NOVEL METHODS FOR GROUNDWATER MONITORING WITHIN REGIONS OF OIL AND GAS PRODUCTION

A University Thesis Presented to the Faculty

of

California State University, East Bay

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Geology

By

Andrew Michael Renshaw

September 2015

ABSTRACT

As California enters its fourth year of drought, protecting shallow groundwater quality for municipal, domestic and agricultural use becomes critically important. Many of California's sedimentary basins host both potable groundwater aquifers and hydrocarbon-producing zones, and Senate Bill 4 (Pavely, 2013) requires groundwater monitoring within hydrocarbon zones that potentially impact shallow aquifers by certain types of oil production. One approach to understanding the relationship between deeply seated fluids associated with oil production and shallow drinking water aquifers is to establish current ambient groundwater quality conditions using available data (both current and historic), then develop a predictive mixing model between ambient groundwater and "produced" waters associated with oil production. In addition, in order to fill datagaps useful for the model, novel methods of analyzing certain chemical analytes in produced waters must be developed. Produced waters typically contain complex hydrogeochemical matrices (e.g., high salinity, TDS, etc.) but also contain effective isotopic tracers such as isotopes of radium. Utilizing produced water samples from three large oil fields in California, a novel method to analyze radium-226 in produced waters was developed. The analytical method utilizes liquid scintillation counting (LSC) to provide accurate and efficient results with a two-week turn around time. Radium analysis in produced water samples will further refine the mixing model and therefore groundwater monitoring programs in oil producing regions.

Data from the Kern County groundwater basin demonstrates the novel groundwater monitoring methods. The San Joaquin Valley's Kern County groundwater

basin contains the majority of California's oil production (~80% of the active wells in California). In addition, the county relies almost entirely on groundwater for its water supply, and is the most productive agricultural county in the US. Considering Kern County's historically stressed groundwater system, the Department of Water Resources ranks the basin as a high priority in the statewide groundwater elevation monitoring (CASGEM) program.

This research consists of a database investigation to define the current water quality conditions within Kern County's groundwater system and to examine the relationship between shallow groundwater and produced waters. An end-member mixing model utilizing multivariate statistics compares the geochemical data between the two distinct waters. Two publically available data sets are used to define the end-members: the California State Water Resources Control Board's Groundwater Ambient Monitoring and Assessment program data are used to define the shallow zone, and the USGS Produced Waters data are used to define the deep formation waters. The multivariate mixing model indicates that six common groundwater chemical analytes (Ca, Cl, Mg, Na, SO_4 and TDS) distinctly segregate shallow groundwater from produced waters. The multivariate model will aid groundwater monitoring programs by providing a statistical test of whether deeply seated basin fluids are mixing with shallow groundwater.

NOVEL METHODS FOR GROUNDWATER MONITORING

WITHIN REGIONS OF OIL AND GAS PRODUCTION

By

Andrew Michael Renshaw

Approved:

Dr. Jean Moran — Date: 27 Aug 2015

Dr. Luther Strayer — Date: 27 August 2015

Dr. Bradley Esser — Date: 27 August 2015

ACKNOWLEDGMENTS

This work would not have been possible without Dr. Jean Moran's tenacity to inspire, teach and pursue new scientific endeavors. Many of the ideas and thoughts throughout this work also would not have germinated without the Cal State East Bay Earth Sciences faculty, especially Dr. Luther Strayer and Dr. Mitchell Craig. Dr. Strayer's passion for structural geology re-illuminated the vastness of geology and helped me develop a more imaginative 3-dimentional way of thinking. In addition, Dr. Bradley Esser, Richard Bibby, Dr. Ate Visser and Lawrence Livermore National Laboratory as a whole provided me the opportunity to intellectually explore and grow, for which I am immensely grateful. Lastly, thanks to my wife, Lindsey, for putting up with all the jargon, late nights and pretending to care about radium isotopes in oil field produced waters (although I think she has a new found appreciation for California's hydrogeology). Thanks to you all.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SCRIPTS

PREFACE

My research started as an attempt at understanding the distribution, activities and analytical capabilities of radium isotopes in shallow groundwater and oil field produced waters in California. However, because of a lack of radium data and lack of co-located GAMA and USGS produced waters data points, this thesis morphed into research that used newly developed skills, R Statistical Software and multivariate statistics. Two interrelated pieces of research are discussed to evaluate the following questions:

1. Can qualitative and quantitative spatial relationships be generated using currently available groundwater and oil field data?
2. More importantly, can a multivariate mixing model utilizing produced waters data and ambient groundwater data in the vicinity of oil and gas development identify useful geochemical and isotopic tracers? Furthermore, can the model 'predict' whether a shallow groundwater sample shows a statistical relationship with the produced waters hydrogeochemical signature? If so, the waters may be currently interacting or have interacted with one another in the past. If there is evidence of mixing, how do the spatial and physical (faults, geologic structures, stratigraphy, etc.) relationships correlate with the potentially mixed waters?
3. Lastly, high radium activities have been shown to exist in hydraulic fracturing wastewaters in other parts of the United States. Even further, isotopes of radium have been successfully used as isotopic tracers in mixing models. However, radium analysis in wastewaters has proven challenging because of

the high dissolved solids matrix. Can an effective and efficient analytical method using liquid scintillation counting be developed? What are the implications of the method on regional scale groundwater monitoring and the mixing model?

CHAPTER I: Introduction and Background

INTRODUCTION

California's water supply heavily relies on groundwater. In 2003 the Department of Water Resources (DWR) updated the California State Groundwater Bulletin 118 stating groundwater use for normal and dry years was approximately 30% and 40%, respectively (DWR, 2003). Twelve years later, DWR estimates that during normal precipitation years shallow groundwater represents approximately 38% of the state's water use. During years of drought the percentage increases to 46% or greater. Although the statewide groundwater use represents less than half of the state's total water supply, some communities rely entirely or almost entirely on groundwater (DWR, 2015). For example, the City of Bakersfield in Kern County (9th largest California city by population) utilizes groundwater for at least 65% of their municipal water supply. Furthermore, Kern County, a large agricultural region, produced 75,637 acre-feet of groundwater in 2014, and agriculture utilized only 0.22% of the total volume, further indicating a strong urban reliance on groundwater (KCWA, 2014). With that, protecting groundwater resources is important.

Both naturally occurring and anthropogenic contaminants degrade California's groundwater quality. Approximately 1,018 community water supply systems contain chemical constituents that exceed the public drinking water standard (SWRCB, 2013). Of the top ten principal contaminants, four of them are almost exclusively naturally occurring: arsenic, gross alpha activity, uranium and fluoride (listed in the order of

frequency of exceedance). Despite arsenic being the most common naturally-sourced contaminant, uranium and the constituents that comprise gross alpha activity (uranium, thorium, radium and radon) radioactively decay creating progeny products, potentially further degrading groundwater quality over time. In addition, remediation of naturally occurring chemical contaminants, including radioactive materials (NORMs), is difficult because they are a product of water-rock interaction.

Groundwater quality degradation generally derives from surface or shallow subsurface activity; however, relatively young shallow drinking water aquifers may be negatively affected by deeply seated geologically confined formation waters or brines (collectively, formation waters). In general, the naturally occurring formation fluids evolve over geologic time in sedimentary basins allowing for characteristic geochemical composition as the water interacts with the surrounding lithology (Kyser, 2007). In turn, the formation fluids, such as entrapped seawater and paleogroundwater, contain high concentrations of salts, trace elements and NORMs.

NORM activities in formation fluids largely exceed ambient activities in shallow drinking water aquifers (Kondash et al., 2014; Rowan et al., 2011; Vengosh et al., 2013; Warner et al., 2013a; Warner et al., 2013b). Of particular concern, radium-226 (^{226}Ra) activities are well above the EPA drinking water Maximum Contaminant Limit (MCL; 5 pCi/L) and the Industrial Effluent Discharge Limit (60 pCi/L). However, accurate quantification of ^{226}Ra within the formation waters proves difficult due to the complex matrices of the formation waters, which contain high concentrations of dissolved solids and potentially organic material. Sedimentary basins contain the largest oil and gas

reservoirs on the planet. As a result, the formation fluids in regions containing accumulations of oil and natural gas also have a hydrocarbon signature further complicating radium analysis. Considering the stringent standards associated with utilizing groundwater for human consumption and the large contrast in geochemistry of formation fluids and shallow aquifers, developing new methods of monitoring groundwater quality will be useful.

Generating an end-member mixing model utilizing multivariate statistics is one way of analyzing the potential mixing of the two geochemically distinct fluids. In the case of formation waters mixing with shallow aquifers the model utilizes produced water or oil field wastewater data, representing the deeply seated formation waters, and shallow drinking water aquifer data as the two end-members. The end-members are statistically derived from publically available data from the United States Geological Survey (USGS) and the California State Water Resources Control Board (SWRCB). Consequently, the model described in this thesis compares common chemical data providing an avenue to predict if a shallow groundwater sample contains a statistical relationship with the produced water data. The predictive capabilities of the model provides insight into how to identify a formation fluids geochemical signature within shallow drinking water aquifers in relation to conventional and unconventional energy extraction in California.

The model provides a geochemical framework for examining potential mixing of the two end-members. In addition, a new analytical method for ^{226}Ra by liquid scintillation counting (LSC) was developed in order to eliminate the sample matrix complications and reduce the amount of time required for each measurement. The new

monitoring methods and tools discussed herein will benefit projects of varying size that require analysis of waters with a wide range of dissolved solid content. The goal of the mixing model and the ^{226}Ra analytical method is to aid regional shallow aquifer groundwater monitoring and oil field operational wastewater characterization.

BACKGROUND

California contains a wealth of natural energy resources. The state contains 125,605 active oil-producing wells, 98.6% of which are conventional extraction wells. The remaining 1.4% (1,752 wells) are hydraulically fractured wells. In September 2013 the California State Government approved Senate Bill 4 (SB4), which defines the regulatory framework of oil and gas well stimulation treatment in California. Since the approval of SB4 nearly 1,500 wells utilizing enhanced well stimulation treatments (WST) have been approved. Figure 1 shows the historical record of hydraulic fracturing and well stimulation treatment in California.

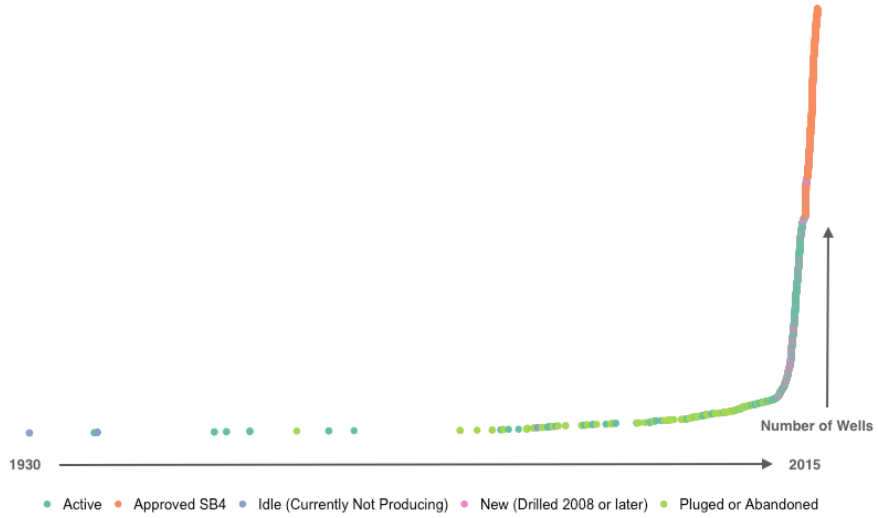


Figure 1: History of hydraulic fracturing in California. History of Oil and Gas Wells Subjected to Hydraulic Fracturing in California.

Along with increased WST activity SB4 has also increased awareness in regards to protecting water resources, including shallow drinking water aquifer quality. With that, SB4 requires groundwater monitoring in regions utilizing hydraulic fracturing and other WST; however, this type and scale of monitoring has not been extensively practiced. Therefore, developing new scientific methods to understand what and how to monitor becomes important. One way to aid the development of monitoring programs is utilizing available data to analyze the current hydrogeologic and geologic systems in regions undergoing conventional and unconventional energy development. Furthermore, the research discussed herein helps to bridge the gap between what currently is known about California's oil fields and shallow aquifer systems in hopes of aiding the development of effective and efficient monitoring programs.

California's Senate Bill 4 (SB4) - Oil and Gas: well stimulation requires all owners or operators (producers) of stimulated oil producing wells to undergo a permitting process, in which the permit is filed and approved by the overseeing agency, the Division of Gas, Geothermal and Oil Resources (DOGGR). Well Stimulation Treatment Notices, the internal title of the filed document, requires the producer to document the location of the well, the type of stimulation (stimulation refers to the type of enhanced oil extraction, i.e. hydraulic fracturing, acid matrix, steam flooding etc.), the targeted formation with approximate depths and orientation of the well, among others. On top of descriptive information about the construction of the stimulated well, SB4 requires an analysis of the regional geologic and hydrogeologic setting (Pavely, 2013). Within the hydrogeologic analysis the producer must establish a groundwater monitoring program in the vicinity of the proposed well. Additionally, SB4 requires the California State Water Resources Control Board to establish regional, basin scale groundwater monitoring programs to temporally observe groundwater conditions in relation to oil and gas production.

As of June 3, 2015 the DOGGR Well Stimulation Treatment (WST) Notice Index contains 1,634 approved WST Notices (the WST Notice Index is a DOGGR open source database containing all the notices in a viewable format). Since December 2013 the WST index has exhibited a sporadic influx of notices. The maximum number of requests were filed in the months of December 2013 and September 2014, at 169 and 171 notices, respectively. Figure 2 summarizes the number of notices received each month since the inception of SB4 and the WST Notice Index.

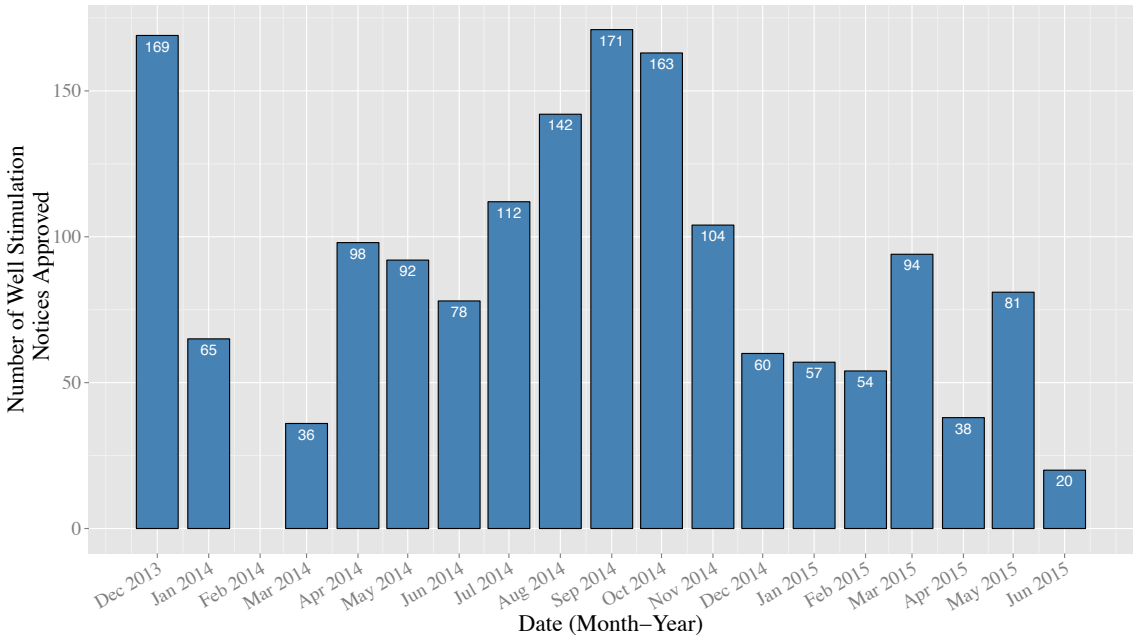


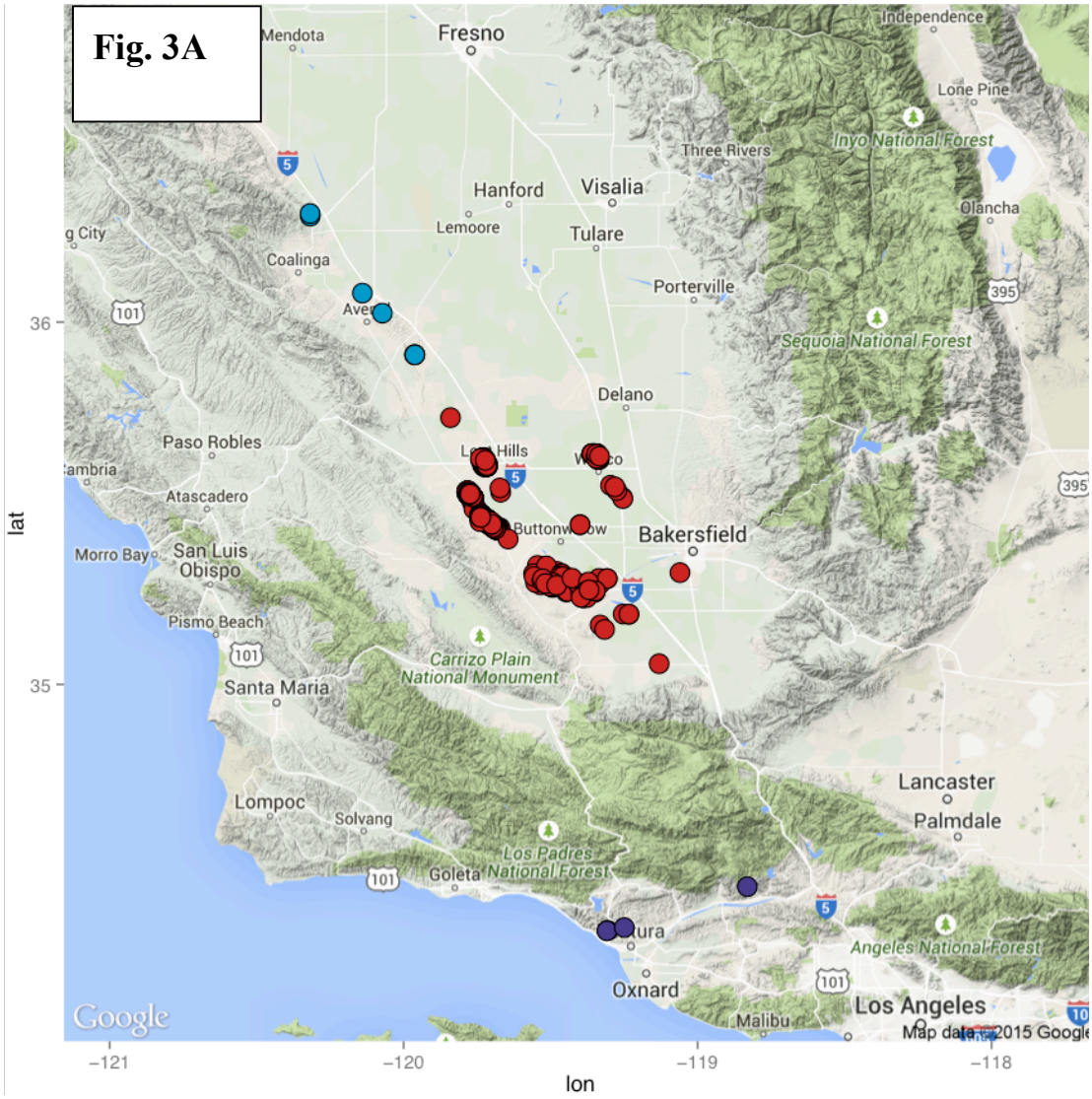
Figure 2: Number of WST Notices approved by DOGGR. Number of WST Notices approved by DOGGR by month since inception on December 2013 (as of June 2015). The California State Legislation signed Senate Bill 4 - Oil and Gas: well stimulation into law on September 2013. The bill requires all producers anticipating utilizing enhanced oil extraction techniques in California's oil fields to receive approval from DOGGR through the submittal of a comprehensive informational and technical package about the requested drilling activities termed WST Notices (DOGGR, 2015).

Nearly all of the WST notices (97%) request drilling activities to take place in Kern County, California (Table 1; Figure 3A). Further, approximately 82% of the notices are requests to stimulate wells within the northern and southern Belridge Oil Fields (collectively Belridge), located in western Kern County (Figure 3B & Figure 3C). Hydraulic fracturing dominates the stimulation type at 96% of all approved notices. The remaining 4% of the notices plan to utilize acid matrix type stimulation techniques.

Although the current regulatory framework highlights the quantity and distribution of well stimulation, hydraulic fracturing and other types of enhanced oil recovery are not new to California. Approximately 944 wells were hydraulically fractured prior to December 2013. In addition, only 3.5% of the 944 hydraulically fractured wells (33 wells) were advanced prior to January 2001. All of the pre-December 2013 hydraulically fractured wells were located in Kern County.

Table 1: Number of WST Notices by County as of June 3, 2015

County	Number of Approved Notices	Percent of Total
Kern	1,623	99%
Kings	4	< 1%
Ventura	4	< 1%
Fresno	3	< 1%



RED = Kern County BLUE = Fresno/Kings County PURPLE = Ventura County

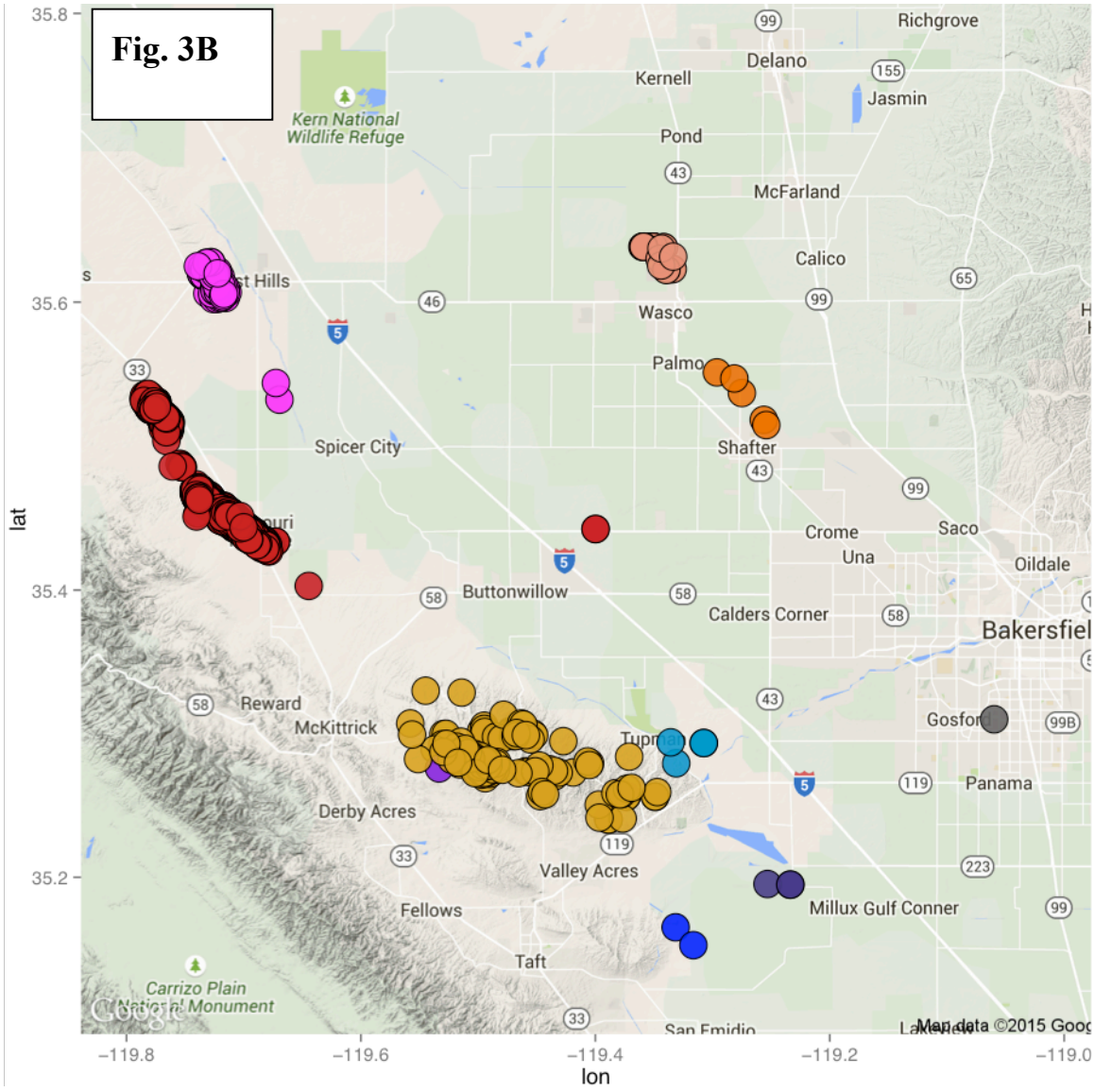
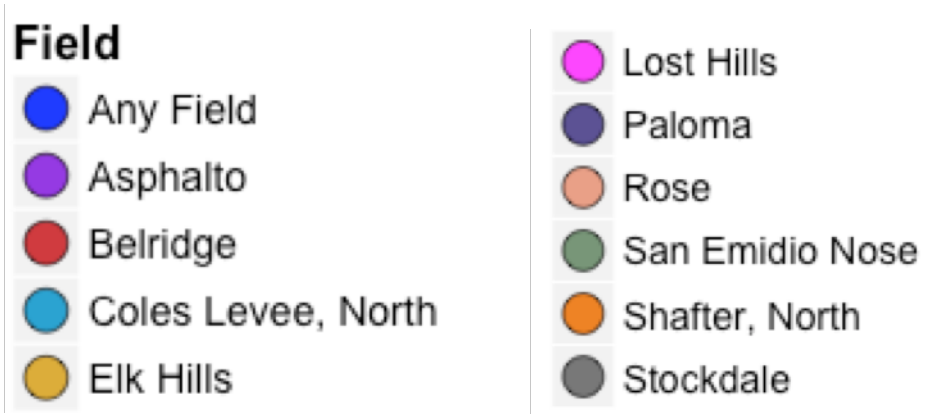


Fig. 3B
Kern National Wildlife Refuge
Richgrove
Delano
Kernell
Jasmin
Pond
McFarland
Calico
Wasco
Palmo
Shafter
Spicer City
Crome
Saco
Una
Oildale
Buttonwillow
Calders Corner
Bakersfiel
Reward
McKittrick
Gosfor
Panama
Derby Acres
Valley Acres
Millux Gulf Conner
Fellows
Taft
Carrizo Plain National Monument
San Emidio
Map data ©2015 Goog
lat
lon
35.8
35.6
35.4
35.2
−119.8
−119.6
−119.4
−119.2
−119.0


Field
Any Field
Asphalto
Belridge
Coles Levee, North
Elk Hills
Lost Hills
Paloma
Rose
San Emidio Nose
Shafter, North
Stockdale

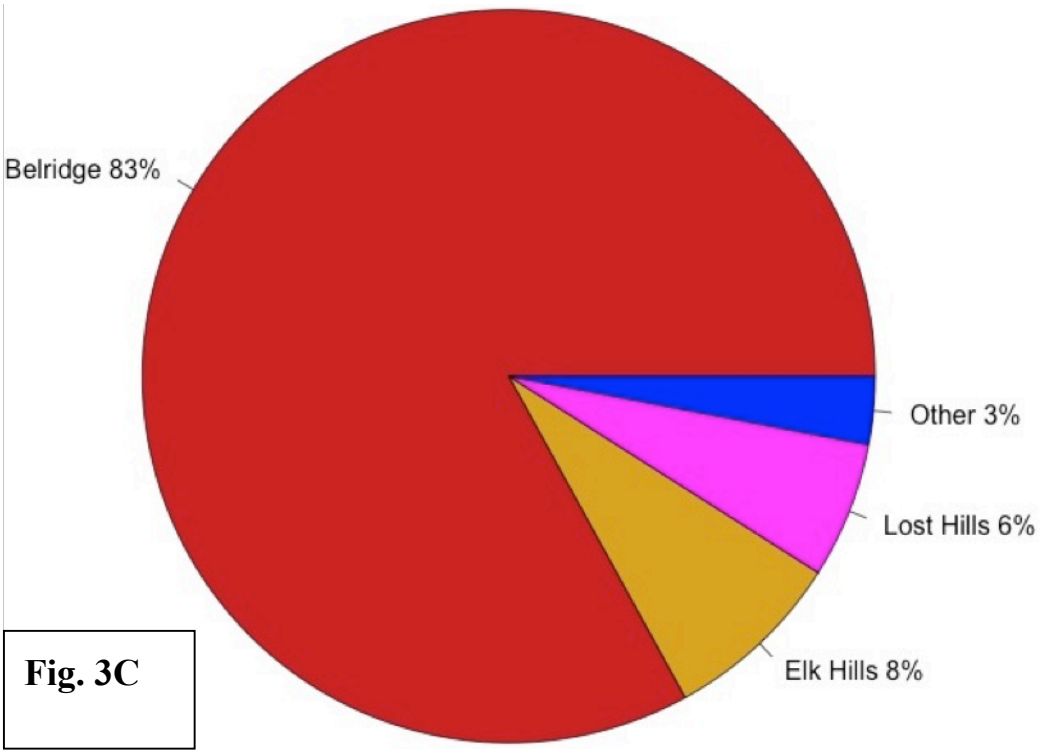


Figure 3A, B & C: WST Notices in California. A) Location of approved WST Notices in California. In other words, locations of a future unconventional oil extraction well. Light blue = Fresno and Kings Counties; Red = Kern County; and, Purple = Ventura County. B) Location of all approved WST Notifications in Kern County distinguished by oil field. C) 97% of the WST Notices are located in Kern County. Of that, 83% of the Kern County wells are located in the north and south Belridge Oil Field. The Lost Hills and Elk Hills oil fields contain a combined 14% of all the wells in Kern County. The distinguishing colors in 3C correspond with the distinguishing colors in 3B.

Well Stimulation Treatment: Hydraulic Fracturing

Hydrofractures occur naturally in rocks with high fluid pressure. Nearly all rocks within the upper crust contain fluids, and the fluids apply pressure to cracks or pores counteracting the compressional lithostatic stress. Classic examples of geologic structures resulting from hydrofracturing include dikes, sills and mineral filled veins. In general, the strength and ductility of a material rises with a rise in mean stress; however, pore fluid pressure reduces the mean stress, therefore weakening rocks (Suppe, 1985).

In order to manipulate the stress regime of a formation, hydraulic fracturing enhanced oil recovery requires the injection of large volumes of pressurized fluid into the source rock to reduce the mean stress of the system. The injection process generally consists of four stages (CCST, 2014b):

Stage 1: Inject fracturing fluid without a proppant (solid material) at a pressure greater than the pressure exerted on the target formation. The fractures extend perpendicularly outward from the well (Figure 4).

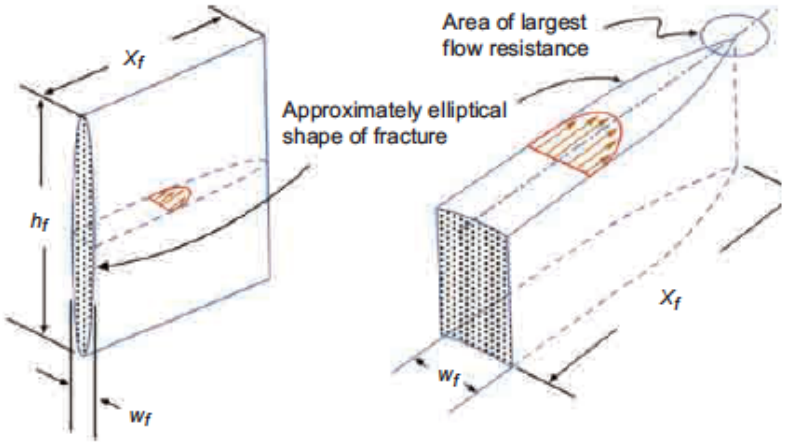


Figure 4: Generalized fracture geometry resulting from hydraulic fracturing.

Generalized fracture geometry resulting from hydraulic fracturing. Injection of pressurized fluid exceeding the lithostatic pressure reduces the normal stress (compressional stress) on the formation causing mode 1 dilatant fractures. The fractures propagate perpendicularly outward from the well: fractures induced from horizontal well on the left and vertical well on the right (Montgomery and Smith, 2010).

Stage 2: Addition of proppant to the injection fluid. Proppants (solid material) typically consist of sand; however, materials such as steel shot, glass beads, plastic pellets and rounded nut shells have been used (Montgomery and Smith, 2010). The proppants fill the hydraulic fractures reducing the chance of fracture collapse once the fluid pressure dissipates. In turn, the proppants effectively increase the permeability of the formation.

Stage 3: Following the addition of proppants, injection of fluids without proppants allows for the residual proppant in the well and formation to flush further into the fractures.

Stage 4: The hydraulic fracture fluid is extracted from the well and formation as 'flowback water,' reducing the fluid pressure in the formation. The flowback water becomes waste material or is recycled for additional well stimulations. Once production of the well begins, naturally occurring fluids in the formation (typically ancient seawater brines) mixes with the hydrocarbons. The oil and formation brine mixture reaches the surface as produced waters. The produced waters continue to flow out of the well for several days or weeks, which is typically stored in onsite waste tanks, holding ponds or disposed of through percolation ponds. The geochemistry of the produced waters varies from region to region but in general largely contrasts the geochemical signature of shallow drinking water aquifers.

To summarize, the injection fluid consists primarily of water (~99%), minor amounts of chemical additives and proppants. The chemical additives reduce mineral scaling on the inside of the wells, act as surfactants stripping oil from the host rock and as acids further dilating fractures. Once the fractures propagate, injection of proppants allows the fractures to remain open, effectively increasing the formation permeability. Next, the system is flushed with proppant-free hydraulic fracturing fluids. Lastly, flowback and produced waters along with the targeted oil return to the surface and undergo treatment, disposal or reuse. The chemical make up of the produced waters is discussed below.

Chemistry of Produced Waters

Produced waters and flowback water typically contain high concentrations of inorganic constituents, creating highly saline brines (Haluszczak et al., 2013). The deeply seated formation brines interact with the injected well stimulation fluids, which eventually reach the surface as hydraulic fracturing wastewater. For illustration, Dresel and Rose (2010), describe the brines associated with the Marcellus Shale as containing total dissolved solids (TDS) concentrations of 35,000 mg/L (roughly equal to seawater salinity) or more. The US EPA's secondary maximum contaminant level (SMCL) for TDS is 500 mg/L. In addition, the primary source of the brines associated with the Marcellus Shale is attributed to the evaporation of trapped seawater to the stage of halite precipitation (Dresel and Rose, 2010) (the interpretation of the origin of Kern County's formations waters is discussed in the Hydrogeochemistry section below). The high salinity of hydraulic fracturing wastewaters poses a problem for treatment techniques and disposal process that do not impair human health or the environment (Gordalla et al., 2013; Nelson et al., 2014; Olsson et al., 2013). Also as previously stated and discussed in more detail in Chapter V, the salinity complicates certain chemical analyses, like ^{226}Ra. Figure 5 shows the relationship between total dissolved solids concentrations in 316 California produced water and 316 randomly selected shallow aquifer samples.

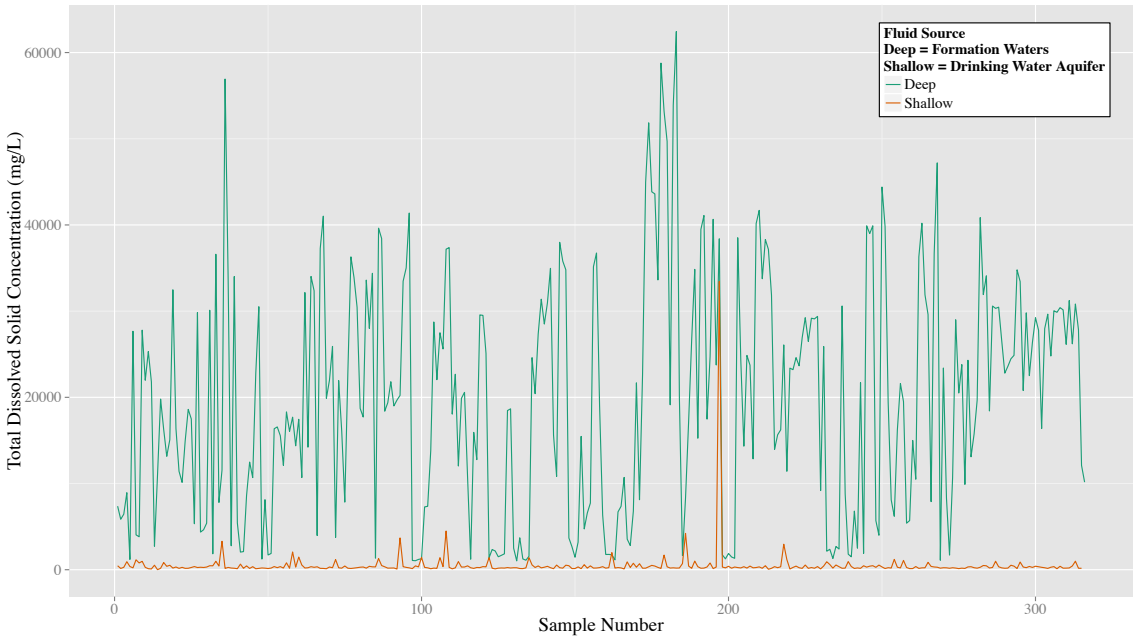


Figure 5: Concentrations of TDS (mg/L) from produced and shallow water. Comparison of randomly selected and organized concentrations of TDS (mg/L) from the 316 USGS produced water samples available for Kern County's oil fields and GAMA shallow drinking water aquifer samples in Kern County (CSWRCB, 2015; USGS, 2015). The produced waters contain a much higher concentration along with more variation compared to the lower concentration of TDS in shallow aquifer water.

Hydraulic Fracturing in California

Hydraulic fracturing for enhanced oil recovery dates back to the late 1940s. The first hydraulically fractured wells in Texas utilized conventional oil and gas wells to effectively discharge oil and gas locked up in low permeability shale layers at depths of up to thousands of feet. The conventional use of the technique, however, did not provide a large lateral dispersion of fractures and joints. Although hydraulic fracturing in

conventional wells provides economically adequate resources, the oil and gas extraction method did not become the driving force in the energy resource industry until the 1990s when horizontal drilling techniques became prominent (Mooney, 2011). In addition, despite the lack of detailed record keeping within the oil and gas industry in regards to hydraulic fracturing, a noticeable increase in development and use of the technique began in the early 2000s resulting in the modern day "gas boom." Recent shale oil extraction largely differs from conventional methods, though. Modern methods extensively utilize horizontal drilling, multi-stage fractures and chemically augmented fracturing fluids to extract oil from low permeability organic rich shale layers at depth (Cherry et al., 2014; Nash, 2010).

Hydraulic fracturing has proven an economically viable method of extracting oil and gas from otherwise inaccessible geologic formations, such as tight oil shale. As noted above, hydraulic fracturing has been extensively utilized in oil fields throughout the United States. California, however, has not seen the same precipitous use of hydraulic fracturing although being one of the country's largest oil producing states (Figure 6). The lack of a hydraulic fracturing boom in California may be attributed to the State's successful oil extraction from conventional methods and other types of WST such as steam flooding. In addition, California's oil fields uniquely differ from other prominent oil fields such as the Barnett Shale in Texas or the Bakken Formation in North Dakota.

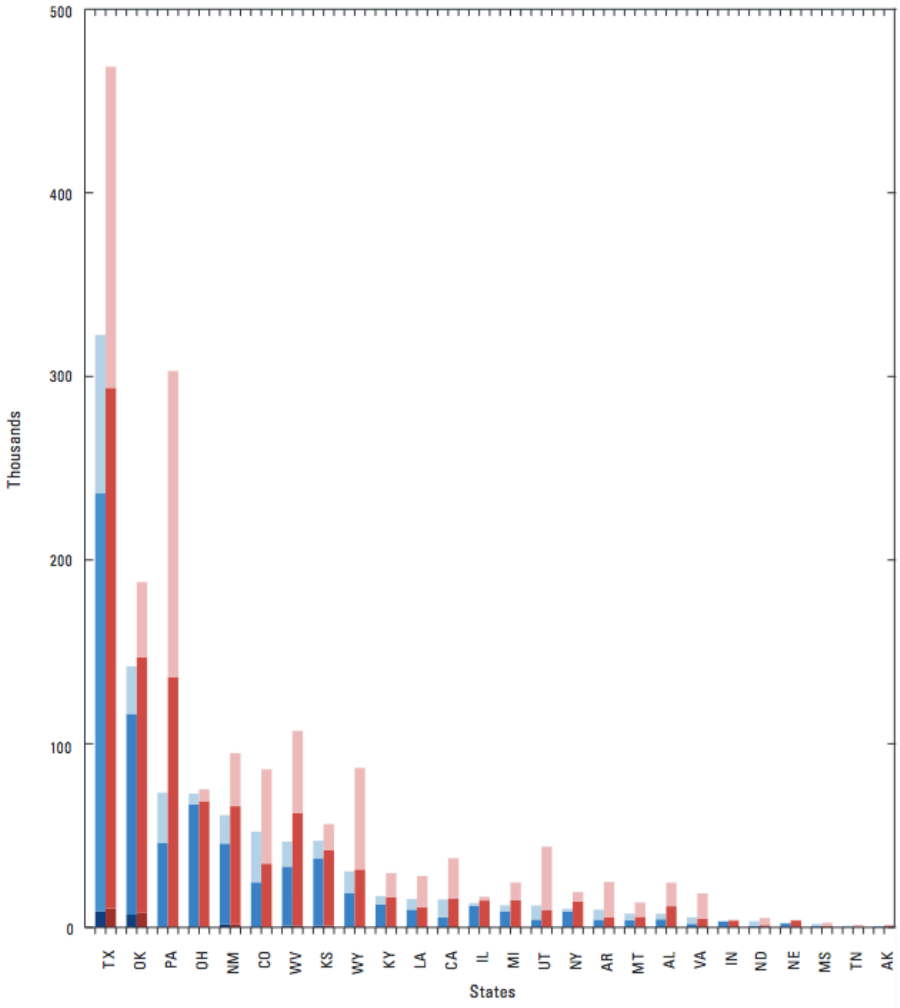


Figure 6: Number of hydraulically fractured wells. Number of hydraulically fractured wells (blue) and treatments (red) between 1947 and 2010 in the United States (Gallegos and Varela, 2015).

Hydraulic fracturing technologies have been utilized in California for several decades; however, hydraulic fracturing in California has significant differences from other active hydraulic fracturing regions, such as Pennsylvania, North Dakota or Texas. In California, hydraulic fracturing consists of shallower vertical wells, less water use,

more chemically diverse hydraulic fracturing fluids and substantially different geology (CCST, 2014a). For example, the Bakken Formation and the Barnett Shale contain long, layered stratigraphic units allowing extensive resource extraction, both vertically and horizontally. In contrast, California's Monterey Formation contains more structural features (such as anticlines where the oil accumulations get trapped), has variable composition and thermal maturity, and offset by numerous faults (Figures 7).

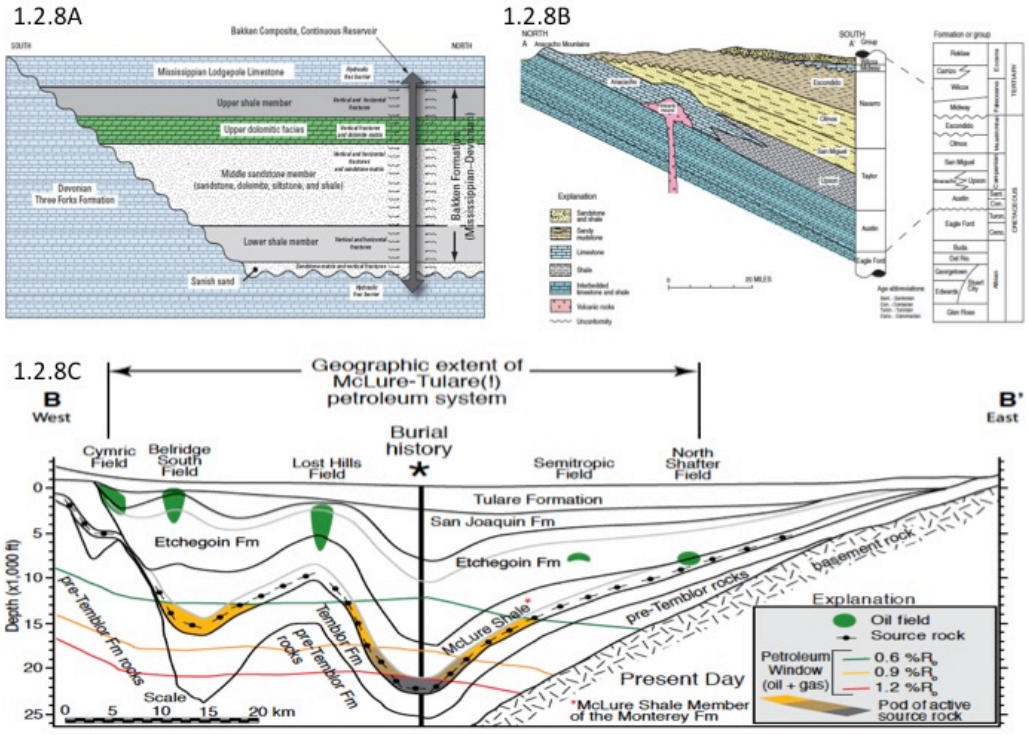


Figure 7: Comparison of three generalized cross-sections of shale oil plays.

Comparison of three generalized cross-sections of shale oil plays from North Dakota, Texas and California. A) Generalized schematic cross-section of the Devonian-Mississippian Bakken Formation in North Dakota. The Bakken formation resides at depths of up to 10,000 feet in parts of North Dakota. B) Cretaceous Maverick Basin,

Eagle Ford/Austin Group, Texas. The Eagle Ford is the second largest shale oil play in the state. C) Geologic cross-section of the McLure-Tulare system in Kern County, CA (Condon and Dyman, 2006; Magoon et al., 2007; Pollastro et al., 2010).

In addition, California is characterized by much younger and more active geology. For example, the Kern County oil fields (Belridge, Elk Hills etc.) are situated near the intersection of the San Andreas fault, White Wolf fault and the Garlock fault, typically referred to as the "Big Bend" (Figure 8).

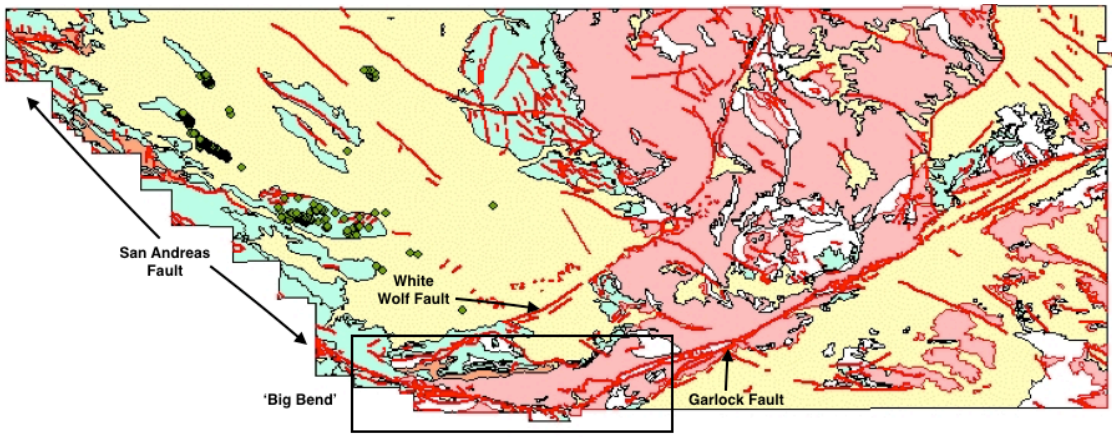


Figure 8: Geologic map of Kern County. Geologic map of Kern County with fault traces (red lines). Geologic units consist of: Holocene surficial sediments (yellow), Pleistocene Tulare Formation (blue) and Mesozoic Sierra Nevada Granitics (red). Three major faults are labeled: San Andreas, White Wolf and Garlock. The DOGGR approved well stimulation treatment wells are depicted as green dots. The San Andreas Fault runs along the county's western border but is not shown due to the stair-stepped shape of Kern County's western boarder.

CHAPTER II: Geology and Hydrogeochemistry of the Site Area

SITE AREA

Kern County, California

Kern County, the 3rd largest county in California by land area, occupies the southern most portion of the San Joaquin Valley (Figure 9). The county encompasses a wide range of environments bounded by the San Andreas Fault and the Temblor Range to the west and extending beyond the southern slope of the Sierra Nevada into the Mojave Desert and the Antelope and Indian Wells Valley to the south and east. Kern County has a population of 874,589, in which 42% reside in the City of Bakersfield, the county seat. Due to Kern County's economy largely relying on natural resources in the Central Valley, a majority of the population resides in the San Joaquin Valley portion of the county.

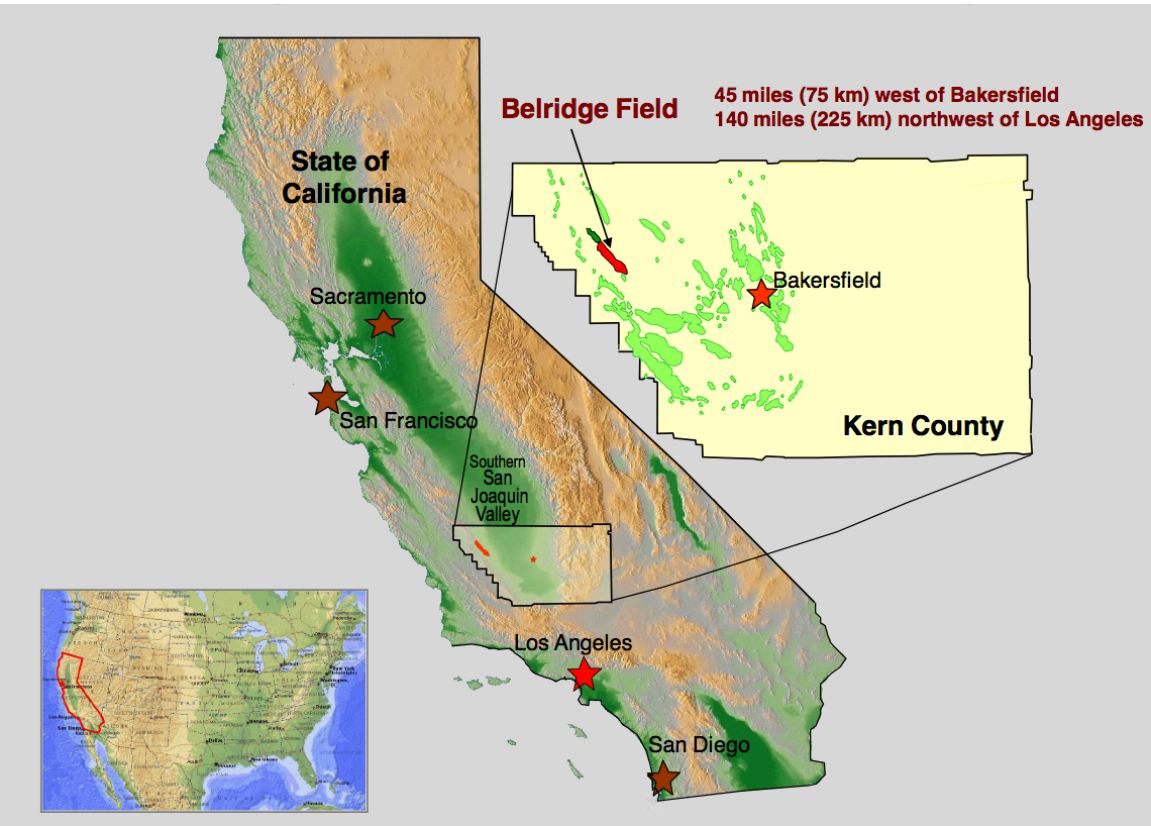


Figure 9: Vicinity map of Kern County, California. Vicinity map of Kern County, California including the major oil fields in green (Smith, 2012). The Belridge Field, shown in red, is the most productive oil field in California.

As of January 2015, natural resource-dependent activities represent nearly 25% of the county's economy with 15% from agriculture and the remaining 10% from mining and oil production. In 2013, Kern County surpassed Fresno County as the most prolific agricultural county in the United States, producing approximately $6.8 billion in agricultural commodities (Arroyo, 2014). In addition, in 2013 Kern County produced the 3rd largest volume of oil (1st largest for inland oil recovery) in the country behind Alaska's Beechey Point Quadrangle and the Gulf of Mexico's Green Canyon Field (Thuot, 2014).

GEOLOGIC SETTING

The geology of Kern County is complex. The county encompasses six of California's eleven geomorphic provinces including three of the six primary mountain ranges (Table 2). The San Joaquin Valley (SJV) is the leading geologic feature in the county. The SJV, the southern most extent of the 700-km-long Great Valley Province, is an asymmetric forearc basin containing upper Mesozoic and Cenozoic sediments up to 9 km thick overlying South Sierran Block crystalline basement (Bartow, 1991). To the west of the SJV the Coast Ranges dominate the geologic setting. To the south the Transverse Ranges, California's only east-west trending range, segregates the SJV portion of the county from the Mojave Desert, Indian Wells Valley and Antelope Valley region (Mojave Desert Province). Lastly, the Sierra Nevada in eastern Kern County separates the SJV from the Basin and Range Province in the east. In addition to three mountain ranges bounding the SJV, three major faults transect the basin and county. Figure 10 depicts a generalized geologic map of the SJV and the three bounding mountain ranges.

Table 2: California Geomorphic Provinces and Mountains in Kern County

Geomorphic Provinces	Mountain Ranges
Coast Ranges	Coast Ranges: Diablo Range & Temblor Range
Transverse Ranges	
Great Valley	Transverse Ranges: Tehachapi Mountains
Sierra Nevada	
Basin and Range	Sierra Nevada: Southern Extent
Mojave Desert	

As alluded to above (see Figure 8), the San Andreas, Garlock and White Wolf faults generally coalesce into the "Big Bend" of the San Andreas fault in Kern County (Figure 8, Figure 11 and Figure 12). The Garlock fault, an uncharacteristically left-lateral fault, distinctly defines the northern boundary of the Mojave Block. The White Wolf fault (WWF), like the Garlock fault, is an east-west trending dextral fault that transects the SJV just north of the Tehachapi Mountains of the Transverse Ranges. The WWF conjoins with the Breckenridge fault and the Kern Canyon fault in the Sierra Nevada and the Pleito thrust near Wheeler Ridge in the west. The WWF ruptured ($M_w = 7.7$) in the 1952 Kern County earthquake near Wheeler Ridge (a star represents the location on Figure 12). The San Andreas fault transects Kern County's southwestern corner at the "Big Bend." The fault enters near Lebec, California and exits the county just south of Taft. The SAF abuts the counties western boarder just west of the Diablo and Temblor Ranges. In addition to the three main faults discussed herein Kern County contains many faults active throughout the Quaternary with many active in the Holocene or within the last 150 years.

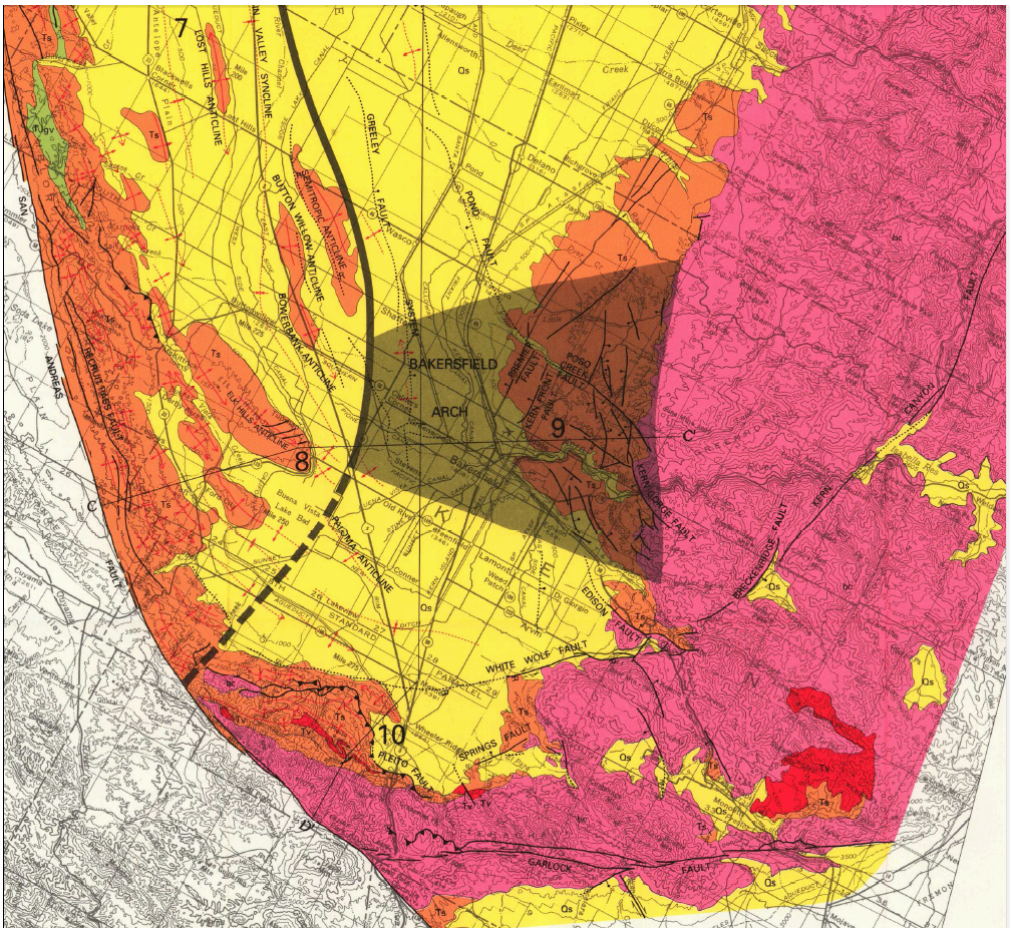


Figure 10: Generalized geologic map of the San Joaquin Valley. Generalized geologic map of the San Joaquin Valley. The lithologic units are as follows: Pink = crystalline basement of the South Sierran Block (Mesozoic and Paleozoic); Green = Tertiary-Jurassic Great Valley Sequence sedimentary rocks; Red = Neogene volcanics; Orange = Tertiary marine and non-marine sedimentary rocks; Yellow = Quaternary alluvial and lacustrine sediments. Black lines indicate the San Andreas, White Wolf, Garlock and Pleito faults, among others. The black shaded area depicts the Bakersfield Arch, westward plunging structural bowing or fold. The Transverse Ranges associated

with the Garlock fault consist of granitics much like the Sierra Nevada whereas the Coast Ranges abutting the San Andreas Fault contains a primarily marine sedimentary lithology. The differing lithologies contribute to a unique depositional environment in the SJV. The Quaternary sediments of the SJV basin are up to 9 km deep.

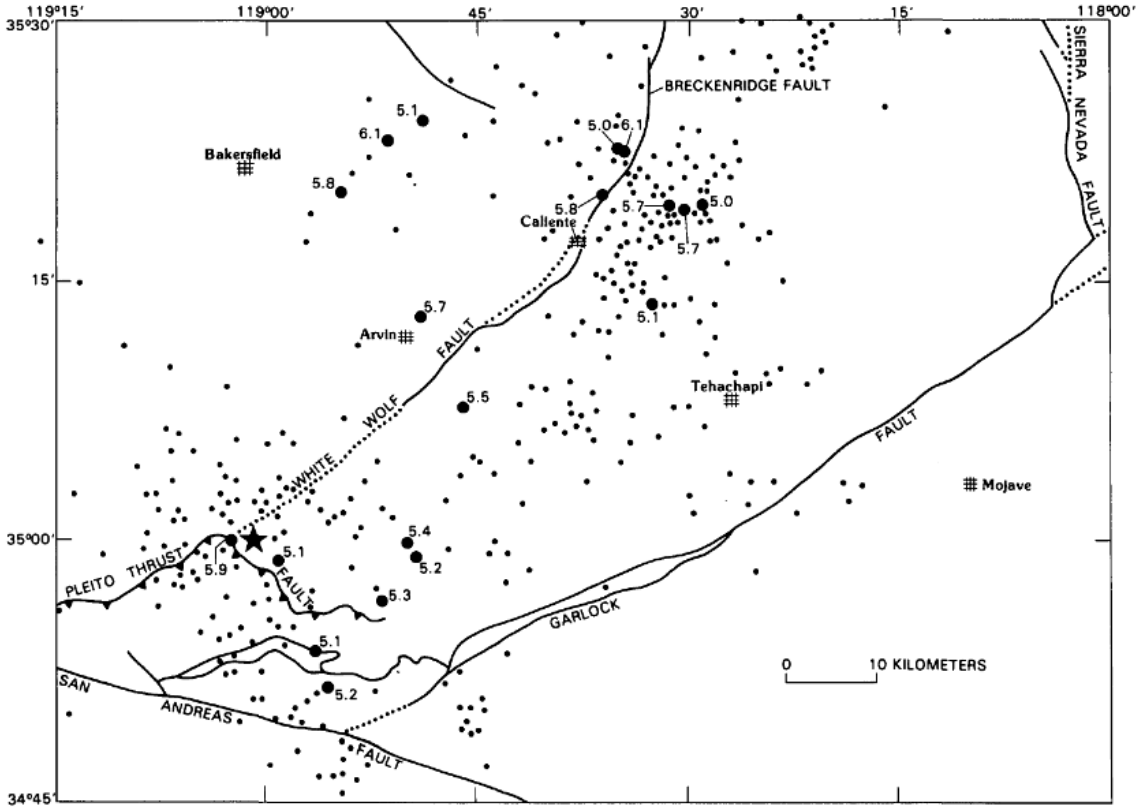


Figure 11: Major faults associated with the southern SJV. Major faults associated with the southern SJV. The Garlock fault is a sinistral fault whereas the San Andreas and White Wolf fault are dextral faults. The Pleito fault is a south-dipping thrust fault potentially associated with the Wheeler Ridge fault. The dotted sections of the faults depict inferred sections of the faults. Black circles represent earthquake epicenters

between 1932 and 1979. The star shows the location of the 1952 M_w 7.7 Kern County earthquake (Ross, 1986).

The San Joaquin Valley (SJV) defines the southern half of the Great Valley of California. The San Joaquin River drains the northern part of the sedimentary basin while the Kings and Kern Rivers generally drain the southern portion of the basin. The two southern rivers (Kings and Kern) extend to the topographic depressions of Tulare Lake and Buena Vista Lake. The Kern River contributes alluvial deposits from the Sierra Nevada in the eastern SJV near Bakersfield and the Kern River Oil Field near Oildale. Although the Kern River provides a significant amount of sedimentation, the SJV's stratigraphy primarily consists of a mixture of late Mesozoic and Cenozoic marine sediments, both biogenic and clastic (CCST, 2014b). Figure 12 depicts a geologic section block diagram of the Central Valley. Table 3 provides lithologic descriptions of primary stratigraphic units in the SJV.

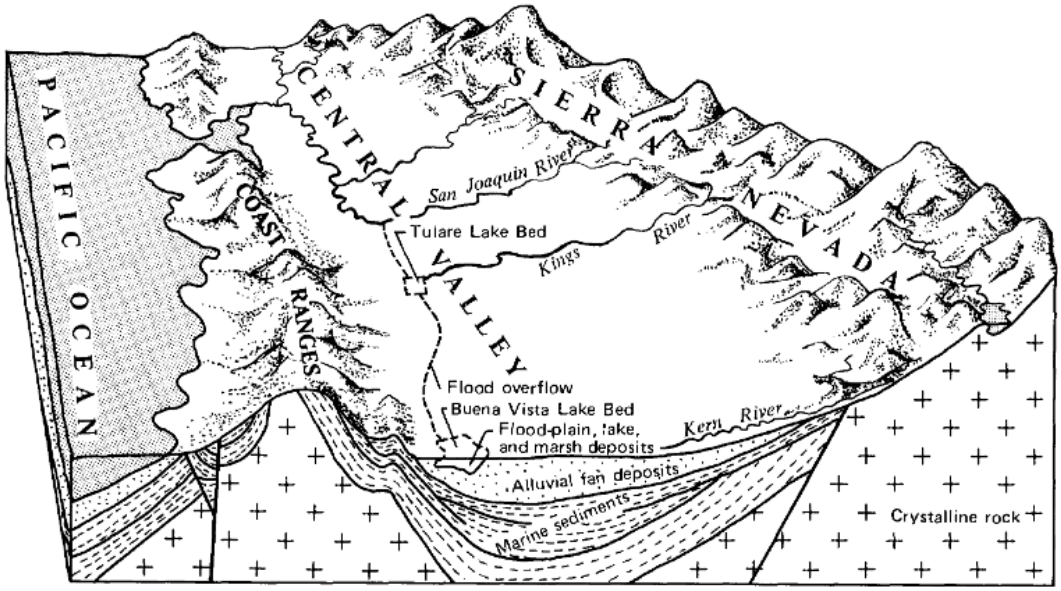


Figure 12: Generalized block diagram of cross-section through the Central Valley.

Generalized block diagram of cross-section through the Central Valley of California from the vantage point of the Transverse Ranges looking north (Page, 1983).

Table 3: Descriptions of primary Lithologic Units in the San Joaquin Valley

Lithologic Unit	Age	Description
Tulare Fm.	Pliocene - Holocene	Continental deposits primarily derived from alluvial-fan, flood-plain, lake and marsh environments. Sediments consist of unconsolidated clay, silt, sand and gravel.
San Joaquin Fm.	Pliocene	Oil bearing, clay-rich, very fine to fine grained sandstones with interbedded siltstone and clay stone. A marine fossil conglomerate exists at the base of the formation.
Etchegoin Fm.	Lower Plioscene – Upper Miocene	Marine fossil rich clay shales and sandstones.
Santa Margarita Fm.	Upper Miocene	Coarse sandstone and conglomerates with beds of chalky diatomaceous sediments.
Monterey Fm.	Miocene	Biogenic fine-grained shale, cherts and diatomites. Organic rich. Discussed in further detail below.
Temblor Fm.	Lower Miocene	Calcareous sandy interval underlying the Monterey. Contains shale layers with intermittent sand layers. Acts as oil reservoir formation.

Geology of the Monterey Formation and Belridge Oil Field

The Miocene Monterey Formation (Monterey) is a complex mixture of siliceous (diatomaceous), calcareous and phosphatic marine sediments. The Monterey is also California's most prominent oil source formation and reservoir. Deposited during the late Cenozoic transition from California's convergent margin to transform system, the Monterey underwent tectonic downwarping and landward transgression. The unit underlies many of California's depocenters, having a typical lithification duration of around 16 – 6 Ma (Behl, 1999). Although the Monterey is generally described as a shale unit, the formation tends to have very little clay mineral content, effectively categorizing a majority of the formation as a mudstone. Three members of the Monterey contain significant oil accumulations: the Tulare, Reef Ridge and Antelope. The primary targeted oil bearing units are largely associated with the Monterey diatomite, as happens in Kern County's largest oil field, the Belridge Oil Field (CCST, 2014b). Figure 13 compares the Monterey diatomite with black shale from the Marcellus Shale.

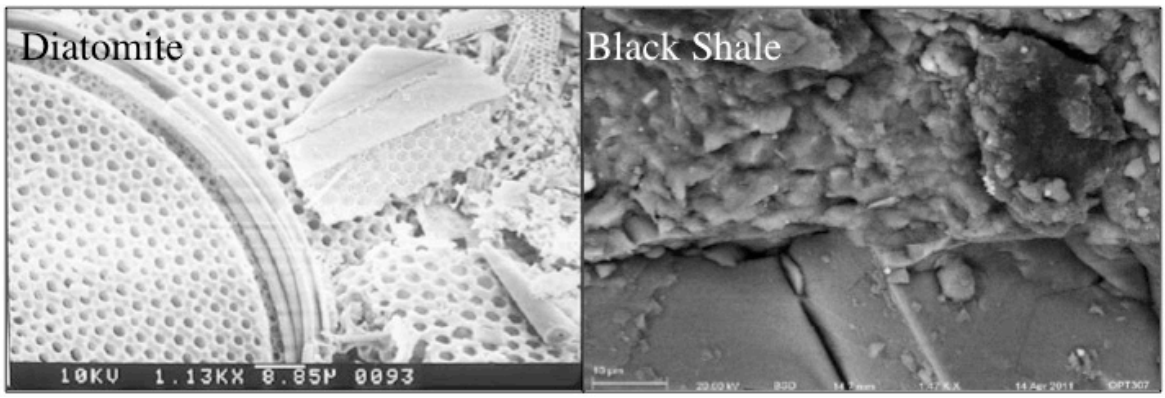


Figure 13: Comparison of diatomite and black shales. Comparison of diatomite found in California's Monterey Formation and black shales typical of the Marcellus Shale

and Bakken Formation. The diatomite contains porosities up to 70% whereas the black shale contains porosities between 1-11%. The differing lithology results in different hydraulic fracturing techniques (Behl, 2012; CCST, 2014b).

HYDROGEOLOGIC SETTING

The San Joaquin Valley – Kern County groundwater basin (DWR groundwater basin number 5-22.14) encompasses 1.95 million acres in the southern most portion of the Great Valley (Figure 14). The basin acts as a drainage region for the Kings, Kaweah, Tule and Kern Rivers. The four rivers drain into the topographic lows associated with the former Tulare, Buena Vista and Kern Lakes. The basin receives an average of 127 mm (5 inches) of precipitation a year (DWR, 2003).

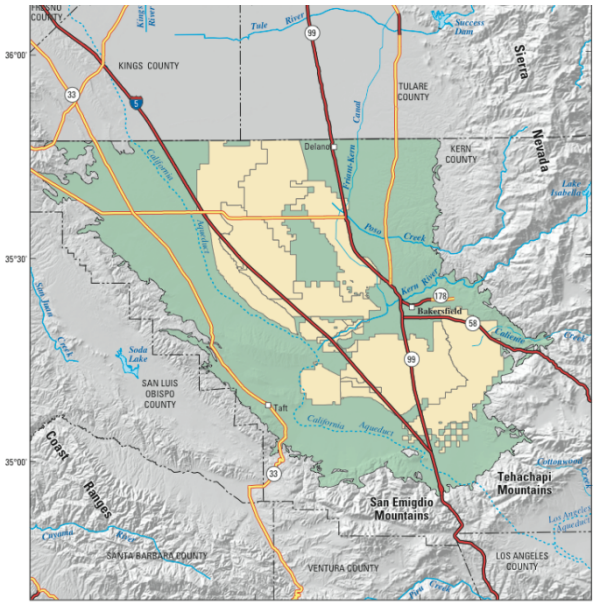


Figure 14: San Joaquin Valley – Kern County groundwater basin (5.22-14). San Joaquin Valley – Kern County groundwater basin (5.22-14) in green and groundwater

banking programs (tan areas) within the county. The groundwater basin is primarily recharged by the Kern River (Shelton et al., 2008).

The Kern River is the primary mode of groundwater recharge; however, Kern County water agencies redistribute water via elaborate artificial recharge groundwater banking facilities. Kern River water recharges the eastern extent of the basin through large alluvial fans. The main water-bearing formations consist of Tertiary to Quaternary continental sediments ranging from gravels to clay. In general, the Tulare and Kern River formations act as the primary aquifers. The Corcoran Clay, a low permeability member of the Tulare Formation, separates the unconfined from the confined aquifer systems. The top of the Corcoran Clay is at depths of between 300 and 650 and wells through the Corcoran Clay extend up to 800 feet deep. Many of the groundwater supply wells in Kern County are in the central or eastern portion of the county for agricultural, municipal and domestic use.

The majority of the groundwater data in Kern County is from the central to eastern portion of the basin because the Kern River and the City of Bakersfield are in the eastern portion of the groundwater basin. Little data exist regarding the physical hydrogeology and groundwater quality in the western extent of the basin. In order to fully understand groundwater movement, recharge and quality, additional information is needed along the western margin of the basin.

Hydrogeochemistry of Kern County Sub-Basin

The groundwater infiltrating aquifers on the east side of the basin derives from the Kern River. The river provides aquifers with water through coarse to fine grained

alluvial fans consisting of materials from the Sierra Nevada. The Kern River's headwaters, in the high Sierra Nevada, deliver snowmelt to the river and alluvial fans provide ample filtration of the water, creating high quality water with low salinity and total dissolved solid (TDS) content. The groundwater in the western portion of the basin likely doesn't receive as much river infiltration as the east, therefore reducing the amount of water circulation. Less water circulation likely increases the duration of water-rock interaction, which effectively increases the total dissolved solid and salinity content. Figure 15 shows a map of TDS from groundwater within the basin indicating lower TDS concentrations in the east compared to TDS concentrations in the west

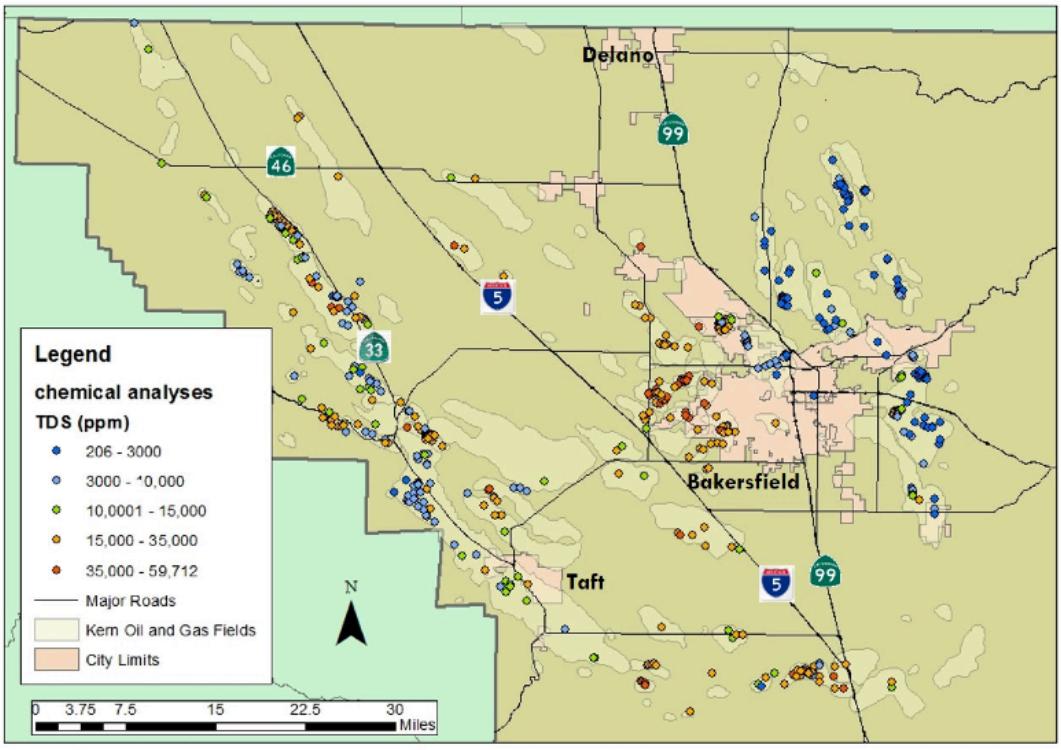


Figure 15: TDS concentrations in the Kern County groundwater basin. TDS concentrations in the Kern County groundwater basin. Cooler colors indicate less TDS

whereas warmer colors indicate a larger TDS concentration. In general, the east side of the basin where the Kern River discharges contains low TDS and the west side of the basin contains high TDS concentrations (Esser et al., 2015).

The deeply seated formation fluids in the Kern County groundwater basin have complex hydrogeochemistry due to the basins dynamic geologic history. To reiterate the Geologic Setting section, the basin underwent the transition from a continental margin to a forearc sedimentary basin within a 23 My period. During the transition, the Mesozoic South Sierran Block was depositing granitic sediments on the continental shelf and coastal plain on the east side of the basin via river systems with headwaters in the high alpine regions of the Sierra. Concurrently during this period, the western extent of the basin largely consisted of a marine depositional environment, which ultimately deposited the principle source rocks (illites and shales) for the large oil fields in the west. Figure 16 provides a schematic of the geologic environment in the Kern County basin during the middle Miocene and late Miocene.

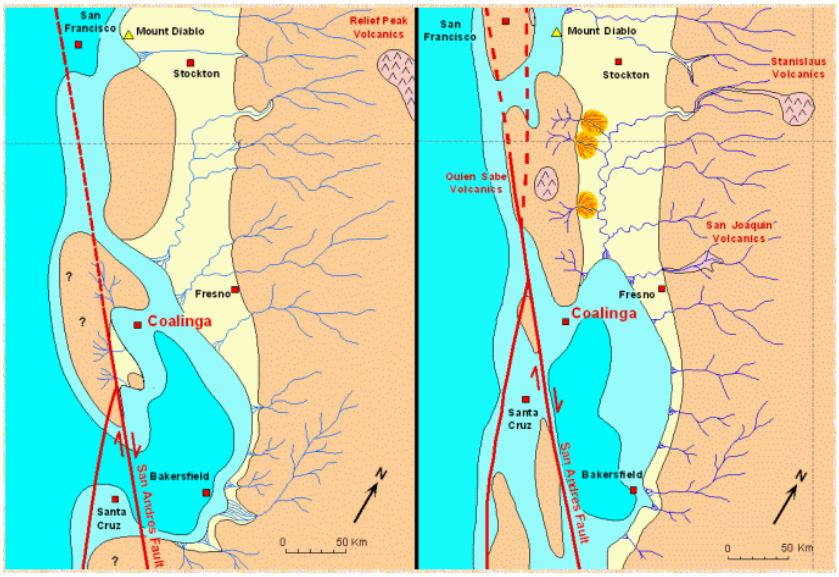


Figure 16: Schematic of the depositional environments of the Kern County basin. Schematic showing the depositional environments of the Kern County basin during the middle Miocene (left) and late Miocene (right). The darker blue (teal) represents a deep-sea basin and the lighter blue represents a continental shelf. The yellow area represents an alluvial depositional environment derived from the highland material represented by orange. The transition of the Kern County basin from a continental margin to a forearc sedimentary basin generated a mix of seawater derived formation fluids which then were largely altered by water-rock interaction (Bartow, 1991; Clark, 2015)

The hydrogeochemistry of the formation waters in the Kern County groundwater basin is affected by:

- Ancient seawater associated with the continental margin transition and sea level regression;

- Water-rock interaction involving South Sierran Block granitics, marine facies rocks and ophiolite-serpentanite rocks associated with the Coast Ranges; and,
- Meteoric infiltration through current and ancient rock formations.

In addition, water from oil-bearing formations may contain an organic signature representative of the hydrocarbons. With that, determining the origin of these formation waters is difficult. To better understand the hydrogeochemistry of these waters, the Kern County oil field produced waters data, which chemically represents the basin formation waters, was investigated (see Chapter I – Background: Chemistry of Produced Waters section for a description of produced waters; the produced waters data is described in further detail in the Methods: Data Acquisition section). By dividing the basin into a western half and eastern half, the formation waters were evaluated under the hypothesis that western water contains a different hydrogeochemical signature than waters on the east (Figure 17). This hypothesis was derived from two ideas:

1. The eastern waters interact with granitic material and middle to early Miocene alluvial deposits. Also, the Kern River drains into the eastern half of the basin supplying the groundwater system with infiltrated river water.
2. The western waters are largely associated with the Miocene marine deposits and sea level regression showing a fossil seawater signature. Also, potentially much less meteoric mixing due to the very low rainfall and lack of surface water bodies.

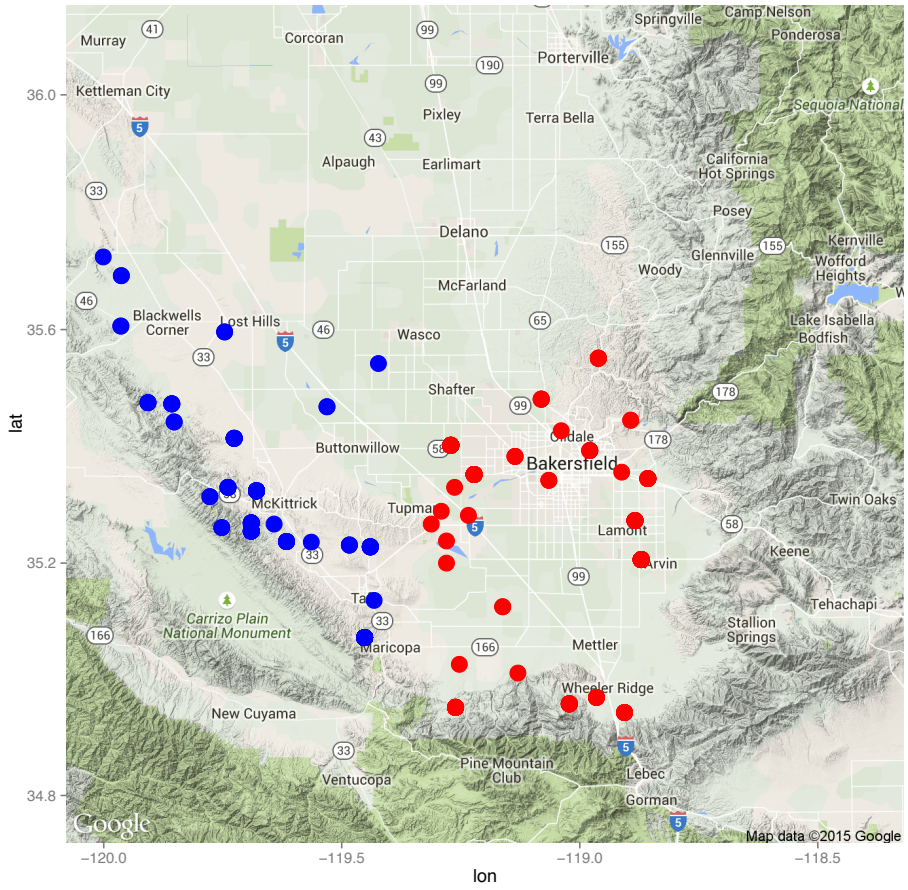


Figure 17: Segregation of produced water samples into west and east. Segregation of produced water samples into west (blue) and east (red) to evaluate the origin of the formation waters on either side of the basin. The produced waters data for the east basin contains 133 samples and the produced waters data for the west contains 134 samples.

In general, both the eastern and western formation waters contain positive correlations between chloride (Cl) and total dissolved solids (TDS), sodium (Na) and TDS and Na to Cl indicating the waters contain a sodium chloride water type or seawater signature. Weddle (1967), also notes the oil field brines in the Midway-Sunset and Elk Hills oil fields in Kern County are of a sodium chloride type. Figure 18 shows the

positive correlation between Na and Cl in both the eastern and western waters. The NaCl type water may signify the formation waters were originally seawater entrapped in the pores of sediments during diagenesis. Although both the western and eastern waters indicate a potential seawater origin they contain differing concentrations of dissolved solids and other chemical constituents. For example, the TDS concentration increases from the Carrizo Plain west of the Temblor Range to central Kern County reaching its maximum concentration within the longitudes that encompass nearly all of the oil fields in western Kern County (i.e. Belridge, Lost Hills, Elk Hills, Cymric, Midway-Sunset, Paloma, etc.). The TDS concentrations distinctly decline east of the oil fields as the longitudes approach the Sierra Nevada foothills (Figure 19).

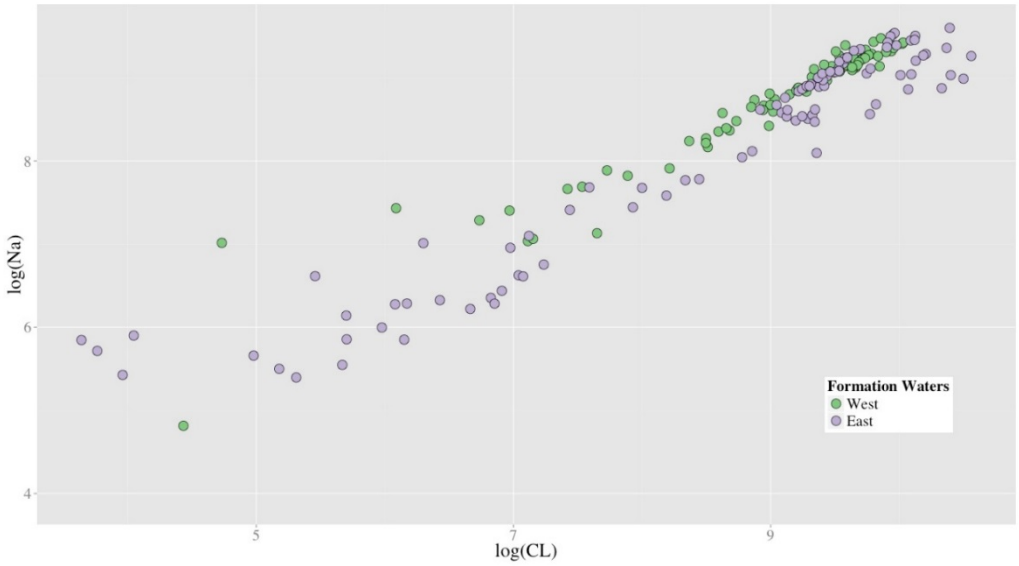


Figure 18: Log-log relationship between Na and Cl. Log-log relationship between Na and Cl concentrations in Kern County formation waters. The plot shows both the eastern

and western waters are of a NaCl type, which indicates a seawater signature. These data agree with the analysis of oil field brines by the Division of Oil and Gas (Weddle, 1967).

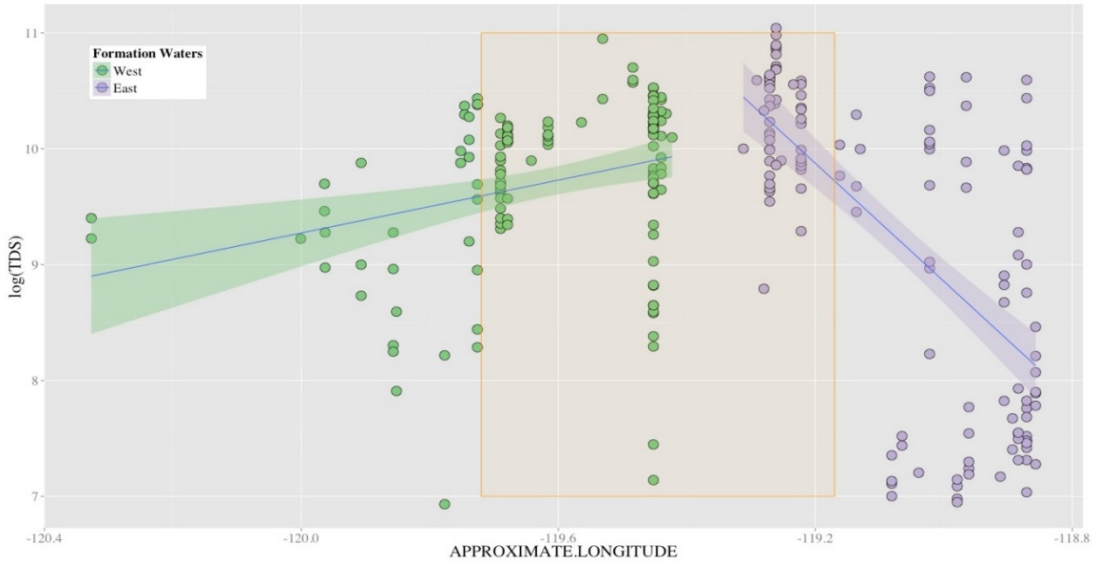


Figure 19: Relationship of log TDS concentration to longitude. Relationship of log TDS concentration to longitude with linear trend lines for both the segregated eastern (purple) and western (green) produced water samples (see Figure 17 for the distinction between eastern and western produced waters). The TDS concentration steadily decreases away from the primary oil field bearing longitudes. The orange square represents the longitudes that encompass nearly all of western Kern County's primary oil fields such as: Belridge, Lost Hills, Elk Hills, Cymric, Midway-Sunset and Paloma. The TDS concentration sharply declines in the eastern formation waters likely due to the influence of meteoric water from the Kern River.

Other constituents such as boron (B), sodium-potassium (K-Na), bicarbonate (HCO_3) and calcium to sodium ratio (Ca/Na) show slight differences between the eastern and western waters. The western waters contain higher boron concentrations than the

east. The high boron concentration in the west likely stems from desorption from the illite dominant soils, serpentinite rocks and marine shales associated with the Coast Ranges (Su and Suarez, 2004). The marine facies associated with the Coast Ranges and the Monterey Formation also contain elevated concentrations of HCO_3. Comparatively, granitic facies, which dominate the eastern portion of the basin, do not contain high concentrations of boron or HCO_3. Granite, however, does contain high concentrations of alkalies, potassium and sodium. K-Na concentrations in the eastern waters vary but contain a higher mean concentration compared to western waters. The high K-Na in the east may be attributed to the original seawater concentration compounded by the addition of alkalis from the weathering of orthoclase and albite in the Sierra Nevada granitics. In addition, Mazor (2004) suggests Ca and Na dominate groundwater associated with granite. The Ca/Na ratio for the eastern waters contains a higher ratio than waters in the west likely due to the weathering of anorthite plagioclase within the South Sierran granitics. In all, the eastern and western waters potentially originated as entrapped seawater with similar chemistry, but after the continental margin to forearc sedimentary basin transition, water-rock interaction altered the chemistry of the waters representing the Sierra Nevada granitics in the east and the Monterey and Etchegoin formations in the west (see Table 4 and Geology of the Monterey Formation section above for descriptions of the Monterey and Etchegoin formations). Figure 20 shows the B, K-Na, HCO_3 and Ca/Na relationships between the western and eastern waters.

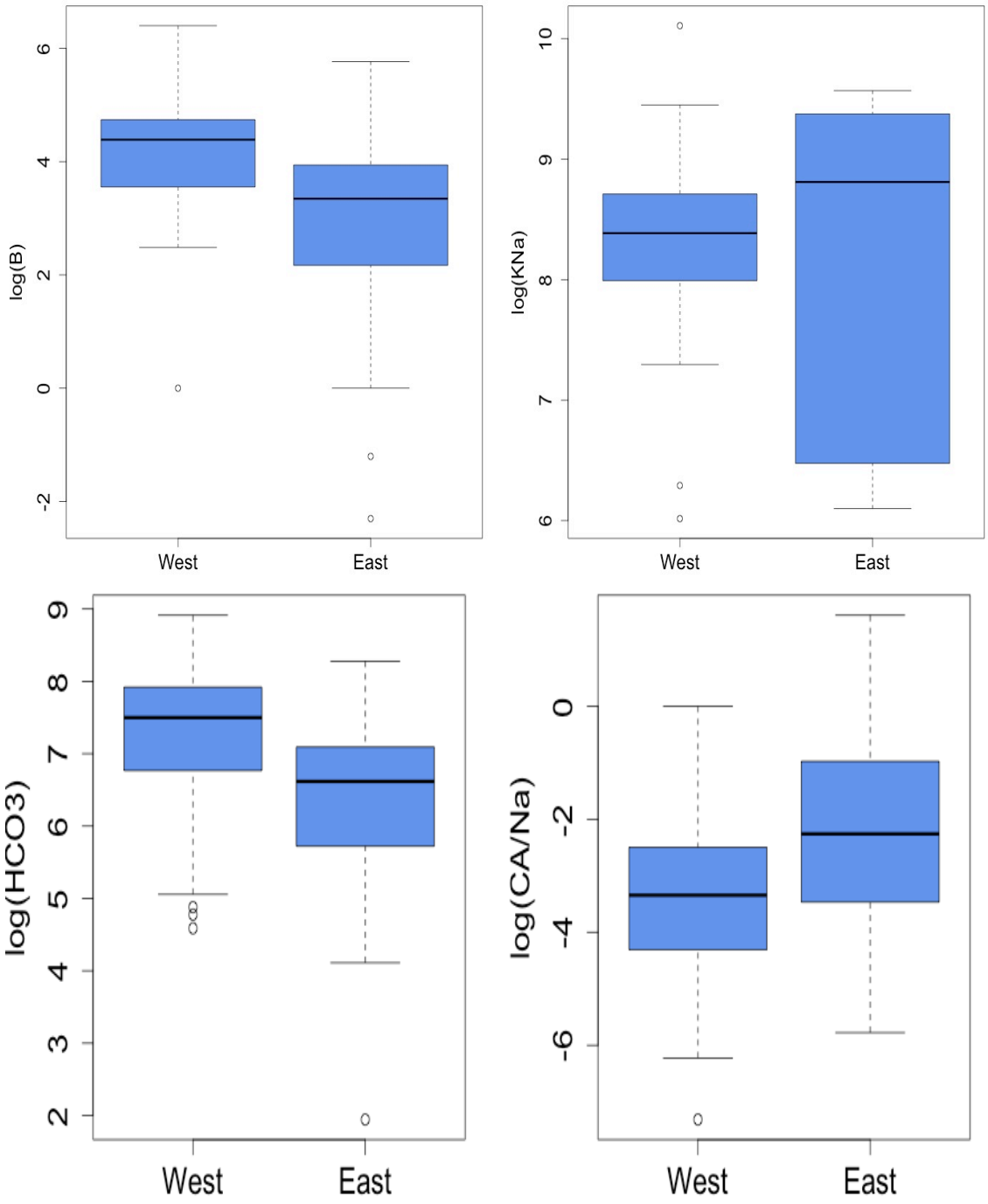


Figure 20: Eastern and western water relationships. Eastern and western water relationships. The western waters contain a higher boron concentration likely due to

sediments derived from the marine rocks desorbing boron from clay and serpentine minerals (top left). The higher KNa concentration in the east may be attributed to the weathering of granite containing orthoclase ($KAlSi_3O_8$) and albite ($NaAlSi_3O_8$) in addition to the original seawater concentration (top right). The marine facies associated with western Kern County probably contributes HCO_3 to the western formation waters resulting in a higher concentration in the west relative to the east (bottom left). The weathering of albite and anorthite plagioclase feldspars in the Sierran granitics likely contributes to the higher Ca/Na ratio in the east relative to the west (bottom right).

Both the eastern and western waters have interacted with meteoric waters. The meteoric water dilutes the formation waters through existing or ancient outcrops and structures and during deformation (Rogers, 1919; Weddle, 1967). In addition, meteoric water from the Kern River likely interacts with eastern formation waters along the basal plain underlying the alluvial fans jetting out from the Sierra Nevada. With that, the eastern waters likely contain a larger meteoric water signature than the western formation waters. To illustrate the meteoric influence on the formation waters the calcium to magnesium ratio (Ca/Mg) was evaluated. In general, seawater has a relatively low Ca/Mg ratio due to a much higher magnesium concentration relative to calcium (magnesium is over 3 times more abundant in seawater than calcium). In contrast, calcium in Kern County's shallow groundwater is approximately 8.5 times more abundant than magnesium. Therefore, shallow groundwater, which is likely affected by meteoric water, has a higher Ca/Mg ratio than seawater. Further, seawater diluted with meteoric water likely would contain a higher Ca/Mg ratio than seawater. Figure 21

shows the produced waters and shallow groundwater in the east contains nearly the same Ca/Mg ratio whereas the shallow groundwater and produced waters in the west do not. These data along with the geological environment in the east basin provide very generalized evidence that the eastern produced waters show a stronger meteoric signature than the west.

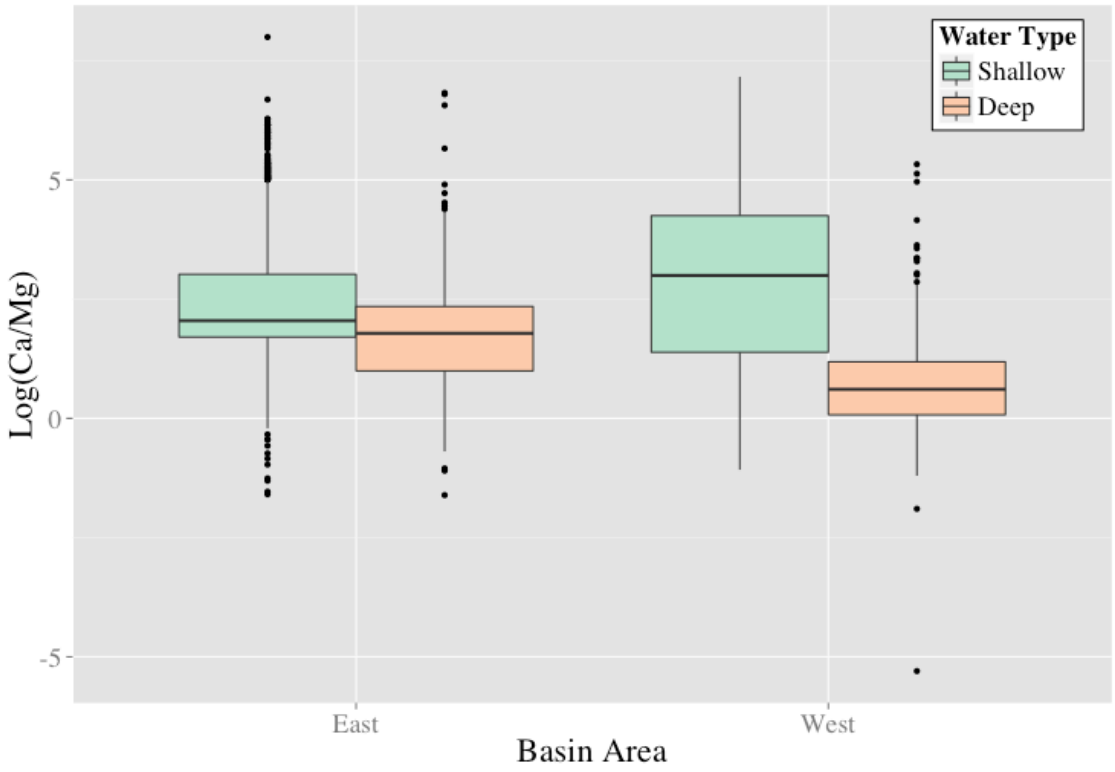


Figure 21: Comparison of Ca/Mg ratios. Comparison of Ca/Mg ratios for eastern produced waters and shallow groundwater relative to the western produced waters and shallow groundwater. The eastern produced waters contain nearly the sratio value indicating a potential meteoric signature in the produced waters. Note: the units on the y-axis are not the same.

The produced water in Kern County likely originated as entrapped seawater, but since the enclosure of the basin from the Pacific Ocean water-rock interaction dominates the hydrogeochemistry of the waters. The waters also show evidence of mixing with meteoric water. The eastern waters likely contain a stronger meteoric influence relative to the west due to the Kern River discharging from the Sierra Nevada into large alluvial fans overlaying basement material. In all, the produced waters contain complex hydrogeochemistry but are distinctly different from east to west.

CHAPTER III: Statistical Model Methods

DATA ACQUISITON

Multiple public databases were utilized in this research. In addition, data from analytical measurements at Lawrence Livermore National Laboratory (LLNL) aid the discussion of the data. Data quality control and data gaps are discussed in Chapter IV: Data Gaps section. A description of the data is below:

- Groundwater Ambient Monitoring and Assessment Program (GAMA): The GAMA database is a compilation of groundwater chemistry data reported by multiple entities including the California Department of Public Health, USGS, and LLNL. All of the data discussed herein came from public drinking water supply wells rather than groundwater monitoring wells. The data is accessible to the public through the SWRCB's GeoTracker website: http://geotracker.waterboards.ca.gov/gama/data_download.asp
- California's Division of Oil, Gas & Geothermal Resources (DOGGR): DOGGR regulates all oil and gas operations in California, which includes the implementation of the California Senate Bill 4 Oil and Gas: Well Stimulation. DOGGR's role includes reviewing and permitting well stimulation applications. The data gathered from the permit applications and approved wells are compiled in DOGGR's

Well Stimulation Treatment database. The data consist of well locations (latitudes and longitudes), well depth, targeted formation, etc. In addition, DOGGR provides a dataset with all of the active and inactive oil and gas wells in California. The databases are accessible through the following URLs:

http://maps.conservation.ca.gov/doggr/iwst_index.html

http://www.conservation.ca.gov/dog/maps/Pages/GISMapping2.aspx

- United States Geological Survey (USGS): Data from the USGS was acquired from the USGS National Produced Waters Geochemical Database (USGS_PWGD). The USGS_PWGD is a national initiative to compile geochemical data from produced waters gathered from oil plays throughout the US. In addition, USGS analyzes some samples in its laboratories and these data, from California oil fields, are in the database. The USGS_PWGD database can be found at the following URL:

http://energy.usgs.gov/EnvironmentalAspects/EnvironmentalAspectsofEnergyProductionandUse/ProducedWaters.aspx#3822349-data

- Lawrence Livermore National Laboratory (LLNL): Four samples were provided to LLNL from four oil wells in three different oil fields in California. Although the analysis of the four produced waters provides useful information (see Chapter V), the data were not incorporated into the statistical model.

Data Reduction

Although the State of California makes an abundance of chemical groundwater data available to the public, the data exist in a form difficult to analyze. A data reduction process helps to transform the downloadable raw data into a working data frame useful for analysis and interpretation. The SWRCB's GeoTracker (GAMA) database, in particular, requires significant processing prior to use. After processing the data, statistical and spatial analysis was carried out within R-Project Statistical Software (R) and Microsoft Excel.

The following procedure outlines the steps that were carried out to create a California statewide groundwater quality database useable in R. First, extract all of the California groundwater quality data from GeoTracker by selecting the '[All Data]' hyperlink for each county (Figure 22). Place the text file (e.g.,, gama_all_alameda.txt) from each zip file into a folder directory. Utilizing the folder directory path, input the following script (Script 1) to load the folder contents into R and generate a 'dataset.'

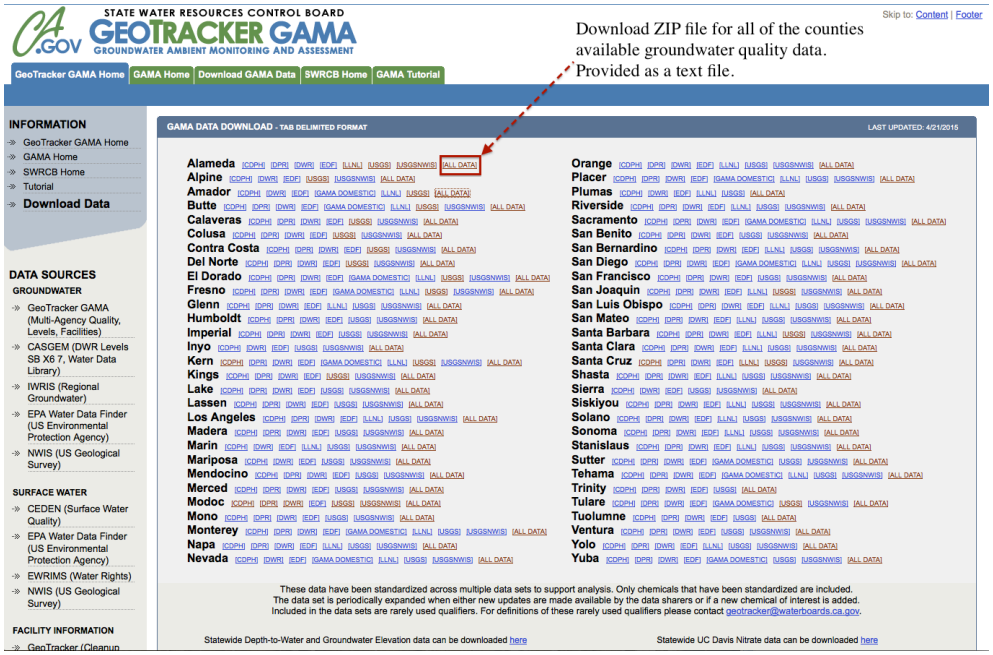


Figure 22: The California State Water Resources Control Board's GeoTracker.

The California State Water Resources Control Board's GeoTracker GAMA data page houses chemical data for all 58 California counties. The downloadable data is provided as a text file within a ZIP file. The data include groundwater monitoring data from private industry groundwater cleanup sites and California Department of Public Health, USGS and LLNL groundwater data from water supply wells. The database is the repository for data collected under the statewide Groundwater Ambient Monitoring and Assessment program overseen by the SWRCB and implemented by the USGS and LLNL.

Script 1: Inputting Text Files (Godwin, 2011)

```
setwd("/Users/renshaw2/Documents/0_RESEARCH/California_GW_Quality/GamaData/data/TEXT_FILE")

file_list <- list.files()

for (file in file_list){

    # if the merged dataset doesn't exist, create it
    if (!exists("dataset")){
        dataset <- read.table(file, header=TRUE, sep="\t")
    }

    # if the merged dataset does exist, append to it
    if (exists("dataset")){
        temp_dataset <-read.table(file, header=TRUE, sep="\t")
        dataset<-rbind(dataset, temp_dataset)
        rm(temp_dataset)
    }

}
```

Inputting the county text files yields a California statewide groundwater quality dataset (termed dataframe in R) of 47,774,808 data points (or observations) and 12 variables. The 12 descriptive variables, such as the latitude and longitude of the sampled well, are presented in Table 4.

Table 4: Descriptive Variables within the Statewide Dataset

Variable	Description
WELL.NAME	Well identification number and/or name
APPROXIMATE.LATITUDE	Approximate latitude of the well sampled
APPROXIMATE.LONGITUDE	Approximate longitude of the well sampled
CHEMICAL	Chemical constituent analyzed
QUALIFIER	Identifier from laboratory analysis, such as non-detect (ND) or sample below detection limit (<)
RESULT	Numerical value from analysis
UNITS	Unit of analytical measurement (mg/L)
DATE	Date the sample was collected
DATASET_CAT	Type of well the sampled
DATASET	Sampling agency (USGS, LLNL, CDPH)
COUNTY	County the sample was collected in
GW_BASIN_NAME	DWR defined groundwater basin

To ensure a complete upload of the data, run a summary function to provide the frequency of data points by county. The output allows for an assessment of whether the data input was successful at compiling all 58 counties data (Script 2).

Script 2: California County Summary Data

```
summary(all$COUNTY)

##                      ALAMEDA          ALPINE          AMADOR
##                      2472788            9328           85317
##           BUTTE       CALAVERAS          COLUSA    CONTRA COSTA
##          364820           40651           69588          677222
##       DEL NORTE       EL DORADO          FRESNO           GLENN
##           40710          276021         1078739           58048
##        HUMBOLDT        IMPERIAL            INYO            KERN
##          245638          111143           61120          966271
##           KINGS            LAKE          LASSEN     LOS ANGELES
##          157655           96543           51977        10101677
##          MADERA           MARIN        MARIPOSA       MENDOCINO
##          182450          137074           48573          316273
##          MERCED           MODOC            MONO        MONTEREY
##          451026            7515           50358          881685
##            NAPA          NEVADA          ORANGE          PLACER
##          175646           93880         6574528          321456
##          PLUMAS       RIVERSIDE      SACRAMENTO      SAN BENITO
##           52384         2441366         1736121          167731
##  SAN BERNARDINO       SAN DIEGO   SAN FRANCISCO     SAN JOAQUIN
##         2913449         2359319          173619         1708461
## SAN LUIS OBISPO       SAN MATEO   SANTA BARBARA     SANTA CLARA
##          461582          222837         1746809         1944374
##      SANTA CRUZ          SHASTA          SIERRA        SISKIYOU
##          235325          149482           12345           82832
##          SOLANO          SONOMA      STANISLAUS          SUTTER
##          517315         1214280          887476           99973
##          TEHAMA         TRINITY          TULARE        TUOLUMNE
##          189471           32783          522651          113304
##         VENTURA            YOLO            YUBA
##          964456          396282          223060
```

After building a statewide dataframe, subsetting the data to accommodate particular questions becomes possible. Subsetting the California data for Kern County provides a dataframe useable for spatial and chemical analysis of shallow groundwater quality in relation to oil and gas production in the county. Kern County, however, contains 20 groundwater basins identified in Bulletin 118 (DWR, 2003). Of the 20 basins, the San Joaquin Valley – Kern County subbasin (DWR basin number 5-22.14)

has the highest priority ranking in the DWR CASGEM program for Kern County and contains the most abundant amount of groundwater quality data and nearly all of the oil and gas production in the county. With that, a San Joaquin Valley – Kern County subbasin subset from the Kern County dataframe provides the most applicable data for groundwater quality assessment. Script 3 shows the subsetting process.

Script 3: Subsetting California Dataframe

```
kern <- all[all$COUNTY %in% c("KERN"), ]
kern <- kern[kern$GW_BASIN_NAME %in% c("SAN JOAQUIN VALLEY - KERN
COUNTY (5-22.14)"), ]
```

The Kern County subbasin dataframe contains 675,129 observations and the analysis of over 100 different chemicals. One idiosyncratic issue is in regards to the how the element sodium is identified in the database. In the dataframe, sodium is represented as 'NA' for the chemical variable; however, R responds to observations with 'NA' as missing data. So, in order to prevent R from misclassifying sodium as missing data, the chemical observation 'NA' is substituted with 'Na.' Further, approximately 20% of the data come from environmental monitoring wells, which typically come from shallow (<30 ftbgs) aquifers and do not act as a drinking water resource. Therefore, removing environmental monitoring wells (Script 4) creates a more refined dataframe containing only groundwater supply wells within the Kern County groundwater basin. The new Kern County dataframe consists of 561,026 observations.

Script 4: Removal of Unwanted Data

```
kern <- subset(kern, !(DATASET_CAT %in% c("ENVIRONMENTAL MONITORING
(WELLS)")))
```

Note: the '!' in the script equates to 'NOT' in the code.

In order to compare the GAMA data with the DOGGR data, the DOGGR data must be loaded into the R-project workspace. The well stimulation treatment (WST) data and all active and inactive oil and gas well data is accessible on the DOGGR website at the URLs provided above. Downloading the data and transforming it into a comma separated value spreadsheet allows importation into R. The raw WST data contains 12 variables and is in a form useable for data analysis with the exception of the latitude and longitude of the wells. The latitudes and longitudes in the WST data exist as one variable (separated by a comma); however, separating the coordinates into two variables makes spatial analysis possible.

After separating the coordinates the data are ready for upload into the R-project workspace. In regards to the DOGGR all oil and gas wells database (conventional or not; active or inactive), the downloadable data exist in a form useable in the R-project workspace requiring no data reduction. The reduction and data input process create a R-project workspace useful for spatial and statistical comparisons between the shallow groundwater data in GAMA and the deeply seated formation waters associated with oil and gas operations. In all, three dataframes reside in the R-project workspace: GAMA (shallow groundwater data), WST (well stimulation treatment data) and produced (oil and gas well produced waters data).

SPATIAL ANALYSIS PROCESSING

Two R packages, `ggmap` (from `ggplot2`) and `SpatStat`, provide plots and numerical assessments of data to interpret spatial relationships of the three datasets.

ggmap

Utilizing the GAMA and WST datasets, map view plots provide an avenue for analyzing where groundwater supply well locations overlap future well stimulation treatment and current oil and gas production wells in Kern County. The `ggmap` package first extracts a Google Map using a central coordinate within the `get_map` function and creates a context layer from the output. For the Kern County analysis the following decimal degree coordinates provides sufficient maps: 35.450, -119.375. The `get_map` function allows for zooming in and out within the map area, black and white or color and different output styles such as terrain or road map. Once the `get_map` layer exists, `ggmap` along with `ggplot` facilitates the overlaying of different layers of data from within the R-project workspace utilizing latitudes and longitudes of the data points (Kahle and Wickham, 2013). In addition, `ggplot` allows for the input of well-specific data such as the TDS concentration. The well-specific data utilizes `ggplot`'s color or fill function to create heatmap style plots of continuous variables. The `ggplot` function allows for the control of many aesthetic attributes of the plot such as: color, size and opaqueness of the points, shape of the points, titles and legend descriptions, among others. Utilizing the annotate function in conjunction with `ggplot` provides the functionality of adding text and shapes (lines, arrows and geometric shapes) to the plots. Script 5 provides the detailed script for plotting the GAMA data against the WST data with continuous variable TDS and annotated text and shapes. The Script 5: Mapping is viable for all dataframes as long as the data contain a latitude and longitude and the code specifies the desired dataframe.

Script 5: Mapping

```
library(ggmap)

## Loading required package: ggplot2

#Create Map
HF <- get_map(location = c(lon = -119.375, lat = 35.45),
color = "color",
source = "google",
maptype = "terrain",
zoom = 9);

## Map from URL :
http://maps.googleapis.com/maps/api/staticmap?center=35.45,-
119.375&zoom=9&size=640x640&scale=2&maptype=terrain&language=en-
EN&sensor=false

ggmap(HF,
extent = "device",
ylab = "Latitude",
xlab = "Longitude");

HF <- ggmap(HF) +
geom_point(aes(x = K3$APPROXIMATE.LONGITUDE, y =
K3$APPROXIMATE.LATITUDE, colour=TDS),
data = K3, size=7, shape = 15, alpha = 0.75) +
theme(legend.title = element_text(size=16, face="bold"),
legend.text=element_text(size=14)) +
scale_colour_gradientn(name="Total Dissolved Solids (mg/L)",
colours=c("blue", "aquamarine", "springgreen", "gold", "goldenrod1",
"chocolate1", "magenta", "firebrick1"));

y <- HF + geom_point(aes(x = Wells_hf$Long, y = Wells_hf$Lat,
fill=Field),
data = Wells_hf, size = 7, shape = 21, alpha = 0.88) +
theme(legend.title = element_text(size=16, face="bold"),
legend.text=element_text(size=14)) +
scale_fill_discrete(name="Approved WST\nLocations");
```

The spatial map view plots provide location and frequency of GAMA samples and location and spatial relationships of WST well data. Although visualizing data in a geographical context is useful, quantitative spatial analysis can be used to confirm spatial correlations between populations of data.

SpatStat

The `SpatStat` package allows for numeric spatial analysis primarily utilizing spatial point patterns (Baddeley et al., 2015). The `SpatStat` function `ppp` in conjunction with the `quadratcount` function generates point pattern plots containing the density of points within gridded spatial matrices defined by confining coordinates. Before applying the point pattern analysis, the coordinate systems of the three dataframes need to be transformed (the `ppp` function does not read negative numbers in decimal degree formatted latitude and longitude) from latitude and longitude to North American Datum 1983 (NAD83). The transformation process requires conditioning the data against a known datum. A shape file of the Kern County boundary (from the Kern County Development Services Agency website: http://esps.kerndsa.com/gis/gis-download-data), which utilizes NAD83, acts as the baseline datum. The `proj4string` and `spTransform` functions transform the GAMA, oil well and produced water data to NAD83 (Script 6).

Script 6: Coordinate Transformation

```
library(rgdal)

library(ggplot2)

#Read in Kern County GAMA Data
water <-
read.csv("~/Documents/0_RESEARCH/Interim_WST_Data/KernCounty_WaterWells
.csv")
mapdata <- water

#Read in Kern County Boundary shapefile
kern <- readOGR(dsn =
"/Users/renshaw2/Documents/0_RESEARCH/Interim_WST_Data/data/KernCounty_
Shapes/KernBoundary.shp", layer = "KernBoundary");

#Transform Latitude and Longitude to NAD83 conditioned to the Kern
County shapefile
coordinates(mapdata) <- ~Long+Lat
```

```
proj4string(mapdata)<-CRS("+proj=longlat +datum=NAD83")
mapdata<-spTransform(mapdata, CRS(proj4string(kern)))
wells<-data.frame(mapdata)
names(wells)[names(wells)=="Long"]<-"x"
names(wells)[names(wells)=="Lat"]<-"y"

##proj4string(mapdata)

[1] "+proj=lcc +lat_1=34.03333333333333 +lat_2=35.46666666666667
+lat_0=33.5 +lon_0=-118 +x_0=2000000 +y_0=500000.0000000001
+datum=NAD83 +units=us-ft +no_defs +ellps=GRS80 +towgs84=0,0,0"
```

The coordinate transformation produces spatial location data that correspond with the `ppp` function. Further, generating three `ppp` plots for the three dataframes within the R-project workspace allows for spatial ratio analysis of the density matrices. The three spatial matrices consist of an 8 by 6 grid, in which the grid numbers start at 1 in the upper left corner and end with 48 in the lower right corner (see Figure 29 in Chapter IV). The entire area of the grid equals approximately 4520 square miles (mi^2). Each square in the grid represents approximately 94 mi^2. The area of the grid and squares are calculated using the confining coordinates in the `owin` and `area.owin` functions in `spatstat`. The `owin` function produces a spatial window using the confining coordinates derived from the extent of the Kern County data. The spatial window contains the geometric data of the area produced, in which the area can be calculated using the `area.owin` function. Dividing the `area.owin` output by 48 and multiplying by 3.58701e-8 mi^2 per ft^2 provides the area in mi^2 of each square in the spatial grid. The units are obtained from the summary output of the coordinate transformation process (highlighted in Script 6). Script 7A shows the script for analysis of the spatial area and grid. The ratios look at the number of GAMA samples (or data points) relative to number of oil wells within one of

the grids in the spatial matrix; therefore, a ratio below 1.0 indicates a larger abundance of oil wells within the particular square. Even if one of the ratios results in 1.0, further investigation into the distribution within that particular square would be required considering each square consists of 94 mi^2. Considering the small population of produced water samples (~300) compared to the GAMA and oil well data, the produced water data are not incorporated into the ratio. However, visual comparison of the density grid for the produced water data to the other two dataframes proves useful. Script 7B depicts the general use of the `ppp` function in `spatstat`.

Script 7A: Spatial Point Pattern Analysis – Area Calculation

```
library(spatstat)

##
## spatstat 1.42-2       (nickname: 'Barking at Balloons')
## For an introduction to spatstat, type 'beginner'

#Generate spatial window using Kern County bounding coordinates
#Calculate the area in sq ft then convert to sq mi
s <- owin(c(5935000, 6339000), c(2168000, 2480000))
plot(s)

area.owin(s)

## [1] 1.26048e+11

(area.owin(s)/48)*3.58701e-8

## [1] 94.19488

##Area of each square in the 8 x 6 – 48 grid plotted in spatial window
‘s’.  See Script 7B.
```

Script 7B: Spatial Point Pattern Analysis

```
library(spatstat)

water <-
read.csv("~/Documents/0_RESEARCH/Interim_WST_Data/data/GAMA_Data/KernCo
unty/KernCounty_WaterWells.csv")
```

```
##Remove duplicate data points and run point pattern analysis for GAMA
data
water <- unique(water)
s <- ppp(water$x, water$y, c(5935000, 6339000), c(2168000, 2480000))

Q <- quadratcount(s, nx = 8, ny = 6)
plot(s)
plot(Q, add=T, cex=2, col="blue")
```

Comparing the output of the spatial grids from the three dataframes provides quantitative data in regards to spatial relationships between the samples and wells. The quantitative spatial information is useful in conjunction with the statistical analysis of the data.

STATISTICAL ANALYSIS PROCESS

The chemometric statistical analysis for comparing the GAMA data and produced waters data consists of five steps:

1. Check for common variables (i.e. chemical constituents) between the two dataframes;
2. Check the common variables for normality, if the data are not normally distributed log transform the data;
3. Run the variables through an Analysis of Variance (ANOVA) test to assess the individual variable's significance in regards to the variance within the two datasets and assess the frequency of data points for the significant common variables from the ANOVA tests;
4. Apply a Principle Component Analysis (PCA) on the most populated variables to assess which variables most strongly characterize each type of groundwater; and,

5. Take the data from the PCA analysis and run it through a Partial Least Squares – Discriminate Analysis (PLS-DA) to predict if the GAMA samples show a produced water signature. The purpose of the PLS-DA analysis is to determine where relationships within the two dataframes may exist.

The statistical model runs completely within the R-project workspace and provides dynamic graphics and plots to aid in the interpretation of the data. Shallow and deep represent the two types of waters, GAMA and produced waters, respectively, within the statistical model. Therefore, the preceding discussion illustrates the five steps necessary to run the multivariate statistical analysis on the two types of waters, shallow and deep.

(1) Common Variables

In order to statistically compare the two types of water, the two datasets must contain common variables. To facilitate the comparison, applying a general summary function to the chemical constituents in the dataframe prints a list of the different chemical analysis for each type of water (see Script 2 for the summary function and output). The constituent summaries are then evaluated for constituents that they have in common. Only 24 common variables are found between the two dataframes. A subset of the main dataframes containing the common constituents generates two new dataframes useful for the statistical analysis; however, the two dataframes contain a different number of observations. For example, the GAMA dataframe contains 4,773 sampled wells whereas the produced waters dataframe contains only 316 wells. Of those wells, not every well was sampled for all 24 common variables. Basically, the produced waters

dataset is the limiting dataset, in that the dataset contains a weak variety of observations. For illustration, 75% of the variables were sampled in less than 40% of the 316 wells. So, using one of the variables within that 75% population means decreasing the number of observations substantially. Decreasing the number of observations by choosing individual variables also applies to the GAMA dataset. Furthermore, constituents containing similarly high observation frequencies within both datasets provide the most reliable statistical information. Table 5 and Figure 23 show the results of the common variable analysis along with the frequencies of each variable in the two datasets.

Table 5: GAMA and Produced Waters Common Variables and Number of Samples

Variable	Produced Waters	GAMA	Variable	Produced Waters	GAMA
TL	1	30	SR	59	72
NO3	9	4280	I	62	237
BR	10	105	FE	64	1322
F	13	1916	BA	71	769
CO	18	21	K	91	1745
LI	24	86	B	123	744
MO	28	93	Na	256	2132
V	38	308	SO4	267	2101
CR	46	204	CL	267	2140
CU	47	345	MG	312	2056
MN	47	1022	TDS	316	2094
AL	48	410	CA	316	2174
		Produced	**GAMA**		
	Total Wells	316	4773		

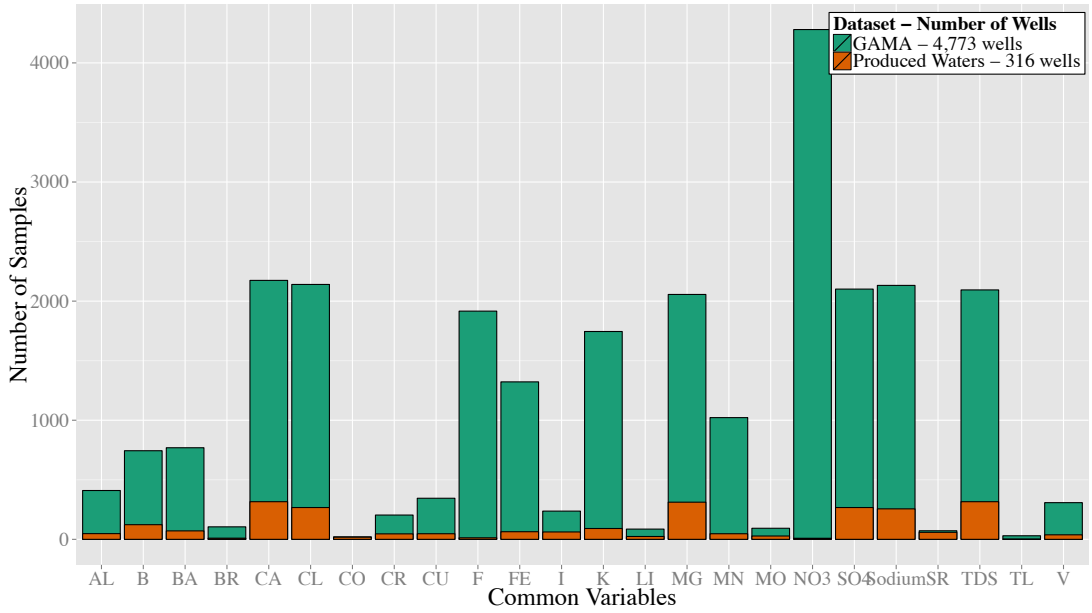


Figure 23: Frequency of samples for each common variable. Frequency of samples for each common variable from the 316 produced water wells sampled and the 4,773 GAMA wells sampled. The plot clearly indicates the smaller number of produced water samples as well as the inconsistency in sampling. Only six variables (Na, SO_4, Cl, Mg, TDS and Ca) are sufficiently populated for both produced water and GAMA samples (see Table 5; variables are ordered from least number of produced water samples to most, meaning all wells were sampled).

(2) Check for Normality

Many statistical tests require normally distributed data. To ensure the data comes from a normal distribution the random errors associated with the measurements of the data must be evaluated. The random errors represent statistical fluctuations in the analyzed samples for each measurement. In other words, the random errors represent the inability of producing the exact same result from replicate measurements of the same

sample. With that, the data were subject to three tests to evaluate the normality of the random errors. First, a Shapiro-Wilk test (`shapiro.test` function) computes the p-value, or significance level, for each of the constituents. The p-value helps determine rejection or acceptance of the null hypothesis that the random errors in the data are normally distributed. Generally, if the p-value equates to less than or equal to 0.05 then the null hypothesis is rejected and the random errors likely do not come from a normal distribution. As the p-value becomes smaller the significance becomes stronger, which in this case would indicate stronger evidence the data do not come from a normal distribution. Next, a Q-Q plot (or quantile-quantile plot; `qqnorm` function) predicts the constituents' random error normality by comparing two probability distributions, one of a sample of the analyzed data (sample quantile) and the other a statistical population of the data (theoretical quantile). A Q-Q plot showing a linear relationship indicates that the analyzed data come from a normal distribution. Lastly, generating a frequency histogram provides visual evidence as to whether the data show a Gaussian distribution or not. In addition, due to the large variance in the data, clipping the tail off the data (removing outliers) shows a more refined frequency histogram. Script 8 shows the normality test on magnesium within the GAMA dataframe.

Script 8: Check for Data Normality

```
library(ggplot2)
GAMA <-
read.csv("~/Documents/0_RESEARCH/MixingModels/Kern_MixingModel/data/CSV
s/kern_GAMA.csv")

#Test Data For Normality: Three tests --
#1)Shapiro-Wilk Normality Test - if p-value >0.05 data is likely normal
#2)QQNorm Plot - the more linear the plot the more normal
#3)Histogram Plot - visual distribution
```

```
x <- GAMA$MG

shapiro.test(x)

##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.14509, p-value < 2.2e-16

qqnorm(x)

q <- ggplot(GAMA, aes(x=MG))
q + geom_histogram(aes(fill=..count..)) +
scale_fill_gradient("count", low = "magenta", high = "cornflowerblue"

q + geom_histogram(aes(fill=..count..)) +
  scale_fill_gradient("count", low = "magenta", high =
"cornflowerblue") +
  xlim(0, 250)
```

Applying the normality tests to each of the common variables within the GAMA and produced waters dataframes indicates which variables need log-transformation prior to statistical analysis. Basically, the log-transformation normalizes the random errors associated with the data by stabilizing the variability in the data. Figure 24 shows the frequency histogram for the log-transformed magnesium data. After checking the normality of the data more sophisticated tests may be run.

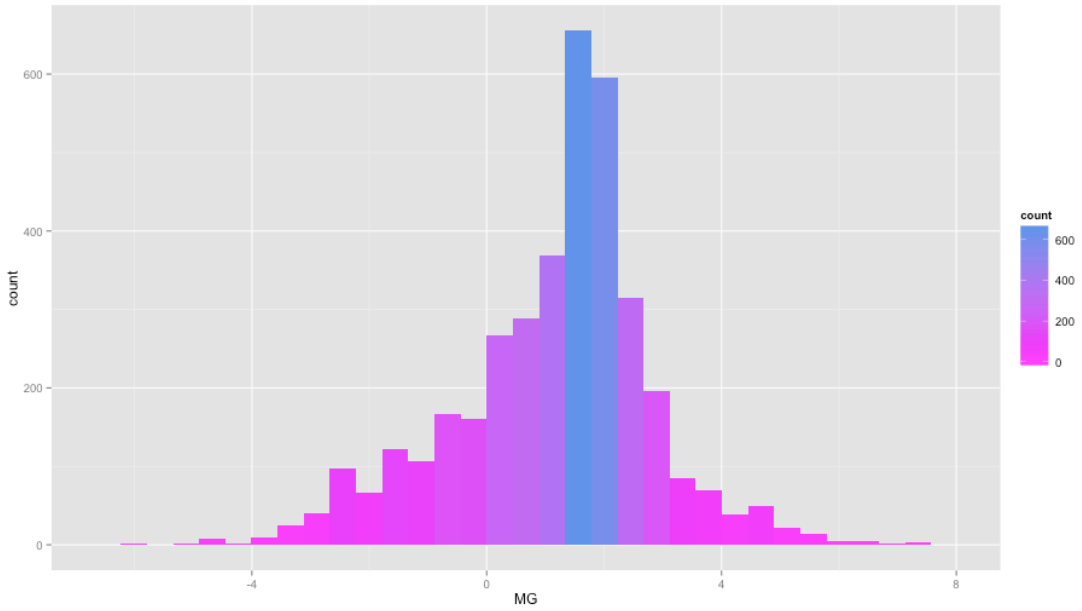


Figure 24: Frequency histogram for the log-transformed magnesium data.

Frequency histogram for the log-transformed magnesium data. The histogram shows that the log-transformation generates a more normally distributed dataset compared to the original data.

(3) Analysis of Variance (ANOVA)

Analysis of Variance separates the total variance of a population of data into components or sources. The test hypothesizes that each of the variables within the dataframe has the same mean value and alternatively that one or more of the variables has a mean below or exceeding the rest. Accordingly, the ANOVA tests the differences in a combined dataframe of the deep and shallow data. The ANOVA makes three assumptions about the data prior to analysis: the analysis is conducted by choosing random samples from the population, the random errors associated with the data comes from a normal distribution (or has been transformed), and that the variances of different

populations are equal (Davis, 2002). To evaluate the variance of a particular variable within the dataframe a one-way ANOVA produces a p-value indicating the significance level of variance for that particular variable. If the p-value equates to less than or equal to 0.05, the null hypothesis is rejected indicating the variable is significantly different between the produced water mean value and the GAMA data mean value. The lower the p-value the more significant the difference in the means. With that, R's `aov` function tests the significance between the difference in means between the produced waters and GAMA data for a particular variable.

As noted above, the first step is to create a combined shallow and deep dataframe and to log-transform the data. Next, inputting one of the variables into the script produces a box and whisker plot representing the mean value and measures of variance for each type of water, along with the ANOVA results. In addition, a table of means indicates the grand mean and the mean values for each type of water. Script 9 shows the ANOVA results for chloride.

Script 9: ANOVA

```
GAMA <-
read.csv("~/Documents/0_RESEARCH/MixingModels/Kern_MixingModel/data/CSV
s/kern_GAMA.csv")
produced <-
read.csv("~/Documents/0_RESEARCH/MixingModels/Kern_MixingModel/data/CSV
s/kern_produced.csv")

## Determination of ANOVA P-Value for shallow (GAMA)
## and deep (produced) water data for Kern County:
## Step 1: Define constituent of Concern
## Step 2: Generate Box-Wisker plot to evaluate the difference in Mean
## Step 3: Run ANOVA stats and print summary

g <- GAMA
ID.Number <- g$ID.Number
```

```
type <- g$type
CA <- g$CA
CL <- g$CL
MG <- g$MG
SO4 <- g$SO4
Na <- g$Na
TDS <- g$TDS
g <- data.frame(type, ID.Number, CA, CL, SO4, Na, MG, TDS)

p <- produced
ID.Number <- p$ID.Number
type <- p$type
CA <- p$CA
CL <- p$CL
MG <- p$MG
SO4 <- p$SO4
Na <- p$Na
TDS <- p$TDS
p <- data.frame(type, ID.Number, CA, CL, SO4, Na, MG, TDS)

z <- rbind(p,g)
type <- z$type
z <- log(z[, 3:8])
z <- cbind(type, z)
plot(CA ~ type, data=z)

results = aov(CA ~ type, data=z)
summary(results)

##               Df Sum Sq Mean Sq F value Pr(>F)
## type           1    966   966.1   793.7 <2e-16 ***
## Residuals   4508   5487     1.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 111 observations deleted due to missingness

print(model.tables(results,"means"),digits=3)

## Tables of means
## Grand mean
##
## 3.610632
##
##  type
##       Deep Shallow
##       5.29    3.48
## rep 319.00 4191.00
```

Evaluating the summary results of the ANOVA indicates the p-value (presented as Pr (>F)) in the printout shows a very low value (<2e-16 represents the lowest p-value R will report). The very low p-value means rejection of the null hypothesis and acceptance of the alternative: chloride's mean value is significantly different between deep waters and shallow waters. The ANOVA analysis provides a means of deciphering differences in important variables between deep waters and shallow groundwaters. After the ANOVA analysis, the most significantly distinct variables with the most abundant number of samples are used to generate a new dataframe for multivariate statistical tests. Six variables were included in the new dataframe: calcium, chloride, magnesium, sodium, sulfate and total dissolved solids.

(4) Principle Component Analysis (PCA)

Principle component analysis is one of the premier tools in multivariate statistics. One purpose of PCA is to depict the statistically significant information within the dataset while downsizing and simplifying the data. Prior to performing PCA the data are pre-processed, in which the data are centered and scaled (or standardized). Centering the data consists of setting the mean value of each variable equal to 0 and scaling adjusts for differences in units or order of magnitude of observations (e.g., TDS concentrations are upwards of 60,000 mg/L and magnesium concentrations are as low as 3 mg/L). Centering the data allows for the extraction of the principle components. The principle components serve as the new variables for the dataset and provide structure to a group of observations described by inter-correlated dependent variables, which can be plotted in the PCA space (Abdi and Williams, 2010).

The structure emphasizes the covariation within the dataset. For example, standard deviation and variance represent 1-dimensional statistics, which describe a variable independently from the rest of the data, whereas covariance is a measure of the spread of the data in 2-dimensions, ultimately evaluating how much the two dimensions vary with respect to one another. To illustrate this, consider the deep water and shallow water dataframes as one two-dimensional dataset. To better understand relationships between the two waters, a covariance matrix of the chemical components of the waters is produced. The matrix provides evidence about which subsets of chemicals help distinguish the two types of waters. The covariance matrix produces eigenvectors and eigenvalues useable for plotting the data's relationship in principle component space (Shlens, 2014).

Prior to the PCA a summary plot of the log-transformed data aids in understanding the data structure before applying the multivariate statistics. Function `ggpairs` generates the box and whisker plots, frequency curves and scatter plots for the six variables of interest. Next, function `prcomp` generates the PCA standard deviations, principle components with the percent of variation and eigenvalues. Considering principal component 1 (PC1), PC2 and PC3 contain over 93% of the variance they are the only PCs interpreted. A variance-covariance matrix plot (`corrplot` function), a biplot (`ggbiplot` function) and a principle component bar graph (`ggplot` function) aid in visualizing the PCA results. The biplot represents the principle component space described above containing the eigenvectors for all six variables and distinguishing

between the two dimensions (GAMA-shallow and produced waters-deep). Script 10 shows the steps to run a PCA on the water data.

Script 10: PCA

```
library(ggplot2)
library(GGally)
library(ggbiplot)

library(corrplot)
library(reshape2)

GAMA <-
read.csv("~/Documents/0_RESEARCH/MixingModels/Kern_MixingModel/data/CSVs/kern_GAMA.csv")
produced <-
read.csv("~/Documents/0_RESEARCH/MixingModels/Kern_MixingModel/data/CSVs/kern_produced.csv")

#PCA preparation and execution
#Generate DataFrame
#Conduct PCA and graph results

g <- GAMA
type <- g$type
CA <- g$CA
CL <- g$CL
MG <- g$MG
SO4 <- g$SO4
Na <- g$Na
TDS <- g$TDS
q <- data.frame(type, CA, CL, SO4, Na, MG, TDS)

p <- produced
type <- p$type
CA <- p$CA
CL <- p$CL
MG <- p$MG
SO4 <- p$SO4
Na <- p$Na
TDS <- p$TDS
w <- data.frame(type, CA, CL, SO4, Na, MG, TDS)

pca <- rbind(w, q)
pca <- pca[complete.cases(pca),]
rm(q, w, CA, CL, MG, SO4, Na, TDS, type)
```

```
type <- pca$type
pca <- log(pca[, 2:7])

pca1 <- cbind(type, pca)
ggpairs(data=pca1, colour="type")

ir.pca <- prcomp(pca, center = TRUE, scale. = TRUE)
print(ir.pca)

## Standard deviations:
## [1] 2.0241162 0.9257072 0.8114259 0.4834290 0.3291297 0.2134896
##
## Rotation:
##         PC1        PC2         PC3          PC4         PC5        PC6
## CA  -0.4192247 -0.4110629  0.03333832 -0.715722609 -0.36220372  0.1035187
## CL  -0.4459155  0.3640233  0.07493804 -0.225923147  0.67117450  0.409483
## SO4 -0.3255831 -0.3197449 -0.83371023  0.284180951  0.11578249  0.0502165
## Na  -0.4204548  0.4889655  0.01954099  0.324566236 -0.62047896  0.3056326
## MG  -0.3414037 -0.5597294  0.54185862  0.500593915  0.12134325  0.1059119
## TDS -0.4754650  0.2095073  0.06486507  0.00945831   0.07217588 -0.8488930

summary(ir.pca)

## Importance of components:
##                             PC1    PC2    PC3     PC4     PC5    PC6
## Standard deviation     2.0241 0.9257 0.8114 0.48343 0.32913 0.2135
## Proportion of Variance 0.6828 0.1428 0.1097 0.03895 0.01805 0.0076
## Cumulative Proportion  0.6828 0.8257 0.9354 0.97435 0.99240 1.0000

mkc <- cor(var(pca));
round(mkc, digits=2);

##        CA    CL   SO4    Na    MG   TDS
## CA   1.00 -0.01 -0.46 -0.27  0.85  0.05
## CL  -0.01  1.00 -0.41  0.94 -0.14  0.98
## SO4 -0.46 -0.41  1.00 -0.38 -0.45 -0.50
## Na  -0.27  0.94 -0.38  1.00 -0.25  0.94
## MG   0.85 -0.14 -0.45 -0.25  1.00  0.02
## TDS  0.05  0.98 -0.50  0.94  0.02  1.00

corrplot(mkc, type="lower", method="pie", tl.srt=45);

h <- ggbiplot(ir.pca, obs.scale = 1, var.scale = 1,
groups = type, ellipse = TRUE,
circle = TRUE)
h <- h + scale_color_discrete(name = '')
h <- h + theme(legend.direction = 'horizontal',
legend.position = 'top')
print(h)
```

```
Constituent <- c("CA", "CL", "SO4", "Na", "MG", "TDS")
melted <- cbind(Constituent, melt(ir.pca$rotation[,1:6]))
barplot <- ggplot(data=melted) +
geom_bar(aes(x=Var1, y=value, fill=Constituent), stat="identity") +
facet_wrap(~Var2); print(barplot)
```

(5) Partial Least Squares – Discriminant Analysis (PLS-DA)

The goal of partial least squares – discriminant analysis (PLS-DA) is to derive a straight line separating a 2-dimensional dataset into two distinct regions. The data plotted in the PLS-DA space will reside within one of the two regions indicating a sample belongs to either of the two groups. Basically, PLS-DA combines PCA and multiple regression to suggest where relationships may or may not exist between a set of dependent variables (shallow water versus deep water) from a large set of independent variables (chemical constituents) called predictors. Much like PCA described above, prior to analysis the data undergo centering and standardization; however, in PLS-DA the central value is not determined by calculating the variables variance but by training and testing the data (Brereton and Lloyd, 2014). To ensure the data validation and training successfully distinguish the two groups, a classification error rate and confusion matrix is produced from the PLS-DA. The classification error rate and confusion matrix of the PLS-DA utilize approximately 66.7% of the data to train the data against the remaining 33.33% and produce a report. The report indicates the percentage of observations misclassified (classification error rate) in the training along with the confusion matrix, which indicates how the misclassification happened. Training the data provides assurance that the model sufficiently and accurately segregates the data into the two regions.

In R, the `pslda` function from the `mixOmics` package produces the PLS-DA data and the `DA.confusion` function from the `RVAideMemoire` package trains the data and tests the validity of the PLS-DA. In addition, the `plotIndiv` function plots the data to visualize the discrimination between the two waters. Prior to analysis each produced water sample and each GAMA sample are given an ID number. During the PLS-DA each observation is also given an ID number, which is plotted on the `plotIndiv` plot. After the PLS-DA and plotting the data, the script produces a CSV file of the PLS-DA data, which contains the PLS-DA results and sample ID numbers. After the analysis the ID numbers allow for finding the complete, original information about the samples within the GAMA or produced waters dataframes. In addition, the `PLSDA.VIP` function produces a plot indicating the variables of importance in projection, meaning the most important variables in discriminating between the deep and shallow water. In general, if a variable contains a VIP score of > 1.0 that variable is very important in projecting the model. Script 11 shows the code for the PLS-DA of the deep and shallow groundwater.

Script 11: PLS-DA

```
library(ggplot2)
library(mixOmics)

## Loading required package: MASS
## Loading required package: lattice

library(RVAideMemoire)

## *** Package RVAideMemoire v 0.9-45-2 ***

GAMA <-
read.csv("~/Documents/0_RESEARCH/MixingModels/Kern_MixingModel/data/CSV
s/kern_GAMA.csv")
produced <-
read.csv("~/Documents/0_RESEARCH/MixingModels/Kern_MixingModel/data/CSV
s/kern_produced.csv")
```

```r
# PLS-DA

g <- GAMA
ID.Number <- g$ID.Number
type <- g$type
CA <- g$CA
CL <- g$CL
MG <- g$MG
SO4 <- g$SO4
Na <- g$Na
TDS <- g$TDS
g <- data.frame(type, ID.Number, CA, CL, SO4, Na, MG, TDS)

p <- produced
ID.Number <- p$ID.Number
type <- p$type
CA <- p$CA
CL <- p$CL
MG <- p$MG
SO4 <- p$SO4
Na <- p$Na
TDS <- p$TDS
p <- data.frame(type, ID.Number, CA, CL, SO4, Na, MG, TDS)

pca <- rbind(p, g)
pca <- pca[complete.cases(pca),]
rm(p, g, CA, CL, MG, SO4, Na, TDS, type, ID.Number)

type <- pca$type
ID.Number <- pca$ID.Number
pca <- log(pca[, 3:8])

plsda <- plsda(pca, type)
DA.confusion(plsda)

##
##  Classification error rate of a PLS-DA
##
## Distance criterion: Mahalanobis distance
## 2518 individuals used for training (among 3776, proportion = 66.7%)
## Classification error rate: 0.2%
##
## Confusion matrix:
##            Predicted group
## Real group Deep Shallow
```

```
##    Deep      56       0
##    Shallow    3    1199

palette(c("red", "blue"))
col.breast <- as.numeric(as.factor(type))
plotIndiv(plsda, ind.names = TRUE, col = col.breast)
legend(c("bottomleft"), c("Produced Water", "Shallow Groundwater"), pch
= c(16, 16),
       col = unique(col.breast), cex = 1, pt.cex = c(1.2, 1.2),
       title = "Water Type")

palette("default")

PLSDA.VIP(plsda, graph=T)

##           VIP
## SO4 1.3432288
## CL  1.0449843
## TDS 1.0405633
## Na  1.0352814
## CA  0.7625414
## MG  0.6063788

all <- cbind(type, ID.Number, pca)
write.csv(all,file="PLS_DA.csv")
```

Additional dataframes can be added to the PLS-DA to predict their relationship with the other two dataframes. To further ensure the validity of the model, the data were redundantly analyzed through PLS-DA in MATLAB. The MATLAB output consists of a confusion matrix and a correlation coefficient. Unlike R's PLS-DA classification error rate, the correlation coefficient represents the total percentage of correctly classified observations. The MATLAB PLS-DA plot rotates the data so that the Deep dimension is located in quadrant 3; however, the spatial orientation of the data does not matter considering that PLS-DA aims to create a straight line between the two groups separating them into arbitrary regions within the plot. The MATLAB plot and output data agree with the R plot and data providing evidence that two independent statistical software packages give the same result. Figure 25 depicts the MATLAB results and PLS-DA plot.

```
MODEL RESULTS
Confusion Matrix:
    Class:       TP           FP           TN           FN
   Deep          0.99441      0.00976      0.99024      0.00559
   Shallow       0.99024      0.00559      0.99441      0.00976
Matthew's Correlation Coefficient =  0.985

Confusion Table:
                                Actual Class
                              Deep      Shallow
Predicted as Deep           178         19
Predicted as Shallow          1       1927

CV RESULTS
Confusion Matrix (CV):
    Class:       TP           FP           TN           FN
   Deep          0.99441      0.00976      0.99024      0.00559
   Shallow       0.99024      0.00559      0.99441      0.00976
Matthew's Correlation Coefficient =  0.985

Confusion Table (CV):
                                Actual Class
                              Deep      Shallow
Predicted as Deep           178         19
Predicted as Shallow          1       1927
```

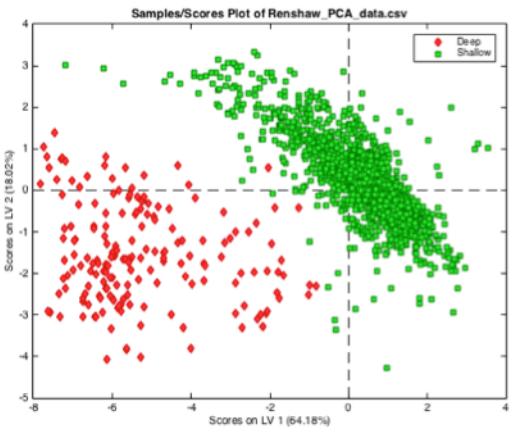


Figure 25: PLS-DA results from MATLAB analysis. PLS-DA results from MATLAB analysis. The results show good agreement with the R-project results indicating the model is accurate and works as intended. The MATLAB plot looks nearly identical to the R plot with the exception of the spatial orientation of the data points, which does not affect the interpretation of the plot or data analysis.

CHAPTER IV: Multivariate Mixing Model to Identify Oil Field Produced Waters In Shallow Drinking Water Aquifers

RESULTS

Techniques and methods of monitoring groundwater quality consistently evolve. Since the early 1990s techniques such as: multi-level sampling, contaminant plume transects, groundwater tracers and more recently groundwater age dating, etc., attempt to refine hydrogeologists' ability to understand groundwater movement and quality. Considering the constant evolution of groundwater science novel methods and techniques continually provide new insight into understanding groundwater. Likewise, new groundwater quality concerns constantly arise. For example, recent evidence shows that natural contaminants (i.e. natural occurring radioactive materials, metals, etc.) within California's groundwater basins effect groundwater quality more negatively than anthropogenic contaminants (Belitz et al., 2015). Concurrently, the development of unconventional energy resources extraction within California's groundwater basins also poses a threat to groundwater quality.

A novel approach utilizing multivariate statistics in conjunction with standard hydrogeochemical analyses provides a new method of monitoring natural contaminants that threaten groundwater used as drinking water resource in basins where oil and gas extraction takes place. In addition to the development of a novel groundwater quality monitoring tool, data gaps and spatial correlations within California's groundwater data and well stimulation treatment data were evaluated.

Data Gaps

California contains a wealth of publically available data regarding groundwater quality and oil and gas production; however, the two data resources do not coincide geographically. The easily accessible produced waters data for California contains samples from 316 conventional hydrocarbon extractions wells. The dataset contains 99 variables, 46 of which are chemical constituents. The other 53 variables consist of descriptive information such as the well depth, sample date and sample depth, latitude and longitude, data source, among others. Although the dataset appears to contain a large amount of information, there are many missing observations. To illustrate the extent of incompleteness of the dataset, Table 6 provides a summary breakdown of missing observations for some of the more useful variables.

Table 6: Missing Data in Produced Waters Dataset

Variable	**Percent of Missing Observations (n) (n/316*100)**
Latitude and Longitude	23%
American Petroleum Institute (API) Well Number	99%
Well Name or Field ID Number	18%
Sample Date	14%
Sample Depth	88%

Of the 273 produced water samples with sample dates, over 50% come from the 1950's. The most recent samples were collected in April of 1975. With that, the lack of recent results poses a concern about the sampling and analytical quality along with the accuracy of the data reporting. Figure 26 shows the distribution of the 273 by decade.

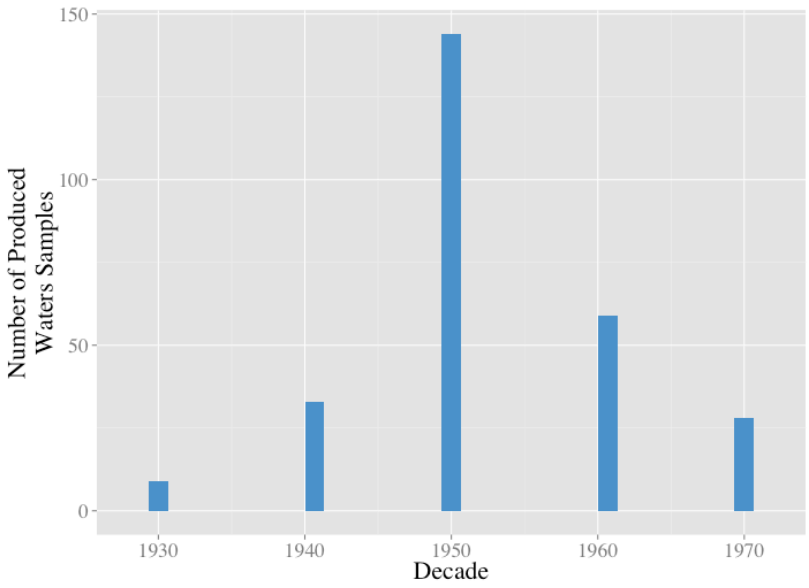


Figure 26: Statistical distribution of produced waters samples. Statistical distribution of produced waters samples by data of collection. Over 50% of the samples were collected in the 1950s and the most recent samples were collected in April 1975.

In regards to the 46 chemical variables, only 8 are populated for at least 50% of the produced waters samples and only three chemical analytes were measured in all 316 produced water samples. In fact, only 12 of the variables were measured in more than 25% (79 of 316) of the produced water samples. Table 7 shows the frequency of analysis for the 8 chemical analytes measured in at least half (158 of 316) of the produced water samples.

Table 7: Measured Analytes in at Least 50% of the Produced Water Samples

Chemical Analyte	Number of Samples the Analyte was Measured In Out of 316 Total
Total Dissolved Solids (TDS)	316
Carbonate (CO_3)	255
Bicarbonate (HCO_3)	267
Calcium (Ca)	316
Chloride (Cl)	267
Magnesium (Mg)	316
Sodium (Na)	256
Sulfate (SO_4)	267

The missing data in the produced waters data causes issues when applying the data to statistical tests, in that the tests require no missing observations. In other words, the tests require complete cases, as previously described in Chapter III. To illustrate, the ANOVA, PCA and PLS-DA scripts (Script 9, Script 10 and Script 11 in the Methods section) contain a call, `pca <- pca[complete.cases(pca),]` which effectively removes all rows of data that contain missing values. In general, the produced waters dataframe is the limiting factor in regards to interpreting interaction between deep formation water and shallow groundwater. The GAMA dataset, however also does not provide easily accessible carbonate and bicarbonate data. In order to gather bicarbonate concentrations within the Kern County GAMA data additional data analysis processes are necessary. The data analysis in this thesis did not incorporate or calculate the GAMA bicarbonate concentrations; however, including those calculations in the data analysis

may prove useful in future analysis. So, out of the 46 chemical analytes measured in the produced water samples and the over 50 analytes measured in the GAMA dataset, only 6 analytes (TDS, Ca, Cl, Mg, Na and SO_4) contain a sufficient amount of data in both datasets usable in the statistical tests.

The GAMA data also contains many non-detect analytical results and results below detection limits. For the 6 analytes described above the non-detected and below detection limit results do not exceed 5% of the total data set. In other words, 95% of all the data for the 6 analytes was reported above the detection limit. So, due to the small population of non-detect and below detection limit data, these data were not incorporated into the statistical analysis. On top of missing data, the GAMA wells and oil wells do not spatially overlap.

Spatial Analysis

The Kern County GAMA data do not spatially correlate with oil and gas development in Kern County. Basically, the majority of Kern County's population lives in the eastern portion of the groundwater basin and nearly all of the agricultural land exists in the central to eastern portions of the basin. Comparatively, most of the oil and gas production in the county takes place on the western side of the basin (large oil fields such as the Kern River field exist on the eastern edge of the basin; however, by volume, the majority of the oil and gas produced in Kern County is from fields on the western edge of the basin). All of the approved well stimulation treatment under SB4 is in oilfields on the western side of the basin. In addition to geographical differences, oil and gas production and groundwater data do not coincide because until recently, as a result of

SB4, the oil producers were not required to monitor groundwater. To evaluate the spatial relationship between oil and gas production and GAMA groundwater data, two types of plots were produced: vicinity maps depicting the locations of groundwater supply wells, oil wells and produced water samples and gridded maps of the groundwater basin with the total number of wells located within each grid. From the gridded maps, ratios of GAMA data points to oil wells are evaluated.

The vicinity maps clearly indicate the lack of groundwater supply wells and therefore groundwater data within the western portion of the basin. Considering that the most populous city, Bakersfield, is in the eastern side of the basin and the majority of the agriculture is located in the central portion of the county, nearly all of the groundwater data exist in the central and eastern portion of the groundwater basin. As a result of the lack of groundwater wells located in the oil fields there is very little information in regards to the aquifer characteristics in the western basin. Advancing groundwater monitoring wells in the western portion of the basin will allow for the analysis of groundwater flow and gradient, depth and age, sources and areas of recharge and, of course water quality and potential impacts from oil and gas operations. In all, the addition of groundwater monitoring well networks in the western basin will aid groundwater sustainability in a county with historically intense groundwater use. Figure 27 depicts the distribution of oil and gas wells relative to groundwater supply wells for the Kern County groundwater basin. Figure 28 depicts the groundwater supply wells in relation to produced water samples in Kern County.

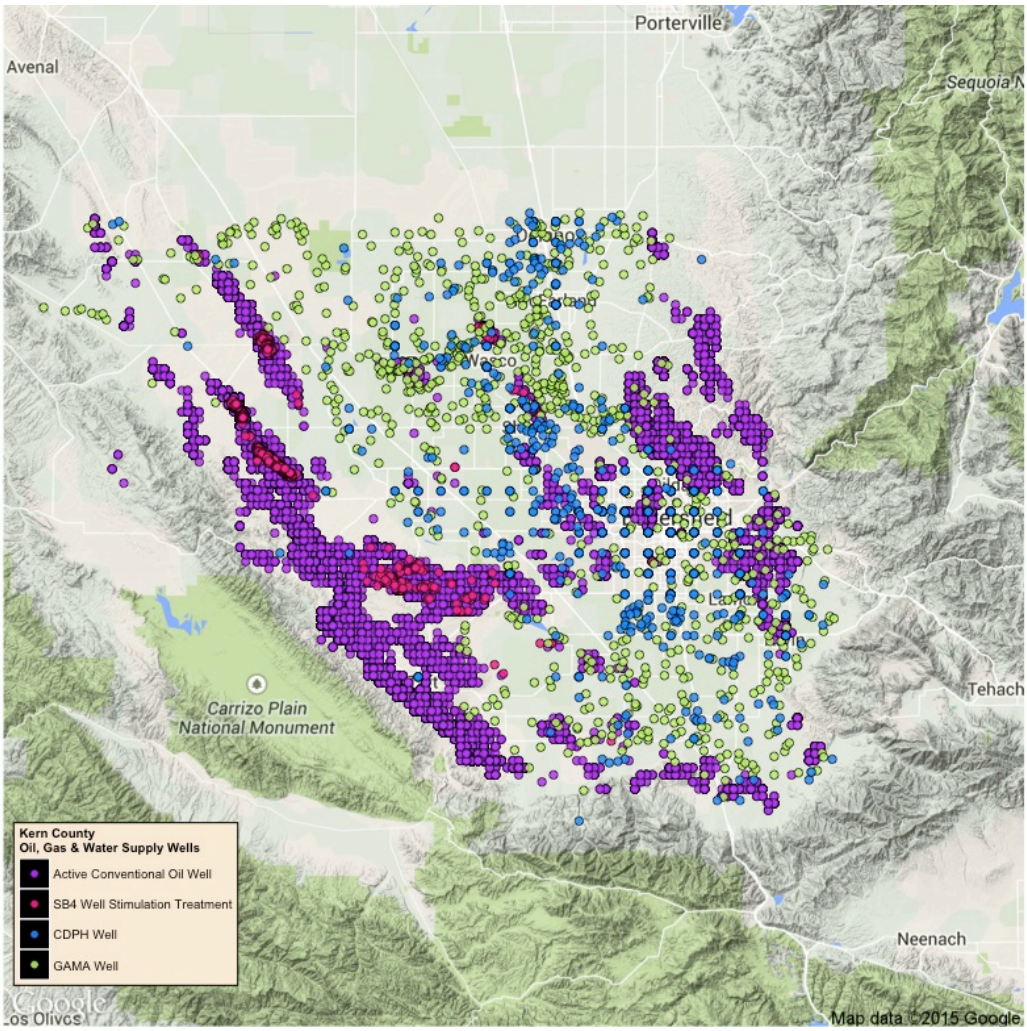


Figure 27: Zoomed in vicinity map on the Kern County groundwater basin.

Zoomed in vicinity map on the Kern County groundwater basin depicting active conventional oil wells (purple), SB4 WST approved wells (red), overlain by California Department of Public Health and United States Geological Survey groundwater supply wells (blue and green, respectively). In general, the oil and gas wells reside on the outer

edge of the groundwater basin within the foothills of the Sierra Nevada and Coast Range. Many of the GAMA wells are located in largely agricultural land and in the vicinity of Bakersfield. The lack of blue and green dots on top of the purple and red dots in the west punctuates the absence of groundwater information in this portion of the basin.

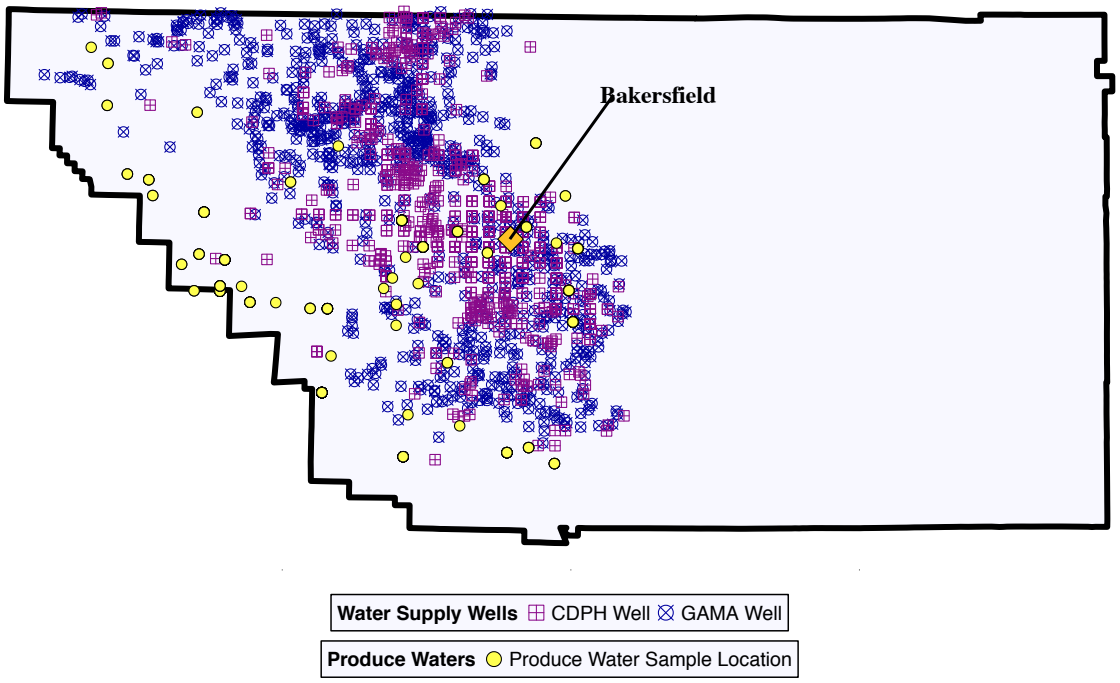
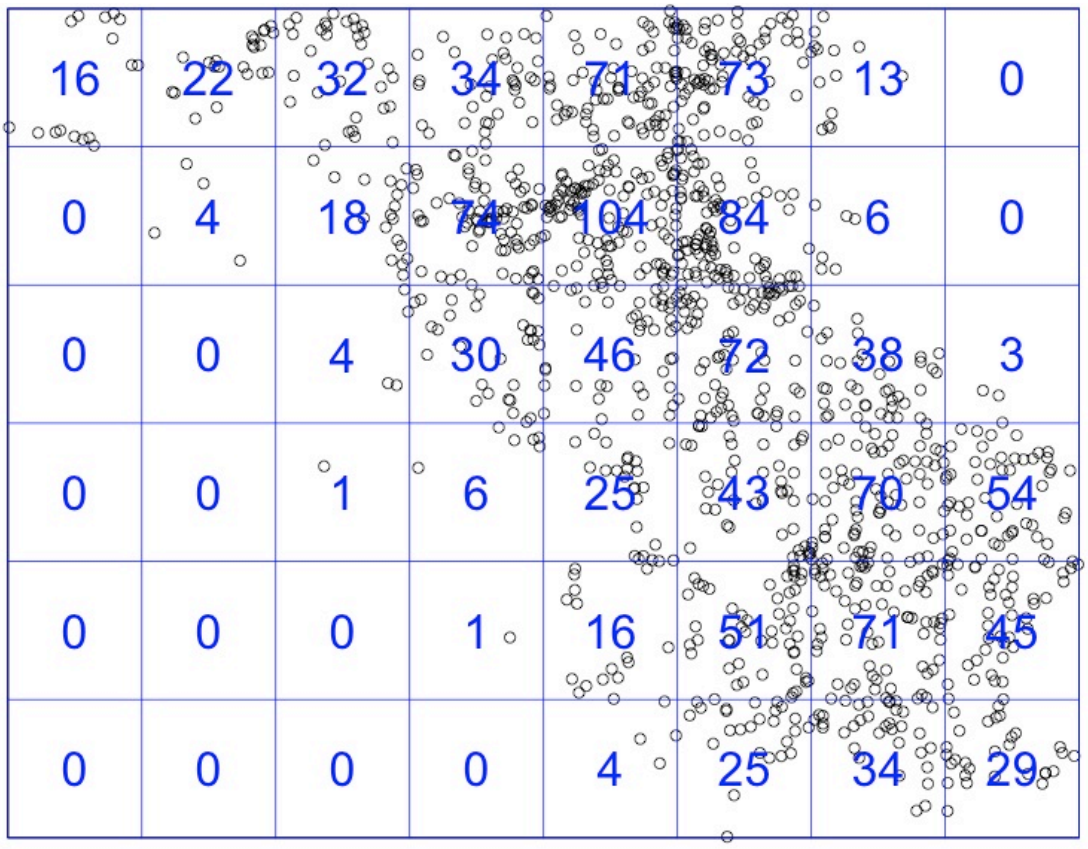


Figure 28: Vicinity map depicting water supply wells and produced water samples. Vicinity map depicting water supply wells (blue and purple) relative to Kern County produced water samples (yellow). The map shows the absence of groundwater data in the western portion of the county where many of the produced water samples are located. It also shows the random distribution of the produced water samples.

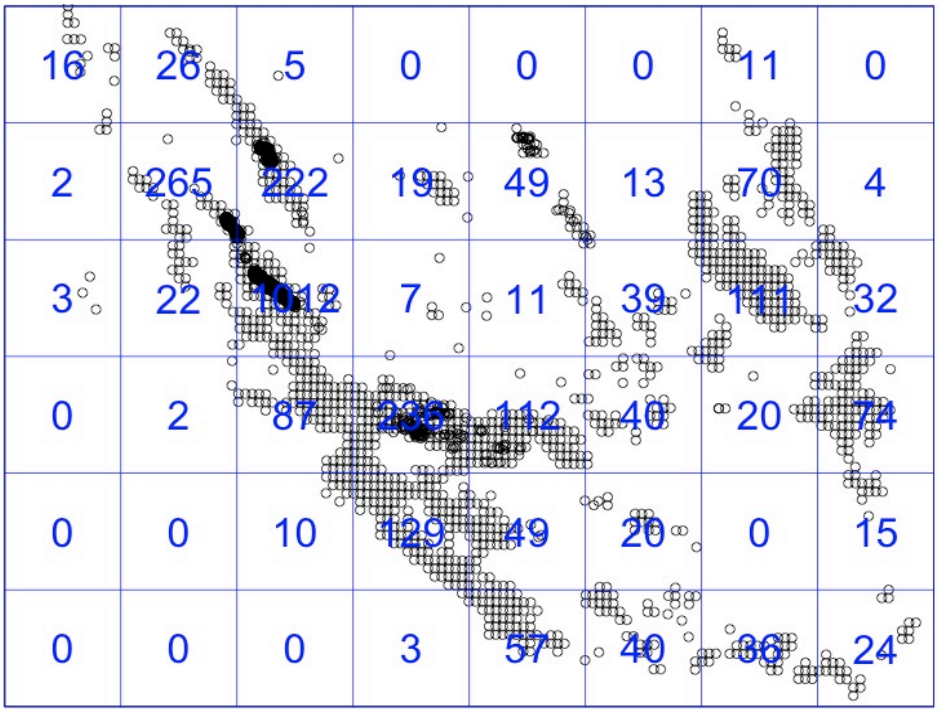
The gridded maps allow for quantitative analysis of the spatial relationships between oil and gas wells with GAMA data points and produced water data points. The

spatial areas contain an 8 by 6 grid in which each square represents approximately 94 mi^2 (see Chapter III: Spatial Analysis Processing). Three plots show the number of data points in each grid for GAMA data points, oil wells and produced water samples, respectively. Figures 29A, B and C show the gridded maps. From the gridded maps, ratios of the number of GAMA data points relative to the number of oil wells were evaluated. The squares are numbered from the upper left (square 1) to bottom right (square 48) and the first 4 columns from left are considered the western half of the grid while columns 5 through 8 comprise the eastern half of the grid.

A



B



C

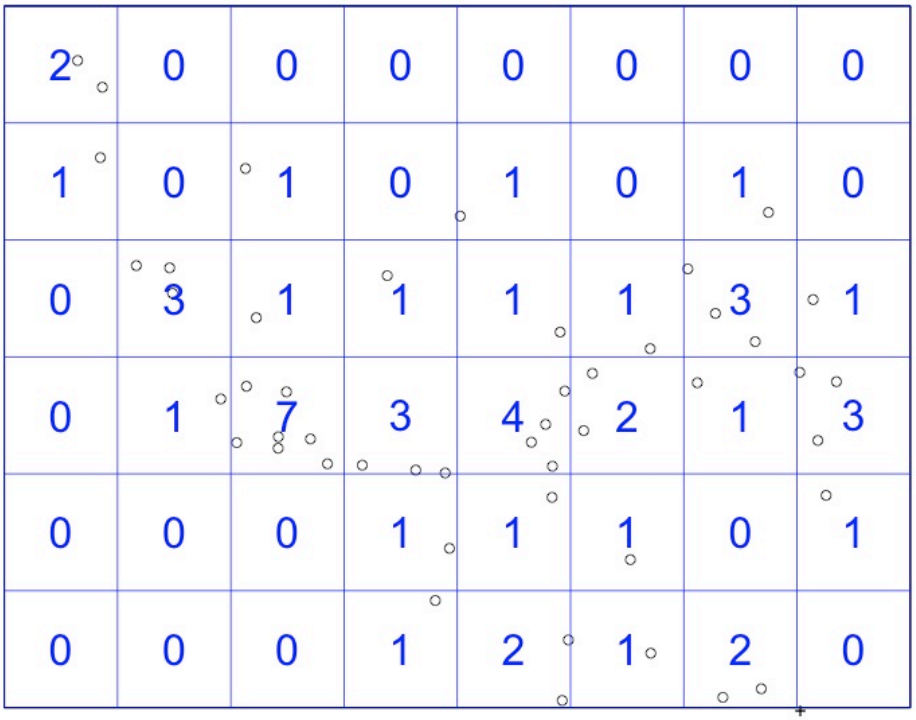


Figure 29A, B and C: Spatial grid analysis. Spatial grid analysis. A spatial grid was placed over the GAMA, oil well and produced waters data within the Kern County

groundwater basin. Each square in the 8 by 6 grid encompasses approximately 94 mi^2. A) The plot contains a total of 1219 GAMA groundwater data point locations (in other words, 1219 groundwater supply wells), in which nearly all are within the northeast-southeast squares. The grid with the greatest number of data points (grid 13) contains 104 samples. This grid is in an area of agricultural activity and the second largest city in Kern County, Delano. B) The oil well plot consists of 2793 oil well locations. The majority of the oil wells are in the southwest-northwest portion of the grid. The oil well locations are dominated by the giant Belridge, Lost Hills and Elk Hills oil fields. C) The produced waters data contains 48 locations, generally in random locations. The largest number of samples (7 in square 27) is located in the Midway-Sunset oil field, between the Belridge oil field and Elk Hills oil field.

The most abundant square in the oil well grid (square 19) contains 1,012 wells and covers majority of the South Belridge Oil Field. The same square in the GAMA grid contains 4 samples. The GAMA wells to oil wells ratio for square 19 results in 1 groundwater sample for every 253 oil wells within that 94 mi^2 area. Likewise, square 11 covers the southern portion of the Lost Hills oil field and contains 222 oil wells and 18 GAMA data samples. The ratio in the southern Lost Hill 94 mi^2 improves to 1 groundwater sample for every 12 oil wells; however, looking a little closer at the square reveals that the GAMA samples are mostly located in the northeastern-eastern portion of the square, whereas the oil wells generally reside in the western half of the square indicating the shallow groundwater samples were not collected within the extent of the oil field. The square with the largest number of oil wells in the eastern half of the grid

(square 23) contains 111 oil wells and 38 GAMA data points resulting in a 1 GAMA sample to every 3 oil wells. Further, dividing the grid down the middle to create an eastern half with 24 squares and a western half with 24 squares reveals the eastern extent of the plot contains 80% of the GAMA data points and only 30% of the oil wells, which in turn, results in 70% of the oil wells and 20% of the GAMA data points residing in the western half of the basin. The general spatial separation of the GAMA wells from the oil wells limits the ability of researchers to examine localized relationships between groundwater used for drinking water or irrigation relative to oil and gas production in California. Due to the much smaller number of produced water samples they were not incorporated into the ratio analysis but provide further affirmation that relatively few data are available for comparison of oil field waters and shallow groundwater.

The gridded spatial analysis in conjunction with review of geographic vicinity maps show the lack of spatial correlation between the GAMA data, oil wells and produced waters data. To further illustrate the inability to conduct spatial analysis of the three datasets, an attempt to utilize kriging (geostatistical tool that interpolates observed values against other observed values in a given area) to interpolate the activity of naturally occurring radioactive materials (NORM), such as radium-226 (^{226}Ra), in shallow groundwater relative to oil and gas wells, failed due to a lack of sufficient data density. As described in the Introduction and Chapter V of this thesis, ^{226}Ra is potentially a good tracer to monitor the mixing of produced waters with shallow groundwater; however, the GAMA dataset does not contain a sufficient amount of ^{226}Ra data to

interpret ^{226}Ra activities in shallow groundwater relative to oil and gas development. Figure 30 shows available radium data in the Kern County groundwater basin.

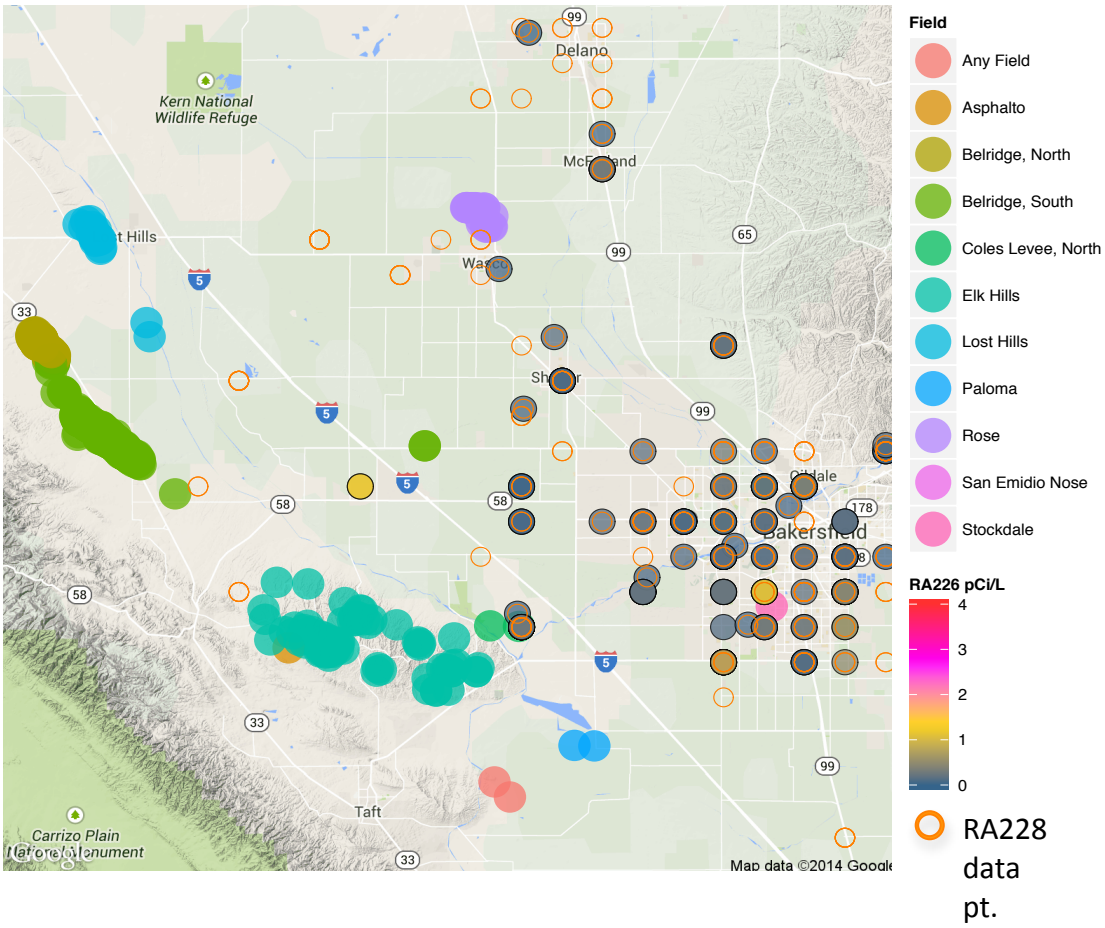


Figure 30: Comparison of radium-226, 228 samples relative to oil and gas production. Comparison of radium-226, 228 samples relative to oil and gas production in Kern County groundwater basin. The GAMA dataset does not contain a sufficient number of radium analyses to interpret spatial correlation of radium to oil and gas wells.

In general, the GAMA data do not spatially correlate with the oil and gas wells in Kern County's groundwater basin. This spatial analysis shows clearly that, if regulators

and other stakeholders want to understand the interaction between oil and gas extraction activities and the shallow aquifer system, groundwater monitoring networks need to be created in the western portion of the basin. Also, analysis of chemical constituents such as radium should be incorporated into the development and implementation of the monitoring programs.

Groundwater studies from the Marcellus Shale, Bakken Formation and Barnett Shale show the usefulness of using isotopic analyses to investigate the sources and mechanisms of shallow groundwater contamination in oil fields. Briefly, Rowan et al. (2011) described the relationship between ^{226}Ra and TDS and the value of using ^{226}Ra to ^{228}Ra ratio as a method of characterizing produced waters in shallow groundwater samples in the Northern Appalachian Basin. Although the ^{226}Ra to ^{228}Ra ratio acts as a useful produced waters tracer, Nelson et al. (2014) describe the difficulties in measuring ^{226}Ra in the chemically complex produced waters. With that, Chapter V of this thesis provides a novel method of analyzing ^{226}Ra in saline waters in hopes of incorporating the isotope into groundwater monitoring programs, but the incorporation of isotopes should not be limited to radium.

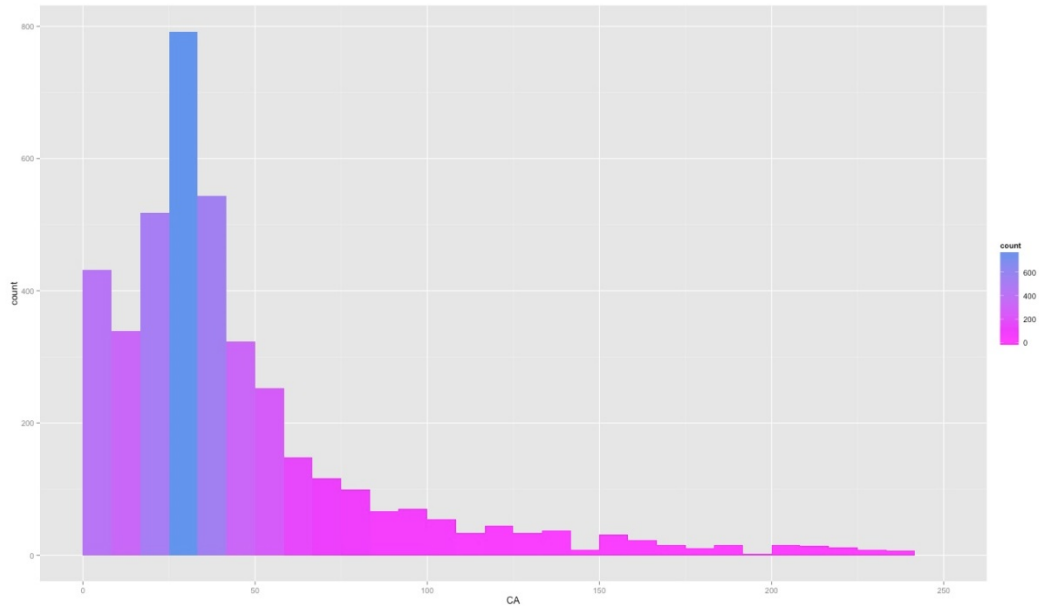
Darrah et al. (2014) analyzed 133 groundwater samples from drinking-water wells within the Marcellus and Barnett Shales for methane, carbon isotopes and isotopes of noble gases (e.g., ^{4}He, ^{20}Ne, ^{36}Ar). The isotopic data proved useful at identifying seven to eight groundwater wells near hydraulic fracturing activity that exhibited fugitive gas contamination. Also, McMahon et al. (2015) utilized a combination of inorganic chemistry, isotopes of water, noble gases, tritium, strontium isotopes, carbon-13 and 14,

among other analytes to determine the age and characterize the ambient quality of shallow groundwater domestic wells in the Bakken Formation. The data will aid in early detection of groundwater contamination from oil and gas activity in the Bakken Formation area. As previously stated and further described above, conducting comprehensive groundwater studies utilizing isotopic compositions and inorganic and organic chemistry will aid groundwater monitoring programs by providing more depth to hydrogeochemical and chemometric analysis.

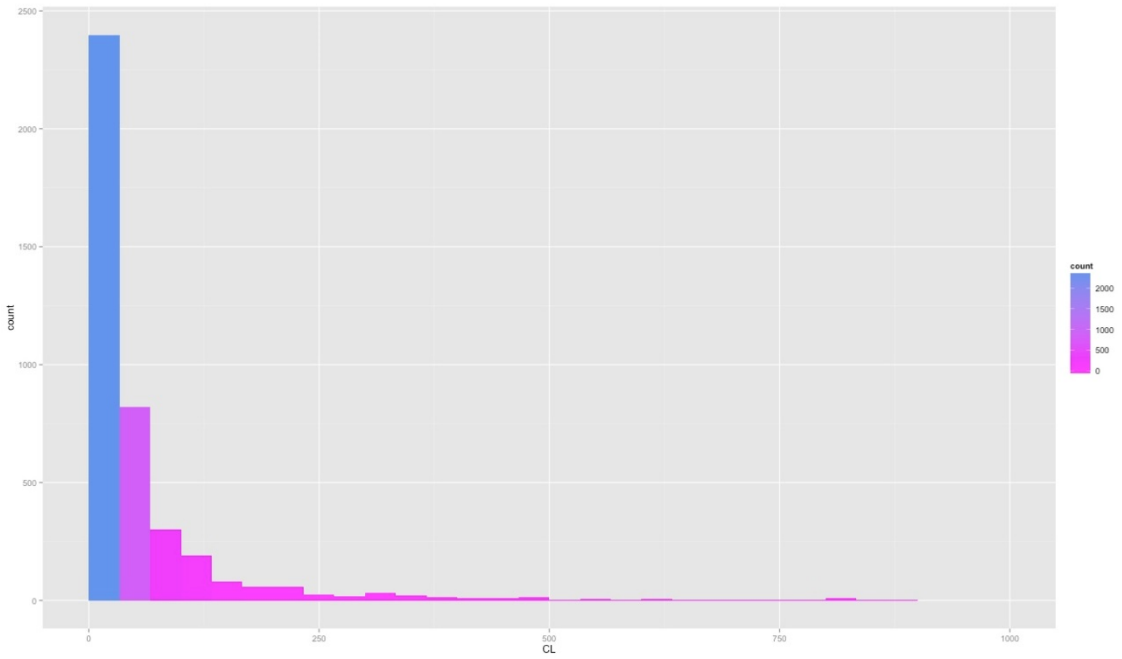
Statistical Analysis

Due to the lack of more extensive, overlapping data (discussed above in the Data Gaps section) only 6 common chemical analytes (Ca, Cl, Mg, Na, SO_4 and TDS) were evaluated in the statistical tests; however, in addition to the 6 analytes two ratios (Ca/Na and Cl/SO_4) also proved useful. To reiterate, in the statistical analysis the Kern County groundwater data and produced water data are described as the shallow population and deep population, repectively. The shallow and deep waters were tested for normality prior to any statistical tests using three normality tests: the Shapiro-Wilk test, Q-Q plots and frequency histograms. None of the analytes is normally distributed in any of the tests. To illustrate, Figure 31A, B, C, D, E and F display the histograms with statistical outliers removed for each of the 6 analytes from the shallow data. Figure 32A, B, C, D, E and F display the histograms with statistical outliers removed for each of the 6 analytes from the deep data.

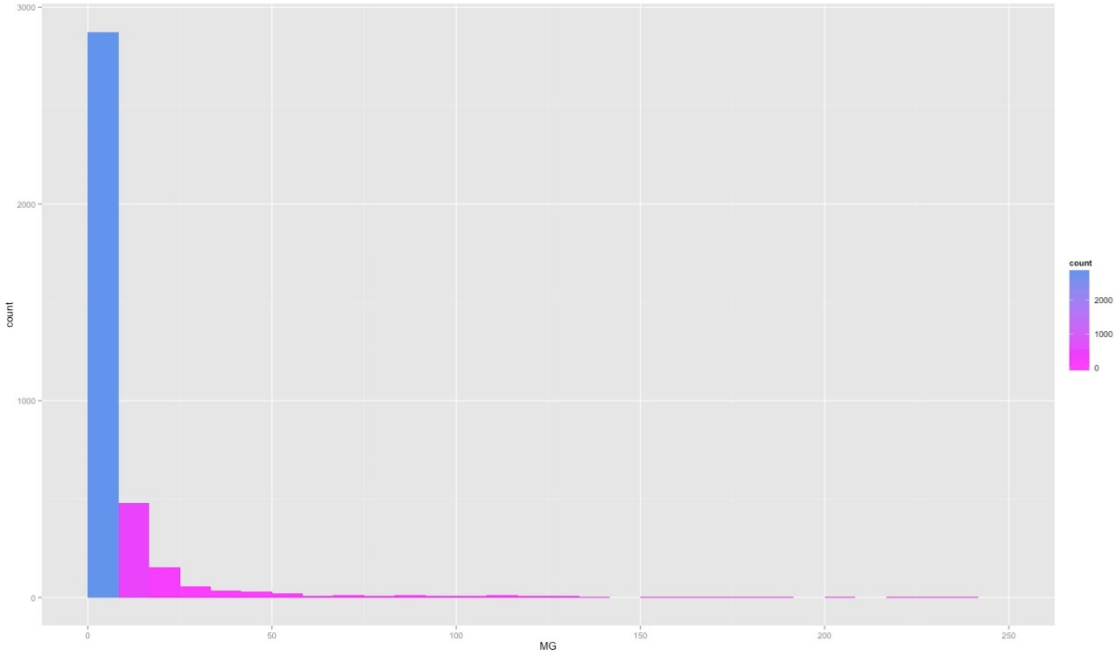
A) Calcium



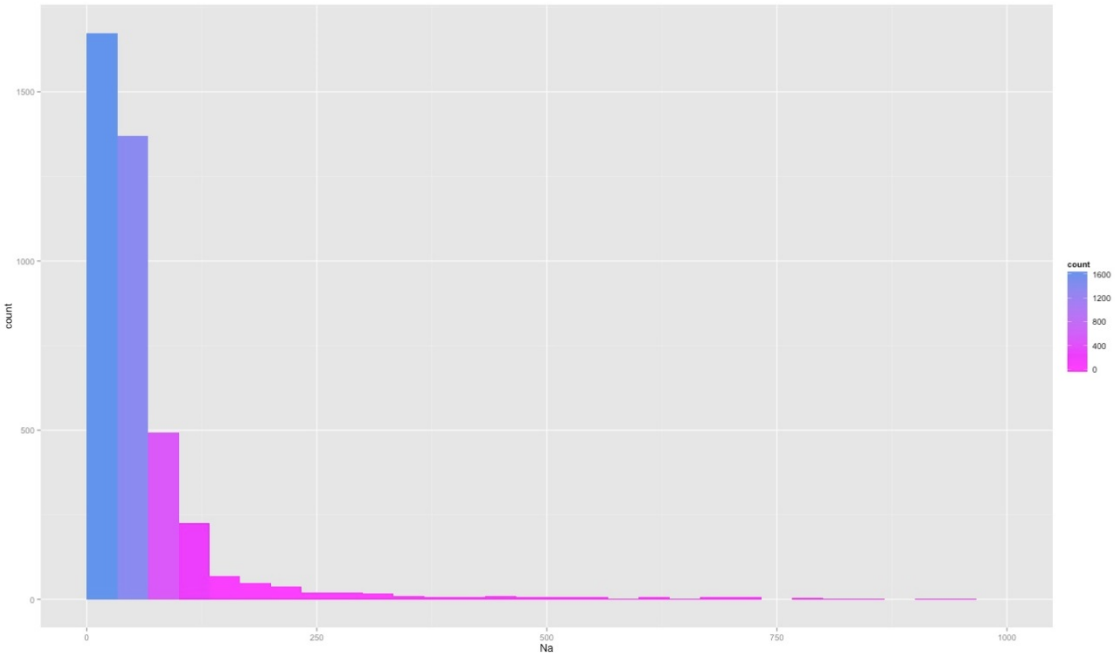
B) Chloride



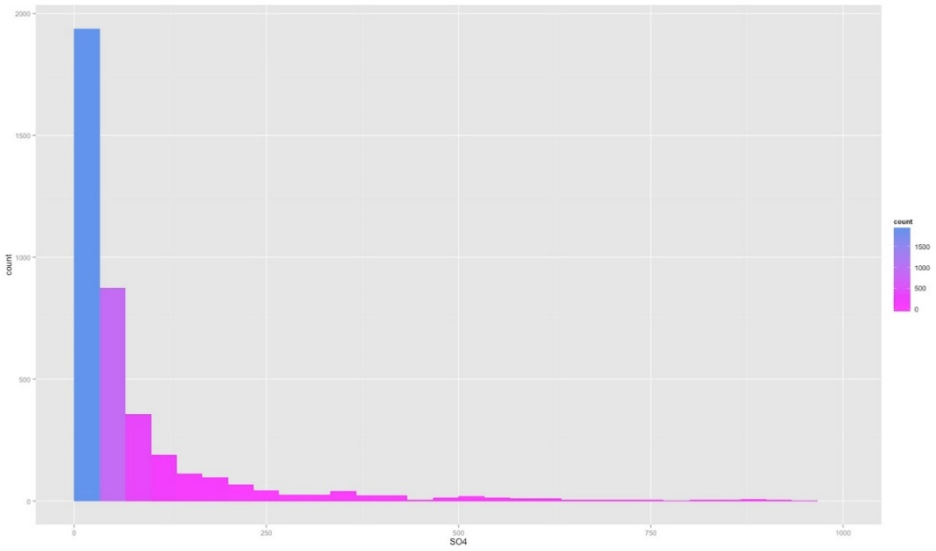
C) Magnesium



D) Sodium



E) Sulfate



F) Total Dissolved Solids

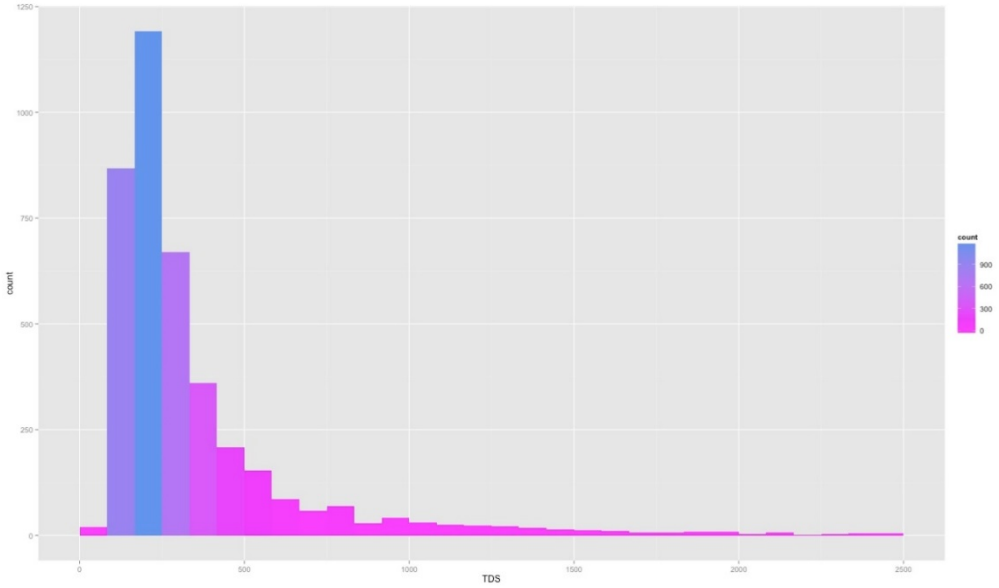
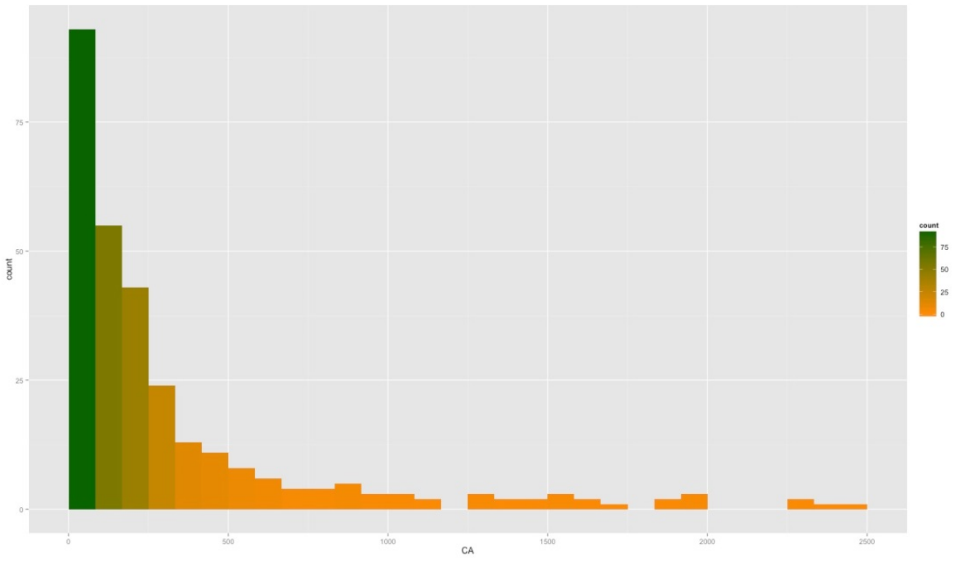


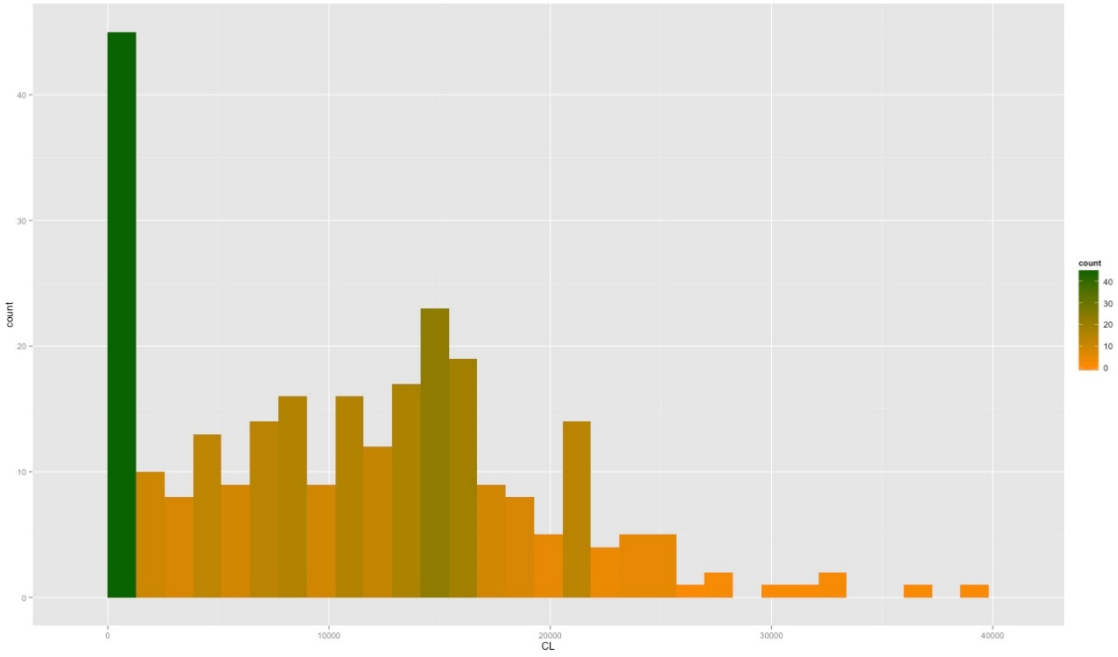
Figure 31A, B, C, D, E and F: GAMA (shallow) sample frequency histograms.

GAMA (shallow) sample frequency histograms with statistical outliers removed indicating none of the 6 analytes come from a normal distribution. All of the analytes also failed to demonstrate a normal distribution in the Shapiro-Wilk test and Q-Q Plot.

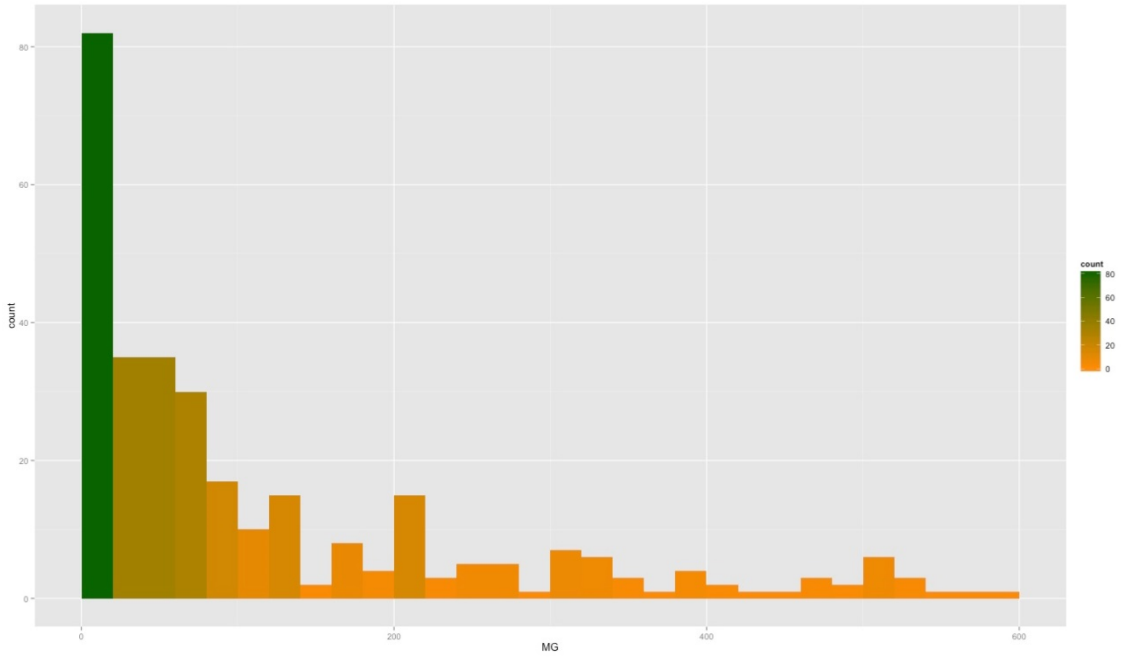
A) Calcium



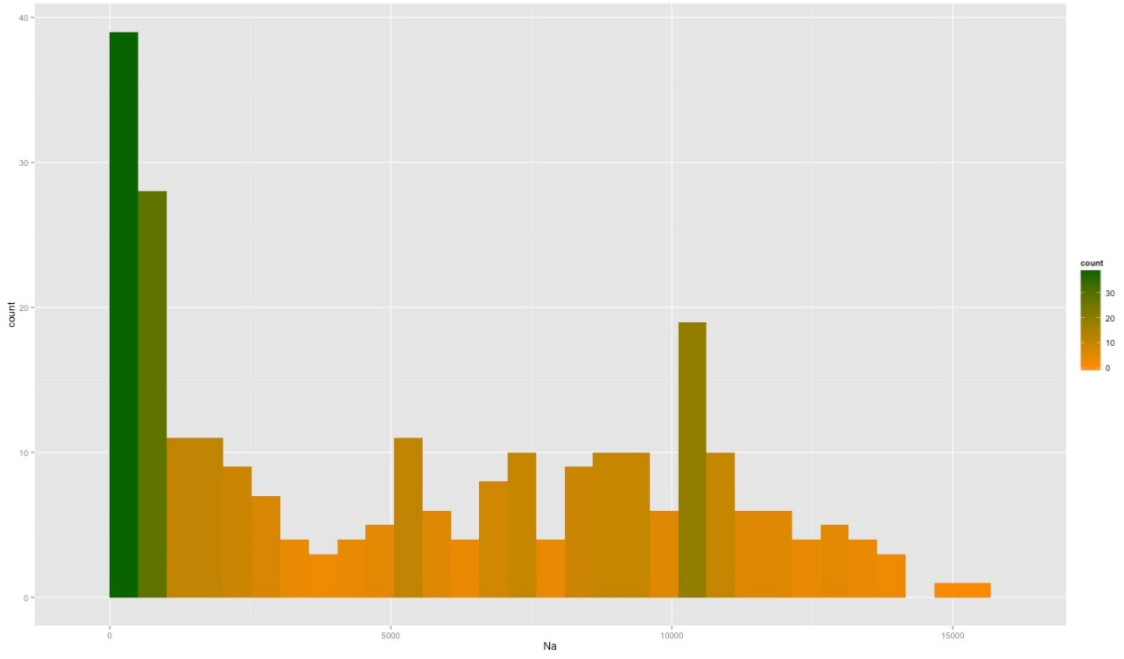
B) Chloride



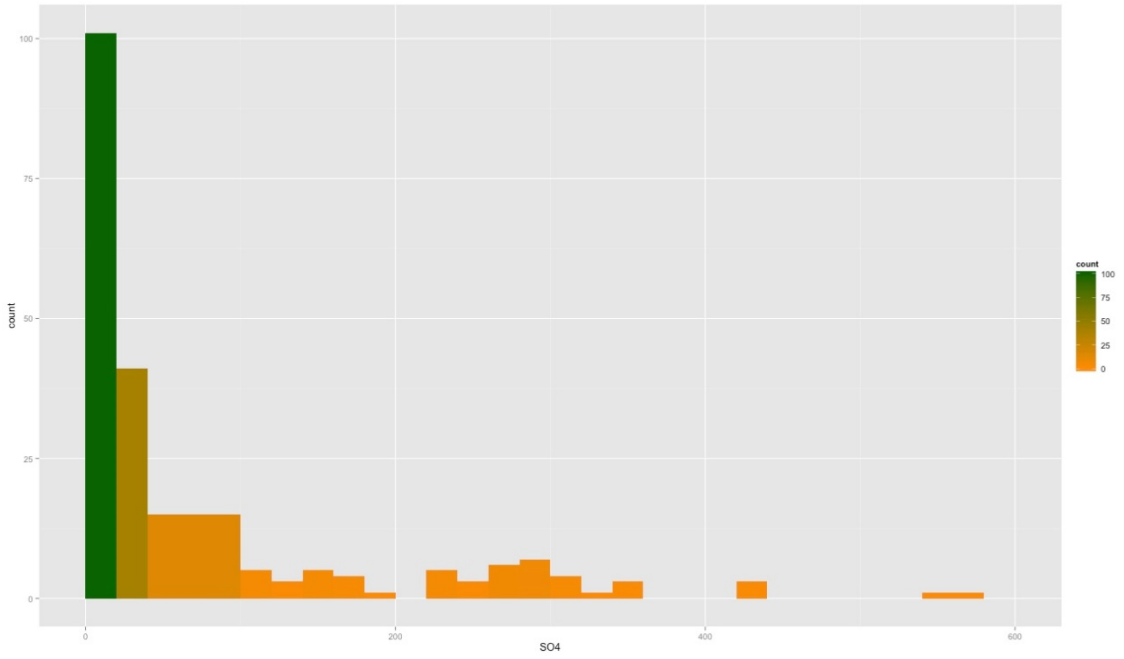
C) Magnesium



D) Sodium



E) Sulfate



F) Total Dissolved Solids

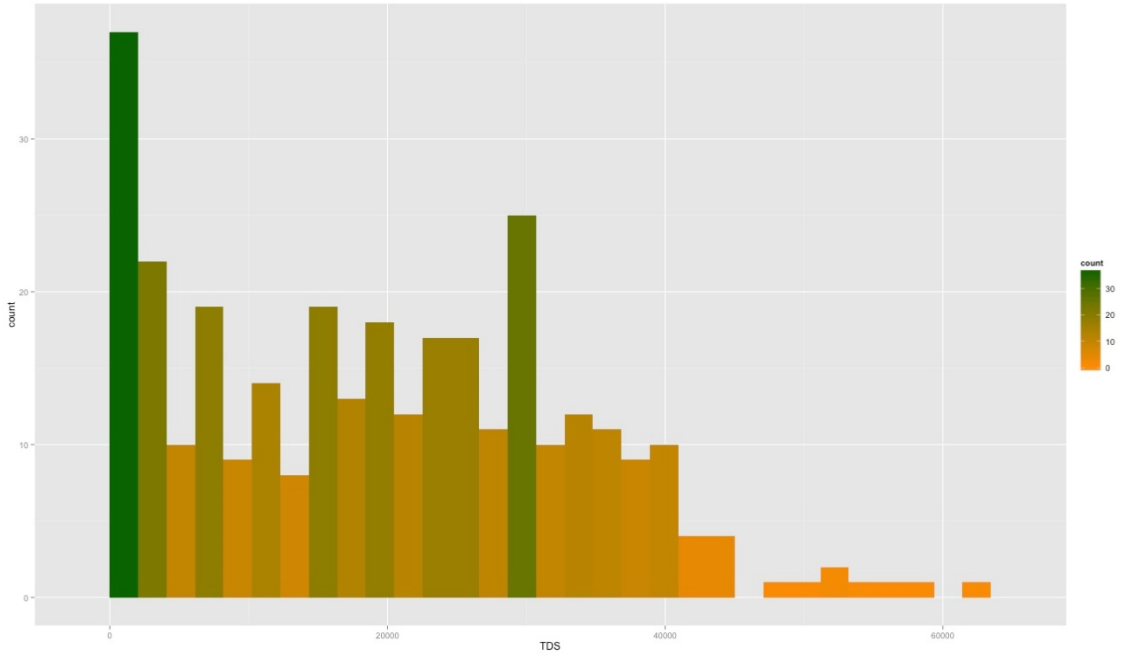


Figure 32A, B, C, D, E and F: Produced waters (deep) sample frequency histograms. Produced waters (deep) sample frequency histograms with statistical

outliers removed indicating none of the 6 analytes comes from a normal distribution. All of the analytes also failed to demonstrate a normal distribution in the Shapiro-Wilk test and Q-Q Plot.

Due to the non-normal distribution, the data undergo a log-transformation prior to statistical analysis. Using the methods described in Chapter III: Statistical Analysis Process, all 24 common variables between the shallow data and deep data were log-transformed and run through an ANOVA test to identify the significance of each variance (p-value) between the two waters. Although all 24 were run through the ANOVA test, the focus of the analysis was on the 6 primary constituents; however, analyzing all 24 proves that other variables, such as boron (B), barium (BA), iodine (I) and potassium (K), for example, provide strong significance of variance and should be further evaluated for inclusion in future groundwater monitoring programs (Table 8). All of the analytes showed high significance of variance, with the exception of nitrate (NO_3), indicating that they are good chemical indicators for segregating the produced waters from the shallow groundwater, but the 6 analytes of interest show extremely strong significance of variance. Table 8 shows the ANOVA p-value results for all 24 common analytes. Figure 33 depicts box plots for the 6 primary analytes indicating the variance between the shallow and deep populations.

Table 8: ANOVA Results

Constituent	P Value
AL	2.00E-16
B	2.00E-16
BA	2.00E-16
BR	1.03E-15
CA	2.00E-16
CL	2.00E-16
CO	5.15E-05
CR	2.00E-16
CU	3.12E-04
F	2.78E-11
FE	2.00E-16
I	2.00E-16
K	2.00E-16
LI	8.82E-08
MG	2.00E-16
MN	2.00E-16
MO	2.00E-16
NO3	5.05E-01
Na	2.00E-16
SO4	3.41E-13
SR	2.00E-16
TDS	2.00E-16
TL	1.90E-02
V	2.00E-16

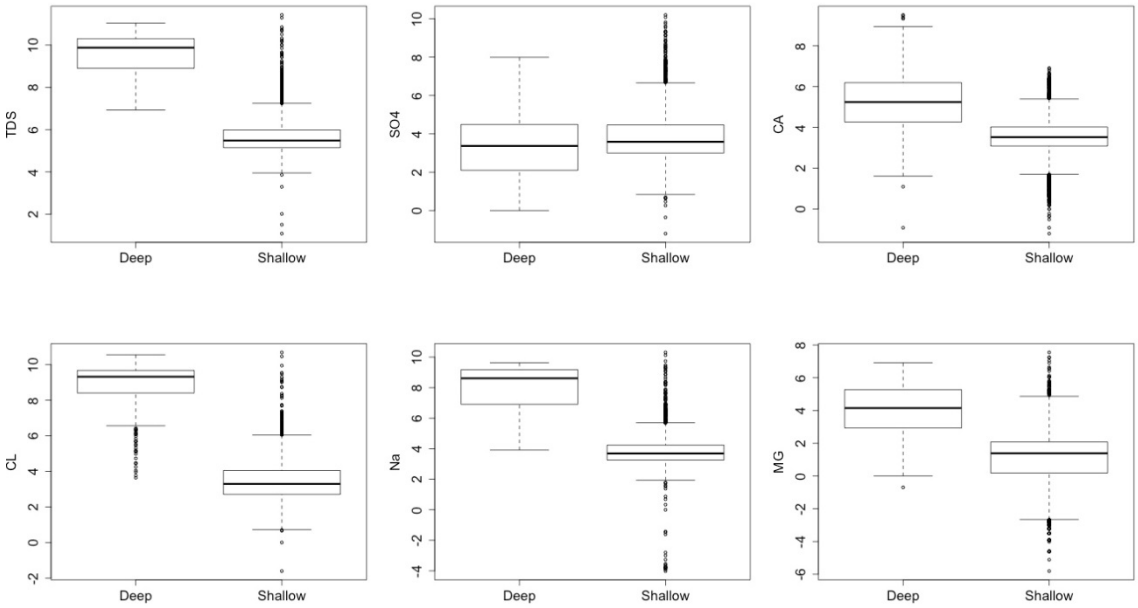


Figure 33: Box plots showing the variance between the two water populations. Box plots showing the variance between the two water populations, deep and shallow, for the 6 primary analytes. TDS and Cl have the largest differences in mean values.

The ANOVA results provide further evidence that the 6 analytes of interest will effectively segregate the two types of water due to their low p-values indicating a very high significance of variance. To further evaluate the 6 analytes on the two data sets PCA was carried out on the data to test the structure of the data, as described in the PCA section of Chapter III. Prior to the PCA a summary plot of the log-transformed data provides descriptive information about the data (Figure 34). The summary plot indicates Na and Cl are strongly, positively correlated for the shallow, deep and combined populations (the deep water contains the strongest correlation at .951). Both Na and Cl correlate strongly with TDS with stronger correlations in the deep water than in the shallow water. The positive relationship between Na, Cl and TDS was also noted in

Chapter II: Hydrogeochemistry section. SO4 is strongly correlated with Cl in the shallow samples, but weakly correlated in the deep and combined water. The only negative correlation is between SO4 and Mg in the deep water. Figure 35 provides a variance-covariance matrix of the combined deep and shallow data used for the PCA. The variance-covariance matrix depicts which variables show similar behavior in relation to each other in the combined dataset. For example, the positive correlation between Cl and Na indicates a very similar behavior within the combined dataset whereas SO_4 and TDS are negatively correlated meaning they do not show similar characteristics within the combined dataset. The SO_4 and TDS relationship in the variance-covariance matrix indicates that these two constituents are important for determining the principal components in the PCA. Thus, constituents with positive correlations in the variance-covariance matrix will show similar relationships between the two waters in the PCA whereas negatively correlated do not represent the same data population in the PCA.

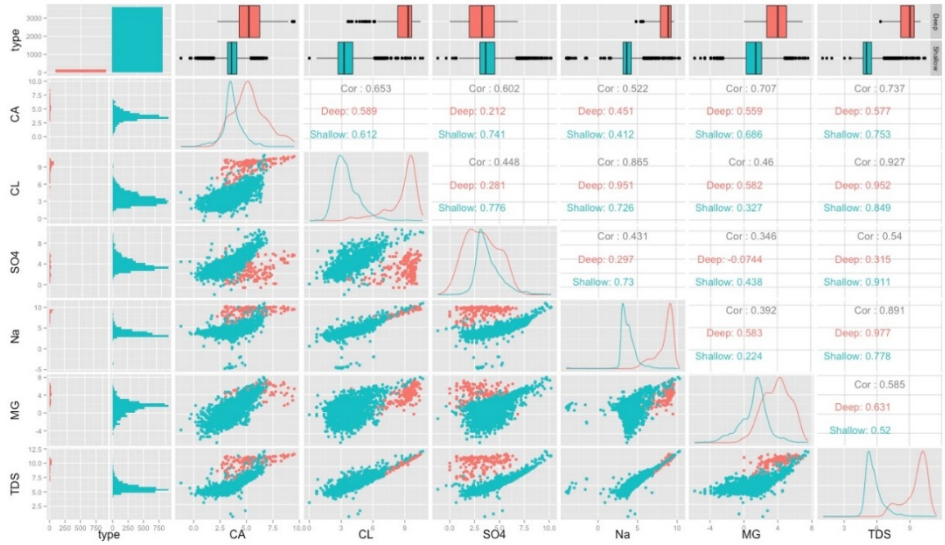


Figure 34: Summary of log-transformed data for the 6 analytes of interest.

Summary of log-transformed data for the 6 analytes of interest prior to PCA. The plot

provides an overview of the correlations between each analyte including the difference in mean values for each analyte in the deep (orange) and shallow (blue) waters. Box and whisker plots on the top row show the variation in the data from each water population for each constituent. The black line represents the mean concentration. The bar graph in upper left corner shows the number of samples for the deep water and shallow water, indicating a large difference in sample quantity. Below the number bar graph in the left column a sample frequency plot is provided to show distribution of the sample population. The remaining figures (scatter plot and line graph) show the relationships between the constituents distinguished by the water type. The text boxes provide three correlations between the constituents: black = correlation between the two combined populations, orange = correlation within the deep water population and blue = correlation within the shallow water population.

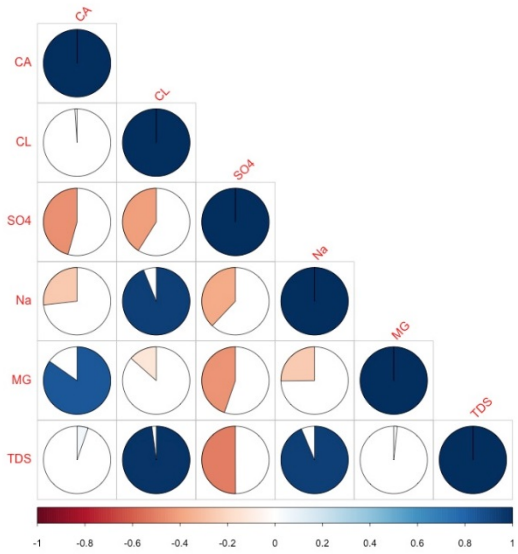


Figure 35: Variance-covariance matrix. Variance-covariance matrix showing statistical similarities and relationships between the constituents. Positively correlated

relationships (e.g., Ca-Mg, Na-Cl, Cl-TDS, Na-TDS) indicate the two constituents show similar behavior within the dataset whereas negatively correlated constituents (e.g., SO_4-TDS etc.) show opposite behavior. These correlations provide insight into which analytes likely have the same source and same chemical significance in the two types of waters. The matrix provides insight into how the constituents will act with respect to each other during the PCA. This figure and Figure 36 below are evaluated together.

In conjunction with the variance-covariance plot, a biplot (Figure 36) further aids the interpretation of the PCA results by representing the observations and variables of multivariate data in the same space. As described in the PCA Methods section (Chapter III), the PCA reduces the data into 6 principal components, in which the first three (PC1, PC2 and PC3) represent over 93% of the total variance in the dataset. With that, the biplot plots the data in a PC1-PC2 space and applies the variance-covariance relationships seen in Figure 35 to the data. The biplot of the 6 analytes indicates the deep water (produced waters) has a stronger affinity for PC2 whereas the shallow water (GAMA data) shows a weaker affinity toward PC2 and plots mostly near the origin. The eigenvectors (represented as red arrows on the plot) for chloride, Na and TDS indicate a strong relationship with PC2 in the deep water. As seen in the variance-covariance matrix, the Ca and Mg eigenvectors show similar behavior and show less of a relationship with PC2 and therefore the deep water. In addition, the SO_4 eigenvector (bottom arrow on plot) shows the most dissimilarity with all of the other 5 analytes and the deep water (also observed in Figure 35). The biplot information tends to agree with the basic hydrogeochemical interpretation of the data, in that produced waters typically

consist of formation waters derived from concentrated seawater brines of very high salinity (see Chapter II: Hydrogeochemistry section). Also, shallow groundwater typically is Ca-Mg-bicarbonate water type and, due to the anoxic nature of the formation waters, sulfur is expected to be largely in reduced form in the deep waters (Esser et al., 2015).

Figure 36: Biplot of the PCA results. Biplot of the PCA results with PC1 on the x-axis and PC2 on the y-axis. PC1 represents 68.3% of the total variance in the dataset and PC2 represents 14.3% of the total variance equaling a total of 82.6% of the variance in the shallow and deep water data. The red data points represent the deep water (produced waters) and the blue points represent the shallow water (GAMA data). Ellipses representing the 95% confidence interval (CI) help visualize the separation of the two populations. The variation within the deep water data tends to reside more heavily in PC2 whereas PC1 contains most of the shallow water variance. The top three

eigenvectors (Na, Cl and TDS, respectively) show a stronger affinity toward PC2 and the deep water. The calcium and magnesium eigenvectors are nearly identical and show less of a relationship with PC2. The bottom eigenvector (SO_4) indicates the most dissimilar behavior with regards to the deep water.

To further examine the PCA results, a PCA bar graph indicates the structure and helps provide meaning to the principle components. Analyzing Figure 37 along with Figure 36 and the PCA results indicates the variation in PC1 comes from the same source (as shown by all variables in the negative), which is likely the hydrogeochemisty related to the shallow groundwater. Further, the PC2 bar graph shows a segregation between the data, in that TDS, Cl and Na are positive whereas SO_4, Ca and Mg are negative. The difference in the bar graph of PC2 indicates TDS, Cl and Na likely come from a different source than SO_4, Ca and Mg. Therefore, in conjunction with the hydrogeochemisty analysis presented in Chapter II, PC2 helps distinguish between deep (TDS, Cl and Na) and shallow (SO_4, Ca and Mg) water. Lastly, PC3 shows SO_4 and Mg aid in understanding the remaining variance, which may be associated with produced waters (formations waters) containing more Mg relative to SO_4. With that, the plot provides further evidence that Cl, Na and TDS (olive green, teal and pink, respectively) correlate strongly with PC2 and therefore with the deep water. The bar plot also indicates that PC2 contains unique information in regards to the variance because all of the analytes show a negative correlation with PC1.

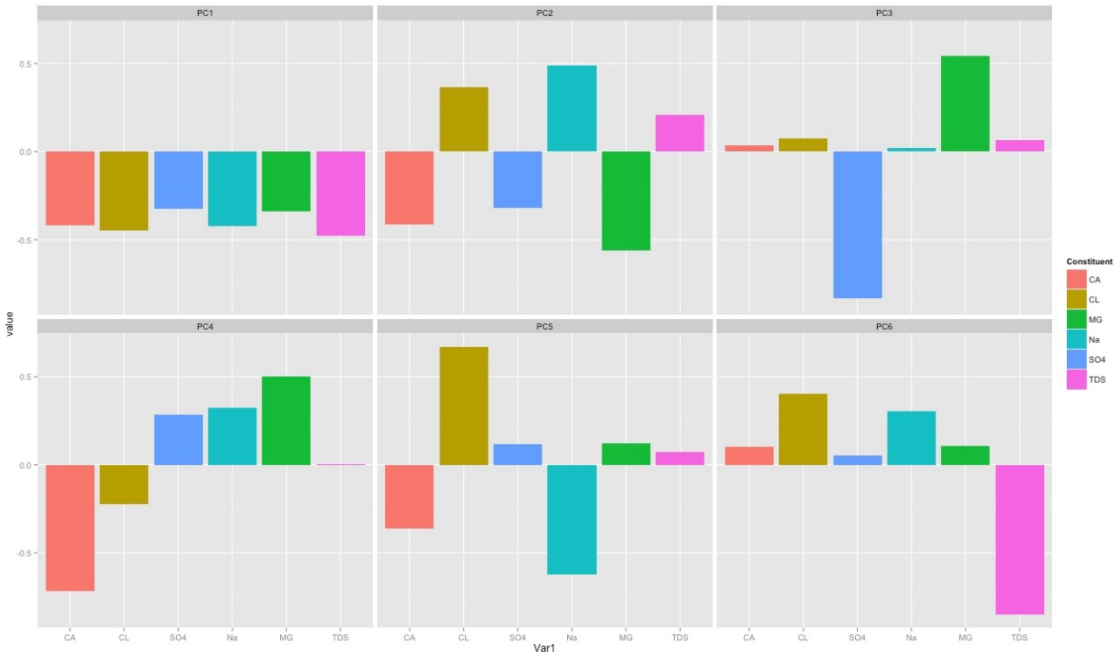


Figure 37: Bar plot of PCA results. Bar plot of PCA results. The plot shows the relationship between the principal components and the analytes. The bar graph for PC2 (top middle) indicates Cl, Na and TDS' strong correlation with PC2. Comparing these results with the biplot, which showed the deep water contained a strong correlation with PC2 provides evidence that the 6 analytes of interest can segregate the two waters.

The four plots provide strong evidence that the 6 analytes of interest effectively segregate the two waters. Running PCA on a set of ratios (Ca/Na and Cl/SO_4) along with SO_4 and TDS also provides a strong means of segregating the two waters. Figure 38 shows the biplot from the PCA from the ratio analysis. The results indicate the deep water does not correlate with PC1 and the two waters are distinct based on the ratio analysis. In addition, the 95% CI for the shallow data is much tighter confirming that the combination of the two ratios and two analytes distinguishes the shallow groundwater

from the deep groundwater. The ratio analysis provides additional evidence that PCA effectively segregates the two waters.

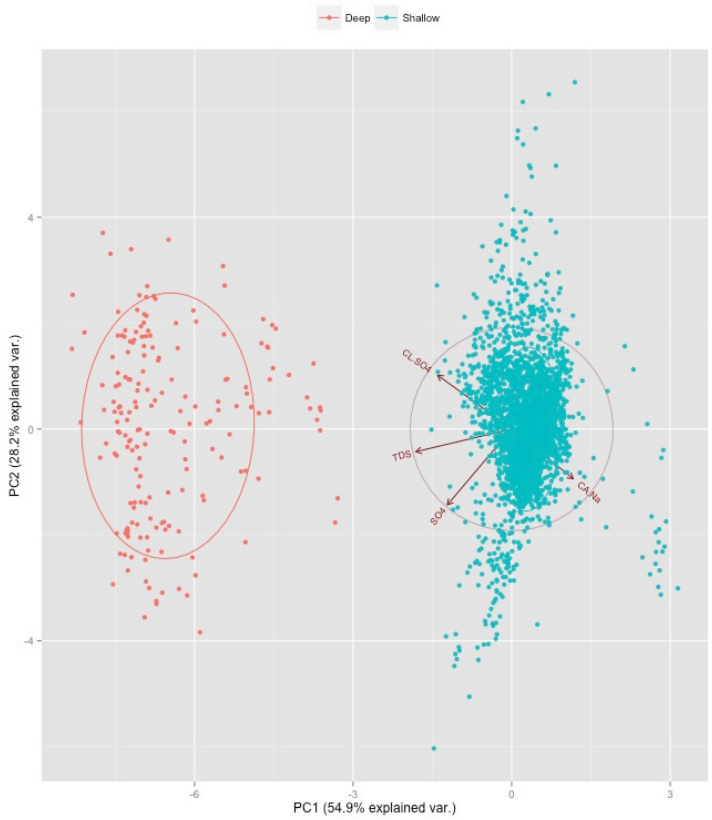


Figure 38: Biplot of PCA for the ratios Ca/Na and Cl/SO_4. Biplot of PCA for the ratios Ca/Na and Cl/SO_4 along with SO_4 and TDS. PC1 represents 54.9% of the variance and PC2 represents 28.2% for a total of 83.1% of the total variance in the data. The four variables readily distinguish the two waters. The ratios Ca/Na and Cl/SO_4 are strongly anti-correlated and likely represent the shallow water and deep water, respectively.

To predict whether an individual sample or group of samples show a produced water chemical signature, a logistic regression was performed on the data. Prior to evaluating the logistic regression model results, the model was tested for its accuracy at properly distinguishing the two waters. To test the model two subsets of the data were

generated, a testing subset and a training subset. The training subset consisted of approximately 75% of the total dataset and the testing subset consisted of the remaining 25% of the dataset. The logistic regression model was fit using the training data and the fitted model was then used to generate predictions for the testing data. After predicting the testing data a confusion matrix and classification error rate, as described in the PLS-DA section in Methods, were generated. The classification error rate was above 50% indicating the logistic regression model does not effectively fit the model and generates erroneous predictions. The logistic model likely did not fit the data due to the lack of intermediate data points. If the dataset contained water data from a deep source, shallow source and an intermediate source the model may have fit better.

After the failure of the logistic regression the data was analyzed through a partial least squares – discriminant analysis (PLS-DA) as described in the PLS-DA Methods section. The PLS-DA effectively predicts whether a GAMA sample shows evidence of a produced water chemical signature with less than a 1% error rate. The PLS-DA utilizes PCA and multiple regression to statistically distinguish the two waters by independent variables such as the 6 analytes or ratio analysis. To reiterate, the ratio analysis consists of SO_4, TDS, Cl/SO_4 and Ca/Na acting as the variables. The results clearly show the statistical test discriminates between the two waters and provides insight into which GAMA samples may show a component of the produced waters. Figure 39 provides the results of the PLS-DA for the 6 analytes of interest.

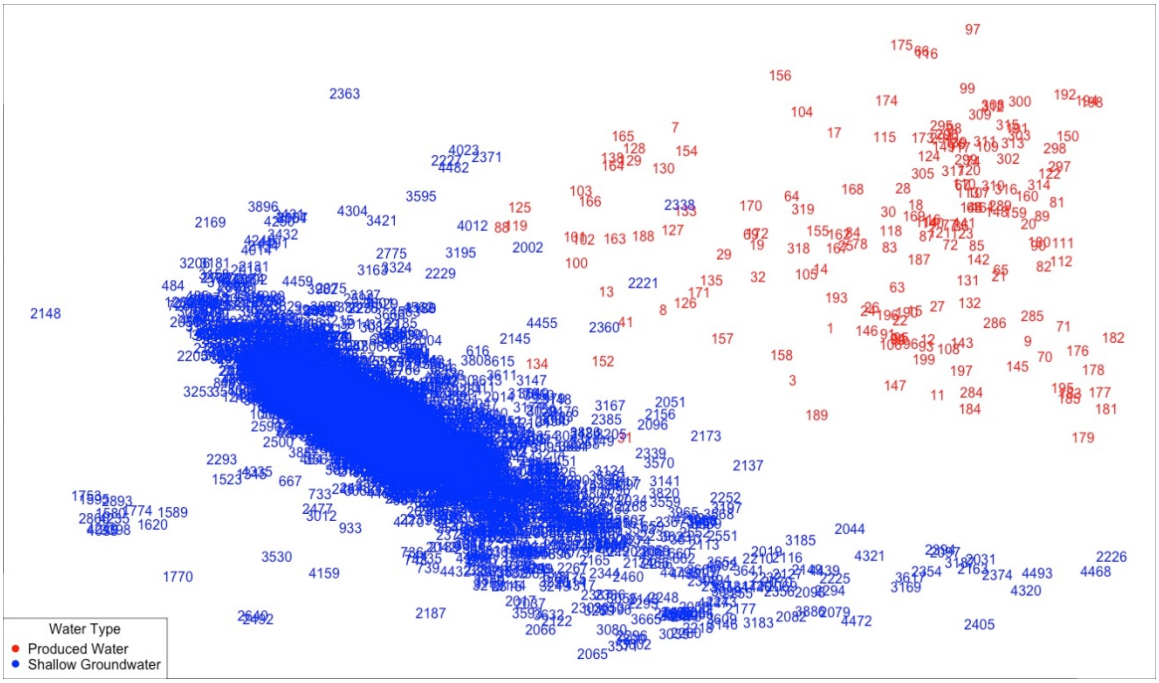


Figure 39: PLS-DA plot. PLS-DA plot effectively discriminating between the deep (red) and shallow (blue) water. The plot shows 3 to 5 (ID numbers: 2338, 2221, 2360, 2002 and 4012) samples that suggest a relationship with the produced waters, especially sample 2338, which is largely embedded in the produced waters region. Using the ID numbers, the GAMA samples can be further investigated within the GAMA dataset. In the GAMA dataset sample information such as the sample depth, coordinates and sample date can be interpreted.

When the PLS-DA is performed on the ratio analysis variables (i.e. SO_4, TDS, CL/SO_4 and CA/Na), a slightly different plot is generated. Like the PCA biplot, the results for the ratio analysis appear to refine the resolution of the two data regions, resulting in tighter data populations. In the results for the ratio analysis approximately 10 samples encroach on the produced waters region of the plot, including the 5 samples from

the original 6 analyte PLS-DA plot. The encroachment of these 10 samples on the deep region indicates they have a relatively strong statistical relationship with the produced waters data, and therefore, likely have a hydrogeochemical relationship with the deep water. As stated above, from the plot the ID numbers provide an avenue for further investigation of the samples within the GAMA dataset and spatial correlations with oil fields. To illustrate, Figure 40 provides the PLS-DA results for the ratio analysis showing the shallow water in relation to the deep water and the 'unknown' samples (in green) with an identification number representing the data point.

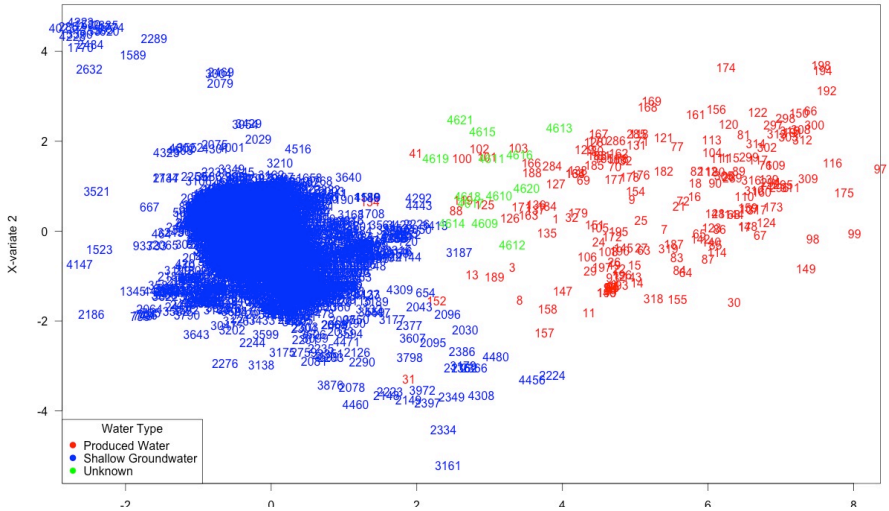


Figure 40: PLS-DA results for the ratio analysis. PLS-DA results for the ratio analysis. As seen in the PCA, the ratio analysis consists of two ratios (Ca/Na and Cl/SO_4) along with TDS and SO_4. The results plot indicates up to 10 samples that suggest a relationship with the deep water and approximately 5 to 7 of those 10 show a fairly strong relationship with the deep water (shown in green on the plot). The results for the 10 samples warrant further investigation into the sample location, sample depth and other descriptive information within the GAMA dataset.

In the ratio analysis plot, shallow water samples 2338, 2221, 2360, 2002, 4012 and approximately 5 others suggest a relationship with the deep water potentially indicating a produced water signature. After evaluating the samples in the GAMA database and plotting the points on a map we see the five GAMA data points demonstrate a spatial correlation with the Lost Hills oil field (Figure 41). Therefore, the results of the PLS-DA show these GAMA samples contain a statistical relationship with the produced waters and a spatial relationship with one of the more prominent oil fields in Kern County. With that, the statistical model provides evidence of potential produced waters mixing with shallow groundwater; however, additional investigation into the physical setting of the samples is needed.

Figure 41: Map of samples suggesting a relationship with the produced waters. Google Earth map showing the locations of the 5 GAMA samples that suggest a statistical relationship with the produced waters. The 5 GAMA samples are located downgradient of and near the large Lost Hills oil and gas field (outlined in red).

DISCUSSION AND IMPLICATIONS

The statistical analysis of the produced waters and GAMA data indicate 6 common chemical analytes (Ca, Cl, Mg, Na, SO_4 and TDS) effectively segregate the deeply sourced formation waters from the shallow groundwater used for drinking water. The statistical results coincide with general hydrogeochemical interpretation of the two waters, as shown by the PCA results. Even further, the PLS-DA results show if relationships and similarities exist between the dependent variables (the deep and shallow water) utilizing the independent variables (chemical analytes). The PLS-DA also provides the opportunity to evaluate individual shallow samples that may suggest a statistical relationship with the deep water as seen in Figure 40. In turn, the PLS-DA segregates the data into individual samples or groups of samples that warrant more in depth investigation as seen with the 5 samples in Figure 41. In all, the statistical analysis provides a method of tying the chemical data, spatial data and physical data together. To further illuminate the implications of this novel method on groundwater monitoring programs the 5 possibly mixed samples are discussed in detail below.

Figure 39 and Figure 40 show 5 shallow groundwater samples contain a statistical relationship with the deep waters indicating potential mixing of deep and shallow groundwater. Plotting the samples on a map shows a spatial relationship with the Lost Hills oil field, which is one of the largest oil fields by volume of oil produced in Kern County. In addition, the Lost Hills oil field contains the third most approved well stimulation treatment permits in California. With that, investigating the physical setting of the 5 wells and the Lost Hills may provide more detailed information in regards to the

mixing of deep and shallow water. The Introduction section briefly discusses methods of geofluid migration, such as: density heterogeneities, aseismic or coseismic fault activity, subsurface fracture systems, stratographic juxtaposition and pressure release or buildup. The Lost Hills oil field likely has undergone or currently undergoes all of the physical processes conducive to geofluid migration.

The Lost Hills oil field contains a blind thrust fault that extends from the Kettleman Hills in the north to the southern extent of the Lost Hills anticline. The blind thrust is associated with the Idria-Lost Hills fold belt that runs parallel to the San Andreas fault accommodating the normal shortening associated with the fault. The 110-km-long blind thrust has ruptured at least 3 times since 1980 ranging between $5.4 \leq Mw \leq 6.5$ and is associated with the Coalinga, Kettleman Hills and Lost Hills anticlines (CCST, 2014b; Stein and Ekstrom, 1992). Figure 42 shows a map view of the fold belt and thrust fault along with geologic cross-sections.

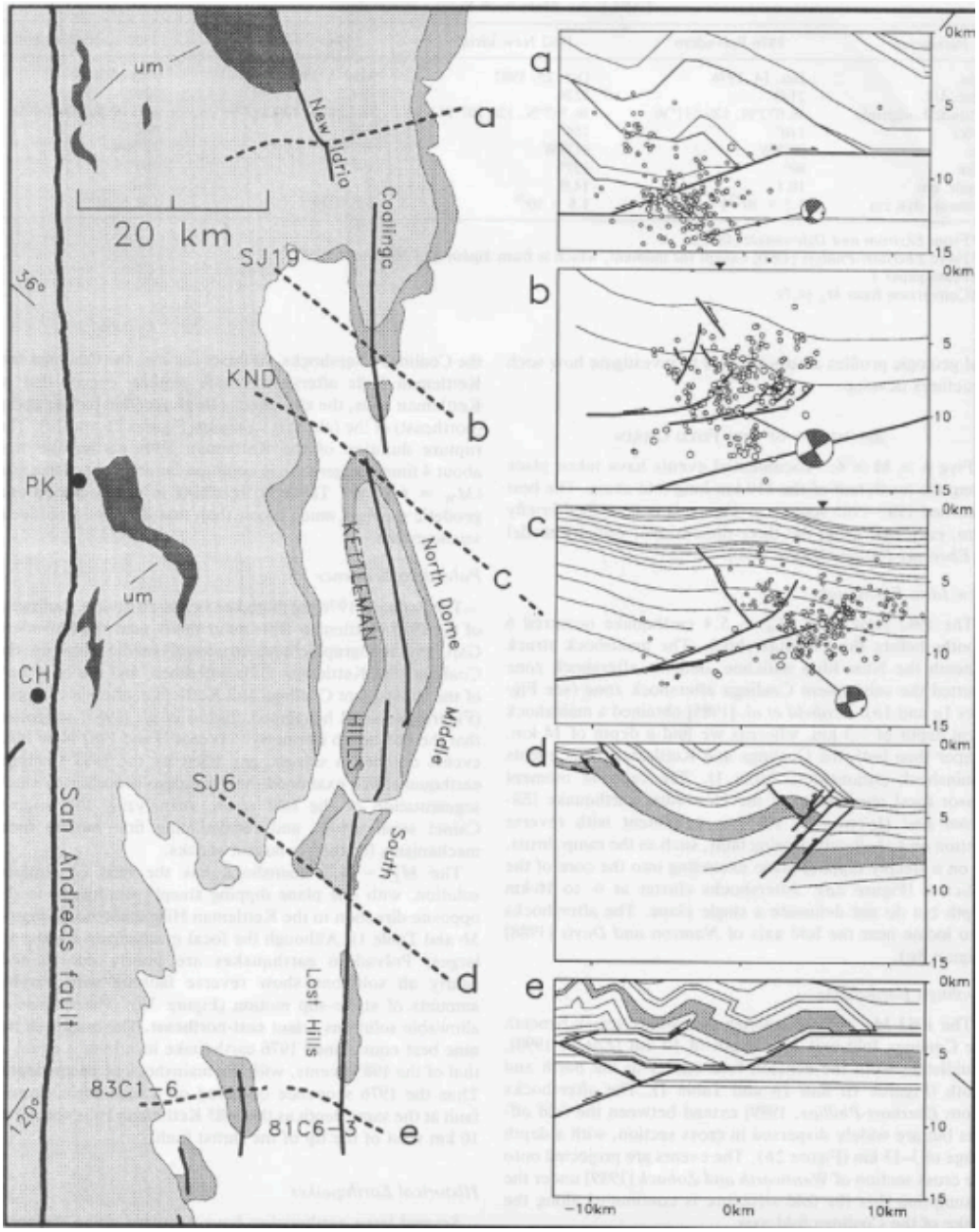


Figure 42: Lost Hills blind thrust. Lost Hills blind thrust. Map and geologic cross-sections of the blind thrust within the Idria-Lost Hills fold belt. The fold belt accommodates normal stress on the San Andreas fault. The thrust fault generated a

sequence of earthquakes in the early 1980s. The cross-sections show the seismic activity associated with the earthquake sequence (Stein and Ekstrom, 1992).

Fault-controlled hydrocarbon migration along fault zones can be an effective method of transporting deeply seated geofluids to shallower geologic units and even to the surface. The fluid migration along faults is largely associated with fault brecciation and seismic pumping during coseismic activity (Aydin, 2000; Boles et al., 2004; Dholakia et al., 1998; Eichuble and Boles, 2000; Jung et al., 2014). Strayer et al. (2001) discusses the effects of deformation on fluid-flow within fold and thrust belts using numerical models, concluding fluid-flow in thrust belts occurs due to plastic deformation even without dilation. Although Strayer et al. (2001) investigated larger structures the results imply that the 110 km blind thrust fault within the Kettleman Hills and Lost Hills oil fields may act as a preferential pathway for formation waters to interact with shallow groundwater.

In addition, the 5 identified GAMA wells and Lost Hills oil wells are located at similar depths. The shallow water wells are at depths of between 420 and 720 feet below grade (fbg) and the Lost Hills oil field contains at least 600 oil wells at depths shallower than 750 fbg. Further, the Corcoran Clay, the prominent aquitard segregating the upper unconfined aquifers from the lower confined aquifers, is located at 400 to 500 fbg in this area (CCST, 2014b) suggesting the GAMA wells are screened in the confined aquifer. This indicates the oil wells and groundwater wells may be perforated in the same geologic unit or in strata that is juxtaposed to each other. The juxtaposition of stratagraphic layers may promote fluid migration from the low permeability oil-bearing

rock to adjacent high permeability sediment layers. Hydraulic fracturing may enhance the process of lateral fluid migration into juxtaposed layers. Hydraulic fracturing wells in the Lost Hills are predominantly vertical with the fracture system extruding horizontally from the well shaft (Figure 4 in the Background section explains the horizontal fracturing process in more detail). The horizontal fractures conceivably could increase the likelihood of formation waters traveling toward the juxtaposed high permeable sediment layers.

To summarize, the statistical model indicates 5 GAMA samples show a produced water signature in two plots using two different sets of analytes. Using their ID numbers the GAMA data was further investigated and they were spatially mapped. The map shows the sample locations are in the vicinity of the Lost Hills oil field, which contains a known blind thrust fault and likely contains unknown conjugate faults that have not been mapped. Lastly, approximately 600 oil wells in the Lost Hills oil field are at depths less than 750 fbg and the 5 GAMA wells are at corresponding depths. In all, the statistical model has provided a baseline interpretation of these data, which allows for a more in-depth analysis of spatial and physical relationships between the wells and oil production. These wells should be further investigated, in which, the groundwater is analyzed for more in-depth chemical indicators like isotopes of radium, iodine, water, boron etc. These additional analyses can be applied to the statistical model to refine the statistical relationships.

For illustration, Figure 43 provides the analysis of the shallow and deep water along with a subset of data collected in the southern Salinas Valley. The plot shows the two water populations as seen before, however now the Salinas data (green) is incorporated into the plot as a third independent groundwater sampling event. This will be useful for groundwater monitoring programs that aim to see if the groundwater samples from the monitoring program show a shallow GAMA water signature or a produced or formation water signature. The Salinas Valley subset clearly shows a shallow water signature.

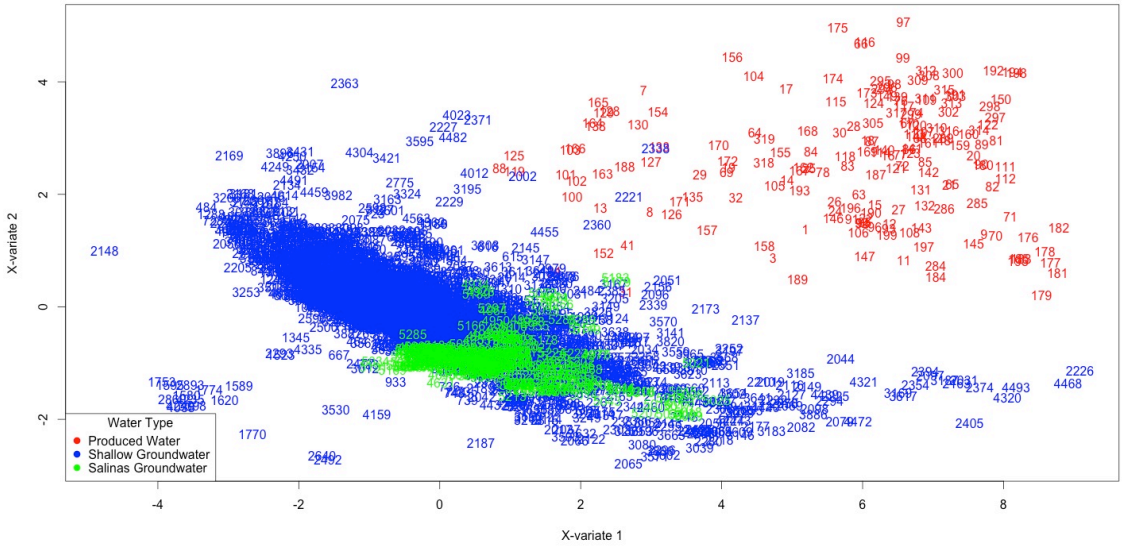


Figure 43: PLS-DA incorporating a subset south Salinas Valley groundwater data. PLS-DA incorporating a subset of GAMA groundwater data from the south Salinas Valley. The plot illustrates the usefulness of PLS-DA at allowing the interpretation of newly gathered groundwater data from regional groundwater monitoring programs against the historic GAMA data and produced water data in the Kern County groundwater basin.

CONCLUSIONS AND FUTURE WORK

The statistical analysis, in conjunction with hydrogeochemical analysis, proves useful for segregating produced water samples from shallow drinking water aquifer samples. These tests clearly indicate that the two waters contain chemical constituents that show strong variation between the two waters, can be statistically segregated and can predict if a sample or group of samples shows a relationship between the two waters within a 1% error rate. This novel method will provide evidence to whether deep formation fluids associated with oil and gas production negatively affect shallow groundwater utilized as a drinking water resource. Due to the lack of cohesive data (discussed above in the Data Gaps section) only 6 common chemical analytes were evaluated in the statistical tests. Although only 6 analytes were available for analysis, they still produce strong results of segregating the two types of water. On top of the 6 analytes, ratios of analytes were utilized in the statistical tests to evaluate their usefulness at segregating the waters. The results of the statistical test will only grow stronger with the addition of more detailed analyses of waters in the Kern County groundwater basin. For example, inputting data from analytes that hydrogeochemically distinguish produced waters, such as isotopes of radium, iodine, and water isotopes (deuterium and ^{18}O), among others will provide additional tools for examining mixing between deep and shallow waters.

This method is not limited to Kern County and should be applied to other oil producing regions in California such as the Los Angeles Basin and Ventura County basin. Even further, this method may prove useful for other oil and gas producing regions in the

United States such as the Marcellus Shale, Williston Basin and Barnett Shale. Other future work ought to include application of the United States Geological Survey's open source PHREEQC software to examine the evolution of subsurface waters along long flow paths. PHREEQC provides an avenue for incorporating reversible and irreversible reactions, including mineral, gas, ion-exchange equilibria, kinetically controlled reactions and mixing of solutions (Parkhurst and Appelo, 2013).

CHAPTER V: Radium-226 Analysis by Liquid Scintillation Counting (LSC): Dilute and Saline Matrices

Naturally occurring radioactive material (NORM) activities in hydraulic fracturing produced waters largely exceed ambient activities in shallow drinking water aquifers (Kondash et al., 2014; Rowan et al., 2011; Vengosh et al., 2013; Warner et al., 2013a; Warner et al., 2013b). Of particular concern, produced waters (a mixture of hydraulic fracturing fluid, hydrocarbons and naturally occurring formation waters; see Chapter II: Chemistry of Produced Waters subsection for a more in depth description of produced waters) typically contain activities of radium-226 (^{226}Ra) well above the EPA drinking water Maximum Contaminant Limit (MCL; 5 pCi/L) and the Industrial Effluent Discharge Limit (60 pCi/L). The high ^{226}Ra activities generally derive from parent product uranium-238 (^{238}U) favoring the solid phase in the typically anoxic conditions of the targeted formation waters, allowing for the continual production of ^{226}Ra. For instance, produced waters from the Marcellus Shale in Pennsylvania contain reported ^{226}Ra activities of up to 18,000 pCi/L (Rowan et al., 2011). The elevated radium content creates wastewater disposal complications, in which proper waste characterization becomes important. However, accurate quantification of ^{226}Ra within the wastewaters proves difficult due to the complex matrices of the produced waters.

In addition to NORMs, produced waters contain high salinity and high concentrations of cations and anions such as chloride, bromide, sodium, strontium and barium (Cl, Br, Na, Sr and Ba), among others (see Chapter II: Chemistry of Produced Waters subsection). Due to the complex chemistry of the produced waters the standard EPA analytical method for ^{226}Ra (method 903.1) may not produce reliable results (Nelson

et al., 2014). The EPA method requires the co-precipitation of radium with chemically similar barium; however, the exceedingly high total dissolved solids concentration and excess barium in the samples results in an overabundance of precipitate, complicating the separation of radium. Further, High Purity Germanium (HPGe) gamma spectroscopy provides reliable results for ^{226}Ra but typically requires a larger volume of sample (~3L), a knowledgeable gamma spectroscopy technician, prior knowledge of sample chemical makeup, and a sufficient amount of time (~17 hr per sample).

A new analytical method for ^{226}Ra by liquid scintillation counting (LSC) was developed at Lawrence Livermore National Laboratory. The LSC method was developed in order to eliminate the sample matrix complications and reduce the amount of time required for each measurement. The LSC method provides accurate and efficient results for ^{226}Ra activities above the environmental background level in both saline and dilute waters while requiring minimal sample preparation and no wet chemistry. In addition, only 10 mL of sample is needed to turn around samples within 2-weeks. The method will benefit projects of varying size that require analysis of waters with a wide range of dissolved solid content concentrations. The goal of the method is to provide ^{226}Ra analysis for waters associated with regional shallow aquifer groundwater monitoring and oil field operational wastewater characterization.

EXPERIMENTAL

Liquid scintillation counting utilizes the energy from radioactive decay events (alpha and beta decay) to generate photons of light. The photons are detected and

converted into electrical pulses by photomultiplier tubes (PMTs) producing a spectral output. Natural radioactive background variation is minimized by optical isolation of the sample detector and the built-in active sample guard (Wallac, 2002). The sample guard, made of lead, efficiently protects the sample by providing passive shielding from background radiation. In order to detect ionizing radiation from the sample, a scintillation cocktail with a fluor agent emits light when excited by alpha and beta radiation energy. The scintillator is water insoluble and designed to remove the targeted nuclide from the aqueous phase. The sample is placed in the sample guard with the PMTs which convert alpha and beta decay events from light pulses to electrical pulses (McKlveen and McDowell, 1984). The LSC utilizes a Pulse Shape Analyzer (PSA) to discriminate pure alpha events from pure beta events. Alpha particles count at a ~100% efficiency and beta particles count at a ~70% efficiency. In addition, the PSA reduces the alpha background, which generally consists of beta type pulses. Two counting configurations exist for the Quantulus LSC: Alpha/Beta Discrimination (ABD) and High Energy Beta (HEB).

Briefly, the ABD counting configuration utilizes the PSA to segregate alpha ionizing radiation events from beta events. The PSA separates alpha and beta radiation into their specified spectra by the length of the emitted light pulse, which in turn, effectively reduces the alpha-background. Therefore, as compared to the HEB configuration, the ABD configuration provides greater sensitivity when counting alpha-particle decay events for alpha emitting radionuclides such as ^{226}Ra and ^{222}Rn (Kaihola, 2000). The ABD configuration provides efficient results for alpha emitting nuclides, but

chemical composition and color potentially reduce the energy output or lead to misclassification of alpha and beta events from the sample (collectively called sample quench). In comparison, the HEB configuration allows for the quantification of all alpha and beta ionizing radiation events. By analyzing the entire alpha and beta spectra the sample matrix does not affect the efficiency or quench of the sample but does contain a higher background count rate. Although exhibiting a higher background, the HEB configuration proves useful when analyzing samples known to contain above ambient NORM activities and complex chemical compositions.

Sample Geometry

The ^{226}Ra method utilizes LSC High-Energy Beta (HEB) counting configuration to measure the short-lived (<4 day half-life) alpha and beta emitting daughter products of ^{226}Ra (^{222}Rn, ^{218}Po, ^{214}Pb, ^{214}Bi and ^{214}Po). Figure 44 shows the uranium-238 (^{238}U) decay series, which includes ^{226}Ra. Daughter product ^{222}Rn, a noble gas, preferentially partitions from the sample into the mineral oil scintillator floating atop the sample and grows into secular equilibrium with ^{226}Ra in the LSC vial. Further decay of ^{222}Rn allows for the accumulation of the four remaining short-lived daughter products of ^{226}Ra in the mineral oil. Secular equilibrium of ^{226}Ra and the daughter products is established after approximately 30 days; however, the ^{226}Ra LSC method provides viable results after approximately 7 days of ingrowth (77% ^{222}Rn ingrown). The LSC, as described above, counts all photons from the daughter nuclides after partitioning into the mineral oil scintillator, generating a photon count rate [in counts per minute (CPM)] after a 60-minute count. Figure 45 depicts the sample geometry.

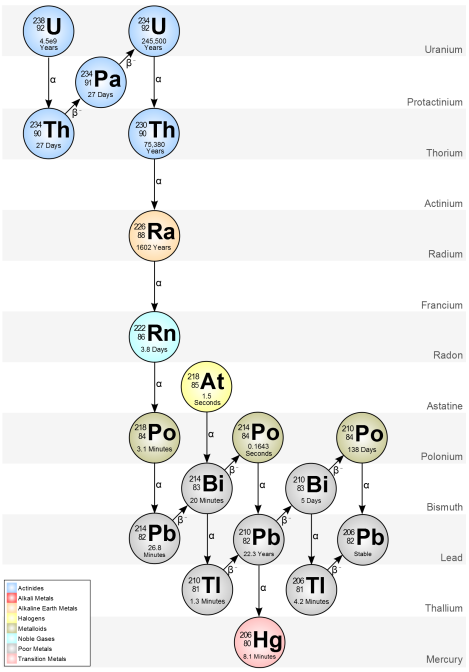


Figure 44: Uranium-238 decay series. Uranium-238 decay series. ^{238}U has a half-life of 4.468 billion years and decays by alpha decay. Daughter product ^{226}Ra has a half-life of ~1602 years and decays into the noble gas ^{222}Rn. The five short-lived (<4 day half-lives) daughter products are used to measure ^{226}Ra by counting photons generated during alpha and beta ionizing radiation decay events (UCB, 2015).

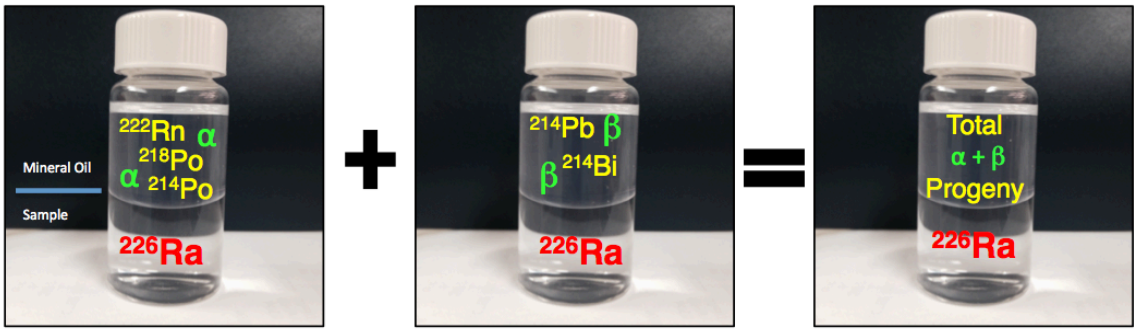


Figure 45: Schematic of sample geometry of ^{226}Ra LSC sample. Schematic of sample geometry of ^{226}Ra LSC sample. The LSC high-energy beta counting configuration

counts all photons produced by short-lived (< 4 day half-life) alpha and beta emitting progeny of ^{226}Ra in the sample.

Sample Matrix

The sample matrix does not complicate ^{226}Ra analysis by HEB LSC, unlike HPGe gamma spectroscopy and ABD LSC. HPGe gamma spectroscopy provides quality results for ^{226}Ra; however, knowledge of the general sample matrix is needed. In order to effectively quantify the ^{226}Ra activity in a sample, the gamma spectroscopy library must contain all suspected radionuclides in the sample. For illustration, a standard spiked with 0.1 mL of a 13 decays per minute (DPM) ^{226}Ra standard solution (~60 pCi/L activity) and a known natural uranium activity (~2400 pCi/L) was measured by gamma spectroscopy and LSC. The analysis indicated the ^{226}Ra may be classified as uranium-235 (^{235}U) on the gamma spectroscopy if the library does not discriminate between the two radionuclides due to ^{226}Ra and ^{235}U having very similar gamma ray emission energies (186.10 and 185.71 keV, respectively). In contrast, the uranium did not affect the LSC spectra due to the LSCs low-level alpha and beta detection. Although gamma spectroscopy analysis typically would differentiate between ^{226}Ra and ^{235}U, the results must be carefully interpreted; whereas, the LSC method eliminates the need for evaluation of uranium contamination altogether. Over and above the issue of uranium contamination, the HEB LSC method provides reliable results for chemically complex waters.

Chemical composition and color of the sample potentially affect the ability to differentiate between alpha and beta events using LSC. The chemical and color effect on

the sample analysis is collectively called the sample quench. Sample quench occurs when the efficiency of the energy transfer is reduced or when the sample matrix absorbs photons rather than emitting them (Wallac, 2002). The Quantulus LSC measures the quench in a sample against an internal standard and reports the sample quench as the standard quench parameter (SQP). Generally, as the sample quench increases, the samples SQP value decreases. To evaluate the efficiency of analyzing ^{226}Ra by the two counting configurations, 5 standards spiked with different known activities of ^{226}Ra were measured. The standards were counted for differing numbers of replicate counting intervals and an average percent recovery (or efficiency) was calculated. The percent recovery represents the efficiency of measuring the known activities of the standards for a given counting configuration. Table 9 and Figure 46 provide the experimental setup parameters and results.

Table 9: Percent Recovery of Spiked Standards - Experimental Data

Standard ID*	Activity (pCi/L)	Number Replicate of Counts (n)		Average Percent Recovery+ (%)		Relative Standard Deviation (%)	
		HEB	ABD	HEB	ABD	HEB	ABD
A	11.89	8	15	92.51± 53.4	176.87±50.5	15.62	14.53
B	59.43	12	7	87.13±13.6	112.71±24.8	6.69	13.86
C	297.15	8	5	89.37±4.9	95.66±8.8	7.08	4.65
D	594.29	8	5	93.23±3.4	99.85±6.1	6.76	3.93
E	14796.17	18	37	99.83±0.7	103.95±0.9	0.71	10.19

*The spiked standards consist of differing volumes of ^{226}Ra standard with an activity of approximately 3317.9 DPM/ml (which is temporally decay corrected) and varying volumes of deionized water in a 10 ml sample. Mineral oil (10 ml) is floated on top.
+The percent recovery error represents the average counting error determined from the 95% confidence interval of the standards in counts per minute.

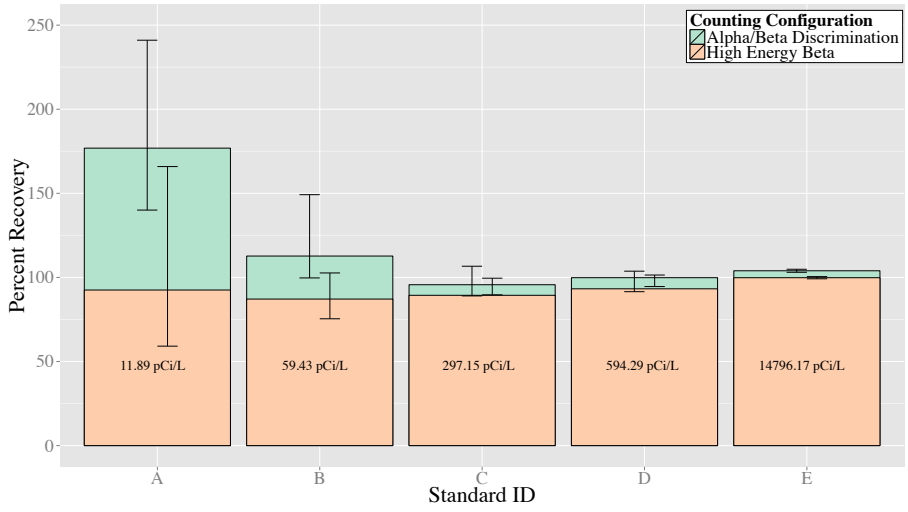


Figure 46: Five standards ranging in known ^{226}Ra activity. Five standards ranging in known ^{226}Ra activity (lowest activity on the left to highest activity on the right; A=11.89

pCi/L, E=14796.17 pCi/L) plotted against the average percent recovery of 'n' replicate counts (see data table). Error bars represent the LSCs average counting error for each standard.

The multiple standards data indicate that the efficiencies for both configurations generally are about 100%, within the counting error. One exception, however, is the lowest activity standard (Standard A = 11.89pCi/L) under the alpha-beta discrimination configuration, which averaged 176.87% recovery over 15 tests. The elevated recovery is likely due to beta events being counted as alpha events. As a whole, the counting error decreases and the percent recovery approaches 100% as the ^{226}Ra activity increases.
To ensure no loss of counts between the two counting configurations, the total counts from each counting configuration for a Standard E counting interval were analyzed. To calculate the total counts from the ABD configuration, the alpha emission count rate (counts per minute; CPM) is added to the beta emission CPM since the method discriminates between the two ionizing radiation events. In the HEB configuration, the output CPM represents all alpha and beta counts. To show that the HEB configuration captures all alpha and beta events, the total counts for each method are plotted against each other to show a highly correlated linear relationship with a slope of 1.02 (Figure 47).

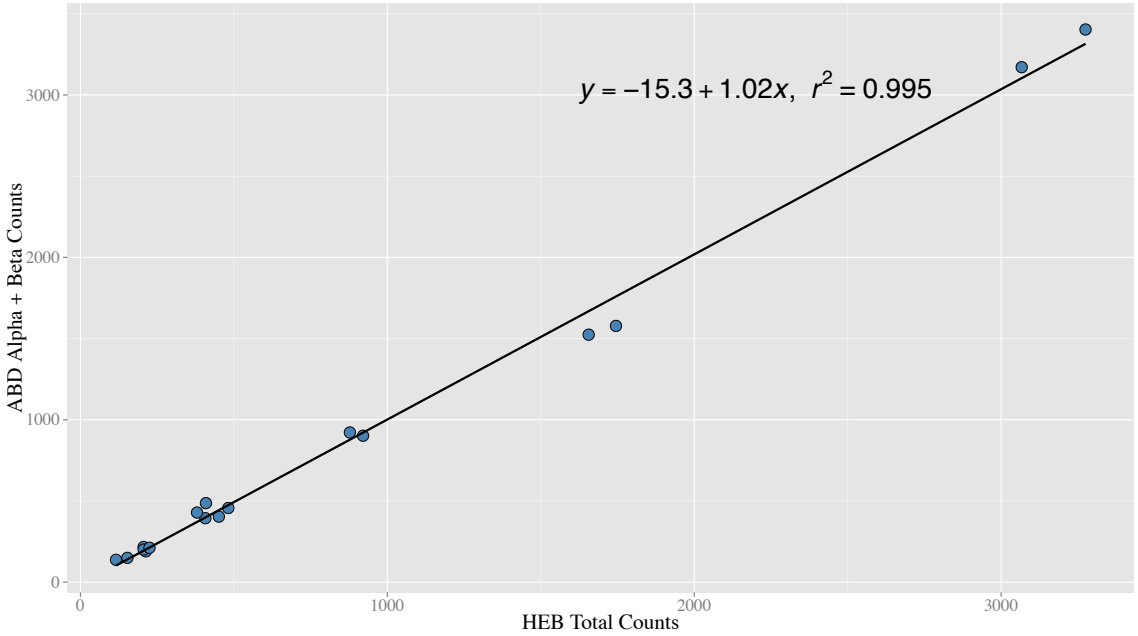


Figure 47: High-energy beta counts versus alpha-beta discrimination. Total high-energy beta counts (HEB) versus alpha-beta discrimination (ABD) total alpha counts plus total beta counts indicating no loss of counts when utilizing HEB.

Background and Minimum Detectable Activity

The background count rates (BCR) and minimum detectable activities (MDA) for a 60-minute count differ between the two configurations. High background radiation (cosmic and naturally occurring nuclides) is more easily detected in the HEB configuration than in pure alpha determination mode, so the BCR for HEB tends to be higher than the BCR for ABD. Although the BCR is higher for HEB (see Table 10) the configuration's ability to sufficiently provide accurate standard activity recoveries for samples slightly above the drinking water standard (5 pCi/L) is encouraging. Along with the BCR, the MDA is affected by the counting configuration and count time. The MDA

is a statistical approach to determine whether a sample count rate exceeds the background count rate, which largely depends on the total number of counts of the sample. With that, as the count duration increases, the LSC registers more counts, allowing for a better distinction between the sample count rate and the background, ultimately lowering the MDA.

Table 10:

Minimum Detectable Activity and Background Count Rate

Counting Configuration	Count Duration (minutes)	MDA (pCi/L)	BCR (CPM)
Alpha/Beta Discrimination	60	5.06	0.65
High Energy Beta	60	5.96	2.50
Alpha/Beta Discrimination	200	2.57	0.65
High Energy Beta	200	3.13	2.50

Counting Standard A under a 200-minute HEB configuration resulted in a decrease in the MDA (3.13 pCi/L) while producing a percent recovery of 93.31%. Conversely, the MDA for the ABD configuration decreased to 2.57 pCi/L. The results from the longer count time data imply that the ^{226}Ra LSC method is applicable for dilute waters that contain ^{226}Ra at or slightly below the drinking water standard.

RESULTS AND DISCUSSION

Produced Waters

Lawrence Livermore National Laboratory (LLNL) received four produced water samples from the United States Geological Survey. The samples were collected from

four different hydraulic fracturing wells located in three different oil fields in Kern County, California. Two, 1-liter samples were non-filtered and collected in pre-cleaned HDPE plastic bottles with no preservative. The samples range in color from a yellowish brown to a dark murky brown. All of the samples contain visible particulates and slight oil sheens. Sample AA30851 from the Lost Hills oilfield visually contains the most particulate material. The well and sample information are provided in Table 11. Figure 48 provides a picture of the four samples.

Table 11: Produced Water Sample Information

LLNL Sample ID	API	Oil Field	Stimulation Type*	Total Depth (m)	Lithology
AA30848	3003251	Belridge, South	Unknown	335.28	Shale
AA30849	3048836	Belridge, North	HF	475.64	Diatomite
AA30850	3051253	Belridge, North	HF	490.73	Diatomite
AA30851	3026538	Lost Hills	HF	740.66	Diatomite
*HF = Hydraulic Fracture					

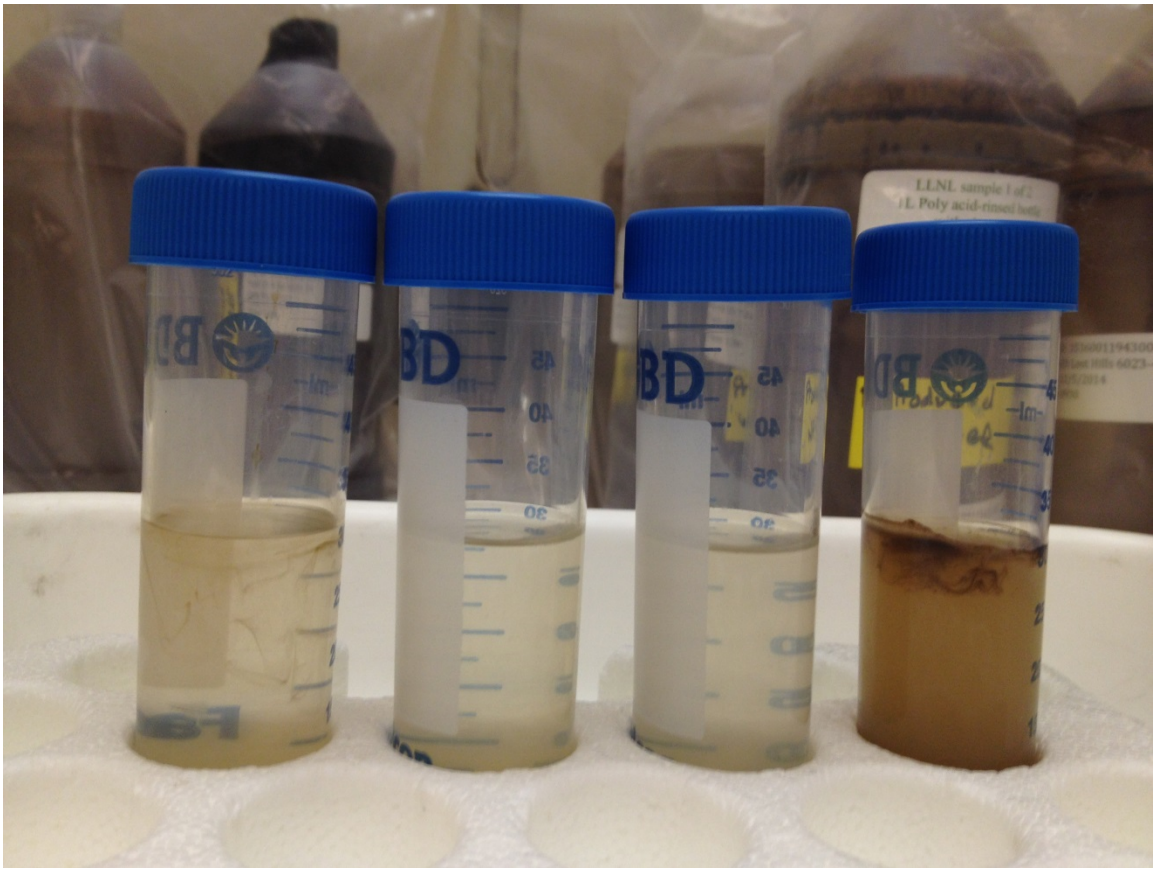


Figure 48: Photo of the four produed water samples. Photo of the four produed water samples in centrifuge tubes. Sample AA30851 (last on the right) contains the largest amount of visible particulates and oil sheen and is darkest in color.

The samples were analyzed for ^{228}Ra and ^{228}Th by high purity germanium (HPGe) gamma spectroscopy and ^{226}Ra by HPGe gamma spectroscopy and ABD LSC at LLNL. The results are presented in Table 12. In addition, the samples were evaluated for sample quench by reviewing the LSC SQP values for each sample.

Sample Quench Evaluation

Analysis of four Kern County, California oil field produced water samples (AA30848, AA30849, AA30850 and AA30851) indicate a considerable decrease in the

Standard Quench Parameter (SQP) relative to standards and blanks. The SQP measures sample quench against an internal standard; however, to determine a maximum SQP for the sample geometry, blank samples were used. Sample quench creates a loss of counts for beta particles or a shift in the spectral output for alpha particles due to attenuation. To assess sample quench on the spectral output of the produced waters, 9.5 mL of the four samples were spiked with 0.5 mL of 3317.9 DMP (~74000 pCi/L) ^{226}Ra standard and left for approximately 6.5 days to allow for the ingrowth of ^{222}Rn. After ingrowth the samples were counted on the LSC as described in the Experimental section. The SQP of the spiked produced waters was evaluated against a standard solution containing 9.9 mL of DI water and 0.1 mL of 3317.9 DPM (~14800 pCi/L) ^{226}Ra standard. The SQP analysis indicated that the spectral output of AA30851 showed the largest amount of distortion and statistically contained the lowest SQP (highest quench) of the four produced water samples. Figure 49 shows the spectral output of the initial sample analysis including the reported SQP values from least amount of quench to most quench. The spectra peak heights for the standards are lower because they have lower ^{226}Ra activity.

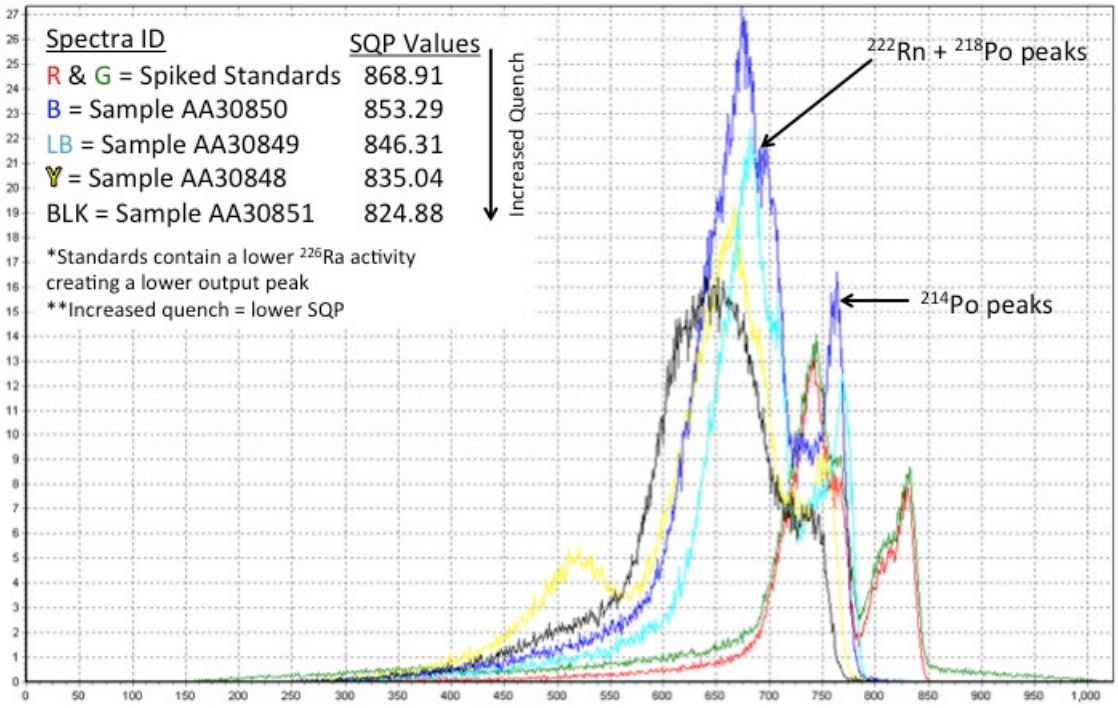


Figure 49: Spectral output for four produced waters. Spectral output for four produced waters spiked with 74,000 pCi/L activity of ^{226}Ra. The distortion of the black spectrum (sample AA30851) indicates the largest amount of quench relative to the sample with the least amount of quench (AA30850) and the standards (red and green spectra). The standards have lower ^{226}Ra activity than the samples resulting in the smaller peaks. The spectrum shows the shift of the produced waters spectra relative to the standards due to quench. To compensate for the shift, the spectral output windows must be appropriately selected to capture all viable counts.

The samples were analyzed for total petroleum hydrocarbons as gasoline, deasel and motor oil range organics (TPH-g,d,mo; carbon range C6 – C36) to evaluate the quenching effect of residual oil. The samples TPH-g concentration ranged from 140 to

2100 micrograms per liter (µg/L) and the TPH-d and –mo was generally consistant between the samples with the exception of sample AA30851. The results of the TPH analysis is provided as Table 12.

Table 12: Total Petroleum Hydrocarbons (µg/L)

LLNL ID Number	TPH-g (C6-C12)	TPH-d (C10-C23)	TPH-mo (C18-C36)
AA30848	2,100	14,000	21,000
AA30849	540	29,000	32,000
AA30850	140	22,000	15,000
AA30851	660	2,500	3,400

Sample AA30851 contained the largest amount of quench but low concentrations of TPH, relative to the other samples. The TPH concentrations in the produced water samples indicate quench likely does not stem from residual oil but rather other organic content in the waters, color or total dissolved solids concentration. In order to reduce the amount of quench, two pre-analysis preparation experiments were conducted: 1) centrifuge the sample prior to emplacing 10 mL in LSC vial in attempt to separate the oil and water; and, 2) gently stir in approximately 0.20 g of emulsifying agent to break up the oil and other organics. Over the counter de-greasing dish soap was utilized for the emulsifying agent. The experiment utilized sample AA30851, in which the centrifuged AA30851 and emulsified AA30851 samples were measured against a regular sample with no sample preparation. The results show an increase in the SQP for both experiments; however, emulsifying AA30851 provides better resolution, tighter relative standard deviation and a greater increase in the percent of maximum (the percent of

maximum is calculated from the average of the standard derived SQP and average SQP from the sample). Sample AA30851 was utilized for the experiment because it showed the largest amount of quench. Using other samples with differing sample matrices may provide different results. Tables 13 and 14 provide the statistical data from all the regular samples and the AA30851 experiments, respectively. Figure 50 and Figure 51 graphically show the SQP results for both experiments from Table 13 and Table 14.

Table 13: SQP Results for the Regular Samples After 16 Replicate Counts

Sample ID	Average SQP	Relative Standard Deviation	Percent from Max
Standard Derive Max	864.59	0.47%	NA
AA30848	829.22	0.56%	4.09%
AA30849	838.24	0.67%	3.05%
AA30850	840.89	1.03%	2.74%
AA30851	822.35	0.57%	4.89%

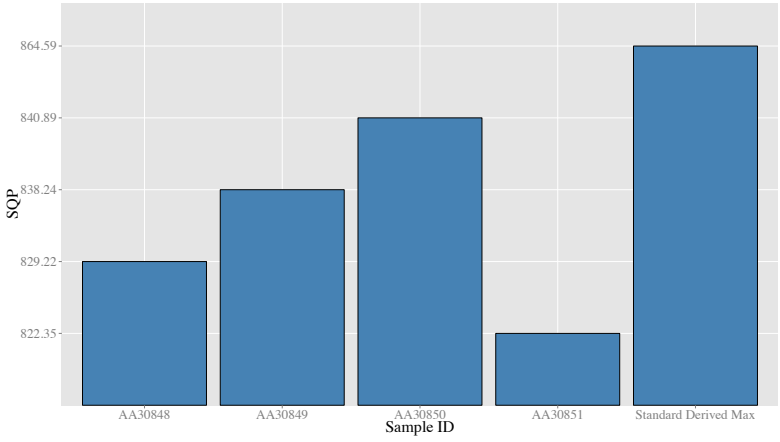


Figure 50: SQP results for the non-treated produced water samples. SQP results for the non-treated produced water samples compared to the standard derived max with an

SQP of 864.59. Sample AA300851 contains the lowest SQP value indicating it has the highest amount of sample quench.

Table 14: SQP Results for Sample AA30851 Experiments After 16 Replicate Counts

AA30851 Pre-Analysis Prep Type	Average SQP	Relative Standard Deviation	Percent from Max
Standard Derive Max	864.59	0.47%	NA
Regular (no prep)	834.92	1.76%	3.41%
Centrifuged	841.51	2.00%	2.64%
Emulsified	845.04	0.81%	2.23%

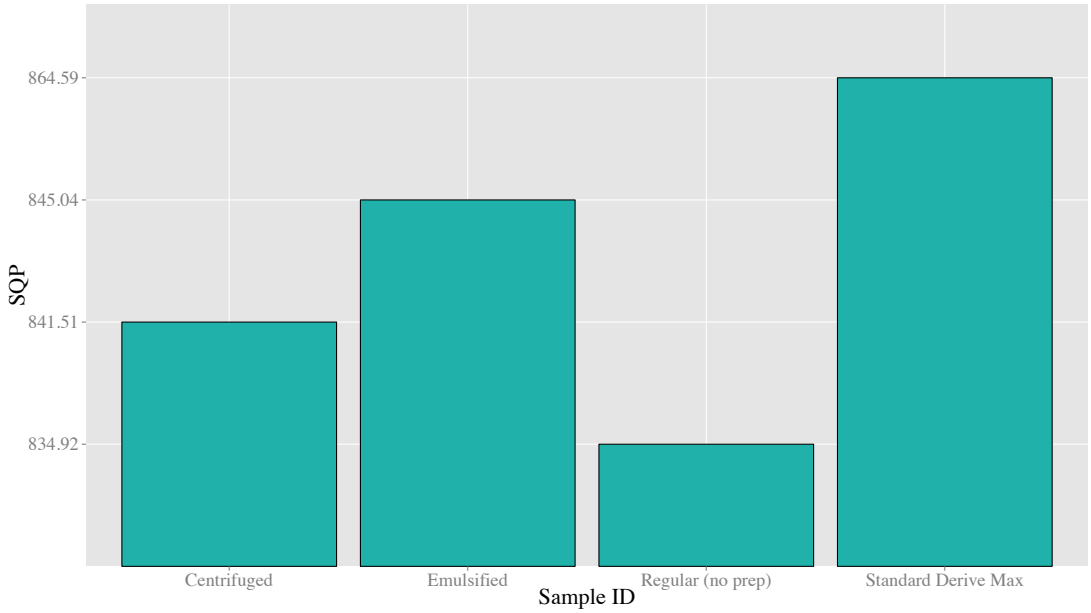


Figure 51: SQP results for the sample AA30851 pre-treatment experiments. SQP results for the sample AA30851 pre-treatment experiments compared to a regular AA30851 sample with no sample preparation and a standard derived maximum SQP

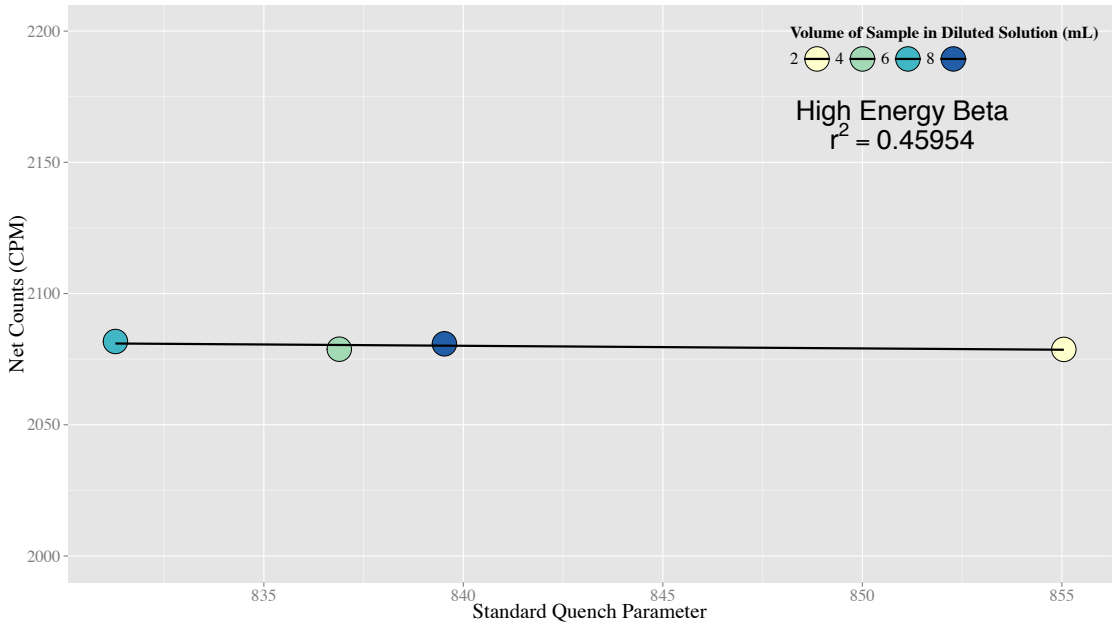
value. Emulsifying the sample aids in alleviating sample quench more than centrifuge sample preparation.

To further investigate the effects of quench on the recovery of counts, aliquots of sample AA30851 were spiked with approximately 30,000 pCi/L of activity and diluted with varying volumes of DI water to a total volume of 10 mL. Diluting the samples effectively reduces quench as the sample volume to DI water volume ratio becomes smaller. In the experiment, four spiked sample to DI solution ratios were tested: 2:8, 4:6, 6:4 and 8:2 (ratio = sample:DI water in mL; least quench to most quench). To demonstrate the effect of quench on both counting configurations, the samples were counted according to the methods described previously for both HEB and ABD. No emulsifying agent was added to the solutions. The solutions were counted twice and the mean values were calculated for interpretation. Table 15 shows the results of the experiment. Figures 52A and 52B represent the results of the dilution tests.

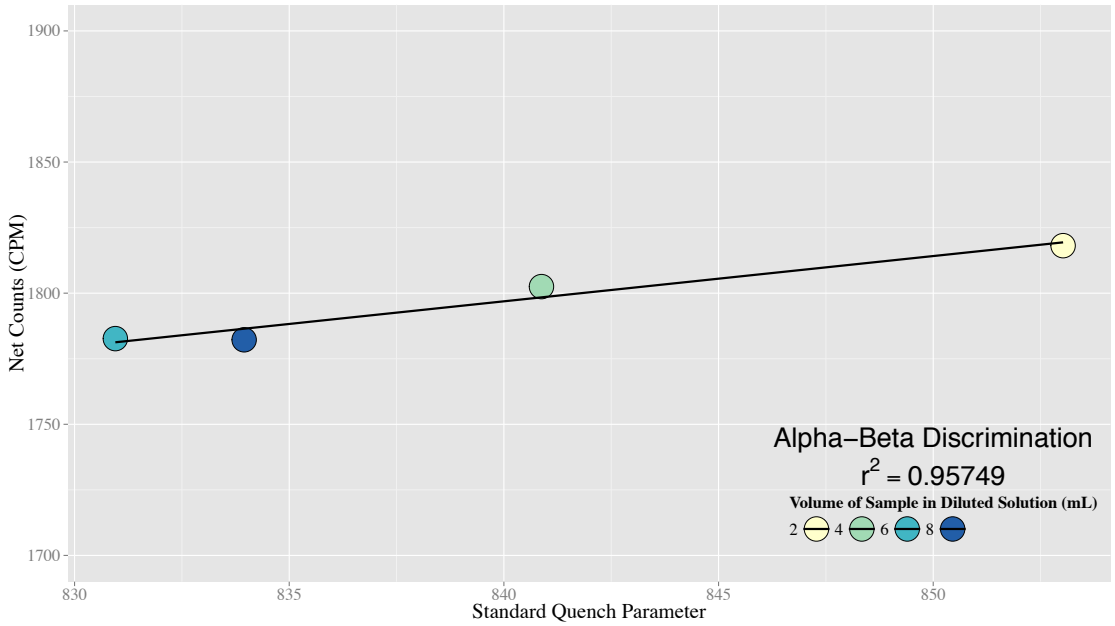
Table 15: Diluted AA30851 Quench Experiment Data

Sample ID	Sample (mL)	DI Water (mL)	Activity (pCi/L)	HEB			ABD		
				Counts (CPM)	SQP	Percent Recovery	Counts (CPM)	SQP	Percent Recovery
851 x1	2	8	30,000	2078.72	855.05	91.71%	1818.11	853.03	119.59%
851 x2	4	6	30,000	2078.80	836.89	91.37%	1802.50	840.88	118.17%
851 x3	6	4	30,000	2081.73	831.28	91.17%	1782.67	830.95	116.51%
851 x4	8	2	30,000	2080.85	839.53	90.82%	1782.22	833.95	116.12%

52A



52B



Figures 52A & B: Relationship between net counts and SQP. Relationship between the net counts and SQP of 4 diluted solutions with varying amounts of spiked sample.

The HEB configuration (5A) does not show a relationship between SQP and net counts, whereas the ABD configuration (5B) shows a strong positive correlation between the SQP and net counts. Utilizing the HEB configuration allows for minimal effect of quench in the analysis. The ABD configuration will lose counts or skew the spectral output as quench becomes more prominent.

The dilution tests show that the quenching agent in the sample affects the ABD configuration more than HEB. Figure 52B shows as the volume of sample increases, the SQP and the net counts decrease. Conversely, the HEB data indicate little to no relationship between SQP and net counts. The dilution experiments provide further affirmation that the HEB LSC method is a reliable analytical method for high quench samples.

Produced Waters Matrix Spike

The four produced water samples from Kern County were spiked with known activities of ^{226}Ra and measured to evaluate the accuracy of both counting configurations. Table 16 shows the results from the spiked sample tests.

Table 16: Spiked Sample Recoverability Data

Sample ID	Spiked Activity (pCi/L)	SQP		Percent Recovery (%)	
		HEB	ABD	HEB	ABD
AA30848	14,793.42	859.06	852.64	99.59 ± 0.73	105.51 ± 0.83
AA30849		860.11	853.12	95.99 ± 0.71	110.02 ± 0.85
AA30850		868.34	871.57	100.68 ± 0.73	113.01 ± 0.86
AA30851		842.51	839.95	98.69 ± 0.72	107.74 ± 0.84

The produced water samples were spiked with 14,793.42 pCi/L of activity and counted on both the ABD and HEB configuration. The HEB results indicate recoveries between 96% and 101%; whereas, the recoveries for the ABD configuration all exceed 105% likely due to misclassification of beta particles as alpha particles as a result of sample quench. In addition, the HEB results contain a slightly tighter counting error in comparison to the ABD results.

CONCLUSIONS

High-energy beta liquid scintillation counting provides efficient and accurate results for analyzing ^{226}Ra in saline and dilute waters. In addition, when sample quench is reduced, alpha/beta discrimination LSC provides viable results. The LSC method will prove useful for field studies of varying size with diverse water matrices. Due to the minimal chemistry involved the method can be applied to a large number of samples with a rapid turn-around-time relative to the standard EPA method or HPGe gamma spectroscopy. Although the

method provides accurate results for low activity samples, the method is more effective at determining ^{226}Ra activities above the drinking water maximum contaminant limit (MCL) of 5 pCi/L.

Analyzing samples from a natural environment with a predicted activity of ^{226}Ra above the MCL will lead to further confidence in the method. With that, future work may involve measuring additional produced waters or wastewaters from oil fields that are known to contain high ^{226}Ra activities. Additionally, measuring ambient groundwater where produced waters potentially interact with the shallow groundwater will prove the usefulness of the method in monitoring ^{226}Ra activities in groundwater associated with oil and gas development.

REFERENCES CITED

Abdi, H., and Williams, L. J., 2010, Principal component analysis: Wiley Interdisciplinary Reviews: Computational Statistics, v. 2, no. 4, p. 433-459.

Arroyo, R. J., 2014, 2013 Kern County Agricultural Crop Report, *in* Standards, D. o. A. a. M., ed.: Bakersfield, California, County of Kern, p. 15.

Aydin, A., 2000, Fractures, Faults, and Hydrocarbon Entrapment, Migration and Flow: Marine and Petroleum Geology, v. 17, p. 797-814.

Baddeley, A., Turner, R., and Rubak, E., 2015, Getting Started with SpatStat, Comprehensive R Archive Network.

Bartow, J. A., 1991, The Cenozoic Evolution of the San Joaquin Valley, California: United States Geological Survey, Professional Paper 1501.

Behl, R. J., 1999, Since Bramlette (1946): The Miocene Monterey Formation of California Revisited, *in* Moores, E. M., Sloan, D., and Stout, D. L., eds., Classic Cordilleran Concepts: A View from California, Volume Special Paper 338: Boulder, Colorado, Geological Society of America, p. 301-313.

Behl, The Monterey Formation of California: New Research Directions, *in* Proceedings AAPG Annual Convention and Exhibition, Long Beach, California, 2012, AAPG.

Belitz, K., Fram, M. S., and Johnson, T. D., 2015, Metrics for Assessing the Quality of Groundwater Used for Public Supply, CA, USA: Equivalent-Population and Area: Environmental Science & Technology.

Boles, J. R., Eichuble, P., Garven, G., and Chen, J., 2004, Evolution of a Hydrocarbon Migration Pathway Along Basin-Bounding Faults: Evidence From Fault Cement: The American Association of Petroleum Geologists Bulletin, v. 88, no. 7, p. 947-970.

Brereton, R. G., and Lloyd, G. R., 2014, Partial least squares discriminant analysis: taking the magic away: Journal of Chemometrics, v. 28, no. 4, p. 213-225.

CCST, 2014a, Advanced Well Stimulation Technologies in California: An Independent Review of Scientific and Technical Informaiton, *in* Technology, C. C. o. S. a., Laboratory, L. B. N., and Institue, P., eds.: Sacramento, CA, California Council on Science and Technology.

CCST, 2014b, Advanced Well Stimulation Technologies in California: An Independent Review of Scientific and Technical Informaiton, *in* Technology, C. C. o. S. a., Laboratory, L. B. N., and Institue, P., eds.: Sacramento, CA, California Council on Science and Technology, p. 396.

Cherry, J., Ben-Eli, M., Bharadwaj, L., Chalaturnyk, R., Dusseault, M. B., Goldstein, B., Lacoursiere, J.-P., Matthews, R., Mayer, B., Molson, J., Munkittrick, K., Oreskes, N., Parker, B., and Young, P., 2014, Environmental Impacts of Shale gas Extractin in Canada, *in* Academics, C. o. C., ed.

Clark, M. S., 2015, Geology of the San Joaquin Valley, Volume 2015, San Joaquin Geological Services, Inc.

Condon, S. M., and Dyman, T. S., 2006, 2003 Geologic Assessment of Undiscovered Conventional Oil and Gas Resources in the Upper Cretaceous Navarro and Taylor Groups, Western Gulf Province, Texas, *in* Survey, U. S. G., ed., Volume DDS-69-H: Reston, VA, U.S. Geological Survey.

CSWRCB, 2015, State Water Resources Control Board: GeoTracker GAMA - Groundwater Ambient Monitoring and Assessment, Volume 2015.

Darrah, T. H., Vengosh, A., Jackson, R. B., Warner, N. R., and Poreda, R. J., 2014, Noble gases identify the mechanisms of fugitive gas contamination in drinking-water wells overlying the Marcellus and Barnett Shales: Proceedings of the National Academy of Sciences, v. 111, no. 39, p. 14076-14081.

Davis, J. C., 2002, Statistics and Data Analysis in Geology, John Wiley & Sons, Inc., 638.

Dholakia, S. K., Aydin, A., Pollard, D. D., and Zoback, M. D., 1998, Fault-Controlled Hydrocarbon Pathways in the Monterey Formation, California: American Association of Petroleum Geologists Bulletin, v. 82, no. 8, p. 1551-1574.

DOGGR, 2015, Interim Well Stimulation Treatment Notices Index, Volume 2015, Division of Oil, Gas and Geothermal Resources.

Dresel, P. E., and Rose, A. W., 2010, Chemistry and Origin of Oil and Gas Well Brines in Western Pennsylvania: Pennsylvania Geological Survey.

DWR, 2003, California's Groundwater: Bulletin 118, *in* Resources, D. o. W., ed.: Sacramento, California, Department of Water Resources, p. 265.

DWR, 2015, Groundwater: Sacramento, California, Department of Water Resources.

Eichuble, P., and Boles, J. R., 2000, Focused Fluid Flow Along Faults in the Monterey Formation, Coastal California: GSA Bulletin, v. 112, no. 11, p. 1667-1679.

Esser, B. K., Beller, H. R., Carroll, S. A., Cherry, J. A., Gillespie, J., Jackson, R. B., Jordan, P. D., Madrid, V., Morris, J. P., Parker, B. L., Stringfellow, W. T., Varadharajan, C., and Vengosh, A., 2015, Recommendations on Model Criteria for Groundater Sampling, Testing, and Monitoring of Oil and Gas Development in California: Lawrence Livermore National Laboratory, LLNL-TR-669645.

Gallegos, T. J., and Varela, B. A., 2015, Trends in Hydraulic Fracturing Distributions and Treatment Fluids, Additives, Proppants, and Water Volumes Applied to Wells Drilled in the United States from 1947 through 2010—Data Analysis and Comparison to the Literature: U.S. Geological Survey, Scientific Investigation Report 2014-5131.

Godwin, H., 2011, Merge All Files in a Directory Using R Into a Single Dataframe, Volume 2015, Psychwire.

Gordalla, B. C., Ewers, U., and Frimmel, F. H., 2013, Hydraulic Fracturing: A Toxicological Threat for Groundwater and Drinking-water?: Environ. Earth Sci., v. 70, p. 3875-3893.

Haluszczak, L. O., Rose, A. W., and Kump, L. R., 2013, Geochemical Evaluation of Flowback Brine from Marcellus Gas Wells in Pennsylvania, USA: Applied Geochemistry, v. 28, p. 55-61.

Jung, B., Garven, G., and Boles, J. R., 2014, Effects of episodic fluid flow on hydrocarbon migration in the Newport-Inglewood Fault Zone, Southern California: Geofluids, v. 14, no. 2, p. 234-250.

Kahle, D., and Wickham, H., 2013, ggmap: Spatial Visualization with ggplot2: The R Journal, v. 5, no. 1, p. 144-161.

Kaihola, L., 2000, Radionuclide Identification in Liquid Scintillation Alpha-Spectroscopy: Journal of Radioanalytical and Nuclear Chemistry, v. 243, no. 2, p. 313-317.

KCWA, 2014, Improvement District No. 4: Report on Water Conditions 2014, *in* Agency, K. C. W., ed.: Bakersfield, California, Kern County Water Agency, p. 73.

Kondash, A. J., Warner, N. R., Lahav, O., and Vengosh, A., 2014, Radium and Barium Removal Through Blending Hydraulic Fracturing Fluids with Acid Mine Drainage: Environmental Science & Technology, v. 48, no. 2, p. 1334-1342.

Kyser, T. K., 2007, Fluids, basin analysis, and mineral deposits: Geofluids, v. 7, no. 2, p. 238-257.

Magoon, L. B., Lillis, P. G., and Peters, K. E., 2007, Petroleum Systems Used to Determine the Assessment Units in the San Joaquin Basin Province, California, *in* Interior, D. o. t., ed., United States Geological Survey, p. Ch. 8.

Mazor, E., 2004, Chemical and Isotopic Groundwater Hydrology, New York, New York, Marcel Dekker, Inc., 451.

McKlveen, J. W., and McDowell, W. J., 1984, Liquid Scintillation Alpha Spectrometry Techniques: Nuclear Instruments and Methods in Physics Research v. 223, p. 372-376.

McMahon, P. B., Caldwell, R. R., Galloway, J. M., Valder, J. F., and Hunt, A. G., 2015, Quality and Age of Shallow Groundwater in the Bakken Formation Production Area, Williston Basin, Montana and North Dakota: Groundwater, v. 53, no. S1, p. 81-94.

Montgomery, C. T., and Smith, M. B., 2010, Hydraulic Fracturing: History of an Enduring Technology: Journal of Petroleum Technology, v. 62, no. 12, p. 26-32.

Mooney, C., 2011, The Truth About Fracking: Scientific America, v. 305, no. 5, p. 80-85.

Nash, K. M., 2010, Shale Gas Development, New York: Nova Science Publishers, 2010, 174.

Nelson, A. W., May, D., Knight, A. W., Eitrheim, E. S., Mehrhoff, M., Shannon, R., Litman, R., and Schultz, M. K., 2014, Matrix Complications in the Determination of Radium Levels in Hydraulic Fracturing Flowback Water from Marcellus Shale: Environmental science & Technology Letters, v. 1, p. 204-208.

Olsson, O., Weichgrebe, D., and Rosenwinkel, K.-H., 2013, Hydraulic Fracturing Wastewater in Germany: Composition, Treatment, Concerns: Environ. Earth Sci., v. 70, p. 3895-3906.

Page, R. W., 1983, Geology of the Tulare Formation and Other Continental Deposits, Kettleman City Area, San Joaquin Valley, California, With a Section on Ground-Water Management Considerations and Use of Texture Maps: United States Geological Survey, Water-Resources Investigations Report 83-4000.

Parkhurst, D. L., and Appelo, C. A. J., 2013, Description of Input and Examples for PHREEQC Version 3—A Computer Program for Speciation, Batch-Reaction, One-Dimensional Transport, and Inverse Geochemical Calculations: United States Geological Survey, U.S. Geological Survey Techniques and Methods, book 6, chap. A43.

Pavely, F., 2013, Senate Bill No. 4 - SB 4, Pavely. Oil and Gas: Well Stimulation, Volume 2015, California State Legislative Branch.

Pollastro, R. M., Roberts, L. N. R., and Cook, T. A., 2010, Assessment of Undiscovered Oil and Gas Resources of the Williston Basin Province of North Dakota, Montana, and South Dakota, *in* Survey, U. S. G., ed., Volume DDS-69-H: Reston, VA, U.S. Geological Survey.

Rogers, G. S., 1919, The Sunset-Midway Oil Field Californias: Geochemical Relations of the Oil, Gas, and Water: United States Geological Survey, Professional Paper 117.

Ross, D. C., 1986, Basement-Rock Correlations Across the White Wolf-Breckenridge-Southern Kern Canyon Fault Zone, Southern Sierra Nevada, California: United States Geological Survey, Bulletin 1651.

Rowan, E. L., Engle, M. A., Kirby, C. S., and Kraemer, T. F., 2011, Radium Content of Oil-and Gas-Field Produced Waters in the Northern Appalachian Basin (USA): Summary and Discussion of Data: U.S. Geological Survey, 2011-5135.

Shelton, J. L., Pimentel, I., Fram, M. S., and Belitz, K., 2008, Ground-Water Quality Data in the Kern County Subbasin Study Unit, 2006—Results from the California GAMA Program: United States Geological Survey, Data Series 337.

Shlens, J., 2014, A Tutorial on Principal Component Analysis, Volume 2015, Google Research.

Smith, T., 2012, The Many Lives of Belridge, Volume 2015: GEO ExPro, GEO ExPro.

Stein, R. S., and Ekstrom, G., 1992, Seismicity and Geometry of a 110-km-Long Blind Thrust Fault: Synthesis of the 1982-1985 California Earthquake Sequence: Journal of Geophysical Research, v. 97, no. B4, p. 4865-4883.

Strayer, L. M., Hudleston, P. J., and Lorig, L. J., 2001, A numerical model of deformation and fluid-flow in an evolving thrust wedge: Tectonophysics, v. 335, no. 1–2, p. 121-145.

Su, C., and Suarez, D. L., 2004, Boron Release from Weathering of Illites, Serpentine, Shales, and Illitic/Palygorskitic Soils: Soil Science Society of America Journal, v. 68, p. 96-105.

Suppe, J., 1985, Principles of Structural, Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 537.

SWRCB, 2013, Communities that Rely on a Contaminated Groundwater Source for Drinking Water, *in* Board, S. W. R. C., ed.: Sacramento, California, State Water Resources Control Board, p. 181.

Thuot, K., 2014, Half of US Oil Production Comes from These 20 Counties, Volume 2015, Drilling Info.

UCB, 2015, Nuclear Forensics: A Scientific Search Problem, Volume 2015, University of California, Berkeley.

USGS, 2015, Energy Resoures Program: Produced Waters Database, Volume 2015.

Vengosh, A., Warner, N., Jackson, R., and Darrah, T., 2013, The Effects of Shale Gas Exploration and Hydraulic Fracturing on the Qualifty of Water Resources in the United States: Procedia Earth and Planetary Science, v. 7, p. 863-866.

Wallac, 2002, Quantulus: Measuring Extremely Low Levels of Environmental Alpha and Beta Radiation, PerkinElmer Life Sciences, Inc., p. 12.

Warner, N. R., Christie, C. A., Jackson, R. B., and Vengosh, A., 2013a, Impacts of Shale Gas Wastewater Disoposal on Water Quality in Western Pennsylvania: Environmental Science & Technology, v. 47, p. 11849-11857.

Warner, N. R., Kresse, T. M., Hays, P. D., Down, A., Karr, J. D., Jackson, R. B., and Vengosh, A., 2013b, Geochemical and Isotopic Variations in Shallow Groundwater in Areas of the Fayetteville Shale Development, north-Central Arkansas: Applied Geochemistry, v. 35, p. 207-220.

Weddle, J. R., 1967, Oilfield Waters in Southwestern San Joaquin Valley, Kern County, California: Division of Oil and Gas, Summary of Operations: California Oil Fields Fifty-Third Annual Report.