Original Paper

Machine Learning Can Assign Geologic Basin to Produced Water Samples Using Major Ion Geochemistry

Jenna L. Shelton ^{1,6} Aaron M. Jubb ¹ Samuel W. Saxe ² Emil D. Attanasi ¹ Alexei V. Milkov ³ Mark Engle ⁴ Philip A. Freeman ¹ Christopher A. Shaffer,⁵ and Madalyn S. Blondes ¹

Received 26 April 2021; accepted 15 September 2021
Published online: 30 September 2021

Understanding the geochemistry of waters produced during petroleum extraction is essential to informing the best treatment and reuse options, which can potentially be optimized for a given geologic basin. Here, we used the US Geological Survey's National Produced Waters Geochemical Database (PWGD) to determine if major ion chemistry could be used to classify accurately a produced water sample to a given geologic basin based on similarities to a given training dataset. Two datasets were derived from the PWGD: one with seven features but more samples (PWGD7), and another with nine features but fewer samples (PWGD9). The seven-feature dataset, prior to randomly generating a training and testing (i.e., validation) dataset, had 58,541 samples, 20 basins, and was classified based on total dissolved solids (TDS), bicarbonate (HCO_3), Ca, Na, Cl, Mg, and sulfate (SO_4). The nine-feature dataset, prior to randomly splitting into a training and testing (i.e., validation) dataset, contained 33,271 samples, 19 basins, and was classified based on TDS, HCO_3 , Ca, Na, Cl, Mg, SO_4 , pH, and specific gravity. Three supervised machine learning algorithms—Random Forest, k-Nearest Neighbors, and Naïve Bayes—were used to develop multi-class classification models to predict a basin of origin for produced waters using major ion chemistry. After training, the models were tested on three different datasets: Validation7, Validation9, and one based on data absent from the PWGD. Prediction accuracies across the models ranged from 23.5 to 73.5% when tested on the two PWGD-based datasets. A model using the Random Forest algorithm predicted most accurately compared to all other models tested. The models generally predicted basin of origin more accurately on the PWGD7-based dataset than on the PWGD9-based dataset. An additional dataset, which contained data not in the PWGD, was used to test the most accurate model; results suggest that some basins may lack geochemical diversity or may not be well described, while others may be geochemically diverse or are well described. A compelling result of this work is that a produced water basin of origin can be determined using major ions alone and, therefore, deep basinal fluid compositions may not be as variable within a given basin as previously thought. Applications include predicting the geochemistry of produced fluid prior to drilling at different intervals and assigning historical produced water data to a producing basin.

KEY WORDS: Machine learning, Produced water, Chemistry, Random Forest, Basinal brines.

¹U.S. Geological Survey, Reston, VA 20192, USA.

²U.S. Geological Survey, Lakewood, CO 80225, USA.

³Colorado School of Mines, Golden, CO 80401, USA.

⁴University of Texas at El Paso, El Paso, TX 79968, USA.

⁵U.S. Fish and Wildlife Service, Alamosa, CO 81101, USA.

⁶To whom correspondence should be addressed; e-mail: jishelton@usgs.gov

INTRODUCTION

The production of hydrocarbons from the subsurface can result in the co-production of large volumes of water. This water is generally referred as

produced water and is typically geochemically representative of the in-place formation water with some potential contribution from fluid injected for well completion and/or stimulation. These produced waters can either be disposed of through injection into the subsurface or treated in some manner and reused for anthropogenic needs. Therefore, knowing the chemical composition of the produced water is critical when planning how to treat or dispose of these produced large waters from a specific oil or gas field. Produced water is geochemically diverse when comparing across producing basins of the USA (e.g., Fakhru'l-Razi et al., 2009) or age of the well (e.g., Cluff et al., 2014; Rowan et al., 2015). For example, the first few days of production from a hydraulically fractured well are geochemically distinct from water produced 90 days later (Engle & Rowan, 2014). Understanding the chemical composition of produced waters across the continental USA to optimize treatment or reuse methods to specific geographic regions could potentially save stakeholders time, money, and resources. Although produced water geochemistry summary statistics can be determined for various producing basins using existing data, it is unclear whether past sampling efforts have fully captured the geochemical diversity of a given basin (i.e., the statistical shortcomings of previous sampling strategies was not evaluated).

Many producing formations or basins across the continental USA produce waters that have been previously well characterized, such as the Appalachian Basin (Blondes et al., 2020; McDevitt et al., 2020; Osborn et al., 2012; Tasker et al., 2020), the Permian Basin (Engle et al., 2016; Nicot et al., 2020), the Denver Basin (Rosenblum et al., 2017), the Williston Basin (Iampen & Rostron, 2000; McMahon et al., 2020; Varonka et al., 2020), and the Gulf Coast (Engle et al., 2020; McIntosh et al., 2010; Nicot et al., 2018). Although these regions produce waters that are geochemically well understood, few studies have analyzed trends in produced water chemistry across multiple basins or the continental USA (Birkle et al., 2019; Kharaka et al., 2019; Scanlon et al., 2020). Understanding basinal, regional, or national trends in produced water geochemistry could help ascertain trends in salinity, naturally occurring radioactive materials, barite, and other parameters essential to understanding best practices for wastewater treatment and disposal. For example, knowing that two geographically similar basins also produce water with a similar geochem-

istry could aid in developing regional water treatment plans (e.g., Scanlon et al., 2020).

Machine learning algorithms have been applied to many earth science-related problems, such as classifying lithology type using geophysical data (Bressan et al., 2020), understanding soil properties (Zhao & Wang, 2020), landslide assessments (Marjanovic et al., 2011), and in oil and gas production applications (Attanasi et al., 2020; Gaurav, 2017; Mohaghegh, 2020; Snodgrass & Milkov, 2020). Additionally, machine learning algorithms have been specifically developed to predict the class label of a given sample based on known features; this is known as multi-class classification (Farid et al., 2014). However, multi-class classification machine learning applications related to produced waters are sparse (e.g., Engle & Brunner, 2019). There are various potential useful applications combining produced water geochemical data and machine learning, including determining source of salinity for basinal fluids, assisting with drilling in frontier areas, determining if biogenic methane is occurring in a given reservoir based on produced water chemistry, fingerprinting produced water to formations in spills or multi-zone completions, and, as investigated in this manuscript, using major ion chemistry to determine basin of origin for a produced water sample. Applying machine learning to determine the comparability of water chemistry across different producing basins is beneficial because basin chemistry distinctness could be inferred based on the accuracy of a given algorithm; if a model consistently and accurately classifies a sample correctly to a given basin, it can be assumed that the geochemistry of that basin is distinct (compared to other basins analyzed) and may require unique water treatment methods compared to surrounding basins.

In this study, we used the US Geological Survey's National Produced Water Geochemical Database (USGS PWGD; Blondes et al., 2018) coupled with multiple supervised machine learning algorithms to determine if basin of origin could be predicted for a given water sample using basic geochemical parameters. The number of basins represented in the USGS PWGD database was reduced to those having at least 1000 samples and had a pre-specified suite of geochemical measurements. The resulting dataset was used to train and validate supervised multi-class classification machine learning algorithms. We discuss the applicability of these results to future applications of machine learning

approaches related to complex processes associated with energy development.

METHODS

Data Preparation

The US Geological Survey's National Produced Waters Geochemical Database version 2.3 (USGS PWGD; Blondes et al., 2018) was downloaded and used as the initial database. The original database contains 114,943 distinct samples across 88 producing basins. The database has a total of 190 features, 38 of which are geochemical parameters of the sampled fluids. We encourage readers to view and download the USGS PWGD and the associated data dictionary, which will provide a better understanding of the data provided in the database. Detailed methods for database culling and the data used for all the analyses in this paper are available through Shelton et al. (2021); however, methods will be discussed concisely below.

The raw USGS PWGD was initially culled by sample type. Because this study only includes waters associated with hydrocarbon production that would also be representative of formation water within the basin, samples with a WELLTYPE attribute of geothermal ($n = 689$), injection ($n = 13$), or undefined ($n = 867$) were removed from the dataset. Additionally, any sample with data in the attribute TIMESERIES was also removed, as these samples are likely not representative of actual formation water (e.g., Cluff et al., 2014; Rosenblum et al., 2017). The resulting dataset containing 113,173 samples can be found in Shelton et al. (2021) as PWGDStep1.csv.

Next, all features that were unlikely to be used for any QA/QC purposes or by a model itself were removed (e.g., OPERATOR, LAB, ELEVATION). Empty cells for each attribute across all remaining samples were then calculated. If data were missing from over 97% of samples of a given attribute, that attribute was eliminated from the dataset (107 features were removed); this step was done to ensure that the resulting dataset had as much data as possible for every sample. The resulting dataset containing 56 features are publicly available through Shelton et al. (2021) as PWGDStep3.csv.

The remaining dataset was then checked to ensure that all assigned location data were self-consistent. For example, if a sample had an assigned

API number, that number was checked against the features LAT, LONG, STATE, or other location information. This step was performed to check that any samples were not incorrectly assigned their BASIN attribute. Next, any assigned BASIN, which is the American Association of Petroleum Geology (AAPG) Geologic Province (Meyer, ; Meyer et al., 1991) (see USGS PWGD documentation), was removed if it had fewer than 1000 samples to investigate the most prominent and data-rich basins. This step was performed to ensure that the training and validation sets would have a robust number of samples for each BASIN included after additional culling (e.g., Chu et al., 2012; Figueroa et al., 2012). The resulting dataset includes 20 different BASIN assignments across 102,571 samples (Fig. 1) and is available as part of Shelton et al. (2021) as PWGDStep4.csv.

As missing data or NAs (i.e., data were not available) in a dataset must be either eliminated or estimated in order to perform multi-class classification modeling with machine learning (e.g., Batista & Monard, 2003; Bertsimas et al., 2017; Ma et al., 2020), additional culling was required to remove all missing data from the PWGDStep4 dataset. Using existing data to estimate these missing data would introduce basinal bias (i.e., using data from one basin to estimate data from that basin would defeat the purpose of this study). Therefore, we chose to eliminate any sample with missing data. First, any remaining attribute with missing data for greater than 75% of the remaining sample set was removed. The resulting dataset contains 102,109 samples across 20 BASINs [PWGDStep5.csv of Shelton et al. (2021)]. This step left nine different features: specific gravity (SG), pH, total dissolved solids (TDS), bicarbonate (HCO_3), calcium (Ca), chloride (Cl), magnesium (Mg), sodium (Na), and sulfate (SO_4). All concentrations are in mg/L. Next, any sample that did not have data for all nine of these given features was removed from the sample set. This step resulted in a complete dataset containing 33,271 samples with nine geochemical features across 19 different BASINs (after this culling step, the Raton BASIN had 0 samples and was therefore eliminated). This dataset will be referred to as PWGD9 throughout this text and can be found in Shelton et al. (2021) as PWGDStep6_9Features.csv.

However, a second final dataset was also curated to increase the sample number included in the analysis. The features SG and pH were eliminated, and any samples without data for all remaining se-

Figure 1. Map of the American Association of Petroleum Geologists Provinces (Meyer, ; Meyer et al., 1991) included in this study (black outlines). Samples used for the training and testing of the three algorithms are depicted as red circles. Only those samples with geographic information as provided in the US Geological Survey's National Produced Waters Geochemical Database (Blondes et al., 2018) are displayed (90% of the samples).

ven features (TDS, HCO_3 , Ca, Cl, Mg, Na, and SO_4) were removed from the dataset. This culling step produced a dataset with 58,541 samples across seven geochemical features and 20 different BASINs. This dataset will be referred to as PWGD7 throughout this text and is publicly available via Shelton et al. (2021) as PWGDStep6_7Features.csv. The number of samples per basin for each dataset and other accessory data can also be found in Shelton et al. (2021).

Outlier Removal

Outliers were removed (e.g., John, 1995) in R (R Core Team, 2020) using R version 4.0.3 (Bunny-

Wunnies Freak Out). Outliers were examined for each BASIN so that an equal percentage of samples would be removed from each BASIN. For example, if outliers for the attribute TDS were removed across all BASINs simultaneously instead for each individual BASIN, there would be BASINs with no samples assigned as outliers. Outliers were assigned at a $< 99\%$ confidence interval for each individual attribute for each individual BASIN. Any outlier identified was removed from each respective dataset. R code for outlier removal can be found in Shelton et al. (2021) as “Outlier Removal Analysis PWGD.R.” It is important to note that no culling was done based on the quality of the data. The USGS PWGD offers some information on data quality, such as charge balance data. However, we

did not consider these quality parameters to limit additional bias and to present a more realistic dataset.

Additional Pre-processing

After outlier removal, the datasets PWGD7 and PWGD9 were each randomly reordered to ensure no bias was introduced by the order of samples in each respective database. The two randomly reordered datasets were then each split into a model training set and a model validation set. A 90/10 split was used for PWGD7 to produce the training and validation sets, respectively; an 80/20 split was also tested for PWGD7, but the 90/10 split produced a more accurate model. The training and validation datasets generated using PWGD7 will be referred to as TrainPWGD7 and Validation7 within, respectively. An 80/20 split was used to divide PWGD9 into training and validation sets (a 90/10 split was also tested for increased accuracy). The training and validation datasets generated using PWGD9 will be referred to as TrainPWGD9 and Validation9 within, respectively. The training and validation sets can be found in Shelton et al. (2021).

Model Selection

The R caret package (Kuhn, 2008) version 6.0–86 (and all dependencies) was implemented to build multi-class classification models in R. The caret package is advantageous to use due to the ability to create custom tuning grids and optimize the tuning of model parameters such as cross-validation and hyperparameters (see package documentation, Kuhn, 2008). Three supervised machine learning algorithms were optimized for the TrainPWGD7 and TrainPWGD9 datasets: k-Nearest Neighbors, Naïve Bayes, and Random Forest. These algorithms were selected as there are a limited number of multi-class classification algorithms, and these three are capable of classification beyond binary. These three algorithms have been used broadly in the literature and for many earth science applications with examples provided below.

Briefly, the k-Nearest Neighbors algorithm uses labeled input data (in this case, labeled by basin) and produces a distance matrix to calculate the distance between new data and the training data; any new data are classified based on the class of the nearest

known data points (e.g., Laaksonen & Oja, 1996). The Naïve Bayes classification method is a simple probabilistic classifier which uses Bayes' Theorem and assumes each of the dataset features is independent and contributes equally to the model outcome (e.g., Zhang et al., 2009). The Naïve Bayes method classifies new samples based on the maximum a posteriori decision rule. The Random Forest algorithm uses a collection of relatively uncorrelated decision trees in which each predicts the classification of an unknown sample; the class that most of the decision trees produce for an unknown sample is ultimately the final prediction (e.g., Breiman, 2001; Liaw & Wiener, 2002). For more information about the Random Forest algorithm used for classification tasks, readers are referred to Strobl et al. (2007), Wright et al. (2020), and references within. Because there are a limited number of algorithms that can be used for multi-class classification, it was reasonable to test these three to determine if any performed better than another in order to find the most accurate model possible.

The train() function in the caret package was used to implement these models, calling method = "ranger" for Random Forest, "naive_bayes" for Naïve Bayes, and "knn" for k-Nearest Neighbors. The train function can use a custom tuning grid to optimize repeated k-fold cross-validation and other model hyperparameters (e.g., impurity scores, mtry) by examining the results from every possible combination of criteria and outputting the most accurate combination (see Table 1 for PWGD7 results, see Table 2 for PWGD9 results). For the Random Forest algorithms, the splitting criterions "gini" and "extratrees" were tested. The splitting criterion "gini" uses the probabilities of the number of classes to assign an impurity score, between 0 and 1, where the lower the score, the better the split within a tree (see Breiman et al., 1984; Breiman, 2001). The splitting criterion "extratrees" implements the extremely randomized trees ensemble method (see Geurts et al., 2006) where both cutting point choice and attribute choice are extremely randomized when splitting a tree node. Custom tuning grids were developed to find optimum parameters for each model, such as the mtry value, resulting in the greatest accuracy (see Kuhn, 2008). Model characteristics and optimum hyperparameters can be found in Tables 1 and 2, and all associated code can be found in Shelton et al. (2021).

Overall model performance was determined using the results of the confusion matrix produced

Table 1. Results from model optimization using the TrainPWGD7 dataset and model accuracy results using the Validation7 dataset

Model without outliers	Random Forest “Gini” PWGD7	Random Forest “extratrees” PWGD7	k-Nearest Neighbors PWGD7	Naïve Bayes PWGD7
Model type	Random Forest	Random Forest	k-Nearest Neighbors	Naïve Bayes
Training dataset	TrainPWGD7	TrainPWGD7	TrainPWGD7	TrainPWGD7
Best-performing tuning parameters	Split Rule = “gini”; mtry = 2	Split Rule = “extratrees”; mtry = 6	k = 13	laplace = 0; usekernel = TRUE; adjust = 0.75
Cross-validation	10 fold	10 fold	10 fold	10 fold
Accuracy (%)	66.55	67.12	56.22	42.19
95% confidence interval	(0.6533, 0.6794)	(0.6581, 0.6841)	(0.5485, 0.5759)	(0.4083, 0.4356)
No information rate	0.1759	0.1751	0.1729	0.2079
<i>P</i> value [Accuracy > No Information Rate]	< 2.20E−16	< 2.20E−16	< 2.20E−16	< 2.20E−16
Kappa	0.6323	0.6375	0.5174	0.3647

Table 2. Results from model optimization using the TrainPWGD9 dataset and model accuracy results using the Validation9 dataset

Model without outliers	Random Forest “Gini” PWGD9	Random Forest “extratrees” PWGD9	k-Nearest Neighbors PWGD9	Naïve Bayes PWGD9
Model type	Random Forest	Random Forest	k-Nearest Neighbors	Naïve Bayes
Training dataset	TrainPWGD9	TrainPWGD9	TrainPWGD9	TrainPWGD9
Best-performing tuning parameters	Split Rule = “gini”; mtry = 3	Split Rule = “extratrees”; mtry = 5	k = 5	laplace = 0; usekernel = TRUE; adjust = 1.5
Cross-validation	10 fold	10 fold	10 fold	10 fold
Accuracy (%)	72.97	73.47	23.47	42.71
95% confidence interval	(0.7180, 0.7411)	(0.7231, 0.7460)	(0.2238, 0.2458)	(0.4143, 0.4400)
No information rate	0.2781	0.2816	0.8135	0.2092
<i>P</i> value [Accuracy > No Information Rate]	< 2.20E−16	< 2.20E−16	1	< 2.20E−16
Kappa	0.6840	0.6892	0.0527	0.3557

for each algorithm along with “one-versus-all” results for each algorithm to calculate the accuracy and Kappa statistics. The overall accuracy and Kappa statistics for each algorithm are calculated using the default performance function “postRresample.” Variable importance scores were calculated using model-based approaches. The Naïve Bayes and k-Nearest Neighbor algorithms used the absolute value of the t-statistic to determine the importance score for each feature (Kuhn, 2008). The Random Forest algorithm records the prediction accuracy on the out-of-bag portion of the dataset for each decision tree in the Random Forest. After

permuting each predictor variable, the process is repeated, and the difference between these two accuracies is averaged across all decision trees and normalized by standard error (Kuhn, 2008). For more information about the caret package, along with equations to calculate model performance, we refer readers to the R documentation for caret (Kuhn, 2008). Other R packages used in this analysis were tidyverse (Wickham et al., 2019), reshape (Wickham, 2007), dplyr (Wickham et al., 2020), caTools (Tuszynski, 2020), bnlearn (Mihaljevic et al., 2018), klaR (Weihs et al., 2005), and vegan (Oksanen et al., 2020).

Development and Testing of “Unknown” Dataset

In addition to developing training sets with data in the USGS PWGD, we also used recently published data (Blondes et al., 2020; Engle, 2019; Engle et al., 2020) that is not currently included in the USGS PWGD as additional testing data for the models. The data are produced waters from the Eagle Ford ($n = 39$; Engle, 2019; Engle et al., 2020) and the Appalachian Basin, specifically the Utica Shale and Clinton sandstone ($n = 52$; Blondes et al., 2020). Because the data used in the validation sets may have originated from the same sampling campaigns as those data used in the training sets, i.e., both were from the same database, this may introduce bias into the TestPWGD7 and TestPWGD9 results. Therefore, we deemed it essential to test the best-performing model on these “unknown” data sources as data from these sampling campaigns were not used to also train the models. The nine geochemical features used to train the models were used to generate this unknown sample set which included 91 different samples. The unknown training dataset can be found in Shelton et al. (2021).

RESULTS

Initial USGS PWGD Analysis

The culled versions of the USGS PWGD, PWGD9, and PWGD7 were examined for any obvious trends in the geochemical data. Prior to outlier removal, TDS values ranged from 1.91 to 420,970 mg/L across all samples (i.e., PWGD7 dataset), SG ranged from 0 to 1.48 (PWGD9 dataset), pH ranged from 0 to 12.3 (PWGD9 dataset), HCO_3 ranged from 0.201 to 241,592 mg/L (PWGD7 dataset), Ca ranged from 0.01 to 77,630 mg/L (PWGD7 dataset), Na ranged from 0.4 to 166,746 mg/L (PWGD7 dataset), Mg ranged from 0.003 to 46,656 mg/L (PWGD7 dataset), Cl values ranged from 0.25 to 321,643.8 mg/L (PWGD7 dataset), and SO_4 concentrations ranged from 0.01 to 150,000 mg/L (PWGD7 dataset). After outlier removal and prior to splitting datasets into testing and training datasets, TDS values ranged from 246 to 389,722 mg/L across all samples (i.e., all 20 basins), SG ranged from 1 to 1.48 (across 19 basins), pH ranged from 1.133 to 12.21 (across 19 basins), HCO_3 ranged from 3 to 14,886 mg/L, Ca ranged from 3 to 69,825 mg/L, Na ranged from 9 to 139,886.4 mg/L,

Mg ranged from 0.71 to 11,652.55 mg/L, Cl values ranged from 1.3 to 240,205.5 mg/L, and SO_4 concentrations ranged from 0.12 to 12,287 mg/L. When the outlier-removed datasets are presented as box plots (PWGD9 for SG and pH and PWGD7 for remaining 7 features), differences and similarities for each of the nine features can be easily visualized across all 20 basins (Fig. 2).

To further understand potential challenges with both datasets, a non-metric multidimensional scaling (NMDS) plot was used to visualize a dissimilarity matrix of the scaled geochemistry data (Fig. 3). The PWGD9 was used for pH and SG values, while the PWGD7 dataset was used for all other features. Samples from the Anadarko and Sedgwick basins, for example, may be difficult to distinguish using machine learning due to the similarity of their initial geochemical data. Similarly, distinction may be difficult between samples originating from the Central Kansas Uplift and Illinois basins, and between the Denver, Powder River, and Wind River basins (Fig. 3). Alternatively, samples from some basins, like the Williston, may be easy to distinguish from other basins. It is not surprising that the Michigan, Appalachian, and Illinois basins plot close to each other (along one axis of the NMDS) given formation waters have been shown to have similar origins and depositional histories (e.g., McIntosh & Walter, 2005; McIntosh et al., 2002, 2011); similarly, this would also explain the similarity of the Anadarko and Permian basins due to their geographic proximity and similar producing formations (Bein & Dutton, 1993; Engle et al., 2016; Sorenson, 2005).

Optimum Model Parameters

Each of the three tested algorithms allowed for the automatic tuning of specific hyperparameters so that the most accurate model could be generated. For the Random Forest algorithm, we found that the split-variable randomization method “extra-trees” produced a marginally more accurate model for both datasets than the splitting rule “gini” (Tables 1 and 2). The number of predictor variables that are sampled at each split in the Random Forest, or the *mtry* value, varied across datasets and split rules. However, the optimum numbers of parameters used at each split in the most accurate Random Forest model were equal to 5 (Tables 1 and 2). Both models performed best when cross-validated ten times.

Figure 2. Box plots by geochemical attribute for each basin included in this study after outlier removal. Basins are listed on the x axis with attribute values on the y-axis. The whiskers extend to the maximum and minimum values for each respective basin after outlier removal. Concentrations (in mg/L) are provided in log scale for all features except pH and specific gravity, both of which are unitless. Note that the Raton basin is absent in the SG and pH plots, as the Raton basin had zero samples with all nine features. Data for SG and pH are based on PWGD7; all other features are based on PWGD9. SG, Specific Gravity; TDS, Total Dissolved Solids; App, Appalachian; CKU, Central Kansas Uplift; CP, Chautauqua Platform; E. Texas, East Texas; GR, Green River; GC, Gulf Coast; PR, Powder River; SO, Southern Oklahoma; WR, Wind River. Basins are differentiated along the y-axis and also by the colors present.

Figure 3. A non-metric multidimensional scaling (NMDS) plot based on a dissimilarity matrix of the geochemical parameters, averaged by producing basin, based on the PWGD7 dataset. Axes are unitless, and colors are meant to distinguish between closely spaced basins. MI, Michigan; App, Appalachian; CP, Chautauqua Platform; SO, Southern Oklahoma; CKU, Central Kansas Uplift; ET, East Texas; GC, Gulf Coast; BH, Big Horn; SJ, San Juan; PR, Powder River; WR, Wind River; GR, Green River.

The Naïve Bayes algorithms required tuning of the Laplace smoothing and parameter adjustment (see documentation for caret package in Kuhn, 2008). The most accurate models generated for both training datasets (TrainPWGD7 and TrainPWGD9) both were optimized with a Laplace smoothing value equal to 0 (Tables 1 and 2), and the optimum adjust value was equal to 0.75 for the TrainPWGD7 dataset and equal to 1.5 for the TrainPWGD9 dataset. The k-Nearest Neighbors algorithm produced the most accurate model when “k” was set to 13 for the TrainPWGD7 dataset and 5 for the TrainPWGD9 dataset. All four models performed best when cross-validated ten times.

Optimized Models and Prediction on Validation Datasets

The optimized Random Forest algorithm using split rule “extratrees” predicting on the 20% split from PWGD9, Validation9, produced the highest accuracy at 73.5% (Table 2). For the validation set generated with a 10% split of PWGD7, Validation7, the most accurate model produced was also the Random Forest with split rule “extratrees.” The

Random Forest algorithm performed better on average than the k-Nearest Neighbors and Naïve Bayes algorithms when using both Validation7 and Validation9 as inputs (Tables 1 and 2). The Naïve Bayes algorithm predicted at almost identical accuracy for both training sets, 42.2% accuracy for Validation7 and 42.7% accuracy for Validation9. However, the models produced using the k-Nearest Neighbors algorithm had the greatest disparity in prediction accuracy when comparing the two different validation datasets. The k-Nearest Neighbors model predicted on Validation7 with 56.2% accuracy but predicted on Validation9 at only 23.5% accuracy (Tables 1 and 2). The confusion matrices for the best-performing model for each Validation set can be found in SI Table S2 (Validation7) and SI Table S3 (Validation9).

Input variable importance was similar across the Random Forest algorithms tested (Table 3). Variable importance ranged from 100 to 0%, with SO_4 and Mg producing importance values of 100%, and HCO_3 and pH producing importance values of 0.00%. Besides SO_4 and Mg, other input variables of importance were Cl, Na, and Ca (importance values greater than 50%) for the Validation7 dataset, and Cl, Na, Ca, and TDS (importance values greater than 50%) for the Validation9 dataset. The k-Nearest Neighbors and Naïve Bayes algorithms produced models that predicted the validation sets at much lower overall accuracies; however, the balanced accuracies for each individual basin were sometimes comparable across the three different algorithms (Fig. 4).

Individual balanced accuracies (SI Table S1) across the 20 different basins ranged from 48.0% (Powder River basin; k-Nearest Neighbors on Validation9) to 100% (Raton basin; Random Forest “extratrees” on Validation7). The Raton basin had the highest average balanced accuracy across all models (trained on the PWGD7 dataset) at 93.4%, while the Powder River had the lowest balanced accuracy across all models at 65.4%. This result indicates that Raton basin water samples in the validation dataset were normally correctly assigned to the Raton basin, and that Powder River samples were the least likely to be correctly assigned to the Powder River basin. Additionally, based on the previous NMDS analysis, it was expected that the Raton, Williston, and Michigan basins were most distinct and may therefore produce the highest accuracies—this was observed and is reflected in the balanced accuracy results (Fig. 4).

Table 3. Input variable importance scores for the Random Forest algorithms

Variable importance	Random Forest “Gini” PWGD7 (%)	Random Forest “extratrees” PWGD7 (%)	Random Forest “Gini” PWGD9 (%)	Random Forest “extratrees” PWGD9 (%)
TDS	36.92	43.99	51.10	51.96
HCO ₃	0.00	0.00	24.34	21.72
Ca	58.49	62.61	55.02	49.80
Na	69.20	79.08	54.96	55.46
Cl	79.61	82.94	68.67	60.75
Mg	100.00	99.45	92.55	82.37
SO ₄	87.26	100.00	100.00	100.00
pH	NA	NA	0.00	0.00
SG	NA	NA	29.13	33.33

Prediction on Unknown Data

The Random Forest “extratrees” model used on the unknown data (SI Table S4) predicted at 97% accuracy for the new Eagle Ford data, but only correctly classified 1.9% of the Appalachian produced water samples (see Supporting Information). Fifty-one of the 52 produced water samples that originated from the Appalachian basin (Blondes et al., 2020) were classified as Arkla waters. Only one sample from this 52 sample dataset was predicted to be from the Appalachian basin. It is important to note that the Arkla and Appalachian basins plotted close to each other using an NMDS analysis (Fig. 3); therefore, it is not surprising that these Appalachian samples were incorrectly classified as Arkla given the geochemical characteristics of those two basins’ waters are similar. All but one of the produced water samples from the Eagle Ford (Engle, 2019; Engle et al., 2020) were correctly classified as producing from the Gulf Coast, while the outlier was assigned to the Permian.

DISCUSSION

Data Quality and Model Results

As any machine learning model is only as accurate as the data used to train it, it is important that the USGS PWGD contained high quality data. Unfortunately, it is impossible to find references for every sample in the USGS PWGD and to check authors’ QA/QC procedures. Although any published geochemical data should have undergone a strenuous QA/QC procedure prior to publication, we still removed outliers in this dataset to improve the dataset quality. We acknowledge this step may have eliminated some

natural, true variability present within produced waters; however, all three algorithms predicted at higher accuracies on the datasets with the outliers removed as opposed to the datasets prior to outlier removal (see Shelton et al., 2021 for input files). It has also been shown that some data in the USGS PWGD are likely erroneous and/or contain samples from waterflooded oil fields, both of which could add error to these results (Engle & Blondes, 2014).

Another variable to be considered is sample distribution. We could not control whether the samples in the USGS PWGD represented a breadth of producing depths, lithologies, formations, etc. For example, most of the data in the USGS PWGD are from conventional oil and gas wells (Blondes et al., 2018). Therefore, bias may exist if every sample in the USGS PWGD from that basin originated from one specific environment in each basin. This potential bias would likely skew the results and suggest less geochemical diversity for a given basin than is present. However, by initially excluding any producing basin with fewer than 1000 samples (see Methods), this bias can be minimized.

It is unclear whether the basins that had high balanced accuracies were geochemically distinct, or if the samples included in the USGS PWGD from those basins were not geographically or geologically diverse. Additionally, some basins may have a larger variance in the data due to characteristics of the reservoirs or complexity of fluid history. For example, the Raton Basin, predicting at 93.7% accuracy across all tested models, is not a striking outlier when viewing its position on the NMDS plot (Fig. 3). The Raton Basin does not cluster with many other basins, such as the Anadarko and Sedgwick, potentially suggesting geochemical distinction, but it is not a noticeable outlier like the Michigan Basin. Conversely, the Powder River,

Figure 4. Balanced accuracies for each model generated and tested across each predicted basin. Note that PWGD9 model prediction results could not be calculated for the Raton basin, as that basin was not included in the PWGD9 dataset.

which had the lowest average balanced accuracy, does appear to be geochemically similar to two other basin, the Denver and Wind River, both of which also generally produced lower average balanced accuracies than observed for other basins (Fig. 4).

To improve the quality and applicability of the model presented within, not only would additional diverse produced water samples from the included basins need to be considered but adding additional data for the ca. 60 basins absent from this modeling

effort should be prioritized. Additionally, as suggested by the variable importance scores, collecting SO_4 and Mg data should also be prioritized, as those data appear to be the most important when identifying basin of origin. Other data, such as pH, SG, HCO_3 , and TDS, appear to be less important due to their lower variable importance scores. Increasing the representation of specific basins in the USGS PWGD, capturing a diverse set of samples from a given basin, and prioritizing the inclusion and

examination of SO_4 and Mg data in the USGS PWGD would increase all the robustness of the model, and potentially increase applicability.

The importance of the input variables should not be overlooked. Previous research suggests that the major ion composition of saline basinal fluids is predominantly a function of depositional origin or the dissolution of subsurface halites, with other contributions to water chemistry originating from processes such as dissolution and precipitation of other minerals, water rock reaction with different lithology-dependent phases, interaction with organic matter, and compaction and orogenic history that can cause fluids of different origins to mix, (e.g., Carpenter, 1978; Hanor, 2001; Kharaka & Hanor, 2003; Lowenstein et al., 2003; Hanor & McIntosh, 2006). Furthermore, these processes cause intra-basin variability (i.e., waters produced from different formations within the same basin can be geochemically different) to generally be high, which is also highlighted by the data in the USGS PWGD (Fig. 2). The results of this study suggest that it is indeed possible to distinguish fluids from some basins even without accounting for these water chemistry-controlling processes during sample selection, meaning intra-basinal variability may not be as dramatic as previous suggested, or that basin-specific processes are distinctive enough to discriminate basin of origin. Because we can generally predict and distinguish produced waters from 20 different producing basins—some with much better accuracy than others—given major ion chemistry alone, this result suggests distinguishable and distinct geochemical characteristics when comparing produced fluids originating from different basins, mostly dependent on Mg and SO_4 compositions. As suggested by the references listed above, SO_4 concentrations in a reservoir are controlled by microbial sulfate reduction, H_2S oxidation, and gypsum solubility, while Mg concentrations are controlled by dolomitization/de-dolomitization reactions, presence and original composition of paleo-seawater contribution, and relative Ca abundance.

The data presented in Figure 2 suggest major differences between the Raton and Big Horn basins that are driving differences in SO_4 concentration, while differences must exist between the Michigan and Green River basins driving differences in Mg concentrations. The Raton Basin is a prolific coalbed methane producer, with hydraulically fracturing assisting in the natural gas development (e.g., US EPA, 2016); Włodarczyk (2016) indicated that

associated brines are likely sourced from meteoric recharge and that halite dissolution has only slightly developed the groundwater. Because SO_4 concentrations greater than 1 mM limit biogenic methane generation (e.g., Löffler and Sanford, 2005), it is not surprising that SO_4 concentrations are low in this basin. According to De Bruin (1997), there are eight main oil and gas producing fields in the Big Horn Basin, which produce from both sandstone and carbonate reservoirs. A study by Ulmer-Scholle and Scholle (1994) on the Park City Formation within the Big Horn Basin found extensive evidence for silicification of evaporites, which would influence SO_4 concentrations in formation fluids.

Hydrocarbons produced from the Michigan Basin are generally sourced from the Antrim or the Utica Shale, while basinal brines are thought to be derived from the evaporation of Paleozoic seawater, dissolution of evaporites (which could easily influence Mg concentrations in a developing brine, e.g., Feng et al. (2018)), and general water-rock interactions (Wilson & Long, 1993; Hanor & McIntosh, 2006; Swezey, 2002; McIntosh et al., 2011). The Green River basin produces hydrocarbons from mostly Cretaceous formations (Toner et al., 2018) and was thought to have developed under a playa-lake model, which may drive the Mg distinction observed here, as calcite precipitation during playa-lake formation drives Mg/Ca ratios in the developing brine (Eugster & Surdam, 1973). These previous studies provide some evidence that formation waters from these basins could differ, supporting the results of this study.

Insights from an Unknown Dataset

None of the unknown testing data (Blondes et al., 2020; Engle, 2019; Engle et al., 2020) were included in the training and validation sets generated from the USGS PWGD. As hoped, 38 of the 39 Eagle Ford Shale produced water samples were correctly classified to the Gulf Coast basin. This result indicates that either (1) the model is robust enough to, at a minimum, accurately assign unknown produced water samples accurately to the Gulf Coast basin, meaning the data fed into the training set were diverse enough to capture any variability across the Gulf Coast, (2) the new Eagle Ford produced water data and the data already present in the USGS PWGD are similar enough to be accurately related, or (3) data discrepancies (i.e., uneven sample distribution across basins) have led

to better sample predictability in some basins compared to others. However, the outlier assigned to the Permian was also a geochemical outlier: it had a much higher TDS concentration coupled with a much lower SO_4 concentration compared to the other 38 Eagle Ford samples. This sample would have been culled had outlier removal been performed on this unknown sample set.

The Appalachian basin sample set produced the opposite outcome, where only one sample was correctly classified. This sample was also an outlier; it had a much lower HCO_3 concentration and a higher SO_4 concentration than the remaining 51 samples. Although we do think that it is promising that the majority of the unknown Appalachian produced water samples were grouped and assigned to one basin, these results suggest that either (1) the current training set for the Appalachian basin (based on the USGS PWGD) is lacking data from the Utica Shale, which is somewhat geochemically distinct from the major producing unit in the Appalachian, the Marcellus Shale (e.g., Blondes et al., 2020; Tasker et al., 2020), or that (2) the Appalachian and Arkla basins are geochemically similar enough to cause misclassification. However, if this were the case, it would be more likely that more than one sample would have been classified to the Appalachian Basin and fewer classified to the Arkla.

Ultimately, these results suggest that the model is likely well trained in some basins and lacking in others. Increasing sampling efforts in these “untrained” basins may not only help better understand the geochemical distribution of waters from those producing formations but also further enhance this modeling effort. Additionally, certain basins may be dominated by samples from a specific formation, lithology type, well type (e.g., hydraulically fractured shale versus coalbed methane), or age, and therefore, the USGS PWGD may not fully capture the potential diversity of samples within a given basin. The results from this study, when combined with new testing data, could help determine which current basins are understudied relative to others. For instance, it appears the Gulf Coast is well defined, while the Appalachian basin data included in the USGS PWGD may be biased.

Applications

Correctly and comprehensively understanding the chemistry of waters produced during hydrocarbon extraction is important if these waters are to be

treated and reused. Specifically, understanding which treatment methods would be most beneficial requires an understanding of the geochemical range of a producing formation (e.g., Al-Ghouti et al., 2019; Chang et al., 2019; Scanlon et al., 2020). As suggested in this study, the comprehensive USGS PWGD may still be lacking sample diversity (e.g., water samples from a variety of producing formations within a given basin, different well types, well ages) for many of the 88 basins currently represented, which implies prioritizing diverse sampling efforts from these basins. The results from this study could help screen basins for those with a dearth of data, or those that are missing data from certain producing formations that host waters with a different composition.

Using machine learning to understand produced fluids from hydrocarbon reservoirs has applicability beyond determining basin of origin for produced waters. For example, recent efforts by Snodgrass and Milkov (2020) used gas geochemistry data to train a model that accurately predicts whether the produced gas is microbial in origin. Because microbially generated natural gas has the potential to be stimulated (e.g., Shelton et al., 2014; Ritter et al., 2015; Davis et al., 2018), identifying an efficient and accurate method to screen for reservoirs that already produce natural gas would streamline the enhancement process. Additionally, water chemistry is known to be closely related to the methanogenic activity of a reservoir (e.g., Head et al., 2014; Oren, 2011; Shelton et al., 2014); therefore, using this technique to also predict the methanogenic potential of a reservoir based on water chemistry would be useful in prospecting formations and for developing methanogenic reservoirs. The models could be further refined through efforts to determine if machine learning could accurately predict origin of the brine (e.g., paleo-seawater, meteoric water) based on geochemical or reservoir parameters, formation of origin within a given basin, oil versus gas wells, and lithology type. Additionally, we have identified specific basins that may have less variability across all producing formations; the reasons why this may be occurring warrant further investigation.

CONCLUSIONS

This paper presents evidence that machine learning can be used to predict, at > 70% accuracy,

producing basin of origin for a produced water sample given seven major geochemical parameters. Three different machine learning algorithms were trained on two different datasets originating from the USGS PWGD; one dataset contained more produced water samples with fewer geochemical features, while the other dataset contained fewer produced water samples but had more geochemical features for those samples. A Random Forest model cross-validated 10 times using the “extratrees” split rule and an mtry number equal to 5 predicted on a validation set with the most accuracy, at 73.5%. The most important input parameters for this model were SO_4 and Mg.

It is unclear if the current dataset contains a geochemically diverse set of samples for each producing basin, and therefore, basins with larger individual balanced accuracies, such as the Raton basin, may either be geochemically distinct and easy to classify, or the data used for training the model did not represent the geochemical diversity of the entire basin. It is important to note that over 60 geologic basins represented in the USGS PWGD were missing from this analysis, and therefore, the models produced may not necessarily be comprehensive enough to cover the continental USA, though most of the major petroleum producing basins of the USA were included in this analysis, such as the Permian, Williston, and Appalachian. Results from this work suggest that it is possible to distinguish between different basins with a high degree of accuracy (i.e., balanced accuracy) using only major element geochemistry even though that chemistry varies widely within individual basins and contains data from multiple well types and lithologies. The results presented in this manuscript suggest that machine learning can be applied successfully to classify produced waters to basin of origin, and similar workflows could potentially be applied to a variety of other energy-related systems.

ACKNOWLEDGMENTS

This work was funded by the USGS Energy Resources Program. We thank C. Özgen Karacan for his exceptional insight. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US Government.

OPEN ACCESS

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

DATA AVAILABILITY

The datasets generated during and/or analyzed during the current study are available as a USGS Data Release (Shelton et al., 2021).

CODE AVAILABILITY

The code generated during the current study is available as a USGS Data Release (Shelton et al., 2021).

SUPPLEMENTARY INFORMATION

The online version contains supplementary material available at <https://doi.org/10.1007/s11053-021-09949-8>.

REFERENCES

- Al-Ghouti, M. A., Al-Kaabi, M. A., Ashfaq, M. Y., & Da'na, D. A. (2019). Produced water characteristics, treatment and reuse: A review. *Journal of Water Process Engineering*, 28, 222–239.
- Attanasi, E. D., Freeman, P. A., & Coburn, T. C. (2020). Comparison of machine learning approaches used to identify the

- drivers of Bakken oil well productivity. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. <https://doi.org/10.1002/sam.11487>.
- Batista, G. E., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5–6), 519–533.
- Bein, A., & Dutton, A. R. (1993). Origin, distribution, and movement of brine in the Permian Basin (USA): A model for displacement of connate brine. *Geological Society of America Bulletin*, 105(6), 695–707.
- Bertsimas, D., Pawlowski, C., & Zhuo, Y. D. (2017). From predictive methods to missing data imputation: An optimization approach. *The Journal of Machine Learning Research*, 18(1), 7133–7171.
- Birkle, P., Zouch, M., Alzaqebah, M., & Alwohaibi, M. (2019). Machine learning-based approach for automated identification of produced water types from conventional and unconventional reservoirs. In *EarthDoc: Conference Proceedings, Petroleum Geostatistics 2019*, September 2019, Vol. 2019, pp. 1–5. <https://doi.org/10.3997/2214-4609.201902193>.
- Blondes, M. S., Gans, K. D., Engle, M. A., Kharaka, Y. K., Reidy, M. E., Saraswathula, V., Thorndsen, J. J., Rowan, E. L., & Morrissey, E. A. (2018). U.S. Geological Survey National Produced Waters Geochemical Database (ver. 2.3, January 2018): U.S. Geological Survey data release. <https://doi.org/10.5066/F7J964W8>.
- Blondes, M. S., Shelton, J. L., Engle, M. A., Tremblay, J. P., Doolan, C. A., Jubb, A. M., Chenault, J. C., Rowan, E. L., Haefner, R. J., & Mailot, B. E. (2020). Utica shale play oil and gas brines: Geochemistry and factors influencing wastewater management. *Environmental Science & Technology*, 54(21), 13917–13925.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.
- Bressan, T. S., de Souza, M. K., Girelli, T. J., & Junior, F. C. (2020). Evaluation of machine learning methods for lithology classification using geophysical data. *Computers & Geosciences*, 139, 104475.
- Carpenter, A. B. (1978). Origin and chemical evolution of brines in sedimentary basins. In *SPE Annual Fall Technical Conference and Exhibition*. Society of Petroleum Engineers.
- Chang, H., Li, T., Liu, B., Vidic, R. D., Elimelech, M., & Crittenden, J. C. (2019). Potential and implemented membrane-based technologies for the treatment and reuse of flowback and produced water from shale gas and oil plays: A review. *Desalination*, 455, 34–57.
- Chu, C., Hsu, A. L., Chou, K. H., Bandettini, P., Lin, C., & Initiative, A. D. N. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage*, 60(1), 59–70.
- Cluff, M. A., Hartsock, A., MacRae, J. D., Carter, K., & Mouser, P. J. (2014). Temporal changes in microbial ecology and geochemistry in produced water from hydraulically fractured Marcellus shale gas wells. *Environmental Science & Technology*, 48(11), 6508–6517.
- Davis, K. J., Lu, S., Barnhart, E. P., Parker, A. E., Fields, M. W., & Gerlach, R. (2018). Type and amount of organic amendments affect enhanced biogenic methane production from coal and microbial community structure. *Fuel*, 211, 600–608.
- De Bruin, R. H. (1997). An overview of Bighorn Basin oil and gas fields, with emphasis on Badger Basin Field. In *1997 Bighorn Basin Symposium Guidebook, MTGS-AAPG* (pp. 7–13).
- Engle, M. A. (2019). Chemical and isotopic composition of produced waters from the lower Eagle Ford Group, south-central Texas: U.S. Geological Survey data release. <https://doi.org/10.5066/P9KUH0F6>.
- Engle, M. A., & Blondes, M. S. (2014). Linking compositional data analysis with thermodynamic geochemical modeling: Oilfield brines from the Permian Basin, USA. *Journal of Geochemical Exploration*, 141, 61–70.
- Engle, M. A., & Brunner, B. (2019). Considerations in the application of machine learning to aqueous geochemistry: Origin of produced waters in the northern US Gulf Coast Basin. *Applied Computing and Geosciences*, 3, 100012.
- Engle, M. A., Doolan, C. A., Pitman, J. A., Varonka, M. S., Chenault, J., Orem, W. H., McMahon, P. B., & Jubb, A. M. (2020). Origin and geochemistry of formation waters from the lower Eagle Ford Group, Gulf Coast Basin, south central Texas. *Chemical Geology*, 550, 119754.
- 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. (2016). Geochemistry of formation waters from the Wolfcamp and “Cline” shales: Insights into brine origin, reservoir connectivity, and fluid flow in the Permian Basin, USA. *Chemical Geology*, 425, 76–92.
- Engle, M. A., & Rowan, E. L. (2014). Geochemical evolution of produced waters from hydraulic fracturing of the Marcellus Shale, northern Appalachian Basin: A multivariate compositional data analysis approach. *International Journal of Coal Geology*, 126, 45–56.
- Eugster, H. P., & Surdam, R. C. (1973). Depositional environment of the Green River Formation of Wyoming: A preliminary report. *Geological Society of America Bulletin*, 84(4), 1115–1120.
- Fakhru'l-Razi, A., Pendashteh, A., Abdullah, L. C., Biak, D. R. A., Madaeni, S. S., & Abidin, Z. Z. (2009). Review of technologies for oil and gas produced water treatment. *Journal of Hazardous Materials*, 170(2–3), 530–551.
- Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, 41(4), 1937–1946.
- Feng, C., Gao, C., Yin, Q. Z., Jacobsen, B., Renne, P. R., Wang, J., & Chang, S. C. (2018). Tracking physicochemical conditions of evaporite deposition by stable magnesium isotopes: A case study of late Permian langbeinites. *Geochemistry, Geophysics, Geosystems*, 19(8), 2615–2630.
- Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12(1), 8.
- Gaurav, A. (2017). Horizontal shale well EUR determination integrating geology, machine learning, pattern recognition and multivariate statistics focused on the Permian basin. In *SPE Liquids-Rich Basins Conference-North America*. Society of Petroleum Engineers.
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42.
- Hanor, J. S. (2001). Reactive transport involving rock-buffered fluids of varying salinity. *Geochimica Et Cosmochimica Acta*, 65(21), 3721–3732.
- Hanor, J. S., & McIntosh, J. C. (2006). Are secular variations in seawater chemistry reflected in the compositions of basinal brines? *Journal of Geochemical Exploration*, 89(1–3), 153–156.
- Head, I. M., Gray, N. D., & Larter, S. R. (2014). Life in the slow lane; biogeochemistry of biodegraded petroleum containing reservoirs and implications for energy recovery and carbon management. *Frontiers in Microbiology*, 5, 566.
- Iampen, H. T., & Rostron, B. J. (2000). Hydrogeochemistry of pre-Mississippian brines, Williston Basin, Canada–USA. *Journal of Geochemical Exploration*, 69, 29–35.
- John, G. H. (1995). Robust decision trees: Removing outliers from databases. In *KDD*, Vol. 95, pp. 174–179.

- Kharaka, Y., Gans, K., Rowan, E., Thordsen, J., Conaway, C., Blondes, M., & Engle, M. (2019). Chemical composition of formation water in shale and tight reservoirs: A basin-scale perspective. *Shale: Subsurface Science and Engineering*, 245, 27–43.
- Kharaka, Y. K., & Hanor, J. S. (2003). Deep fluids in the continents: I. Sedimentary Basins. *Trgeo*, 5, 605.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1–26.
- Laaksonen, J., & Oja, E. (1996). Classification with learning k-nearest neighbors. In *Proceedings of International Conference on Neural Networks (ICNN'96)* (Vol. 3, pp. 1480–1483). IEEE.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22.
- Löffler, F. E., & Sanford, R. A. (2005). Analysis of trace hydrogen metabolism. *Methods in Enzymology*, 397, 222–237.
- Lowenstein, T. K., Hardie, L. A., Timofeeff, M. N., & Demicco, R. V. (2003). Secular variation in seawater chemistry and the origin of calcium chloride basinal brines. *Geology*, 31(10), 857–860.
- Ma, J., Cheng, J. C., Jiang, F., Chen, W., Wang, M., & Zhai, C. (2020). A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. *Energy and Buildings*, 216, 109941.
- Marjanović, M., Kovačević, M., Bajat, B., & Voženilek, V. (2011). Landslide susceptibility assessment using SVM machine learning algorithm. *Engineering Geology*, 123(3), 225–234.
- McDevitt, B., Cavazza, M., Beam, R., Cavazza, E., Burgos, W. D., Li, L., & Warner, N. R. (2020). Maximum removal efficiency of barium, strontium, radium, and sulfate with optimum AMD-Marcellus flowback mixing ratios for beneficial use in the Northern Appalachian Basin. *Environmental Science & Technology*, 54(8), 4829–4839.
- McIntosh, J. C., Garven, G., & Hanor, J. S. (2011). Impacts of Pleistocene glaciation on large-scale groundwater flow and salinity in the Michigan Basin. *Geofluids*, 11(1), 18–33.
- McIntosh, J. C., & Walter, L. M. (2005). Volumetrically significant recharge of Pleistocene glacial meltwaters into epicratonic basins: Constraints imposed by solute mass balances. *Chemical Geology*, 222(3–4), 292–309.
- McIntosh, J. C., Walter, L. M., & Martini, A. M. (2002). Pleistocene recharge to midcontinent basins: Effects on salinity structure and microbial gas generation. *Geochimica Et Cosmochimica Acta*, 66(10), 1681–1700.
- McIntosh, J. C., Warwick, P. D., Martini, A. M., & Osborn, S. G. (2010). Coupled hydrology and biogeochemistry of Paleocene-Eocene coal beds, northern Gulf of Mexico. *GSA Bulletin*, 122(7–8), 1248–1264.
- McMahon, P. B., Galloway, J. M., Hunt, A. G., Belitz, K., Jurgens, B. C., & Johnson, T. D. (2020). Geochemistry and age of groundwater in the Williston Basin, USA: Assessing potential effects of shale-oil production on groundwater quality. *Applied Geochemistry*, 125, 104833.
- Meyer, R. F. (Ed.). (1968). *AAPG-CSD geological provinces code map: American Association of Petroleum Geologists Bulletin*, 1 sheet, scale 1:5,000,000.
- Meyer, R. F. (1970). Geologic provinces code map for computer use. *AAPG Bulletin*, 54(7), 1301–1305.
- Meyer, R. F., Wallace, L. G., & Wagner, F. J., Jr. (Eds.). (1991). *AAPG-CSD geological provinces code map: Association of Petroleum Geologists Bulletin*, Vol. 75, No. 10, pp. 1644–1651, 1 sheet, scale 1:7,500,000.
- Mihaljevic, B., Larranaga, P., & Bielza, C. (2018). bnlassify: Learning Bayesian Network Classifiers. *The R Journal*, 10(2), 455–468.
- Mohaghegh, S. D. (2020). Subsurface analytics: Contribution of artificial intelligence and machine learning to reservoir engineering, reservoir modeling, and reservoir management. *Petroleum Exploration and Development*, 47(2), 225–228.
- Nicot, J. P., Darvari, R., Eichhubl, P., Scanlon, B. R., Elliott, B. A., Bryndzia, L. T., & Fall, A. (2020). Origin of low salinity, high volume produced waters in the Wolfcamp Shale (Permian), Delaware Basin, USA. *Applied Geochemistry*, 122, 104771.
- Nicot, J.-P., Gherabati, A., Darvari, R., & Mickler, P. (2018). Salinity reversal and water freshening in the Eagle Ford Shale Texas, USA. *ACS Earth and Space Chemistry*, 2, 1087–1094.
- Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2020). *vegan: Community Ecology Package. R package version 2.5-7*. <https://CRAN.R-project.org/package=vegan>.
- Oren, A. (2011). Thermodynamic limits to microbial life at high salt concentrations. *Environmental Microbiology*, 13(8), 1908–1923.
- Osborn, S. G., McIntosh, J. C., Hanor, J. S., & Biddulph, D. (2012). Iodine-129, 87Sr/86Sr, and trace elemental geochemistry of northern Appalachian Basin brines: Evidence for basinal-scale fluid migration and clay mineral diagenesis. *American Journal of Science*, 312(3), 263–287.
- R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Ritter, D., Vinson, D., Barnhart, E., Akob, D. M., Fields, M. W., Cunningham, A. B., Orem, W., & McIntosh, J. C. (2015). Enhanced microbial coalbed methane generation: A review of research, commercial activity, and remaining challenges. *International Journal of Coal Geology*, 146, 28–41.
- Rosenblum, J., Nelson, A. W., Ruyle, B., Schultz, M. K., Ryan, J. N., & Linden, K. G. (2017). Temporal characterization of flowback and produced water quality from a hydraulically fractured oil and gas well. *Science of the Total Environment*, 596, 369–377.
- Rowan, E. L., Engle, M. A., Kraemer, T. F., Schroeder, K. T., Hammack, R. W., & Doughten, M. W. (2015). Geochemical and isotopic evolution of water produced from Middle Devonian Marcellus shale gas wells, Appalachian basin, Pennsylvania Geochemistry of Produced Water from Marcellus Shale Water, PA. *AAPG Bulletin*, 99(2), 181–206.
- Scanlon, B. R., Reedy, R. C., Xu, P., Engle, M., Nicot, J. P., Yoxtheimer, D., & Ikonnikova, S. (2020). Can we beneficially reuse produced water from oil and gas extraction in the US? *Science of the Total Environment*, 717, 137085.
- Shelton, J., Jubba, A., Saxe, S., Attanasi, E., Freeman, P., Blondes, M., & Croke, M. (2021). Input Files and Code for: Machine learning can accurately assign geologic basin to produced water samples using major geochemical parameters. *U.S. Geological Survey Data Release*. <https://doi.org/10.5066/P95G2SZC>.
- Shelton, J. L., McIntosh, J. C., Warwick, P. D., & Yi, A. L. Z. (2014). Fate of injected CO₂ in the Wilcox Group, Louisiana, Gulf Coast Basin: Chemical and isotopic tracers of microbial-brine-rock-CO₂ interactions. *Applied Geochemistry*, 51, 155–169.
- Snodgrass, J. E., & Milkov, A. V. (2020). Web-based machine learning tool that determines the origin of natural gases. *Computers & Geosciences*, 145, 104595.
- Sorenson, R. P. (2005). A dynamic model for the Permian Panhandle and Hugoton fields, western Anadarko basin. *AAPG Bulletin*, 89(7), 921–938.
- Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1), 25.

- Swezey, C. (2002). Regional stratigraphy and petroleum systems of the Appalachian Basin, North America. *Geologic Investigations Series Map I-2768*.
- Tasker, T. L., Warner, N. R., & Burgos, W. D. (2020). Geochemical and isotope analysis of produced water from the Utica/Point Pleasant Shale, Appalachian Basin. *Environmental Science: Processes & Impacts*, 22, 1224–1232.
- Toner, R. N., Lynds, R. L., Stafford, J. E. & Pisel, J. R. (2018). *Interactive oil and gas map of Wyoming*, Wyoming State Geological Survey. <http://wsgs.maps.arcgis.com/apps/webappviewer/index.html?id=-3f7ab99343c34bd3ac5ae6ac8c04d95a> (revised May 2018).
- Tuszynski, J. (2020). *caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc. R package version 1.18.0*. <https://CRAN.R-project.org/package=caTools>.
- U.S. EPA. (2016). *Hydraulic fracturing for oil and gas: Impacts from the hydraulic fracturing water cycle on drinking water resources in the United States* (Final Report). U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-16/236F.
- Ulmer-Scholle, D. S., & Scholle, P. A. (1994). Replacement of evaporites within the Permian Park City Formation, Bighorn Basin, Wyoming, USA. *Sedimentology*, 41(6), 1203–1222.
- Varonka, M. S., Gallegos, T. J., Bates, A. L., Doolan, C., & Orem, W. H. (2020). Organic compounds in produced waters from the Bakken Formation and Three Forks Formation in the Williston Basin, North Dakota. *Heliyon*, 6(3), e03590.
- Weihs, C., Ligges, U., Luebke, K., & Raabe, N. (2005). klaR Analyzing German Business Cycles. In D. Baier, R. Decker, & L. Schmidt-Thieme (Eds.), *Data analysis and decision support* (pp. 335–343). Springer.
- Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), 1–20.
- Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A grammar of data manipulation. R package version 1.0.2*. <https://CRAN.R-project.org/package=dplyr>.
- Wickham, et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>.
- Wilson, T. P., & Long, D. T. (1993). Geochemistry and isotope chemistry of CaNaCl brines in Silurian strata, Michigan Basin, USA. *Applied Geochemistry*, 8(5), 507–524.
- Włodarczyk, M. A. (2016). *Shallow aquifer geochemistry in areas of intense coal bed methane recovery within the Raton Basin, Colorado* (Doctoral dissertation, California State Polytechnic University, Pomona).
- Wright, M. N., Wager, S., & Probst, P. (2020). *Ranger: A fast implementation of Random Forests, R package version 0.12.1, 2020*. Available at <https://CRAN.R-project.org/package=ranger>.
- Zhang, M. L., Peña, J. M., & Robles, V. (2009). Feature selection for multi-label naive Bayes classification. *Information Sciences*, 179(19), 3218–3229.
- Zhao, T., & Wang, Y. (2020). Interpolation and stratification of multilayer soil property profile from sparse measurements using machine learning methods. *Engineering Geology*, 265, 105430.