UC Riverside
UC Riverside Electronic Theses and Dissertations

Title
Visualizing Arsenic Groundwater Contamination in Coachella Valley, California Through Spatial Analysis and Modeling

Permalink
https://escholarship.org/uc/item/63t0v0fh

Author
Hope, Monica Jean

Publication Date
2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Visualizing Arsenic Groundwater Contamination in Coachella Valley, California Through Spatial Analysis and Modeling

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Environmental Sciences

by

Monica Jean Hope

June 2022

Thesis Committee:
Dr. Samantha Ying, Chairperson
Dr. Hoori Ajami
Dr. Ying-Hsuan Lin

Copyright by
Monica Jean Hope
2022

The Thesis of Monica Jean Hope is approved:

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

I would like to acknowledge all the people who have been on this journey with me. First of all, to the amazing women I have had the privilege to be mentored by. Not the least of which being my graduate advisor, Dr. Samantha Ying, who made all of this possible for me. Also, a big thank you to my lab mates who have always been there to lend a hand or ear when graduate life got rough. Another thanks to my collaborators, especially Nataly Garcia and Alandra Lopez, who helped me develop my project. A tremendous thank you to my family who made time to call and visit me after I moved to Riverside for school. And last, to my dog, Minnehaha, who was always there for me after moving here alone during the pandemic.

DEDICATION

This thesis is dedicated to my father, Bruce Alan Hope. He established and nurtured my love for the natural world. Without his constant cheerleading and belief in my abilities I wouldn't be here today. While he didn't live to see me achieve this accomplishment, I know he would be so ridiculously proud of me for pursuing my Master's Degree.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

1.1 Geochemistry of Arsenic Groundwater Contamination

Arsenic is a naturally occurring metal contaminant that is pervasive in global groundwater sources (Fendorf et al., 2010) and commonly found in drinking water supply wells (Lombard et al., 2021). Groundwater is a critical source of drinking water for more than 2 billion people worldwide (Thomas & Famiglietti, 2015) and the World Health Organization (WHO) estimates that more than 200 million people worldwide may have chronic arsenic exposure through drinking water at concentrations above the WHO safety standard of 10 µg L^{-1} (Naujokas et al., 2013). Chronic exposure to arsenic can lead to skin lesions, cardiovascular diseases, diabetes, and the development of various cancers including skin and lung cancers (Magalhães, 2002). Due to its potential to contribute to public health hazards and its high frequency of occurrence in drinking water sources, arsenic is ranked number one on the Agency for Toxic Substances and Disease Registry (ATSDR) substance priority list (Naujokas et al., 2013).

Arsenic contaminated groundwater generally occurs in locations that are characterized as inland or closed basins in arid areas or strongly reduced aquifers derived from alluvium (Smedley & Kinniburgh, 2001). Both types of environments have groundwater flow that tends to be sluggish and held within poorly flushed aquifers which allows arsenic released from the sediments to accumulate. In the United States, the southwestern states have been identified as some of the most affected by arsenic contamination in groundwater due the arid climate and the presence of both oxidizing and reducing subsurface geochemical conditions (Smedley & Kinniburgh, 2001).

Geogenic metals contamination of groundwater can occur via weathering of bedrock followed by oxidation-reduction (redox) reactions (Ritter et al., 2002). These metals include both primary and secondary contaminants, including manganese, that have been shown to lead to health effects (Kondakis et al., 1989). Oxyanions, such as uranium (U), chromium (Cr), and vanadium (V), are generally more mobile under arid and oxidizing conditions, while arsenic (As), iron (Fe), and manganese (Mn) are generally more mobile under reducing conditions in their reduced forms as As(III), Fe(II), and Mn(II) (when sulfide concentrations are relatively low). Arsenic, and other anion-forming elements can be released due to the development of alkaline pH (>8.5) conditions in arid (and generally oxidizing) environments resulting from the combined effects of mineral weathering and high evaporation rates. Arsenic tends to be less strongly adsorbed to clay mineral surfaces as solution pH increases due to being a hydrolyzing metal that forms oxyanions, where the deprotonation of coordinated oxygens under alkaline conditions decreases the favorability of forming a surface complex on negatively charged mineral binding sites. Due to similar geochemical characteristics, the most common trace metal contaminants in groundwaters that co-occur with arsenic also tend to be oxyanion forming elements (Cr, As, U, and Se).

Arsenic occurs in the environment in several oxidation states, but in groundwater and most terrestrial systems As is mostly found as one of two inorganic oxyanions: trivalent arsenite [As(III)] under predominantly reducing conditions or pentavalent arsenate [As(V)] under oxidizing conditions. Arsenic is among the most problematic oxyanion-forming elements because of its relatively high mobility over a wide range of redox conditions. Under oxidizing conditions, the mono- and di-protonated forms of As(V) ($HAsO_4^{2-}$ and

$H_2AsO_4^-$) are dominant at low pH (less than about pH 6.9, below the second pK_a), whereas at highly alkaline conditions (pH>11), AsO_4^{3-} becomes dominant. In contrast, As(III) species dominate under reducing, anoxic conditions. As(III) is generally more mobile than As(V) due to higher selectivity in its adsorption to aluminum hydroxides and iron (oxyhydr)oxides (Goldberg, 1986). In reducing conditions at neutral pH, the uncharged As(III) species, $H_3AsO_3^0$, dominates; deprotonation of $H_3AsO_3^0$ does not occur until pH is greater than 9.

Arsenic can also be mobilized from the solid phase at large scales when strongly reducing conditions developed under neutral conditions. Reducing conditions, often fueled by the presence of ample organic carbon, can fuel microbial reductive dissolution of Fe (oxyhydr)oxides and Mn oxides, which can lead to the release of previously adsorbed As. Arsenic can also be mobilized through direct reduction of As(V) to As(III) by anaerobic metals-respiring bacteria (Tufano et al., 2008). Highly reducing conditions can lead to accumulation of bicarbonate, a by-product of microbial respiration, which can competitively desorb As from aquifer minerals like Fe (oxyhydr)oxides (Schaefer et al., 2020).

1.2 Arsenic Contamination in Coachella Valley, California

Groundwater As contamination is also pervasive throughout California aquifers. Over 370 million residents in the state are estimated to be exposed to groundwater contaminated with As, nitrate, or hexavalent chromium [Cr(VI)] (Pace et al., 2022). Although a grown number of studies have focused on elucidating the mechanisms responsible for groundwater arsenic in the Central Valley (e.g., Smith et al., 2018; Ayotte et al., 2016; Gao et al., 2004), much less is known about the spatial distribution of and mechanisms responsible for groundwater arsenic occurrence in the Coachella Valley groundwater basin. Coachella Valley is located within the Imperial Valley in southeastern California, which is characterized by an arid climate averaging only 6 inches of precipitation per year; runoff from surrounding mountains from these infrequent rain events serve as the primary source of recharge for the groundwater basin along with direct infiltration of irrigation waters. Similar to the Central Valley, agriculture is a major industry in the region which employs a large migrant workforce particularly within eastern Coachella Valley composed of mostly Mexican immigrants who reside in mobile home parks known as *polancos* which are predominantly dependent upon groundwater accessed through private domestic wells.

The California State Water Resources Control Board (SWRCB) along with several other state and federal agencies including the U.S. Geological Survey created the Groundwater Ambient Monitoring and Assessment (GAMA) Program in 2020 to establish a platform for groundwater monitoring and data repository to inform the public of groundwater quality issues in California. In 2021, the SWRCB released the Aquifer Risk Map for Domestic and State Small Water Systems, which is an interactive web-based map for residents, engineers, and policy makers to identify priority areas for Safe and Affordable Drinking Water (SADW) funds based on Safe and Affordable Funding for Equity and Resilience (SAFER) program assessments. The interactive map was created to fulfill one of the requirements of SB-200 known as the Human Rights to Water Act. This map identifies many areas in California that are at risk of arsenic contamination, including Coachella Valley.

Although a paucity of geochemical research has focused on deciphering mechanisms responsible for As release in Coachella Valley groundwater, instances of As contamination in the region has been reported in several publicly available white papers released by Coachella Valley Water District and non-peer reviewed publications including in the popular press. Recently, Coachella Valley's groundwater quality issues were brought to the national spotlight when the U.S. Environmental Protection Agency (USEPA) issued a third emergency order on drinking water safety issues to owners of Oasis Mobile Home Park (MHP), located on Torres Martinez Tribe's lands in Thermal, California. The emergency order required the Park owners, "to provide alternative drinking water, reduce the levels of arsenic in the system's water and monitor the water for contamination" (EPA notice, 2019, 2021). Oasis MHP, which houses approximately 1,900 residents, was found to have concentrations of As of up to 97 parts per billion (ppb), which is nearly 10 times above the federally set maximum contaminant level (MCL) of 10 ppb (EPA notice, 2019). Oasis MHP is just one of hundreds of small informal mobile home communities in the area that have been located and mapped (unpublished data, Leadership Council for Justice and Accountability), many of which provide drinking water to residents through private wells that are potentially drawing from contaminated portions of the basin. Although there are currently monitoring wells installed in the region, their distribution is sparse and the chemistry of water withdrawn by these MHPs is unknown. The U.S. EPA and other California and federal agencies only require monitoring and regulating drinking water for sources that serve 15 or more service connections or greater than 25 people (EPA SDWA) and recommend testing private water supplies for contamination on a periodic basis every 1 to 3 years (Ayotte et al., 2015); these private water supplies are otherwise not monitored or regulated. And because the *polancos* generally have less than 15 service connections, there is no monitoring of the water quality. In fact, a major reason high drinking water arsenic contamination at Oasis Mobile Home Park was detected is because it is a large MHP located on tribal lands which therefore has federal (EPA) oversight. Given how pervasive small *polancos* are in this region, there is reason to believe that more communities in Eastern Coachella Valley are likely facing similar water quality issues.

In the 1980s, the environmental justice movement brought attention to the fact that communities of different race, ethnic, and socioeconomic backgrounds were disparately impacted by environmental pollution. A study that "retooled" the California Community Environmental Health Screening Tool (CalEnviroScreen) created by the California Office of Environmental Health Hazard Assessment (OEHHA) found that the extent of Latina/o presence had the strongest correlation with cumulative pollution burden (Lievanos, 2018). Indigenous communities in the U.S. have and continue to face many environmental justice issues. When the west was being settled many Indigenous tribes were neglected in discussions of water access and design and construction of water infrastructure did not include their considerations. This has led to many tribes facing persistent water quality issues on their land. Since many migrant latina/o workers live on tribal lands in Eastern Coachella Valley, both communities are simultaneously adversely affected by the poor water quality.

1.3 Inverse Distance Weighting (IDW)

Spatial distribution of groundwater contamination has been estimated through creating maps applying interpolation methods to existing groundwater chemical data (Figure S1). A number of spatial interpolation methods can be used to generate a continuous surface map using known data points. One of the most used deterministic models in spatial interpolation is inverse distance weighting (IDW) method (Lu & Wong, 2008). It is a commonly applied local spatial interpolation method in geographic information science because it is relatively fast, easy to compute, and results are straightforward to interpret (Lu & Wong, 2008). The IDW method is quite straightforward and not computationally intensive, and as succinctly stated by Lu & Wong (2008), "the general premise of this method is that the attribute values of any given pair of points are related to each other, but their similarity is inversely related to the distance between two locations…[the] value of an unsampled point is the weighted average of known values within the neighborhood, and the weights are inversely related to the distances between the prediction location and the sampled locations."

Many past studies have applied IDW to describe the spatial distribution of groundwater contaminants (e.g., Elumalai et al., 2017; Mirzaei et al., 2016). Nath et al. (2018) developed spatial distributions maps of various heavy metal contaminants, including arsenic, across a river valley using the IDW approach to estimate the spatial variability in the groundwater chemistry. Given these factors, we applied the IDW method to describe the magnitude and distribution of groundwater metal contamination in the Coachella Valley groundwater basin.

1.4 Random Forest Classification (RFC) Modeling

Random forests "are an ensemble learning method for classification and regression that constructs a number of randomized decision trees during the training phase and predicts by averaging the results" (Scornet. et al., 2015). Breiman (2001) published a groundbreaking study that showed that random forests (RF) are an effective tool in prediction. Since its publication, RFs have become a significant data analysis tool that are successfully used to solve a variety of practical problems (Scornet et al., 2015). Breiman attributes their effectiveness to the fact that the Law of Large Numbers keeps RFs from overfitting and the injected randomness makes RFs accurate classifiers and regressors. The rising popularity of RF method can be attributed to the fact that it can be applied to a wide range of prediction problems and is generally easily parameterized. In addition, the RF method is recognized for its accuracy, ability to handle small sample sizes, high-dimensional feature spaces, and complex data structures (Scornet et al., 2015).

Researchers have recently begun using RF methods to identify areas that are likely to have high arsenic concentrations in groundwater to create arsenic groundwater prediction maps based on publicly available geospatial and subsurface geochemical datasets (Podgorski & Berg, 2020; Lopez et al., 2020; Lombard et al., 2021). In these studies, RF is used to predict the spatial distribution of groundwater contamination because it has proven to have improved predictive power compared to the more commonly used logistic or linear regression models, while also being able to identify important explanatory variables (Lopez et al., 2020). Using these previous studies as an example, we used RF modeling to predict areas in the Coachella Valley groundwater basin that have arsenic concentrations at levels relevant to human health and identify the environmental factors important in prediction.

1.5 Overarching Objective

In order to investigate the extent of As contamination, as well as Cr and Mn, we performed spatial analyses of contaminant data. To investigate where arsenic concentrations were at levels relevant to human health, we applied random forest modeling using existing geochemical data from the Coachella Valley groundwater basin. Spatial interpolation mapping and predictive modeling of As concentrations within Coachella Valley has not been performed using these methods. In this study, we aim to highlight areas in the region that are most likely impacted by groundwater arsenic contamination approaching the MCL. We also use spatial interpolation to map the distribution of potential co-contaminants to identify areas that are affected by multiple geogenic groundwater contaminants. We hypothesized that 1) the highest concentrations of arsenic would be found in the eastern portion of Coachella Valley, the region where Oasis Mobile Home Park and many other MHPs are located; 2) Mn contamination would coincide in locations with As contamination due to similar geochemical controls on their mobility; and 3) Cr groundwater contamination would be in more oxic areas and would not be co-located with As and Mn contamination.

In order to highlight areas that may be impacted by concentrations of As that can lead to chronic exposure, we performed random forest modeling on arsenic and environmental data within the Coachella Valley groundwater basin. We did so to have a stronger predictive model than that of a spatially interpolated map, and to elucidate which geochemical and environmental factors are best predictors of As contamination in the region. We hypothesized 1) RF model results will also highlight the eastern part of the region as being at risk of groundwater As contamination approaching the primary MCL and 2) soil texture would be a strong predictor of As contamination because fine textured soils (i.e. high clay content) can enhance and sustain anoxic conditions that lead to As release through reductive dissolution (Fendorf et al., 2010).

We chose to create both spatially interpolated inverse distance weighting maps and random forest model maps because each provides different information regarding distribution of groundwater contamination. The spatially interpolated maps show the spatial distribution of metal contaminants and the predicted concentrations across a continuous surface. They are relatively easy to create and only require groundwater metal concentration data. The random forest model shows areas that are approaching the primary MCL of As, specifically by calculating the probability that a location will draw groundwater As concentrations above half the primary MCL (5 ppb). The RF map can also identify the explanatory environmental factors that best predict locations likely to have As contamination. Because of this additional information resulting from RF modeling, it is a more complex process that requires many additional datasets that represent potential predictive environmental factors.

2. Material and Methods

2.1 Study Area: Coachella Valley Groundwater Basin

The Coachella Valley groundwater basin is located in Southern California just northwest of the Salton Sea, between 34.1° and 33.3° north latitude and 117.0° and 115.9° west longitude, and has a surface area of 1,360 km^2 (Figure 1). The climate in Coachella

Valley is typical of an arid region in the southwestern United States, with annual precipitation averaging about 15 centimeters yr^{-1} and a wide temperature range throughout the year. Groundwater recharge, also typical of arid regions, is primarily from the induced recharge of losing streams as well as valley-edge recharge (Thomas & Famiglietti, 2015). Between 1936 and 1967, Coachella Valley increased its groundwater withdrawals ten-fold, resulting in aquifer elevation declines of up to 1.5 m yr^{-1} and subsistence of up to 6 mm yr^{-1} (Thomas & Famiglietti, 2015). In more recent years, in order to counteract the effects of groundwater pumping, Coachella Valley has developed methods for groundwater replenishment and introduced Colorado River allocations to use for irrigation (Thomas & Famiglietti, 2015).

There are large socioeconomic disparities between the western and eastern portions of Coachella Valley. The western side is well known for its wealthy recreational resorts in Palm Springs, Palm Desert, and Indian Wells (Mukhija & Mason, 2014). The western region is urbanized and residents receive water through public municipal water systems which provide treated, consistent, and safe drinking water. On the other hand, Eastern Coachella Valley is an expansive rural and agricultural area mostly populated by low-income immigrant families (Mukhija & Mason, 2014). Because the eastern portion of the valley is more rural with low-density housing, a majority of residents in the eastern region obtain their drinking water from private wells that are unregulated and potentially untreated. Many treatment options including point of use or consolidation is cost-prohibitive for communities in Eastern Coachella Valley.

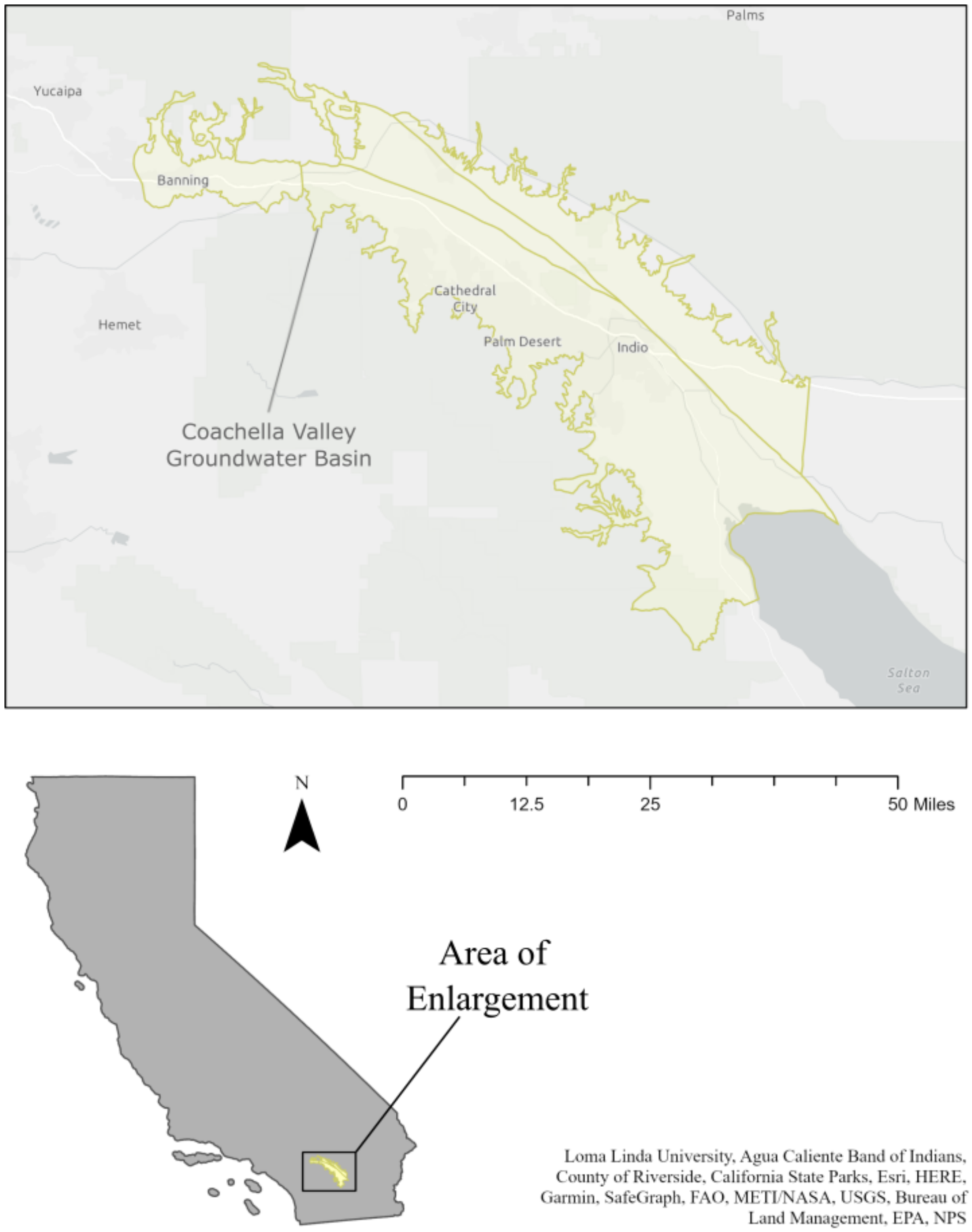

Figure 1. Map showing the location of Coachella Valley within the state of California as well as the boundaries of the Coachella Valley Groundwater Basin and major cities within the basin.

2.2 Inverse Distance Weighting Data Sources

Groundwater quality data were obtained from the California State Water Resources Control Board, Groundwater Ambient Monitoring and Assessment Program (GAMA) (Table S1 ; Figure S1). To construct this database GAMA compiled well-sample data that were collected and analyzed to comply with state regulations. Samples were collected from wells used for a variety of purposes, including as domestic use, irrigation, and monitoring. All water samples were collected from raw groundwater, so measurements represent water quality within the aquifer.

2.3 Inverse Distance Weighting Analyses

Inverse distance weighting (IDW) analyses were performed using ArcGIS Pro software (2.7.2, ESRI). Groundwater data was drawn from the GAMA dataset and clipped to the Coachella Valley Aquifer Region (Figure 1). IDW's were performed to interpolate concentrations of As, Mn, and Cr in the region's groundwater with samples taken from 343 wells for all analyses. The GIS software package ArcGIS Pro 2.7.2 was used in conjunction with MATLAB to map, query, and analyze the data in this study. Since choosing the number of nearest neighbors determines IDW precision (Yao et al., 2013), we compared results obtained from a range of closest neighbors (3 to 12), and then performed 5-fold cross validation. The number of nearest neighbors used in the IDWs that produced the least amount of error was then chosen as the optimal IDW parameters for interpolation; in the current study, the optimal number of nearest neighbors was 6.

2.4 Random Forest Data Sources and Preprocessing

We compiled publicly available predictors, or explanatory variables, and groundwater arsenic concentrations in Coachella Valley, CA in ArcGIS Pro for RF modeling. (Table S1). Groundwater As concentrations were again compiled from the GAMA dataset and clipped to the Coachella Valley Aquifer Region as mentioned in the methods for IDW interpolation (Figure 1). Like IDW methods, concentrations from a total of 343 well locations that had groundwater metals chemistry from 1984 to 2021 were used (Figure S1). Groundwater chemical characteristics used included aqueous As, Cr, hardness, Fe, Mn, nitrate (as N), sulfate, and pH values. Since chemical concentrations were highly variable and contained outliers, all concentrations were log-transformed for statistical modeling. Samples that were below detection limit were adjusted to be a random number chosen from a uniform distribution between 0 and the chemical detection limit. Since individual wells were sampled variable number of times over the study period, the median of sample concentrations for each analyte from an individual well was retained. Apart from chemistry, other groundwater or well characteristics considered included well location (latitude, longitude), well depth, and data source; however, very few wells had well depth information (~50%) so this factor was excluded from the final model. There was a prevalence of wells with As concentrations reported as less than the MCL, therefore, threshold models were developed to estimate the probability of exceeding a concentration threshold instead of regression models that estimate concentration values (Lombard et al., 2021).

To create the simplest model, an initial set of 76 potential relevant environmental predictor variables was reduced based on consideration of their relative importance and potential impact on the accuracy of the random forest models. The final selection of 21 predictor variables (Table S1) includes soil geochemistry (C horizon bulk density, clay content, electrical conductivity, pH, sand content, and soil organic carbon content), well depth (both domestic and public top open well depth and bottom open well depth), hydrologic landscape (Percent total flat land, relief, minimum elevation, percent sand, slope, and aquifer permeability), climate (temperature and precipitation), groundwater recharge, and land use. While groundwater chemistry is likely to be an important predictor for groundwater As concentrations, we were unable to use the data in our RF models as they were only available at well locations and not mapped across the whole region.

The final 21 predictor variables were used to predict the occurrence of threshold exceedances of As in groundwater with random forest classification (RFC) modeling. The model input predictor or explanatory variables explain soil geochemistry, well depth, hydrologic landscape, climate, groundwater recharge, and land use (Table S1). All variables were mapped across the study, compiled within 500 m radius circular buffers around the well locations, and either the mean value or majority value of the variable raster was taken, depending on if the data was continuous or categorical. GIS processing of explanatory variables was performed using ArcGIS Pro 2.7.2. This data was then compiled into a large dataframe to use for modeling.

2.5 Random Forest Methods

Using the R computing environment, version 4.1.2, we developed our final RF models with the GAMA dataset of As concentrations in Coachella Valley along with our chosen 21 predictor variables. We initially set out to develop two RFC models that either estimated the probability of As concentrations above the MCL of 10 ppb or above half the MCL of 5 ppb. In the end, we were only able to develop a RFC model predicting occurrences of As above 5 ppb; our preliminary RFC model with a As concentration threshold of 10 ppb was not pursued due to poor model performance metrics likely resulting from limited data availability.

The RFC model was developed using the *caret*, *randomForest*, and *caTools* packages. We started by first refining the number of predictor variables with recursive feature elimination (RFE) in order to improve model performance and delete uninformative predictors that might bias the results (Bahl et al., 2019). The model was further refined by tuning, which involved using a training dataset and 10-fold cross validation 10-times repeated to produce the most accurate model using accuracy as the metric. The tuning consisted of altering the number of variables randomly sampled at each decision tree split (mtry hyper-parameter) from 1 to 10 while the number of trees to grow (ntree) was maintained at 500.

Given that the As dataset was imbalanced (less than 10% of samples were above 5 ppb) we chose to set a new probability threshold for the model that would maintain model specificity and total accuracy, while also increasing model sensitivity (Erickson et al., 2021) . The final model was used to make maps of model estimates within Coachella Valley groundwater basin. A 500 m^2 raster grid was created and clipped to the same extent as the predictor variables in ArcGIS Pro. As some predictor rasters had grid cells with missing data, the RFC model could not calculate estimates in those cells. A majority of missing grid values were around the edges of the region and were excluded from the model. The model estimates were calculated from the raster files in R using the *stats* package and output with latitude and longitude to be imported into ArcGIS Pro. The maps created show the probability of exceeding 5 ppb of As in Coachella Valley groundwater.

3. Results and Discussion

3.1 Application of Inverse Distance Weighting

Since the distribution of groundwater contaminants is primarily controlled by geochemical heterogeneity, spatial interpolation such as inverse distance weighting is often used to estimate groundwater concentrations and provide a comprehensive representation of groundwater concentration distribution (Nath et al., 2018). We subjected our sample dataset to spatial interpolation using the IDW approach to illustrate the magnitude and distribution of metal contamination in the Coachella Valley Aquifer. Using the publicly available GAMA dataset, we utilized data from 343 wells in the region to use as groundwater samples. The data was spatially interpolated using IDW to generate maps that represent the distribution of arsenic, chromium, and manganese in the region at 500 m spatial resolution (Figures 2-4).

Datasets for each chemical of interest were divided into training and testing samples, with 20% of data reserved for testing, and then subjected to 5-fold cross validation. To evaluate the error in our interpolated maps, we calculated mean absolute error (MAE), root-mean-square error (RMSE), and G-value (G), which can be found in Table 1. There was variation in calculated MAE and G between the resulting IDW maps, ranging from 1.5-4.0 ppb and -0.5-0.5, respectively. The RMSE was approximately 4.0 ppb for all interpolated maps. MAE was likely smallest for the As IDW because it is a measure of the difference between the predicted value and the actual value, and As concentrations in the region were much less variable than Cr and Mn. The G error statistic was also best for As, with 1 being a perfect prediction and negative values being worse predictions than taking the dataset average. Mn likely had a negative value because concentrations in the regions were highly variable and the IDW couldn't predict well due to the limited data in the region. All the contaminates had RMSE around 4.0 ppb, with a great model having a value between 0.2 - 0.5, so the IDWs are overall not great predictors for contamination in the region but they do show us the general distribution trends for As, Cr, and Mn.

In the raw well data, As concentrations ranged from 0 ppb to 100 ppb, with a majority of the samples being under the MCL for As of 10 ppb. Arsenic distribution within the IDW output map shows most areas predicted to have groundwater concentrations near or above 10 ppb are located within the eastern portion of Coachella Valley (Figure 2). There are also hotspots of higher As concentrations predicted along the northern portion of the aquifer boundary as seen in the IDW.

IDW maps show that Cr groundwater distribution generally follows an inverse relationship to As concentrations, where areas of high arsenic concentrations have low Cr concentrations and vice versa (Figure 3). This result is consistent with our hypothesis that opposing redox conditions favor As versus Cr mobilization within soils and sediments, with As predominantly being mobilized in suboxic to mixed redox conditions and Cr release being promoted under oxic conditions (Coyte & Vengosh, 2020). Manganese groundwater distribution in Coachella Valley also appears to also have an inverse relationship with As distribution, in contrast to our hypothesis (Figure 4). Although Mn and As are both mobilized under suboxic to anoxic conditions, they may be occurring in contrasting areas in our interpolated map because well-depth has been shown to separate As versus Mn contamination in wells, with Mn contamination occurring at significantly shallower depths than As contamination (Ying et al., 2017).

Table 1. Error statistics for 5-fold cross-validated spatially interpolated IDW maps. Equations for error statistics found in supplementary information.

Chemical Mapped	MAE	RSME	G
Arsenic	1.55	4.05	0.49
Chromium	3.24	4.45	0.20
Manganese	4.16	4.58	-0.51

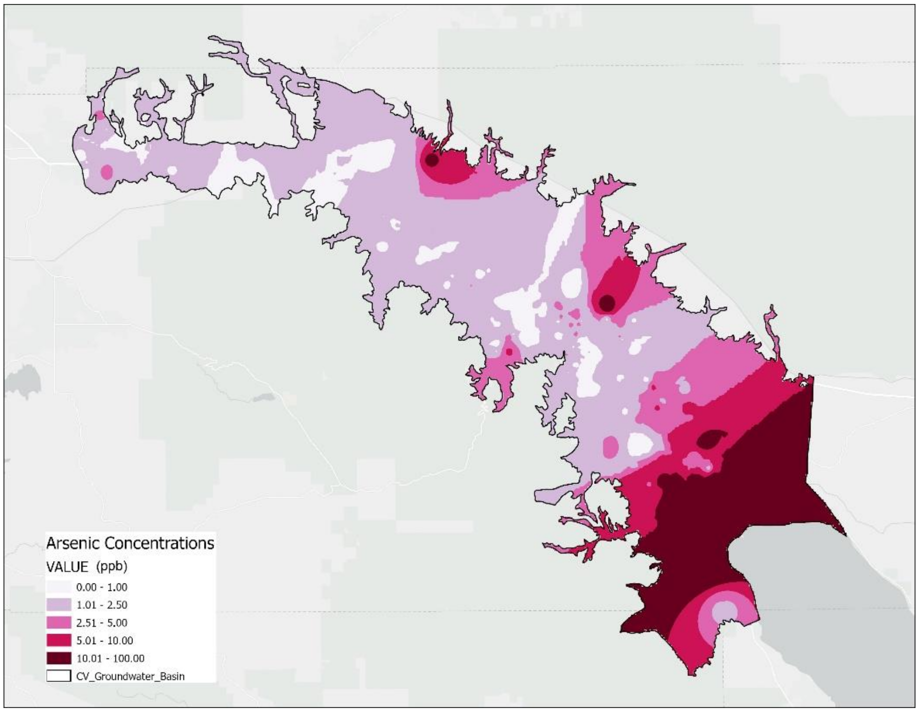

Figure 2. Spatially interpolated map using IDW showing groundwater arsenic distribution in Coachella Valley, concentrations of arsenic.

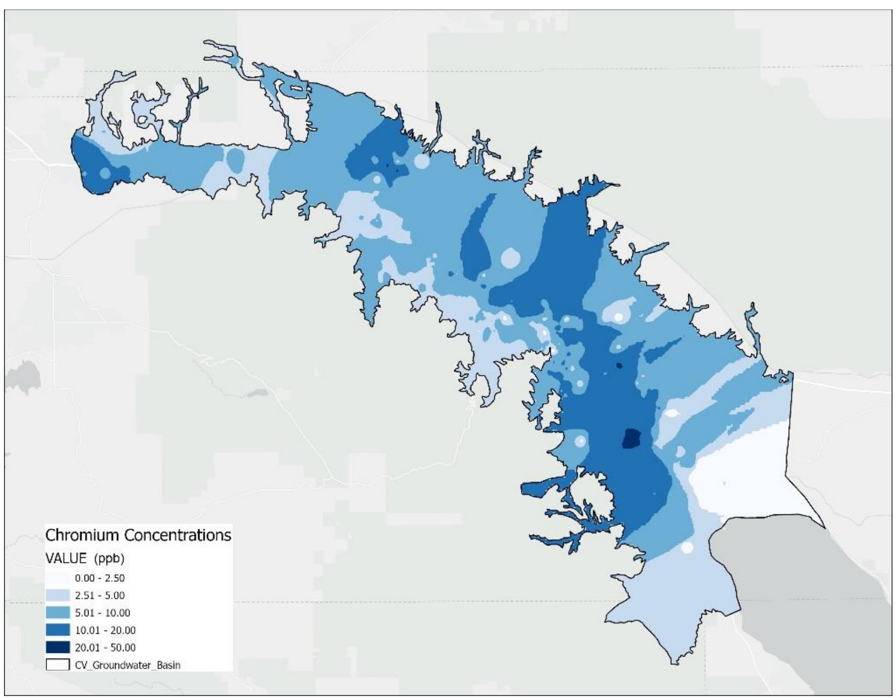

Figure 3. Spatially interpolated map using IDW showing groundwater chromium distribution in Coachella Valley, concentrations of chromium.

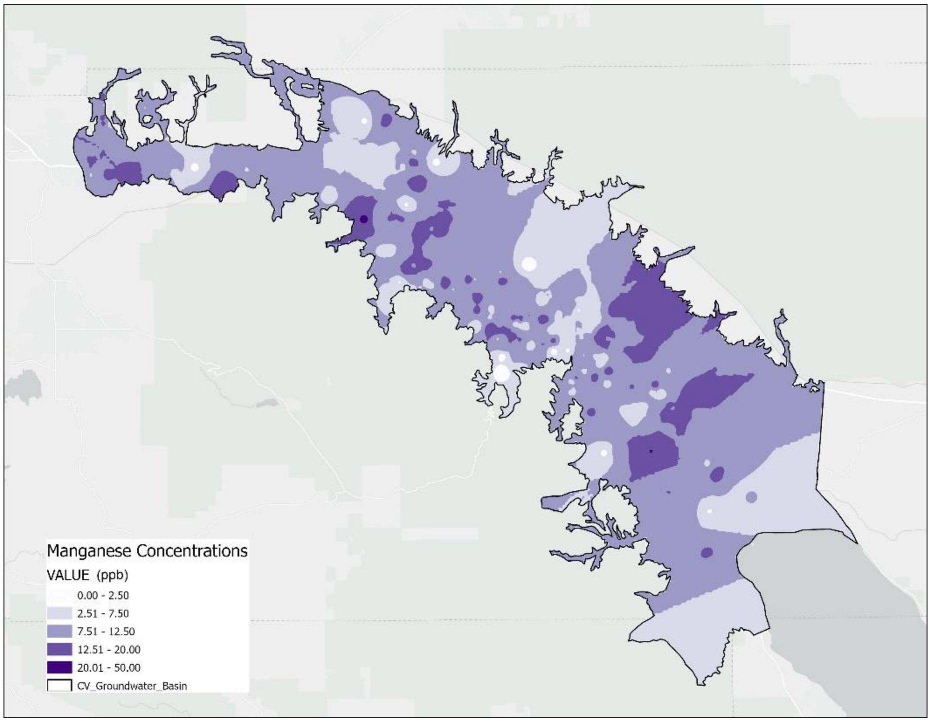

Figure 4. Spatially interpolated map using IDW showing groundwater manganese distribution in Coachella Valley, concentrations of manganese.

3.2 Predicting Groundwater Arsenic through Random Forest Modeling

Although IDW interpolation provides an easily applied method to assess the potential distribution of groundwater chemistry, it does not provide information on the factors likely controlling arsenic distribution contamination in Coachella Valley. The development of a random forest model can both predict high arsenic concentrations in the region along with identifying and ranking predictor variables. Here, we explored the relationship between groundwater arsenic concentration across Coachella Valley, CA using 21 environmental parameters in a RFC modeling approach that has proven high accuracy and superiority with imbalance datasets (More & Rana, 2017).

We utilized recursive feature elimination (RFE) to reduce the number of predictor variables and simplify the model. RFE selected the variables that were most important in predicting arsenic above 5 ppb and selected the number of variables that had the best accuracy. Using RFE on our model resulted in only selecting 2 variables for our model, resulting in an accuracy of 0.9334. Reducing the number of variables in the RFC model from 21 to 2 increased total model accuracy from 0.9296 to 0.9346 and increased *kappa* from 0.4618 to 0.4992.

We continued tuning the model with cross-validation after reducing the number of predictor variables. The main tuning parameter for RF models is mtry, or the number of explanatory variables evaluated at each decision tree split. After ten-times repeated, 10-fold cross validation, the optimal *mtry* value was 1, with an accuracy of 0.9346 and an error rate of 7.08% which are great for a RF model of groundwater contamination. After optimization, a map was generated which is displayed in Figure 5. The performance of the RF model on the test dataset (30% of the data, which was randomly selected while maintaining the relative distribution of high and low values) is summarized in the confusion matrix in Table 2. The probability threshold cutoff was set at 0.50 when applying the *randomForest* package in R, which may not be suitable for our imbalanced dataset (Luo et al., 2019). Not surprisingly, the low prevalence of high values (defined as As concentrations greater than 5 ppb) in the dataset caused the model to perform well in areas with low values with a specificity of 0.97, but poorly in areas with high values with a sensitivity of 0.5.

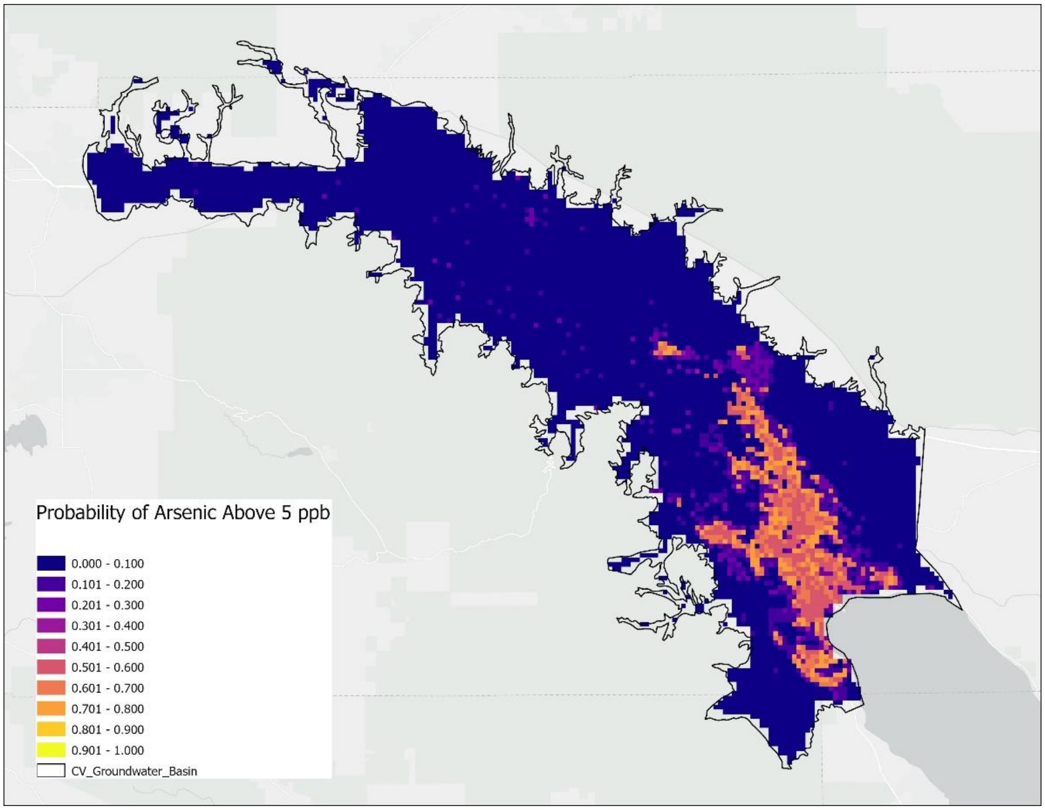

Figure 5. Map of probability of groundwater arsenic concentrations exceeding 5 ppb in Coachella Valley groundwater basin, CA, created with RFC modeling.

Table 2. Confusion matrix for RFC model of As arsenic concentrations above 5 ppb in Coachella Valley.

Model Output	Value before Thresholding	Value after Thresholding
Measured Arsenic > 5ppb		
Predicted Arsenic > 5ppb	10	20
Predicted Arsenic < 5ppb	10	0
Measured Arsenic < 5ppb		
Predicted Arsenic < 5ppb	213	209
Predicted Arsenic > 5ppb	7	11
Sensitivity	0.500	1.000
Specificity	0.968	0.949

The imbalanced data included relatively few values above 5 ppb which lead to the model making a large number of <5 ppb predictions in its classification predictions since there are very few >5 ppb events within the dataset (Yao et al., 2020). To account for the influence of our imbalanced data on classification predictions, a new probability threshold cutoff was selected to improve our model predictions (Yao et al, 2020). We used receiver operating characteristic (ROC) curves to select the best threshold value for our model; a threshold value of 0.179 was selected with a specificity of 0.949 and a sensitivity of 1.000. Application of this revised threshold value greatly improved the model's prediction of As >5 ppb, as shown in the confusion matrix calculated from the revised model output (Table 2). The confusion matrix also revealed that there was a decrease in the model's ability to predict As <5 ppb after the threshold adjustment; however, we focused on optimizing model predictions of areas with As >5 ppb given the main objective is to identify areas of public health concern. To highlight areas where communities may be at risk of arsenic exposure from groundwater contamination, it is more important to optimize for true positives than false negatives, making the current thresholding appropriate. The new map created with the revised ROC determined threshold is provided in Figure 6.

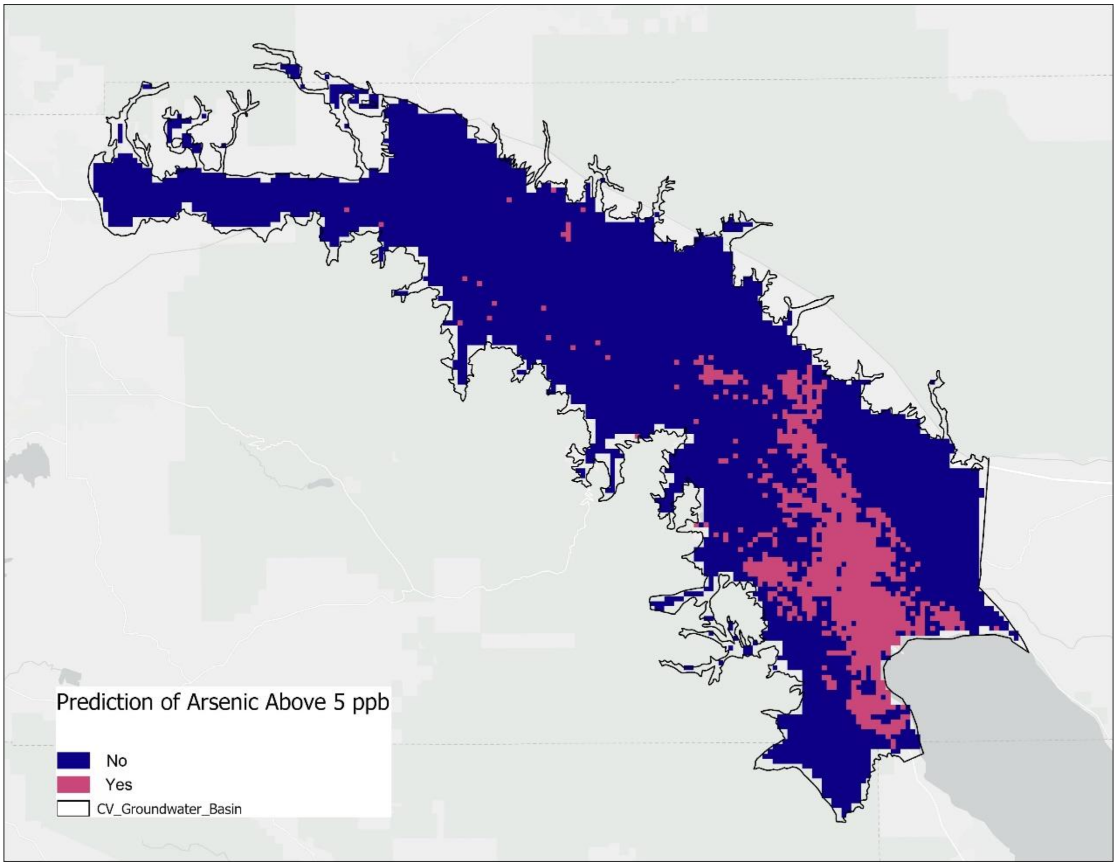

Figure 6. Map of RF classification showing locations where groundwater arsenic concentrations are expected to exceed 5 ppb in Coachella Valley groundwater basin, CA. Receiver operating characteristic (ROC) curves were used to select a model threshold value of 0.179, with a specificity of 0.949 and a sensitivity of 1.000.

The final 2 predictors selected after applying RFE included land use and soil c-horizon bulk density (Figure 7). Groundwater arsenic concentrations above 5 ppb were negatively correlated with c-horizon bulk density, or where bulk density was low, arsenic was high. Bulk density of soils is dependent on factors such as mineral composition, organic matter content, and the size of soil grains (Brogowski et al., 2014). It could be that the soils with low bulk density have mineral compositions relevant to arsenic mobilization and explain this relationship.

Land use, particularly whether or not the land was agricultural use, was also an influential factor for predicting As contamination above 5 ppb in the model. Historically, Coachella Valley was an expansive agricultural area for the state of California, and between the years of 1936 and 1967 the aquifer elevation declined with subsidence up to 6 mm yr^{-1} (Thomas & Famiglietti, 2015). Smith et al. (2018) reported that overpumping of groundwater in the Central Valley led to substantial land subsidence that resulted in the increased As mobilization from subsurface sediments within the aquifer. Although subsidence in Coachella Valley has been halted because of recent managed aquifer recharge efforts, historic overpumping for agricultural use may explain the reason the agricultural land use is a strong predictor of arsenic contaminations above 5 ppb in this region.

It was unexpected that no hydrologic landscape predictors were retained in the final model after RFE. This is likely due to the low spatial resolution of data available for those predictors. Similarly, groundwater chemical composition (e.g., groundwater nitrate, sulfate, Mn, and Fe concentrations) is an important predictor of As contamination, however, groundwater monitoring both spatially and temporally in the Coachella Valley is sparse and data was inconsistently collected for various chemicals. The high level of data missingness required us to remove them from the model as predictors. In general, there are large gaps in the publicly available hydrologic and geochemical datasets for Coachella Valley which greatly constrained the number and type of predictor variables that could be included in the RFC model. Nevertheless, this study demonstrates the results of creating a RFC model using available data to predict As concentrations above 5 ppb in Coachella

Valley, the results of which highlight not only areas that have groundwater approaching the As MCL, but also where additional hydrological and geochemical data collection efforts should be focused.

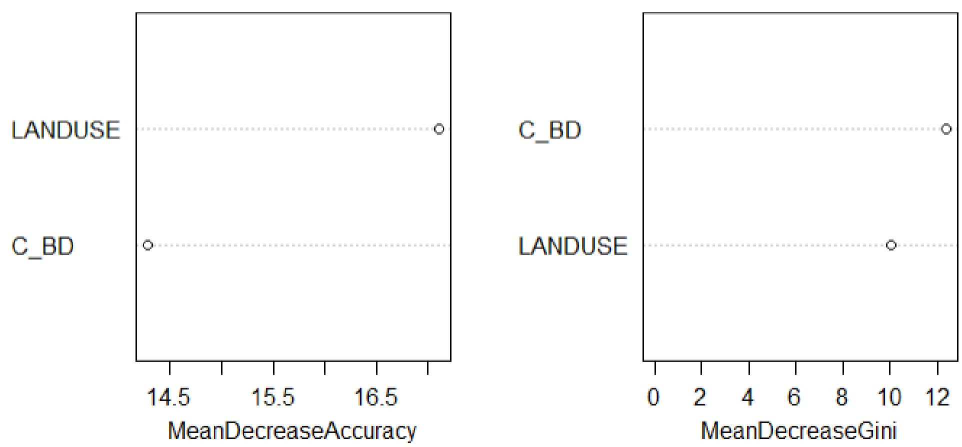

Figure 7. Variable importance plots for RFC of Coachella Valley, CA arsenic concentrations above 5 ppb. LANDUSE is land use of natural, urban, or agricultural, and C_BD is the soil C-horizon bulk density.

4. Summary and Conclusions

Our study utilized random forest classification modeling, a powerful machine learning tool, to model available environmental data and visualize the spatial distribution of potential As groundwater quality hazard on a regional scale. Our study also applied spatial interpolation methods to supplement RFCs which displayed the magnitude and distribution of metals contamination in Coachella Valley by creating maps of the spatial distribution of other groundwater contaminants in the region. Three IDW maps were created that individually show As, Cr, and Mn concentration distributions in the region (Figures 2-4); subsequently, a random forest classification model was created to predict spatial distribution of arsenic concentrations in the region greater than half the As MCL (5 ppb) as a potential proxy for regions that have arsenic approaching the primary MCL (Figures 5-6). Additionally, RF modeling provided a ranking of predictor variables that are best able to explain predicted groundwater distribution (Figure 7) which cannot be achieved with interpolation methods.

The IDW maps supported our hypothesis that groundwater As concentrations would be higher in Eastern Coachella Valley, where the Oasis Mobile Home Park and many others are located. IDW maps also showed that concentrations of Mn and Cr, two other common geogenic groundwater metal contaminants harmful to human health, were inversely related to groundwater As concentrations. This result implies that residents confronted with groundwater As contamination may be relieved from having to also monitor and treat other geogenic metals including Cr and Mn.

The RFC model confirmed that groundwater As contamination is found in concentrations relevant to human health primarily in eastern Coachella Valley. The best predictors for high arsenic concentrations were land use and soil C-horizon bulk density. As with any model, a primary limitation of our study was data availability and quality (Erickson et al., 2021). Furthermore, the factors and conditions that control arsenic mobility can differ significantly between aquifer systems, which prevents the direct application of RFC models developed in one region to be applied to another. The RFC modeling method, however, can be applied to other aquifers by using publicly available predictor variables and groundwater chemical (response variable) data. Nevertheless, the RFC model presented in the current study was reasonably accurate at predicting arsenic concentrations above 5 ppb in Coachella Valley and can be used to identify and prioritize communities that would benefit from routine water quality monitoring.

5. Future Directions

There is evidence that the aquifer conditions within the Coachella Valley groundwater basin are conducive to arsenic mobilization leading to groundwater contamination, yet there is limited spatial and temporal data available on the extent of groundwater arsenic contamination in the region. We recommend that extensive groundwater quality monitoring should be performed in the region to assess potential exposure of residents to groundwater arsenic poisoning, particularly for residents within Eastern Coachella Valley, where there is predicted high arsenic concentrations correlated with larger percentage of low-income families composed of people of color. The eastern portion of the region is also more rural, and there are more residents dependent upon unmonitored, potentially untreated, private groundwater wells to supply drinking water. Due to the sparse spatial distribution of groundwater monitoring wells, there are likely high As concentration hotspots that cannot be identified with the current well network. Instead, efforts and funds should be placed into collecting private well chemistry data, particularly from wells serving *polancos* and similarly vulnerable communities. In parallel to geochemical assessment of the pervasiveness of As contamination in the region, the public health impact of chronic arsenic consumption should be assessed simultaneously, with a particular focus on adverse health effects in children.

To this end, the RFC model produced in this study highlighting regions with higher probability of having arsenic above 5 ppb may be referenced by policy makers and researchers to identify where further studies and funding should be dedicated.

References

Ayotte, J. D., Belaval, M., Olson, S. A., Burow, K. R., Flanagan, S. M., Hinkle, S. R., & Lindsey, B. D. (2015). Factors affecting temporal variability of arsenic in groundwater used for drinking water supply in the United States. *Science of The Total Environment*, *505*, 1370–1379. https://doi.org/10.1016/j.scitotenv.2014.02.057

Ayotte, J. D., Nolan, B. T., & Gronberg, J. A. (2016). Predicting Arsenic in Drinking Water Wells of the Central Valley, California. *Environmental Science & Technology*, *50*(14), 7555–7563. https://doi.org/10.1021/acs.est.6b01914

Bahl, A., Hellack, B., Balas, M., Dinischiotu, A., Wiemann, M., Brinkmann, J., Luch, A., Renard, B. Y., & Haase, A. (2019). Recursive feature elimination in random forest classification supports nanomaterial grouping. *NanoImpact*, *15*, 100179. https://doi.org/10.1016/j.impact.2019.100179

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Coyte, R. M., & Vengosh, A. (2020). Factors Controlling the Risks of Co-occurrence of the Redox-Sensitive Elements of Arsenic, Chromium, Vanadium, and Uranium in Groundwater from the Eastern United States. *Environmental Science & Technology*, *54*(7), 4367–4375. https://doi.org/10.1021/acs.est.9b06471

Elumalai, V., Brindha, K., & Lakshmanan, E. (2017). Human Exposure Risk Assessment Due to Heavy Metals in Groundwater by Pollution Index and Multivariate Statistical Methods: A Case Study from South Africa. *Water*, *9*(4), 234. https://doi.org/10.3390/w9040234

Erickson, M. L., Elliott, S. M., Brown, C. J., Stackelberg, P. E., Ransom, K. M., Reddy, J. E., & Cravotta, C. A. (2021). Machine-Learning Predictions of High Arsenic and High Manganese at Drinking Water Depths of the Glacial Aquifer System, Northern Continental United States. *Environmental Science & Technology*, *55*(9), 5791–5805. https://doi.org/10.1021/acs.est.0c06740

Fendorf, S., Michael, H. A., & van Geen, A. (2010). Spatial and Temporal Variations of Groundwater Arsenic in South and Southeast Asia. *Science*, *328*(5982), 1123–1127. https://doi.org/10.1126/science.1172974

Gao, S., Fujii, R., Chalmers, A. T., & Tanji, K. K. (2004). Evaluation of Adsorbed Arsenic and Potential Contribution to Shallow Groundwater in Tulare Lake Bed Area, Tulare Basin, California. *Soil Science Society of America Journal*, *68*(1), 89–95. https://doi.org/10.2136/sssaj2004.8900

Goldberg, S. (1986). Chemical Modeling of Arsenate Adsorption on Aluminum and Iron Oxide Minerals. *Soil Science Society of America Journal*, *50*(5), 1154–1157. https://doi.org/10.2136/sssaj1986.03615995005000050012x

Kondakis, X. G., Makris, N., Leotsinidis, M., Prinou, M., & Papapetropoulos, T. (1989). Possible Health Effects of High Manganese Concentration in Drinking Water. *Archives of Environmental Health: An International Journal*, *44*(3), 175–178. https://doi.org/10.1080/00039896.1989.9935883

Liévanos, R. (2018). Retooling CalEnviroScreen: Cumulative Pollution Burden and Race-Based Environmental Health Vulnerabilities in California. *International Journal of Environmental Research and Public Health*, *15*(4), 762. https://doi.org/10.3390/ijerph15040762

Lombard, M. A., Bryan, M. S., Jones, D. K., Bulka, C., Bradley, P. M., Backer, L. C., Focazio, M. J., Silverman, D. T., Toccalino, P., Argos, M., Gribble, M. O., & Ayotte, J. D. (2021). Machine Learning Models of Arsenic in Private Wells Throughout the Conterminous United States As a Tool for Exposure Assessment in Human Health Studies. *Environmental Science & Technology*, *55*(8), 5012–5023. https://doi.org/10.1021/acs.est.0c05239

Lopez, A. M., Wells, A., & Fendorf, S. (2021). Soil and Aquifer Properties Combine as Predictors of Groundwater Uranium Concentrations within the Central Valley, California. *Environmental Science & Technology*, *55*(1), 352–361. https://doi.org/10.1021/acs.est.0c05591

Lu, G. Y., & Wong, D. W. (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Computers & Geosciences*, *34*(9), 1044–1055. https://doi.org/10.1016/j.cageo.2007.07.010

Luo, H., Pan, X., Wang, Q., Ye, S., & Qian, Y. (2019). Logistic Regression and Random Forest for Effective Imbalanced Classification. *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, 916–917. https://doi.org/10.1109/COMPSAC.2019.00139

Magalhães, M. C. F. (2002). Arsenic. An environmental problem limited by solubility. *Pure and Applied Chemistry*, *74*(10), 1843–1850. https://doi.org/10.1351/pac200274101843

Mirzaei, R., & Sakizadeh, M. (2016). Comparison of interpolation methods for the estimation of groundwater contamination in Andimeshk-Shush Plain, Southwest of Iran. *Environmental Science and Pollution Research*, *23*(3), 2758–2769. https://doi.org/10.1007/s11356-015-5507-2

More, A. S., & Rana, D. P. (2017). Review of random forest classification techniques to resolve data imbalance. *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, 72–78. https://doi.org/10.1109/ICISIM.2017.8122151

Mukhija, V., & Mason, D. R. (2015). Resident-Owned, Informal Mobile Home Communities in Rural California: The Case of Rancho Don Antonio, Coachella Valley. *Housing Policy Debate*, *25*(1), 179–194. https://doi.org/10.1080/10511482.2014.921220

Nath, B. K., Chaliha, C., Bhuyan, B., Kalita, E., Baruah, D. C., & Bhagabati, A. K. (2018). GIS mapping-based impact assessment of groundwater contamination by arsenic and other heavy metal contaminants in the Brahmaputra River valley: A water quality assessment study. *Journal of Cleaner Production*, *201*, 1001–1011. https://doi.org/10.1016/j.jclepro.2018.08.084

Naujokas, M. F., Anderson, B., Ahsan, H., Aposhian, H. V., Graziano, J. H., Thompson, C., & Suk, W. A. (2013). The Broad Scope of Health Effects from Chronic Arsenic Exposure: Update on a Worldwide Public Health Problem. *Environmental Health Perspectives*, *121*(3), 295–302. https://doi.org/10.1289/ehp.1205875

Pace, C., Balazs, C., Bangia, K., Depsky, N., Renteria, A., Morello-Frosch, R., & Cushing, L. J. (2022). Inequities in Drinking Water Quality Among Domestic Well Communities and Community Water Systems, California, 2011–2019. *American Journal of Public Health*, *112*(1), 88–97. https://doi.org/10.2105/AJPH.2021.306561

Podgorski, J., & Berg, M. (2020). Global threat of arsenic in groundwater. *Science*, *368*(6493), 845–850. https://doi.org/10.1126/science.aba1510

Ritter, Keith Solomon, Paul Sibley, L. (2002). SOURCES, PATHWAYS, AND RELATIVE RISKS OF CONTAMINANTS IN SURFACE WATER AND GROUNDWATER: A PERSPECTIVE PREPARED FOR THE WALKERTON INQUIRY. *Journal of Toxicology and Environmental Health, Part A*, *65*(1), 1–142. https://doi.org/10.1080/152873902753338572

Schaefer, M. V., Plaganas, M., Abernathy, M. J., Aiken, M. L., Garniwan, A., Lee, I., & Ying, S. C. (2020). Manganese, Arsenic, and Carbonate Interactions in Model Oxic Groundwater Systems. *Environmental Science & Technology*, *54*(17), 10621–10629. https://doi.org/10.1021/acs.est.0c02084

Scornet, E., Biau, G., & Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, *43*(4). https://doi.org/10.1214/15-AOS1321

Smedley, P. L., & Kinniburgh, D. G. (2002). A review of the source, behaviour and distribution of arsenic in natural waters. *Applied Geochemistry*, *17*(5), 517–568. https://doi.org/10.1016/S0883-2927(02)00018-5

Smith, R., Knight, R., & Fendorf, S. (2018). Overpumping leads to California groundwater arsenic threat. *Nature Communications*, *9*(1), 2089. https://doi.org/10.1038/s41467-018-04475-3

Thomas, B. F., & Famiglietti, J. S. (2015). Sustainable Groundwater Management in the Arid Southwestern US: Coachella Valley, California. *Water Resources Management*, *29*(12), 4411–4426. https://doi.org/10.1007/s11269-015-1067-y

Yao, H., Li, X., Pang, H., Sheng, L., & Wang, W. (2020). Application of random forest algorithm in hail forecasting over Shandong Peninsula. *Atmospheric Research*, *244*, 105093. https://doi.org/10.1016/j.atmosres.2020.105093

Yao, X., Fu, B., Lü, Y., Sun, F., Wang, S., & Liu, M. (2013). Comparison of Four Spatial Interpolation Methods for Estimating Soil Moisture in a Complex Terrain Catchment. *PLoS ONE*, *8*(1), e54660. https://doi.org/10.1371/journal.pone.0054660

Ying, S. C., Schaefer, M. V., Cock-Esteb, A., Li, J., & Fendorf, S. (2017). Depth Stratification Leads to Distinct Zones of Manganese and Arsenic Contaminated Groundwater. *Environmental Science & Technology*, *51*(16), 8926–8932. https://doi.org/10.1021/acs.est.7b01121

Supplemental Information

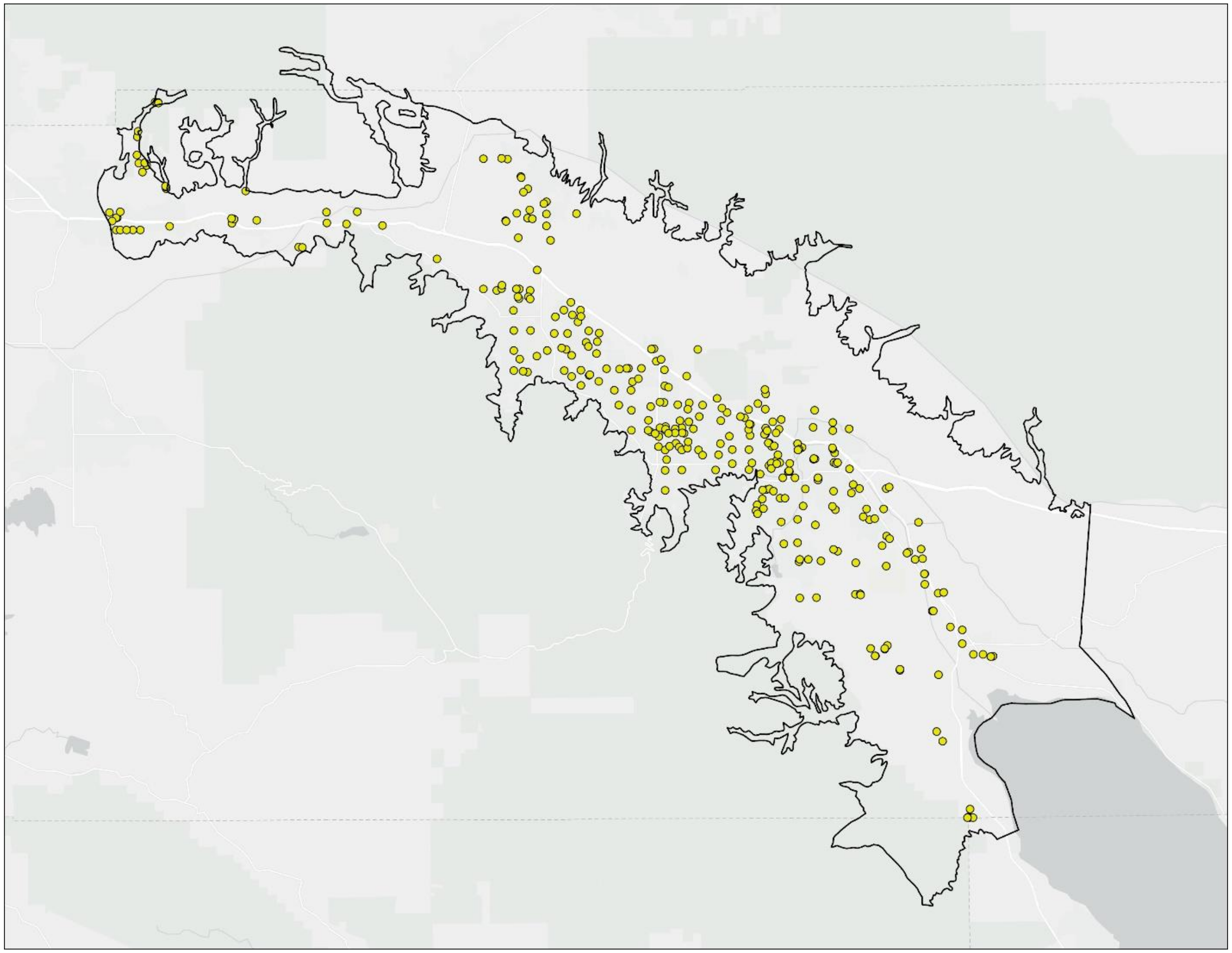

Figure S1. Map of Coachella Valley showing where our 343 wells were located that were used for random forest modeling and inverse distance weighting taken from the California State Water Boards GAMA program.

Table S1. Data Sources for IDWs and RFC model.

Variable	**Description**	**Source**
ARSENIC	Groundwater arsenic concentration, ug/L	HTTPS://GAMAGROUNDWATER.WATERBOARDS.CA.GOV/GAMA/GAMAMAP/PUBLIC/
CHROMIUM	Groundwater chromium concentration, ug/L	HTTPS://GAMAGROUNDWATER.WATERBOARDS.CA.GOV/GAMA/GAMAMAP/PUBLIC/
HARDNESS	Groundwater hardness, mg/L	HTTPS://GAMAGROUNDWATER.WATERBOARDS.CA.GOV/GAMA/GAMAMAP/PUBLIC/
IRON	Groundwater iron concentration, ug/L	HTTPS://GAMAGROUNDWATER.WATERBOARDS.CA.GOV/GAMA/GAMAMAP/PUBLIC/
MANGANESE	Groundwater manganese concentration, ug/L	HTTPS://GAMAGROUNDWATER.WATERBOARDS.CA.GOV/GAMA/GAMAMAP/PUBLIC/
NITRATE	Groundwater nitrate (as N) concentration, mg/L	HTTPS://GAMAGROUNDWATER.WATERBOARDS.CA.GOV/GAMA/GAMAMAP/PUBLIC /
PH	Groundwater pH, pH units	HTTPS://GAMAGROUNDWATER.WATERBOARDS.CA.GOV/GAMA/GAMAMAP/PUBLIC/

SULFATE	Groundwater sulfate concentration, mg/L	HTTPS://GAMAGROUNDWATER.WATERBOARDS.CA.GOV/GAMA/GAMAMAP/PUBLIC/
C_BD	Soil bulk density, C-horizon, g/cm3	https://acsess.onlinelibrary.wiley.com/doi/10.2136/sssaj2017.04.0122
C_CLAY	Soil clay content, C-horizon, weight percent	https://acsess.onlinelibrary.wiley.com/doi/10.2136/sssaj2017.04.0122
C_EC	Soil electrical conductivity, C-horizon, σ	https://acsess.onlinelibrary.wiley.com/doi/10.2136/sssaj2017.04.0122
C_PH	Soil pH, C-horizon, pH units	https://acsess.onlinelibrary.wiley.com/doi/10.2136/sssaj2017.04.0122
C_SAND	Soil sand content, C-horizon, weight percent	https://acsess.onlinelibrary.wiley.com/doi/10.2136/sssaj2017.04.0122
C_SOC	Soil organic carbon concentration, C-horizon, weight percent	https://acsess.onlinelibrary.wiley.com/doi/10.2136/sssaj2017.04.0122
WellDepth_Dom_TopOpen	Domestic-supply well depth to the top of the open interval	https://www.sciencebase.gov/catalog/item/5e43efc3e4b0edb47be84c3d
WelDepth_Dom_BottomOpen	Domestic-supply well depth to the bottom of the open interval	https://www.sciencebase.gov/catalog/item/5e43efc3e4b0edb47be84c3d

WellDepth_Pub_ TopOpen	Public-supply well depth to the top of the open interval	https://www.sciencebase.gov/catalog/item/5e43efc3e4b0edb47be84c3d
WellDepth_Pub_ BottomOpen	Public-supply well depth to the bottom of the open interval	https://www.sciencebase.gov/catalog/item/5e43efc3e4b0edb47be84c3d
PFLATTOT	Total percent flat land (slope less than 1 percent) in watershed	https://water.usgs.gov/GIS/metadata/usgswrd/XML/hlrus.xml#stdorder
RELIEF	Maximum elevation in watershed minus minimum elevation in watershed in meters	https://water.usgs.gov/GIS/metadata/usgswrd/XML/hlrus.xml#stdorder
MINELE	Minimum elevation in watershed in meters	https://water.usgs.gov/GIS/metadata/usgswrd/XML/hlrus.xml#stdorder
PSAND	Percent sand in soil in watershed	https://water.usgs.gov/GIS/metadata/usgswrd/XML/hlrus.xml#stdorder
SLOPE	Slope in percent rise in watershed	https://water.usgs.gov/GIS/metadata/usgswrd/XML/hlrus.xml#stdorder
AQPERM	Aquifer permeability class, 1-7 low to high in watershed	https://water.usgs.gov/GIS/metadata/usgswrd/XML/hlrus.xml#stdorder
HLR	Hydrologic landscape region identification number in watershed	https://water.usgs.gov/GIS/metadata/usgswrd/XML/hlrus.xml#stdorder

TMEAN	Mean temperature, °F	https://prism.oregonstate.edu/normals/
PPT	Precipitation, in.	https://prism.oregonstate.edu/normals/
GW_R	Mean annual groundwater recharge, BFI	https://water.usgs.gov/GIS/metadata/usgswrd/XML/rech48grd.xml
LANDUSE	Landuse, numerical 1-3 for natural, urban, and agricultural	https://databasin.org/datasets/f12a901528c0498ba63ca291b8e6627b/

Equations of Error Statistics for IDWs

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Where y_i is the predicted value, x_i is the true value, and n is the number of data points.

$$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2]^{1/2}$$

Where n is the number of validation points, p_i is the predicted value at point 1, o_i is the overserved value at point i.

$$G = 1 - [\sum_{i=1}^{n}(p_i - o_i)^2 / \sum_{i=1}^{n}(o_i - \overline{o})^2]$$

Where n is the number of validation points, p_i is the predicted value at point 1, o_i is the overserved value at point i, and $\bar{o}$ is the sample arithmetic mean.