



Article

Predicting Rare Earth Element Potential in Produced and Geothermal Waters of the United States via Emergent Self-Organizing Maps

Mark A. Engle ^{1,*} , Charles W. Nye ² , Ghanashyam Neupane ³, Scott A. Quillinan ², Jonathan Fred McLaughlin ², Travis McLing ³ and Josep A. Martín-Fernández ⁴

¹ Department of Earth, Environmental and Resource Sciences, University of Texas at El Paso, 500 West University Ave., El Paso, TX 79930, USA

² Center for Economic Geology Research, University of Wyoming, Laramie, WY 82071, USA; cnye3@uwyo.edu (C.W.N.); scottyq@uwyo.edu (S.A.Q.); derf1@uwyo.edu (J.F.M.)

³ Idaho National Laboratory, Idaho Falls, ID 83415, USA; hanashyam.neupane@inl.gov (G.N.); travis.mcling@inl.gov (T.M.)

⁴ Department of Computer Science, Applied Mathematics and Statistics, University of Girona, 17003 Girona, Spain; josepantoni.martin@udg.edu

* Correspondence: maengle@utep.edu; Tel.: +1-915-747-5503



Citation: Engle, M.A.; Nye, C.W.; Neupane, G.; Quillinan, S.A.; McLaughlin, J.F.; McLing, T.; Martín-Fernández, J.A. Predicting Rare Earth Element Potential in Produced and Geothermal Waters of the United States via Emergent Self-Organizing Maps. *Energies* **2022**, *15*, 4555. <https://doi.org/10.3390/en15134555>

Academic Editors: Renato Somma and Alban Kuriqi

Received: 30 April 2022

Accepted: 20 June 2022

Published: 22 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: This work applies emergent self-organizing map (ESOM) techniques, a form of machine learning, in the multidimensional interpretation and prediction of rare earth element (REE) abundance in produced and geothermal waters in the United States. Visualization of the variables in the ESOM trained using the input data shows that each REE, with the exception of Eu, follows the same distribution patterns and that no single parameter appears to control their distribution. Cross-validation, using a random subsample of the starting data and only using major ions, shows that predictions are generally accurate to within an order of magnitude. Using the same approach, an abridged version of the U.S. Geological Survey Produced Waters Database, Version 2.3 (which includes both data from produced and geothermal waters) was mapped to the ESOM and predicted values were generated for samples that contained enough variables to be effectively mapped. Results show that in general, produced and geothermal waters are predicted to be enriched in REEs by an order of magnitude or more relative to seawater, with maximum predicted enrichments in excess of 1000-fold. Cartographic mapping of the resulting predictions indicates that maximum REE concentrations exceed values in seawater across the majority of geologic basins investigated and that REEs are typically spatially co-associated. The factors causing this co-association were not determined from ESOM analysis, but based on the information currently available, REE content in produced and geothermal waters is not directly controlled by lithology, reservoir temperature, or salinity.

Keywords: emergent self-organizing maps; Critical Minerals; compositional data analysis; neural networks; brines

1. Introduction

Rare earth elements (REE) are considered strategic mineral commodities of the United States [1]. One relatively unexplored source for REEs is high-salinity brines produced from sedimentary basins, often associated with oil and gas production or geothermal exploration and development [2,3]. Recent work, funded by the U.S. Department of Energy, led to development of a new analytical method by Idaho National Laboratory for the low-level quantification of REEs in saline waters [4]. This new method, combined with a sampling campaign led by the University of Wyoming and access to a large catalog of produced waters sample housed by the U.S. Geological Survey, allowed for publication of the most comprehensive dataset currently available on REE concentrations in produced and deep geothermal waters in the United States [4]. As information about REEs in produced waters

generated during oil and gas production and geothermal fluids is significantly lacking, any insight gleaned on the abundance and distribution in waters from other geologic basins informs the potential for future exploration and comparison of various possible REE sources. In an attempt to provide further insight, the objectives of this work are to leverage additional information from this new dataset in two ways: (1) examine patterns in the distribution of REEs in produced and geothermal waters through the use of multivariate data analysis techniques and (2) develop and implement a technique allowing for the estimation of REE potential in produced and geothermal waters of the United States.

Machine learning techniques are becoming more frequent in their use to study mineral resource potential using geochemical data [5,6] due to their ability to find complex patterns and their ability to handle large datasets. Within the field of deep groundwater chemistry, for example, machine learning methods have recently been shown to accurately predict the composition and quantity of produced waters [7], determine the basin of origin for samples of unknown source based solely on major ion chemistry [8], estimate the origin of water for samples which lack traditional geochemical data (e.g., Br, $\delta^{18}\text{O}$, $\delta^2\text{H}$, etc.) to make such determinations [9], determine the spatial and vertical extent of deep groundwaters of various origins from historic datasets [10], and identifying groundwater flow paths and areas of CO_2 sequestration by unmixing end-members [11]. Despite such flexibilities, myriad machine learning algorithms exist and appropriate methods must be identified and utilized for each specific use. Broadly speaking, machine learning methods fall into three categories [12]: supervised (where the grouping of each sample in the training data is known by the algorithm), unsupervised (where the grouping of each sample in the training data is not known by the algorithm), and reinforcement learning (primarily for decision-making, based on trial and error). An excellent overview of various supervised and unsupervised machine learning techniques as applied to geochemistry and other earth science applications is provided by Zuo [5].

In the case of this investigation, we examined various machine learning methods to accomplish project goals. The emergent self-organizing map (ESOM), a subclass of the more general self-organizing map (SOM) algorithm, was selected as the tool to meet both goals of this work. The SOM is an unsupervised system of competitive learning used to sort multivariate data based on similarity (e.g., distance) and structure. The result is typically a one or two-dimensional “map” that captures the variability and patterns in the training data, the dataset used to generate the SOM. The SOM is not a literal cartographic map of sample or data locations in the physical world; rather, it is a representation of structure of the input samples and the associated parameters. Competitive learning refers to the process by which neurons, which make up the map, “compete” for each sample; the process allows for similar data to form into clusters. The SOM is particularly useful for working with high-dimensional datasets that can be more easily interpreted in a lower dimensional visualization, much like principal component analysis (PCA) or cluster analysis. However, the SOM algorithm is more relaxed than PCA in that it operates on variables with missing data and, unlike cluster analysis, allows for visualization of both samples and parameters from the same mapping. Additionally, the SOM allows for mapping of new data onto a pre-existing map, which is potentially useful for statistical modeling. Unsurprisingly, the SOM is increasingly used for a variety of data analysis topics in the earth sciences [10,13–18]. Creation of the SOM begins with generating a lattice of neurons that make up the map, typically arranged in a rectangular or hexagonal geometry. Associated with each neuron is a codebook vector whose length matches the number of dimensions in the dataset being mapped. The size of the map is selected arbitrarily but a traditional SOM is usually rather small, typically smaller than the number of samples in the dataset. For instance, Sun et al. [17] used a 3×3 map for a training dataset of nearly 3000 samples. The ESOM is mathematically nearly identical to the SOM except that it contains thousands of neurons (many more than the number of data used in its creation). Utsch [19] observed that when the number of neurons in the SOM exceeds a certain threshold (~ 4000), it exhibited “emergence”, the concept that a system can generate a higher order through internal

cooperation of its parts. He and his students went on to demonstrate that by exhibiting emergence (even if the number of neurons greatly outnumbered the number of samples in the training dataset), the ESOM are able to separate much more complex patterns than the conventional SOM [20], at the cost of lower computational efficiency [21]. As a tool for predicting missing or unknown values in a large dataset from a smaller, more complete dataset, the ESOM is ideal because it provides a larger range in available predicted values.

For the purposes of this work, the intent was to apply the information gained on REE distribution from a relatively small dataset (the input dataset) to a database that covers a much larger geographic area, through application of the ESOM. One such large dataset is the U.S. Geological Survey Produced Waters Geochemical Database (Version 2.3; herein referred to as the USGS database), which is publicly available and downloadable (available at: <https://eerscmap.usgs.gov/pwapp/>, accessed on 12 October 2018) containing geochemical and related data for roughly 115,000 water samples collected from deep, geologic reservoirs in the United States [22]. The database primarily consists of data of samples from oil and gas wells, but does include a number of data from geothermal reservoirs. While the USGS database is a useful tool for a variety of studies, it contains no data on the concentration of REEs. The USGS database also contains a significant proportion of missing parameters for most samples. Thus, it represents significant challenges for application in any standard estimation or modeling technique (e.g., multivariate regression). Moreover, the input dataset used to create a predictive model contains subpopulations on account of inclusion of data from a broad range of geologic settings. Traditionally, statistical modeling requires pre-segregation of all subpopulations which would mathematically inhibit the ability to perform the mathematical functions required [23]. However, the ESOM algorithm does not contain these limitations, making it an ideal approach for the analysis at hand. The allowance for samples with missing values increases estimation uncertainty, but such samples cannot even be utilized in traditional statistical modeling methods.

As a final comment, the vast majority of data utilized in this section are compositional, meaning that they are relative parts (i.e., concentration data). As such, special care was taken to apply proper techniques of so-called compositional data analysis (CoDA) to prevent the development of spurious or induced correlations [24]. Brines are particularly prone to such issues and have been previously shown to generate unrealistic results or results which lack internal consistency if CoDA methods are not utilized [25–27].

2. Materials and Methods

2.1. Description of Input Data

Two primary sources of data were used in this investigation. The first data source, the input dataset, was used in the creation, training, and cross-validation of the ESOM (a generalized flow diagram of data processing is shown in Figure 1). The second data source is a subset of the USGS database used for estimating REE potential across the United States (see Section 2.6). The input dataset consisted of water quality parameters, REE concentrations, and concentration of other dissolved constituents including ions, for 105 samples of produced water collected from across the United States and 119 samples (no field duplicates or laboratory replicates) of shallow geothermal groundwater and springs in the Eastern Snake River Plain analyzed by Idaho National Laboratory (Table 1). The dataset is available through the U.S. Department of Energy Geothermal Data Repository (Submission 1125). Produced and geothermal waters used as input data come from a broad spatial area of the U.S. and cover a large span of salinities (<1000 to >300,000 mg/L total dissolved solids (TDS)) and origin.

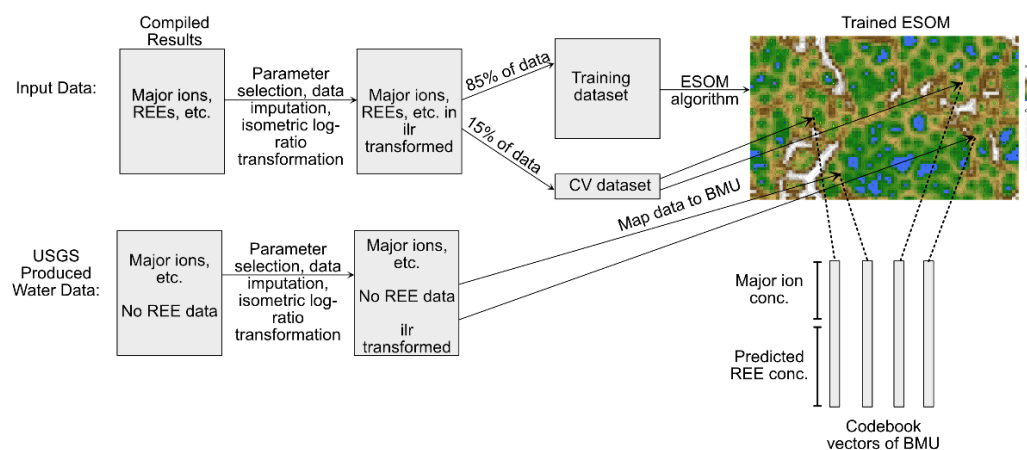


Figure 1. Flow diagram showing steps in processing and analyzing data and making REE concentration predictions using the approach utilized in this research.

Table 1. Source and number of water samples used in the input dataset for the REE prediction potential modeling.

Water Type	Area/Basin	Sample Collector/Source of Data	Number of Samples
Produced/Formation Water	Wind River Basin	Univ. Wyoming	17
Produced/Formation Water	Powder River Basin	Univ. Wyoming	10
Produced/Formation Water	Washakie Basin	Univ. Wyoming	6
Produced/Formation Water	Green River Basin	Univ. Wyoming	6
Produced/Formation Water	Williston Basin	U.S. Geol. Survey	18
Produced/Formation Water	Appalachian Basin	U.S. Geol. Survey	13
Produced/Formation Water	Permian Basin	U.S. Geol. Survey	12
Produced/Formation Water	Kevin Dome	Idaho Nat'l Lab.	23
Geothermal Waters	Eastern Snake River Plain	Idaho Nat'l Lab.	119

2.2. Processing of Input Data

Sixty quantitative parameters and constituents are present in the input data. Selection of variables available for use in machine learning methods is critical. Theodoridis and Koutroumbas [28] argue for including as much data as possible to inform the goal of the process while minimizing redundant variables. Given the nature of studying elements leached from the rock matrix, included variables include those derived from the source (pathfinder elements) and ligands to ensure that solubility is maximized. Review of major ligands for REEs at geothermal temperatures and brine salinities suggested that Cl, OH, and F are dominant [29], so Cl, F, and pH (proxy for OH) were included. Because no ore deposits model exists for REEs in brines, pathfinder elements are unknown. Conventional thinking for prediction models (i.e., regression) is to reduce highly correlated variables, either through variable removal or dimension reduction (e.g., principal component analysis or partial least squares regression). However, CoDA methods were used in this analysis; variables were transformed into ratios and the ratio between other highly correlated variables indicated reactions or processes. Given a complete lack of understanding of pathfinder elements combined with the utility that variations in ratios between correlated elements plays in the approach, a decision was made to include as many variables as possible, which contained a reasonable proportion of uncensored results. Approximately 44% of the data in these 60 parameters were either missing (i.e., not measured or reported; 38%) or censored (6%). Censored data are those in which the value is above or below a certain threshold; in this case censored data correspond to concentrations below the method

or instrument detection limits (i.e., nondetects). To minimize errors caused by inclusion of constituents with few data, all constituents with >50% missing data were removed from the input model. In addition, NO₃ was removed as it is not generally present in produced and formation waters, specific conductance and total dissolved solids were removed as they are highly correlated with the sum of major ions already in the dataset, and oxidation-reduction potential was removed as it is not known to be a reliable measurement in brines. In addition, several parameters were virtually non-existent (As, Al, Th, and U) in the U.S. Produced Waters Database and thus, would not be useful for prediction. The final list of 31 parameters for the input dataset included (Table 2):

1. Water quality parameters—pH (as H⁺) and reservoir temperature;
2. Major, minor, and trace constituents—alkalinity as HCO₃, B, Ba, Br, Ca, Cl, F, K, Li, Mg, Na, Si, SO₄, and Sr;
3. REEs—Sc, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, and Lu.

Table 2. List of parameters used in the input dataset indicating units, and relative proportion of present, censored, and missing data.

Constituent	Units	% Present	% Censored	% Missing
Water Quality Parameters				
Reservoir Temp.	°C	71.4%	0.0%	28.6%
pH	pH units	90.2%	0.0%	9.8%
Major, minor, and trace constituents				
Alkalinity as HCO ₃	mg/L	75.9%	0.0%	24.1%
B	mg/L	92.4%	3.1%	4.5%
Ba	mg/L	87.0%	5.4%	7.6%
Br	mg/L	41.5%	47.3%	11.2%
Ca	mg/L	99.6%	0.5%	0.0%
Cl	mg/L	96.9%	0.5%	2.7%
F	mg/L	71.9%	15.2%	13.0%
K	mg/L	98.7%	1.3%	0.0%
Li	mg/L	88.8%	2.7%	8.5%
Mg	mg/L	95.5%	4.5%	0.0%
Na	mg/L	99.1%	0.0%	0.9%
Si	mg/L	87.2%	2.2%	9.8%
SO ₄	mg/L	84.4%	14.3%	1.3%
Sr	mg/L	92.0%	0.5%	7.6%
Rare Earth Elements				
Sc	ng/L	57.6%	0.0%	42.4%
La	ng/L	97.8%	2.2%	0.0%
Ce	ng/L	98.2%	1.8%	0.0%
Pr	ng/L	96.4%	3.6%	0.0%
Nd	ng/L	96.0%	4.0%	0.0%
Sm	ng/L	98.7%	1.3%	0.0%
Eu	ng/L	99.1%	0.9%	0.0%
Gd	ng/L	97.8%	2.2%	0.0%
Tb	ng/L	96.9%	2.7%	0.5%
Dy	ng/L	99.6%	0.5%	0.0%
Ho	ng/L	98.7%	1.3%	0.0%
Er	ng/L	98.7%	1.3%	0.0%
Tm	ng/L	97.8%	2.2%	0.0%
Yb	ng/L	98.2%	1.8%	0.0%
Lu	ng/L	97.8%	2.2%	0.5%

All 31 parameters except for temperature were converted to units of mg/L (an activity coefficient of 1 was assumed for hydrogen in the conversion of pH).

The resulting dataset contains 4.2% censored data and 4.8% missing data. Missing concentration and pH data were assumed to be missing completely at random (MCAR). Censored data were also missing, but not at random as their value was below a certain

threshold (typically the method detection limit or reporting limit). Nearly 70% of the censored data were associated with five ions: Ba, Br, F, Mg, and SO₄. While the ESOM algorithm does not strictly require imputation of missing values during the initial training, the implementation that was utilized does not operate with null values. The presence of censored and missing data was handled using the following approach:

1. All censored data were initially replaced with a nominal value of 0.65 times the detection limit [30]. Due to high detection limits for samples at higher salinity, two sets of detection limits were used based on the threshold of 1000 mg/L Cl. High- and low-Cl detection limits for Ba, Br, F, Mg, and SO₄ were chosen based on information from the labs, depending on the Cl concentration of the sample (Table 3). For the remaining elements (only 1.3% of the data), the lowest detected concentration for that group (Cl above vs. below 1000 mg/L) was used as the detection limit.
2. Using the censored values from step 1, MCAR values were imputed using a maximum likelihood estimation log-ratio algorithm (lrEM) in the zCompositions package for R [31].
3. Using the estimates of the missing values from step 2, the original estimates of the censored values were discarded and new values were imputed using a maximum likelihood estimation log-ratio algorithm (lrEM) in the zCompositions package in R [31].
4. Steps 2 and 3 were repeated until the algorithm converged. This was determined by converting the data to isometric log-ratios (discussed in Section 2.3) and comparing the difference in the covariance matrix and the mean values of the coordinates between steps in the iteration. Convergence was defined as having a tolerance < 0.0001 for both sets of differences [31].

Table 3. Maximum detection limits used in the imputation of censored data. All results in units of mg/L.

Element	Samples with Cl \geq 1000 mg/L	Samples with Cl < 1000 mg/L
Ba	25	0.05
Br	50	5
F	5	0.5
Mg	25	1
SO ₄	300	10

Given its role in reaction kinetics and thermodynamics, specific attention was given to reservoir temperature. In the case of produced waters, estimated temperatures within the reservoirs were used where available. In the case of the geothermal waters, the sample temperatures were used but assumed to be minimum values due to cooling during travel to and at the surface. Comparison between reported reservoir temperature and the Li-Mg chemical geothermometer estimated temperature [32] suggests that the available reservoir temperatures were consistent with the Li-Mg chemical geothermometer (Figure 2). Any missing reservoir temperatures from the produced water samples were estimated from the Li-Mg chemical geothermometer calculated value. In the case of surface geothermal waters, it was assumed that they have traveled from depth and the Li-Mg chemical geothermometer is a more accurate temperature for reactions that control their composition. For this reason, temperatures for geothermal waters were replaced by those estimated from the Li-Mg geothermometer.

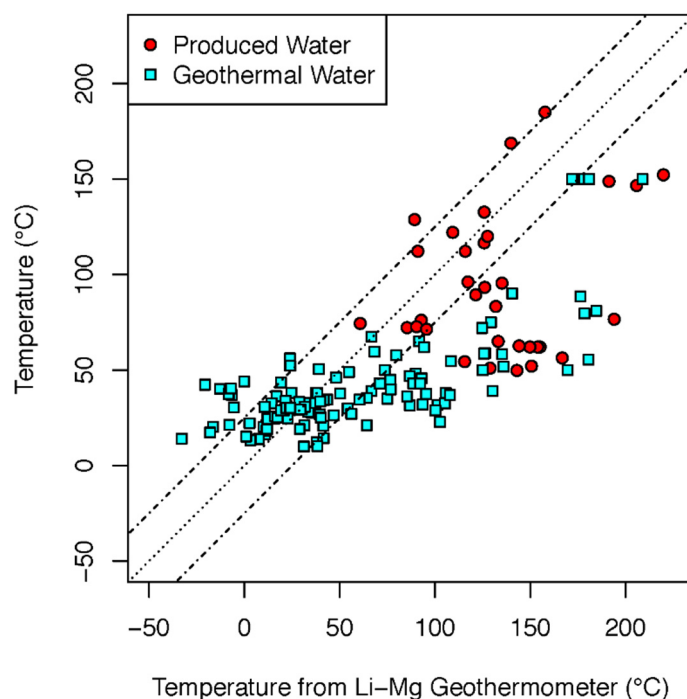


Figure 2. Reservoir and surface temperatures from produced and geothermal waters, respectively, from the input dataset versus temperatures estimated using the Li-Mg geothermometer [32]. Dotted line shows 1:1 line and dashed-dotted lines show 25 °C offsets from the 1:1 line.

2.3. Applying Principles of Compositional Data Analysis

Of the parameters in the input dataset, the clear majority were so-called compositional data, meaning they were relative parts; all concentration data are compositional. Compositional data lie within a lower dimensional subspace in positive real space, called the simplex. Despite being part of positive real space, data within the constraints of the simplex do not follow the standard rules of Euclidean geometry. Thus, all measures based on distances and angles (e.g., correlation, similarity, etc.) using compositional data are incorrect when using standard data analysis methods. Martín-Fernández et al. [33] developed a specific scheme for the treatment of compositional data in the SOM and that approach is used here. Specifically, in the case of brines, previous workers have shown that large variations in salinity produce apparent co-associations between elements which are mathematically induced and, in some cases, completely spurious [26,27,34]. To avoid such problems, compositional data defined as $\mathbf{x} = (x_1, \dots, x_D)$ consisting of D -parts (e.g., constituents) were converted to $D - 1$ isometric log-ratio (ilr) coordinates [27] which follow rules of Euclidean geometry, using:

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{\sqrt[D-j]{\prod_{l=j+1}^D x_l}}{x_j} \text{ for } j = 1, \dots, D-1. \quad (1)$$

Due to the presence of a geometric mean in the numerator of the Equation (1), missing data in the pre-transformed data can generate significant data loss in the ilr coordinates; the geometric mean of a set of samples with missing values cannot be calculated. The USGS database contains a significantly higher proportion of missing data than the input dataset (Table 4). To minimize data loss, all non-REE constituents with >99% of values missing were removed from the input dataset (Al, As, Th, and U). Of the remaining compositional parameters, the data were reordered according to the proportion of missing data in each row (Sc, Nd, Pr, Tb, Lu, La, Gd, Tm, Ce, Yb, Sm, Ho, Er, Eu, Dy, F, alkalinity as HCO_3^- , Si, B, Br, Li, Ba, Sr, SO_4 , H^+ , K, Mg, Ca, Cl, and Na) and then converted to ilr coordinates

using Equation (1) [35]. The ilr coordinates are scaled natural logs, where each element in the list is taken as the denominator to the geometric mean of all the elements to the right of it. That is, the first ilr coordinate (z_1) contains the geometric means of Nd to Na divided by Sc, the second coordinate (z_2) contains the geometric mean of Pr to Na divided by Nd, and so on until and the final ilr coordinate (z_{29}), which contains the ratio of Na to Cl. Reordering of the variables in this manner reduces the effect of missing data in that the likelihood of having all the elements to calculate the geometric mean in the numerator of the transformation increases as the number of missing elements goes down. By using this approach (originally proposed by [36]), conversion of the input data to ilr coordinates increased the number of blank cells only modestly, from 2.38% to 3.51%.

Table 4. Number of concentration data present from the USGS Produced Waters Geochemical Database (Version 2.3).

Constituent	Number of Data	% Present
pH	86,630	75%
Major, minor, and trace constituents		
Al	680	0.6%
Alkalinity as HCO ₃	1691	1.5%
As	493	0.4%
B	4618	4.0%
Ba	12,498	11%
Br	6548	5.7%
Ca	107,478	94%
Cl	108,646	95%
F	1127	1.0%
K	31,550	27%
Li	6126	5.3%
Mg	103,240	90%
Na	96,432	84%
Si	3708	3.2%
SO ₄	93,104	81%
Sr	7812	6.8%
Th	0	0.00%
U	21	<0.01%
Rare Earth Elements		
Sc	0	0.0%
La	0	0.0%
Ce	0	0.0%
Pr	0	0.0%
Nd	0	0.0%
Sm	0	0.0%
Eu	0	0.0%
Gd	0	0.0%
Tb	0	0.0%
Dy	0	0.0%
Ho	0	0.0%
Er	0	0.0%
Tm	0	0.0%
Yb	0	0.0%
Lu	0	0.0%

The version of the ilr transformation used in this case was designed with a focus on minimizing data loss rather than trying to maximize the interpretative ability of the coordinates. Thus in order to provide interpretable information, all results (z) produced using these coordinates were back transformed ($\text{ilr}^{-1}(z)$) into the original units using:

$$x_1 = \exp\left(-\sqrt{\frac{D-1}{D}}z_1\right), \quad (2)$$

$$x_j = \exp\left(\sum_{l=1}^{j-1} \frac{1}{\sqrt{(D-l+1)(D-l)}}z_l - \sqrt{\frac{D-j}{D-j+1}}z_j\right), \text{ for } j = 2, \dots, D-1 \quad (3)$$

and

$$x_D = \exp\left(\sum_{l=1}^{D-1} \frac{1}{(D-l+1)(D-1)}z_l\right). \quad (4)$$

2.4. Univariate and Multivariate Outlier Investigation and Detection

To examine for potentially problematic data (poor imputation methods, invalid analytical results, or erroneous entries), univariate and multivariate outlier analysis methods were applied to the ilr-transformed data. Univariate investigation included generation of an exploration data analysis plot following Reimann et al. [37] for each ilr coordinate and subsequent inspection. Multiple populations were observed for ilr coordinates z_{19} (B in the denominator), z_{24} (SO₄ in the denominator), and z_{27} (Mg in the denominator), and appear to represent clear geochemical differences between geothermal and produced waters. Visual examination of the EDA plots for all ilr-transformed data showed no evidence of significant outliers or extreme univariate values. An adaptive chi-square distance threshold of minimum covariance determinant-based Mahalanobis distance of ilr coordinates [38] was applied to complete cases of the REE data in the input dataset. Examination of the results shows a handful of groups of samples with large robust Mahalanobis distances, but no particularly unusual samples. Because the ESOM algorithm is nonlinear and the analytical reports from Idaho National Laboratory indicated no apparent analytical problems with these samples, the data were kept in the dataset.

2.5. Model Cross-Validation

Cross-validation methods were applied in an attempt to quantify uncertainty through the application of the ESOM to estimate REE values in unknown samples. Rows from the input dataset were randomly split between the training dataset (85% of rows; $n = 190$) and the cross-validation dataset (15% of rows, $n = 34$). The former was used to train the ESOM model and is the basis for all prediction (Figure 1). The latter was used to check the ability of the ESOM model to accurately predict REE concentrations for “blind” samples. The cross-validation data were mapped to trained ESOM (described in the next section) to generate predicted REE concentrations (Figure 1). The predicted concentrations were compared against the known REE concentrations for the cross-validation dataset, allowing for calculation of model error.

2.6. Creation of a Trained Emergent Self-Organizing Map

As the ESOM algorithm relies on distance as its metric of similarity, the ilr-transformed data (z) (except for temperature, which was kept in its original units) were normalized by the mean and the sample standard deviation of each ilr coordinate (\hat{z}_i):

$$\hat{z}_j = \frac{z_j - \bar{z}_j}{s_j}, \text{ for } j = 1, \dots, D-1, \quad (5)$$

where \bar{z}_j is the arithmetic mean of the j th coordinate and s_j is the sample standard deviation of the j th coordinate. Normalization provides an equal “weight” of every single coordinate and prevents individual ilr coordinates from having an unusually large impact on mapping. Data produced from ESOM results can be un-normalized through the inverse transformation:

$$z_j = \hat{z}_j s_j + \bar{z}_j. \quad (6)$$

The values for \bar{z}_j and s_j from the training dataset also were used for normalization of the data mapped onto the trained ESOM in later steps. Data for temperature, which is not compositional, were also normalized using Equation (5).

The ESOM lattice applied here consists of 82×50 neurons (4100 total), which greatly outnumbers the 190 samples in the training dataset. In addition, unlike the SOM, the ESOM is typically created so that the top of the map is connected to the bottom and the right-hand side connected to the left, as if one has unpeeled the outer layer from a donut and laid it out flat on a table.

Starting with our normalized data for temperature and the ilr coordinates (\hat{z}) (a total of m columns), a codebook vector of length m was created for each neuron and filled with random values. The SOM algorithm was then run as follows (modified from [19]):

1. An input vector \hat{z}_i is selected at random from the training dataset and the Euclidean distances between it and all codebook vectors for all the neurons on the map are computed. Note that the Euclidean distance of ilr-transformed variables is equal to the Aitchison distance of untransformed compositional data.
2. The input vector \hat{z}_i is assigned to the codebook vector of the neuron that is closest to it. This neuron is known as the best matching unit (BMU). A Gaussian function is used to define the neighborhood of nearby neurons around the BMU. With each iteration of the algorithm, radius of the neighborhood decreases.
3. The codebook vectors of those neurons within the neighborhood are reweighted to be more similar to \hat{z}_i using one of several possible functions. Typically, codebook vectors of neurons closer to BMU are more heavily reweighted than those more distal. The amount of re-weighting (learning weight) also decreases over time.
4. The next input vector, $\hat{z}_i + 1$, is randomly selected and steps 1–3 are repeated. Once all the input vectors have been mapped (1 epoch), they are removed from the map and the process is repeated starting back at step 1. The learning is continued for a set number of epochs. Because the neighborhood and reweighting function both decrease with each epoch, the map stabilizes with an increasing number of iterations.

The exact parameters applied in the generation of the ESOM shown here are provided in Table 5. These are the default parameter values suggested by Ultsch and Hermann [39] and were developed by testing values for a complex set of benchmark datasets, called the Fundamental Clustering Problems Suite, and thus are considered suitable for most datasets. All ESOM calculations were made using Version 3.1 of the Umatrix package in R [40]. In the example presented here, weights for a neuron's codebook vector at the next step in the iteration (\mathbf{W}_{t+1}) are calculated from the corresponding weight at time t (\mathbf{W}_t) via:

$$\mathbf{W}_{t+1} = \mathbf{W}_t + N_t L_t (\mathbf{X}_{it} - \mathbf{W}_t), \quad (7)$$

where L_t is the learning rate at time t , and N_t is the neighborhood function.

Table 5. Parameters used in the ESOM algorithm.

Parameters	Method or Setting
Number of rows in map	82
Number of columns in map	50
Number of training epochs	24
Initialization method	Uniform random values from mean ± 2 standard deviations
Shape	Toroidal
Neighborhood function	Gaussian
Starting neighborhood radius	24
Ending neighborhood radius	1
Neighborhood cooling function	Linear
Starting learning rate	0.5
Ending learning rate	0.1
Learning rate cooling function	Linear

The neighborhood function is defined as:

$$N_t = \exp\left(\frac{\|r_c - r_i\|}{2\sigma_t^2}\right), \quad (8)$$

where the numerator is the Euclidean distance between the codebook vector of the BMU (r_c) and that of any random neuron in the neighborhood (r_i), and σ_t is the radius of neighborhood at time t . Both L_t and σ_t decrease linearly throughout the run (Table 5). This decrease in L_t and σ_t ensures that in the early stages of the process, changes in the map are dramatic but near the end of the run, they become less impactful and the map starts to stabilize. When the trained ESOM is created (i.e., after algorithm has run through its defined number of epochs), the final weights from the codebook vectors for all the neurons in the ESOM are un-normalized using Equation (6) and the mean and standard deviation. For all compositional variables, the weights are converted from ilr coordinates back into the original variables using the inverse transformation (Equations (2)–(4)). Note that these back-transformed data are in units of proportion (0–1). Compositional data analysis does not distinguish between units of scale because corresponding data are of the same equivalence class.

2.7. Predicting Rare Earth Element Potential Using the Trained Emergent Self-Organizing Map

Prediction using trained SOM analysis works by taking a new set of data points (not using in training) that are missing one or more values (the parameters to be predicted) and assigning them a BMU based on the multivariate distance for the elements that are present (Figure 1). Dickson and Giblin [13] used this same approach to predict uranium concentrations in groundwater. Our problem was more complex in that not only does the USGS database contain no data on REEs, it also contains missing values for many other constituents (as summarized in the documentation by Blondes et al. [22]). Using the same parameters as those used to create the trained ESOM using the training dataset, the compositional data from the USGS database were converted to ilr coordinates using Equation (1) using the same singular binary partition. The ilr coordinates and reservoir temperature data, estimated using the Li-Mg chemical geothermometer, were normalized using the mean and standard deviation from the training dataset (Equation (5)). In a balance between retaining enough data to be effective and requiring enough non-missing parameters to be meaningful, all data containing fewer than 7 of the ilr-variables were removed from the analysis, shrinking the total dataset from roughly 115,000 to 3688 data points. If a larger cutoff were used, the number of points dropped off precipitously; only 826 samples (~0.7% of the total database) contained non-missing data for more than 8 ilr coordinates. The normalized and transformed data in this abridged version of the USGS database were mapped to the trained ESOM by finding the neuron whose codebook vector had the shortest distance to the input vector z_i for each sample (i.e., the BMU). The missing values for each input vector were then taken from the corresponding element in the codebook vector of the respective BMU. The resulting data were converted to un-normalized values via Equation (6), using the mean and standard deviation from the training dataset, and the compositional parameters were back-transformed into the original variables using Equations (2)–(4). The resulting compositional data are proportional; to convert them back into units of mg/L, each row was multiplied by the sum of the non-missing compositional data in the original USGS dataset.

3. Results

3.1. Trained Emergent Self-Organizing Map

Arrangement of the training data using the ESOM algorithm showed that geothermal waters were distinguished from produced waters based on their multivariate geochemical structure (Figure 3). Data for samples from the Washakie and Wind River Basins exhibited overlap with those from the Williston and a portion of samples from the Appalachian

Basin. Samples from the Permian and Appalachian basins showed less similarity among themselves than the other sampling areas.

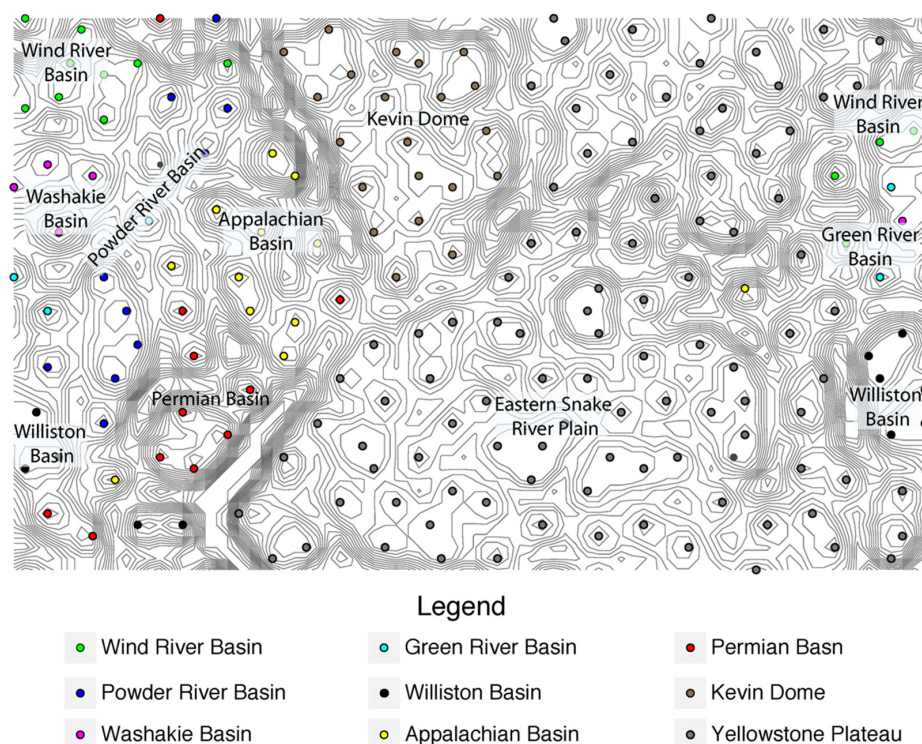


Figure 3. Foreground: Visualization of BMUs for the input data on the trained ESOM. Background: Contours of U-heights, the distance of each neuron and its neighbors. Points color-coded to indicate general sampling location.

The U-matrix map (Figure 3) displays contours of the U-heights, the sum of all distances between each neuron and its nearest neighbors, for each neuron within in the trained ESOM [41]. In this visualization of the data, features mapped as “mountain ranges” can be thought of as boundaries between clusters of data while those that appear as “valleys” represent the center of data clusters. Thus, the data from the Eastern Snake River Plain appear as their own cluster, distinct from the produced waters that plot around them. Similarly, samples from the Kevin Dome region in Montana also appear as their own cluster. Several of the Permian Basin samples, noted by features with a sink-hole appearance around several of the samples in the lower left corner of the map, suggest these samples are dissimilar to most other samples in the dataset. However for the most part, the produced water samples generally tended to overlap among basins. This is not surprising because similar processes tend to dominate the composition of produced waters, regardless of geographic location [42].

Patterns in the relative distribution of the constituents used in the training dataset can be examined through visualization of the back-transformed weights from the codebook vectors of the trained ESOM (Figure 4). Mapped compositional parameters are in units of mass proportion and have been rescaled over the range [0, 1]. In the case of reservoir temperature, the map was simply been rescaled over the range [0, 1]. Comparison of mapped weights for Pr (a light REE) and Yb (a heavy REE) demonstrate that the REEs are distributed similarly among the samples in the training dataset. This is remarkable given that samples come from both geothermal and hydrocarbon waters and cover a diverse range of geologic systems (Table 1). Because the ilr coordinates in this analysis were normalized using Equation (5), the apparent co-association between the REEs appears to not simply be an artifact due to them exhibiting a larger log-ratio variance than the other constituents included in the input dataset. Europium, which unlike the other REEs

exhibits both 2+ and 3+ oxidation states, shows a somewhat different pattern in its mapped weights, which is expected given its unique geochemical behavior relative to the other REEs. Maps for weights of reservoir temperature, Cl (a proxy for salinity), and H (a proxy for pH) suggest that none appear to independently control REE distribution among the samples. Previous studies have shown pH is a significant control on REE concentrations, with elevated concentrations at and below ~ 3.5 [43]. Only $\sim 1.1\%$ of the data in the abridged USGS database exhibited pH values below 3.5, minimizing any apparent control by pH. Of the remaining elements, two that most closely followed the patterns of the REE weight were F and Si. Thermodynamic calculations suggest that F can be an important ligand for REEs in F-rich, high-temperature aqueous systems [29]. Comparison of the weights for the REEs versus an overlay of the F and reservoir temperature maps do not overlap well, suggesting that if F complexation does play a role in controlling REE abundances, it is not limiting. Similarly, Si concentrations in deep geothermal and hydrocarbon reservoirs are strongly controlled by temperature-dependent reactions with silica polymorphs [32,44]. However, comparison of the mapped weights for Si versus reservoir temperature suggests that Si is not entirely temperature-controlled in these systems. Alternatively, Si may be supply-limited in non-clastic or mafic systems. In this scenario, the positive relationship between REE abundances and Si may indicate the significant role that specific clastic- or felsic-rich reservoirs play as the source for REEs.

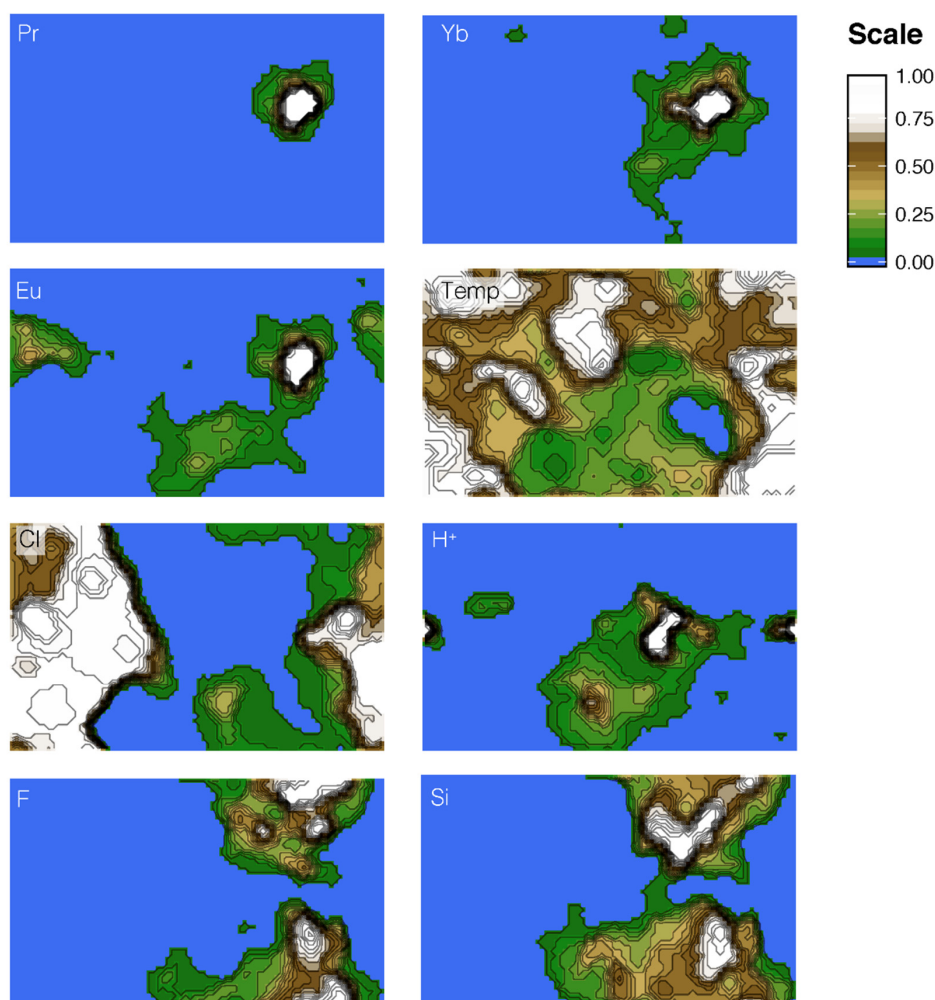


Figure 4. Visualizations of weights from the codebook vectors of the trained ESOM for Pr (proxy light REE), Yb (proxy heavy REE), Eu, estimated reservoir temperature, Cl, H^+ , F, and Si. Compositional data are in units of relative proportion and were rescaled over the range [0, 1] for mapping. Temperature in units of $^{\circ}C$ (from ~ 12 to 230 $^{\circ}C$; Figure 2) were rescaled over the range [0, 1] for mapping.

3.2. Cross-Validation Results

Each ilr-transformed (except for temperature) and normalized data point in the cross-validation set (as described in Section 2.4) was mapped to the neuron on the trained ESOM with the smallest distance to itself (analogous to the BMU during the ESOM training; Figure 1). To emulate the worst-case scenario, only major ion data and reservoir temperature data were used for prediction (ilr coordinates z_{23} to z_{29}). The REE values from these neurons were then assigned to each data point. All data were then un-normalized via Equation (6), and the compositional parameters were back-transformed via Equations (2)–(4). The resulting data are proportional; to convert them back into units of mg/L each row was multiplied by the sum of the compositional data in the corresponding row of the input dataset. Cross-validation performance (CV_{error}) was determined by normalizing the data to seawater (NPDW, North Pacific Deep Water) [45] and then taking the ratio of the ESOM-predicted values to the known values (Figure 5):

$$CV_{\text{error}} = \frac{x_{\text{ESOM}}}{x_{\text{known}}}, \quad (9)$$

where x_{ESOM} is the concentration of a given REE as predicted from the ESOM and x_{known} is the known concentration of the given REE for that sample from the input database. Results suggest that even for this simple model built using fewer than 200 data points for training, the ESOM can generally predict REE concentrations within an order of magnitude. Results were modestly better for La, Gd, and Yb (i.e., smaller spread) and, in general, results tended to under-predict concentrations, especially for Ce, Sm, Eu, and Tb, suggesting the estimates are conservative.

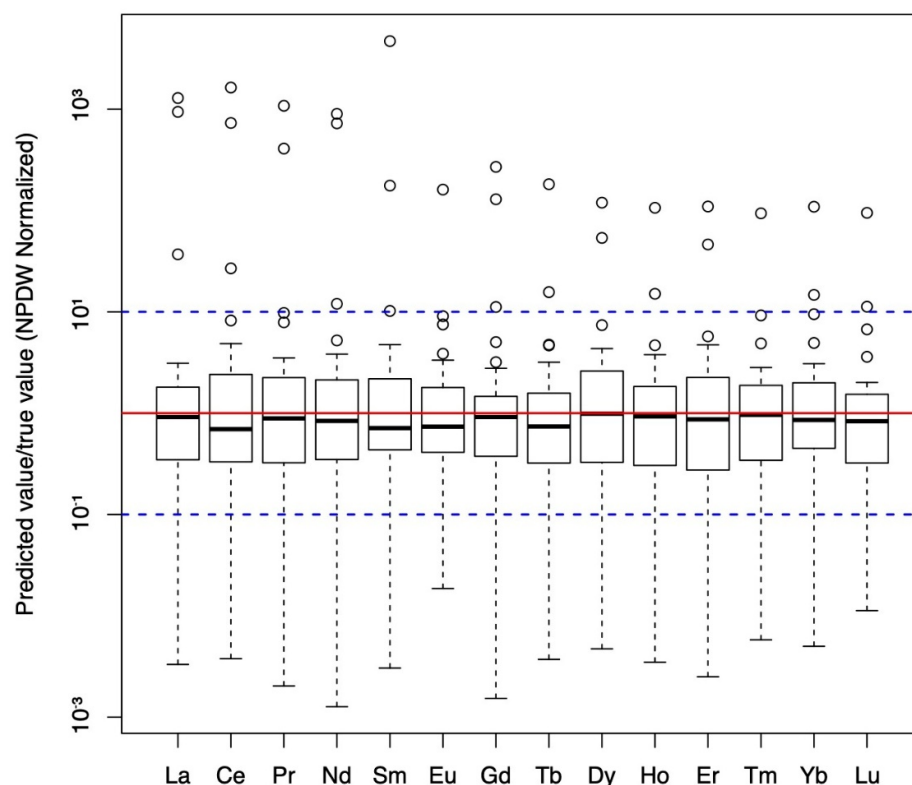


Figure 5. Prediction results for REEs in the cross-validation dataset based only from major ions (z_{23} to z_{29}) and reservoir temperature. Upper and lower boundaries of each box denote the 3rd and 1st quartiles, respectively, thick black line denotes median, and whiskers are drawn out to 1.5 times the interquartile range beyond the 3rd and 1st quartiles. Circles indicate values beyond the whiskers. Red line ($y = 1$) indicates perfect prediction and blue dashed lines ($y = 0.1, 10$) indicate under- and over-prediction of NPDW-normalized data by an order of magnitude, respectively.

3.3. Rare Earth Element Potential Prediction

In general, results from REE concentration prediction for the abridged USGS database show that produced and deep geothermal waters of the United States are enriched in REEs compared to seawater (Figure 6). The markedly high predicted enrichments in Ce and Eu are due to depletion of these elements in seawater due to Ce oxidation and scavenging in the upper seawater column [45] and release of Eu from feldspars into associated formation waters present in clastic reservoirs during weathering and diagenetic reactions. In general, the largest average predicted enrichments in produced and geothermal waters relative to NPDW were observed in light REEs (La to Gd), with generally less enrichment in the heavy REEs (Tb to Lu). This pattern suggests influence from sedimentary rock interaction, which has a higher light REE to heavy REE ratio than seawater.

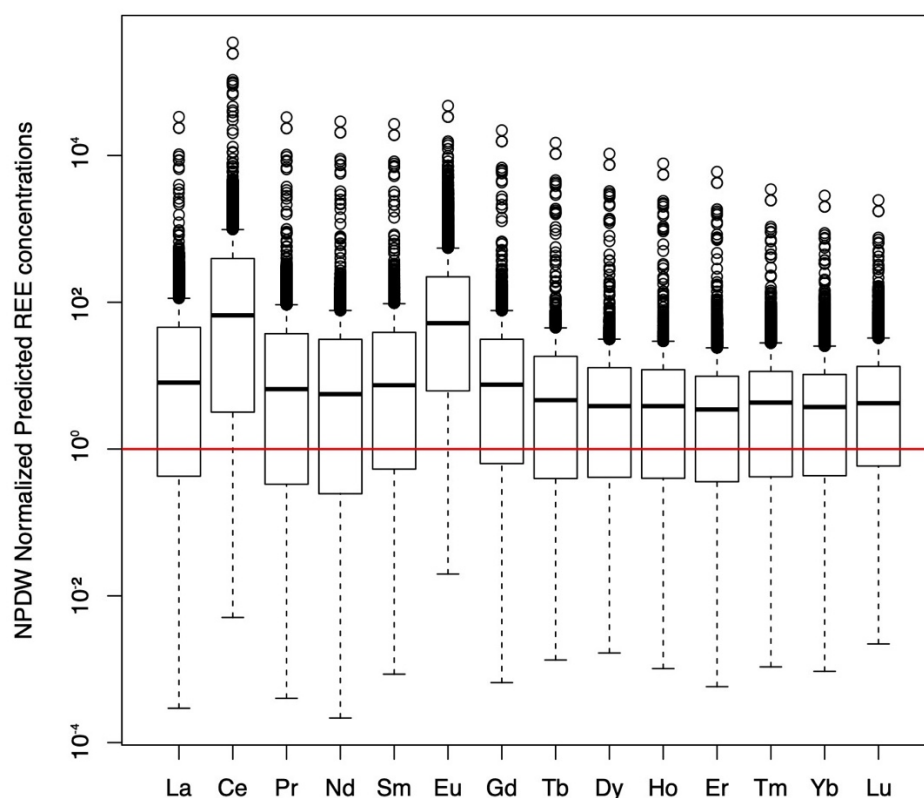


Figure 6. Predicted REE concentrations in produced and geothermal waters from the abridged USGS database relative to seawater (NPDW normalization). Thick red line denotes median and whiskers are drawn out to 1.5 times the interquartile range beyond the 3rd and 1st quartiles. Circles indicate values beyond the whiskers.

Comparison of ESOM weights shows that similar patterns are observed for light REEs other than Eu versus heavy REEs, with Eu behaving independently due to its 2+ and 3+ redox states. To understand spatial controls and patterns in predicted REE concentrations from the ESOM model, cartographic maps were created showing predicted NPDW-normalized concentrations (i.e., predicted potential). Praseodymium was used as a surrogate for light REE behavior, Yb was used a surrogate for heavy REE behavior, and Eu was also mapped due to its individual behavior. The data were categorized into low, medium, high, and highest categories based on the relative magnitude of the predicted potential values. Categorical ranges were developed for each surrogate individually, to account for variations in the ranges of predicted potential (Figure 6). Note that 290 samples in the abridged USGS dataset did not contain latitude and longitude information and could not be mapped.

Maps of predicted mineral potential for Pr, Yb, and Eu (Figures 7–9) are largely similar with the highest values focused primarily in the Permian Basin (southeast New Mexico and west Texas), Anadarko Basin (Oklahoma), Texas-Louisiana Salt Basin, Appalachian Basin (eastern Ohio and western Pennsylvania), Illinois Basin (southern Illinois and southern Indiana), Michigan Basin (Michigan), Powder River Basin (Wyoming) and a few points spread out among basins in southern Utah. In general, these basins exhibit higher salinity values and all, except those in Wyoming and Utah, contain rocks and produced waters of marine origin [42,46]. Notable differences for Eu include highest predicted potential also in the Michigan and Williston (North Dakota and Montana) basins. In terms of reservoir lithology, samples associated with the 25 highest predicted Pr and Yb potential largely overlap and include clastic sandstone and shale reservoirs (e.g., Clinton sandstone of the Appalachian Basin, Wilcox and Bromide sands of the Anadarko Basin, Aux Vases and Tar Springs sandstones of the Illinois Basin, Yeso, Artesia, and Abo groups of the Permian Basin, and Muddy sandstone in the Wind River Basin) and carbonate reservoirs (Smackover Formation of the Texas-Louisiana Salt Basin, Ste. Genevieve limestone of the Illinois Basin, Wolfcamp and Cisco limestone of the Permian Basin, and Morrow limestone in the Anadarko Basin). Additionally, no consistent trends with depth (a proxy for temperature) were observed with depths ranging from ~60 m to over 3500 m, which agrees with the ESOM weight maps showing no clear relationship between temperature and relative REE proportion (Figure 4). Total dissolved concentrations also range substantially in the 25 samples with the highest predicted Pr and Yb potentials, from ~2160 to 345,000 mg/L, suggesting that salinity does not directly appear to serve as a control. Slightly different patterns are observed with respect to Eu; the 25 samples with highest predicted Eu mineral potential are all very saline (range = 42,200 to 341,000 mg/L; median = 238,000 mg/L) but of variable depth (~300 to ~3000 m below ground surface). This finding does not immediately identify lithology as the sole control on predicted REE potential in such waters.

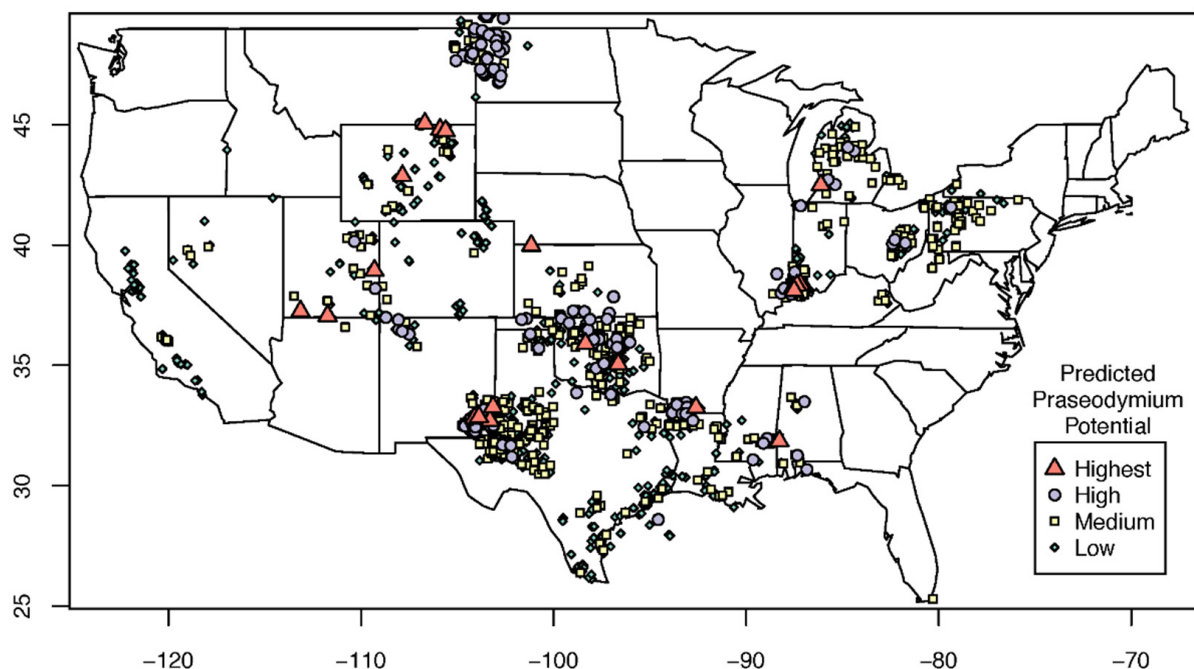


Figure 7. Predicted Pr potential of produced and geothermal waters. Predicted potential categories: highest ≥ 500 , high = 100–499, medium = 10–99, low < 10 .

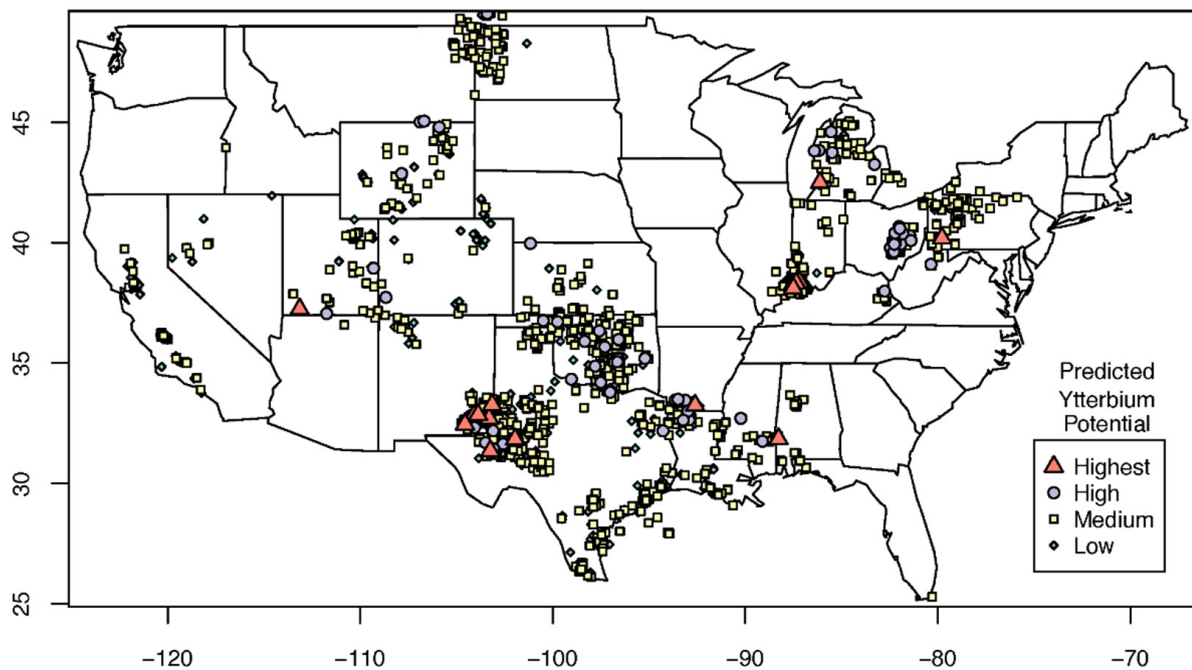


Figure 8. Predicted Yb potential of produced and geothermal waters. Predicted potential categories: highest ≥ 125 , high = 50–124, medium = 1–49, low < 1 .

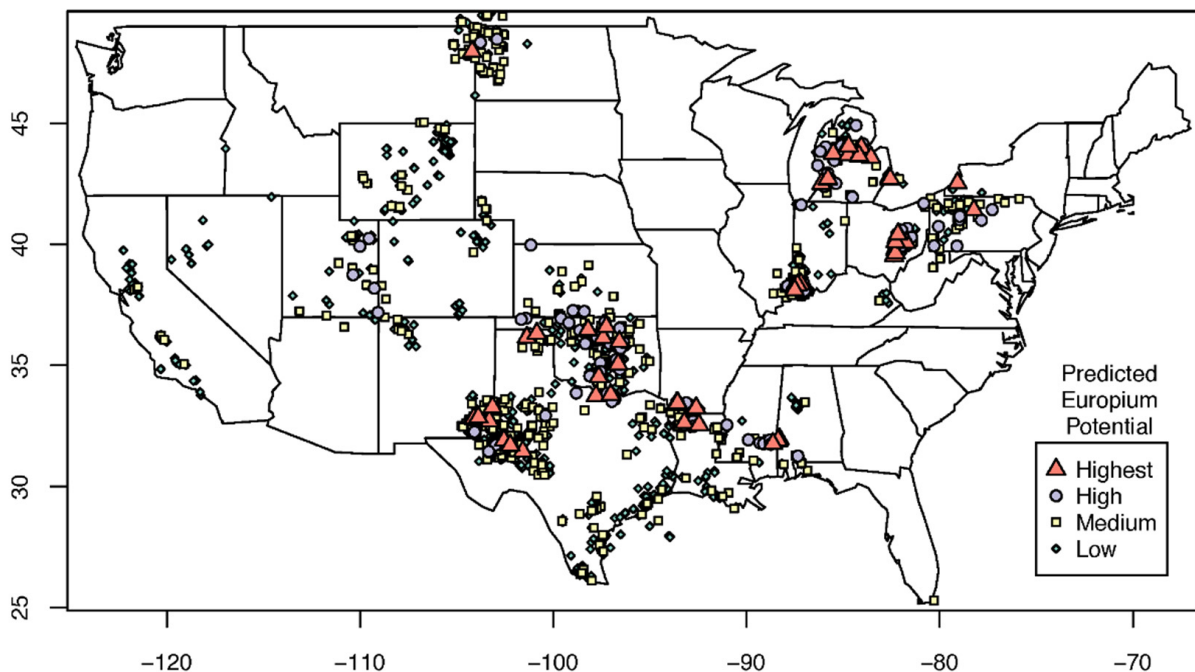


Figure 9. Predicted Eu potential of produced and geothermal waters. Predicted potential categories: highest ≥ 125 , high = 50–124, medium = 1–49, low < 1 .

These findings indicate that predicted REE potential is high across many oil and gas producing basins in the United States and is not directly controlled by lithology, reservoir temperature, or salinity (except for perhaps Eu). However, the strong co-association among most of the REEs except for Eu (Figure 4) suggests that samples that contain elevated concentration of any single REE also contain elevated concentrations of most, if not all, remaining REEs. This knowledge reduces the analytical effort required in potential future geochemical prospecting because analysis of only one of two REEs should provide enough

information to inform the behavior of the remaining elements. Similar observations of co-association have been reported for REE accumulations in other media, including seafloor muds [47]. Co-association of REEs informs extraction approaches because potentially economic sources would necessitate processes to separate the REEs from one another. To that effect, efforts to improve extraction and separation of REEs from geological sources are ongoing.

In terms of possible economic benefit, the U.S. Geological Survey summarizes information on values for many mineral commodities [48]. Its 2017 estimates suggest that Dy and Tb are by far the most individually economic REEs at roughly USD 180–190 per kg for Dy_2O_3 and USD 470–480 per kg for Tb_2O_3 . Assuming 100% removal of both Dy and Tb from produced or geothermal water samples, none of the samples is predicted to be worth more than a USD 0.01/barrel (1 barrel = 159 L). However, such calculations are made entirely from mathematically estimated results and should be taken with extreme caution. At a minimum, economic extraction would require mineral commodity values in excess of the disposal costs for the remaining waste brine. Disposal costs for brines within the United States are not well quantified and vary by region, but generally span the range of roughly USD 0.10 to more than USD 4.00 per barrel [49].

4. Discussion

Results from this investigation demonstrate that the combined approach of ESOM training using a training dataset following by mapping data for which REE concentrations are unknown to the nearest BMU and taking predicted values from the corresponding codebook vector is successful in producing first-order estimates of these values (typically within an order of magnitude). While the method is not optimized necessarily for prediction, this is an impressive feat in that trace element concentrations are being estimated from major constituents entirely through higher-level structures found in the data.

Beyond prediction itself, analysis of the trained ESOM shows that concentrations of REEs, except Eu, in produced and geothermal waters are strongly co-associated with one another and no single parameter used in the model serves as a strong pathfinder or predictor of REEs. Co-associations between high relative abundances of REEs and F were observed, suggesting a role for REE–F complexation, but F concentration does not appear to be limiting. Co-associations between Si and REEs were also observed, suggesting that specific clastic- or felsic-rich reservoirs serve as a lithologic source, suggesting the potential for further investigation.

In general, our prediction results indicate that produced and geothermal waters are enriched in REEs by an order of magnitude or more relative to seawater, up to and exceeding 1000 times. The largest average predicted enrichments in produced and geothermal waters relative to NPDW were observed in light REEs (La to Gd), with generally less enrichment in the heavy REEs (Tb to Lu). Cerium and Eu exhibited the highest predicted potentials due to Ce depletion in seawater and release of Eu from reservoir materials during weathering and diagenetic reactions. The economic worth of the two most valuable REEs, Dy and Tb, is estimated to be significantly less than typical costs required to dispose of the waste brines remaining after REE removal. Spatial mapping shows predicted elevated enrichments across many geologic basins of the United States, rather than being concentrated in specific regions. The REEs are typically spatially co-associated, but based on the information currently available, their concentrations appear to not be directly controlled by lithology, reservoir temperature, or salinity, suggesting that further work is required to better understand control on their behavior at such large scales.

In terms of future work, new produced and geothermal water data containing REE concentrations can be added to the input dataset used here, to generate an even more robust ESOM. In turn, this can allow for better examination on the controlling variables for REEs in waters from deep sedimentary basins. Moreover, the trained ESOM and a table containing the predicted REE potential of samples in the USGS database created as part of this work are available as a digital download through the U.S. Department of Energy

Geothermal Data Repository. Other datasets in need of REE prediction can be mapped to the ESOM to predict values using the methods described here. With increasing data and more flexible techniques such as the ESOM, methods to estimate resource commodities will only improve.

Author Contributions: Conceptualization, M.A.E. and C.W.N.; methodology, M.A.E. and J.A.M.-F.; formal analysis, M.A.E. and J.A.M.-F.; data production and curation, G.N. and T.M.; writing—original draft preparation, M.A.E., C.W.N. and J.A.M.-F.; writing—review and editing, M.A.E.; visualization, M.A.E.; project administration, S.A.Q.; funding acquisition, S.A.Q., J.F.M. and T.M. All authors have read and agreed to the published version of the manuscript.

Funding: Funding for this work was provided by U.S. Department of Energy Geothermal Technologies Office under Award DE-EE0007603 and the U.S. Geological Survey Energy Resources Program (Engle).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The codebook vectors of the trained ESOM and a table of predicted REE potential for the USGS produced waters database are available for download as a USGS Data Release (<https://doi.org/10.5066/P9GCKYKG0>). The raw geochemical data used in this investigation are available from the U.S. Department of Energy’s Geothermal Data Repository (<https://gdr.openet.org/submissions/1125>, accessed on 1 June 2019).

Acknowledgments: The authors would like to thank our DOE project managers Holly Thomas and Josh Mengers and also our Technical Monitoring Team for their support, advice, and insight. Madalyn Blondes and Ricardo Olea (both USGS) and anonymous journal reviewers provided helpful feedback and suggestions on an earlier version of this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. National Research Council. *Minerals, Critical Minerals, and the US Economy*; National Academies Press: Washington, DC, USA, 2008; 264p.
2. Gosselin, D.C.; Smith, M.R.; Lepel, E.A.; Laul, J. Rare earth elements in chloride-rich groundwater, Palo Duro Basin, Texas, USA. *Geochim. Cosmochim. Acta* **1992**, *56*, 1495–1505. [[CrossRef](#)]
3. Tian, L.; Chang, H.; Tang, P.; Li, T.; Zhang, X.; Liu, S.; He, Q.; Wang, T.; Yang, J.; Bai, Y.; et al. Rare earth elements occurrence and economical recovery strategy from shale gas wastewater in the Sichuan Basin, China. *ACS Sustain. Chem. Eng.* **2020**, *8*, 11914–11920. [[CrossRef](#)]
4. Quillinan, S.; Nye, C.; Engle, M.A.; Bartos, T.; Neupane, G.; Brant, J.; Bagdonas, D.; McLing, T.; McLaughlin, J.F. *Assessing Rare Earth Element Concentrations in Geothermal and Oil and Gas Produced Waters: A Potential Domestic Source of Strategic Mineral Commodities*; Project Final Report for U.S. Department of Energy; U.S. Department of Energy: Washington, DC, USA, 2019; 189p.
5. Zuo, R. Machine learning of mineralization-related geochemical anomalies: A review of potential methods. *Natl. Resour. Res.* **2017**, *26*, 457–464. [[CrossRef](#)]
6. Brown, W.; Groves, D.; Gedeon, T. Use of fuzzy membership input layers to combine subjective geological knowledge and empirical data in a neural network method for mineral-potential mapping. *Natl. Resour. Res.* **2003**, *12*, 183–200. [[CrossRef](#)]
7. Jiang, W.; Pokharel, B.; Lin, L.; Cao, H.; Carroll, K.C.; Zhang, Y.; Galdeano, C.; Musale, D.A.; Ghurye, G.L.; Xu, P. Analysis and prediction of produced water quantity and quality in the Permian Basin using machine learning techniques. *Sci. Total Environ.* **2021**, *801*, 149693. [[CrossRef](#)] [[PubMed](#)]
8. Shelton, J.L.; Jubb, A.M.; Saxe, S.W.; Attanasi, E.D.; Milkov, A.V.; Engle, M.; Freeman, P.A.; Shaffer, C.A.; Blondes, M.S. Machine learning can assign geologic basin to produced water samples using major ion geochemistry. *Natl. Resour. Res.* **2021**, *30*, 4147–4163. [[CrossRef](#)]
9. Engle, M.A.; Chaput, J.A. Groundwater origin determination in historic chemical datasets through supervised compositional data analysis: Brines of the Permian Basin, USA. In *Advances in Compositional Data Analysis*; Filzmoser, P., Hron, K., Martín-Fernández, J.A., Palarea-Albaladejo, J., Eds.; Springer: Cham, Switzerland, 2021; pp. 265–283. [[CrossRef](#)]
10. Engle, M.A.; Brunner, B. Considerations in the application of machine learning to aqueous geochemistry: Origin of produced waters in the northern U.S. Gulf Coast Basin. *Appl. Comput. Geosci.* **2019**, *3–4*, 100012. [[CrossRef](#)]
11. Shaughnessy, A.R.; Gu, X.; Wen, T.; Brantley, S.L. Machine learning deciphers CO₂ sequestration and subsurface flowpaths from stream chemistry. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 3397–3409. [[CrossRef](#)]
12. Russell, S.; Norvig, P. *Artificial Intelligence—A Modern Approach*, 3rd ed.; Pearson Education: London, UK, 2010; 1132p.

13. Dickson, B.L.; Giblin, A.M. An evaluation of methods for imputation of missing trace element data in groundwaters. *Geochem. Explor. Environ. Anal.* **2007**, *7*, 173–178. [CrossRef]
14. Lacassie, J.P.; Roser, B.; Solar, J.R.D.; Herve, F. Discovering geochemical patterns using self-organizing neural networks: A new perspective for sedimentary provenance analysis. *Sediment. Geol.* **2004**, *165*, 175–191. [CrossRef]
15. Lacassie, J.P.; Solar, J.R.D.; Roser, B.; Herve, F. Visualization of volcanic rock geochemical data and classification with artificial neural networks. *Math. Geol.* **2007**, *38*, 697–710. [CrossRef]
16. Liu, Y. Patterns of ocean current variability on the west Florida Shelf using the self-organizing map. *J. Geophys. Res.* **2005**, *110*, C06003. [CrossRef]
17. Sun, X.; Deng, J.; Gong, Q.; Wang, Q.; Yang, L.; Zhao, Z. Kohonen neural network and factor analysis based approach to geochemical data pattern recognition. *J. Geochem. Explor.* **2009**, *103*, 6–16. [CrossRef]
18. Žibret, G.; Šajn, R. Hunting for geochemical associations of elements: Factor analysis and self-organising maps. *Math. Geosci.* **2010**, *42*, 681–703. [CrossRef]
19. Ultsch, A. Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. In *Kohonen Maps*; Elsevier: New York, NY, USA, 1999; pp. 33–45.
20. Ultsch, A.; Mörchen, F. *ESOM-Maps: Tools for Clustering, Visualization, and Classification with Emergent SOM*; Technical Report No. 46; Department of Mathematics and Computer Science, University of Marburg: Marburg, Germany, 2005; pp. 1–7.
21. Ultsch, A. Emergence in Self Organizing Feature Maps. In Proceedings of the Workshop on Self-Organizing Maps (WSOM '07), Bielefeld, Germany, 3–6 September 2007; pp. 1–7.
22. Blondes, M.S.; Gans, K.D.; Engle, M.A.; Kharaka, Y.K.; Reidy, M.E.; Saraswathula, V.; Thordsen, J.J.; Rowan, E.L.; Morrissey, E.A. *U.S. Geological Survey National Produced Waters Geochemical Database*; Version 2.3; U.S. Geological Survey: Reston, VA, USA, 2018.
23. Varmuza, K.; Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics*; CRC Press: Boca Raton, FL, USA, 2009; ISBN 9781420059472.
24. Aitchison, J. *The Statistical Analysis of Compositional Data*; Chapman & Hall: London, UK, 1986; 416p.
25. Engle, M.A.; Rowan, E.L. Interpretation of Na-Cl-Br systematics in sedimentary basin brines: Comparison of concentration, element ratio, and isometric log-ratio approaches. *Math. Geosci.* **2013**, *45*, 87–101. [CrossRef]
26. Engle, M.A.; Rowan, E.L. Geochemical evolution of produced waters from hydraulic fracturing of the Marcellus Shale, northern Appalachian Basin: A multivariate compositional data analysis approach. *Int. J. Coal Geol.* **2014**, *126*, 45–56. [CrossRef]
27. Engle, M.A.; Reyes, F.R.; Varonka, M.S.; Orem, W.H.; Ma, L.; Ianno, A.J.; Schell, T.M.; Xu, P.; Carroll, K.C. Geochemistry of formation waters from the Wolfcamp and “Cline” Shales: Insights into brine origin, reservoir connectivity, and fluid flow in the Permian Basin, USA. *Chem. Geol.* **2016**, *425*, 76–92. [CrossRef]
28. Theodoridis, S.; Koutroumbas, K. *Pattern Recognition*; Elsevier: New York, NY, USA, 2009; 961p.
29. Migdisov, A.; Williams-Jones, A.E.; Brugger, J.; Caporuscio, F.A. Hydrothermal transport, deposition, and fractionation of the REE: Experimental data and thermodynamic calculations. *Chem. Geol.* **2016**, *439*, 13–42. [CrossRef]
30. Martín-Fernández, J.A.; Barceló-Vidal, C.; Pawlowsky-Glahn, V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math. Geol.* **2003**, *35*, 253–278. [CrossRef]
31. Palarea-Albaladejo, J.; Martín-Fernández, J.A. ZCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemomet. Intell. Lab. Syst.* **2015**, *143*, 85–96. [CrossRef]
32. Kharaka, Y.K.; Mariner, R.H. Chemical geothermometers and their application to formation waters from sedimentary basins. In *Thermal History of Sedimentary Basins*; Naeser, N.D., McCulloh, T.H., Eds.; Springer: New York, NY, USA, 1989; pp. 99–117.
33. Martín-Fernández, J.A.; Engle, M.A.; Ruppert, L.F.; Olea, R.A. Advances in Self-Organizing Maps for Their Application to Compositional Data. *Stoch. Environ. Res. Risk Assess.* **2019**, *33*, 817–826. [CrossRef]
34. Engle, M.A.; Blondes, M.S. Linking Compositional Data Analysis with Thermodynamic Geochemical Modeling: Oilfield Brines from the Permian Basin, USA. *J. Geochem. Explor.* **2014**, *141*, 61–70. [CrossRef]
35. Egozcue, J.; Pawlowsky-Glahn, V.; Mateu-Figueras, G.; Barceló-Vidal, C. Isometric logratio transformations for compositional data analysis. *Math. Geol.* **2003**, *35*, 279–300. [CrossRef]
36. Hron, K.; Templ, M.; Filzmoser, P. Imputation of missing values for compositional data using classical and robust methods. *Comput. Stat. Data Anal.* **2010**, *54*, 3095–3107. [CrossRef]
37. Reimann, C.; Filzmoser, P.; Garrett, R.; Dutter, R. *Statistical Data Analysis Explained: Applied Environmental Statistics with R*; Wiley: Hoboken, NJ, USA, 2008; 384p.
38. Filzmoser, P.; Hron, K.; Reimann, C. Interpretation of multivariate outliers for compositional data. *Comput. Geosci.* **2012**, *39*, 77–85. [CrossRef]
39. Ultsch, A.; Hermann, L. *Automatic Clustering with U*C*; Technical Report; Department of Mathematics and Computer Science, Philipps-University of Marburg: Marburg, Germany, 2006; pp. 1–22.
40. Lerch, F.; Thrun, M.; Pape, F.; Paebst, R.; Ultsch, A. Umatrix: Visualization of Structures in High-Dimensional Data; R Package Version 3.1; Comprehensive R Archive Network. 2017. Available online: <https://cran.r-project.org> (accessed on 19 June 2022).
41. Ultsch, A. *U*-Matrix: A Tool to Visualize Clusters in High Dimensional Data*; Technical Report No. 36; Department of Mathematics and Computer Science, University of Marburg: Marburg, Germany, 2003; p. 10.
42. Kharaka, Y.K.; Hanor, J.S. 7.14 Deep Fluids in Sedimentary Basins; In *Surface and Groundwater, Weathering and Soils*, 7th ed.; Elsevier Ltd.: New York, NY, USA, 2014; pp. 471–515.

43. Lewis, A.J.; Palmer, M.R.; Sturchio, N.C.; Kemp, A.J. The rare earth element geochemistry of acid-sulphate and acid-sulphate-chloride geothermal systems from Yellowstone National Park, Wyoming, USA. *Geochim. Cosmochim. Acta* **1997**, *61*, 695–706. [[CrossRef](#)]
44. Fournier, R.O.; White, D.E.; Truesdell, A.H. Geochemical indicators of subsurface temperature-Part 1, basic assumptions. *J. Res. U.S. Geol. Surv.* **1974**, *2*, 259–262.
45. Alibo, D.S.; Nozaki, Y. Rare earth elements in seawater: Particle association, shale-normalization, and Ce oxidation. *Geochim. Cosmochim. Acta* **1999**, *63*, 363–372. [[CrossRef](#)]
46. Hanor, J.S. Origin of Saline Fluids in Sedimentary Basins. In *Geofluids: Origin, Migration and Evolution of Fluids in Sedimentary Basins*; Parnell, J., Ed.; Geological Society: London, UK, 1994; Volume 78, pp. 151–174.
47. Takaya, Y.; Yasukawa, K.; Kawasaki, T.; Fujinaga, K.; Ohta, J.; Usui, Y.; Nakamura, K.; Kimura, J.-I.; Chang, Q.; Hamada, M.; et al. The tremendous potential of deep-sea mud as a source of rare-earth elements. *Sci. Rep.* **2018**, *8*, 5763. [[CrossRef](#)]
48. U.S. Geological Survey National Minerals Information Center. *Mineral Commodity Summaries 2020*; U.S. Geological Survey: Reston, VA, USA, 2020; 204p.
49. Ray, S. National Evaluation for Development and Exploration Potential of Mineral Commodities in Produced Waters. Master's Thesis, University of Texas at El Paso, El Paso, TX, USA, 2016.