothermal Fields: A Case-Study Using 3D Geologic Data from Brady Geothermal Field and NMFk*. 1344. Available at: https://doi.org/10.15121/1832133.

Smith, C.M. *et al.* (2023) "Exploratory analysis of machine learning techniques in the Nevada geothermal play fairway analysis," *Geothermics*, 111, p. 102693. Available at: https://www.sciencedirect.com/science/article/pii/S0375650523000470 (Accessed: February 20, 2026).

USGS (2022) *GeoDAWN: Airborne magnetic and radiometric survey*. Available at: http://doi.org/10.5066/P93LGLVQ.

Vesselinov, V. *et al.* (2021) "Discovering the Hidden Geothermal Signatures of Southwest New Mexico," *World Geothermal Congress 2021*. *World Geothermal Congress 2021*, Reykjavik, Iceland. Available at: https://gitlab.com/monty/monty.gitlab.io/raw/master/papers/Vesselinov et al 2021 Discovering the Hidden Geothermal Signatures of Southwest New Mexico.pdf.

Vesselinov, V. *et al.* (2022) "GeoThermalCloud: Machine Learning for Discovery, Exploration, and Development of Hidden Geothermal Resources," *Stanford Geothermal Workshop*. *Stanford Geothermal Workshop*, Stanford, CA. Available at: https://pangea.stanford.edu/ERE/db/GeoConf/papers/SGW/2022/Vesselinov.pdf.

Vesselinov, V.V. *et al.* (2019) "Unsupervised machine learning based on non-negative tensor factorization for analyzing reactive-mixing," *JCP* [Preprint]. Available at: https://doi.org/10.1016/j.jcp.2019.05.039.

Vesselinov, V.V. *et al.* (2022) "Discovering hidden geothermal signatures using non-negative matrix factorization with customized k-means clustering," *Geotermics* [Preprint]. Available at: https://doi.org/10.1016/j.geothermics.2022.102576.

Vesselinov, V.V. (2023) *GeoThermalCloud.jl: Machine Learning for Geothermal Exploration*. Available at: https://github.com/SmartTensors/GeoThermalCloud.jl.

Vesselinov, Kliphuis.

Vesselinov, V.V. *et al.* (2025) “GeoDAWN & GeoFLIGHT to GeoTGo: From complex data to defensible decisions related to geothermal prospectivity.” Available at: http://doi.org/10.13140/RG.2.2.14194.82881.

Vesselinov, V.V. (2025) *MADS: Model Analysis & Decision Support*. Available at: https://github.com/madsjulia.

Vesselinov, V.V., Alexandrov, B.S. and O’Malley, D. (2018) “Contaminant source identification using semi-supervised machine learning,” *Journal of Contaminant Hydrology* [Preprint]. Available at: https://doi.org/10.1016/j.jconhyd.2017.11.002.

Vesselinov, V.V., Alexandrov, B.S. and O’Malley, D. (2019) “Nonnegative tensor factorization for contaminant source identification,” *Journal of Contaminant Hydrology* [Preprint]. Available at: https://doi.org/10.1016/j.jconhyd.2018.11.010.

Vesselinov, V.V. and O’Malley, D. (2016) “Model Analysis of Complex Systems Behavior using MADS,” *AGU Fall Meeting*. San Francisco, CA.

Walsh, J. and Spandler, C. (2022) “Anomalously high rare earth element (REE) concentrations in zircon: Implications for mineralisation in unconformity-related REE deposits,” *Goldschmidt2022 abstracts*. *Goldschmidt2022*, Honolulu, HI, USA: European Association of Geochemistry. Available at: https://doi.org/10.46427/gold2022.11539.

Young, K. and Akar, S. (2015) “Assessment of New Approaches in Geothermal Exploration Decision Making.” *40th Workshop on Geothermal Reservoir Engineering*. Available at: https://research-hub.nrel.gov/en/publications/assessment-of-new-approaches-in-geothermal-exploration-decision-m/ (Accessed: February 20, 2026).

Zhou, J. *et al.* (2021) “Graph Neural Networks: A Review of Methods and Applications.” arXiv. Available at: https://doi.org/10.48550/arXiv.1812.08434.