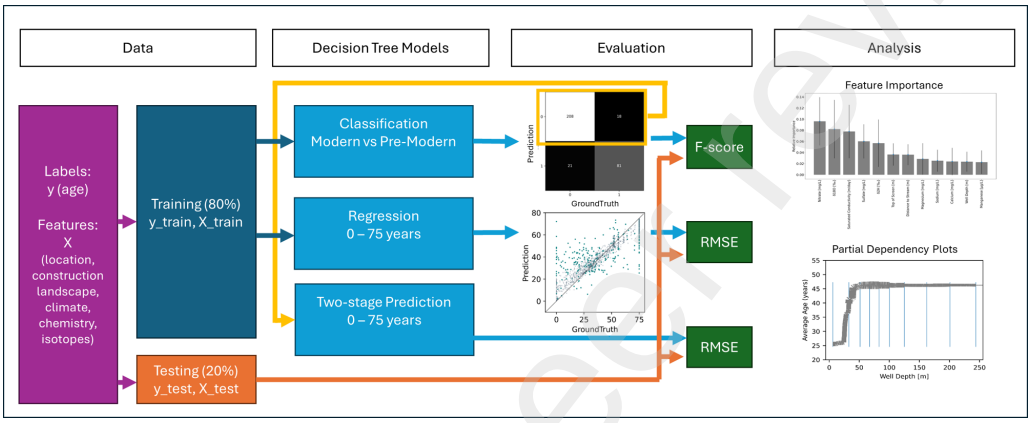
Graphical Abstract

Machine Learning Prediction of Tritium-Helium Groundwater Ages in the Central Valley, California, USA

Abdullah Azhar, Indrasis Chakraborty, Ate Visser, Yang Liu, Jory Chapin Lerback, Erik Oerter



Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

Highlights

Machine Learning Prediction of Tritium-Helium Groundwater Ages in the Central Valley, California, USA

Abdullah Azhar, Indrasis Chakraborty, Ate Visser, Yang Liu, Jory Chapin Lerback, Erik Oerter

- Six machine learning (ML) decision tree (DT) models were trained on tritium, groundwater age and 63 features.
- The Bagging Regressor DT model accurately classifies groundwater as either modern or pre-modern.
- ML prediction of tritium-helium groundwater ages explains $\sim 30\%$ of variability.
- Most important features are related to recharge source, contaminants, aquifer materials, well depth, and geochemical reactions.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

Machine Learning Prediction of Tritium-Helium Groundwater Ages in the Central Valley, California, USA

Abdullah Azhar[a], Indrasis Chakraborty[a], Ate Visser[a], Yang Liu[a], Jory Chapin Lerback[a], Erik Oerter[a]

[a]*Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, 94550, CA, USA*

Abstract

Groundwater ages provides insight into recharge rates, flow velocities, and vulnerability to contaminants. The ability to predict groundwater ages based on more accessible parameters via Machine Learning (ML) would advance our ability to guide sustainable management of groundwater resources. In this study, ML models were trained and tested on a large dataset of tritium concentrations ($n = 2410$) and tritium-helium groundwater ages ($n = 1157$) from the California Central Valley, a large groundwater basin with complex land use, irrigation, and water management practices. The ML models were trained on 63 features, including location, well construction information, landscape characteristics, and climate variables, water chemistry, and stable isotopes. The Bagging Regressor ML method can accurately classify (F1-score = 0.91) groundwater samples as either modern or pre-modern whereas the accuracy of the ML prediction of continuous tritium-helium groundwater ages is limited and explains only $\sim 30\%$ of the variability in this dataset. In general, ML groundwater age prediction relies mostly on features related to (1) the source of groundwater recharge, (2) contaminant history, (3) aquifer materials, (4) well construction, and (5) geochemical reactions along flow paths.

Keywords: machine learning, ensemble learning, decision tree, groundwater age, tritium-helium, California, USA

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

1. Introduction

Groundwater supplies nearly half of all drinking water, approximately 40% of water used for irrigation, and a third of industrial water use [1]. Globally, groundwater is impacted by both depletion and pollution, threatening the water security of people around the world [2]. Protecting groundwater resources and guiding sustainable use requires a deep understanding of groundwater fluxes (recharge rates, flow velocities) and the fate and transport of contaminants in groundwater. One of the tools to advance our understanding of groundwater systems is the age of groundwater, which is typically derived from environmental tracers [3] or model simulations [4] of groundwater flow. The age of groundwater relates directly to groundwater recharge rates and flow velocities and can serve to calibrate numerical flow and transport models [5], or calculate aquifer turnover times [6]. Understanding the relationship between groundwater quality and groundwater age also underpins evaluations of well vulnerability to contaminants [7, 8]. Groundwater age can be derived from environmental tracers such as radioactive isotopes (e.g., tritium [9], tritium-helium [10], carbon-14 [11] or transient anthropogenic gases or isotopes (e.g., CFCs [12, 13] and SF6 [14], krypton-85 [15]). Each of these environmental tracers requires an expensive laboratory analysis to calculate a groundwater age. The ability to predict groundwater ages based on more accessible parameters (e.g., well characteristics, ion chemistry, climate, and/or land use information) would advance our ability to guide sustainable management of groundwater resources. Machine Learning (ML) approaches can discern patterns in "large" and "complex" datasets to make more accurate predictions of variables of interest (referred to as "labels") based on other independent variables (referred to as "features"). ML techniques have been applied to predict changes in total water storage [16] as well as groundwater quality [17]; Singha, [18]. For example, ML has been applied to predict nitrate concentrations in the Central Valley [19] or groundwater ages in the Great Lakes Basin [20]. In addition to producing a prediction of groundwater age in the absence of environmental tracer information, examining which features are predictive of groundwater age provides deeper insights into the groundwater recharge sources, groundwater flow system, contaminant history, and geochemical reactions along flow paths. Due to the expense and labor limitations constraints, few groundwater age datasets exist that are sufficiently large to train a ML model. One prior study predicted groundwater ages in the Great Lakes Basin (GLB) using a gradient boosted regression

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

approach [20] based on 961 tracer-based groundwater ages. The dataset collected by the California Waterboards for the Groundwater Ambient Monitoring and Assessment (GAMA) program includes 2410 tritium measurements and 1157 tritium-helium groundwater ages, with associated well construction information and analyses of groundwater contaminants and ion chemistry. We used this dataset to train various ML models to address the following research questions:

- How accurately can groundwater age be predicted using various ML models? Here we focus on how well these models can classify whether a groundwater sample is pre-modern (recharged before 1950) or modern (recharged after 1950) and (b) how well a regression model can predict groundwater age, including pre-modern and modern samples. Addressing this question will guide future applications of ML to predict groundwater ages in their choice of ML model and a priori expectations of accuracy.

- Which features are most important to classify or regress groundwater age? Addressing this question will enhance our understanding of the Central Valley groundwater system and guide future studies to focus on key aspects of groundwater systems that control groundwater age.

2. Methods

2.1. Machine Learning and Decision Tree Models

The idea of using machine learning to predict groundwater ages was first introduced by [20], where the authors used a gradient boosted decision trees (GBDTs) to predict groundwater ages in the Great Lakes Basin Region. Decision trees (DTs) make predictions by making a series of binary decisions (splits) in a single dimension that together bisects the d-dimensional feature space into axis-aligned sub-spaces (nodes). Predictions are made from leaf nodes where further splitting is not possible (e.g. due to depth limit or only a single sample is in the node). In addition to good performance, DTs are considered inherently interpretable as the decisions can be written as simple logic rules. While decision trees were initially introduced for classification tasks, they can also be used for predicting continuous outputs by estimating piece-wise constant functions. For a DT regressor, a split in the tree is made in a **single feature dimension** at a **specific threshold** such that it

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

maximizes the RMSE reduction of the whole training set. For a training set with n samples of d dimensions, the training time complexity is $O(n \log(n) d)$ and test time complexity is proportional to the tree depth ($O(depth)$).

Similar to many ML algorithms, DTs may suffer from significant overfitting. For instance, a DT that has n leaf nodes (1 sample per node) memorizes the training set and achieves perfect training performance but usually does not generalize well. To address this, DT-based ensemble methods combine weak DTs to reduce overfitting: Bagging [21] and Random forests (RF) [22] aggregate weak learners in parallel, whereas AdaBoost [23] and Gradient boosting (GB) [24] build the ensemble sequentially. In this paper, the base DT model as well as the aforementioned DT-based ensemble models are benchmarked for the age prediction problem.

The core ML challenge of this study can be summarized as a classic supervised learning problem:

$$y = f(X)$$

where an ML model $\mathbf{f}$ is trained to predict ground-truth ages $\mathbf{y}$ computed by tritium-helium groundwater age dating using features $\mathbf{X}$ including chemical and isotopic analyses.

2.2. Tritium-helium Groundwater Age Dating

Multiple tracer systems have been developed to gain information about groundwater age based on chemical or isotopic analyses. Tritium-helium groundwater dating relies on the decay of tritium (a radioactive isotope of hydrogen with a 12.32 year half-life [25]) to helium-3 (a stable noble gas isotope), and the measurement of both tritium and helium-3 in groundwater samples. The apparent tritium-helium groundwater age (τ) is calculated as $\tau = \ln\left({}^3\mathrm{He}_{\mathrm{trit}}/{}^3\mathrm{H} + 1\right)/\lambda$, where the tritium decay-constant $\lambda = 0.05626a^{-1}$. Tritium-helium groundwater dating has been applied to understand groundwater flow velocities, examine trends in groundwater quality [7, 8], and basin characterization. The apparent tritium-helium age generally reflects the mean of the distribution of ages in a groundwater sample, but it can be affected by mixing of flow paths with different initial tritium concentrations. This particularly affects ages that contain a component of water that recharged during the "bomb-pulse" period (1950s-1970s) when tritium levels in precipitation were several orders of magnitude higher due to above-ground nuclear weapons testing [26, 27]. In aquifers with highly contrasting permeabilities, tritium and helium can diffuse into stagnant zones at different rates,

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

resulting in apparent age biases of several years [28]. The 12.32 year half-life and the bomb-pulse in precipitation practically limit the $^3H/^3He$ dating method to groundwater that recharged since 1950. The absence of tritium in groundwater (below 1 pCi/L) indicates that the sampled groundwater recharged entirely before 1950 and can be classified as "**pre-modern**" for the binary ML classification.

2.3. Study Area and Data Sources

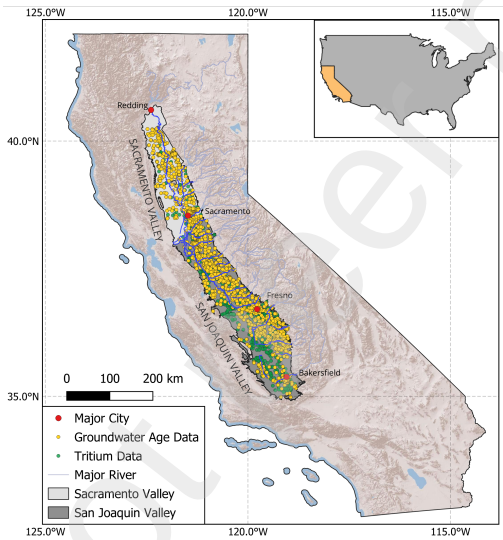


Figure 1: Map of the Central Valley, indicting major cities, and the locations of groundwater age data and tritium data.

The Central Valley is a 700 km long and 80 km wide valley in California between the Coast Ranges to the west, the Cascades to the North, and the Sierra Nevada mountains to the east. The Central Valley has a mediterranean climate with a mean annual precipitation varying from 150 mm in the south to 830 mm in the north [29]. Average daily summer (July) high temperatures range from 31°C to 37°C and daily low temperatures are between 2°C and 5°C in winter (January) [29]. Potential annual evapotranspiration ranges from 706 to 957 mm/year, with the highest rates in areas with lowest precipitation [30]. Potential ET exceeds precipitation in 90% of the Central Valley. The Central Valley is a highly productive agricultural area thanks

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

to an intricate systems of dams and reservoirs on the major rivers draining the Cascades and Sierra Nevada and canals to deliver river water for agricultural irrigation. River water irrigation is supplemented with groundwater pumping, in particular in dry years when surface water deliveries are limited. Groundwater is the only source of irrigation in 35% of the Central Valley. Besides agriculture, major cities, small communities, and private domestic well owners rely on groundwater for drinking water needs.

The Central Valley groundwater system is subdivided into the Sacramento groundwater basin north of the San Francisco Bay Delta and San Joaquin groundwater basin south of the Delta [31]. Groundwater basins are recharged by infiltration of local precipitation and river water, via direct riverbank infiltration, dedicated managed aquifer recharge (MAR) facilities, and infiltration of agricultural irrigation. In the San Joaquin Valley, a regional analysis of groundwater isotopic data showed that river water recharge accounts for nearly 50% of all groundwater that recharged since 1950 [32].

Groundwater basins, in particular in the San Joaquin Valley, are threatened by declining groundwater levels due to pumping in excess of recharge rates (overdraft) [33]. The decline of groundwater levels [34] has caused shallower domestic wells to go dry [35], particularly during periods of drought [36]. Since 2014, 3894 domestic groundwater wells in the San Joaquin Valley have been reported to the voluntary statewide Dry Well Reporting Tool [37]. It is estimated that nearly 20% of wells in the Central Valley has gone dry [38].

In addition, several contaminants threaten groundwater quality in the Central Valley, such as nitrate [39, 40], arsenic [41, 42], and uranium [43].

In response to concerns over groundwater quality, the California Waterboards started a comprehensive groundwater quality monitoring program (Groundwater Ambient Monitoring and Assessment, GAMA) in 2000 [44]. The GAMA Program is based on interagency collaboration with the State and Regional Water Boards, Department of Water Resources, Department of Pesticide Regulations, U.S. Geological Survey, and Lawrence Livermore National Laboratory, University of CA and CA State University, and cooperation with local water agencies and well owners. As part of the GAMA Program, 3170 public supply wells and 1147 domestic wells have been sampled and the analysis are available online [45]. Ongoing efforts focus on shallow aquifer water quality and detecting long term water quality trends. In addition to water quality analyses (such as nitrate, arsenic, and uranium), the GAMA program collected full ion chemistry, trace element, trace organic

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

contaminants, and isotopic tracers, including the stable hydrogen and oxygen isotope ratios of water (referred to here as "stable isotopes"), tritium (a radioactive isotope of hydrogen with a 12.3 year half-life), dissolved noble gases including the helium isotope ratio, and carbon isotopes (both stable carbon-13 and radioactive carbon-14). This effort has produced an unparalleled dataset which has been examined in numerous reports and scientific publications [46], including a spatial analysis of tritium and groundwater ages across the state.

2.4. Dataset

Tritium-helium Groundwater Age (**y**): All tritium and noble gas data collected from wells in the Central Valley groundwater basins from the GAMA dataset [47] were included for this study. Additional tritium, noble gas, and stable isotope analyses were obtained from the USGS National Water Information System (NWIS) [48] and prior LLNL studies that were not in the GAMA database [19, 49]. Tritium-helium ages were calculated following the procedures developed for the geospatial analysis of California groundwater ages [50]. Due to the limitation in tritium-helium dating, a few pre-processing steps are applied to ensure only samples with reliable age are included. 1) Age uncertainties were derived from propagated measurement uncertainties and samples with age uncertainties greater than 10 years or 25% of the value were excluded. 2) Ages that correspond to a recharge date prior to 1950, with a tritium concentration above 1 pCi/L, were also excluded. 3) Pre-modern samples with a tritium concentration of less than 1 pCi/L were assigned an age of 75 years. The result is a dataset of 1640 groundwater ages, including 1157 samples with reliable tritium-helium groundwater ages (¡75 years) and 483 samples which were assigned an age of 75 years. For the binary classification task, pre-modern samples were assigned a label of 1 ($n = 483$) and modern samples a label of 0 ($n = 1927$).

Features (X): Table 1 summarizes 64 features divided into 3 groups: 1) features related to location (latitude and longitude), well construction information (total well depth, top and bottom depth of perforations), landscape characteristics, land use, and climate (n=32), 2) features related to water chemistry (n=29), and 3) features related to water stable isotopes (n=3).

Groundwater ages were associated with chemical and isotopic analyses collected from the sample well on the same date (available for most GAMA samples) or with the analysis that was conducted closest in time within a 10 years of the age sample. Chemical analyses which were reported with a

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

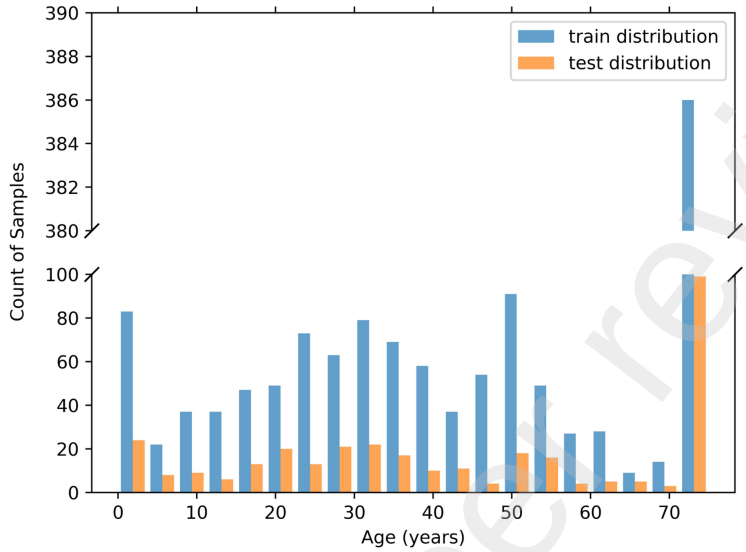


Figure 2: Age Distribution of training and validation sets. As tritium-helium dating is unable to estimate groundwater recharged before 1950, these "pre-modern" data points are assigned a fixed age of 75. The continuous age distribution was binned with 20 equal-sized bins from 0-75.

value below the laboratory detection limit were set to zero. If the reported detection exceeded a reasonable threshold (listed in Suppl. Info.), the data was removed. Missing values were imputed with the mean value of the dataset for each feature.

Table 1: Table 1: List of Labels and Features used in the ML classification of age category (based on tritium) and ML regression of tritium-helium (^{3}H/^{3}He) groundwater ages. A detailed list of Labels and Features with analytical detection limits and data sources is presented in the Supplementary Information. The tritium-helium age Label includes 483 pre-modern samples (^{3}H ¡ 1 pCi/L) that were set to 75 years for the age regression.

Name	**Number**	**Group**
Labels		
3H/3He age [years]	1640	Label for age regression

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

Tritium [pCi/L]	2410	Label for binary classification
Features		
Latitude [°]	2410	1
Longitude [°]	2410	1
Top of Screen [m]	1059	1
Bottom of Screen [m]	928	1
Nitrate [mg/L]	2410	2
Tetrachloroethylene [μg/L]	1653	2
Trichloroethylene [μg/L]	1654	2
Dibromochloropropane [μg/L]	1206	2
Uranium [pCi/L]	1676	2
1,2,3,-Trichloropropane [μg/L]	1598	2
Specific Conductivity [μS/cm]	2410	2
pH	2410	2
Alkalinity [mg/L]	2238	2
Bicarbonate Alkalinity [mg/L]	878	2
Calcium [mg/L]	2410	2
Magnesium [mg/L]	2410	2
Sodium [mg/L]	2410	2
Potassium [mg/L]	2410	2
Chloride [mg/L]	2410	2
Sulfate [mg/L]	2410	2
Fluoride [mg/L]	2201	2
Boron [mg/L]	2410	2
Total Dissolved Solids [mg/L]	2252	2
Arsenic [μg/L]	2399	2
Manganese [μg/L]	2196	2
Perchlorate [μg/L]	1233	2
Hexavalent Chromium [μg/L]	1099	2
Iron [μg/L]	2381	2
Radon [pCi/L]	339	2
Temperature [°C]	2039	2
δ^2H [‰]	2410	3
δ^{18}O [‰]	2410	3

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

Dissolved Oxygen [mg/L]	888	2
Nitrate+Nitrite [mg/L]	782	2
Carbonate Alkalinity [mg/L]	1765	2
Well Depth [m]	2371	1
Ground Elevation [m]	2410	1
Groundwater Elevation [m]	2410	1
Depth to Water [m]	2410	1
Depth to Water minus Total Well Depth [m]	2371	1
above Corcoran Clay (0: no CC, 1: below CC, 2: above CC, NA: no depth)	2375	1
Noble Gas Recharge Temperature [°C]	1603	4
"Excess Air" [% ΔNe]	1640	4
Terrigenic Helium [cm3STP/g]	1639	4
Deuterium Excess [‰]	2410	3
Precipitation [mm/year]	2410	1
Mean Annual Air Temperature [°C]	2410	1
Distance to Stream [m]	2410	1
% water within 500m radius	2410	1
% developed within 500m radius	2410	1
% semideveloped within 500m radius	2410	1
% crops within 500m radius	2410	1
% pasture within 500m radius	2410	1
% grazing within 500m radius	2410	1
% lowuse within 500m radius	2410	1
Runoff Coefficient	2410	1
Evapotranspiration [mm/year]	2410	1
Effective Recharge [mm/year]	2410	1
Saturated Conductivity [m/day]	2410	1
Added Irrigation [mm/year]	2410	1
Slope [°]	2410	1
Mean Annual Air Temperature [°C]	2410	1
Precipitation [mm/year]	2410	1
Potential Evapotranspiration [mm/year]	2410	1
Soil Sand Fraction [%]	2410	1

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

Precipitation minus PET [mm/year]	2410	1
Relief [m]	2410	1

Following the above-mentioned pre-processing including log-normalization, the final dataset included 1,640 samples and 63 features which were then split into train and test sets with an 80/20 split ratio. Figure 2 shows the age distribution for both the train and the test dataset, the close match of the shapes of the distribution of the train and test set. The curated dataset has a mean age of 44.69 years with a standard deviation of 24.85 years. The latter also serves as a baseline root-mean-squared error (RMSE) from a zeroth-order predictor where the average age is predicted for all samples.

2.5. ML Prediction Tasks and Analysis

In this study, we tested the performance of six DT techniques: (a) DT Regressor with cross-validation, (b) DT Regressor with cost-complexity pruning, (c) DT Regressor with Bagging Predictors, (d) DT Regressor with Random Forests, (e) DT Regressor with AdaBoost, (f) DT Regressor with Gradient Boosting. The six methods are desribed in detail in the Supplementary Information. All six techniques were applied to two problems: (1) classification of groundwater age categories (”modern” vs ”pre-modern” based on the tritium concentration), and (2) regression of tritium-helium groundwater ages (including pre-modern samples as 75 years old). The classification of pre-modern and modern samples serves two purposes: 1) examining the important features for classification provides insights on groundwater recharge patterns, 2) enabling a two-stage approach where a second-stage regressor only predicts the ages of samples classified as ”modern” by the first stage. For each task, we analyze the feature importance based on the Bagging Regressor approach, based on the performance and the ability to extract feature importance directly. We also analyze Partial Dependence Plots (PDPs) for the twelve most important features. PDPs visualize the marginal average of the predicted values in relation to the feature of interest. The underlying function as mentioned by [51] works by represented the effect of the feature of interest after accounting for the average effects (expectations) of all the other variables on the predictions.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

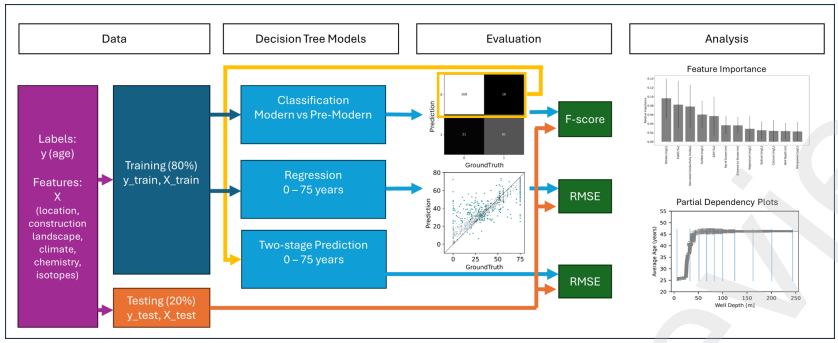


Figure 3: Schematic of the data, training, evaluation, and analysis process.

3. Results

3.1. Binary Water Age Classification

For the classification problem, ensemble learning methods to perform well in general with an average F1 score of 0.91 on the test set. The confusion matrix in Figure 4 represents model predictions on the test set for a Bagging classifier trained on all the sample points with a total of 2000 estimators. The final classification for a sample point is the majority vote taken for all the 2000 estimators' predictions. The Bagging classifier accurately (F1-score = 0.91) classified groundwater samples as either modern or pre-modern. This result is promising for applications where the risk to anthropogenic contaminants stems from their application since the 1950s.

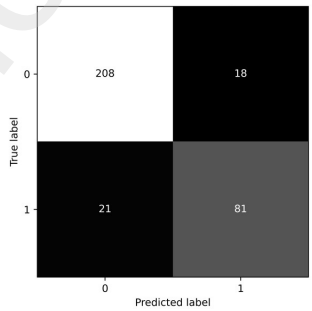


Figure 4: Confusion matrix for the binary classification model predictions on the validation dataset. Modern and pre-modern samples are labeled as 0 and 1, respectively. The binary Bagging classifier achieves a precision of 0.90, a recall of 0.92, and an F1 score of 0.91.

Feature Importance for Classification To examine which features the DT methods rely on, Figure 5 shows the Feature Importance plot for

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

the binary Bagging classifier. Feature Importance is computed as an average of 2000 estimators that were used to train the model. While decision trees allow for greater feature and model interpretability, we observe a trade-off between model accuracy and interpretability when we move from single tree based models to ensemble learning approaches. For this plot, we computed the average for all the trees with the understanding that this process is highly stochastic in nature especially as the number of estimators increase. This is also evident from the large error bars in the feature importance for all the features.

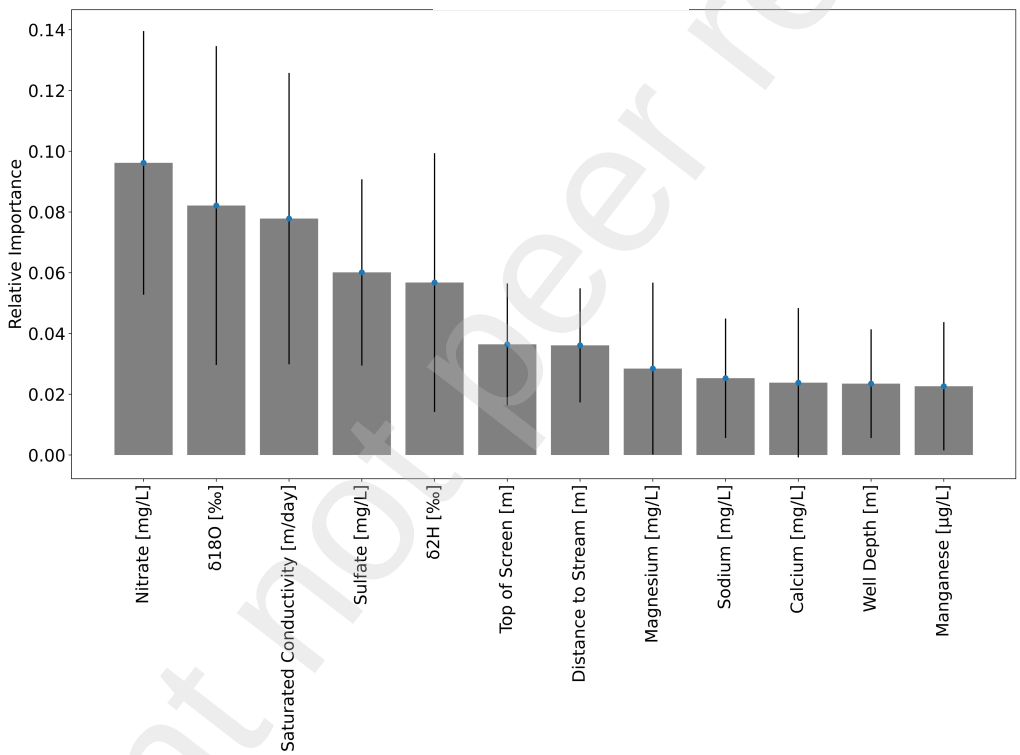


Figure 5: Feature importance plot for the binary Bagging classifier computed as an average of 2000 estimators used to train the model.

Figure 5 shows that for this binary classification, nitrate concentration was the most important feature. In our study, the nitrate leaching history in the Central Valley informs the age classification, not vice versa. In comparison, nitrate ranked third in the Gradient Boosting Regression (GBR) model of continuous groundwater ages in the Great Lakes Basin (GLB) [20]. The

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

relationship between nitrate and age predictability is discussed in detail in the context of the continuous tritium-helium age prediction below.

The high ranking of features associated with river water recharge (oxygen-18: 2nd; deuterium: 5th; distance to streams: 7th) suggests that river water is both an important source and indicator of modern recharge. Lower oxygen-18 and deuterium values, reflecting river water signatures, as well as shorter distances to streams are associated with a lower probability of pre-modern groundwater. This confirms a prior finding that the contribution of river water recharge nearly doubled from 24% in pre-modern groundwater to 46% in modern groundwater in the San Joaquin Valley [32]. Well construction features (top of screen: 6th; total depth: 11th) were less important for age classification than aquifer material (saturated conductivity: 3rd). In contrast, bottom of screen ranked 1st in the GLB-GBR prediction of groundwater ages [20].

Partial Dependency Plots (PDP) in Figure 6 show that, besides nitrate, higher sulfate, magnesium, and calcium concentrations are associated with modern water, indicating a commonality with modern land use practices. In contrast, higher sodium and manganese concentrations were associated with pre-modern water, possibly as a result of diffusion of sodium from marine depositional environments or geochemical reactions, specifically anoxic conditions releasing manganese, along longer flow paths. Fluoride, acidity, and iron concentrations were similar indicators of geochemical reactions in GLB groundwater. These associations are consistent with trends in pollutant concentrations and the effects of reactive transport at the groundwater body scale [8].

Most features show a gradual transition over the entire range of feature values. In contrast, nitrate, sulfate, total well depth, and distance to streams show a distinct transition from pre-modern to modern over a short range of feature values. For nitrate, this transition occurs between 0.6 and 1 mg/L, from the 30th to the 40th percentile of the data, indicating that concentrations below 1 mg/L are a strong indicator for pre-modern groundwater. For well depth, the transition occurs from 30 m to 50 m depth, between the 20th and 30th percentile. Shallower wells are more likely to produce modern water, deeper wells more likely to produce pre-modern water. The distance to the nearest stream only affects wells within 2 km of the stream, with little effect at larger distances. A similar threshold was found for the relationship between recharge source and nitrate in domestic wells in the San Joaquin Valley [49].

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

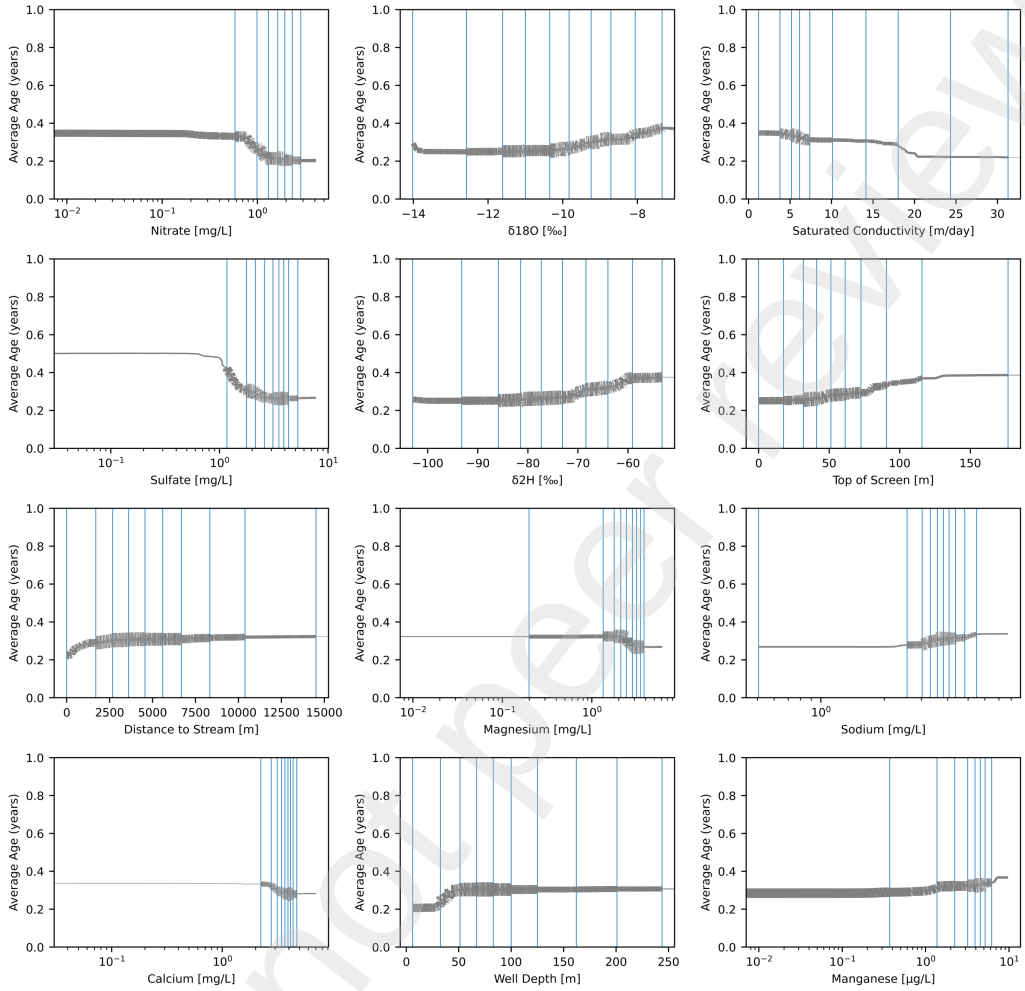


Figure 6: Partial dependence plots for the top 12 ranked features for the binary Bagging classifier (2000 estimators). Each subplot shows the effect of that feature (x-axis) on average age (y-axis). The blue lines on the plots are drawn according to the deciles of the input variable and the width of the grey line represents the relative number of training points for each decile. The relative number of training points for each deciles are computed after multiplying the number of data points per decile with the ratio of the decile to total range for a corresponding feature.

3.2. Water Age Regression

Single DT models: Simple DT models were implemented with early stopping technique with cross-validation. Hyper-parameter tuning resulted in a maximum tree depth of 7, and minimum samples per leaf of 10. A train

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

RMSE of 16.11 years whereas the RMSE on the test set was 22.43. Figure 7 (a) shows the train and test set performances. Compared with the baseline zeroth-order regressor with an RMSE of 24.85, the DT only achieves a 9.7% improvement. One popular regularization method to mitigate overfitting is pruning [52], which works by pruning back an over-grown DT till the validation performance improves. Pruning was tuned by further subdividing our train set by an 80/20 split to get a validation set which was then used for cost complexity pruning of the tree depth. Pruning resulted in a train RMSE of 18.48 years and a test RMSE of 21.74 years, a slight improvement of 3% over the vanilla DT above and an overall improvement of 12.5% over the zeroth-order baseline. Comparing (a) and (b) of Figure 7 shows that pruning further reduced the tree depth as well as the number of unique predictions.

Parallel ensemble models: To further improve performance, two parallel ensemble models, Bagging and RF are trained. Bagging with 3000 estimators achieved the best train and test RMSE of 6.36 and 17.58 years, respectively. It correspond to a 29.2% error reduction over the baseline zeroth-order regressor. Figure 7 (c) shows the train and test set performance. Recalling our label distribution from Figure 2, we observes that our model struggled with predicting samples clipped at 75 years. However, for samples in the mid-range distribution, ensemble learning resulted in an improved accuracy. As shown in Figure 7 (d), RF performed on par with Bagging: with $d/3$ features and 2000 estimators, RF achieves train and test RMSE of 6.37 and 17.42 years, respectively.

Sequential ensemble models: While the previous two ensemble learning techniques build independent estimators in parallel, AdaBoost and GBDT ensembles weak learners sequentially, the latter being used by [20]. Figure 7 (e) and (f) show the performance of AdaBoost and Gradient boosting regressors. Both models slightly under-performed the parallel ensemble models: AdaBoost achieved a train RMSE of 13.04 years and a test RMSE of 18.35 years, GB 8.54 years on train and 18.12 years on test data set.

Two-stage model: A two-stage approach that leverages the binary age classier was also implemented as an alternative. Samples classified as pre-modern are assigned a fixed value of 75, modern samples are then fed into the second stage regressor. While we experimented with different architectures in our arsenal of regressors, the best test RMSE of 18.22 years still under-performed Bagging and RF. It should be noted that this performance includes the error of the samples that were misclassified by the Bagging classifier.

Table 2 summarizes the Train and Test performance in terms of RMSE of

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

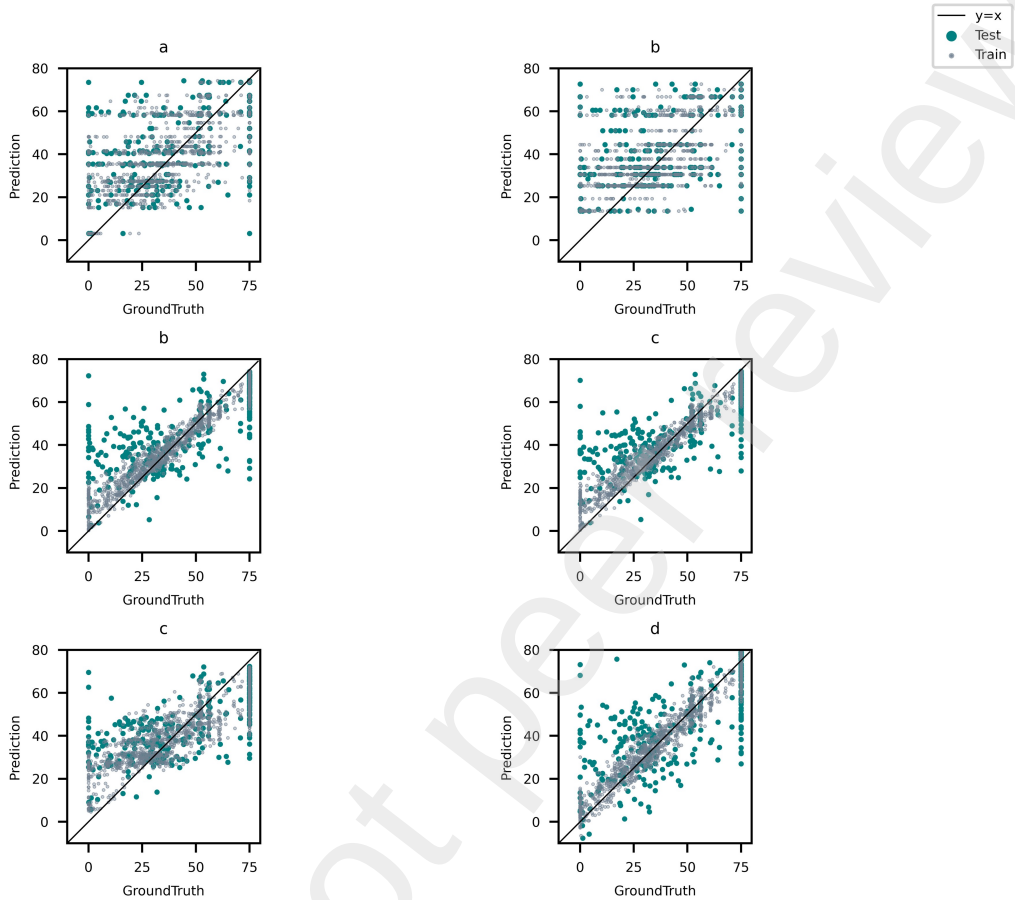


Figure 7: Train and test set performances 6 approaches: (a) DT Regressor with cross-validation, (b) DT Regressor with cost-complexity pruning, (c) DT Regressor with Bagging Predictors, (d) DT Regressor with Random Forests, (e) DT Regressor with AdaBoost, (f) DT Regressor with Gradient Boosting.

the six DT methods examined, including their improvement over the zeroth-order mean prediction. Overall, we observed that Random Forests resulted in the lowest root-mean-squared error on the held-out data set followed by the Bagging regressor. Moreover, we also observed the strength of the ensemble learning approaches in terms of observing higher model performances than single estimator based approaches.

Feature Importance for Regression

Figure 8 shows the Feature importance plot for the Bagging regressor

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

Table 2: Summary of performance between different ML models

Method	RMSE		
	Train	Test	Improvement
DT Regressor			
.. with cross-validation	16.11	22.43	9.7%
.. Regressor with cost-complexity pruning	18.48	21.74	12.5%
.. Regressor with Bagging Predictors	6.36	17.57	29.3%
.. Regressor with Random Forests	6.37	17.42	29.9%
.. Regressor with AdaBoost	13.04	18.35	26.2%
.. Regressor with Gradient Boosting	8.54	18.12	27.1%
Two-stage model		18.22	26.7%

computed as an average of 3000 estimators used to train the model. For modern water age regression, oxygen-18 is the most important feature, confirming the importance of river water recharge in modern groundwater. Total well depth ranks second (with a similar importance score as oxygen-18) and top of screen ranks third, indicating that vertical stratification of groundwater ages is still an important concept underlying patterns in groundwater age.

The Partial Dependency Plot (Figure 9) for the regression shows that total well depth is particularly important for wells with a depth up to 50 m (representing the shallowest 20% of all wells in this dataset), indicated by the largest transition (25 years, from 20 to 45 years) in all partial dependence plots. The depth of the top of the screen is particularly important between 75 and 125 meters. This suggests that the gradient of modern groundwater ages (0-75 yrs) is present in this depth range and that deeper parts of these aquifers are dominated by pre-modern groundwater.

Notably, nitrate is not among the 12 most important features for continuous tritium-helium groundwater age prediction. While nitrate was important to distinguish between pre-modern and modern groundwater, nitrate concentrations contain little information on the age of modern groundwater. We attribute this to spatial and/or temporal variability in nitrate concentrations in recharging groundwater and geochemical reactions (i.e. denitrification) along groundwater flow paths. The high ranking of nitrate in the GLB-GBR prediction suggests that trends in groundwater in relation to groundwater age may be more pronounced in the GLB than in modern groundwater in

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

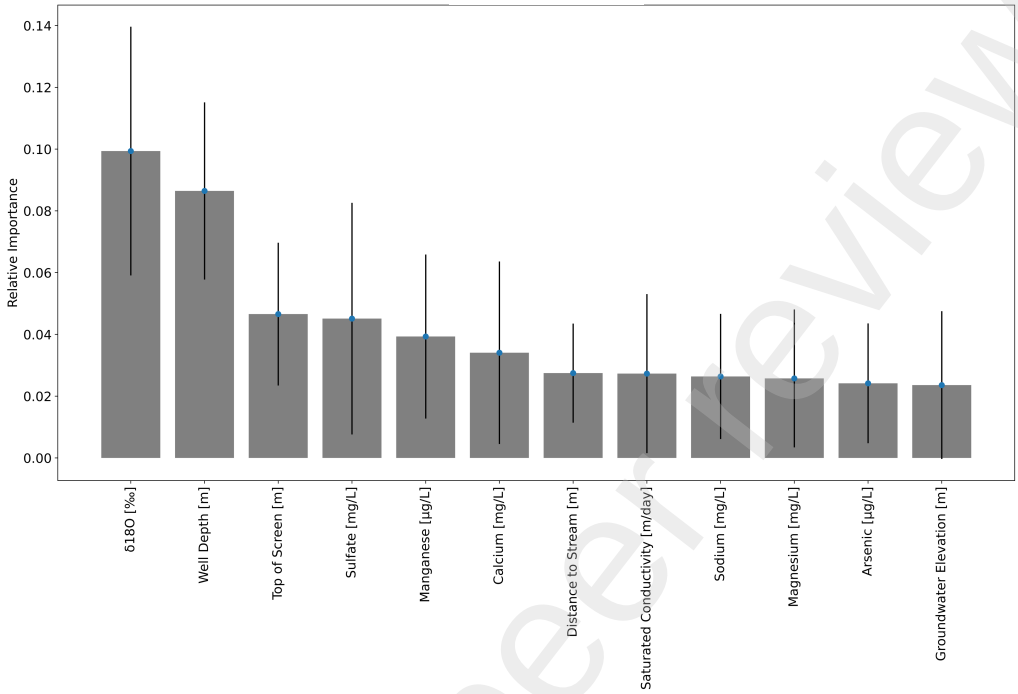


Figure 8: Feature importance plot for the Bagging regressor computed as an average of 3000 estimators used to train the model. While decision trees allow for greater feature and model interpretability, we observe a trade-off between model accuracy and interpretability when we move from single tree based models to ensemble learning approaches. For this plot, we computed the average for all the trees with the understanding that this process is highly stochastic in nature especially as the number of estimators increase. This is also evident from the deep error bars in the feature importance for all the features similar to the feature importance plot in Figure 8

the Central Valley. Additionally, because the GBR model of GLB groundwater ages was trained on continuous ages up to 150 years, prediction errors for pre-modern groundwater were more heavily penalized than in our continuous groundwater age prediction where pre-modern ages were set at 75 years.

Instead of nitrate, the concentrations of the same ions that were important to the binary age classification are also important to the continuous age prediction, although in a different order. High sulfate, magnesium, and calcium are associated with younger water, suggesting a common source in modern recharge. Similarly, high chloride concentrations in the GLB groundwater were associated with younger groundwater due to road salt application in recent history. As with the age classification, fossil marine water (sodium)

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

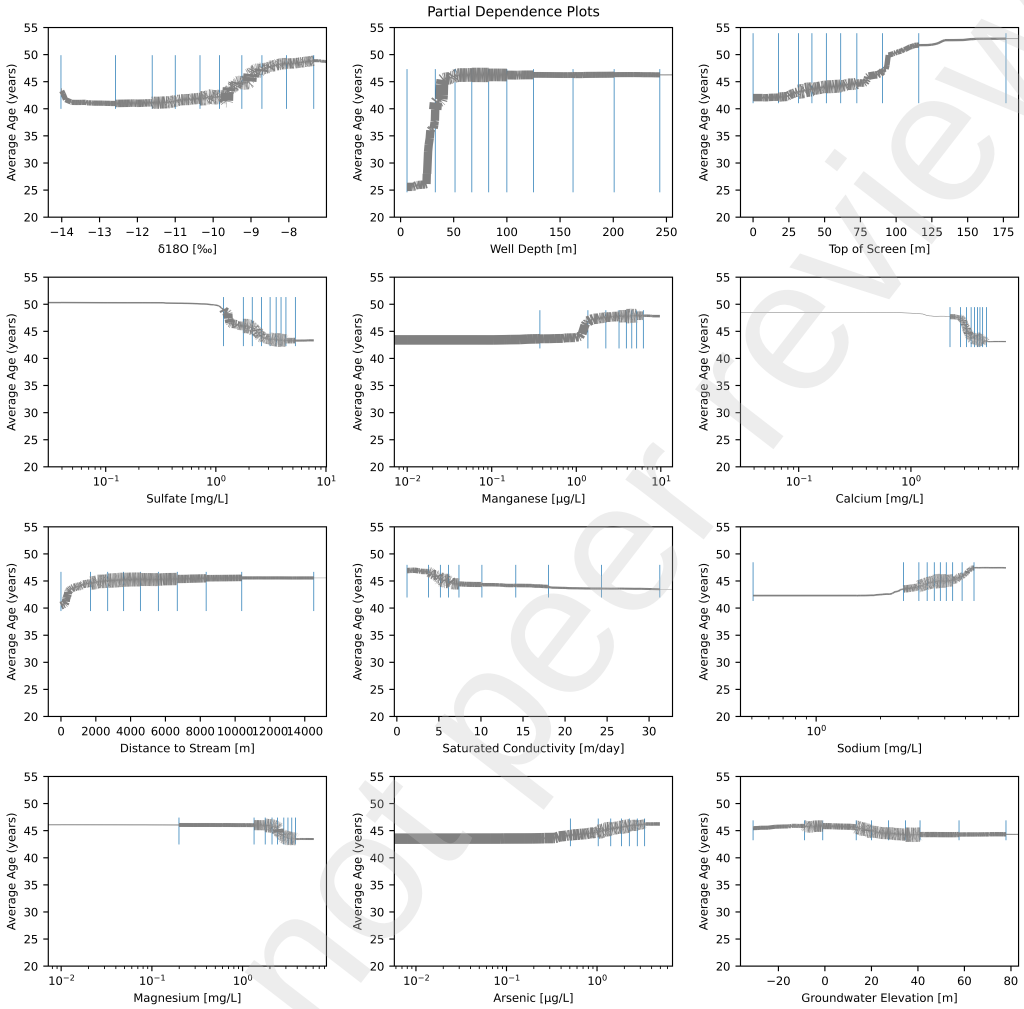


Figure 9: Partial Dependence Plots (PDPs) for the top twelve variables estimated by the feature importance plots for the bagging regressor approach. For each plot, the y-axis denotes the marginal change in the predicted outcomes from the training data as a feature of interest is varied.

and redox sensitive (manganese) indicators are associated with older water.

Comparison with Other Predictions of Groundwater Ages The performance of ML models in this study is not directly comparable with Green et al. (2021) [20] for two reasons: 1) the data distribution between two datasets are different. 2) Green et al. (2021) [20] predicts log-transformed

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

age while this study predicts ages directly. To address this, predictions from Green et al. (2021) were transformed back into linear years and clipped at 75 years. The test RMSE of the model is 20.6 years, a 32% RMSE reduction from 30.3 years of its corresponding zeroth-order baseline. By this metric, the performance of both ML predictions are remarkably similar. It should be noted that the GLB-GBR prediction would likely perform better on this metric if it had been trained to an age dataset that was limited to 75 years. The RMSE of the best ML models in the Central Valley groundwater basin is also similar to the short-distance spatial variability (14 years) observed in the same dataset including tritium-helium ages in all California groundwater basins. These residuals suggest that groundwater ages are difficult to predict for large datasets, collected from public supply, domestic, and monitoring wells. Patterns in groundwater ages are complex due to localized recharge from rivers or managed aquifer recharge, large scale variations in recharge due to agricultural irrigation, and temporal variation due to interannual climate variability. While clear groundwater age trends with depth have been observed in short-screened monitoring wells [53, 8], interaction between pumping wells on shorter distances can mutually influence the age distribution of pumped groundwater [54, 55, 56].

4. Conclusion

A machine learning method (Bagging Regressor) can accurately classify (F1-score = 0.81) groundwater samples as either modern or pre-modern in a large groundwater basin with complex land use, irrigation, and water management practices. The classification was based on 63 features, including location, well construction information, landscape characteristics, and climate variables, water chemistry, and stable isotopes. The accuracy of the ML prediction of continuous tritium-helium groundwater ages is limited and explains $\sim 30\%$ of the variability in this dataset. The RMSE of the hold-out test data in this study is 17.5 years. This value is similar to the short-distance spatial variability (14 years) derived from a geostatistical analysis of groundwater ages in California [50] and the RMSE (20 years) of a ML prediction of groundwater ages in the Great Lakes Basin [20]. An analysis of feature importance and partial dependency plots show the impact of river water recharge on the modern groundwater flow system. In general, ML groundwater age prediction relies on five different types of features: (1) the recharge source, (2) contaminant history, (3) aquifer materials, (4) well

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

construction, and (5) geochemical reactions along flow paths. While ML predictions are not sufficiently accurate to replace laboratory isotopic age tracer analyses, they can predict the distinction between pre-modern and modern groundwater. This can refine dynamic penetration depths of post-1950s water in unconfined aquifers [57], based on a larger number of wells where no tracer data is available, to better constrain recharge rates and groundwater budgets. The distinction between modern and pre-modern water can also support well owner decisions regarding well vulnerability to contamination and sustainable use.

5. Acknowledgements

We thank the California Waterboards Groundwater Ambient Monitoring and Assessment (GAMA) team for their assistance acquiring and curating the GAMA data. We thank Jake Harm for creating Figure 1. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and supported by Laboratory Directed Research and Development (LDRD) funding (23-ER-039). LLNL-JRNL-861411

References

[1] U. Nations, Groundwater overview. making the invisible visible (pp. 1–60), United Nations Water, International Groundwater Resources Assessment Centre (2018) 1–60.

[2] U. Lall, L. Josset, T. Russo, A snapshot of the world's groundwater challenges, Annual Review of Environment and Resources 45 (2020) 171–194.

[3] P. Cook, D. Solomon, Recent advances in dating young groundwater: chlorofluorocarbons, 3h3he and 85kr, Journal of hydrology 191 (1-4) (1997) 245–265.

[4] D. W. Pollock, Semianalytical computation of path lines for finite-difference models, Groundwater 26 (6) (1988) 743–750.

[5] W. Sanford, Calibration of models using groundwater age, Hydrogeology Journal 19 (1) (2011) 13–16.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

[6] P. Małoszewski, A. Zuber, Determining the turnover time of groundwater systems with the aid of environmental tracers: 1. models and their applicability, Journal of hydrology 57 (3-4) (1982) 207–231.

[7] A. Visser, H. P. Broers, M. F. Bierkens, Dating degassed groundwater with 3h/3he, Water Resources Research 43 (10) (2007).

[8] A. Visser, H. P. Broers, R. Heerdink, M. F. Bierkens, Trends in pollutant concentrations in relation to time of recharge and reactive transport at the groundwater body scale, Journal of Hydrology 369 (3-4) (2009) 427–439.

[9] A. Nir, On the interpretation of tritium 'age'measurements of groundwater, Journal of Geophysical Research 69 (12) (1964) 2589–2595.

[10] P. Schlosser, M. Stute, H. Dörr, C. Sonntag, K. O. Münnich, Tritium/3he dating of shallow groundwater, Earth and Planetary Science Letters 89 (3-4) (1988) 353–362.

[11] J. Vogel, Investigation of groundwater flow with radiocarbon., pp 355-69 of Isotopes in Hydrology. Vienna, International Atomic Energy Agency, 1967. (1968).

[12] G. M. Thompson, J. Hayes, Trichlorofluoromethane in groundwater—a possible tracer and indicator of groundwater age, Water Resources Research 15 (3) (1979) 546–554.

[13] E. Busenberg, L. N. Plummer, Use of chlorofluorocarbons (ccl3f and ccl2f2) as hydrologic tracers and age-dating tools: The alluvium and terrace system of central oklahoma, Water resources research 28 (9) (1992) 2257–2283.

[14] E. Busenberg, L. N. Plummer, Dating young groundwater with sulfur hexafluoride: Natural and anthropogenic sources of sulfur hexafluoride, Water Resources Research 36 (10) (2000) 3011–3030.

[15] H. H. Loosli, B. Lehmann, W. Balderer, Argon-39, argon-37 and krypton-85 isotopes in stripa groundwaters, Geochimica et cosmochimica acta 53 (8) (1989) 1825–1829.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

[16] E. A. Hussein, C. Thron, M. Ghaziasgar, A. Bagula, M. Vaccari, Groundwater prediction using machine-learning tools, Algorithms 13 (11) (2020) 300.

[17] R. Haggerty, J. Sun, H. Yu, Y. Li, Application of machine learning in groundwater quality modeling-a comprehensive review, Water Research 233 (2023) 119745.

[18] S. Singha, S. Pasupuleti, S. S. Singha, R. Singh, S. Kumar, Prediction of groundwater quality using efficient machine learning technique, Chemosphere 276 (2021) 130265.

[19] K. M. Ransom, B. T. Nolan, J. A. Traum, C. C. Faunt, A. M. Bell, J. A. M. Gronberg, D. C. Wheeler, C. Z. Rosecrans, B. Jurgens, G. E. Schwarz, et al., A hybrid machine learning model to predict and visualize nitrate concentration throughout the central valley aquifer, california, usa, Science of the Total Environment 601 (2017) 1160–1172.

[20] C. Green, K. M. Ransom, B. T. Nolan, L. Liao, T. Harter, Machine learning predictions of mean ages of shallow well samples in the great lakes basin, usa, Journal of Hydrology 603 (2021) 126908.

[21] L. Breiman, Bagging predictors, Machine learning 24 (1996) 123–140.

[22] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[23] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of computer and system sciences 55 (1) (1997) 119–139.

[24] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

[25] L. L. Lucas, M. P. Unterweger, Comprehensive review and critical evaluation of the half-life of tritium, Journal of research of the National Institute of Standards and Technology 105 (4) (2000) 541.

[26] R. L. Michel, Tritium deposition over the continental united states, 1953-1983, in: Atmospheric deposition, 1989.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

[27] R. Michel, Tritium in the hydrologic cycle, in: Isotopes in the Water Cycle: Past, Present and Future of a Developing Science, Springer, 2005, pp. 53–66.

[28] E. M. LaBolle, G. E. Fogg, J. B. Eweis, Diffusive fractionation of 3h and 3he in groundwater and its impact on groundwater age estimates, Water resources research 42 (7) (2006).

[29] O. S. University, Prism climate group, http://prism.oregonstate.edu (2020).

[30] D. M. Wolock, Hydrologic landscape regions of the united states, Tech. rep., US Geological Service (2003).

[31] C. D. of Water Resources, California's groundwater update 2020, Bulletin 118 (2020).

[32] A. Visser, J. E. Moran, M. J. Singleton, B. K. Esser, Importance of river water recharge to the san joaquin valley groundwater system, Hydrological Processes 32 (9) (2018) 1202–1213.

[33] A. Escriva-Bou, B. Gray, S. Green, T. Harter, R. Howitt, D. MacEwan, N. Seavy, Water stress and a changing san joaquin valley, Public Policy Institute of California. https://www. ppic. org/content/pubs/report/R_0317EHR. pdf (2017).

[34] C. C. Faunt, M. Sneed, J. Traum, J. T. Brandt, Water availability and land subsidence in the central valley, california, usa, Hydrogeology Journal 24 (3) (2016) 675.

[35] D. Perrone, S. Jasechko, Dry groundwater wells in the western united states, Environmental Research Letters 12 (10) (2017) 104002.

[36] R. Pauloo, A. Escriva-Bou, H. Dahlke, A. Fencl, H. Guillon, G. Fogg, Domestic well vulnerability to drought duration and unsustainable groundwater management in california's central valley, Environmental Research Letters 15 (4) (2020) 044010.

[37] Dry well reporting tool.
URL `https://mydrywatersupply.water.ca.gov/report/`

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

[38] S. Jasechko, D. Perrone, California's central valley groundwater wells run dry during recent drought, Earth's Future 8 (4) (2020) e2019EF001339.

[39] T. Harter, J. R. Lund, Addressing nitrate in California's drinking water: With a focus on Tulare Lake Basin and Salinas Valley groundwater: Report for the State Water Resources Control Board report to the Legislature, Center for Watershed Sciences, University of California, Davis, 2012.

[40] K. Lockhart, A. King, T. Harter, Identifying sources of groundwater nitrate contamination in a large alluvial groundwater basin with highly diversified intensive agricultural production, Journal of contaminant hydrology 151 (2013) 140–154.

[41] R. Smith, R. Knight, S. Fendorf, Overpumping leads to california groundwater arsenic threat, Nature communications 9 (1) (2018) 2089.

[42] E. A. Haugen, B. C. Jurgens, J. A. Arroyo-Lopez, G. L. Bennett, Groundwater development leads to decreasing arsenic concentrations in the san joaquin valley, california, Science of the Total Environment 771 (2021) 145223.

[43] B. C. Jurgens, M. S. Fram, K. Belitz, K. R. Burow, M. K. Landon, Effects of groundwater development on uranium: Central valley, california, usa, Groundwater 48 (6) (2010) 913–928.

[44] K. Belitz, Framework for a ground-water quality monitoring and assessment program for California, Vol. 3, US Department of the Interior, US Geological Survey, 2003.

[45] Gama groundwater information system.
URL `https://gamagroundwater.waterboards.ca.gov/gama/gamamap/public/`

[46] Gama publications.
URL `https://mydrywatersupply.water.ca.gov/report/`

[47] Gama database.
URL `https://gamagroundwater.waterboards.ca.gov/gama/datadownload`

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

[48] Usgs nwis database.
URL https://gamagroundwater.waterboards.ca.gov/gama/datadownload

[49] G. Castaldo, A. Visser, G. E. Fogg, T. Harter, Effect of groundwater age and recharge source on nitrate concentrations in domestic wells in the san joaquin valley, Environmental science & technology 55 (4) (2021) 2265–2275.

[50] A. Visser, J. Moran, D. Hillegonds, M. Singleton, J. T. Kulongoski, K. Belitz, B. Esser, Geostatistical analysis of tritium, groundwater age and other noble gas derived parameters in california, Water research 91 (2016) 314–330.

[51] W.-Y. Loh, Classification and regression trees, Wiley interdisciplinary reviews: data mining and knowledge discovery 1 (1) (2011) 14–23.

[52] D. P. Bovet, P. Crescenzi, D. Bovet, Introduction to the Theory of Complexity, Vol. 7, Prentice Hall London, 1994.

[53] H. P. Broers, The spatial distribution of groundwater age for different geohydrological situations in the netherlands: implications for groundwater quality monitoring at the regional scale, Journal of Hydrology 299 (1-2) (2004) 84–106.

[54] H. Peter Broers, F. C. Van Geer, Monitoring strategies at phreatic wellfields: A 3d travel time approach, Groundwater 43 (6) (2005) 850–862.

[55] A. Visser, H. P. Broers, R. Purtschert, J. Sültenfuß, M. de Jonge, Groundwater age distributions at a public drinking water supply well field derived from multiple age tracers (85kr, 3h/3he, and 39ar), Water Resources Research 49 (11) (2013) 7778–7796.

[56] A. Massoudieh, A. Visser, S. Sharifi, H. P. Broers, A bayesian modeling approach for estimation of a shape-free groundwater age distribution using multiple tracers, Applied Geochemistry 50 (2014) 252–264.

[57] K. Faulkner, B. Jurgens, S. Voss, D. Dupuy, Z. Levy, Modeling the dynamic penetration depth of post-1950s water in unconfined aquifers using environmental tracers: Central valley, california, Journal of Hydrology 616 (2023) 128818.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

[58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

Appendix A. Supplementary Information

Appendix A.1. Decision Tree Models

In this study, we tested the performance of six DT models that can be described in three different categories: Single DT Models, Parallel Ensemble Models, and Sequential Ensemble Models.

Single DT Models: We have performed regression on a single decision tree using two different approaches. Firstly, we decide to treat the tree depth as a hyper-parameter and perform an exhaustive hyper-parameter search (280 combinations) with 5-fold cross-validation to choose the optimal tree architecture. Hyperparameter tuning and subsequent model training were all done using scikit-learn package [58] in Python version 3.10.10. These choices of hyperparameters include the tree depth (number of levels or feature splits), minimum samples of splits required to create a split, and minimum samples in a leaf node (prediction node). The idea behind tuning the tree depth or level is to balance the bias-variance trade-off while evaluating the model performance on a held-out dataset. However, building an ideal decision tree is an NP-hard problem [52] given the heuristic nature of model training where hyperparameter tuning could be considered as stopping early. A better approach for training decision trees in general is to perform pruning which is characterized by growing the tree too large and then trimming it till validation performance of the model improves. This technique was first introduced by [52] which penalizes the size of the tree while minimizing the cost of the tree.

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|,$$

where $R(T)$ is the traditionally defined classification/regression loss that is minimized with a deeper tree and in addition there is regularizing term that penalizes the number of terminal nodes in T. The parameter α is computed using hyper-parameter tuning. While single decision trees using cross-validation and cost complexity pruning may not yield the highest predictive outcome in terms of predicting water ages, they offer the greatest insights on the predictive power of each feature based on computing the mean decrease in impurity at each level of the tree.

Parallel Ensemble Models: We tested two parallel ensemble models that rely on aggregating predictions from several 'weak learners' with low biases to predict an output that has a low variance. First, we tested

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017

a technique called bagging [21] that generates multiple versions of decision trees by picking random sub-samples from the same training dataset with replacement. For the case of a regression, the predicted outcome is the average of all the individual predictors. A total of 10 different hyperparameter combinations were tested where the number of learners were varied and cross-validated using 5-fold cross validation. While aggregating several predictors often results in improved model performance, it often comes at the cost of reduced feature interpretability given the randomness in selecting a subset of the training data and using the average of several weak learners to predict an outcome. Second, we tested Random Forests [22], which follow a similar approach as bagging however, in addition to extracting a randomly drawn subset of the training data with replacement, it also picks a random subset of the features within that training sample. For this algorithm, we performed hyperparameter tuning on a total of 70 combinations pertaining to number of estimators and number of features per estimator. While varying the number of features, we observed peak performance on a dataset with one third of the features extracted. While random forests are expected to generate a stronger predictive power than bagging classifiers, it often comes at the cost of reduced model interpretability since the number of features for each estimator are different hence, we lose confidence in the average importance of each feature per estimator.

Sequential Ensemble Models: While bagging and random forests build independent estimators based on random subsampling of training data and number of features selected, sequential learning algorithms where an estimator is grown using information from the previous estimator. We tested both the AdaBoost [24] and Gradient Boosting Regressors [24]. Adaboost pertains to allocating more weights to those samples that have a higher root-mean-squared-error from the previous learner whereas gradient boosting also builds learners in a stage-wise sequential fashion but generalizes better by allowing optimization of an arbitrary differentiable loss function.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4754017