A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA

Bernard T. Nolan [a,*], Michael N. Fienen [b], David L. Lorenz [c]

[a] U.S. Geological Survey, National Center, 12201 Sunrise Valley Drive, Reston, VA 20192, USA
[b] U.S. Geological Survey, Wisconsin Water Science Center, 8505 Research Way, Middleton, WI 53562, USA
[c] U.S. Geological Survey, Minnesota Water Science Center, 2280 Woodale Drive, Mounds View, MN 55112, USA

SUMMARY

We used a statistical learning framework to evaluate the ability of three machine-learning methods to predict nitrate concentration in shallow groundwater of the Central Valley, California: boosted regression trees (BRT), artificial neural networks (ANN), and Bayesian networks (BN). Machine learning methods can learn complex patterns in the data but because of overfitting may not generalize well to new data. The statistical learning framework involves cross-validation (CV) training and testing data and a separate hold-out data set for model evaluation, with the goal of optimizing predictive performance by controlling for model overfit. The order of prediction performance according to both CV testing R^2 and that for the hold-out data set was BRT > BN > ANN. For each method we identified two models based on CV testing results: that with maximum testing R^2 and a version with R^2 within one standard error of the maximum (the 1SE model). The former yielded CV training R^2 values of 0.94–1.0. Cross-validation testing R^2 values indicate predictive performance, and these were 0.22–0.39 for the maximum R^2 models and 0.19–0.36 for the 1SE models. Evaluation with hold-out data suggested that the 1SE BRT and ANN models predicted better for an independent data set compared with the maximum R^2 versions, which is relevant to extrapolation by mapping. Scatterplots of predicted vs. observed hold-out data obtained for final models helped identify prediction bias, which was fairly pronounced for ANN and BN. Lastly, the models were compared with multiple linear regression (MLR) and a previous random forest regression (RFR) model. Whereas BRT results were comparable to RFR, MLR had low hold-out R^2 (0.07) and explained less than half the variation in the training data. Spatial patterns of predictions by the final, 1SE BRT model agreed reasonably well with previously observed patterns of nitrate occurrence in groundwater of the Central Valley.
Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

We evaluated three off-the-shelf machine learning methods for their ability to predict nitrate concentration in shallow groundwater of the Central Valley, California: boosted regression trees (BRT), artificial neural networks (ANN), and Bayesian networks (BN). We developed the models within a statistical learning framework (Hastie et al., 2009) to optimize predictive performance. The Central Valley is an intensive agricultural region and produces 8% of U.S. agricultural value on 1% of the U.S. farmland (Reilly et al., 2008) (Fig. 1). Decadal increases in groundwater nitrate concentrations have been observed in portions of the Central Valley, particularly in the eastern fans (shown as light green on the map), which typify younger, oxic conditions (Burow et al., 2013). Competition

for groundwater resources in the region calls into question whether the aquifer can remain a viable source of supply to drinking water wells (Faunt, 2009).

Suitability of groundwater for drinking depends both on quantity and quality. Statistical models are commonly used at large spatial scales to identify areas with high contamination potential and to understand factors that increase contamination risk. However, modeling groundwater contaminants derived mainly from the land surface is challenging because of numerous processes that influence solute transport and fate in soils and groundwater. Transport processes frequently are nonlinear and are complicated by the spatial variability of hydraulic and geochemical conditions in aquifers. Linear regression and classification methods have been popular choices for estimating nitrate impacts on groundwater (Ayotte et al., 2006; Boy-Roura et al., 2013; Frans, 2008; Gardner and Vogel, 2005; Gurdak and Qi, 2012; Huebsch et al., 2014; Jang and Chen, 2015; Ki et al., 2015; LaMotte and Greene, 2007; Liu et al.,
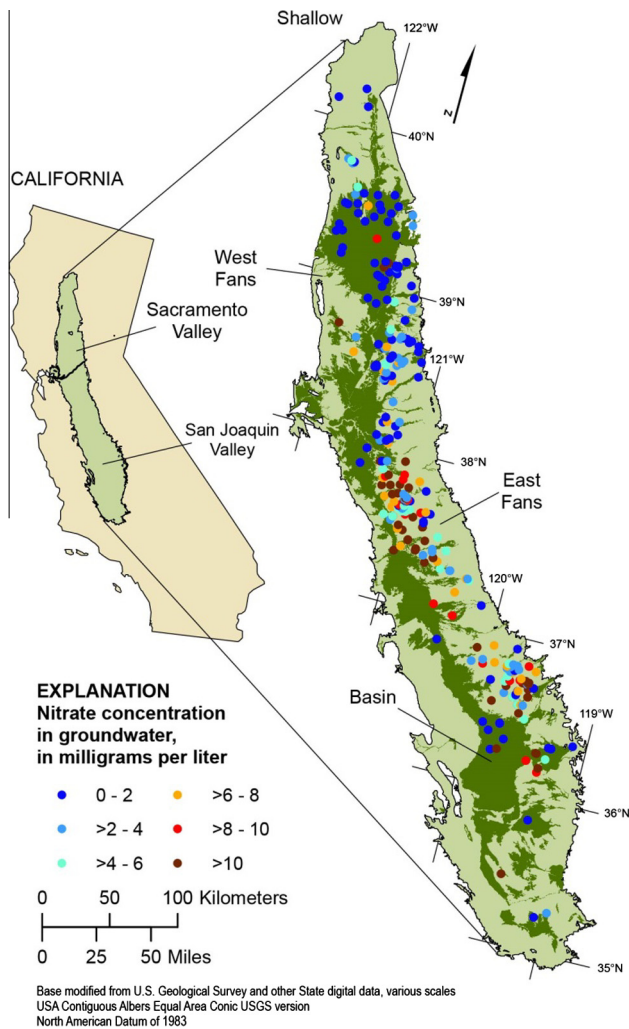
Fig. 1. Locations of shallow wells used to develop the models (modified from Nolan et al., 2014). The east and west fans are shown in light green and the basin subregion in dark green. Units of groundwater nitrate concentration are mg/L as N.

2005, 2013; Nolan et al., 2002; Rupert, 2003; Warner and Arnold, 2010). Although such methods are straightforward to apply at large spatial scales, hypothesis testing assumptions (linear and monotonic responses, assumed distributions of model residuals) are difficult to satisfy. For example, logistic regression assumes that the log odds ratio (logit) of observing some condition, such as exceeding a threshold nitrate concentration, is linearly related to a set of predictor variables.

Machine learning methods are promising alternatives that dispense with traditional hypothesis testing. For example, tree-based methods do not require data transformation, can fit nonlinear relations, and automatically incorporate interactions among predictor variables (Elith et al., 2008). Random forest regression (RFR), an ensemble tree method, was previously applied to shallow and deep wells of the Central Valley and yielded a pseudo R^2 of 0.90 for training data (Nolan et al., 2014). Random forest produces many classifiers (decision trees) and aggregates the predictions (Liaw and Wiener, 2002). The method employs bootstrap aggregating (bagging) to average the predictions over many trees, which reduces the variance of the prediction (Hastie et al., 2009). Random forest has only recently been applied to water resources data; other examples include nitrate and arsenic in aquifers of the southwestern U.S. (Anning et al., 2012), nitrate in an unconsolidated

aquifer in southern Spain (Rodriguez-Galiano et al., 2014), and nitrate in private wells in Iowa (Wheeler et al., 2015).

A perceived disadvantage of machine learning methods is their "black box" nature; without estimated coefficients it is difficult to show significant relations between the response and predictor variables. However, individual classification trees can be extracted from BRT models and are easy to interpret. BRT also yields variable importance rankings and partial dependence plots. The latter can be used to infer the direction and degree of influence of predictor variables, and can provide additional insight by revealing nonlinear and non-monotonic responses. Nolan et al. (2014) used partial dependence plots to show that increasingly negative, MODFLOW-simulated vertical water fluxes (i.e., increasing downward) were related to increasing RFR-predicted groundwater nitrate concentration, particularly for deep wells during the irrigation season (see Fig. S2 in the Supporting Information of Nolan et al., 2014). Use of MODFLOW outputs as predictor variables in the RFR models constituted a multi-model, hybrid modeling approach. Variables with a high importance ranking by RFR included the depths to the top and midpoint of a well's screened interval. The first depth was a useful proxy for travel time from the land surface to the well, and the latter was a proxy for the groundwater age distribution. Bayesian networks are directed acyclic graphs comprising nodes (output and predictor variables) and edges (correlated connections between nodes) (Fienen et al., 2013). The graphic depiction of a BN is quite interpretable because the user draws the connections between predictor and response variables.

In the present study we evaluated BRT, ANN, and BN using the same data set as Nolan et al. (2014). The objective was to compare the predictive performance of the methods in the context of statistical learning, described in more detail below. The three models were then compared with the RFR model of Nolan et al. (2014) and multiple linear regression (MLR).

2. Material and methods

2.1. Data set

The Central Valley data set comprised 318 shallow domestic wells, and another 119 wells lacking screened interval data were held out for model evaluation (Nolan et al., 2014). Groundwater nitrate concentration data are summarized in Table 1. In the present study, the modeled response variable was the natural log of groundwater nitrate concentration (mg/L NO_3^- as N) in sampled shallow wells (i.e., domestic wells with depth below water table ⩽46 m). The log transform reduced the influence of very high nitrate values (up to 74.7 mg/L) on model predictions. The 41 predictor variables represented soils, land use, groundwater age surrogates, and aquifer texture and MODFLOW-simulated vertical water fluxes from previous textural and numerical models of the Central Valley (Faunt, 2009) (Appendix A). All predictor variables were compiled within 500-m radius circular well buffers.

Table 1
Summary statistics of nitrate concentration in groundwater from shallow wells (from Nolan et al., 2014).

Variable	Nitrate concentration, mg/L as N
Minimum	<0.5
Maximum	74.7
Mean	6.38
Standard deviation	8.20
Median	3.61
Interquartile range	7.47
Number of observations	318

2.2. Machine-learning

Machine-learning methods evaluated here included BRT, ANN, and BN. BRT differs from RFR in that it weights (boosts) the contribution of each new tree while minimizing a loss function. A link function is specified, and the final model is a linear combination of all of the trees. The additive boosted model is defined as (Hastie et al., 2009)

$$f(\mathbf{x}) = \sum_{m=1}^{M} \beta_m b(\mathbf{x}; \gamma_m) \tag{1}$$

where β_m are expansion coefficients corresponding to each of the M boosting iterations; $\mathbf{x}$ is the set of predictor variables; γ parameterizes splitting variables and split levels at internal nodes, and predictions at terminal nodes; and b is a basis function that represents an individual tree. We used stochastic gradient boosting, which enhances the general form of boosting as follows. Estimation is stagewise such that β_m and γ_m are estimated sequentially from $m = 1$ to M. After an initial tree is trained, subsequent trees are fitted to the residuals of the previous tree rather than to the data directly. The loss function is minimized by driving each tree to focus on the worst performance of the previous tree, and the β values are the predictions at terminal nodes (De'ath, 2007). We used the squared error loss function, described as

$$L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2 \tag{2}$$

Gradient boosting modifies the above by adding steepest-descent minimization (Friedman, 2001). Stochastic gradient boosting adds randomness at each sequential step through bagging, which increases accuracy and computational efficiency and is robust against overfitting (Friedman, 2002).

Artificial neural networks are nonlinear regression models comprising an input layer, output layer and unobserved intermediate (hidden) layers defined as (Günther and Fritsch, 2010):

$$o(\mathbf{x}) = f\left(w_0 + \sum_{j=1}^{J} w_j \bullet f\left(w_{0j} + \sum_{i=1}^{n} w_{ij}x_i\right)\right) \tag{3}$$

where $o(\mathbf{x})$ is an output neuron, J is the number of nodes in the hidden layer, w_0 is the intercept of the output neuron, w_j is the weight corresponding to the jth hidden unit, w_{0j} is the intercept of the jth hidden unit, w_{ij} is the weight corresponding to the ith predictor variable feeding the jth hidden unit, and x_i is the ith predictor variable. The hidden units are linear combinations of the predictor variables $\mathbf{x}$, and these are transformed by the nonlinear activation function f. In the current work, f was the hyperbolic tangent sigmoid function. The unknown parameters w_0, w_{0j}, w_j, and w_{ij} were initially set to random values, then adjusted by a back-propagation algorithm to minimize the loss function, which is the mean square error (Limas et al., 2010):

$$MSE = \frac{R(\theta)}{n} \tag{4}$$

where n = the number of observations and $R(\theta)$, the sum of squared errors, is given by Hastie et al. (2009)

$$R(\theta) = \sum_{h=1}^{H} \sum_{l=1}^{L} (y_{lh} - o_{lh})^2 \tag{5}$$

where y is the observed value and the subscripts refer to the lth observation and the hth output node (Günther and Fritsch, 2010). In the current work the ANNs had a single output (groundwater nitrate concentration), so Eq. (5) reduced to a single summation over L observations.

The BN is a nonlinear classifier that yields a probability distribution for each class. We specified up to 10 nitrate concentration classes for the BN models, using cutpoints of 0.25, 1–6, 8, 13, 21, and 35 mg/L, which is tantamount to a semi-continuous response variable. The BN allows for dependencies among predictors through specification of joint probability distributions. The Bayesian approach involves estimating a posterior probability of a class (C_l). Using Bayes' theorem we estimate the posterior probability that an outcome is in a class based on the predictors that have been observed ($\mathbf{x}$), expressed as (Kuhn and Johnson, 2013)

$$Pr[y = C_l|\mathbf{x}] = \frac{Pr[\mathbf{x}|y = C_l]Pr[y]}{Pr[\mathbf{x}]} \tag{6}$$

where $Pr[y]$ is the prior probability of an outcome, $Pr[\mathbf{x}]$ is the probability of the predictor variables, and $Pr[\mathbf{x}|y = C_l]$ is the likelihood function, or the conditional probability of obtaining the predictor variables given the observed data for the lth class.

BNs grow dramatically both in computational expense and computer memory footprint with an increasing number of parent nodes. The parent nodes are all those directly connected to a response node. In this dataset with a single response (nitrate concentration) and 41 predictors, the BN rapidly became computationally impractical with increasing numbers of bins for each node. To mitigate this, and to take advantage of correlation among predictors of similar types, latent nodes were implemented. Latent nodes represent correlated combinations of several other nodes with values learned both from the correlations among their parent nodes and the correlation of the latent node with the output.

The BNs comprise nodes and their correlated connections (edges), and the correlations are aggregated to form conditional probability tables. Conditional probabilities were calculated using Bayes' theorem above, and the posterior predictions were updated based on conditioning to the observed data (Fienen and Plant, 2014).

2.3. Statistical learning

All three machine-learning methods can detect and simulate complex patterns in the data and as a result are prone to overfitting, which decreases predictive performance. Here we developed the models within a statistical learning framework to optimize prediction performance by controlling for overfit. Specifically, we used cross-validation (CV) to evaluate predictive performance for increasing levels of model complexity with the objective of minimizing "expected" prediction or test error. The expected test error includes randomness in CV training data sets (Hastie et al., 2009). We varied model complexity by changing values of the following metaparameters: tree interaction depth in the case of BRT; the number of hidden layer nodes in an ANN; and the number of bins corresponding to predictor and latent variables that compose a BN. We refer to the metaparameters as CV-tuning parameters and these are indicated for each method in Table 2. Before the cross validations, we made initial model runs on training data to determine reasonable values of variables other than the CV tuning parameters, such as learning rate in the case of BRT and ANN. During CV runs, only the tuning parameters were varied to isolate the effect of increasing model complexity. Following CV, we further evaluated selected models using hold-out data that were withheld from the CV process. Lastly, we compared the predictive performance of the CV-tuned models to the RFR model of Nolan et al. (2014) and MLR.

We used 10-fold CV to generate testing data sets; models were trained on 90% of the 318 observations and tested on 10% for each level of complexity. In this work, "testing" refers to 10% data subsets, "training" refers to the 90% data subsets and also to re-fitting of CV-tuned models to all 318 observations, and "evaluation" refers to application of final, CV-tuned models to the 119 hold-out wells.

Table 2
Model variables for boosted regression trees, artificial neural networks, and Bayesian networks.

Variable	Description	Value
Boosted regression trees		
interaction.depth	Tree depth, or number of layers in each tree (*cross-validation tuning parameter*)	1–16
n.trees	Total number of trees	800
shrinkage	Learning rate; determines the contribution of each new tree to the model	0.04
bag fraction	Proportion of data selected for each new tree	0.5
Artificial neural networks		
n.neurons	Number of neurons on hidden layers (*cross-validation tuning parameter*)	5–20
n.neurons	Number of neurons on input and output layers	41 (input), 1 (output)
learning.rate.global	Controls degree to which weights are adjusted based on change in partial derivative of error function	0.02
momentum.global	Adds fraction of previous weight change to current update to mitigate convergence on local minimum	0.05
hidden.layer	Function for nonlinear transformation of linear combination of predictor variables	tansig (hyperbolic tangent sigmoid)
error.criterium	Loss function	LMS (mean square error)
show.step × n.shows	Number of training epochs	10,000
Bayesian networks		
Number of input bins	Number of bins used to discretize input nodes in the BN (*cross-validation tuning parameter*)	2–6
Number of latent bins	Number of bins assigned to latent variables; each latent variable is evenly discretized from 0 to 100 with specific values representing correlations among parent nodes (*cross-validation tuning parameter*)	4–6
Number of output bins	Number of bins assigned to output nodes in the BN	4–10

"Estimation" refers to model simulated values for training samples, and "prediction" refers to simulated values for CV testing subsets and hold-out data. Sixteen configurations of each model were ordered from least to most complex, and we evaluated both maximum R^2 and one standard error (1SE) rule models. The 1SE models were the simplest configurations with CV testing R^2 within one standard error of the maximum testing R^2, where R^2 is the square of Pearson's correlation coefficient, or "model R^2." The latter statistic varies between 0 and 1, with higher values indicating better fit. Testing R^2 means of final models were compared using two-sided confidence intervals given by

$$\overline{R^2} \pm t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} \tag{7}$$

where $\overline{R^2}$ is the mean R^2 of the CV resamples, s is the standard deviation, $n = 10$, and $t_{0.975}$ at 9 degrees of freedom is 2.26 for a 95% confidence interval.

Final, CV-tuned models were refitted to all training data, applied to the hold-out wells, and evaluated for "generalization error" (Hastie et al., 2009), which indicates model performance with new data. In addition to MSE and R^2, we computed average bias and the variance of predictions. Average bias was computed as the difference between the sums of the predictions and hold-out observations divided by the number of hold-out observations (119). Hold-out wells lacked data on depths to the top and midpoint of the screened interval, therefore we used kriged estimates of these depths. As was mentioned above, these data are useful proxies for groundwater travel time and age distribution. Adding kriging uncertainty to these proxy variables likely decreased predictive performance for hold-out wells, but the same limitation applied to all three methods.

A map of predicted groundwater nitrate concentration was obtained by applying the final BRT model to gridded predictor data comprising over 50,000 cells that were 1 km² in size. Geographic Information System (GIS) data layers were made for each predictor variable in the data set and combined into an input file supplied to the BRT model object. Because depths to the top and midpoint of the perforated interval were not available at unsampled locations,

we used median values of these predictors (31 and 32 m, respectively) at all grid cells. We exponentiated the predictions and used smearing (Duan, 1983) to correct for bias during transformation back into original units of mg/L of nitrate.

2.4. Modeling software

We used the gbm package for BRT (Ridgeway, 2013) and the AMORE package for ANN (Limas et al., 2010) within R's computing environment (R, 2014). Cross validation was performed for BRT and ANN using R's crossval package (Strimmer, 2014), and for BN we used the Python module CVNetica (Fienen and Plant, 2014), which is a driver for Netica Bayesian network software (Norsys Software Corp., 2014). We performed MLR using the stepAIC function in R's MASS package, and the stepwise search was run forwards and backwards (Ripley, 2014). GIS processing of predictor variables and model predictions for mapping was performed using ArcGIS 10.2.2 for Desktop (ESRI, 2014).

3. Results and discussion

3.1. Cross validation tuning of models

Training R^2 by all three methods generally increased as model complexity increased (Fig. 2A), underscoring the flexibility and learning ability of machine-learning methods. Training R^2 was highest for BRT and was essentially 1 for model complexities greater than 6, followed by ANN and BN for which maximum training R^2 values (0.98 and 0.94) occurred at model complexities of 8 or more. The error bands in Fig. 2 are ±1SE of the mean R^2 of the 10 CV resamples, and are comparatively narrow for training data. However, training R^2 is a poor indicator of predictive performance by the models.

Testing R^2 values indicate predictive performance and were lower than training values for all three methods (Fig. 2B). Model fit to testing data plateaued or, in the case of ANN, decreased with increasing model complexity, indicating overfit. Overfit involves fitting noise in the training data, with the result that the models
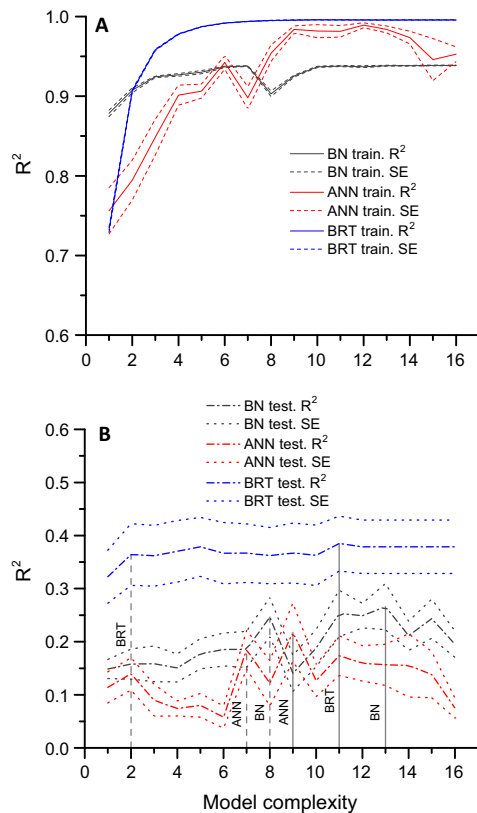
Fig. 2. Performance of boosted regression trees (BRT), artificial neural networks (ANN), and Bayesian networks (BN) for cross validation (A) training and (B) testing data subsets. Error bands are ± one standard error (1SE). Solid vertical lines in (B) show the maximum testing R^2 models, and dashed vertical lines show simpler models obtained by the 1SE rule.

do not generalize well to new data (Kuhn and Johnson, 2013). Similar degradation of predictive performance was seen for BN models of onshore ocean wave height and the percent of pumped groundwater derived from surface water (Fienen and Plant, 2014), and a BN model of mean depth to groundwater (Fienen et al., 2013). In the present study, ANN error bands indicated significant improvement in testing R^2 after model complexity 6, but performance degraded after complexity 11 (Fig. 2B). Model 9 (13 hidden nodes) had maximum ANN testing R^2 (0.22), and model 7 (11 hidden nodes) satisfied the 1 SE rule (testing $R^2 = 0.19$). Model 7 had a lower training R^2 (0.90) than model 9 (0.98), but performance with hold-out data improved (see next section).

Applying the 1SE rule to BN resulted in model 8, which had latent variables containing 4 bins, 4 bins on the predictor variables (which feed the latent variables), and 4 response bins (testing

$R^2 = 0.25$, training $R^2 = 0.90$) (Fig. 2B). The maximum-R^2 version (BN model 13) had 5 bins on the predictor variables, 6 bins on latent variables, and 10 response bins, and yielded training and testing R^2 values of 0.94 and 0.27, respectively.

In the case of BRT, the maximum testing R^2 was obtained with model 11 which had interaction depth = 11 (testing $R^2 = 0.39$, training $R^2 = 1.0$), and the 1SE rule yielded model 2 with interaction depth = 2 (testing $R^2 = 0.36$, training $R^2 = 0.91$) (Fig. 2B). Beyond interaction depth = 1, testing R^2 was comparable through the range of BRT model complexities.

3.2. Model evaluation

Following the cross validations, the maximum R^2 and 1SE models were retrained on all of the training data and applied to hold-out data to evaluate how well the models generalized to new data. All MSE values are in units of $(\ln(\text{mg/L NO}_3^- \text{ as N}))^2$, which in the case of ANN involved rescaling the model estimates from tangent sigmoid space to log space. When maximum testing R^2 was used as the model selection criterion, the order of models with hold-out data was BRT > BN > ANN ($R^2 = 0.01$–0.23, and MSE = 1.85–6.58) (Table 3). These results suggested that ANN was more susceptible to overfit than BRT and BN. Based on the 1SE rule, the order of the models by hold-out MSE was the same as the above (BRT > BN > ANN) (MSE = 1.75–3.08). However for BRT and ANN the lowest MSE and highest R^2 values for hold-out data were obtained with the simpler 1SE models. The superior predictive performance of the 1SE BRT model is consistent with the concept of using trees as weak learners in an additive model. Classification trees can be made into weak learners simply by limiting interaction depth, and they can be easily combined and can be generated quickly (Kuhn and Johnson, 2013). Limiting interaction depth reduces the number of tree nodes and the number of parameters γ in Eq. (1). An aggregate model comprising a number of simple trees is more accurate than a single complex tree model with many parameters, and BRT model 2 (interaction depth = 2) is much simpler than model 11.

Scatterplots of groundwater nitrate predictions versus hold-out observations provided insight into degree of model fit, model bias, and the variance of predictions reported in Table 3. The bias indicates the average difference between observed and predicted values, and the variance indicates the spread or degree of scatter of the predictions. Figs. 3–5 compare models selected according to 1SE and maximum-R^2 criteria. In the following discussion, units of variance are $(\ln(\text{mg/L NO}_3^- \text{ as N}))^2$, and bias units are $\ln(\text{mg/L NO}_3^- \text{ as N})$. BRT had smaller bias (0.03–0.06) than BN (0.23–0.30) and ANN (−0.53 to −0.25), and ANN had the largest prediction variance (2.36–3.55). Fig. 3A shows a moderate amount of scatter in points fitted by BRT model 2 to training data and somewhat more scatter in the hold-out predictions. The cloud of hold-out predictions is oriented along the 1:1 line, which is consistent with the

Table 3
Training and evaluation results for maximum R^2 and one-standard error (SE) cross-validation tuned models re-fitted to all training observations and applied to hold-out data. ANN model 9 with regularization had 1600 training epochs. Units of MSE (mean square error) and variance are $(\ln(\text{mg/L NO}_3^- \text{ as N}))^2$, and bias units are $\ln(\text{mg/L NO}_3^- \text{ as N})$.

| Model	Model selection rule (model no.)	Training ($n = 318$)		Evaluation with hold-out data ($n = 119$)			
		R^2	MSE	R^2	MSE	Bias	Variance
Boosted regression trees	Max. R^2 (11)	1.00	0.01	0.23	1.85	0.03	1.13
	One SE (2)	0.89	0.26	0.26	1.75	0.06	1.02
Artificial neural network	Max. R^2 (9)	0.97	0.06	0.01	6.58	−0.53	3.55
	One SE (7)	0.88	0.26	0.12	3.08	−0.25	2.36
	(9) with regularization	0.89	0.03	0.12	3.72	−0.18	3.33
Bayesian network	Max. R^2 (13)	0.98	0.05	0.18	1.93	0.23	0.26
	One SE (8)	0.94	0.15	0.03	2.39	0.30	0.34
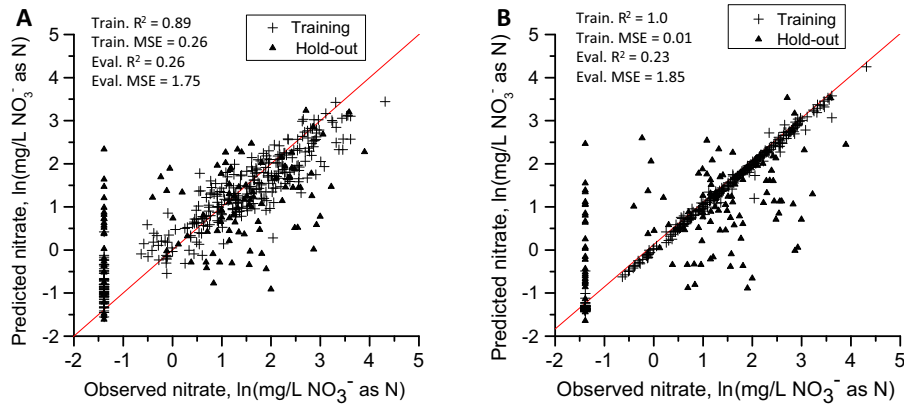
Fig. 3. Observed vs. predicted groundwater nitrate concentrations for boosted regression tree models re-fitted to all training data and evaluated using hold-out data: (A) one standard error rule (complexity = 2); (B) maximum R^2 rule (complexity = 11).
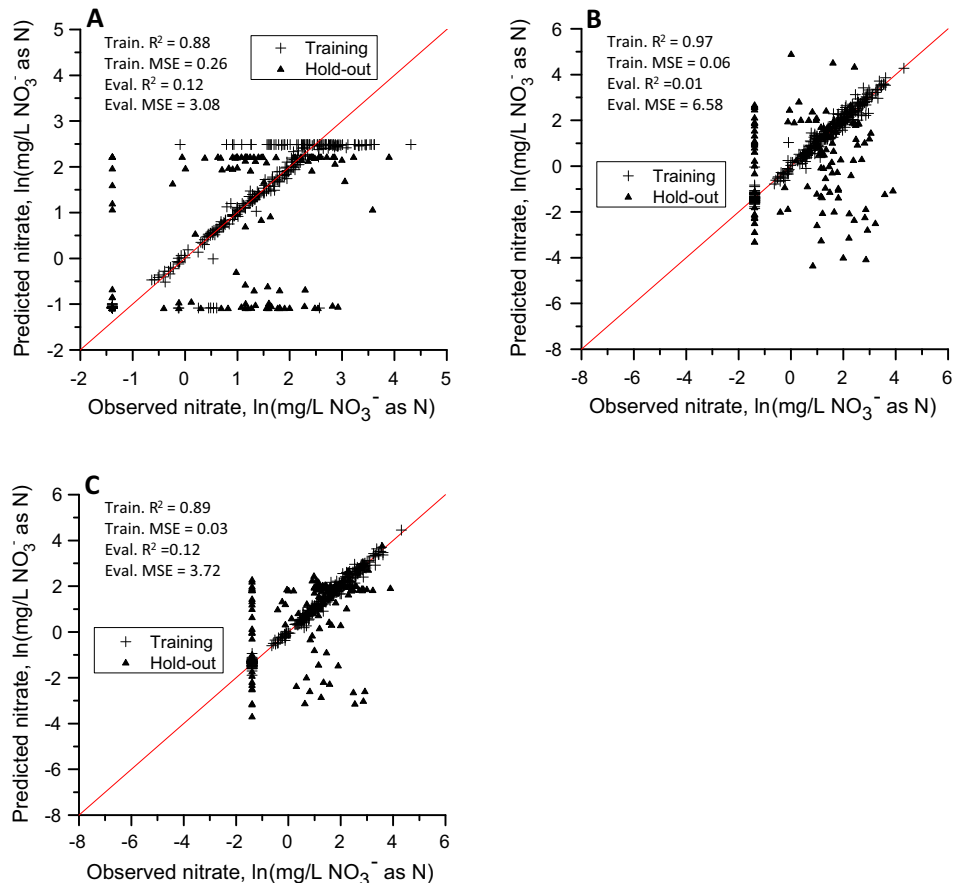

Fig. 4. Observed vs. predicted groundwater nitrate concentrations for artificial neural network models re-fitted to all training data and evaluated using hold-out data: (A) one standard error rule (complexity = 7); (B) maximum R^2 rule (complexity = 9); (C) model (9) with regularization.

low bias of this model (0.06) (Table 3). The lowest observed value (−1.4) evident in the figures corresponds to a common nitrate censoring level of 0.25 mg/L used by Burow et al. (2013) to accommodate multiple reporting levels in the data set.

The maximum-R^2 BRT model (11) had somewhat more scatter in the hold-out predictions compared with BRT model 2 (Fig. 3B), which is expressed as higher prediction variance (1.13) compared with model 2 (1.02) (Table 3). Training estimates by model 11 conformed more closely to the 1:1 line, even for censored NO_3^-

values, such that there was greater disparity between training and predictive model performance. These results illustrate a trade-off wherein the training performance of model 2 ($R^2 = 0.89$) was less than that of model 11 ($R^2 = 1.0$), but the predictive performance for hold-out data was improved ($R^2 = 0.26$ for model 2 vs. 0.23 for model 11).

The 1SE ANN model (7) showed poor fit to training data at the predicted data extremes, and the hold-out predictions do not follow the 1:1 line (Fig. 4A), which is consistent with the high bias
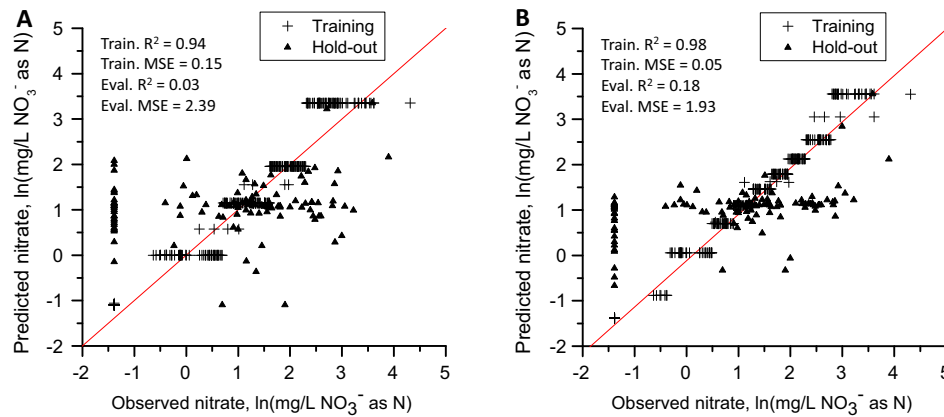
Fig. 5. Observed vs. predicted groundwater nitrate concentrations for Bayesian network models re-fitted to all training data and evaluated using hold-out data: (A) one standard error rule (complexity = 8); (B) maximum R^2 rule (complexity = 13).

(-0.25) (Table 3). The maximum-R^2 ANN model (9) fitted the training data well along the 1:1 line, but the prediction bias (-0.53) and variance (3.55) were the highest of any of the models and the latter was 1.6 × that of the observed data (2.28) (note the difference in vertical scale between Figs. 4A and B). The high bias is consistent with the low hold-out R^2 values by ANN (Table 3). Artificial neural networks with many weights typically overfit the data at minimum values of $R(\theta)$ (Hastie et al., 2009). We attempted to improve the performance of model 9 through a form of regularization that involved limiting the number of training epochs (i.e., early stopping). We focused on this model because the patterns of predictions appeared more reasonable compared with model 7's consistent overpredictions and underpredictions evident in Fig. 4A. We varied the number of training epochs from 500 to 10,000 and the maximum R^2 for hold-out data occurred with 1600 training epochs ($R^2 = 0.12$, MSE = 3.72) (Table 3). Bias (-0.18) was lower and prediction variance (3.33) was less than for model 9 without regularization. However, the version with regularization did not perform as well with hold-out data as did either BRT model or BN model 13.

Among off-the-shelf machine-learning methods, tree methods have several advantages over ANN, including ability to handle mixed data types and missing values, resistance to outliers, and ability to handle irrelevant inputs (Hastie et al., 2009). The data set has numerous predictor variables of different types, each with different scales of measurement (Appendix A). ANN is hampered by the presence of irrelevant predictor variables. Nolan et al. (2014) removed totally irrelevant variables from the data set when developing their RFR model, but the remaining variables varied in importance; and, each predictor variable in the data set requires a weight w_{ij} as shown in Eq. (3), which contributes to ANN model complexity.

Training estimates obtained by BNs generally followed the 1:1 line, although the estimates fell into clusters corresponding to the discrete nature of the output bins (Figs. 5A and B). For both the maximum-R^2 and 1SE models (nos. 13 and 8 in Table 3), the hold-out predictions fell into a narrow range of around 1 ln(mg/L NO_3^-), which contributed to considerable prediction bias (0.23–0.30). The narrow range of predictions resulted in low prediction variance (0.26–0.34), which was substantially less than that of the observed data.

The previously reported RFR model (Nolan et al., 2014) when applied to these same data performed comparably to BRT: hold-out $R^2 = 0.24$, hold-out MSE = 1.74, training $R^2 = 0.93$, and training MSE = 0.22. Nolan et al. (2014) did not use CV but instead screened different values of nodesize from 1 to 8 and monitored out-of-bag

(OOB) MSE. The number of samples in terminal nodes determines tree complexity, with decreasing numbers of nodes resulting in larger, more complex trees. In RFR, each tree is constructed with a subsample of the data, and observations not used are referred to as OOB and reserved for bootstrap estimates of model error. The best RFR predictive performance with OOB data was obtained for nodesize = 5.

We included MLR for context, which based on MSE predicted less well to hold-out data than BRT, BN, RFR, and ANN model 7, and calibrated less well than all of the other methods. For MLR we obtained hold-out $R^2 = 0.07$, hold-out MSE = 3.13, training $R^2 = 0.42$, and training MSE = 1.23. MLR explained less than half the nitrate variation in the training data, whereas BRT, BN, ANN, and RFR explained from 88% to 100%.

The three 1SE models were compared using two-sided confidence intervals on the mean CV testing R^2 values, which ranged from 0.19 to 0.36 (Fig. 6). The order of the models in terms of CV predictive performance was BRT > BN > ANN, which is consistent with hold-out results. The confidence intervals overlap, suggesting that CV predictive performance was not significantly different among the three models.
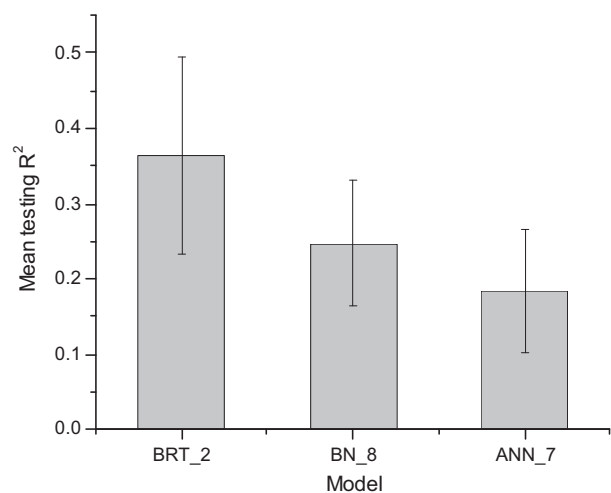

Fig. 6. Comparison of mean cross-validation resample R^2 by boosted regression tree (BRT), Bayesian network (BN), and artificial neural network (ANN) models obtained using the one standard error rule. The error bars are two-sided 95% confidence intervals on the mean testing R^2 values.

Considering all CV results, maximum testing R^2 was as high as 0.39 with an upper confidence interval bound of 0.50 (BRT model 11). The disparity between CV testing and hold-out results (highest $R^2 = 0.26$) may reflect the fact that the former data sets had measured depth to screened interval, whereas the latter relied on kriged estimates of this important variable. In general, prediction of nitrate in drinking water wells is challenging because they commonly are deeper than monitoring wells. For Central Valley wells sampled in the 1980s–2000s (Burow et al., 2013), the median depth of domestic wells was 55 m and median nitrate concentration was 2.1 mg/L. In contrast, monitoring wells had median depth of 11 m, and median nitrate was 5.0 mg/L. Deeper wells commonly have longer travel times, increased likelihood of nitrate-reducing conditions, and age mixtures that reflect recharge that occurred before the intensive use of synthetic N fertilizer (Dubrovsky et al., 2010). These factors may have reduced correlations between groundwater nitrate and land-use variables and limited the predictive ability of the models.

We selected BRT model 2 as final for mapping based on CV testing and model evaluation results with hold-out data. BRT yielded consistently higher testing R^2 across the range of model complexities (Fig. 2B), and model 2 had the highest R^2 and lowest MSE for hold-out data (Table 3). Also, the range of predictions by model 2 for hold-out data reasonably matched the spread of observed nitrate values (Fig. 3A). The resulting map (Fig. 7) closely resembles the groundwater nitrate map previously generated by random forest regression for the Central Valley (Nolan et al., 2014). Spatial patterns of predicted nitrate are consistent with what is known about nitrate occurrence in the region. The map shows a north–south gradient wherein predicted nitrate is generally higher in the San Joaquin Valley (south part of Central Valley) compared with the Sacramento Valley in the north. Prior researchers noted that Fe and Mn were consistently higher in the east fans and basin subregion of the Sacramento Valley compared with the San Joaquin Valley, which indicates reducing conditions less conducive to nitrate (Burow et al., 2013). Fig. 7 also shows low predicted nitrate concentration (0–2 mg/L) in the center of the Valley (dark green area in Fig. 1), which makes sense because groundwater becomes older and more reduced as it migrates toward the center of the basin.

4. Conclusions

A statistical learning approach helped control tendencies of machine-learning methods to overfit training data. Ordered by prediction R^2, performance by the best models was BRT > BN > ANN. Hold-out data augmented the CV approach in two ways. Hold-out data suggested that simpler BRT and ANN models obtained by the 1SE rule would outperform the more complex versions (i.e., those with maximum CV testing R^2) when applied to new data. Additionally, hold-out data provided insight into model prediction bias and variance. Although CV testing R^2 values by the three methods were not significantly different, scatterplots of observed vs. predicted hold-out data revealed considerable bias by ANN and BN and high prediction variance by ANN. ANN may have been hampered by the comparatively large number of predictor variables, which necessitates many weights and contributes to overfitting. The 1SE BRT model was selected for mapping and yielded mapped predictions that reasonably conformed to what is known about nitrate occurrence in shallow groundwater of the Central Valley. Although the hold-out data were useful in model evaluation, the results represent a single, unique data set and results with other independent data sets may differ.

Tree methods are flexible, are resistant to outliers, can handle irrelevant inputs, and seem well suited to the data set. Although none of the three methods adequately predicted censored nitrate data in hold-out data sets, the frequency of censored nitrate values in the training data was comparatively low (18%). However, random forest classification or probability-based BRT (Bernoulli link function) may be more appropriate with higher levels of censoring.

Predictive performance by the models may have been hampered by using screened interval depth as a proxy for groundwater age. Age is among the most important variables controlling groundwater nitrate concentration but is difficult to estimate. Future efforts are focused on improving estimates of groundwater age distributions through particle tracking with numerical groundwater flow models.

Of the off-the-shelf methods tested here, only BN estimates prediction uncertainty. Adding an uncertainty component such as Monte Carlo would benefit BRT, ANN, and RFR. Few studies have applied Monte Carlo to machine learning methods, but a Markov Chain Monte Carlo approach has been used with ANN (Hastie et al., 2009; Neal and Zhang, 2006). Additionally, Monte Carlo methods have been used with ensemble tree methods to estimate the variance of treatment effects (Austin, 2012).
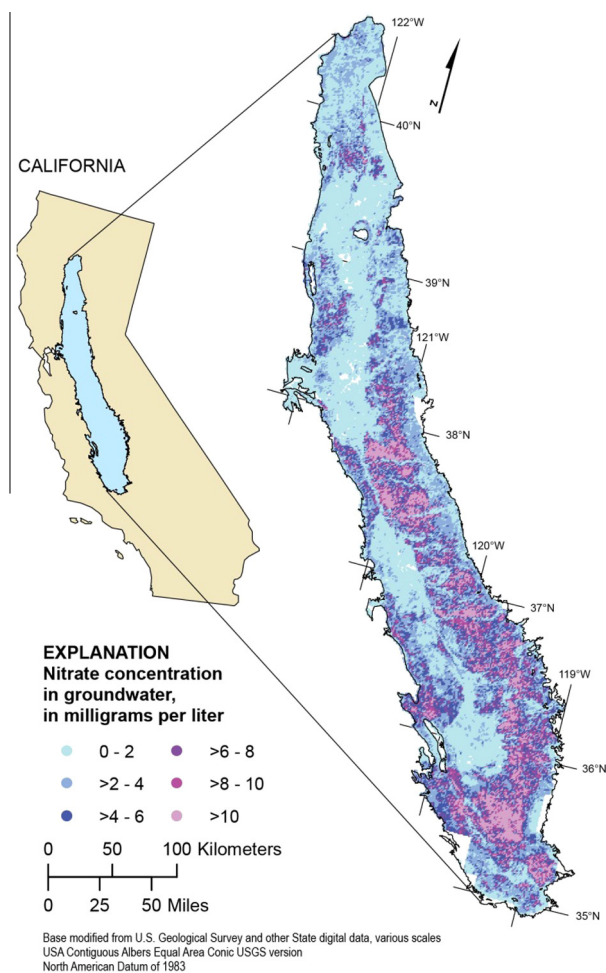
Fig. 7. Predicted groundwater nitrate concentration by boosted regression tree model 2 for the Central Valley, CA. Units of groundwater nitrate concentration are mg/L as N.

Appendix A

Predictor variables for the BRT, BN, and ANN models. CVHM, Central Valley Hydrologic Model, and CVTM, Central Valley Textural Model (Faunt, 2009); DWR, California Department of Water Resources (DWR, 2013); NWIS, National Water Information System (USGS, 2005); SSURGO, Soil Survey Geographic Database (USDA, 2014); STATSGO, State Soil Geographic Database (Wolock, 1997).

Variable	Description	Source
Nitrogen input and land use		
Fert_N_tot	Farm + nonfarm fertilizer N, kg/ha	County N data (Gronberg and Spahr, 2012); farm N was apportioned within well buffers by cropland-pasture and orchard-vineyard lands, and nonfarm N was apportioned by urban-residential land.
LU_crop_pas	Cropland-pasture (2-pasture, 5-idle, 14-grain/hay, 15-vineyard, 16-truck, 17-field, 20-rice)	DWR
LU_orch_vin	Orchard-vineyard (15-vineyard, 18-deciduous fruit and nut tree, 19-citrus)	DWR
LU_urb_res	Urban-residential (7-landscape/golf, 8-residential, 9-urban)	DWR
Pop_density	Population density, people/km^2 × 10	1990 census
Soil properties		
AWC	Available water capacity, fraction	SSURGO
Bulk_den	Bulk density, g/cm^3	SSURGO
Clay	Clay content, percent	SSURGO
Hi_WT_dep	Depth to saturated soil, m	STATSGO
Hydgrp_A	Percent of hydrologic group A	SSURGO
Hydgrp_B	Percent of hydrologic group B	SSURGO
Hydgrp_C	Percent of hydrologic group C	SSURGO
Hydgrp_D	Percent of hydrologic group D	SSURGO
Hydrat_A	Percent of drainage class well-drained	SSURGO
Hydrat_B	Percent of drainage class somewhat poorly drained	SSURGO
Hydrat_C	Percent of drainage class poorly drained	SSURGO
Hydrat_D	Percent of drainage class moderately well drained	SSURGO
Hydrat_E	Percent of drainage class excessively drained	SSURGO
Hydrat_G	Percent of drainage class somewhat excessively drained	SSURGO
Ksat_vert	Depth-integrated Ksat, μm/s	Calculated from SSURGO Ksat for *i* layers and total soil depth
Org_mat	Organic matter content (percent by weight)	SSURGO
Porosity	Porosity, percent	SSURGO
Sand	Sand content, percent	SSURGO
Silt	Silt content, percent	SSURGO
Well-construction data		
Perf_top	Depth to top of perforated interval, ft	NWIS
Scr_mid	Depth to midpoint of perforated interval, ft	NWIS
Central Valley model outputs		
PC_screen	Percent coarse sediment above the well screen	CVTM
PC_upper	Percent coarse sediment in upper active model layer	CVTM
Vel_Oct91	Vertical water flux[a], Oct. 1991, m^3/d	CVHM
Vel_Nov91	Vertical water flux[a], Nov. 1991, m^3/d	CVHM
Vel_Dec91	Vertical water flux[a], Dec. 1991, m^3/d	CVHM
Vel_Jan92	Vertical water flux[a], Jan. 1992, m^3/d	CVHM
Vel_Feb92	Vertical water flux[a], Feb. 1992, m^3/d	CVHM
Vel_Mar92	Vertical water flux[a], Mar. 1992, m^3/d	CVHM
Vel_Apr92	Vertical water flux[a], Apr. 1992, m^3/d	CVHM
Vel_May92	Vertical water flux[a], May 1992, m^3/d	CVHM
Vel_Jun92	Vertical water flux[a], June 1992, m^3/d	CVHM
Vel_Jul92	Vertical water flux[a], Jul. 1992, m^3/d	CVHM
Vel_Aug92	Vertical water flux[a], Aug. 1992, m^3/d	CVHM
Vel_Sep92	Vertical water flux[a], Sept. 1992, m^3/d	CVHM
Wat_lev92	Avg. simulated depth to water, 1992, m	CVHM

[a] Monthly average MODFLOW water flux across the bottom of the upper active model layer, where negative sign represents water moving vertically downward.

References

Anning, D.W., Paul, A.P., McKinney, T.S., Huntington, J.M., Bexfield, L.M., Thiros, S.A., 2012. Predicted Nitrate and Arsenic Concentrations in Basin-fill Aquifers of the Southwestern United States. U.S. Geological Survey Scientific Investigations Report 2012-5065.

Austin, P.C., 2012. Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. Multivar. Behav. Res. 47 (1), 115–135. http://dx.doi.org/10.1080/00273171.2012.640600.

Ayotte, J.D. et al., 2006. Modeling the probability of arsenic in groundwater in New England as a tool for exposure assessment. Environ. Sci. Technol. 40 (11), 3578–3585.

Boy-Roura, M., Nolan, B.T., Menció, A., Mas-Pla, J., 2013. Regression model for aquifer vulnerability assessment of nitrate pollution in the Osona region (NE Spain). J. Hydrol. 505, 150–162.

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.

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

Duan, N., 1983. Smearing estimate: a nonparametric retransformation method. J. Am. Stat. Assoc. 78, 605–610.

Dubrovsky, N.M. et al., 2010. The Quality of our Nation's Waters—Nutrients in the Nation's Streams and Groundwater, 1992–2004. U.S. Geological Survey Circular 1350.

DWR, 2013. Land and Water Use Data Collections. California Department of Water Resources. <http://www.water.ca.gov/landwateruse/> (accessed January 2011).

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77 (4), 802–813. http://dx.doi.org/10.1111/j.1365-2656.2008.01390.x.

ESRI, 2014. ArcGIS for Desktop. <http://www.esri.com/software/arcgis/arcgis-for-desktop/index.html> (accessed June 2014).

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

Fienen, M.N., Plant, N.G., 2014. A cross-validation package driving Netica with python. Environ. Model. Softw. 63, 14–23.

Fienen, M.N., Masterson, J.P., Plant, N.G., Gutierrez, B.T., Thieler, E.R., 2013. Bridging groundwater models and decision support with a Bayesian network. Water Resour. Res. 49 (10), 6459–6473.

Frans, L., 2008. Trends of pesticides and nitrate in ground water of the Central Columbia Plateau, Washington, 1993–2003. J. Environ. Qual. 37 (suppl. 5), S273–S280. http://dx.doi.org/10.2134/jeq2007.0491.

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 (4), 367–378. http://dx.doi.org/10.1016/S0167-9473(01)00065-2.

Gardner, K.K., Vogel, R.M., 2005. Predicting ground water nitrate concentration from land use. Ground Water 43 (3), 343–352. http://dx.doi.org/10.1111/j.1745-6584.2005.0031.x.

Gronberg, J.M., Spahr, N.E., 2012. County-level Estimates of Nitrogen and Phosphorus from Commercial Fertilizer for the Conterminous United States, 1987–2006. U.S. Geological Survey Scientific Investigations Report 2012-5207.

Günther, F., Fritsch, S., 2010. Neuralnet: training of neural networks. R J. 2 (1), 30–38.

Gurdak, J.J., Qi, S.L., 2012. Vulnerability of recently recharged groundwater in principle aquifers of the United States to nitrate contamination. Environ. Sci. Technol. 46 (11), 6004–6012.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, second ed. Springer, New York.

Huebsch, M. et al., 2014. Statistical analysis correlating changing agronomic practices with nitrate concentrations in a karst aquifer in Ireland. WIT Trans. Ecol. Environ. 182, 99–109. http://dx.doi.org/10.2495/WP140091.

Jang, C.S., Chen, S.K., 2015. Integrating indicator-based geostatistical estimation and aquifer vulnerability of nitrate-N for establishing groundwater protection zones. J. Hydrol. 523, 441–451. http://dx.doi.org/10.1016/j.jhydrol.2015.01.077.

Ki, M.G., Koh, D.C., Yoon, H., Kim, H.S., 2015. Temporal variability of nitrate concentration in groundwater affected by intensive agricultural activities in a rural area of Hongseong, South Korea. Environ. Earth Sci. 74 (7), 6147–6161. http://dx.doi.org/10.1007/s12665-015-4637-7.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling, first ed. Springer, New York.

LaMotte, A.E., Greene, E.A., 2007. Spatial analysis of land use and shallow groundwater vulnerability in the watershed adjacent to Assateague Island National Seashore, Maryland and Virginia, USA. Environ. Geol. 52 (7), 1413–1421. http://dx.doi.org/10.1007/s00254-006-0583-8.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.

Limas, M.C. et al., 2010. Package 'AMORE'. The R Project for Statistical Computing. <http://www.r-project.org/> (accessed January 2014).

Liu, A., Ming, J., Ankumah, R.O., 2005. Nitrate contamination in private wells in rural Alabama, United States. Sci. Total Environ. 346 (1–3), 112–120. http://dx.doi.org/10.1016/j.scitotenv.2004.11.019.

Liu, C.W., Wang, Y.B., Jang, C.S., 2013. Probability-based nitrate contamination map of groundwater in Kinmen. Environ. Monit. Assess. 185 (12), 10147–10156. http://dx.doi.org/10.1007/s10661-013-3319-8.

Neal, R., Zhang, J., 2006. High dimensional classification with Bayesian neural networks and Dirichlet diffusion trees. In: Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L. (Eds.), Feature Extraction: Foundations and Applications. Springer, Berlin-Heidelberg, pp. 265–296. http://dx.doi.org/10.1007/978-3-540-35488-8_11.

Nolan, B.T., Hitt, K.J., Ruddy, B.C., 2002. Probability of nitrate contamination of recently recharged groundwaters in the conterminous United States. Environ. Sci. Technol. 36 (10), 2138–2145.

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. http://dx.doi.org/10.1021/es405452q.

Norsys Software Corp., 2014. Netica, Version 5.12. <http://www.norsys.com/> (accessed January 2014).

R, 2014. The R Project for Statistical Computing. <http://www.r-project.org/> (accessed June 2014).

Reilly, T.E., Dennehy, K.F., Alley, W.M., Cunningham, W.L., 2008. Ground-Water Availability in the United States. U.S. Geological Survey Circular 1323, Reston.

Ridgeway, G., 2013. Package 'gbm', The R Project for Statistical Computing. <http://www.r-project.org/> (accessed January 2014).

Ripley, B., 2014. Package 'MASS'. The R Project for Statistical Computing. <http://www.r-project.org/> (accessed January 2014).

Rodriguez-Galiano, V., Mendes, M.P., Garcia-Soldado, M.J., Chica-Olmo, M., Ribeiro, L., 2014. Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (Southern Spain). Sci. Total Environ. 476–477, 189–206. http://dx.doi.org/10.1016/j.scitotenv.2014.01.001.

Rupert, M.G., 2003. Probability of Detecting Atrazine/Desethyl-atrazine and Elevated Concentrations of Nitrate in Ground Water in Colorado. U.S. Geological Survey Water-Resources Investigations Report 02-4269.

Strimmer, K., 2014. Package 'crossval'. The R Project for Statistical Computing. <http://www.r-project.org/> (accessed January 2014).

USDA, 2014. Soil Survey Geographic (SSURGO) Database. USDA Natural Resources Conservation Service. <http://sdmdataaccess.nrcs.usda.gov/> (accessed June 2011).

USGS, 2005. National Water Information System Web (NWISWeb). U.S. Geological Survey. <http://waterdata.usgs.gov/nwis/about> (accessed January 2011).

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. U.S. Geological Survey Scientific Investigations Report 2010-5100.

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. http://dx.doi.org/10.1016/j.scitotenv.2015.07.080.

Wolock, D.M., 1997. STATSGO Soil Characteristics for the Conterminous United States, U.S. Geological Survey Open-File Report 97-656. <http://water.usgs.gov/GIS/metadata/usgswrd/XML/muid.xml#Top> (accessed July 2012).