Science of the Total Environment 601–602 (2017) 1160–1172

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA

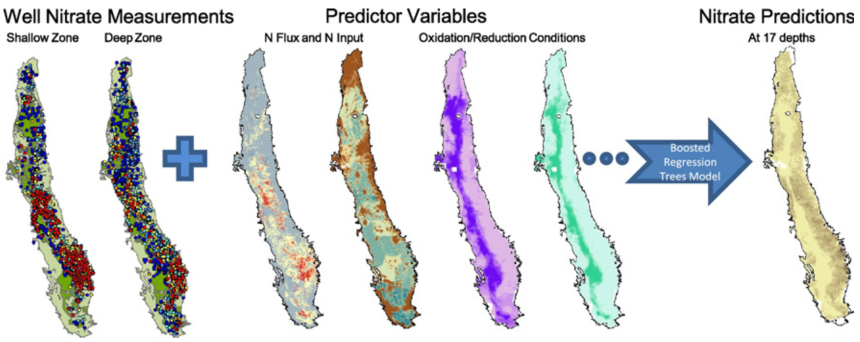
Katherine M. Ransom [a,*], Bernard T. Nolan [b], Jonathan A. Traum [c], Claudia C. Faunt [d], Andrew M. Bell [e], Jo Ann M. Gronberg [f], David C. Wheeler [g], Celia Z. Rosecrans [c], Bryant Jurgens [c], Gregory E. Schwarz [b], Kenneth Belitz [h], Sandra M. Eberts [i], George Kourakos [a], Thomas Harter [a]

[a] University of California, Davis, Department of Land, Air, and Water Resources, United States
[b] U.S. Geological Survey National Water Quality Program, Reston, VA, United States
[c] U.S. Geological Survey California Water Science Center, Sacramento, CA, United States
[d] U.S. Geological Survey California Water Science Center, San Diego, CA, United States
[e] University of California, Davis, Center for Watershed Sciences, United States
[f] U.S. Geological Survey California Water Science Center, Menlo Park, CA, United States
[g] Virginia Commonwealth University, Department of Biostatistics, Richmond, VA, United States
[h] U.S. Geological Survey New England Water Science Center, Northborough, MA, United States
[i] U.S. Geological Survey Ohio Water Science Center, Columbus, OH, United States

HIGHLIGHTS

- Boosted regression tree model produced 3D map of nitrate concentration.
- Hybrid multi-modeling approach used numerical model outputs as predictors.
- Redox characteristics and field scale unsaturated zone N flux were most important.
- Nitrate concentrations <2 mg/L NO_3-N generally conformed to basin subregion.
- Nitrate concentrations >10 mg/L NO_3-N most common in eastern alluvial fans subregion

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 6 March 2017
Received in revised form 19 May 2017
Accepted 20 May 2017
Available online 9 June 2017

Editor: D. Barcelo

Keywords:
Groundwater
Nitrate
Boosted regression trees

ABSTRACT

Intense demand for water in the Central Valley of California and related increases in groundwater nitrate concentration threaten the sustainability of the groundwater resource. To assess contamination risk in the region, we developed a hybrid, non-linear, machine learning model within a statistical learning framework to predict nitrate contamination of groundwater to depths of approximately 500 m below ground surface. A database of 145 predictor variables representing well characteristics, historical and current field and landscape-scale nitrogen mass balances, historical and current land use, oxidation/reduction conditions, groundwater flow, climate, soil characteristics, depth to groundwater, and groundwater age were assigned to over 6000 private supply and public supply wells measured previously for nitrate and located throughout the study area. The boosted regression tree (BRT) method was used to screen and rank variables to predict nitrate concentration at the depths of domestic and public well supplies. The novel approach included as predictor variables outputs from existing physically based models of the Central Valley. The top five most important predictor variables included two oxidation/

[*] Corresponding author.
E-mail address: kmlockhart@ucdavis.edu (K.M. Ransom).

http://dx.doi.org/10.1016/j.scitotenv.2017.05.192

Machine learning
Modeling

reduction variables (probability of manganese concentration to exceed 50 ppb and probability of dissolved oxygen concentration to be below 0.5 ppm), field-scale adjusted unsaturated zone nitrogen input for the 1975 time period, average difference between precipitation and evapotranspiration during the years 1971–2000, and 1992 total landscape nitrogen input. Twenty-five variables were selected for the final model for log-transformed nitrate. In general, increasing probability of anoxic conditions and increasing precipitation relative to potential evapotranspiration had a corresponding decrease in nitrate concentration predictions. Conversely, increasing 1975 unsaturated zone nitrogen leaching flux and 1992 total landscape nitrogen input had an increasing relative impact on nitrate predictions. Three-dimensional visualization indicates that nitrate predictions depend on the probability of anoxic conditions and other factors, and that nitrate predictions generally decreased with increasing groundwater age.

1. Introduction

Nitrate contamination of groundwater is a problem that many agricultural regions around the world share. Nitrate is a naturally occurring form of nitrogen necessary for plant growth, however, decades of farming in agricultural regions underlain by alluvial aquifers has resulted in leaching of excess nitrate to groundwater from nitrogen sources such as animal manure and synthetic fertilizers. Septic tanks in rural, unincorporated areas can also contribute nitrate and atmospheric deposition (from the combustion of fossil fuels) or natural organic matter can contribute relatively minor amounts (Canter, 1996). Nitrate in drinking water above the U.S. Environmental Protection Agency maximum contaminant level (MCL) of 10 mg/L as N has been linked to low infant blood oxygen levels, a condition known as methemoglobinemia. More recently, health effects including cancers and adverse reproductive outcomes have been associated with drinking water nitrate concentrations less than the MCL (Ward et al., 2005). Many factors influence the amount of nitrate that reaches groundwater including current and historical land use, current and historical nitrogen applications/deposition, soil type, depth to groundwater and groundwater recharge rate. Any of these variables can interact in complicated ways to affect the amount of nitrogen that leaches to groundwater.

In the Central Valley of California, nitrate contamination of drinking water wells is of increasing concern due to decadal increases in nitrogen fertilizer, manure, and population, and observed groundwater nitrate trends. In a study of the Central Valley aquifer, Burow et al. (2013) observed increasing trends in the proportion of the aquifer with nitrate concentration >5 mg/L in both shallow and deep zones of the eastern fans subregion from the 1950s through the 2000s (Burow et al., 2013). They separated well nitrate concentration data into shallow and deep zones based on the depths of domestic and public-supply wells, and the increases were approximately four-fold in the shallow aquifer and two-fold in the deep aquifer. However, variability of nitrogen sources hampered interpretation of aquifer proportion trends in the western fans subregion.

Two previous studies of private wells in the Central Valley found between 40 and 50% of wells sampled for nitrate exceeded the drinking water MCL (Lockhart et al., 2013; CSWRCB, 2010). Both of the aforementioned studies were focused in the Central Valley and relatively small (approximately 200 wells each). However, many private wells are in use within the Central Valley; one study has estimated that approximately 92,000 private wells exist in the Central Valley (Johnson and Belitz, 2015); and given that private wells are not regulated in California, relatively few of these wells have been tested. Public supply wells are required to test for nitrate, and over 400 public supply wells throughout California have had nitrate exceeding the MCL, which affects over 600,000 people according to an equivalent-population based metric (Belitz et al., 2015). The highest numbers of high-nitrate wells and equivalent populations relying on such wells occurred in the San Joaquin Valley (the southern two-thirds of the Central Valley) and the Transverse and Selected Peninsular Ranges (the latter province is south of the Central Valley and includes Los Angeles). Among the nine hydrogeologic provinces in California, the San Joaquin Valley had the largest area affected by high groundwater nitrate (Belitz et al., 2015).

Because private wells are not regulated in California, it is important to determine high risk and high priority areas in order to focus monitoring resources as well as future outreach and educational efforts aimed at private well owners. Public supply wells also are at risk, and the cost of treating contaminated groundwater or relocating such wells is high, especially for low-income or disadvantaged communities (Honeycutt et al., 2012). We used boosted regression trees (BRT), a machine learning method, to model groundwater nitrate concentration in response to predictor variables representing land use, climate, soil, and hydrogeologic factors. The algorithm learns the relationship between the response and the predictor variables and does not rely on hypothesis testing assumptions about the data as do more traditional statistical methods (Elith et al., 2008). BRT results then were extended throughout the Central Valley and to all depths of drinking water supplies to provide a three-dimensional (3D) map of nitrate in groundwater at the depths typically tapped by domestic and public-supply wells.

BRT makes use of two algorithms: regression or classification trees and boosting (Elith et al., 2008). Classification and regression tree (CART) methods subdivide the data with multiple nested binary splits, a procedure known as binary recursive partitioning (Hastie et al., 2009; Kuhn and Johnson, 2013). The procedure determines which variable explains most of the variance in the dependent variable and then determines a split point for that variable that most reduces the overall sum of squared errors. These steps are repeated with each branch until some stopping criterion takes effect. The output of the CART algorithm is a decision tree structure (Hastie et al., 2009). Boosting combines a set of predictions from a “weak classifier”, a simple model such as a single split classification tree, in order to form a stronger classifier and improve prediction accuracy (De'ath, 2007; Hastie et al., 2009).

In the case of BRT, an ensemble of smaller decision trees (weak learners) is built and with each successive tree trained on the residuals from the previous tree (De'ath, 2007). The fitted value for each observation is recalculated with the formation of each new tree and the final model is a stagewise sum of all of the trees (Elith et al., 2008). Stochastic BRT improves performance by adding randomness into the model fit by using subsamples of the training data at each iteration (Friedman, 2002). BRT is able to account for many variables and their interactions, allows non-linear and non-monotonic responses, and is robust in the presence of multicollinearity. Interactions among predictor variables are expected for nitrate because multiple conditions can cause elevated nitrate in ground water, including 1) a significant source of nitrate, 2) soil (and aquifer) properties that permit transport, 3) sufficient recharge, and 4) existence of oxic conditions; if one or more of these conditions are not met, a different outcome will occur (Anning et al., 2012).

Machine learning methods have been used to predict nitrate concentrations in groundwater in several studies: one for the U.S. Geological Survey (USGS) Southwest Principal Aquifers, which includes the Central Valley, California (Anning et al., 2012), a National Cancer Institute and USGS model in Iowa (Wheeler et al., 2015), and two Central Valley, California specific models (Nolan et al., 2014, 2015). For the

Southwest Principal Aquifers model, 58 explanatory variables, including county-level nitrogen loading, septic/sewer ratio, geology, soil, aquifer, and geochemical variables were used to develop a Random Forest classification model. Random Forest classification, like BRT, features many trees but the response variable is categorical and model fitting is not stagewise. The authors calculated total landscape nitrogen input from farm and non-farm fertilizer using county-level nitrogen data (for the 1982–2001 time period) and apportioned it to 3 km grid cells based on the amount of agricultural, urban, or residential land use within the cell. Manure nitrogen inputs were similarly estimated based on county-level livestock population estimates from the Census of Agriculture and apportioned to agricultural land uses likely to receive manure (such as pasture and hay crops) (Anning et al., 2012; Ruddy et al., 2006). When analyzed with a validation data set, the model correctly predicted 48.6% of nitrate categories into the correct class and 80.4% of nitrate categories into one class above or below (Anning et al., 2012). Wheeler et al. (2015) developed a Random Forest regression model from 34,084 nitrate measurements from private wells in Iowa sampled between 1980 through 2011. The authors compiled nearly 300 predictor variables including agricultural land use, county-level nitrogen fertilizer input, soil type, climatic variables, and aquifer characteristics. Fertilizer nitrogen input variables were compiled from county level fertilizer sales and assigned to each well based on the surrounding amount of agricultural land use (for seven separate years between 1978 and 2006). Manure nitrogen inputs were derived from county-level Census of Agriculture animal population data for multiple years between 1982 and 2002 and apportioned to relevant land uses (Mueller and Gronberg, 2013; Ruddy et al., 2006). The final model from Wheeler et al. (2015) explained 76.86% of log nitrate variation in the training data, and mean square error (MSE) was 0.97. All MSE units reported in the current paper are $(\ln(mg/L))^2$. For hold-out data, the model explained 38.27% of the variation in log nitrate and the MSE was 2.39.

Nolan et al. (2014) developed Random Forest regression and classification models to predict log nitrate for domestic and public supply wells located in the Central Valley, California. The authors compiled 54 explanatory variables including total farm and non-farm nitrogen inputs (compiled from 1992 county-level farm and non-farm fertilizer inputs and apportioned to agricultural, residential, and urban land uses), soil characteristics, well depth, percent coarse soil texture above the well screen, vertical water flux for the year 1992, and 1990 land use data. Their final Random Forest regression models selected for mapping shallow and deep nitrate had training R^2 of 0.90 (both models), training MSE of 0.22 and 0.14, respectively, out-of-bag R^2 values of 0.39 and 0.40, respectively, and out-of-bag MSE values of 1.28 and 0.83, respectively. Nolan et al. (2015) compared three machine learning techniques: BRT, artificial neural networks, and a Bayesian network (latter two methods not discussed in this paper) using the same shallow well data and predictor variables as Nolan et al. (2014). Using a statistical learning framework, a comparatively simple BRT model was selected as final for mapping based on hold-out data model evaluation results (Nolan et al., 2015). This model yielded an R^2 of 0.89 and an MSE of 0.26 for training data, an R^2 of 0.26 and MSE of 1.75 for hold-out data, and a cross validation testing R^2 of 0.36.

The objectives of the current study were to (1) characterize groundwater nitrate concentration at the depths of drinking water supply in the Central Valley, with emphasis on domestic and public supplies; and (2) extend modeling results valley-wide and top-to-bottom throughout the aquifer using 3D interpolation and visualization techniques. In this study, we used a hybrid multimodeling approach wherein outputs of previous models, including numerical and/or physically based types, were used as inputs to a BRT model. Physically based models included the Central Valley Hydrologic MODFLOW/MODPATH Model (CVHM) and the Central Valley Textural Model (CVTM) (Faunt, 2009), and the Groundwater Nitrogen Loading Model (GNLM) developed for the Central Valley (Viers et al., 2012; Rosenstock et al., 2013). Additionally, interpolated values from lumped parameter groundwater age models (Jurgens et al., 2016, 2012) and BRT groundwater reduction-oxidation (redox) models (Rosecrans et al., 2017), were used as predictor variables. CVHM, CVTM, and the BRT redox models describe conditions in 3D for the Central Valley. The use of field-scale, unsaturated zone nitrogen leaching flux differs from previous nitrate vulnerability models that used county-level nitrogen data apportioned to agricultural land use within well buffers as mentioned above (Nolan and Hitt, 2006; Nolan et al., 2014, 2015). The use of groundwater age and redox variables also differs from previous aquifer nitrate vulnerability models that used proxies for groundwater age and redox potential such as well depth or depth to water (typically available only at sampled wells) (Rupert, 1998; Warner and Arnold, 2010; Nolan and Hitt, 2006; Nolan et al., 2014, 2015). We believe that this study represents the first use of 3D estimates of redox and age in a regional aquifer vulnerability model. Ordinary kriging (OK), universal kriging (UK), and multiple linear regression (MLR) models were also developed for comparison with BRT.

2. Materials and methods

2.1. Sources of nitrate data

We compiled a large database of groundwater nitrate measurements from private supply and public supply wells (Ransom et al., 2017b). Nitrate data came from two main sources, the University of California at Davis (UC Davis) and those previously compiled by the USGS (Burow et al., 2013). Combining the UC Davis and USGS data substantially increased the number of wells available for modeling, compared with Nolan et al. (2014). Domestic well measurements from UC Davis (Ransom et al., 2017a) sampled during 2000–2011 were used. The UC Davis data set included wells from the Groundwater Ambient Monitoring and Assessment domestic well program (GAMA) and the Central Valley Regional Water Quality Control Board dairy monitoring program. The methods described in the paper by Ransom et al. (2017a) for the years 2000–2011 resulted in 2407 domestic wells for use in the BRT model. These methods included randomly selecting a single well from multiple wells with overlapping locations (Ransom et al., 2017a). Well nitrate measurements with a zero value were replaced with the most common detection limit of 0.33 mg/L nitrate as nitrogen (NO_3-N) prior to imputation. Additional nitrate data for the years 2000–2010 came from another 4788 wells comprising the following sources: the California State Water Resources Control Board Division of Drinking Water (SWRCB-DDW) (4307 wells); the USGS National Water Information System (NWIS) (471 wells); and the U.S. Environmental Protection Agency Storage and Retrieval database (STORET) (10 wells) (Burow et al., 2013). The SWRCB-DDW and STORET wells were predominantly used for public supply, and most of the NWIS wells were used for domestic or monitoring purposes. Duplicate wells in the combined data set were removed and a multiple imputation routine (Lubin et al., 2004) was used to impute values of censored nitrate data below the method detection limit (DL). For censored values, the lower bound was set to missing and the upper bound was set to the DL, which varied from 0.05 to 8.9 mg/L depending on the source of the data. Following imputation, median nitrate concentration was computed for each well and was natural log transformed prior to BRT modeling. Wells then were spatially declustered using an equal area grid cell approach (Belitz et al., 2010) to reduce effects on the modeling of oversampling in areas of intensive agricultural land use. We randomly sampled domestic and public supply wells in 100 km^2 grid cells at up to the same rate (i.e., the maximum number of wells taken per grid cell) to maintain the proportions of each well type in the training data set. Wells not selected for training were randomly sampled in the same manner to create an independent (hold-out) data set for model testing. A total of 5170 wells was selected, 3508 of which were used for training and 1662 of which served as hold-out. All nitrate measurements in the final database were converted to NO_3-N. In contrast, a total of 2505 wells were used by Nolan et al. (2014), 1255 of which were selected for training and 1250 of which served as hold-out for model evaluation.

2.2. Predictor variables

A database of 145 predictor variables was compiled from a combination of well construction data, Geographic Information System (GIS) attributes and outputs of previous models, including well characteristics (physical and location based), land use, climate, soil properties, aquifer properties, depth to the water table, and estimates of nitrogen loading based on field-scale and county-scale data (Supporting Materials (SM) Table S1). Well characteristics were either provided in each of the well databases (as was the case for use of water at the well), assigned (e.g. latitude, longitude, well depth zone), or calculated (e.g. distance to nearest major river). Wells were assigned to shallow (private well) or deep (public-supply) zones based on measured depth below ground surface to the bottom of the well screen. The 75th percentile well depth was 81.99 m for domestic wells and the 25th percentile depth was 84.73 m for public supply wells; therefore rounding to an intermediate value (270 ft), a depth of 82.30 m was used to divide the shallow and deep aquifer zones. Wells lacking measured well depth were assigned to shallow or deep zones based on water use, with domestic and monitoring wells designated shallow and public supply and irrigation wells designated deep. Depth to top and bottom of well screened interval was estimated by Empirical Bayesian Kriging (EBK) for these wells based on information from wells with construction data so that well construction characteristics were available for each well and could be tested as predictor variables in the BRT model. The predicted surfaces produced by this approach (Voss and Jurgens, 2017) are spatial averages of depths to top and bottom of well screens across the Central Valley. Empirical semivariances were developed from ensembles of 100 semivariograms generated by the restricted maximum likelihood method for subsets of the data (ArcGIS, 2016). The power semivariogram model was used to fit the empirical semivariances. EBK-predicted screen depths were extracted to all wells based on their assigned depth zone (shallow or deep). Historical and current land use data are from the California Augmented Multi-Source Land use (CAML) 50 m resolution maps based on five time periods at 15-year intervals over the past 70 years. Land use surrounding wells was classified for each of the five CAML maps each representing a 5-year time period centered on: 1945, 1960, 1975, 1990, and 2005 (Viers et al., 2012). The CAML land use categories were reclassified into 13 groups based on similarity in land use or crop type. Temperature and rainfall data are from the WorldClim BioClim 1 km resolution layers generated through interpolation of average monthly climate data from weather stations, representative of 1950–2000 (Hijmans et al., 2005). The BioClim data set contains 19 variables including annual precipitation and precipitation of the wettest and driest month. Soil characteristics, hydrologic group, and drainage class predictor variables are from the Soil Survey Geographic database (SSURGO) (U.S. Department of Agriculture, 2014; Wieczorek, 2014). Estimates of nitrogen loading from farm and non-farm fertilizer were based on county-wide fertilizer expenditures (Gronberg and Spahr, 2012), and groundwater recharge was derived from base-flow index and mean annual runoff values (Wolock, 2003).

In addition to the above attributes, model outputs that included estimations of aquifer physical characteristics were used as predictor variables, as part of the hybrid modeling approach. Model-estimated physical aquifer properties include, from CVHM at 1 mile resolution, monthly vertical groundwater flux in the upper active CVHM layer (for water year 1999–2000), and depth below ground surface to the water table, and from CVTM, average estimated percent coarse material in the upper active layer (Faunt, 2009). Additionally, average groundwater age at the well screen for all wells was calculated based on CVHM/MODPATH simulations performed for this study (SM S2.0). Three particles were used for each well (at the top, middle, and bottom of well screen) and the average of the three travel times was assigned to each well (SM S2.0). GNLM was used at 50 m scale to estimate historical and current field-scale unsaturated zone nitrogen leaching fluxes for five time periods at 15-year intervals (Viers et al., 2012; Rosenstock et al., 2013). From these, we calculated a normalized field-scale nitrogen mass balance for each of the time periods where GNLM data were available (SM Eqs. (1) and (2)).

We also used model estimates from previous statistical and geostatistical models as predictor variables. These include BRT-predicted probability of anoxic groundwater and probability of high groundwater manganese concentration ("redox" variables indicating reducing conditions) (Rosecrans et al., 2017), and kriged depth below ground surface to 60 year old groundwater at the well, based on lumped parameter modeling (Jurgens et al., 2012, 2016) (SM Table S1). We used the redox variable predictions from Rosecrans et al. (2017) for the 45.72 m or 91.44 m depths below ground surface (SM Table S1) for both redox variables. Wells, designated as shallow or deep, were assigned the values for the Rosecrans et al. (2017) prediction depth closest to the midpoints of measured top and bottom of well screen for domestic and pubic supply wells in this current study, which were 48.77 and 89.92 m, respectively.

All predictor variables were processed in ArcGIS (version 10.3.1), SAS (version 9.4), or in the R computing environment (R Core Team, 2016) for assignment to wells by either point extract or by statistics within an approximate well source area (SM Table S1). A 500 m radius circular well buffer was used for variables that required an approximate well source area for attribution. Well buffers are reasonable surrogates for contributing areas to wells in the Central Valley when the actual contributing area is unknown, and spatial analysis of land use within a range of buffer sizes has indicated that a 500 m radius buffer size is appropriate for attributing land use data to wells in the region (Johnson and Belitz, 2009). Missing predictor variable values at wells were estimated using the R package caret (Kuhn, 2016) by the "bag impute" method which fits a bagged tree model for each predictor as a function of all the others. The gbm package (Ridgeway et al., 2015) was used to perform BRT modeling.

Predictor variables selected for the final model (Section 3) were assigned to 1 km grids as Arc GIS raster layers. Based on the gridded predictor variables and final model, nitrate predictions were made using the R raster package (Hijmans, 2016) for 17 depth zones spaced throughout the aquifer (at 15.24, 30.48, 45.72, 60.96, 76.20, 91.44, 106.68, 121.92, 152.40, 182.88, 213.36, 243.84, 274.32, 304.80, 365.76, 426.72, and 487.68 m below ground surface) to create input layers for 3D mapping in Oasis Montaj. Redox variables, average groundwater age at the well screen, use of water at the well, screen length, and depth to bottom of screened interval varied with depth zone. Other predictors, such as soil properties and total N input at the land surface, selected for the final model did not vary with depth. Based on measured well depths of private and public supply wells before data declustering, screen length was set to 12.19 m for depth increments <82.30 m (shallow) and to 64.01 m for depths >82.30 m (deep). A hypothetical depth to bottom of screened interval was assumed by centering the average screen length for each respective depth zone (either shallow or deep) on the vertical centroid location of each prediction grid depth (SM Table S2). A groundwater age estimate was made for each of the 17 depth zones based on the hypothetical wells centered on 1 km grid cells throughout the Central Valley (SM S2.0). To produce the MODFLOW/MODPATH age estimates used to create the raster layers used as a part of the nitrate predictions, a particle was placed at the top, middle, and bottom of well screen for each hypothetical well (Table S2). Water use was set to H (private) and P (public supply), respectively, for BRT predictions at shallow and deep depths. Two raster layers were used to represent the redox variables: either a shallow or deep zone representation for each variable (SM Table S1).

2.3. Machine learning

Prior research used regularization to improve the predictive performance of BRT through adjustment of selected parameters such as the number of trees in the ensemble (Elith et al., 2008). Along these lines, we used BRT within a statistical learning framework (Nolan et al., 2015; Hastie et al., 2009) that systematically adjusts tuning parameters (Kuhn and Johnson, 2013) referred to here as "metaparameters," to control for potential overfit by the BRT models and to maximize prediction accuracy. Adjusted metaparameters included the number of trees, interaction depth, and shrinkage. The minimum number of observations in each tree's terminal nodes was held constant at 10. Ten-fold cross validation tuning was performed on the training data set with the full list of 145 predictor variables for each combination of a range of each of the metaparameter values (Table 1) using the R package caret (Kuhn, 2016). The selected ranges for each metaparameter yielded 900 possible combinations and the "best" combination of metaparameters was selected based on minimum cross validation (CV) testing root mean square error (RMSE), which represents the expected testing error. One standard-error models (Nolan et al., 2015) were tested but performed less well with hold-out data compared with the minimum CV testing RMSE model. Cross validation tuning run time was approximately 16 h on an Intel Core I-5 4670 processor with 16 GB of 1600 MHz DDR3 RAM.

The full BRT model consisting of all 145 predictor variables was then re-fitted with the best CV-tuned metaparameters to all 3508 training observations. Predictor variables were sorted in descending order by relative importance score, which is an estimate of predictor variable influence in the model (Friedman, 2001; Elith et al., 2008). Predictors were removed from the minimum RMSE model one-at-a-time, in ascending order of predictor variable influence beginning with the least influential variable, until the percent difference in RMSE for hold-out data, compared to the full model with all 145 predictors, consistently exceeded about 1%. We further refined the model to remove redundant variables (such as atmospheric nitrogen input for 1992, which is included in the total landscape nitrogen input for 1992 variable) and to include others desirable for interpretation purposes, while maintaining fit to hold-out data. We performed Sobol' sensitivity analysis on the final model using the soboljansen function in R package sensitivity (Pujol et al., 2017) to show the degree of variability of BRT model output for changes in the values of predictor variables. Sobol' sensitivity analysis apportions the variance of the model predictions to each of the prediction variables (Saltelli et al., 2010).

We obtained BRT prediction intervals by bootstrapping for the purpose of mapping prediction uncertainty at the depths of domestic and public groundwater supplies. Bootstrapping involved sampling the training data with replacement and fitting the final model to each of 199 bootstrap samples. Whereas the final metaparameters (interaction.depth, n.trees, shrinkage) were held constant, the true model parameters (tree splitting variables, split levels at internal tree nodes, and predictions at terminal nodes) changed with each bootstrap sample. We calculated an additive error component by subtracting a random sample of the model residuals from each bootstrap sample BRT prediction ($\tilde{t}$) for each grid cell in the Central Valley. Subtraction of the model residual ensured that any skew in the residuals had an opposite effect to skew created by uncertainty in the parameters, which is appropriate given that the empirical distribution underlying the interval is based on the difference of predicted and observed concentrations, the former being uncertain due to sampling error in the estimated parameters and the latter due to the model residual. For each grid cell we determined the quantiles for the lower (1 - α/2) and upper (α/2) indices of the distribution of the error component. These indices were 10 and 189, respectively, for 199 bootstrap samples. Both the quantiles and the gridded BRT predictions in log space based on all of the training data were back transformed by exponentiation, and relative prediction intervals were computed for each grid cell as (Schwarz et al., 2006):

$$PI_{lower} = \frac{\hat{t}^2}{Q_{\tilde{t}}\left(1-\frac{\alpha}{2}\right)}$$

$$PI_{upper} = \frac{\hat{t}^2}{Q_{\tilde{t}}\left(\frac{\alpha}{2}\right)}$$

where PI is the lower or upper prediction interval, $\hat{t}$ is the BRT prediction based on all of the training data for the grid cell in real space (mg/L of nitrate), $Q_{\tilde{t}}$ is the quantile of the distribution of the model error component in real space for the grid cell, and α is the significance level (0.10). We made prediction uncertainty maps by plotting prediction interval widths, $PI_{upper} - PI_{lower}$, corresponding to shallow and deep wells for the Central Valley grid cells.

BRT was compared with traditional geospatial extrapolation methods consisting of ordinary kriging (OK), universal kriging (UK), and multiple linear regression (MLR). OK models were cross-validated with 10 folds to determine the optimal number of nearest neighbor observations (100) subsequently used in both OK and UK. The latter used four depth-related variables as regressors in a linear trend component. Both forward and backward selection methods were used in the fitting of MLR models, and predictor variables were retained based on the Akaike Information Criterion. Kriging was performed using the R gstat package (Pebesma, 2004) and MRL was performed in the R MASS package (Venables and Ripley, 2002).

2.4. 2D and 3D mapping

The 1 km nitrate prediction grids for each of the 17 depth zones (Section 2.2) were imported into the Oasis Montaj 9.0.2 mapping environment software (Geosoft, Inc., 2016) for 3D interpolation and visualization. Each grid was assigned a vertical thickness of 1 m and linear interpolation was used between each of the layers at a vertical resolution of 1 m to produce a complete representation of predicted nitrate concentration at depth throughout the Central Valley. For visualization purposes, nitrate predictions were extracted from the interpolated model at 54.86 m and at 121.92 m deep. These depths correspond to the median depths of private and public wells for the training wells (Table 2). Back transformed, BRT-predicted nitrate concentrations were corrected for bias using a smearing estimator (Helsel and Hirsch, 2002).

CVHM/MODPATH average groundwater age at each of the 17 depth layers was also imported into Oasis Montaj software and interpolated according to the methods described above for nitrate predictions. The resulting 3D map of groundwater age indicates the intrinsic susceptibility of the Central Valley aquifer to contaminants introduced at the land surface.

Table 1
Summary of boosted regression tree metaparameter ranges used for cross validation tuning (RMSE, root mean square error).

Parameter	Description	Range	Minimum RMSE model
interaction.depth	Tree depth, or number of layers in each tree.	2–16, by 1	16
n.trees	Total number of trees in the additive model.	500–3000, by 500	1000
shrinkage	Learning rate; determines the contribution of each new tree to the model.	0.002–0.02, by 0.002	0.016

Table 2
Summary of groundwater nitrate concentration in sampled wells, water level, and well depth data for shallow and deep wells used to train the models.

Variable	Shallow wells	Deep wells
No. of wells	1400	2108
Well nitrate concentration (mg/L NO_3-N)		
Minimum	0.004	0.007
Maximum	74.7	49.2
Mean	6.96	3.35
Standard deviation	8.86	4.30
Median	3.84	2.03
Interquartile range	8.62	4.00
Median simulated depth to water (m)[a]	11.74	12.47
Median depth to top of perforated interval (m)[b]	37.49	70.10
Median depth to bottom of perforated interval (m)[b]	54.86	121.92

[a] Based on MODFLOW predicted depth to water (MFDTWSpr2000Faunt).
[b] Based on all well types classified as shallow or deep that had measured depth data, before data declustering.

3. Results and discussion

3.1. Summary statistics for sampled wells

Well nitrate concentrations used for model training ranged from 0.004 mg/L to 74.7 mg/L, over seven times the MCL (Table 2). Shallow training wells tended to have greater nitrate concentrations and more MCL exceedances than deep training wells (Table 2 and Fig. 1). Training well MCL exceedances were concentrated in the central and in the southeastern part of the Central Valley (within the San Joaquin Valley), particularly in the eastern alluvial fans (Fig. 1). Well nitrate concentrations below 2 mg/L NO_3-N were concentrated in the northern third of the Central Valley (the Sacramento Valley) and in the basin subregion throughout the Central Valley (Fig. 1). Training and hold-out wells had near identical distributions of nitrate measurements (SM Fig. S1).

3.2. Modeling results

3.2.1. Model training and testing

Best CV-tuned metaparameters according to the minimum RMSE criterion were 16, 1000, and 0.016 for interaction depth, number of trees, and shrinkage, respectively (Table 3). The final BRT model comprised these metaparameters and 25 predictor variables remaining after the variable reduction described above (Fig. 3). Variables selected for the final model included redox indicators, unsaturated zone nitrogen leaching flux, total landscape nitrogen input, groundwater age, well characteristics, soil variables, groundwater depth, groundwater recharge/flux, and climatic variables (Fig. 3, SM Table S1). Based on relative importance score (Friedman, 2001; Elith et al., 2008) for the final model, the top five predictor variables were two redox variables (probability of manganese (Mn) concentration to exceed 50 ppb and probability of dissolved oxygen concentration to be below 0.5 ppm), field-scale adjusted unsaturated zone nitrogen leaching flux value for the 1975 time period, difference between average precipitation and evapotranspiration between 1971 and 2000, and 1992 total landscape nitrogen input amount (Fig. 3, SM Table S1). Predictor variables used in the final model were mapped for visualization and discussion purposes and appear in order of variable importance (SM Figs. S2-S23). Predictor variables which remained constant within aquifer depth zones (depth to bottom of well screen, use of water at well, and screen length) were not mapped.

The final BRT model had $R^2 = 0.83$ and $RMSE = 0.002$ for training data, and had the highest hold-out R^2 (0.44) and lowest hold-out RMSE (1.13) among all of the models tested (Table 3). (All RMSE values are reported as ln(mg/L NO_3-N). Model estimated versus observed log nitrate values (training data set) mostly plotted along a one-to-one line while plotted model predicted versus observed log nitrate values (hold-out data set) were slightly more dispersed, as would be expected for independent data (Fig. 2). Among traditional modeling approaches, OK (hold-out $R^2 = 0.42$ and hold-out $RMSE = 1.16$) was more competitive with BRT than UK or MLR. MLR used a stepwise procedure with all of the same predictor variables as BRT and explained less than half the variation in the training data (training $R^2 = 0.42$) and had $R^2 = 0.31$ and $RMSE = 1.27$ for hold-out data. UK used the probability that Mn exceeds 50 μg/L in groundwater (ProbMn50ppb), the probability that groundwater dissolved oxygen is <0.5 mg/L (ProbDOpt5ppm), the depth to 60 year old groundwater (DTW60YrJurgens), and average age of groundwater at the well screen (Age_yrs) as regressors in the trend component (all predictor variables are defined in SM Table S1). The regressors were natural log transformed to address non-linear relations with groundwater nitrate, and the resulting UK model had $R^2 = 0.39$ and $RMSE = 1.19$ for hold-out data. The superior performance of BRT with independent data is consistent with prior studies that involved comparisons of machine learning and linear regression, linear classifiers, generalized additive models, OK, and/or UK (Ayotte et al., 2016; Nolan et al., 2015; Wheeler et al., 2015). BRT model residuals were mapped for the training wells in each of the shallow and deep aquifer zones (Section 2.2) separately and no spatial patterns were apparent (SM Fig. S24).

3.2.2. Predictor variable influence

Partial dependence plots of predictor variables in the final model show the behavior of each predictor within the model, after accounting for the average effect of each of the other predictors (Elith et al., 2008). Partial plots were highly useful in gaining insights in the relationship between predictor variables and well nitrate concentrations (SM Fig. S25), and show how the BRT model integrates the effects of low redox conditions, older groundwater, and upward groundwater fluxes associated with groundwater discharge areas. According to partial dependence plots, predicted well nitrate concentrations tended to decrease as the probability of anoxic conditions in wells increased (first two plots in SM Fig. S25). The partial plots for the probability of manganese concentration to exceed 50 ppb and dissolved oxygen concentration to be below 0.5 ppm suggested the presence of a threshold probability value near 0.6. Anoxic conditions have been linked to lower well nitrate concentrations: a study focused in the San Joaquin Valley, California (southern two-thirds of the Central Valley) found significantly lower well nitrate concentrations in anoxic groundwater versus oxic or mixed redox groundwater and the authors attributed this mostly to longer residence times (older groundwater ages) of the groundwater classified as anoxic (Landon et al., 2011). In that study, the anoxic groundwater samples were clustered near the valley trough (Landon et al., 2011), where the probability of anoxic conditions has also been estimated to be greatest (above 0.6) (Rosecrans et al., 2017) (SM Figs. S2 and S3). Landon et al. (2011) also found that decreases in well nitrate concentration due to denitrification were mostly small throughout their study region and did not protect wells on a regional scale from nitrate contamination. Results of another regional study focused in the eastern San Joaquin Valley agree with the findings of Landon et al. (2011): based on multi-model residence time distributions, estimated nitrate reduction rates were significant for the denitrification zone; however, well nitrate concentrations will continue to increase given current nitrogen inputs even after accounting for the rates of oxygen and nitrate reduction (Green et al., 2016).

Estimated adjusted field-scale unsaturated zone nitrogen leaching flux to the water table based on components of the GNLM model for the 1975 time period was ranked third in terms of variable relative importance in the final model (Fig. 3). Predicted well nitrate concentration increased rapidly as the normalized adjusted nitrogen flux value for 1975 increased to about 20% (see partial plot Ngw_1975 in SM Fig. S25). Partial dependence plots for other GNLM time periods in the full model (not shown) were more erratic and did not display a clear relationship between well nitrate concentration and adjusted nitrogen flux value. In addition, adjusted GNLM nitrogen leaching flux value for

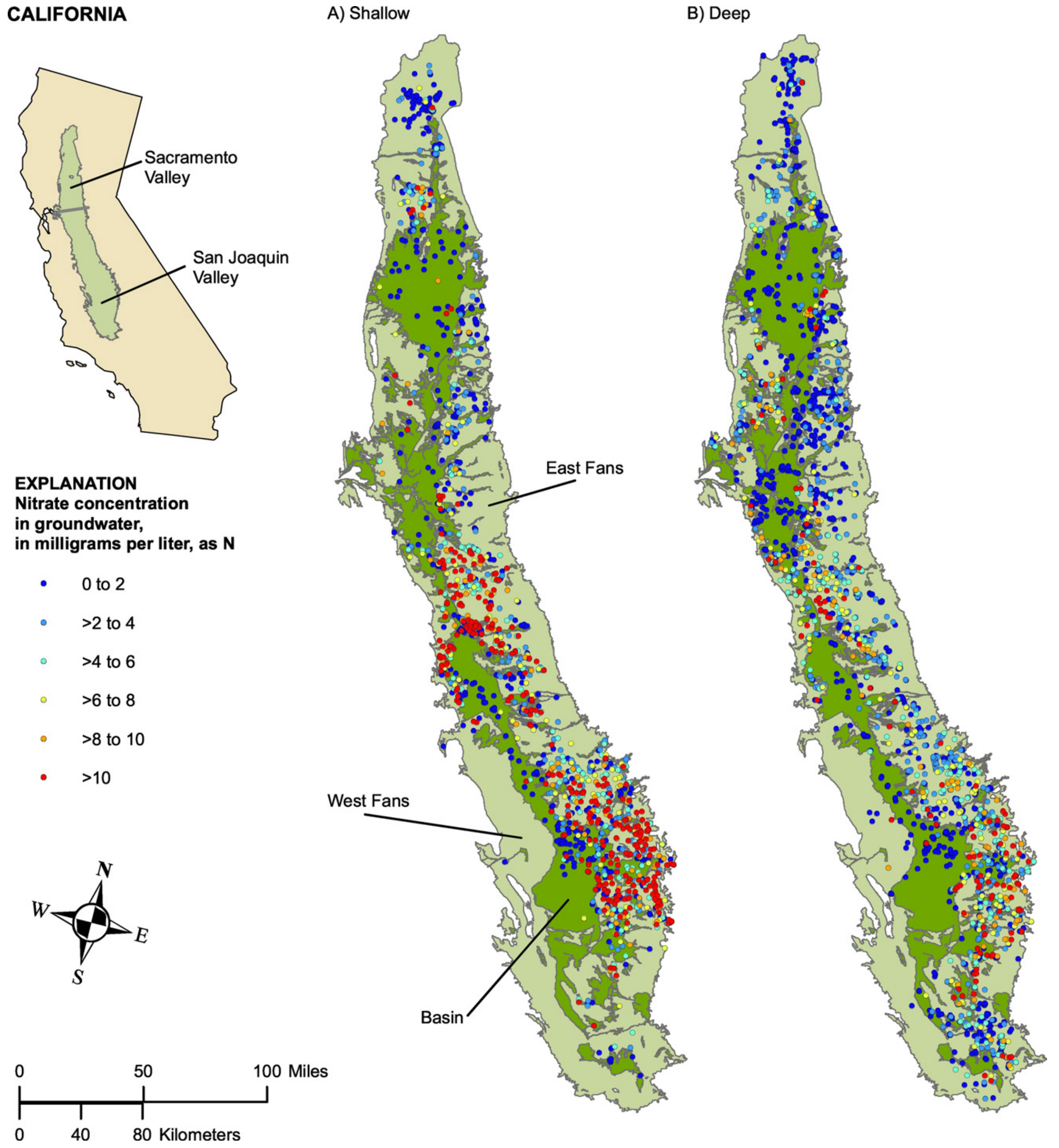


Fig. 1. Training well locations, color coded by well nitrate concentration (3508 wells total) for shallow (1400 wells, mostly private) and deep (2108 wells, mostly public supply) zones.

1975 was ranked third in terms of variable importance in the full model, above the other time periods, which suggested a relationship between this variable and total travel time through the unsaturated zone and aquifer to the sampled wells. Training well MCL exceedances for both shallow and deep wells (shown as red dots in Fig. 1) appeared to be spatially correlated to regions with GNLM adjusted nitrogen flux values above 10% for the 1975 period (compare Fig. 1 with SM Fig. S4). Total landscape nitrogen input to the land surface for the year 1992 was also relevant (relative importance rank fifth) and represented a more conventional estimate of nitrogen input from readily available fertilizers and atmospheric deposition data (SM Fig. S6). Including both total landscape nitrogen inputs and GNLM-based unsaturated zone nitrogen leaching flux in the BRT model allowed us to evaluate the relative importance of these two nitrogen loading estimation methods, and also improved the fit of the final model. Removing either the adjusted field-scale nitrogen leaching flux variable or the total landscape nitrogen input variable from the final model resulted in a near identical decrease in hold-out R^2 value and increase in RMSE value compared to the final model (about 0.438 and 1.137 compared to 0.443 and 1.132 from Table 3). The GNLM model includes additional sources of N, such as septic systems and urban N losses, and also accounts for N removal by harvested crops. BRT predicted nitrate tended to increase as total landscape nitrogen flux for 1992 increased to about 6000 kg within 500 m of a well (partial plot N_total in SM Fig. S25). Total landscape

Table 3
Boosted regression tree (BRT) model training and testing results for full and final models (CV, cross validation; R^2, coefficient of determination; RMSE, root mean square error).

Model	10-fold CV Testing		Training		Hold-out	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
Full BRT model, CV testing, best metaparameters	0.452	1.185	0.873	0.003	0.434	1.142
Final BRT model after variable reduction	NA	NA	0.825	0.002	0.443	1.132
Ordinary kriging	0.423	1.214	NA	NA	0.415	1.164
Universal kriging	0.400	1.238	NA	NA	0.390	1.185
Multiple linear regression	NA	NA	0.419	1.218	0.306	1.266

nitrogen flux for 1992 also tended to appear spatially correlated to training well MCL exceedances (Fig. 1 and SM Fig. S6).

Precipitation minus evapotranspiration was ranked number 4 in the BRT model and the partial plot indicates a steep decrease in well nitrate concentration for values greater than −20 in/yr (PrecipMinusET partial plot in SM Fig. S25). Values of this variable become increasingly negative going from north to south in the Central Valley (SM Fig. S5) and this pattern appears spatially correlated to the training well MCL exceedances (Fig. 1), which also exhibited a north-south gradient. These patterns could be the result of precipitation and sediment texture. The Sacramento Valley has greater precipitation and more fine-grained sediments than the San Joaquin Valley, which contribute to anoxic conditions (Burow et al., 2013).

Well nitrate concentrations increased with increasing average silt content (SM Fig. S18, importance ranking 19th), hydrologic group C (soils with a layer that restricts downward flow of water, or soils with moderately fine or fine texture, SM Fig. S19, importance ranking 21st), and decreasing percent coarse materials in the upper CVHM model layer (SM Fig. S12, importance ranking 11th). This is counter to previous aquifer vulnerability models of the Central Valley, which showed increasing well nitrate concentrations with decreasing poorly drained soils or increasing well-drained soils (Nolan et al., 2014). Previous cluster and principal component analysis has linked the presence of hardpan or fine-textured soils to the presence of pesticides in groundwater of the Central Valley which suggests a contamination pathway other than leaching (Troiano et al., 1994). In the current study, hydrologic group C soils may indicate the possibility of alternative contamination pathways such as cracks or dry wells, which are common in some areas of hardpan soils in the Central Valley (DeMartinis and Royce, 1990).

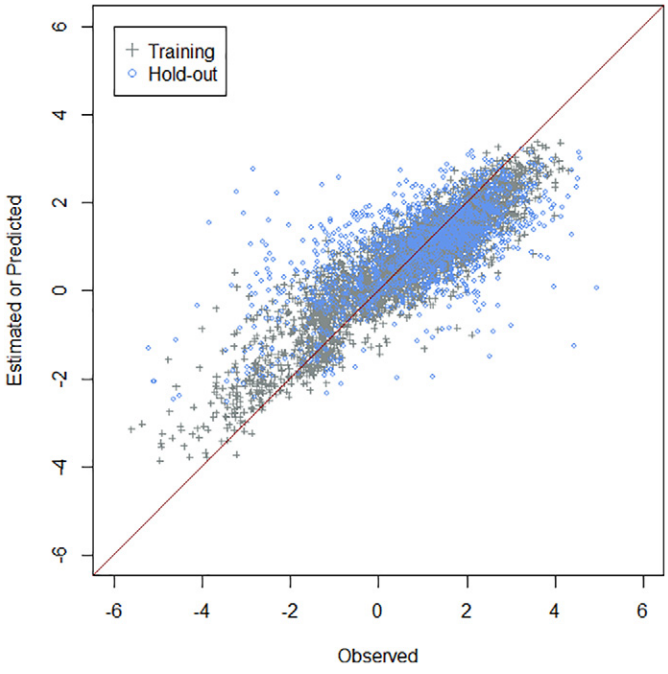


Fig. 2. Final model estimated versus observed log nitrate values (training data set) and final model predicted versus observed log nitrate values (hold-out data set). Plots correspond to the R^2 values of 0.83 for the training data set and 0.44 for the hold-out data set).

3.3. Groundwater nitrate and mean age mapping

Extractions of Oasis Montaj interpolated nitrate predictions at the shallow and deep (private and public supply well depths of 54.86 m and 121.92 m, respectively, Table 2) typically were greater for the private supply well depth versus the public supply well depth as indicated by the more intense colors in the map of the former, particularly in the eastern alluvial fans of the San Joaquin Valley (Fig. 4). The majority of the grid cell predictions for each layer were below the nitrate drinking water standard of 10 mg/L NO_3-N. Empirical cumulative distribution functions show the proportion (p) of grid cells with predicted nitrate concentration less than or equal to an indicated value (Fig. 5). The exceedance rate of an indicated nitrate concentration is 1-p. Based on the gridded predictions, the rate of a domestic well exceeding 10 mg/L of nitrate was 0.02 and that of a public supply well was <0.01. In contrast, the raw exceedance rates based on measured groundwater nitrate concentration were 0.27 and 0.06, respectively. At 5 mg/L of nitrate (half the MCL), the predicted exceedance rate for domestic wells was 0.21 and that of public supply wells was 0.04, whereas raw measured exceedance rates were 0.47 and 0.23, respectively. These differences may reflect well sampling bias towards areas of known nitrate contamination, whereas the model predictions are for grid cells of uniform size (1 km^2) throughout the Central Valley, including extensive areas that lack measured data. As described above, we attempted to compensate for well sampling bias by declustering the data. The differences may also reflect negative bias on the part of the model, which somewhat underestimates high nitrate concentration (Fig. 2). Lastly, linear interpolation in Oasis Montaj resulted in additional smoothing of the BRT-predicted nitrate values. To put these values in a more consequential context, the private well depth layer had the equivalent of a 1021 km^2 and 10,366 km^2 area with predictions greater than the MCL and one half the MCL, respectively, while the public supply well depth had the equivalent of a 25 km^2 and 2123 km^2 area with predictions greater than the MCL and one half the MCL, respectively (compared to a total prediction area of 48,802 km^2).

The predictions are similar to those of Nolan et al. (2014), except in the southern-most portion of the Central Valley, where our current model has generally lower nitrate predictions for both depths (Fig. 4). (Following Burow et al. (2013), Nolan et al. used a well depth of 46 m to designate wells as shallow or deep.) This may be due to the additional predictor variables not included in the Nolan et al. (2014) model, such as depth to 60 year water, which had a strong north to south gradient (SM Fig. S7) and which exhibited a strong response in predicted nitrate (partial plot DTW60YrJurgens in SM Fig. S25). The greatest values for depth to 60 year old water (> 91 m) were located in the southern-most region of the Central Valley and greater values of this variable were related to lower nitrate concentrations suggesting increased travel time in the unsaturated zone in the southern San Joaquin Valley. Unsaturated zone thickness, indicated by depth to water (MFDTWSpr2000Faunt), is greater in this region (SM Fig. S14). As depth to water increases, unsaturated zone travel time increases and there is greater opportunity for nitrogen transformation processes and/or semi confining layers to restrict nitrate transport to groundwater. The depth to 60 year water variable incorporates unsaturated zone travel times of between 0 and 24 years (SM S3.0). Therefore, depth to 60 year water was likely a cause of the lower nitrate predictions when compared to Nolan et al. (2014) for the southern region. For both the shallow and deep zones, the greatest predicted nitrate concentrations were largely located within the eastern and western alluvial

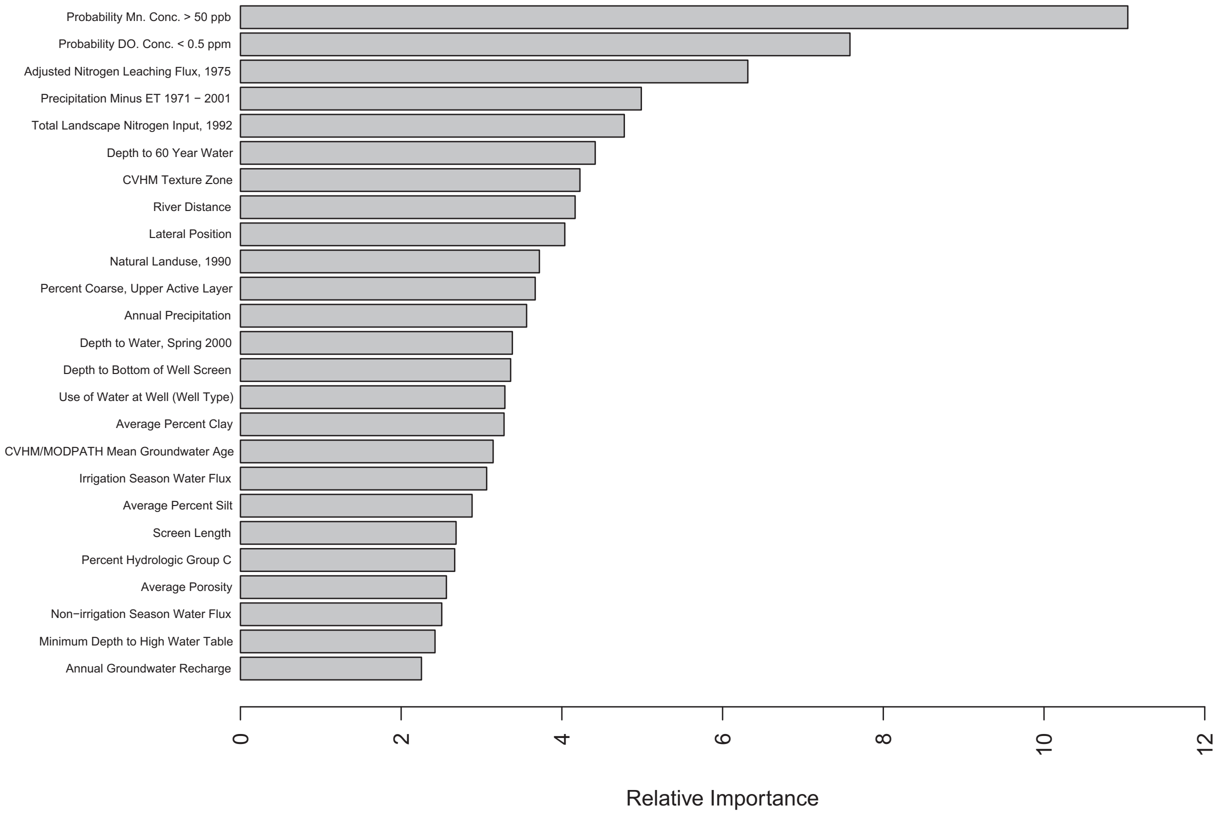


Fig. 3. Relative influence of variables in the final boosted regression tree model (all predictor variables described in Table S1). Mn = Manganese, Conc = Concentration, DO = Dissolved Oxygen, ET = Evapotranspiration.

fans subregions (Fig. 4). As with Nolan et al. (2014), predicted nitrate concentrations at background levels (<2 mg/L NO_3-N) were concentrated in the basin subregion, along the axis of the Central Valley. Predicted concentrations above the MCL (>10 mg/L NO_3-N) were much more common in the shallow well zone than in the deep well zone, and are mainly concentrated in the eastern alluvial fan region for both depths (Fig. 4). Similarly, predicted nitrate concentrations of <2 mg/L largely follow the pattern of predicted high probability of anoxic conditions and low lateral positions, which generally correspond to the basin subregion (SM Figs. S2, S3, and S10). In the current study, in the eastern alluvial fans subregion, predicted nitrate concentrations <2 mg/L spatially correlate with small distance to major rivers (<3000 m) (Fig. 4 and SM Fig. S9). This pattern is apparent in our results due to the river distance variable, which was the eighth most important variable in the final model. Wells nearer to rivers may have lower nitrate concentrations due to low nitrate concentration in infiltrating river water (Boyle et al., 2012). Predicted nitrate >2 mg/L in oxic regions of the valley was variable due to the spatial distribution of 1975 unsaturated zone nitrogen leaching flux and 1992 total landscape nitrogen inputs as well as other properties such as percent silt, percent clay, and hydrologic group C soils (Section 3.2.2 and SM Figs. S4, S6, S15, S18, and S19).

The 3D maps revealed that predicted groundwater nitrate concentrations generally decreased with depth for the entire Central Valley (Fig. 6). However, predicted nitrate concentrations remained above the MCL for portions of the southern eastern alluvial fans subregion and at elevated concentrations (above 4 mg/L NO_3-N) for portions of the northern eastern alluvial fans and western alluvial fans subregions down to the deepest depth for which predictions were made (487.68 m) (Fig. 6). This is likely due to a combination of high nitrogen loading and low probability of anoxic conditions as well as younger relative groundwater age at depth (<250–500 years), especially for the eastern alluvial fans subregion (SM Fig. S16). Portions of the basin subregion had older predicted groundwater ages at shallow depths (SM Fig. S16). This is likely the result of discharge of groundwater with long residence times along regional flow lines that recharged prior to the modern period of high nitrogen applications, and which is also more likely to be anoxic. MODFLOW non-irrigation and irrigation season vertical water fluxes show upward groundwater fluxes primarily in these same areas, and more extensive areas of strong upward fluxes (0 to 10,000 or more m^3/d) during the non-irrigation season (SM Figs. S17 and S21). Otherwise, modeled groundwater age tended to increase with depth throughout the model domain (SM Fig. S16).

According to Sobol' sensitivity analysis, the most sensitive variables in the model were the two redox variables (probability of manganese concentration to be >50 ppb and probability of dissolved oxygen concentration to be <0.5 ppm), percent coarse textured soils in the upper active CVHM model layer, and depth to 60 year old water (SM Fig. S26). Redox/age proxies such as depth to bottom of well screen and screen length had comparatively low sensitivities, indicating the added value of using direct estimates of redox and groundwater age in the BRT model. These results are generally consistent with the relative influence of predictor variables by BRT, and underscore the importance of groundwater redox and age to nitrate occurrence in the Central Valley aquifer.

Maps of BRT prediction interval width show that prediction uncertainty is greater in the alluvial fans than the basin subregion (Fig. 7). The uncertainty is related to the variance of the bootstrap samples, which is greater in the fan subregions owing to the heterogeneity of sediments, variable redox conditions, and localized geologic sources of nitrogen (Burow et al., 2013). In contrast, groundwater travelling to the center of the basin is older, more reduced, and typically has lower nitrate concentration.

3.4. Conclusions

The final BRT model with 25 predictor variables including CVHM/MODPATH modeled mean groundwater age had higher prediction

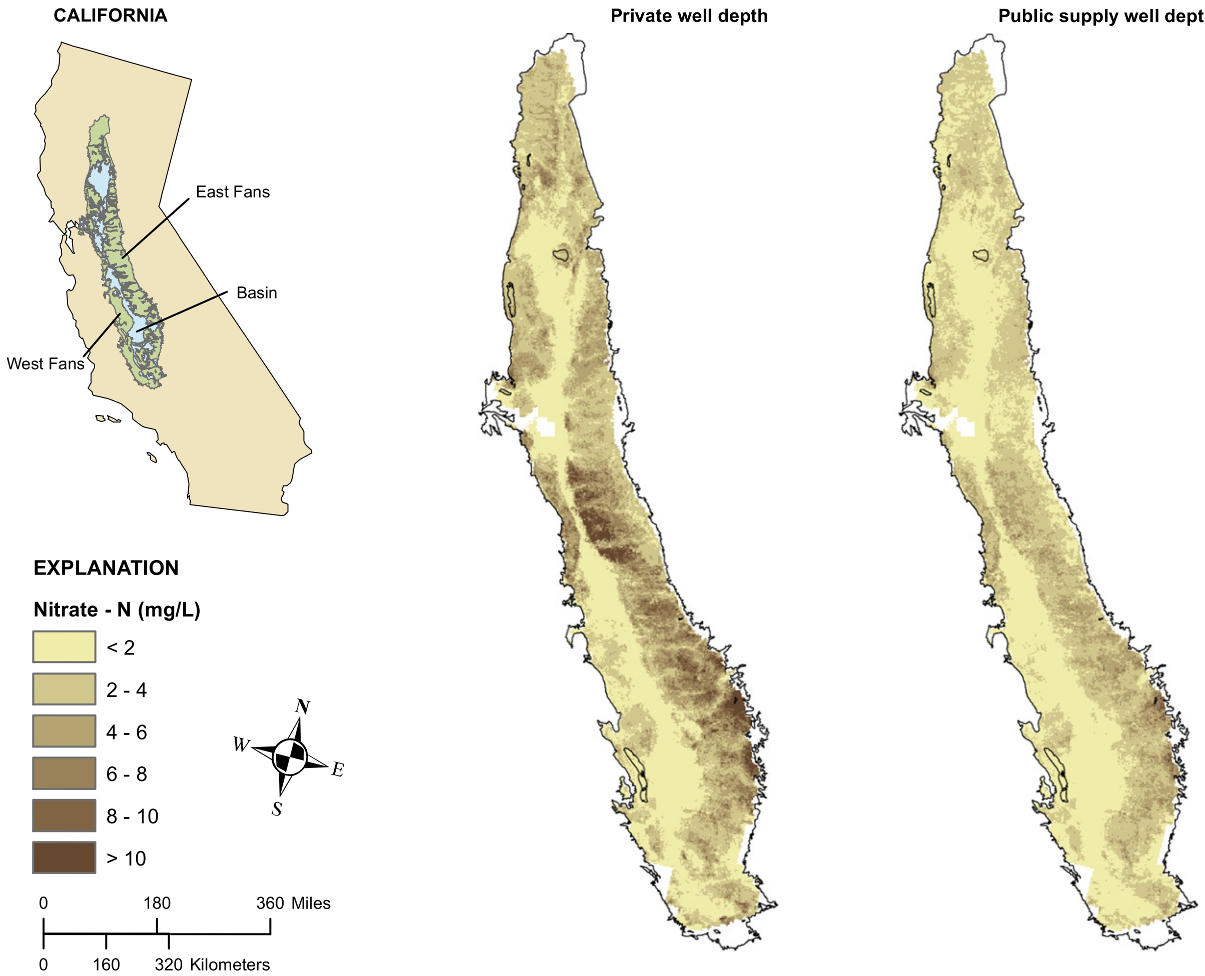


Fig. 4. Oasis Montaj interpolated boosted regression tree prediction of groundwater nitrate at median depths of private and public supply wells (54.86 m and 121.92 m, respectively). Unmapped (white) area within the alluvial bed boundary was due to missing data for one or more predictors in the final BRT model.

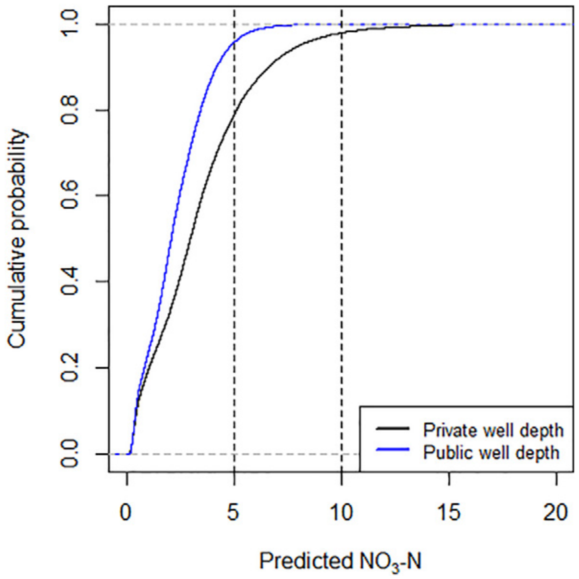


Fig. 5. Empirical cumulative distribution functions of predicted groundwater nitrate concentrations for map grid cells corresponding to the private (black line) and public supply (blue line) well depths depicted in Fig. 4. Vertical lines are the MCL and one half the MCL for nitrate (NO_3-N).

accuracy to hold-out data ($R^2 = 0.44$) than previous groundwater nitrate models of the Central Valley that lacked direct estimates of groundwater age and redox conditions. Cross-validation tuning of metaparameter values within a statistical learning framework optimized model performance with new data, which benefits mapping in unsampled areas. The BRT model had lower overall error rates compared to OK, UK, and MLR. The incorporation of modeled groundwater age at depth was a key component not included in previous models (Nolan et al., 2015, 2014), which relied on age proxies such as depth to the top of the well screen. In the present study, direct estimates of groundwater age enhanced understanding and prediction of nitrate occurrence at all drinking water depths in the Central Valley aquifer. The 3D map of groundwater age complements that of nitrate and also depicts the relative intrinsic susceptibility of the Central Valley aquifer to contaminants originating at the land surface. Other researchers' model outputs included as predictor variables in our BRT model were also highly important predictors of groundwater nitrate including probability of anoxic conditions (Rosecrans et al., 2017) and field-scale unsaturated zone nitrogen leaching flux (Viers et al., 2012; Rosenstock et al., 2013). Predicted nitrate concentration followed a similar pattern as high probability of anoxic conditions in the basin subregion, but in the alluvial fan subregions, which have low probability of anoxic conditions, was shaped by other factors, such as unsaturated zone nitrogen leaching flux, total landscape nitrogen inputs, and soil properties.

The results of our study highlight the usefulness of the hybrid modeling approach and resulted in more accurate predictions of nitrate

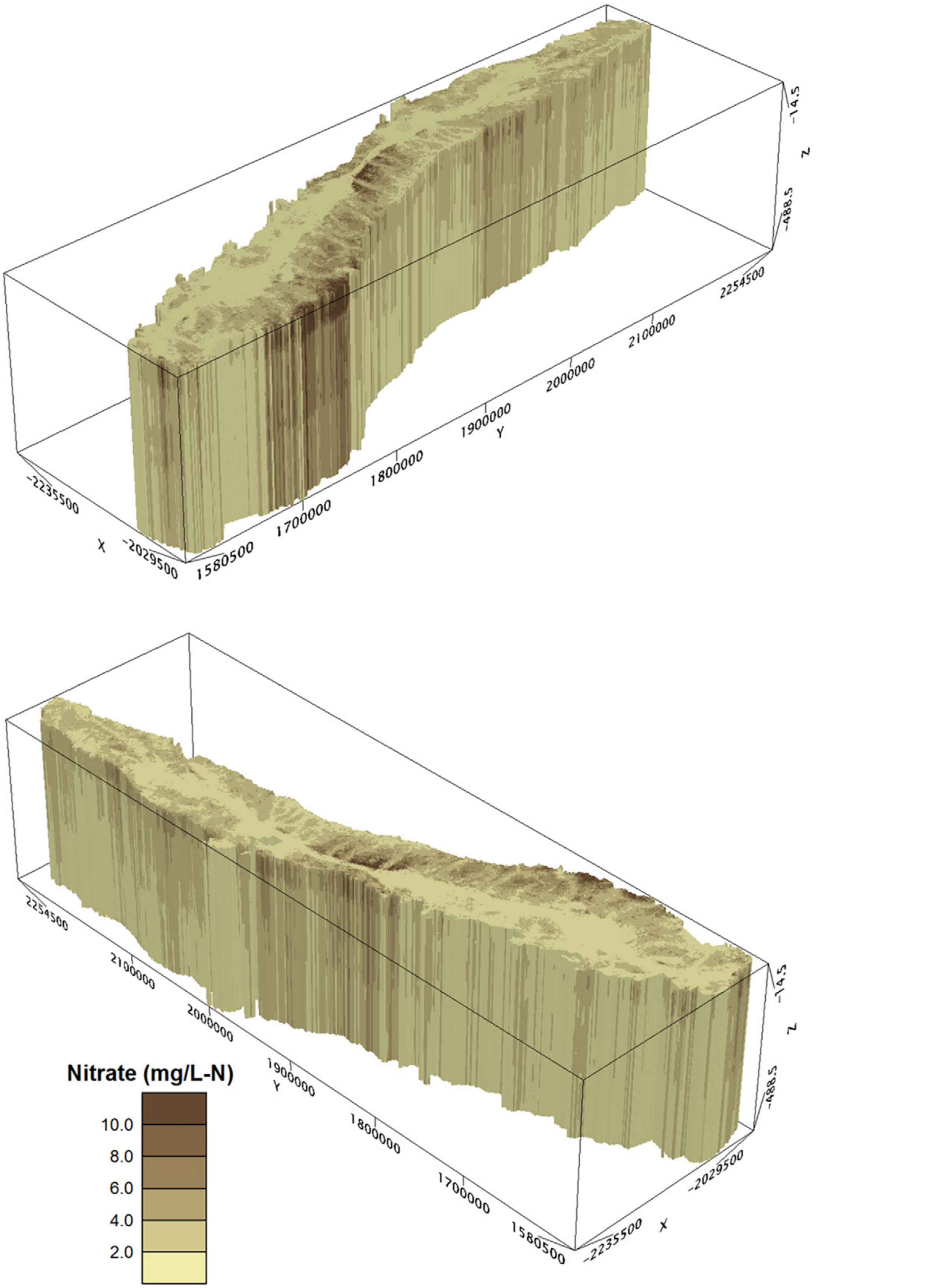


Fig. 6. Oasis Montaj interpolated groundwater nitrate predictions for the Central Valley aquifer (view from southeast on top, view from southwest on bottom). Vertical scale is depth in m.

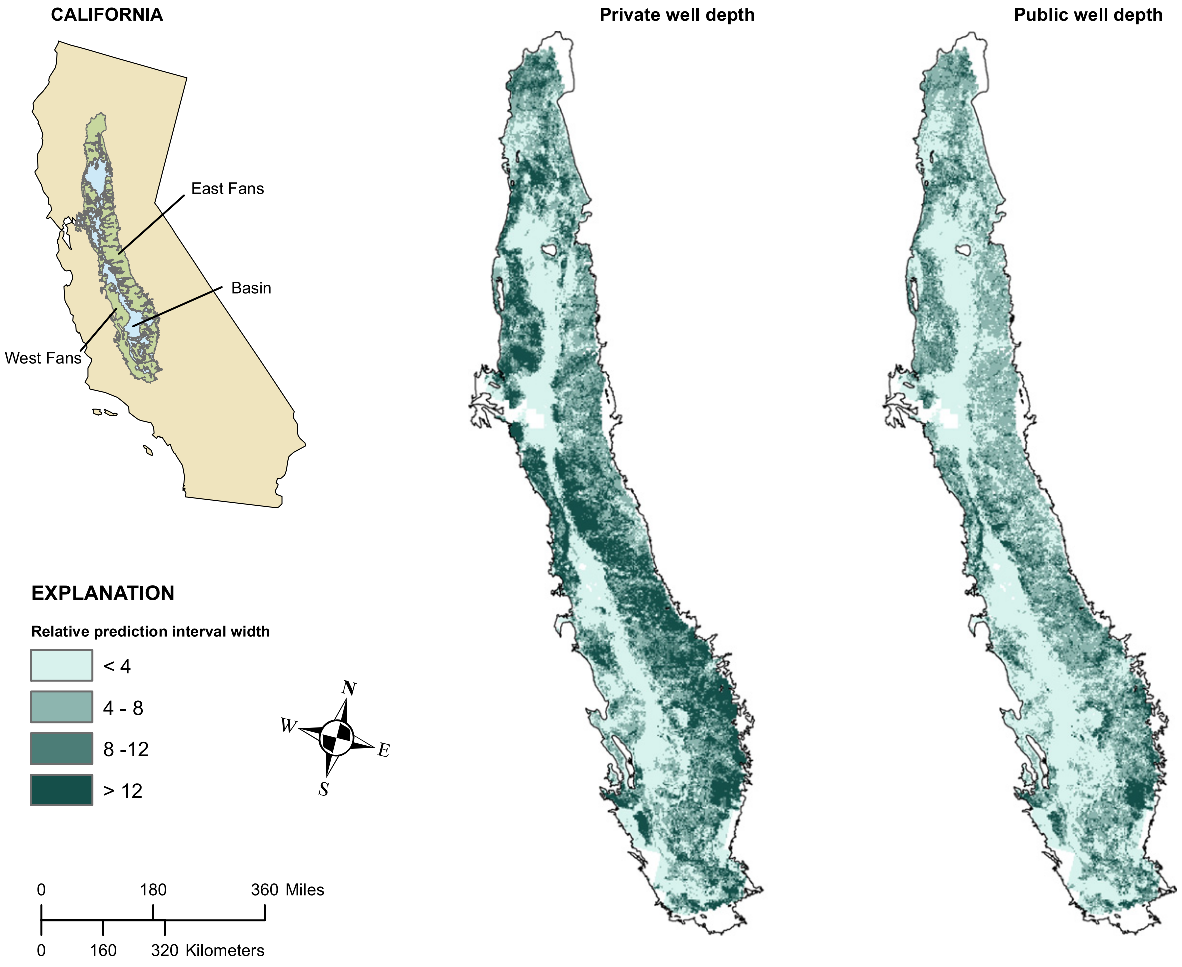


Fig. 7. Relative prediction interval widths at median measured depths of private and public supply wells (54.86 m and 121.92 m, respectively). Interval widths are in mg/L nitrate, NO_3-N. Unmapped (white) area within the alluvial bed boundary was due to missing data for one or more predictors in the final BRT model.

in groundwater in the Central Valley. This model could be updated in the future and potentially improved with estimates of unsaturated zone travel time and geostatistically interpolated estimates of groundwater age from tritium/helium age tracers (Visser et al., 2016).

We anticipate that the model can be used by local agencies developing groundwater management plans in response to California's Sustainable Groundwater Management Act (SGMA, 2014), which stipulates management and use of groundwater without significant degradation of water quality. Reliable models can be used to extend monitoring results in space; to inform the design of cost effective monitoring programs by targeting sampling to the most vulnerable areas; and to forecast future conditions where it is impractical to take decades' worth of samples. In this way, predictions of nitrate concentration at private and public supply well depths can help resource managers protect groundwater quality and support well owners who rely on groundwater for their daily needs.

Acknowledgement

We would like to thank Quinn Barber for his work compiling the UC Davis database of domestic well samples, and Jerry and Donna White for their generous donation (White Family Graduate Student Award 2015), which supported GIS processing of predictor variables. We thank Steven Predmore for compiling MODFLOW/MODPATH estimates of groundwater age, and Tyler Johnson for declustering the sampled wells in GIS. Katherine Ransom was supported in part by the USGS National Water Quality Program via a California Cooperative Ecosystem Studies Unit grant number G15AC00114 and by the California State Water Resources Control Board award number 09-122-250. The authors gratefully acknowledge the financial support of the Center for Watershed Sciences made possible by a gift from the S. D. Bechtel, Jr. Foundation. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Groundwater nitrate data from sampled wells used to train and test the models, raster files of the 25 predictor variables in the final boosted regression tree model, and raster files of predicted groundwater nitrate concentration data at the depths of private and domestic wells are publicly available at https://doi.org/10.5066/F7V40SDN (Ransom et al., 2017b).

Appendix A. Supporting materials

Supporting materials to this article can be found online at http://dx.doi.org/10.1016/j.scitotenv.2017.05.192.

References

Anning, D., Paul, A., McKinney, T., Huntington, J., Bexfield, L., Thiros, S., 2012. Predicted nitrate and arsenic concentrations in basin-fill aquifers of the Southwestern United States. Scientific Investigations Report 2012–5065. U.S. Geological Survey.

ArcGIS, 2016. Empirical Bayesian kriging. Retrieved from. http://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/what-is-emperical-bayesian-kriging-.htm.

Ayotte, J.D., Nolan, B.T., Gronberg, J., 2016. Predicting arsenic in drinking water wells of the Central Valley, California. Environ. Sci. Technol. 50 (14), 7555–7563.

Belitz, K., Jurgens, B., Landon, M.K., Fram, M.S., Johnson, T., 2010. Estimation of aquifer scale proportion using equal area grids: assessment of regional scale groundwater quality. Water Resour. Res. 46 (11), 1–14.

Belitz, K., Fram, M.S., Johnson, T.D., 2015. Metrics for assessing the quality of groundwater used for public supply, CA, USA: equivalent-population and area. Environ. Sci. Technol. 49 (14), 8330–8338.

Boyle, D., King, A., Kourakos, G., Lockhart, K., Mayzelle, M., Fogg, G., Harter, T., 2012. Groundwater nitrate occurrence. Technical report 4 in: 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.

Burow, K.R., Jurgens, B.C., Belitz, K., Dubrovsky, N.M., 2013. Assessment of regional change in nitrate concentrations in groundwater in the Central Valley, California, USA, 1950s-2000s. Environ. Earth Sci. 69, 2609–2621.

Canter, L.W., 1996. Nitrates in Groundwater. CRC Press LLC, Boca Raton, Florida.

CSWRCB, 2010. Groundwater Ambient Monitoring and Assessment (GAMA) Domestic Well Project Groundwater Quality Report, Tulare County Focus Area. California State Water Resources Control Board, Groundwater Protection Section, Sacramento, California.

De'ath, G., 2007. Boosted trees for ecological modeling and prediction. Ecology 88 (1), 243–251.

DeMartinis, J.M., Royce, K.L., 1990. Identification of direct-entry pathways by which agricultural chemicals enter ground water. Proceedings of the 1990 Cluster of Conferences, Agricultural Impacts Ground Water Quality, pp. 51–65 (Kansas City, MO.: National Well Water Association, Westerville, OH).

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77 (4), 802–813.

Faunt, C.C., 2009. Groundwater availability of the Central Valley aquifer, California. Professional Paper 1766. U.S. Geological Survey.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378.

Geosoft, Inc., 2016. Oasis Montaj: Integrated Platform for Earth Exploration, version 9.0.2. http://www.geosoft.com/products/oasis-montaj/overview.

Green, C., Jurgens, B.C., Zhang, Y., Starn, J.J., Visser, A., Singleton, M., Esser, B., 2016. Regional oxygen reduction and denitrification rates estimated from mixed age groundwater samples, San Joaquin Valley, USA. J. Hydrol. 145, 47–55.

Gronberg, J.M., Spahr, N.E., 2012. County-level estimates of nitrogen and phosphorus from commercial fertilizer for the conterminous United States, 1987–2006. Scientific Investigations Report 2012–5207. U.S. Geological Survey.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. second ed. Springer Science + Business Media, LLC, New York.

Helsel, D.R., Hirsch, R.M., 2002. Chapter A3, statistical methods in water resources. Techniques of Water-resources Investigations of the United States Geological Survey, Book 4, Hydrologic Analysis and Interpretation. U.S. Geological Survey, p. 524.

Hijmans, R.J., 2016. Geographic data analysis and modeling R package version 2.5–8. Retrieved from. https://CRAN.R-project.org/package=raster.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978.

Honeycutt, K., Canada, H.E., Jenkins, M.W., Lund, J.R., 2012. Alternative water supply options for nitrate contamination. Technical Report 7. Center for Watershed Sciences, University of California, Davis, California.

Johnson, T.D., Belitz, K., 2009. Assigning land use to supply wells for the statistical characterization of regional groundwater quality: correlating urban landuse and VOC occurrence. J. Hydrol. 370, 100–108.

Johnson, T.D., Belitz, K., 2015. Identifying the location and population served by domestic wells in California. J. Hydrol. Reg. Stud. 3, 31–86.

Jurgens, B.C., Bohlke, J.K., Eberts, S.M., 2012. TracerLPM (version 1): An Excel workbook for interpreting groundwater age distributions from environmental tracer data. Techniques and Methods Report 4-F3. U.S. Geological Survey (60 pp.).

Jurgens, B.C., Bohlke, J.K., Kauffman, L.J., Belitz, K., Esser, B.K., 2016. A partial-exponential lumped parameter model to evaluate groundwater age distributions and nitrate trends in long-screened wells. J. Hydrol. 543, 109–126.

Kuhn, M., 2016. Caret: classification and regression training, R package version 6.0–71. Retrieved from. https://CRAN.R-project.org/package=caret.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer Science + Business Media, LLC, New York.

Landon, M.K., Green, C.T., Belitz, K., Singleton, M.J., Esser, B.K., 2011. Relations of hydrogeologic factors, groundwater reduction-oxidation conditions, and temporal and spatial distributions of nitrate, central-eastside San Joaquin Valley, California, USA. Hydrogeol. J. 19, 1203–1224.

Lockhart, K.M., King, A.M., Harter, T., 2013. Identifying sources of groundwater nitrate contamination in a large alluvial groundwater basin with highly diversified intensive agricultural production. J. Contam. Hydrol. 151, 140–154.

Lubin, J.H., Colt, J.S., Camann, D., Davis, S., Cerhan, J.R., Severson, R.K., Bernstein, L., Hartge, P., 2004. Epidemiologic evaluation of measurement data in the presence of detection limits. Environ. Health Perspect. 112 (17), 1691–1696.

Mueller, D., Gronberg, J., 2013. County-level estimates of nitrogen and phosphorus from animal manure for the conterminous United States, 2002. Open-File Report 2013–1065. U.S. Geological Survey.

Nolan, B.T., Hitt, K.J., 2006. Vulnerability of shallow groundwater and drinking-water wells to nitrate in the United States. Environ. Sci. Technol. 40 (24), 7834–7840.

Nolan, B.T., Gronberg, J.M., Faunt, C.C., Eberts, S.M., Belitz, K., 2014. Modeling nitrate at domestic and public-supply well depths in the Central Valley, California. Environ. Sci. Technol. 48 (10), 5643–5651.

Nolan, B.T., Fienen, M.N., Lorenz, D.L., 2015. A statistical learning framework for groundwater nitrate models. J. Hydrol. 531, 902–911.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691.

Pujol, G., Iooss, B., Janon, A., 2017. Sensitivity: global sensitivity analysis of model outputs R package version 1.14.0. Retrieved from. https://CRAN.R-project.org/package=sensitivity.

R Core Team, 2016. R: A language and environment for statistical computing. 3.3.1. R Foundation for Statistical Computing, Vienna, Austria Retrieved from https://www.R-project.org.

Ransom, K.M., Bell, A.M., Barber, Q.E., Kourakos, G., Harter, T., 2017a. A Bayesian approach to infer nitrogen loading rates from crop and landuse types surrounding private wells in the Central Valley, California. Hydrol. Earth Syst. Sci. Discuss. 2017. http://dx.doi.org/10.5194/hess-2017-39.

Ransom, K.M., Nolan, B.T., Bell, A.M., Gronberg, J.M., 2017b. Groundwater nitrate data and ascii grids of predicted nitrate and model inputs for the Central Valley aquifer, California, USA. U.S. Geological Survey Data Release https://dx.doi.org/10.5066/F7V40SDN.

Ridgeway, G., et al., 2015. gbm: generalized boosted regression models R package version 2.1.1. Retrieved from. https://CRAN.R-project.org/package=gbm.

Rosecrans, C.Z., Nolan, B.T., Gronberg, J.A., 2017. Prediction and visualization of redox conditions in the groundwater of Central Valley, California. J. Hydrol. 546, 341–356.

Rosenstock, T.S., Liptzin, D., Six, J., Tomich, T.P., 2013. Nitrogen fertilizer use in California: assessing the data, trends and a way forward. Calif. Agric. 67 (1), 68–79.

Ruddy, B., Lorenz, D., Mueller, D., 2006. County-level estimates of nutrient inputs to the land surface of the conterminous United States, 1982–2001. Scientific Investigations Report 2006–5012. U.S. Geological Survey.

Rupert, M.G., 1998. Probability of detecting atrazine/desethyl-atrazine and elevated concentrations of nitrate (NO2 + NO3 – N) in ground water in the Idaho part of the upper Snake River basin. Water-resources Investigations Report 98–4203. U.S. Geological Survey.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S., 2010. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Comput. Phys. Commun. 181, 259–270.

Schwarz, G.E., Hoos, A.B., Alexander, R.B., Smith, R.A., 2006. The SPARROW surface water quality model: theory, application, and user documentation. U.S. Geological Survey Techniques and Methods Book 6, Section B, Chapter 3 Retrieved from https://pubs.usgs.gov/tm/2006/tm6b3/.

SGMA, 2014. Sustainable groundwater management act. Retrieved from. http://groundwater.ca.gov/legislation.cfm.

Troiano, J., Johnson, B.R., Powell, S., 1994. Use of cluster and principal component analyses to profile areas in California where ground water has been contaminated by pesticides. Environ. Monit. Assess. 32, 269–288.

U.S. Department of Agriculture, 2014. Soil survey geographic (SSURGO) database, digital data set. Natural Resources Conservation Service. Retrieved January 2016, from. http://soildatamart.nrcs.usda.gov/.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. fourth ed. Springer, New York Retrieved from http://www.stats.ox.ac.uk/pub/MASS4.

Viers, J.H., Liptzin, D., Rosenstock, T.S., Jensen, V.B., Hollander, A.D., McNally, A., King, A.M., Kourakos, G., Lopez, E.M., De La Mora, N., Fryjoff-Hung, A., Dzurella, K.N., Canada, H., Laybourne, S., McKenney, C., Darby, J., Quinn, J.F., Harter, T., 2012. Nitrogen sources and loading to groundwater, Technical Report 2 in: 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.

Visser, A., Moran, J., Hillegonds, D., Singleton, M.J., Kulongoski, J.T., Belitz, K., Esser, B.K., 2016. Geostatistical analysis of tritium, groundwater age and other noble gas derived parameters in California. Water Res. 91, 314–330.

Voss, S.A., Jurgens, B.C., 2017. Spatial point data sets and interpolated surfaces of well construction characteristics for domestic and public supply wells in the Central Valley, California, USA. U.S. Geological Survey Available at http://dx.doi.org/10.5066/F76Q1V9G.

Ward, M.H., deKok, T.M., Levallois, P., Brender, J., Gulis, G., Nolan, B.T., VanDerslice, J., 2005. Workgroup report: drinking-water nitrate and health—recent findings and research needs. Environ. Health Perspect. 113 (11), 1607–1614.

Warner, K.L., Arnold, T.L., 2010. Relations that affect the probability and prediction of nitrate concentration in private wells in the glacial aquifer system in the United States. Scientific Investigations Report 2010–5100. U.S. Geological Survey.

Wheeler, D.C., Nolan, B.T., Flory, A.R., DellaValle, C.T., Ward, M.H., 2015. Modeling groundwater nitrate concentrations in private wells in Iowa. Sci. Total Environ. 536, 481–488.

Wieczorek, M., 2014. Area-weighted and depth-weighted averages of selected SSURGO variables for the conterminous United States and District of Columbia. Data series 866. Retrieved January 2016, from. https://pubs.er.usgs.gov/publication/ds866.

Wolock, D.M., 2003. Estimated mean annual natural ground-water recharge in the conterminous United States, digital data set. Retrieved January 2016, from. http://water.usgs.gov/lookup/getspatial?rech48grd.