HIGH SCHOOL EDITION
Journal of Student Research

Volume 13 Issue 2 (2024)

Examining the Effects of Fracking on Groundwater Using the United States Fracking Well Data

William Lin[1] and Magaly Koch[#]

[1]Belmont High School, USA
[#]Advisor

ABSTRACT

Ever since fracking technology, drilling downward and then horizontally, has been developed to extract oil and gas from underground bedrock, the number of fracking wells has been drilled at an accelerating speed, and the amount of underground water consumed has also increased enormously. The goal of this paper is to address this alarming environmental and climate threat from the proliferation of fracking wells accompanied by a surge in groundwater consumption, through analyzing the up-to-date registry data available as of October 2023 from FracFocus.org on the fracking wells nationwide[1]. The paper starts with a literature review to study the existing research findings on fracking water consumption issues. Next it provides a comprehensive data analysis of the water consumption by fracking wells at both national and state levels. It also explores the relationships between the water consumption with various other factors such as the vertical depth of the wells, the chemical ingredients of the fracking fluid, and the purposes of the additives, and aims to provide insights from the correlation and causality analyses that may offer potential strategies to reduce groundwater consumption in the fracking industry. In addition, the paper employs machine learning techniques such as Random Forest model to explore using predictive models to identify wells that have high likelihood of causing a water contamination issue so that proactive controls can be developed to reduce the occurrence of water quality issues.

Introduction

Hydraulic fracturing, colloquially known as fracking, has emerged as a pivotal technology in the global energy landscape, transforming the extraction of oil and natural gas from unconventional reservoirs. This drilling technique, characterized by its reliance on high-pressure fluid injection to induce fractures in subsurface rock formations, has sparked significant scientific, environmental, and socioeconomic discourse. While fracking has facilitated access to vast reserves of hydrocarbons, its rapid proliferation has prompted intense scrutiny due to concerns regarding its impact on the environment, public health, and social well-being.

During the process of fracking, several steps must be taken. Firstly, the area has to be prepared, which involves clearing the area, building roads and drilling pads, and analyzing the surroundings to prevent spills from damaging the nearby area.[2] Next, a hole is drilled straight down into the ground. A steel pipe known as the surface casing is then installed, and cement is piped in between the walls of the hole and the steel pipe, where it sets.[3] Extra sets of casing and walls may be installed depending on the area. At certain depths known

[1] Frac Focus Chemical Disclosure Registry, "The national hydraulic fracturing chemical disclosure registry.," last modified October 12, 2023, https://fracfocus.org/data-download.
[2] "*Hydraulic fracturing*. Independent Petroleum Association of America. (1970, January 27). https://www.ipaa.org/fracking/ ."
[3] Denchak, M. (2019, April 19). *Fracking 101*. Be a Force for the Future. https://www.nrdc.org/stories/fracking-101#work

as the “kickoff point”, horizontal drilling is done, using a similar process to vertical drilling. Once this is finished, a perforating gun is shot into the wall, creating holes into the rock beyond the wall. A solution of chemicals, sand, and water is pumped at high pressure through these holes, widening the fractures. The sand is left in to widen the cracks, allowing oil and natural gas to flow in easier. Once this process is done, production of oil and natural gas begins as oil and gas flow into the well, and the fracturing fluid is recovered and processed. About 25-75 percent of the fluid is recovered, which is either recycled or disposed of. Once the oil and gas are drained from the area, the wells are filled with cement and the equipment is removed, with the land being returned to what it was before to the best of the landowner and the drilling company’s abilities.[4]

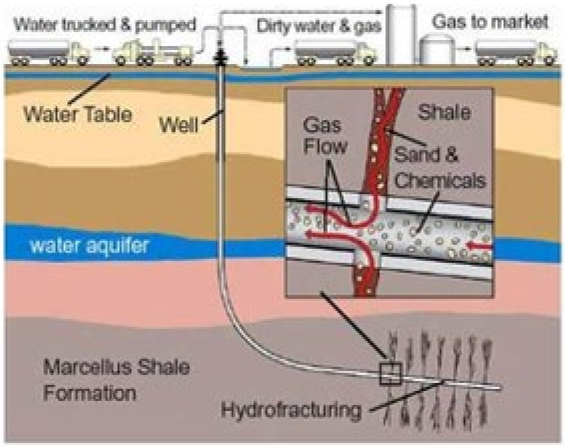
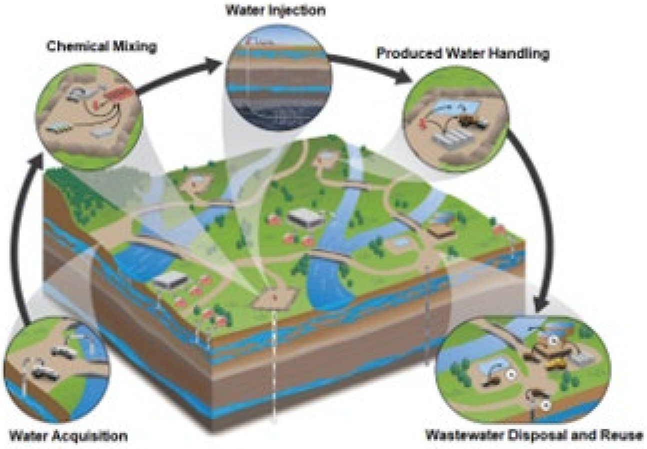


Figure 1. Hydraulic Fracturing, What is it? [5] Source: https://hydrauliceconomics.weebly.com/



[4] *Drilling and the hydraulic fracturing (fracking) process*. UKOOG. (n.d.). https://www.ukoog.org.uk/onshore-extraction/drilling-process

[5] Haley Greenyer, "Hydraulic Fracturing, What is it?," https://hydrauliceconomics.weebly.com/, https://hydrauliceconomics.weebly.com/what-is-fracking.html.

Figure 2. Stages of the hydraulic fracturing water cycle[6] Source: https://www.epa.gov/hfstudy, Pg: ES10

Currently, fracking plays a significant role in the United States oil industry. More than 1.7 million U.S. wells have been completed using the fracking process, producing more than seven billion barrels of oil and 600 trillion cubic feet of natural gas[7]. According to the U.S. Energy Department, up to 95 percent of new wells drilled today are hydraulically fractured, which, according to the EIA in 2018, accounts for two-thirds of total U.S. marketed natural gas production and about half of U.S. crude oil production[8]. This technique is a prominent focus with current natural gas and oil production, with it being more and more used over time.

A large amount of controversy around fracking focuses on its environmental impact. Fracking has a large potential to disrupt habitats due to its large land footprint[9], can increase seismic activity in the area[10], can increase air pollution[11], and has potential hazardous chemical exposure[12]. A prominent focus on these concerns is that of groundwater, with the wastewater disposal and fracking injections potentially allowing hazardous chemicals to enter the groundwater[13]. In this paper we will start with going over the current knowledge of how fracking affects the quality and accessibility of groundwater in the United States. Then we will conduct a comprehensive analysis on the fracking well water consumption data nationwide using the most current registry data available as of October 2023. To accomplish the research goals, we employ various statistical methods and machine learning techniques. We hope the insights gained from this research process can help to identify potential strategies for mitigating underground water consumption in the fracking industry.

Methodology

First, a comprehensive literature review was conducted. We focused on searching from research papers from academic journals such as Environmental Science & Technology, articles from government agencies such as the United States Environmental Protection Agency (EPA), and publications from the environmental science departments in colleges. To identify research papers and websites that have relevant information on fracking,

[6] Office of Research and Development, EPA, "EPA's Study of Hydraulic Fracturing for Oil and Gas and Its Potential Impact on Drinking Water Resources," EPA - United States Environmental Protection Agency, last modified December 2016, https://www.epa.gov/hfstudy.

[7] *Hydraulic fracturing*. Independent Petroleum Association of America. (1970, January 27). https://www.ipaa.org/fracking/

[8] *Hydraulic fracturing*. Energy API. (n.d.). https://www.api.org/news-policy-and-issues/hydraulic-fracturing

[9] *The costs of fracking*. Environment America Research & Policy Center. (2022, September 29). https://environmentamerica.org/center/resources/the-costs-of-fracking/

[10] Quinones, L., DeShon, H. R., Jeong, S., Ogwari, P., Sufri, O., Holt, M. M., & Kwong, K. B. (2019). Tracking induced seismicity in the Fort Worth Basin: A summary of the 2008–2018 north texas earthquake study catalog. *Bulletin of the Seismological Society of America*, *109*(4), 1203–1216. https://doi.org/10.1785/0120190057

[11] Srebotnjak, T., & Rotkin-Ellman, M. (2014, December). *NRDC: Air pollution from hydraulic fracturing threatens public health* ... Fracking Fumes: Air Pollution from Hydraulic Fracturing Threatens Public Health and Communities. https://www.nrdc.org/sites/default/files/fracking-air-pollution-IB.pdf

[12] Endocrine Society. (2020, March 31). *Fracking chemical may interfere with male sex hormone receptor*. https://www.endocrine.org/news-and-advocacy/news-room/2020/fracking-chemical-may-interfere-with-male-sex-hormone-receptor

[13] Landis, J. D., Sharma, M., Renock, D., & Niu, D. (2018). Rapid desorption of radium isotopes from black shale during hydraulic fracturing. 1. source phases that control the release of Ra from marcellus shale. *Chemical Geology*, *496*, 1–13. https://doi.org/10.1016/j.chemgeo.2018.06.013

we used a list of keywords to help to search for sources. The keywords initially used started with simply “fracking”, “groundwater”, “contamination”, “Texas”, and “pollution”. Later, the scope of the paper was expanded, and new keywords introduced were such as “United States”, “health”, “leeching”, and “aquifers”. Data from Government Agencies, Industry Reports, Academic Research, and Non-Governmental Organizations (NGOs) were collected, and sorted into categories for analyses. We then studied the research papers, articles, and data pertaining specifically to the effects on groundwater consumption and contamination in the fracking industry and its environmental implications.

Next we downloaded historical national hydraulic fracturing chemical disclosure data from FracFocus Chemical Disclosure Registry from https://fracfocus.org/data-download.[14] FracFocus provides the release of hydraulic fracturing companies’ disclosure data to the public free of charge and is updated on a daily basis. The data are stored in CSV formats as well as in database table format that can be read using Microsoft SQL server 2019. There are three parts of the data elements that are included. 1) Table contains hydraulic fracturing company’s name, well name, well location, base water volume and total vertical depth. 2) Table contains each disclosure’s additive names and purpose for the additives used. 3) Table contains each disclosure’s chemical ingredients that are used in the additives and jobs. We then imported the downloaded data into a Python environment. The database covers registry data from around 1,800 companies with about 206,000 wells. The detailed registry table with additional data elements such as purposes for additives and ingredients has about 6.1 million rows. Since data may be manually entered into the databases, there are typos in some fields. We did some manual data fixes such as fixing the name of the states prior to analyzing the data. Statistical analysis such as distribution analysis, correlation analysis, regression analysis, and boxplot were used to examine the water consumption in wells and to explore which factors may affect the total water consumed by a well.

Thirdly, we use Natural language Processing (NLP) and Machine Learning techniques such as Cluster Analysis to extract and analyze the data when there is a need. For example, since the additive, purpose and ingredient data are stored in free text forms across over 6 million rows, it is hard to transform the data for analytical needs using traditional query methods. We tried several Natural Language Processing Python packages such as TF-IDF and NLTK to extract top keywords from the data fields for the entities. Then we created dummy variables, 1 or 0, to indicate whether a well has used or has not used certain ingredients identified by keywords such as “methanol”. Transforming the keywords into dummy variables allows us to use them as independent variables in a regression analysis and machine learning models in order for us to test the relationships between these factors and the independent variable, the total volume of water consumed. In addition, Machine learning techniques such as Logistic Regressions and Random Forest models are used to learn from the reported water contamination cases and to build models that may help to predict future violations on water quality. However, machine learning relies on learning from prior knowledge, but we ran into obstacles to obtaining the previous data that can be used as training cases. Since the water contamination violation reporting database is not easily obtained by the public, we searched online and identified several reports with information on companies that had reported violation cases. 1)Violations Per Well Among PA Operators.[15] 2) Fracking’s Environmental Impacts: Water.[16] 3) Fracking’s Most Wanted: Lifting the Veil Oil and Gas Company Spills and

[14] Frac Focus Chemical Disclosure Registry, "The national hydraulic fracturing chemical disclosure registry.," last modified October 12, 2023, https://fracfocus.org/data-download.

[15] Matt Kelso, "VIOLATIONS PER WELL AMONG PA OPERATORS," FracTracker.org, last modified October 29, 2013, https://www.fractracker.org/2013/10/violations-per-well-among-pa-operators/.

[16] Les Stone, "Fracking's Environmental Impacts: Water," https://www.greenpeace.org/, https://www.greenpeace.org/usa/fighting-climate-chaos/issues/fracking/environmental-impacts-water/#:~:text=Contamination%20of%20Water%20Wells%20and,water%20for%20many%20rural%20communities.

Violations.[17] For companies that are mentioned in the reports, we created a water contamination violation flag, 1 being yes for contamination 0 being no for contamination. The violation flag is then joined back with other data from FracFocus at company level, which is then used to build the machine learning models. However, since we only have a limited number of violation cases, we do not have enough training data, therefore, more data needs to be obtained to enhance the accuracy of the models to predict companies that may have higher likelihood of having a water quality violation.

Results and Discussions

Results from Literature Review

For the literature review, the keyword searches identified a couple of hundred research papers and websites, and we ended up reading about 50 articles for this research. The literature review section of this study has several key findings. Firstly, the presence of drilling wells was linked to higher concentrations of methane in drinking wells near the natural-gas wells[18]. On a study done on 68 private water wells in Upstate New York as well as Northeast Pennsylvania, 60 of which were also analyzed for dissolved-gas concentrations of methane and higher-chain hydrocarbons and for carbon and hydrogen isotope ratios of methane, 51 out of those 60 had methane concentrations detected, with methane concentrations were 17-times higher on average (19.2 mg CH4 L−1) in shallow wells from active drilling and extraction areas than in wells from nonactive areas(1.1 mg L−1 on average).[19] In addition, several harmful compounds have been detected in the wells near fracking sites in the Barnett Shale region of Texas, a region known for high fracking activity, such as methanol, ethanol, dichloromethane(DCM), several BTEX class compounds(benzene, toluene, ethylbenzene, xylene), arsenic, and strontium[20,21]. Studies done on the Marcellus Shale formation with samples from Indiana County, Pennsylvania, Chenango County, New York; and Yates County, New York showed that the process of hydraulic fracturing

[17] NRDC (the Natural Resources Defense Council), Fracking's Most Wanted: Lifting the Veil on Oil and Gas Company Spills and Violations, report no. ip:15-01-a, [Page #], April 2015, https://www.nrdc.org/sites/default/files/fracking-company-violations-IP.pdf.

[18] Osborn, S. G., Vengosh, A., Warner, N. R., & Jackson, R. B. (2011). Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing. *Proceedings of the National Academy of Sciences*, *108*(20), 8172–8176. https://doi.org/10.1073/pnas.1100682108

[19] Osborn, S. G., Vengosh, A., Warner, N. R., & Jackson, R. B. (2011). Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing. *Proceedings of the National Academy of Sciences*, *108*(20), 8172–8176. https://doi.org/10.1073/pnas.1100682108

[20] Hildenbrand, Z. L., Carlton, D. D., Fontenot, B. E., Meik, J. M., Walton, J. L., Taylor, J. T., Thacker, J. B., Korlie, S., Shelor, C. P., Henderson, D., Kadjo, A. F., Roelke, C. E., Hudak, P. F., Burton, T., Rifai, H. S., & Schug, K. A. (2015). A comprehensive analysis of groundwater quality in the Barnett Shale Region. *Environmental Science &amp; Technology*, *49*(13), 8254–8262. https://doi.org/10.1021/acs.est.5b01526

[21] Fontenot, B. E., Hunt, L. R., Hildenbrand, Z. L., Carlton Jr., D. D., Oka, H., Walton, J. L., Hopkins, D., Osorio, A., Bjorndal, B., Hu, Q. H., & Schug, K. A. (2013). An evaluation of water quality in private drinking water wells near natural gas extraction sites in the Barnett Shale Formation. *Environmental Science &amp; Technology*, *47*(17), 10032–10040. https://doi.org/10.1021/es4011724 .

could cause radium to begin to leech into nearby areas, potentially contaminating aquifers[22][23]. Lastly, the process of fracking is one that consumes a large amount of water, with the water use per well increased up to 770% from 2011 to 2016[24] and the EPA raising concerns over water withdrawal in areas with limited or declining groundwater resources. [25]

From these results, a correlation between fracking and the lowered quality and quantity of groundwater in certain areas can be drawn. While a link between the direct contamination due to fracking can be formed, it cannot be concretely proven with current data, and we suggest that further studies be done on the matter to gain conclusive evidence. The increased water usage, however, can be concretely shown, and as more than half of the continental United States has experienced drought conditions with 40 of 50 state water managers expected shortages in some portion of their states over the next 10 years in 2014[26]. This issue is one that should be addressed. Steps should be taken to either reduce the water usage done by fracking or to reduce the usage of fracking in areas that experience water shortages. The issue of direct contamination is one that warrants further studies, and as the fracking industry grows in the United States, drought conditions as well as concerns of contamination are likely to make the issue more contentious.

Results from Data Analysis and Statistical Analysis

Using Python code, we analyzed the historical national hydraulic fracturing chemical disclosure data from FracFocus Chemical Disclosure Registry. The goal of this exercise is to examine the relationships between the total base water consumed by a well with other available data elements such as the total vertical depth, purposes for the additives used, and ingredients used in the job and additives. Since water resources are scarce in many regions, using large quantities of water resources in some regions significantly impacts the drinking water availability and even causes drought for the land near the fracking wells. Understanding the factors that may contribute to a surging underground water consumption in fracking wells can offer potential strategies to reduce water usage in wells.

The following chart breaks down by state for the 206,000 wells registered by the 1,800 companies historically. Texas accounts for about half of total fracking wells, followed by Colorado, Oklahoma, North Dakota, New Mexico, and Pennsylvania.

[22] Landis, J. D., Sharma, M., Renock, D., & Niu, D. (2018). Rapid desorption of radium isotopes from black shale during hydraulic fracturing. 1. source phases that control the release of Ra from marcellus shale. *Chemical Geology*, *496*, 1–13. https://doi.org/10.1016/j.chemgeo.2018.06.013

[23] Landis, J. D., Sharma, M., & Renock, D. (2018). Rapid desorption of radium isotopes from black shale during hydraulic fracturing. 2. A model reconciling radium extraction with marcellus wastewater production. *Chemical Geology*, *500*, 194–206. https://doi.org/10.1016/j.chemgeo.2018.08.001

[24] Kondash, A. J., Lauer, N. E., & Vengosh, A. (2018). The intensification of the water footprint of hydraulic fracturing. *Science Advances*, *4*(8). https://doi.org/10.1126/sciadv.aar5982

[25] U.S. EPA. 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, 2016.

[26] "Solutions to address water scarcity in the U.S. The Nature Conservancy. (2020, February 13). https://www.nature.org/en-us/what-we-do/our-priorities/provide-food-and-water-sustainably/food-and-water-stories/solutions-address-water-scarcity-us/

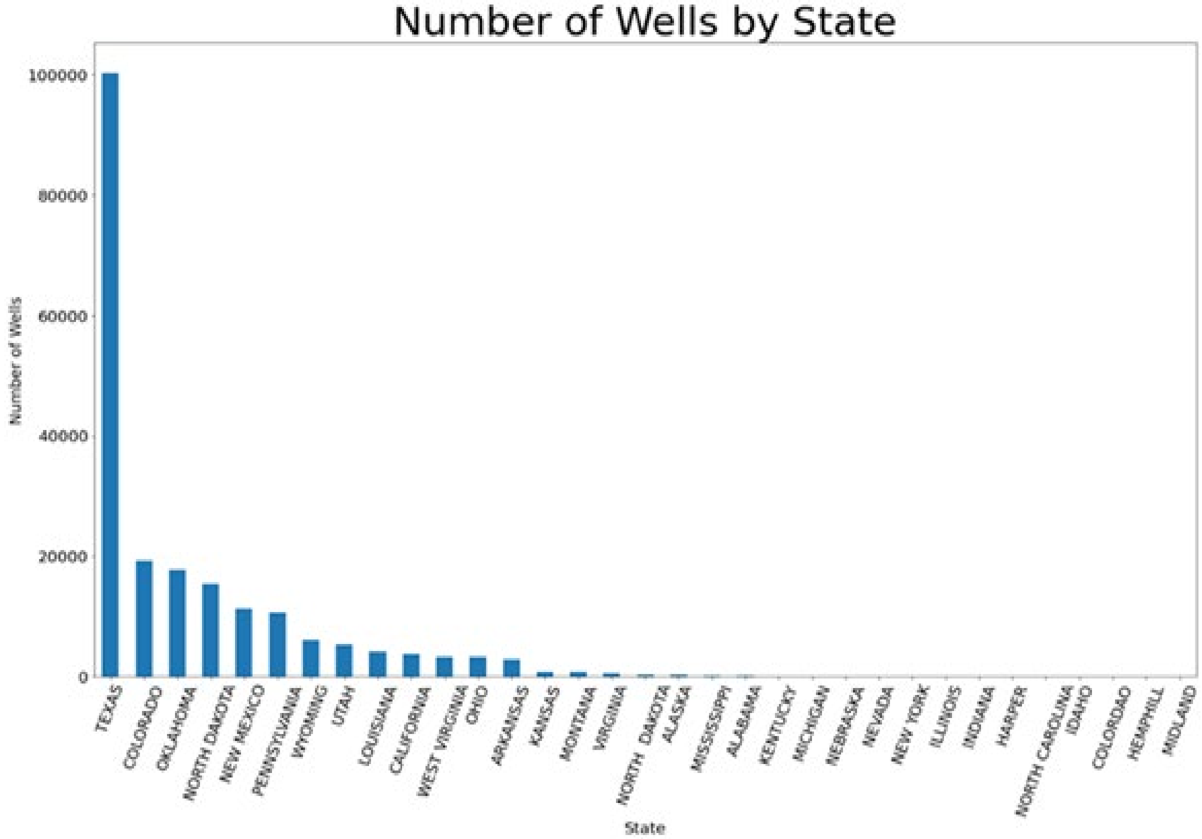


Figure 3. Number of Wells by State

The next chart lists the operating firms with more than 1,000 wells registered historically. These firms tend to be larger operators that may have wells across different states. There are 6 companies which have registered more than 5,000 wells historically, EOC Resources, Anadarko Petroleum, Chesapeake, XTO, Pioneer Natural and Devon Energy. These operating firms constitute a big proportion of the total number of wells and the strategies that they are taking could have a significant impact on the total industry. Government agencies such as the United States Environmental Protection Agency (EPA) and organizations such as the Natural Resources Defense Council (NRDC) can set up on-going watch programs to monitor the practices of these firms to avoid excessive water consumption issues.

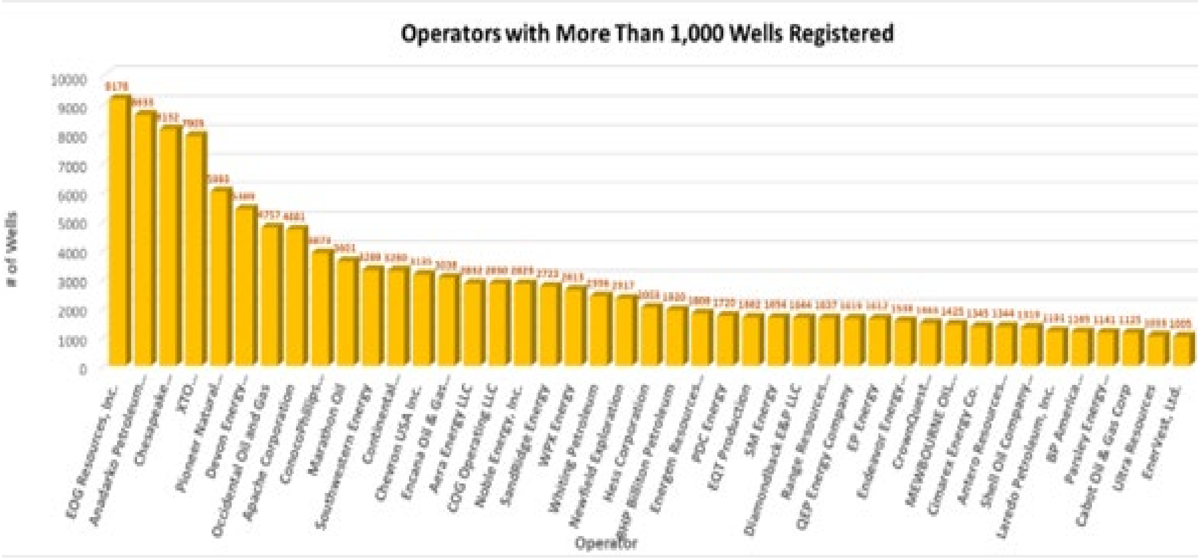


Figure 4. Operators with More Than 1,000 Wells Registered

For the majority of the wells, the database provides data on two import measures, Total Base Water Volume (in gallon) and Total Vertical Depth of the well (in feet). As the first step, we ran some descriptive data analysis for the two measures, and then tried to evaluate whether the Total Vertical Depth of the well has a correlation with the Total Base Water Volume consumed in the well.

Since the Fracking database involves manual data entry that can be error prone, some of the wells showed extremely large TVD and Total Base Water Volume numbers and they are apparently inaccurate in many cases. Therefore, we decided to exclude all wells that had deeper than 25,000 feet in TVD and consumed more than 50 million gallons of water from our analysis. Out of about 206,000 wells, the database has TVD and base water data on about 177,000 wells. We exclude 73 wells with extremely large TVD, 140 wells with more than 50 million Base Water Volume. 2,000 Observations that do not have both measures are also excluded.

Table 1. Median Total Vertical Depth & Total Base Water Volume for the Wells

	Median Total Vertical Depth (Feet)	Median Total Base Water Volume (Gallon)
count	174,685	174,685
median	8,800	6,240,024
mean	8,664	8,525,870
std	2,857	8,336,247
25%	6,928	1,478,442
75%	10,758	13,362,381

Based on the 174,685 wells registered in the US historically, median Total Vertical Depth is about 8,800 feet. The 25 – 75 percentile TVD is between 6,928 and 10,758. Media Total Based Water Volume used is about 6.24 million gallons per well. The 25-75 percentile water consumption per well is between 1.48 million and 13.36 million gallons. These numbers are consistent with what has been published from some other papers from my literature review.

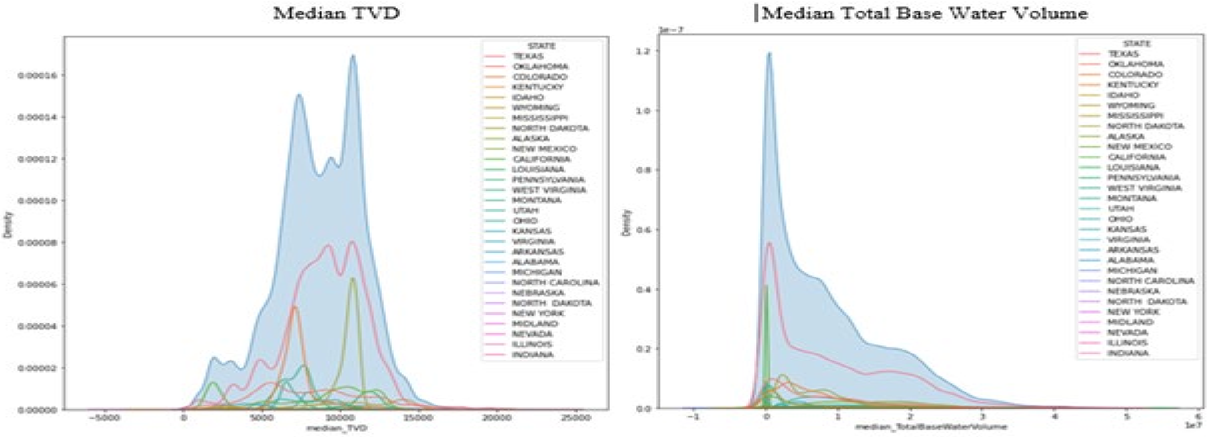


Figure 5. Distribution Curves for Total Vertical Depth & Total Base Volume for the Wells

The above charts show the total and by state distribution curves for the two measures. The TVD distribution curves center in the middle, except that there are two spikes in around 7,000 and 11,000 feet in TVD values. The Total Base Water Volume distribution curve is right skewed, which means that a large proportion of the wells consumed less water but with outliers which had significantly larger water consumption. With the development of more mega fracking projects consuming enormous amounts of water, the distribution curve will be more right skewed.

Next we compared the state-by-state numbers. While each state has shallower and deeper wells, at the Median level, Kentucky, California, Kansas, Virgina, Arkansas, and Alabama have shallower wells, while Wyoming, Mississippi, North Dakota, New Mexico, Louisiana, Montana, North Dakota, and New York have deeper wells. Although the Median numbers fall in the middle, states such as Texas and Oklahoma have a larger number of wells deeper than 20,000 feet.

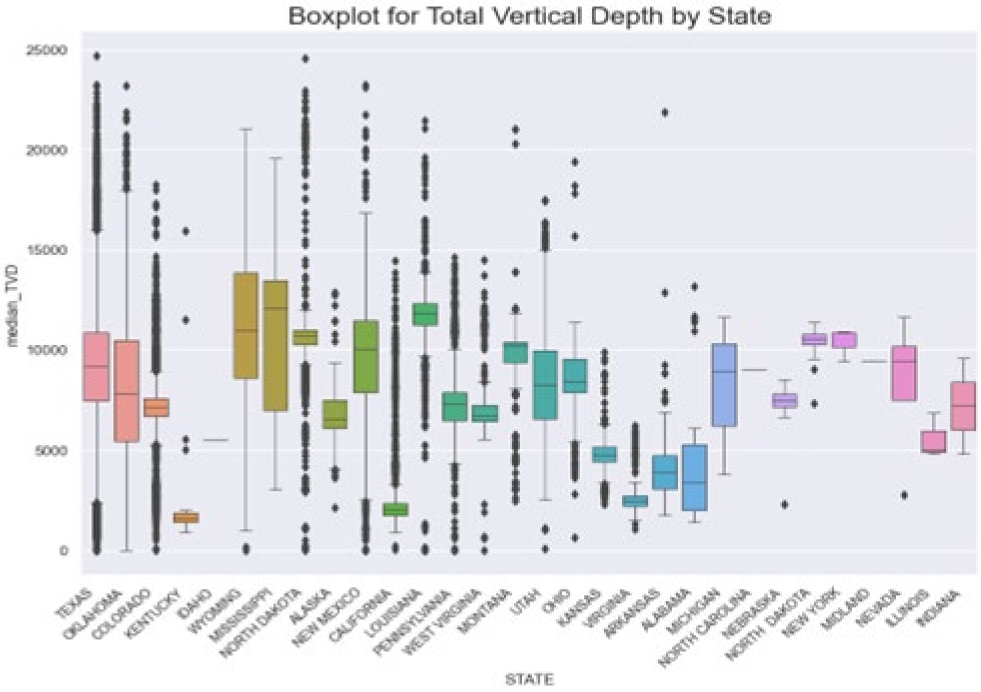


Figure 6. Boxplot for Total Vertical Depth by State

Comparing state by state Total Base Water Volume numbers, Louisiana had the highest Median Total Base Water Volume at over 15 million gallons per well. Other states with a Median Total Base Water Volume above 10 million gallons were New Mexico, West Virginia, Ohio, and New York. Although Texas has a Median water number below 10 million gallons per well, it has the greatest number of wells that consume base water of more than 40 million gallons, followed by Ohio and Pennsylvania. In an interesting article by the New York Times, *'Monster Fracks' Are Getting Far Bigger. And Far Thirstier*, on September 25, 2023[27], refers to these fracking projects consuming enormous amounts of water as "Monster fracks". The article indicates that these "monster fracks" barely existed a decade ago, but now became the industry norm. The article also states that the "monster fracks" constitute almost two out of three fracking wells in Texas nowadays.

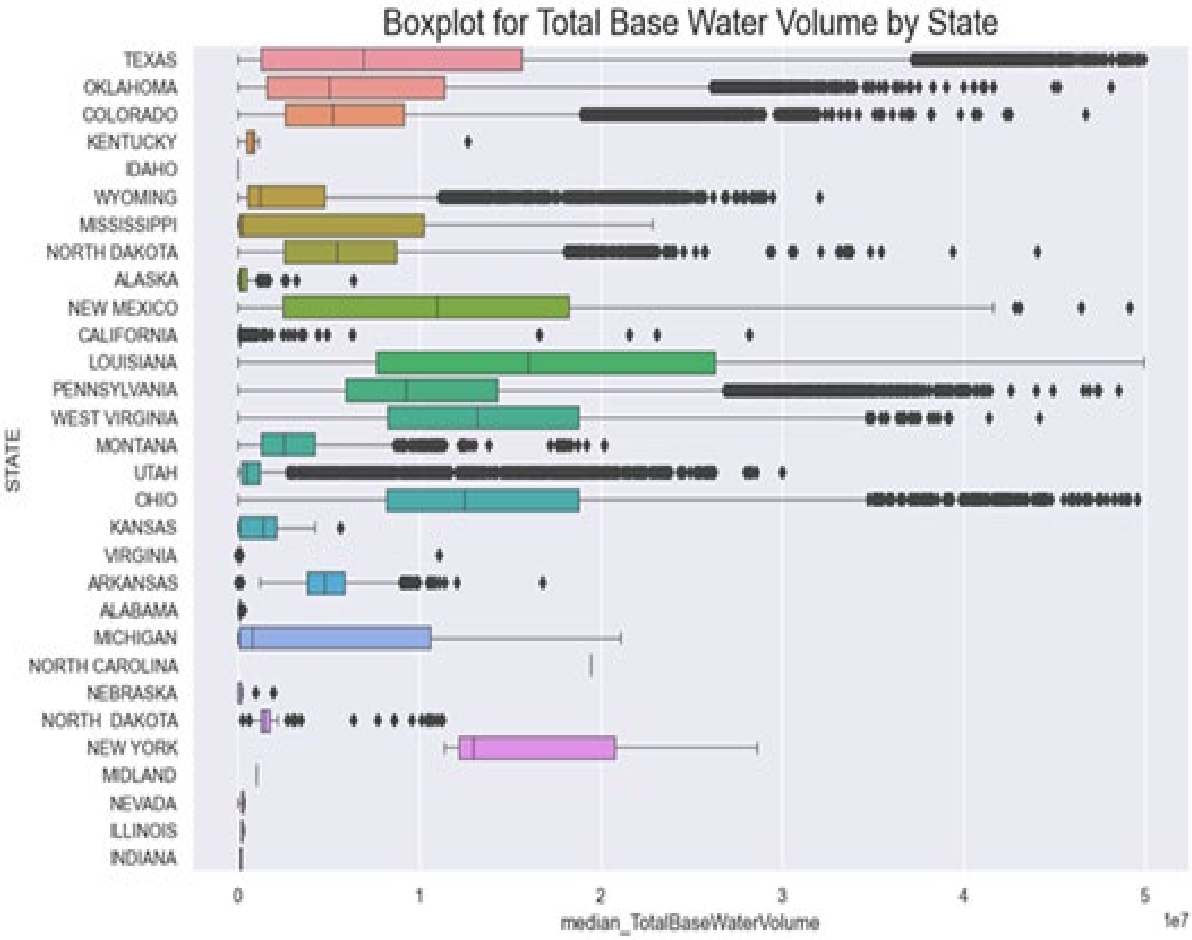
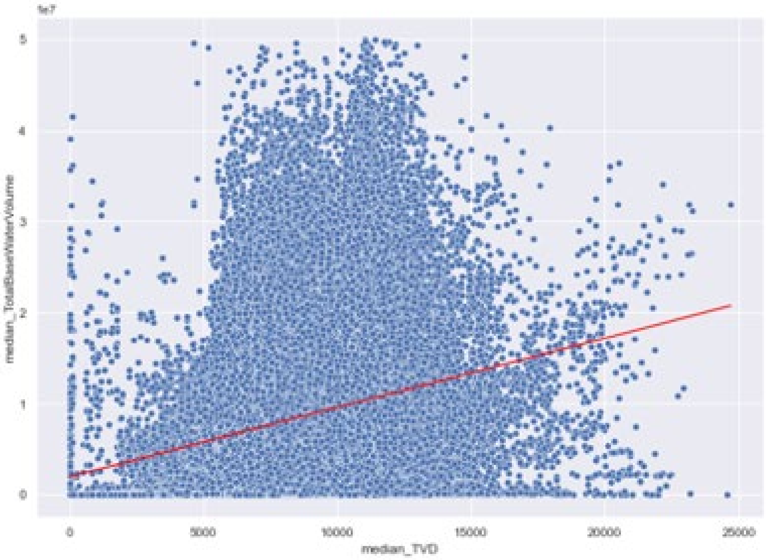


Figure 7. Boxplot for Total Base Water Volume by State

The correlation analysis shows that there is a positive correlation between the two variables (0.26). The simple linear regression analysis using the Ordinary Least Squares (OLS) method shows that the Total

[27] Hiroko Tabuchi and Blacki Migliozzi, "'Monster Fracks' Are Getting Far Bigger. And Far Thirstier.," The New York Times, [Page #], accessed September 25, 2023, https://www.nytimes.com/interactive/2023/09/25/climate/fracking-oil-gas-wells-water.html.

Vertical Depth of a well is a significant factor to explain the Total Base Water Volume consumed by a well. The deeper wells tend to be associated with high level of water consumption. We also took the Z-scores of both variables to standardize the two attributes.



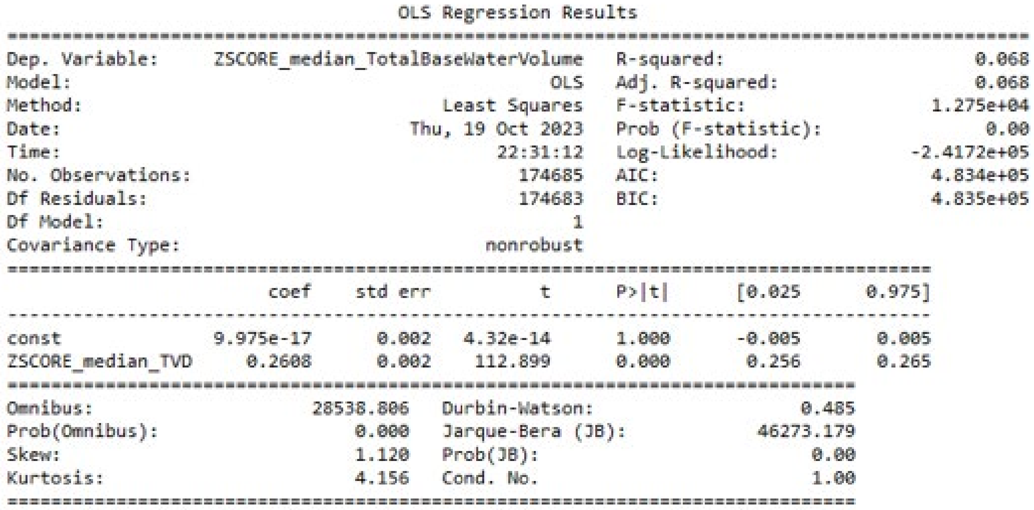
$y = \alpha + \beta x$

y: TotalBaseWaterVolume Zscore

x: TVD (Total Vertical Depth) Zscore

α: Intercept (9.975e-17)

β: Beta Coefficient (0.1608)



```
                                    OLS Regression Results
==============================================================================================
Dep. Variable:     ZSCORE_median_TotalBaseWaterVolume   R-squared:                       0.068
Model:                                            OLS   Adj. R-squared:                  0.068
Method:                                 Least Squares   F-statistic:                 1.275e+04
Date:                                Thu, 19 Oct 2023   Prob (F-statistic):               0.00
Time:                                        22:31:12   Log-Likelihood:            -2.4172e+05
No. Observations:                              174685   AIC:                         4.834e+05
Df Residuals:                                  174683   BIC:                         4.835e+05
Df Model:                                           1
Covariance Type:                            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const              9.975e-17      0.002   4.32e-14      1.000      -0.005       0.005
ZSCORE_median_TVD     0.2608      0.002    112.899      0.000       0.256       0.265
==============================================================================
Omnibus:                    28538.806   Durbin-Watson:                   0.485
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            46273.179
Skew:                           1.120   Prob(JB):                         0.00
Kurtosis:                       4.156   Cond. No.                         1.00
==============================================================================
```

Figure 8. Regression Results

Results from Natural Language Processing and Machine Learning

In addition to the simple regression analysis, we also wanted to run multiple regression analysis to use multiple independent variables in regression. The registry database also provides two additional attributes, PercentHighAdditive and PercentHFjob, that we are going to try to use. In addition, we wanted to include attributes such as ingredients used in fracking fluids as well as purposes for additives. However, the detailed registry table with these data elements are stored in free text forms and has over 6.1 million rows. It is difficult to use simple queries to identify the keywords. Therefore, we researched and tested several Natural Language Process techniques (NLP) in Python such as TF-IDF and NLTK to extract keywords from the data fields. For example, NLTK word_tokenize function can split the text strings into word tokens, and then the NLTK FreqDist function can count the occurrence of certain keywords being used. After the keywords frequencies are calculated, we decided to select the top 40 common keywords for the purposes of additives and ingredients used. The relationships between these key elements and the total base water volume consumed will be examined later. Here are some keyword frequency analyses using NLTK.

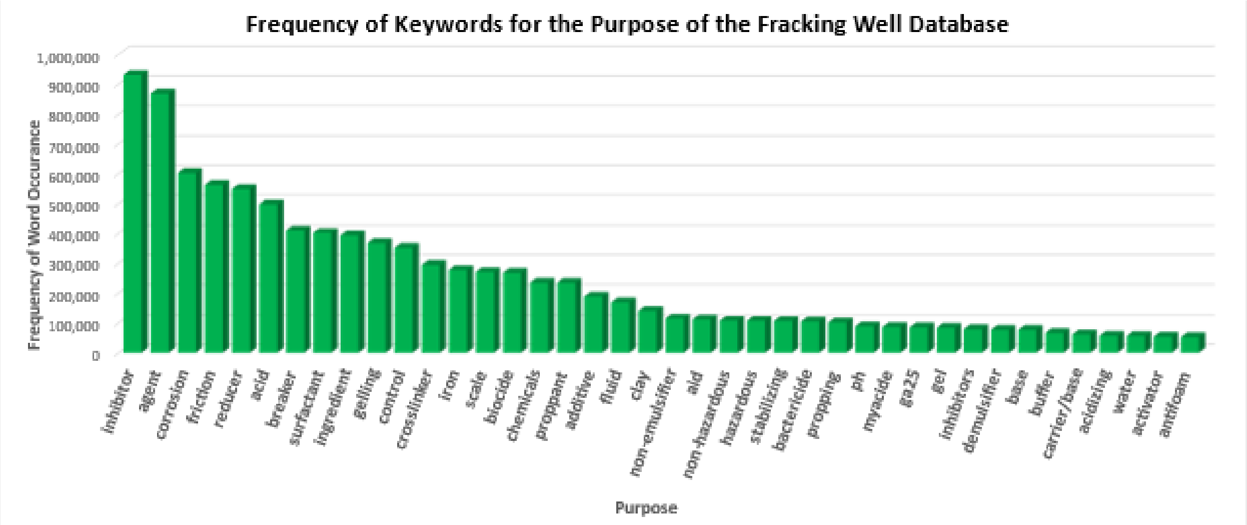


Figure 9. Frequency of Keywords for the Purpose of the Fracking Well Database

The above chart shows the top 40 keywords that have the largest occurrences in the Purpose for additives field in the database. We can see the most frequent words are such as Proppant, Carrier Base Fluid, Friction Reducer, Biocide, Breaker, Scale Inhibitor, Corrosion Inhibitor, Cross Linker, Surfactant, Iron Control, Gelling Agent, Acid, Solvent, Clay Control, Acidizing, etc.

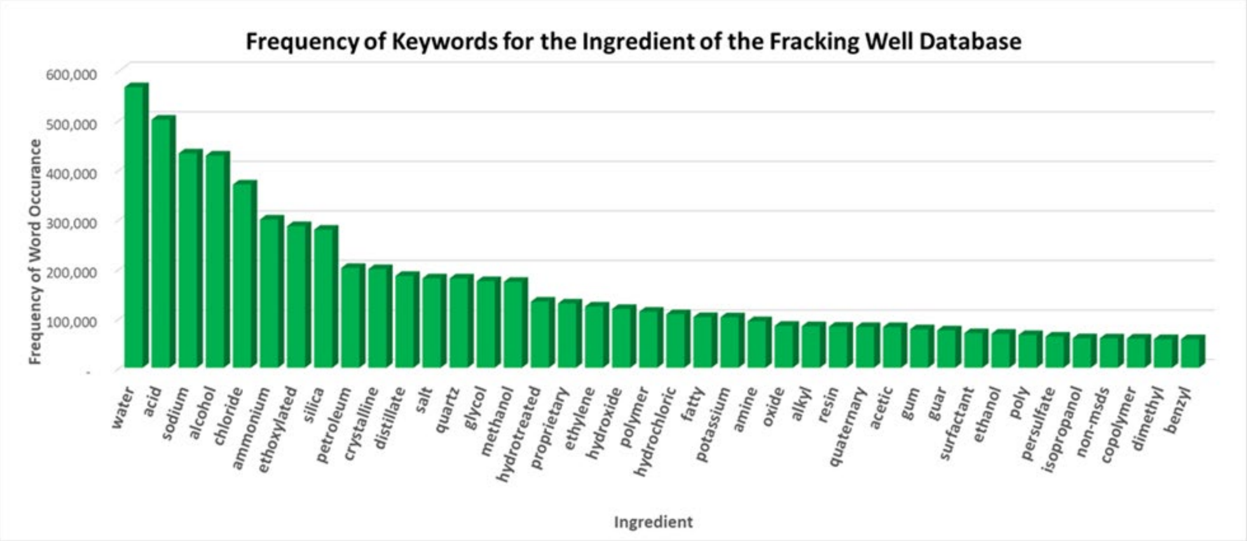


Figure 10. Frequency of Keywords for the Ingredient of the Fracking Well Database

We did a similar analysis to look at the ingredients in fracking fluids for the drilling jobs and additives. The ingredients that are used most often are water, acid, sodium, alcohol, chloride, ammonium, ethoxylated, silica, petroleum, crystalline, distillate, salt, quartz, glycol, methanol, etc. Later in this paper we will evaluate the relationships between these ingredients and the total base water volume consumed in a well. We also created a WordCloud Map to show the frequency of the ingredients. The map has Water in the largest font centered in the middle, clearly indicating that Water is used excessively in fracking and it alerts us to take precaution in the fracking techniques which may bring negative implications to drinking water and natural environment.

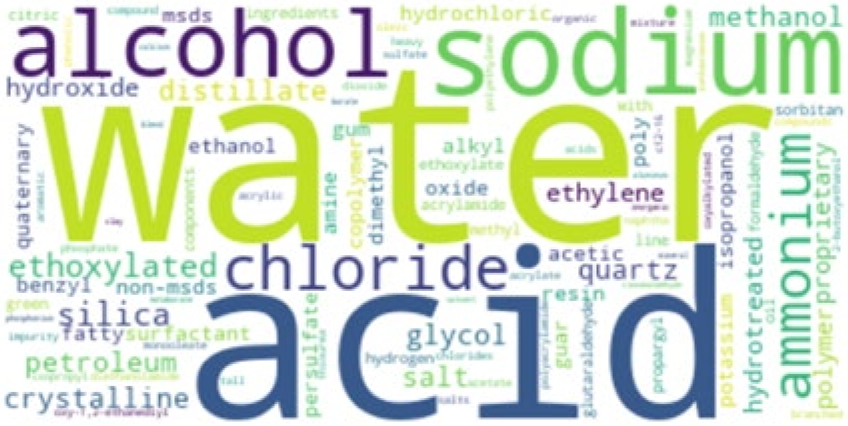


Figure 11. WordCloud Map of Keywords for the Ingredient of the Fracking Well Database

With the keywords for ingredients and Promises being identified using NLP for each entity (Firm/Well/job), we then picked the top keywords and transformed them into a list of dummy variables. For example, if a well uses "ammonium" as an ingredient, then the value of the dummy variable IN_ammonium_ind for this well would be 1. The wells that do not use "ammonium" would have a value of 0 for the dummy variable IN_ammonium_ind. After the transformation, we can use these elements in regressions as well as in feeding them into machine learning models.

With dummy variables created, we first created correlation maps, and then built multiple linear regression models to examine the relationships between the Total Base Water Volume vs. a list of other factors. The following table shows the correlations with the Total Base Water Volume. As we evaluated earlier, there is a positive correlation between the vertical depth of the well and the base water volume consumed by the well. The deeper the well, the more base water is consumed. The following table shows that positive correlations with the base water consumed exist between acid, alcohol, ethoxylated, silica, petroleum, and crystalline. Negative correlations with the base water consumption exist between Percent High Additive, Sodium, salt, quartz, glycol and methanol.

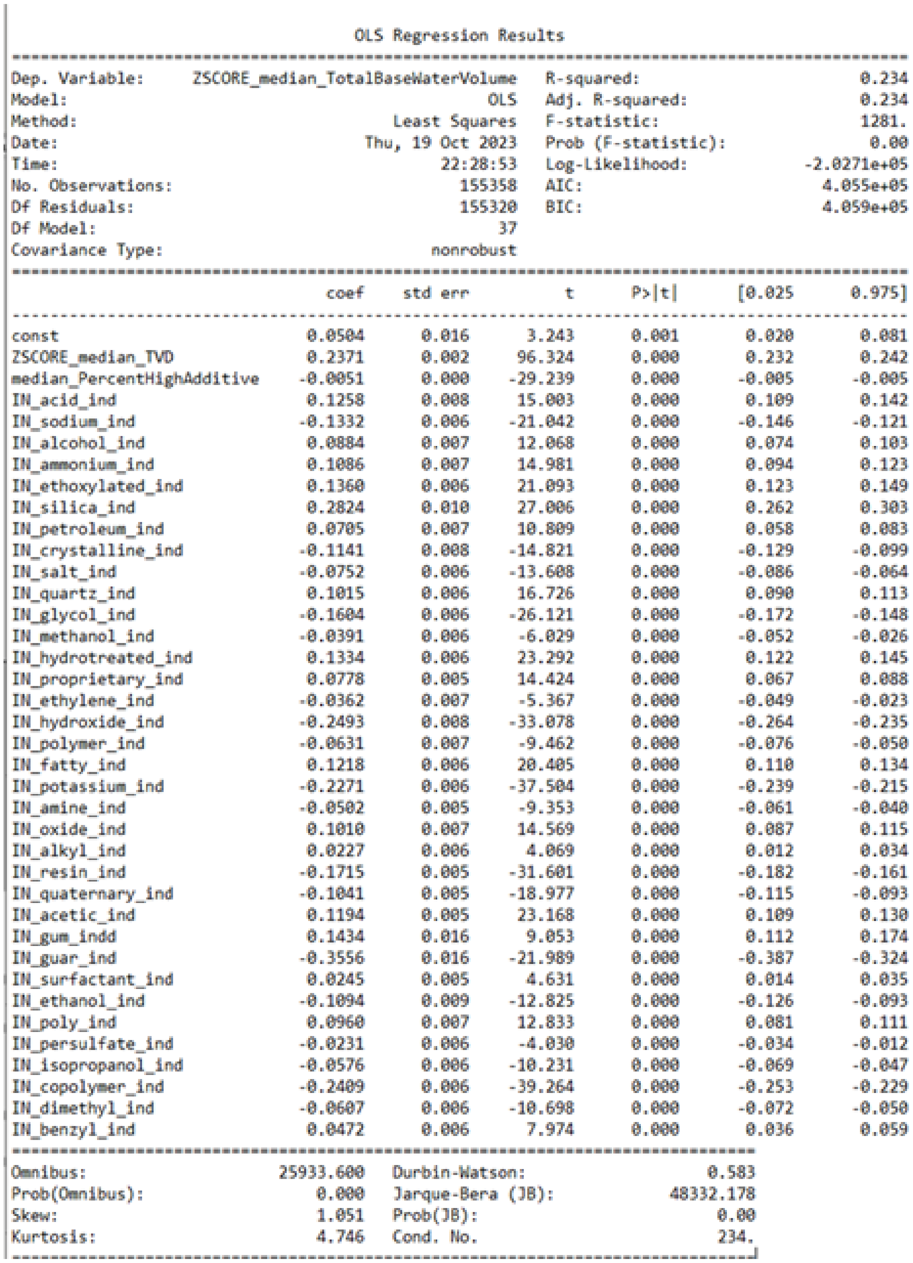
Table 2. Correlations of Total Base Water Volume Consumed with Ingredients

	Zscore Median Total Base Water Volume		Zscore Median Total Base Water Volume		Zscore Median Total Base Water Volume
Zscore Median Total Base Water Volume	1	glycol	-0.182	quaternary	-0.041
Zscore Median TVD	0.259	methanol	-0.026	acetic	0.057
Median Percent High Additive	-0.064	hydrotreated	0.140	gum	-0.214
acid	0.060	proprietary	0.031	guar	-0.252
sodium	-0.169	ethylene	-0.121	surfactant	-0.068
alcohol	0.072	hydroxide	-0.230	ethanol	-0.037
ammonium	0.003	polymer	-0.047	poly	0.002
ethoxylated	0.106	fatty	0.024	persulfate	-0.138
silica	0.080	potassium	-0.236	isopropanol	-0.089
petroleum	0.061	amine	-0.088	copolymer	-0.124
crystalline	0.043	oxide	-0.129	dimethyl	0.004
salt	-0.064	alkyl	0.040	benzyl	0.079
quartz	-0.003	resin	-0.116		

Next we ran a multiple linear regression model. Since there seem to be multicollinearity issues between Ingredients used and purpose of additives, we decided only to include the ingredients in the model. We also ran the regression model a few times, each time dropping the independent variables with coefficient T-value less than 2. Here are the remaining factors. The F-statistics equal to 1281, which indicates that the model as a whole can explain the level of base water volume consumed. However, a further look into each of the factors will help to enhance the model and understand the factor's impact on the total base water consumed.

Equation 1: multiple linear regression

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots + \beta_n x_n$$



```
                                 OLS Regression Results
==============================================================================================
Dep. Variable:     ZSCORE_median_TotalBaseWaterVolume   R-squared:                       0.234
Model:                                            OLS   Adj. R-squared:                  0.234
Method:                                 Least Squares   F-statistic:                     1281.
Date:                                Thu, 19 Oct 2023   Prob (F-statistic):               0.00
Time:                                        22:28:53   Log-Likelihood:            -2.0271e+05
No. Observations:                              155358   AIC:                         4.055e+05
Df Residuals:                                  155320   BIC:                         4.059e+05
Df Model:                                          37
Covariance Type:                            nonrobust
==============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         0.0504      0.016      3.243      0.001       0.020       0.081
ZSCORE_median_TVD             0.2371      0.002     96.324      0.000       0.232       0.242
median_PercentHighAdditive   -0.0051      0.000    -29.239      0.000      -0.005      -0.005
IN_acid_ind                   0.1258      0.008     15.003      0.000       0.109       0.142
IN_sodium_ind                -0.1332      0.006    -21.042      0.000      -0.146      -0.121
IN_alcohol_ind                0.0884      0.007     12.068      0.000       0.074       0.103
IN_ammonium_ind               0.1086      0.007     14.981      0.000       0.094       0.123
IN_ethoxylated_ind            0.1360      0.006     21.093      0.000       0.123       0.149
IN_silica_ind                 0.2824      0.010     27.006      0.000       0.262       0.303
IN_petroleum_ind              0.0705      0.007     10.809      0.000       0.058       0.083
IN_crystalline_ind           -0.1141      0.008    -14.821      0.000      -0.129      -0.099
IN_salt_ind                  -0.0752      0.006    -13.608      0.000      -0.086      -0.064
IN_quartz_ind                 0.1015      0.006     16.726      0.000       0.090       0.113
IN_glycol_ind                -0.1604      0.006    -26.121      0.000      -0.172      -0.148
IN_methanol_ind              -0.0391      0.006     -6.029      0.000      -0.052      -0.026
IN_hydrotreated_ind           0.1334      0.006     23.292      0.000       0.122       0.145
IN_proprietary_ind            0.0778      0.005     14.424      0.000       0.067       0.088
IN_ethylene_ind              -0.0362      0.007     -5.367      0.000      -0.049      -0.023
IN_hydroxide_ind             -0.2493      0.008    -33.078      0.000      -0.264      -0.235
IN_polymer_ind               -0.0631      0.007     -9.462      0.000      -0.076      -0.050
IN_fatty_ind                  0.1218      0.006     20.405      0.000       0.110       0.134
IN_potassium_ind             -0.2271      0.006    -37.504      0.000      -0.239      -0.215
IN_amine_ind                 -0.0502      0.005     -9.353      0.000      -0.061      -0.040
IN_oxide_ind                  0.1010      0.007     14.569      0.000       0.087       0.115
IN_alkyl_ind                  0.0227      0.006      4.069      0.000       0.012       0.034
IN_resin_ind                 -0.1715      0.005    -31.601      0.000      -0.182      -0.161
IN_quaternary_ind            -0.1041      0.005    -18.977      0.000      -0.115      -0.093
IN_acetic_ind                 0.1194      0.005     23.168      0.000       0.109       0.130
IN_gum_indd                   0.1434      0.016      9.053      0.000       0.112       0.174
IN_guar_ind                  -0.3556      0.016    -21.989      0.000      -0.387      -0.324
IN_surfactant_ind             0.0245      0.005      4.631      0.000       0.014       0.035
IN_ethanol_ind               -0.1094      0.009    -12.825      0.000      -0.126      -0.093
IN_poly_ind                   0.0960      0.007     12.833      0.000       0.081       0.111
IN_persulfate_ind            -0.0231      0.006     -4.030      0.000      -0.034      -0.012
IN_isopropanol_ind           -0.0576      0.006    -10.231      0.000      -0.069      -0.047
IN_copolymer_ind             -0.2409      0.006    -39.264      0.000      -0.253      -0.229
IN_dimethyl_ind              -0.0607      0.006    -10.698      0.000      -0.072      -0.050
IN_benzyl_ind                 0.0472      0.006      7.974      0.000       0.036       0.059
==============================================================================
Omnibus:                    25933.600   Durbin-Watson:                   0.583
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            48332.178
Skew:                           1.051   Prob(JB):                         0.00
Kurtosis:                       4.746   Cond. No.                         234.
==============================================================================
```

Figure 12. Multiple Linear Regression Model Output

We then ran an unsupervised cluster analysis by feeding in all the variables into a clustering algorithm. Since we have both numerical variables and categorical variables in the models, we use a clustering algorithm called KPrototypes model that incorporates KMeans model for numerical variables and KMode for categorical data. [28]

The first part of the equation uses Euclidean distance to measure similarities for numerical variables.

Equation 2: KMeans Model Using Euclidean Distance$d(X_i, Z_i\ ) = \sqrt{\sum_{j=1}^{m_r}\left(x_{ij}^{r} - z_{ij}^{r}\right)^2}$

[28] Zhexue Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Mining and Knowledge Discovery* 2 (1998): [Page #].

The second part of the following equation is used to measure similarities for categorical variables.

Equation 3: KMode Model to Measure Similarities for Categorical Data

$$d(X_i, Z_i\ ) = \sqrt{\gamma_i \sum_{j=1+1}^{m_c} \delta\left(x_{ij,}^{c} z_{ij}^{c}\right)}$$

Combining KMeans and KMode models, KPrototypes algorithm enables us to feed both numerical and categorical variables into the clustering model.

Equation 4: KPrototypes Algorithm Equation

$$d(X_i, Z_i\ ) = \sqrt{\sum_{j=1}^{m_r} \left(x_{ij}^{r} - z_{ij}^{r}\right)^2 + \gamma_i \sum_{j=1+1}^{m_c} \delta\left(x_{ij,}^{c} z_{ij}^{c}\right)}$$

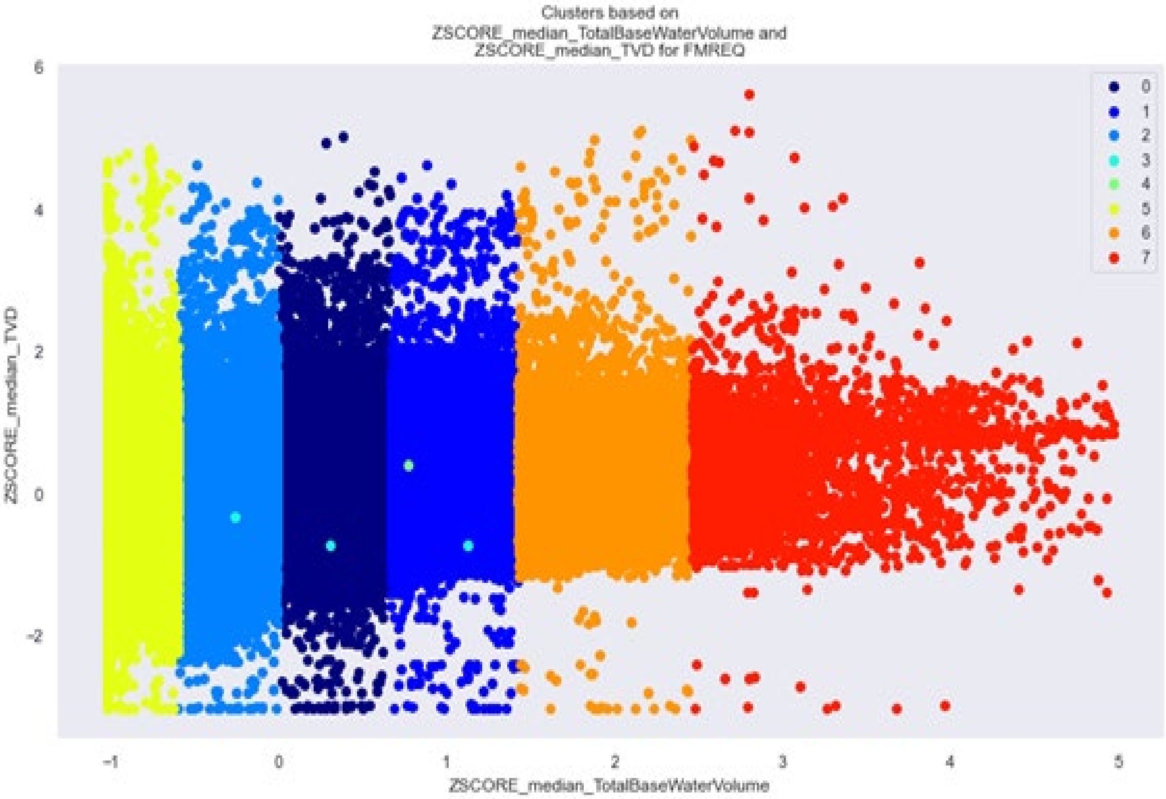


Figure 13. Cluster Analysis Output

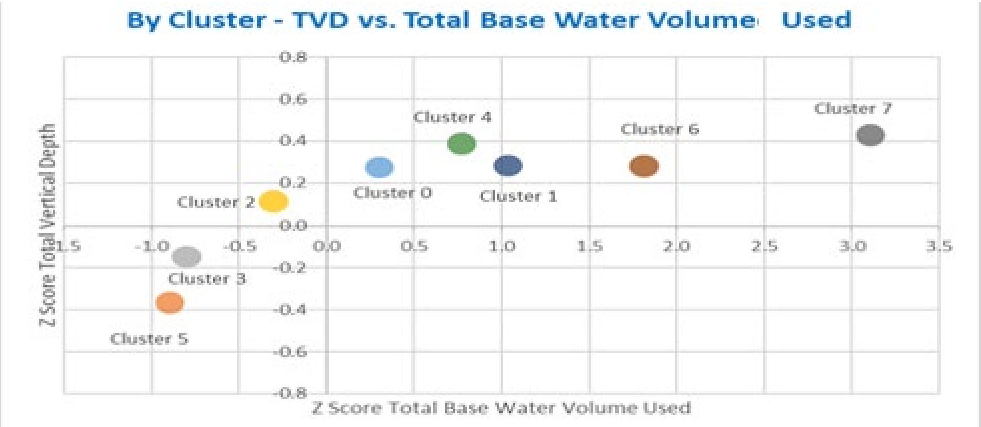
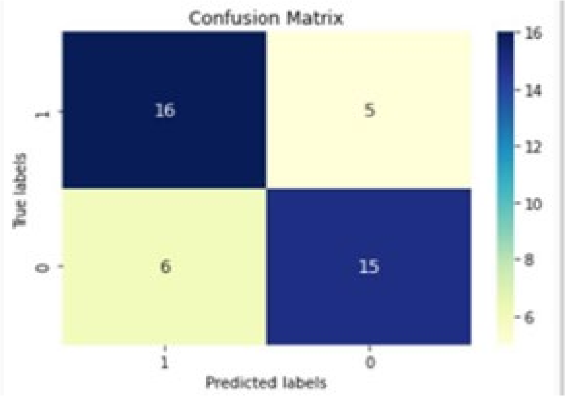
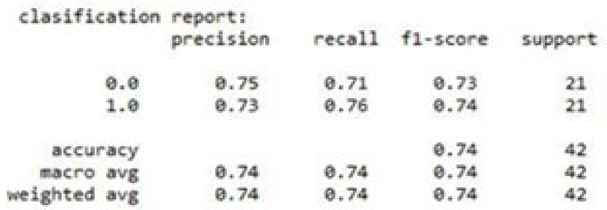


Figure 14. By Cluster – TVD vs. Total Base Water Volume Used

We decided to set K=8 to ask the Clustering algorithm to create 8 clusters based on all the factors fed into the model. The left chart above shows the 8 clusters in a two-dimensional view, Water Volume vs. TVD. The right graph shows the centroid of each cluster, also in a two-dimensional view. The actual clusters are formed using all the factors which are difficult to visualize.

In addition to evaluating the factors that affect the base water volume consumption per well, we also want to evaluate what may be the important factors to affect water quality and cause water contamination in areas close to the fracking wells. We tried to look for the violation reporting database with records on the firms that have had reported cases of water quality violations. Unfortunately, such data is not easily available to the general public free of cost. Therefore, we identified several papers and reports with information on water contamination cases and used these cases as training data. We then translated the data into machine learning ready data format by creating a Boolean variable, where 1 means the firm had reported violation in water contamination, and 0 means no reported water contamination cases. The dataset from the FracFocus database is then joined with this Boolean variable. The goal is to build a machine learning model that can use the registry data provided by FracFocus to predict the likelihood that an operating firm will cause a water contamination incident. We tested both the Logistic Regression model and Random Forest model, and the Random Forest model generated better results in this case. Also, we separated data into two buckets, one, training dataset used to build the model and the other, testing dataset used to test the accuracy of the model. The Confusion Matrix and Classification Report below show the accuracy of the model. When testing the model using a test dataset, the Confusion Matrix on the left below shows that the model had 31 cases of correct prediction out of a total of 42 cases, 16 cases of true positives and 15 cases of true negatives. The model had 11 cases of incorrect prediction, 6 cases of false negative and 5 cases of false positive.




```
 clasification report:
               precision    recall  f1-score   support

         0.0       0.75      0.71      0.73        21
         1.0       0.73      0.76      0.74        21

    accuracy                           0.74        42
   macro avg       0.74      0.74      0.74        42
weighted avg       0.74      0.74      0.74        42
```

Figure 15. Confusion Matrix and Classification Report from the Random Forest Model

The accuracy of the Random Forest model needs to be enhanced. However, we only have a limited number of violation cases available to use as a training dataset. Machine learning relies heavily on learning from prior knowledge. In order to improve the accuracy of this predictive model to identify firms with water contamination probabilities, we need to explore more data sources for further model building needs. In addition, we can test other clustering methods that would perform better with a reduced set of training data.

Conclusions

In conclusion, while fracking has brought significant economic benefits to the United States through providing abundant energy for industrial and personal consumptions, it has undoubtedly created a significant environmental and climate threat from its excessive and surging water consumption and water contamination from chemical ingredients used in fracking fluids. We hope this research using the up-to-date data on fracking wells nationwide can provide people with some insights on the water consumption and its correlating factors in fracking wells. We hope to enhance the predictive model for identifying potential problem wells to empower the regulators and policy makers to take proactive actions. To protect our precious underground water resources and natural environment, we should also advocate for a responsible use of water resources and seek alternative and sustainable energy resources.

Acknowledgments

I would like to thank FracFocus.org for providing the up-to-date registry data on the fracking wells nationwide in the United States.

References

1. Dutzik, T., Ridlington, E., Rumpler, J. (2022, September 29). *The Costs of Fracking, The Price Tag of Dirty Drilling's Environmental Damage.* Environment America Research & Policy Center. https://environmentamerica.org/center/resources/the-costs-of-fracking/.
2. Denchak, M. (2019, April 19). Fracking 101. Be a force for the future. NRDC. (2019, April 19). https://www.nrdc.org/stories/fracking-101#work.
3. UKOOG. (n.d.). *Drilling and the hydraulic fracturing (fracking) process.* https://www.ukoog.org.uk/onshore-extraction/drilling-process.
4. IPAA. Hydraulic fracturing. Independent Petroleum Association of America. https://www.ipaa.org/fracking/.
5. Greenyer, H. Hydraulic Fracturing, What is it? https://hydrauliceconomics.weebly.com/, https://hydrauliceconomics.weebly.com/what-is-fracking.html.
6. Endocrine Society. (2020, March 31). *Fracking chemical may interfere with male sex hormone receptor*. https://www.endocrine.org/news-and-advocacy/news-room/2020/fracking-chemical-may-interfere-with-male-sex-hormone-receptor.
7. Fontenot, B. E., Hunt, L. R., Hildenbrand, Z. L., Carlton Jr., D. D., Oka, H., Walton, J. L., Hopkins, D., Osorio, A., Bjorndal, B., Hu, Q. H., & Schug, K. A. (2013). An evaluation of water quality in private drinking water wells near natural gas extraction sites in the Barnett Shale Formation. *Environmental Science & Technology*, *47*(17), 10032–10040. https://doi.org/10.1021/es4011724.

8. Hildenbrand, Z. L., Carlton, D. D., Fontenot, B. E., Meik, J. M., Walton, J. L., Taylor, J. T., Thacker, J. B., Korlie, S., Shelor, C. P., Henderson, D., Kadjo, A. F., Roelke, C. E., Hudak, P. F., Burton, T., Rifai, H. S., & Schug, K. A. (2015). A comprehensive analysis of groundwater quality in the Barnett Shale Region. *Environmental Science & Technology*, *49*(13), 8254–8262. https://doi.org/10.1021/acs.est.5b01526.
9. Energy API. (n.d.). Hydraulic fracturing. https://www.api.org/oil-and-natural-gas/wells-to-consumer/exploration-and-production/hydraulic-fracturing#sort=date%20descending.
10. Kondash, A. J., Lauer, N. E., & Vengosh, A. (2018). The intensification of the water footprint of hydraulic fracturing. *Science Advances*, *4*(8). https://doi.org/10.1126/sciadv.aar5982.
11. Landis, J. D., Sharma, M., & Renock, D. (2018). Rapid desorption of radium isotopes from black shale during hydraulic fracturing. 2. A model reconciling radium extraction with marcellus wastewater production. *Chemical Geology*, *500*, 194–206. https://doi.org/10.1016/j.chemgeo.2018.08.001.
12. Landis, J. D., Sharma, M., Renock, D., & Niu, D. (2018). Rapid desorption of radium isotopes from black shale during hydraulic fracturing. 1. source phases that control the release of Ra from marcellus shale. *Chemical Geology*, *496*, 1–13. https://doi.org/10.1016/j.chemgeo.2018.06.013.
13. Osborn, S. G., Vengosh, A., Warner, N. R., & Jackson, R. B. (2011). Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing. *Proceedings of the National Academy of Sciences*, *108*(20), 8172–8176. https://doi.org/10.1073/pnas.1100682108.
14. Quinones, L., DeShon, H. R., Jeong, S., Ogwari, P., Sufri, O., Holt, M. M., & Kwong, K. B. (2019). Tracking induced seismicity in the Fort Worth Basin: A summary of the 2008–2018 north texas earthquake study catalog. *Bulletin of the Seismological Society of America*, *109*(4), 1203–1216. https://doi.org/10.1785/0120190057.
15. The Nature Conservancy. (2020, February 13). Solutions to address water scarcity in the U.S. https://www.nature.org/en-us/what-we-do/our-priorities/provide-food-and-water-sustainably/food-and-water-stories/solutions-address-water-scarcity-us/.
16. Srebotnjak, T., & Rotkin-Ellman, M. (2014, December). Fracking Fumes: Air Pollution from Hydraulic Fracturing Threatens Public Health and Communities. https://www.nrdc.org/sites/default/files/fracking-air-pollution-IB.pdf.
17. Frac Focus Chemical Disclosure Registry. The national hydraulic fracturing chemical disclosure registry. (Last modified October 12, 2023). https://fracfocus.org/data-download.
18. Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery.* 2, 283-304.
19. Kelso, M. (Last modified October 29, 2013). Violations Per Well Among PA Operators. Fractracker Alliance https://www.fractracker.org/2013/10/violations-per-well-among-pa-operators/.
20. NRDC (the Natural Resources Defense Council). *Fracking's Most Wanted: Lifting the Veil on Oil and Gas Company Spills and Violations*. Report no. ip:15-01-a. April 2015. https://www.nrdc.org/sites/default/files/fracking-company-violations-IP.pdf.
21. Stone, Les. Fracking's Environmental Impacts: Water. Greenpeace. https://www.greenpeace.org/usa/fighting-climate-chaos/issues/fracking/environmental-impacts-water/#:~:text=Contamination%20of%20Water%20Wells%20and,water%20for%20many%20rural%20communities.
22. Tabuchi, H., and Blacki M. 'Monster Fracks' Are Getting Far Bigger. And Far Thirstier. *The New York Times*. (Accessed September 25, 2023). https://www.nytimes.com/interactive/2023/09/25/climate/fracking-oil-gas-wells-water.html.
23. Office of Research and Development, EPA. EPA's study of hydraulic fracturing for oil and gas and its potential impact on drinking water resources. EPA - United States Environmental Protection

Agency. (Last modified December 2016). https://cfpub.epa.gov/ncea/hfstudy/recordisplay.cfm?deid=332990.

24. Cook, M., Huber, K. L., and Webber, M. E. (2015). Who Regulates It? Water Policy and Hydraulic Fracturing in Texas. *Texas Water Journal.* 6(1), 45-63. https://doi.org/10.21423/twj.v6i1.7021.
25. Charles, D. (2017). Fracking and Environmental Protection: An Analysis of U.S. State Policies. *The Extractive Industries and Society.* 4(1), 63-68. https://doi.org/10.1016/j.exis.2016.12.009.
26. DiGiulio, D. C., and Jackson, R. B. (2016). Impact to Underground Sources of Drinking Water and Domestic Wells from Production Well Stimulation and Completion Practices in the Pavillion, Wyoming, Field. *Environmental Science & Technology.* 50(8), 4524-36. https://doi.org/10.1021/acs.est.5b04970.
27. Fuchs, S., GIS in Water Resources, and Maidment D., *Hydraulic Fracturing in the Barnett Shale*. University of Texas at Austin. (Fall 2015). https://www.caee.utexas.edu/prof/maidment/giswr2015/TermProject/Fuchs.pdf.
28. Gassiat, C., Gleeson T., Lefebvre, R., and McKenzie, J. (2013). Hydraulic Fracturing in Faulted Sedimentary Basins: Numerical Simulation of Potential Contamination of Shallow Aquifers over Long Time Scales. *Water Resources Research.* 49(12), 8310-27. https://doi.org/10.1002/2013wr014287.
29. Heilweil, V. M., Grieve. P. L., Hynek, S. A., Brantley, S. L., Solomon, D. K., and Risser, D. W. (2015). Stream Measurements Locate Thermogenic Methane Fluxes in Groundwater Discharge in an Area of Shale-Gas Development. *Environmental Science & Technology.* 49(7), 4057-65. https://doi.org/10.1021/es503882b.
30. Hildenbrand, Z. L., Carlton D. D. Jr., Fontenot, B. E., Meik, J. M., Walton, J. L., Thacker J. B., Korlie, S., Shelor, C. P., Kadjo, A. F., Clark, A., Usenko. S., Hamilton, J. S., Mach, P. M., Verbeck, G. F., Hudak, P., and Schug, K. A. (2016). Temporal Variation in Groundwater Quality in the Permian Basin of Texas, a Region of Increasing Unconventional Oil and Gas Development. *Science of the Total Environment.* 562, 906-13. https://doi.org/10.1016/j.scitotenv.2016.04.144.
31. Hitaj, C., Boslett, A. J., and Weber, J. G. (2020). Fracking, Farming, and Water. *Energy Policy.* 146, 111799. https://doi.org/10.1016/j.enpol.2020.111799.
32. Jackson, R. B., Lowry, E. R., Pickle, A., Kang, M., DiGiulio, D., and Zhao, K. (2015). The Depths of Hydraulic Fracturing and Accompanying Water Use across the United States. *Environmental Science & Technology.* 49(15), 8969-76. https://doi.org/10.1021/acs.est.5b01228.
33. Kuwayama, Y., Olmstead, S., and Krupnick, A. (2015). Water Quality and Quantity Impacts of Hydraulic Fracturing. *Current Sustainable/Renewable Energy Reports*. 2(1) 17-24. https://doi.org/10.1007/s40518-014-0023-4.
34. Ong B. (2014). The potential impacts of hydraulic fracturing on agriculture. *European Journal of Sustainable Development*. 3(3), 63-72. https://doi.org/10.14207/ejsd.2014.v3n3p63.
35. Renock, D., Landis, J. D., and Sharma, M. (2016). Reductive Weathering of Black Shale and Release of Barium during Hydraulic Fracturing. *Applied Geochemistry*. 65, 73-86. https://doi.org/10.1016/j.apgeochem.2015.11.001.
36. Rodriguez, J., Heo, J., and Kim, K. H. (2020). The Impact of Hydraulic Fracturing on Groundwater Quality in the Permian Basin, West Texas, USA. *Water*, 12(3), 796. https://doi.org/10.3390/w12030796.
37. Rogers, J. D., Burke, T. L., Osborn, S. G., and Ryan, J. N. (2015). A Framework for Identifying Organic Compounds of Concern in Hydraulic Fracturing Fluids Based on Their Mobility and Persistence in Groundwater. *Environmental Science & Technology Letters*. 2(6), 158-64. https://doi.org/10.1021/acs.estlett.5b00090.
38. Soriano, M. A., Deziel, N. C., and Saiers, J. E. (2022). Regional Scale Assessment of Shallow

Groundwater Vulnerability to Contamination from Unconventional Hydrocarbon Extraction." *Environmental Science & Technology*. 56(17), 12126-36. https://doi.org/10.1021/acs.est.2c00470.

39. STANFORD UNIVERSITY. Does living near an oil or natural gas well affect your drinking water? *EurekAlert!* (Last modified February 14, 2016). https://www.eurekalert.org/news-releases/803327.
40. Zhai, G., Shirzaei M., and Manga M. (2021). Widespread Deep Seismicity in the Delaware Basin, Texas, Is Mainly Driven by Shallow Wastewater Injection. *Proceedings of the National Academy of Sciences*. 118(20), https://doi.org/10.1073/pnas.2102338118.