Using Boosted Regression Tree Models to Predict Salinity in Mississippi Embayment Aquifers, Central United States

Katherine J. Knierim, James A. Kingsbury, Connor J. Haugh, and Katherine M. Ransom

Research Impact Statement: Machine learning was used to create spatially continuous and three-dimensional aquifer-scale predictions of groundwater salinity, an important factor with respect to water availability.

ABSTRACT: High salinity limits groundwater use in parts of the Mississippi embayment. Machine learning was used to create spatially continuous and three-dimensional predictions of salinity across drinking-water aquifers in the embayment. Boosted regression tree (BRT) models, a type of machine learning, were used to predict specific conductance (SC) and chloride (Cl), and total dissolved solids (TDS) was calculated from a correlation with SC. Explanatory variables for BRT models included well location and construction, surficial variables (e.g., soils and land use), and variables extracted from a groundwater-flow model, including simulated groundwater ages. BRT model fits ($r^2$) were 0.74 (SC and Cl) and 0.62 (TDS). BRT models provided spatially continuous salinity predictions across surficial and deeper aquifers where discrete water-quality samples were missing. Uncertainty was smaller where salinity was lower, and models tended to underpredict in areas of highest salinity. Despite this, BRT models were able to capture areas of documented high salinity that exceed the TDS secondary maximum contaminant level for drinking water of 500 mg/L. Variables that served as surrogates for position along groundwater flowpaths were the most important predictors, indicating that much of the control on dissolved solids is related to rock-water interaction as residence time increases. BRT models additionally support hypotheses of both surficial and deep sources of salinity.

(KEYWORDS: machine learning; groundwater; water quality; salinity; boosted regression trees.)

INTRODUCTION

Groundwater is a vital resource in the Mississippi embayment physiographic region of the central United States (U.S.) (Figure 1) because it supports an $11.9 billion agricultural economy (Alhassan et al. 2019). Additionally, about 4.3 million people living in the region rely on groundwater for their drinking water (Dieter et al. 2017). Widespread declines in groundwater levels have occurred because of groundwater use, and concerns exist about the quality of available groundwater to support domestic and public supply, industrial uses, and agriculture (Clark et al. 2011; Waldron et al. 2011; Kingsbury et al. 2015). High salinity can limit the use of groundwater as a source of drinking water and irrigation. For example, total dissolved solids (TDS) and chloride (Cl) have secondary maximum contaminant levels (SMCLs) for drinking water of 500 and 250 mg/L, respectively.
Previous studies found that approximately 15% of drinking-water wells in the Mississippi embayment had TDS concentrations greater than the SMCL (DeSimone et al. 2015; Kingsbury et al. 2015). Cl concentrations >100 mg/L adversely affect crops such as rice (Hardke 2018), which is one of the major crops in the Mississippi embayment (Alhassan et al. 2019).

Groundwater salinity mapping has previously been accomplished through geostatistical modeling (kriging) of groundwater samples (Arslan 2012), conversion of airborne electromagnetic surveys of bulk resistivity to groundwater salinity (Delsman et al. 2018), and multivariate regression models of mapable predictor variables (such as land use and surficial geology) (Stanton et al. 2017). The probability of exceeding groundwater salinity thresholds was predicted across the continental U.S. as part of a brackish water assessment, but the model did not capture small-scale features of observed high salinity in the Mississippi embayment and was not intended for depths <152 m (Stanton et al. 2017). Coarse resolution (259 km²) TDS maps were produced for Mississippi embayment aquifers by calculating the median concentration in groundwater wells for each grid cell (Pettijohn 1988) and local-scale maps of Cl are available (Whitfield 1975; Kresse and Clark 2008). However, high-resolution and aquifer-wide predictions of salinity are not available for Mississippi embayment aquifers. Several studies have focused on understanding the sources of salinity in Mississippi embayment aquifers at the local or regional scale (Broom et al. 1984; Morris and Bush 1986; Huff and Bonck 1993; Kresse and Clark 2008; Paul et al. 2018), but aquifer-scale assessments of the factors affecting groundwater salinity are lacking.

Machine learning (ML) methods are increasingly being used to make predictions in hydrologic systems, such as simulating streamflow at ungaged streams (Worland et al. 2018, 2019), predicting hydraulic head in groundwater wells (Michael et al. 2005; Sahoo et al. 2017), and mapping water quality

FIGURE 1. Specific conductance (SC) in µS/cm and chloride (Cl) in mg/L in groundwater wells of the Mississippi River Valley alluvial (MRVA), Middle Claiborne (MCAQ), and Lower Claiborne (LCAQ) aquifers.

(U.S. Environmental Protection Agency 2020). Previous studies found that approximately 15% of drinking-water wells in the Mississippi embayment had TDS concentrations greater than the SMCL (DeSimone et al. 2015; Kingsbury et al. 2015). Cl concentrations >100 mg/L adversely affect crops such as rice (Hardke 2018), which is one of the major crops in the Mississippi embayment (Alhassan et al. 2019).

Groundwater salinity mapping has previously been accomplished through geostatistical modeling (kriging) of groundwater samples (Arslan 2012), conversion of airborne electromagnetic surveys of bulk resistivity to groundwater salinity (Delsman et al. 2018), and multivariate regression models of mapable predictor variables (such as land use and surficial geology) (Stanton et al. 2017). The probability of exceeding groundwater salinity thresholds was predicted across the continental U.S. as part of a brackish water assessment, but the model did not capture small-scale features of observed high salinity in the Mississippi embayment and was not intended for depths <152 m (Stanton et al. 2017). Coarse resolution (259 km²) TDS maps were produced for Mississippi embayment aquifers by calculating the median concentration in groundwater wells for each grid cell (Pettijohn 1988) and local-scale maps of Cl are available (Whitfield 1975; Kresse and Clark 2008). However, high-resolution and aquifer-wide predictions of salinity are not available for Mississippi embayment aquifers. Several studies have focused on understanding the sources of salinity in Mississippi embayment aquifers at the local or regional scale (Broom et al. 1984; Morris and Bush 1986; Huff and Bonck 1993; Kresse and Clark 2008; Paul et al. 2018), but aquifer-scale assessments of the factors affecting groundwater salinity are lacking.

Machine learning (ML) methods are increasingly being used to make predictions in hydrologic systems, such as simulating streamflow at ungaged streams (Worland et al. 2018, 2019), predicting hydraulic head in groundwater wells (Michael et al. 2005; Sahoo et al. 2017), and mapping water quality
(Tesoriero et al. 2017; Knoll et al. 2019; Koch et al. 2019). ML identifies patterns in complex datasets in an attempt to produce an accurate prediction of a response variable, such as groundwater salinity, by minimizing some error term (Kuhn and Johnson 2013). There is a balance between accurate prediction and interpretation of models (Kuhn and Johnson 2013), such that ML has been criticized as a “black box” that ignores physical theory (Molnar 2019). However, approaches for ML explanations (Biecek 2018), preprocessing techniques that decrease redundant explanatory variables (Kuhn and Johnson 2013; Sahoo et al. 2017), and postprocessing tools that explore variable influence (Greenwell 2017; Ransom et al. 2017; Fienen et al. 2018) can aid model interpretation, leading to insights about the hydrologic system from ML models. Tree-based ML models (such as random forest and regression trees) are particularly suited to environmental data because they are able to handle large and diverse sets of explanatory variables, missing data, and collinearity among predictors, and do not assume monotonic relations between predictors and response data (Elith et al. 2008; Hastie et al. 2009; Kuhn and Johnson 2013).

The U.S. Geological Survey (USGS) National Water Quality Assessment Project has compared more traditional statistical techniques (such as linear and logistic regression) with a variety of ML methods (such as boosted regression trees [BRT], random forest, artificial neural networks, and Bayesian networks) for predicting groundwater quality (Nolan et al. 2014, 2015; Ayotte et al. 2016). BRT, which are a type of ML model, have been shown to provide accurate predictive performance and efficacy for modeling groundwater quality across areas of the aquifer not sampled for water quality (Ransom et al. 2017; Rosecrans et al. 2017; Erickson et al. 2018; DeSimone et al. 2020). Individual regression trees partition data into groups based on similarity with respect to the response variable, but can be unstable and may not predict well to newly collected data (Kuhn and Johnson 2013). Ensemble models, such as BRT, create an ensemble (or additive) model through boosting by minimizing a loss function (such as root-mean-squared error, or RMSE) on residuals for a user-specified number of additional trees (Elith et al. 2008). To reduce overfitting, the model is constrained by learning rate (or shrinkage) so that only a fraction of the predicted value is added to the previous iteration (Kuhn and Johnson 2013).

The goal of this paper was to provide spatially continuous and three-dimensional predictions of salinity across Mississippi embayment aquifers. The process of selecting the final models in this study tried to balance accurate predictions of salinity from BRT models with methods that aid model interpretation to gain information about the physical processes driving groundwater salinity. Groundwater salinity predictions can be used to better estimate freshwater volumes in aquifers, more accurately interpret bulk resistivity measurements by quantifying variation in groundwater resistivity (i.e., conductance), and characterize areas of aquifers that exceed benchmark standards for water use. Additionally, understanding the drivers causing changes in salinity concentrations is an important aspect of ensuring the sustainability of groundwater resources for future use.

STUDY AREA

The Mississippi embayment includes two principal regional aquifer systems; the surficial Mississippi River Valley alluvial aquifer (MRVA), and the Mississippi embayment aquifer system, which includes Tertiary-age sediments (Figure 1). Based on the distribution of groundwater use for drinking water, this modeling effort is focused on the MRVA, Middle Claiborne aquifer (MCAQ), and Lower Claiborne aquifer (LCAQ). The MRVA is a highly productive aquifer used mostly for irrigation, although an estimated 69.8 million gallons of groundwater per day (Mgal/d) were withdrawn for public supply in 2000 (Maupin and Barber 2005; Clark et al. 2011). Public-supply groundwater use from the Mississippi embayment aquifer system was 576 Mgal/d in 2000 (Maupin and Barber 2005), much of which was withdrawn from the MCAQ. Although not as regionally important as the MRVA or MCAQ, domestic and public-supply wells withdraw groundwater from the LCAQ, especially on the margins of the Mississippi embayment physiographic region where the LCAQ is near the surface or crops out (Figure 1).

Hydrogeologic units within the embayment are composed of unconsolidated clastic sediments of gravel, sand, silt, and clay of marine and nonmarine origin deposited throughout the Cenozoic Era (Tertiary and Quaternary) during cyclical marine transgressions and regressions (Clark and Hart 2009). In general, sandy deposits constitute aquifer units and clay-rich deposits constitute confining units (Hosman and Weiss 1991; Hart et al. 2008). Groundwater in the MRVA is generally under unconfined conditions because of declining groundwater levels — in some areas over 30 m compared to predevelopment conditions (Clark et al. 2011) — below the base of a regional surficial confining clay layer (Kresse et al. 2014). The median thickness of the MRVA is 40 m, but the unit can be as much as 88 m thick (Hart et al. 2008). Groundwater in the MCAQ generally occurs under
confined conditions because of the regionally extensive middle Claiborne confining unit (MCCU), such that small amounts of pumping have caused large declines in water level (Hart et al. 2008; Clark et al. 2011). The MCAQ consists mostly of the Sparta Sand (Memphis Sand in northeastern Arkansas, western Tennessee, and southern Missouri), which is predominantly sand deposits with interbedded clay (Hosman and Weiss 1991; Clark et al. 2011). The MCAQ has a median thickness of about 244 m, but the unit can be as much as 576 m thick toward the central axis of the embayment along the Mississippi River (Hart et al. 2008). The LCAQ includes predominantly sand deposits with interbedded silt and clay and is primarily composed of the Carrizo Sand in Arkansas and Louisiana, the Meridian Sand in Mississippi, and Meridian Sand Member of the Tallahatta Formation in Alabama (Hart et al. 2008). Similar to the MCAQ, the LCAQ is confined throughout most of its extent because of overlying regional confining units, except along the margins of the Mississippi embayment where the LCAQ crops out (Figure 1). The aquifer is relatively thin, ranging from 15 to 59 m thick with a median thickness of 38 m (Hart et al. 2008). North of approximately the 35th parallel (near the border between Tennessee and Mississippi), the Lower Clairborne confining unit (LCCU) undergoes a facies change such that the Memphis Sand includes the MCAQ, LCCU, and LCAQ. Therefore, the LCAQ does not extend north of the 35th parallel (Hosman and Weiss 1991) and in this area the MCCU provides confinement of underlying aquifer units.

METHODOLOGY

Response Variables

Salinity is defined by the total dissolved ions in groundwater, and typically, TDS >1,000 mg/L is classified as brackish or slightly saline (Stanton et al. 2017). Specific conductance (SC) provides a measurement of the electrical conductivity resulting from the presence of dissolved ions, is often widely measured, and is closely correlated with TDS, but SC does not have a drinking-water standard. Although SC and TDS may be closely correlated with Cl in groundwater systems, other ions can drive SC and TDS values, such as in the Mississippi embayment where water types are dominantly sodium- or calcium-bicarbonate. Therefore, mapped predictions of SC, Cl, and TDS were all included in the assessment of groundwater salinity to capture salinity parameters that are widely sampled and low-cost (SC), have a drinking-water standard (TDS and Cl), and may have independent relationships to other measures of dissolved ions (Cl).

Water-quality data were compiled from several sources to maximize the number and spatial distribution of samples because ML techniques benefit from large datasets (Elith et al. 2008). Natural log of SC and Cl from groundwater wells were the modeled response variables in BRT models. BRT (and other tree-based ML models) are generally insensitive to data transformations of explanatory variables (Elith et al. 2008). However, the loss function (RMSE) for tuning the BRT model can be sensitive to outliers, so transforming the dependent (response) variable is an appropriate method and can produce more accurate models (Ransom et al. 2017; Nolan et al. 2018). Because SC and Cl included larger datasets than TDS, both were modeled directly using BRT. TDS was calculated from a correlation with SC such that TDS served as an independent check of model performance. The primary data source was the National Water Information System of the USGS (U.S. Geological Survey 2017), with additional data from state agencies that maintain Groundwater Ambient Monitoring Networks (GWAM) or the Safe Drinking Water Information System (SDWIS) maintained by the U.S. Environmental Protection Agency (2013). SC data did not include censored results; however, about 10% of the Cl results in the SDWIS dataset were censored or not detected above laboratory reporting limits of 2.5 or 10 mg/L. Concentrations for censored data below the method detection limit were imputed using a multiple-imputation routine (Lubin et al. 2004) using SAS® software because (1) there is still relevant information in censored values (Kuhn and Johnson 2013) and (2) removing low-level concentrations could introduce bias in models.

Data spanned from 1960 to 2018 and included samples from public-supply, domestic, irrigation, industrial, and monitoring wells. Initial models only included data from about 1990 to 2018, but parts of the aquifers with elevated Cl and SC were not adequately modeled because no recent (post-1990) samples were collected in these areas. In order to increase both the spatial coverage and range of constituent values for the training datasets, samples collected since 1960 were included. Because the most recent sample for each well was used (when there was more than one sample) and groundwater residence times are long (>100 years) throughout much of the study area (Kingsbury et al. 2017; Haugh, Knierim, and Kingsbury 2020), the potential influence of changes in water quality over time was considered small compared to the improvement in overall model accuracy.
Explanatory Variables

Explanatory variables for the ML models (Table S1) included attributes associated with well geometry (i.e., location and construction, including hydrogeology), surficial variables (e.g., soils and land use), and variables extracted from a MODFLOW groundwater-flow model for the Mississippi embayment (Haugh, Knierim, and Kingsbury 2020; Haugh et al. 2020a, b). Explanatory variables were acquired from a variety of data sources (Table S1) or constructed using well metadata and the hydrogeologic framework for aquifers of the Mississippi embayment (Hart et al. 2008). Distance to faults (Table S1) was calculated using the ArcMap™ Near tool (Esri 2016) and faults from the Clark and Hart (2009) MODFLOW groundwater-flow model and Zimmerman (1992) (Figure 1). Although there are many more faults in the Mississippi embayment than what was used in the BRT models, the two dataset sources were a starting point for assessing how faults affect salinity predictions because (1) substantial effort went into capturing faults for the MODFLOW groundwater-flow model (Clark and Hart 2009) and (2) Zimmerman (1992) represent regional zones of faulting associated with the formation of the Mississippi embayment. A flag indicating whether a well was located within the lowlands of the alluvial plain (coincident with the footprint of the MRVA) was created (inMRVA) to distinguish alluvial plain wells vs. wells located in upland areas associated with outcrops of Tertiary units (Figure 1). The inMRVA flag was created after initial modeling efforts found that explanatory variables serving as surrogates for landscape position (such as altitude) were important predictors, but the inMRVA flag allowed for more direct inspection and interpretation of whether physiographic boundaries were important predictors of groundwater salinity.

The availability of well-construction information varied by data source: aquifer, well depth, and screened-interval depths were often incomplete for some wells. Rasters representing groundwater-withdrawal zones for the MRVA, MCAQ, and LCAQ were used to assign well depths when other information was not available (Knierim et al. 2019). Wells were assigned to a regional hydrogeologic unit (either MRVA, MCAQ, or LCAQ) using local aquifer codes and by intersecting the bottom screened-interval altitude with the hydrogeologic framework (Hart et al. 2008). Approximately 10% of the wells had a different hydrogeologic-unit assignment from the framework vs. the source data aquifer designation. Many of these wells were either located on the borders of the study area or were wells assigned to the MRVA, but with depths exceeding the typical thickness of the MRVA (~88 m). Wells were assigned to a final hydrogeologic unit as follows: (1) if the screened-interval bottom was within ±30 m of the framework-assigned hydrogeologic unit, then the source data aquifer code was used, (2) if a well was assigned to a confining unit (MCCU or LCCU) using the hydrogeologic framework, then the source data aquifer code was used, or (3) if well depth was >61 m, but a well was assigned to the MRVA in the source data, then the framework-assigned hydrogeologic unit was used (based on published thickness of the MRVA and observed well depths of MRVA wells). Wells located north of approximately the 35th parallel (where the LCCU undergoes a facies change) and deeper than MODFLOW groundwater-flow model Layer 8 (Clark and Hart 2009) have local aquifer designations of the Memphis Sand and MCAQ, but were considered “LCAQ equivalent” because of being screened toward the bottom of the MCAQ aquifer and to maintain consistency with groundwater-flow model layers (Table 1).

Explanatory variables were attributed to wells using either a point, buffer, or grid extraction in Python (Python Software Foundation 2019). Multi-order hydrologic position (Belitz, Arnold, et al. 2019; Belitz, Moore, et al. 2019) was attributed to wells using a point extraction to maintain the high resolution of the source data. Other surficial explanatory variables — such as land use/land cover, recharge,
and soil characteristics (physical and chemical) — were attributed to wells with 500-m buffer using the zonepy package (Clark et al. 2019) to account for spatial variation and to represent a surrogate for contributing recharge area to the well (Johnson and Belitz 2009). MODFLOW-derived variables were attributed to wells by intersecting the bottom of the well with the corresponding groundwater-flow model layer grid cell (Haugh et al. 2020b). MODFLOW variables included variables from selected stress periods throughout the model simulation (from years 1898, 1930, 1950, 1970, 1990, and 2010), summary statistics for the entire model period (including range, median, mean, and standard deviation), and the slope for the period of major groundwater use (after about 1950) to represent change in a variable over time (Table S1). Groundwater age and flowpath lengths were derived from particle tracks (Haugh, Knierim, and Kingsbury 2020) using MODPATH6 (Pollock 2012) following methods from Starn and Belitz (2018). Some explanatory variables were missing at some of the wells: for the 45 explanatory variables with null values, the null values ranged from 0.3% to 61% of data with a median of 3% (see Knierim et al. 2020, for the full dataset). Null values were retained because BRT models can handle missing data and still make predictions (Elith et al. 2008).

Explanatory variable datasets with many related predictors — including MODFLOW groundwater-flow model variables, groundwater-age model variables, soil geochemistry, and physical soil variables — were preprocessed to remove variables with high linear correlation following methods from Kuhn and Johnson (2013). One benefit of BRT methods is that predictive performance is not weakened by correlation of explanatory variables but, in practice, some inspection of correlation among explanatory variables may aid model interpretation (Kuhn and Johnson 2013). For example, if two explanatory variables (A and B) are perfectly and linearly correlated ($r^2 = 1$), either explanatory variable may be selected by the model, but the selection will be random. Therefore, it would be incorrect to interpret that A is more influential on the model response variable than B (or vice versa). A preprocessing routine (Kuhn and Johnson 2013) was used to remove variables with linear correlation coefficients $\geq 0.8$ using the statistical programming software R (R Core Team 2019) and the findCorrelation function in the caret package (Kuhn et al. 2019). This preprocessing reduced the explanatory variable dataset from 227 to 132 variables (Table S1), which also helped decrease run times for model tuning.

**Machine Learning Modeling Routine**

BRT models require the selection of five hyperparameters that control model fit and complexity. The values for the hyperparameters were selected based on a 10-fold cross-validation (CV) tuning scheme following methods from Ransom et al. (2017) and using the R packages caret and gbm (Kuhn et al. 2019; Ridgeway 2019) with 80% of the response-variable data (training data; Table 2). A range of values for each hyperparameter was tested resulting in a total possible 1,260 combinations of hyperparameters: interaction depth (2–18 × 2), minimum observations per node (8 and 10), shrinkage or learning rate (0.002–0.014 × 0.002), and number of trees (500–5,000 × 500); bagging fraction was held at the recommended 0.5. During 10-fold CV tuning, approximately 10% of the training data was used as a testing data-set to measure model performance via RMSE. Hyperparameter combinations were checked to find the combination that produced a model with the lowest CV RMSE (referred to as “best” model). The remaining 20% of the response-variable data was withheld as “holdout” data to evaluate model performance and the average RMSE of all the folds from the “best” CV tuning model tends to approximate the holdout RMSE; therefore, holdout RMSE provided an independent check on model performance. Model tuning was completed on the USGS Yeti supercomputer (USGS Advanced Research Computing 2019).

| Response variable | n (training) | n (hold-out) | Model | no. vars | id | mo | sh | nt |
|-------------------|--------------|--------------|-------|----------|----|----|----|----|
| SC                | 3,498        | 873          | Final | 55       | 10 | 8  | 0.010 | 5,000 |
|                   |              |              | 1 SE  | 131      | 10 | 8  | 0.010 | 5,000 |
|                   |              |              | Best  | 131      | 16 | 10 | 0.010 | 3,500 |
| Cl                | 4,617        | 1,153        | Final | 44       | 14 | 8  | 0.006 | 5,000 |
|                   |              |              | 1 SE  | 131      | 14 | 8  | 0.006 | 5,000 |
|                   |              |              | Best  | 131      | 16 | 10 | 0.012 | 2,500 |

Notes: n, sample size; no. vars, number of explanatory variables; id, interaction depth; mo, minimum observations per node; sh, shrinkage (or learning rate); nt, number of trees.
The combination of hyperparameters resulting in the lowest CV RMSE during the hyperparameter tuning scheme may produce a complex model highly specific to the training data, which can lead to overfitting and poor predictive performance to new data (i.e., holdout data). Therefore, once the hyperparameter combination with the lowest CV RMSE (“best” model) was identified, a 1 SE routine following Nolan et al. (2015) was used to inspect simpler models with reasonable model fit based on RMSEs being within 1 SE of the best model RMSE (referred to as “candidate” models, Figure 2a). During the 1 SE routine, RMSE and $r^2$ values of holdout data were calculated for candidate BRT models (Figure 2b); 243 and 235 models for SC and Cl, respectively, of 1,260 possible tuning models were within 1 SE of the “best” model. Candidate models were sorted by complexity, which was defined by model hyperparameters (in order) interaction depth, minimum observations per node, shrinkage, and number of trees; lower values of interaction depth, shrinkage, and number of trees and higher values of minimum observations per node produce simpler models. All the candidate models identified during the 1 SE routine (Figure 2b) could be chosen as reasonable models to predict water quality. Generally, the best model (lowest training CV RMSE), simplest model (i.e., least complex based on hyperparameters), and one to three other models (simpler and with similar RMSE to the best model) were inspected in detail (Figure 2b). Selecting the “final” model from the 1 SE routine was somewhat subjective, but generally models similar to “model a” or “model b” in Figure 2b were selected because they balanced model simplicity with similar (or lower) holdout RMSE compared to the “best” model. Because all candidate models are within 1 SE of the lowest training RMSE, variation in model predictive performance among candidate models was small.

Following the 1 SE routine, methods from Ransom et al. (2017) were used to reduce the number of explanatory variables in the final model without substantially decreasing predictive performance (Figure 2c). In general, the 1 SE and variable-reduction routines strive to balance predictive performance with finding simpler models, which tend to predict more accurately to new data. The variable-reduction routine removes explanatory variables one-at-a-time starting with the least influential variables and continuing until the model was trained with only one explanatory variable (Figure 2c). The change in RMSE and $r^2$ between model runs was used to make a final selection of explanatory variables. Similar to the 1 SE routine, final model selection is at the discretion of the modeler, and generally explanatory variables were chosen when RMSE and $r^2$ showed little variation (Figure 2c).

**Prediction Mapping**

Once the final model hyperparameters and explanatory variables were selected, water-quality predictions across aquifers were created using gbm predict in R (R Core Team 2019; Ridgeway 2019) and
a text file of values extracted from explanatory variable rasters. The gbm predict package requires that explanatory variable rasters have the same resolution and extent, so source rasters were resampled using Python and the gdal resampling method of bilinear interpolation (for most variables) or nearest neighbor (where interpolation computed values outside of the range of published values; GDAL and OGR Contributors 2019). Water-quality predictions from gbm were mapped to raster cells of the 1-km resolution national hydrologic grid (Clark et al. 2018) using Python. The depth of each raster cell corresponds to the midpoint of each MODFLOW groundwater flow model layer from Clark and Hart (2009) for each hydrogeologic unit (Table 1) and was masked to the appropriate aquifer extent (i.e., active cells in the groundwater flow model). Although units of the LCAQ do not extend into the northern part of the Mississippi embayment, predictions were made following MODFLOW groundwater-flow model layers, such that layers deeper than model Layer 8 (Clark and Hart 2009) were considered “LCAQ equivalent” (Table 1). This produced a continuous surface from north to south across the study area for MCAQ and LCAQ units. Final prediction maps included seven layers (one MRVA, four MCAQ, and two LCAQ) for each water-quality constituent (Table 1) because well-construction and MODFLOW groundwater-flow model explanatory variables changed with depth and among model layers (Knierim et al. 2020). Predictions were back transformed from log space using a bias correction factor (Nolan et al. 2015). TDS maps were created using the correlation between SC and TDS (Figure S1).

To quantify uncertainty of water-quality predictions, BRT prediction intervals (PIs) at an alpha of 0.1 were calculated for SC and Cl with a bootstrapping routine following methods in Ransom et al. (2017). For the bootstrapping, final model hyperparameters were held constant and 199 models were trained on bootstrapped replicates (created with replacement) of the training data. Each bootstrapped training dataset will create a slightly different model due to variation in the dataset as tree structure (such as split variables, split levels at nodes, and prediction at terminal nodes) varies with each of the 199 bootstrap replicates. Each of the models created with the bootstrapped datasets were used to make a spatial prediction of SC and Cl, resulting in a range of predictions for each water-quality parameter at each raster grid cell. At each raster grid cell, the upper (90th percentile) and lower (10th percentile) PIs were calculated from the quantile of the distribution of model error component, which was based on randomly sampled training residuals (Schwarz et al. 2006; Ransom et al. 2017). For SC and Cl, PI width was calculated by subtracting the upper and lower PIs. Upper and lower PIs for TDS were calculated from SC PIs using the same relation between SC and TDS.

RESULTS

Water-Quality Data

Natural log SC ($n = 4,371$) ranged from 2.8 to 9.5 ($17–13,500 \text{ uS/cm}$) and the median concentration in the MRVA (6.4) was higher than the MCAQ (5.8) and LCAQ (5.5) (Figure 3). Natural log Cl ($n = 5,770$) ranged from $–2.8$ (an imputed value for Cl < 2.5 mg/L) to 8.4 (0.06–4,400 mg/L) and similarly the median concentration in the MRVA (3.0) was higher than the MCAQ (2.2) and LCAQ (1.4) (Figure 3). Approximately 6% of Cl samples were above the SMCL for drinking water of 250 mg/L. Median and interquartile ranges between training and holdout data for SC and Cl were nearly identical. Natural log TDS ($n = 2,866$) ranged from $–1.0$ to 9.0 (0.37–8,100 mg/L) with a median of 5.5 (239 mg/L), and approximately 17% of samples was above the TDS SMCL of 500 mg/L. TDS was closely correlated with SC (Figure S1). In the MRVA, SC and Cl were highest in the southwestern area of the aquifer (Figure 1). Generally, for the MCAQ and LCAQ, SC and Cl were lower at the edges of the study area and higher toward the central and southern parts of study area (Figure 1).

BRT Models

BRT models had accurate performance for SC and Cl, where training $r^2$ values ranged from 0.93 to 0.98 and holdout $r^2$ values ranged from 0.73 to 0.75 (Figure 4). There was no loss to accuracy (based on holdout RMSE) throughout the hyperparameter tuning, 1 SE, and variable-reduction routines (Figure 4). Average testing RMSE from CV tuning was 0.45 for SC and 0.82 for Cl. The bias correction factor was 1.02 for SC and 1.19 for Cl. Final BRT model RMSE for holdout data was 0.45 for SC (or approximately 491 μS/cm from back-transformed values) and 0.81 for Cl (or approximately 98 mg/L from back-transformed values).

Final models for SC and Cl included 55 and 44 explanatory variables, respectively, and hyperparameters for BRT models varied for SC and Cl (Table 2). Although final explanatory variables between models differed (Table S1), some variables describing well position within the aquifer (e.g., inMRVA and DepBot), groundwater-flow model variables (e.g., hdt1930
and KH2), and groundwater age variables (e.g., age-Min, age.90, and pathMax) were common to both models (Figure 5). The relative influence of explanatory variables (Figure S2) was calculated in gbm as the reduction in squared error that can be attributed to each variable (Ridgeway 2019).

**Water-Quality Predictions**

SC, TDS, and Cl predictions are presented for the bottom of the MRVA, the uppermost MCAQ (model Layer 5), and the lowermost LCAQ (model Layer 10) in Figures 6–8, respectively, and the remaining prediction layers (MCAQ model Layers 6, 7, and 8 and LCAQ model Layer 9) are available in Knierim et al. (2020). The depth of prediction for each layer varies based on the midpoint altitude of the groundwater flow model layer (Hart et al. 2008) and may represent near-surface conditions at the water table or depths up to approximately 1,024 m (Knierim et al. 2020) (Table 1). Salinity varied in the MRVA, but was generally higher in the southwestern extent of the aquifer where SC, TDS, and Cl were higher (Figures 6–8, respectively). In the MCAQ and LCAQ, salinity was higher toward the center axis of the embayment where hydrogeologic units are deeper (near the Mississippi River) and toward the southern extent of the aquifers. Based on the correlation between SC and TDS, TDS was above the SMCL for drinking water of 500 mg/L in 11% of the MRVA footprint and between 3% and 19% of the footprint for layers representing MCAQ or LCAQ. The area of aquifer footprint above the TDS SMCL varied based on upper and lower PIs, which provides a measure of uncertainty for model predictions (Figure 7). For the MRVA, lower PI modeled 3% of the aquifer area above the TDS SMCL compared to the upper PI of 40%. For MCAQ and LCAQ, lower PIs ranged from 1% to 7% and upper PIs ranged from 10% to 30% for the area above the TDS SMCL. A maximum of 2.7% of the modeled aquifer area was above the brackish groundwater threshold of 1,000 mg/L TDS for all modeled layers. Between 0% and 3% of the aquifer footprints were above the Cl SMCL of 250 mg/L (Figure 8).
Because BRT models are not restricted to modeling linear relations, explanatory variable influence on a response variable can be complex. Partial dependence plots show the average influence of a change in an explanatory variable of interest on the response variable (y-axis) if all other variables in the model are

FIGURE 4. Performance of BRT models for SC and CI for (a) the “best” model from the CV hyperparameter tuning routine, (b) the selected model from the 1 SE routine, and (c) the final model from the variable-reduction routine. Gray dashed lines are the 1:1 lines.
held at their observed/measured values, and are provided for selected explanatory variables (Figures S3–S7). Scatter plots are presented with partial dependence plots for additional reference of data distribution and range by hydrogeologic unit (Figures S3–S7).

DISCUSSION

Important Explanatory Variables

Salinity in aquifers of the Mississippi embayment was largely driven by relative position within the groundwater flow system. Although SC and Cl differed among the MRVA, MCAQ, and LCAQ (Figure 3), hydrogeologic unit (Layer2) was not an influential variable, and position within or outside of the MRVA footprint (inMRVA) was the most important predictor for SC and the 14th most important for Cl (Figure 5). Wells within the alluvial plain (inMRVA = 1) had higher SC and Cl than wells in the upland areas outside of the alluvial plain (inMRVA = 0; Figure 1); therefore, inMRVA captured both hydrogeologic unit (all MRVA wells have the same inMRVA value of 1) and groundwater flowpath for the MCAQ/LCAQ because an inMRVA value of 0 includes the outcrop areas that receive much of the recharge to these aquifers. Similarly, groundwater altitude in 1930 (hdt1930) was an important predictor of salinity because groundwater altitude prior to the onset of large amounts of pumping (beginning in about 1950) reflects land-surface altitude, with higher heads in the MRVA toward the northern portion of the aquifer and in the MCAQ/LCAQ in the recharge area of the uplands. SC and Cl values were higher where hdt1930 was lower (Figure S3), generally corresponding to the southern extent of the study area. Multi-order hydrologic position (MOHP) describes position within the watershed and can be used to predict depth to water (Belitz, Arnold, et al. 2019; Belitz, Moore, et al. 2019). In the Mississippi embayment, MOHP variables for ninth order streams (mohp9DS and mohp9LP) represent the location of a well relative to the Arkansas and Mississippi Rivers, which serve as major groundwater flow boundaries, and were important explanatory variables of SC and Cl (Figure 5; Table S1). Explanatory variables that were surrogates for position along the groundwater flowpath, such as inMRVA, hdt1930, DepBOT, dtwt1930, mohp9DS, and mohp9LP, generally reflect lower SC and Cl where fresh and more recently recharged groundwater occurs in upland areas and dissolved constituent concentrations increase along regional groundwater flowpaths. Location-based explanatory variables were also important water-quality predictors in ML models of other unconsolidated sediment aquifers and were similarly attributed to serving as surrogates for relative position along the flowpath or

FIGURE 5. Relative influence of the top 20 explanatory variables in the (a) SC and (b) Cl models.
capturing other spatially variable geologic attributes, such as sediment texture or transmissivity (Rosecrans et al. 2017; Erickson et al. 2018).

Groundwater-age parameters derived from a MODFLOW groundwater-flow model (Haugh et al. 2020a, b) were also important predictors of salinity and further reflect position along groundwater flowpaths. For example, as maximum groundwater flowpath length (pathMax) increased, SC and (to a lesser extent) Cl generally increased (Figure S4). Groundwater age (ageMin and age.90) showed a similar relation; higher salinity generally occurred with older groundwater for the MCAQ and the LCAQ, although the influence was minimal on very old groundwater (greater than approximately 250,000 years and representing a small portion of the wells; Figure S5). Previous research found that MCAQ groundwater transitioned from calcium-bicarbonate to sodium-bicarbonate water with concomitantly higher TDS as groundwater residence time increased (Pettijohn 1988; Kresse et al. 2014). This pattern is observed in the salinity maps for the MCAQ and LCAQ (Figures 6 and 8) because predicted SC and Cl concentrations were lower along the edges of the study area.

FIGURE 6. Mapped predictions and prediction interval (PI) width of SC for the MRVA, top of the MCAQ (Layer 5), and bottom of the LCAQ (Layer 10). Inactive cells from Haugh et al. (2020b).
FIGURE 7. TDS for the MRVA, top of the MCAQ (Layer 5), and bottom of the LCAQ (Layer 10) based on upper PIs, predicted values, and lower PIs. Inactive cells from Haugh et al. (2020b).
near the outcrop areas (and, subsequently, recharge zones for MCAQ and LCAQ layers) and higher toward the axis and southern extent of the study area. In particular, the southern extent of the study area has highly saline groundwater at depth that precluded groundwater-flow modeling for Tertiary units in Clark and Hart (2009). The effects of this wedge of saline water can be observed in the prediction maps for the MCAQ and LCAQ where Cl is highest near the southern boundary of the active model cells (Figure 8); concentrations in the inactive cells are generally greater than 10,000 mg/L TDS (not shown in Figure 7) owing to dissolution of halite in salt domes (Pettijohn 1996).

SC model predictions were more accurate (higher $r^2$ and lower RMSE) than for Cl. This difference may be largely a result of SC measuring all dissolved ions in groundwater, and thus representing any rock-water interaction process that contributes ions to the dissolved phase. Although Cl is an important driver of SC and TDS in Mississippi embayment aquifers, other dissolved ions, particularly bicarbonate, contribute to SC. Additionally, in areas where shallow groundwater is affected by anthropogenic sources of...
Cl, relatively young, low ionic strength water with elevated Cl concentrations can move into aquifers (Kingsbury et al. 2017). The difference in the partial dependence plots for ageMin between SC and Cl (Figure S5), with some high Cl concentrations where young water is present, but SC is low, likely reflects the anthropogenic contribution of Cl.

Areas of higher salinity in the MRVA have been hypothesized to be controlled by (1) deeper sources of salinity (whether saline groundwater from dissolution of salt domes or connate seawater) moving upward from underlying aquifers (e.g., along faults, wells, or where confining units are thin or absent) or (2) accumulation of salts during evaporation when recharge is slowed by surficial clays (Pettijohn 1996; Kresse and Clark 2008; Paul et al. 2018). Because of the physiography of the alluvial plain footprint, many surficial explanatory variables (especially soils and land use) strongly reflect that footprint. This elucidates why inMRVA was an important explanatory variable for prediction of salinity concentrations: inMRVA captured the effects of both surficial and subsurface variables. Other surficial explanatory variables may further explain some of the heterogeneity in MRVA salinity. None of the soil geomorphology variables from Saucier (1994) remained in the SC or Cl models after the variable-reduction routine, which contrasts with conclusions from Kresse and Clark (2008) that differences in geomorphic units (and associated soil properties) controlled Cl concentrations in the MRVA in areas not influenced by deeper salinity sources. However, both SC and Cl were lower where recharge (rech_WB_RC0013) was higher, indicating that greater recharge (from precipitation) with lower TDS flushes freshwater through the system. Additionally, where soil clay amount (slGC_C_Tot_10A_Clay) was greater, SC and Cl were also greater, suggesting that finer grain textures may slow recharge and concentrate surficial salts during evapotranspiration. Although large geomorphic regions were not important predictors of salinity, soil texture and recharge variables showed relations with SC and Cl that support the hypothesis of surficial controls on salinity. Additionally, soil geomorphology categories (Saucier 1994; Wacaster et al. 2018) were treated as 20 separate explanatory variables (Table S1), which may dilute the effect of geomorphology in BRT models of SC and Cl. It was also difficult to separate the effect of these surficial explanatory variables on the MRVA compared to the influence on MCAQ and LCAQ units and, as stated, many of these properties also vary between the alluvial plain and uplands. It may be beneficial in future modeling efforts to separately model the MRVA to aid explanatory variable interpretation. Overall though, some aspect of surficial control on salinity (and more so on Cl than SC) was observed in the important explanatory variables identified from BRT models (Figure 5).

Deep sources of saline water have been hypothesized to affect both the MRVA and Tertiary units; in some areas the deeper saline water may migrate upward because high groundwater heads at depth induce upward flow either through thin or absent confining units or along regional faults (Broom et al. 1984; Morris and Bush 1986; Huff and Bonck 1993; Kresse and Clark 2008; Paul et al. 2018). Distance to faults from the Clark and Hart (2009) MODFLOW groundwater-flow model (fltDist_M) and from Zimmerman (1992) (fltDist_Z) were important predictors of both SC and Cl (Figure 5; Table S1), although the influence of fault distance was opposite for SC vs. Cl (Figure S6). Cl increased with greater values of fltDist_M (i.e., greater distance), which occur on the eastern and western margins of the southern extent of the study area (Figure 1). SC decreased at fltDist_M distances greater than approximately 300 km (Figure S6), which was relatively coincident with the extent of freshwater in the uplands of the Mississippi embayment (Figure 1). SC was greater for lower values of fltDist_Z (i.e., closer proximity), but Cl increased with distances greater than approximately 100 km (Figure S6), coincident with areas where Cl is elevated in the MRVA along the northwestern edge of the study area (Figure 8). The distance from faults variables may serve as surrogates for positional variables, such as distance to adjacent aquifer systems or relative distance to the southern extent of the study area. Additionally, explanatory variables associated with faulting do not consider the type of fault, associated movement or offset along faults, and three-dimensional planes or distance between wells and faults. As such, it is difficult to ascertain the extent to which faults control salinity predictions in the BRT models, despite faults having been shown to be important controls for Cl concentration in the MRVA in other research (Paul et al. 2018).

Possible groundwater flow through confining units was represented in the BRT models as confining unit thickness and leakage through the MCCU (the major confining unit separating deeper Tertiary units from the MRVA) extracted from the groundwater-flow model (e.g., lk4t1970). Long maximum flowpaths (pathMax) in the MRVA (Figure S4) indicate that water moves from the underlying aquifers and into the surficial MRVA. Explanatory variables representing leakage through the MCCU (lk4t1898 and lk4slp) were important predictors of SC and Cl (Figure 5), and higher SC and Cl associated with upward flow through the MCCU after 1970 (lk4t1970) corresponded to periods of greater groundwater withdrawal. Therefore, upward groundwater flow through
the MCCU was an important predictor of salinity, suggesting that hypotheses of deep saline sources in aquifers are substantiated by the BRT models.

Model Limitations

Mapped predictions of SC and Cl were able to capture areas of documented high salinity, especially in the southern extent of the study area in deeper MCAQ and LCAQ units and in the MRVA where salinity is higher in southeastern Arkansas, northeastern Louisiana, and smaller regions at the northwestern edge of the study area (Figures 6 and 8). Although the spatial extent of predicted higher salinity zones was accurate compared to well observations (Figure 1), the predicted areas of high salinity in raster cells were generally lower than observed concentrations of SC and Cl in groundwater wells. For example, in northeastern Louisiana where some of the highest SC and Cl concentrations occur in the MRVA, observed Cl in wells was >1,000 mg/L, but predicted Cl in raster cells coincident with these wells ranged from 455 to 723 mg/L. Therefore, residuals at groundwater wells (i.e., difference in prediction and observation at the groundwater well) were generally less than the residual between an observed SC or Cl concentration at a groundwater well and the prediction at the coincident raster cell. To create explanatory variable raster datasets with the same resolution and extent, source rasters were resampled, which tended to average out values in high-resolution explanatory variable datasets. Therefore, the resampling may cause smoothing in explanatory variable rasters, which may bias salinity predictions at raster cells low compared to groundwater well observations. The inability of the BRT models to accurately capture highly saline water can be observed in the PI widths for SC and Cl (Figures 6 and 8), such that PIs are disproportionately wider (and uncertainty is greater) where SC and Cl are highest.

ML models produce the most accurate predictions when trained over the full range of the response variable. Data were limited in salinity modeling because most groundwater-quality samples are available from areas where the aquifer is used, which inherently biases observations toward freshwater and limits high-salinity values, as noted in Stanton et al. (2017). Furthermore, because groundwater-flow models tend to model freshwater zones and Clark and Hart (2009) did not model the highly saline portion of Tertiary units, explanatory variables derived from the MODFLOW groundwater-flow model are null where MODFLOW groundwater-flow model cells are inactive, and SC and Cl predictions do not extend into the wedge of highly saline water in the southern extent of the study area (Figures 6 and 8). For these reasons, mapped salinity predictions from BRT models characterize the spatial extent and geometry of higher salinity zones, but may underpredict observed groundwater concentrations where SC is greater than approximately 1,000 μS/cm and Cl is >100 mg/L.

PIs for TDS provide additional context for uncertainty and for evaluating the spatial extent of concentrations that may exceed the SMCL. Large portions of the aquifers exceed the 500 mg/L SMCL for TDS, especially at the upper PI (Figure 7). The BRT models extrapolated to raster cells tend to underpredict SC, which would also lead to an underprediction of TDS. Additionally,因为 PIs are not symmetrical, TDS between the model prediction and upper PI are interpreted as more indicative of observed conditions than the lower PI. The difference in aquifer area above the SMCL between predictions and upper PIs ranged from 7% to 29%, which is equivalent to approximately 2,800 to 22,000 km² of aquifer area. Observed TDS values at wells can also be used as an independent check on the predictive performance of the BRT models, as TDS was not used to train BRT models. Additionally, TDS residuals represent the difference between a well observation and a raster cell prediction (as opposed to a prediction at the groundwater well for SC and Cl, see Figure 4). Although TDS predictions show large (~4,000 mg/L) differences compared to observations at high TDS (Figure 9a), overall model fit was an $r^2$ of 0.64 and RMSE of 250 mg/L. The lower end of TDS predictions (below approximately 500 mg/L) closely follow a 1:1 line with observed values (Figure 9b). Therefore, interpreting location and spatial extent of the aquifer area above the TDS SMCL is an appropriate use of the BRT model predictions within the Mississippi embayment. ML models that predict the probability of exceeding a threshold (or concentration) have been used in other aquifers successfully for similar purposes of identifying areas above maximum contaminant levels (Ayotte et al. 2016; Erickson et al. 2018), but concentration models provide additional benefits that model predictions (i.e., concentrations) can be used directly as input for other models (such as bulk resistivity interpretations or health risk assessments).

CONCLUSIONS

BRT models were used to create spatially continuous and three-dimensional predictions of groundwater salinity across drinking-water aquifers of the Mississippi embayment. Natural log of SC and Cl was predicted directly from BRT models, with $r^2$ on holdout
data (not used to train model) of 0.74 (for both) and RMSE of 0.45 for SC and 0.81 for Cl. TDS was predicted from the linear correlation with SC and serves as an independent check of BRT model performance because TDS data were not used for model training. Overall model performance for TDS was $r^2$ of 0.62 and RMSE of 250 mg/L, where the RMSE represents residuals between TDS at groundwater wells and predictions at 1-km$^2$ raster cells. Generally, models performed better (or PI width, and thus uncertainty, was smaller) where salinity was lower, and models tended to underpredict in areas of highest salinity. Although the models were trained over the full range of measured SC and Cl values, groundwater samples are biased toward where the aquifers are used, which results in few samples in areas with moderate to high salinity. Despite TDS being underpredicted in areas of high observed salinity (greater than about 1,000 mg/L TDS), mapped areas of TDS exceeding the SMCL drinking-water standard of 500 mg/L corresponded to observed samples. Explanatory variable importance and influence on the response variable (via partial dependence plots) was used to explore drivers of salinity in the Mississippi embayment. Variables that served as surrogates for position along the groundwater flowpaths were most important for predicting SC and Cl, indicating that much of the control on dissolved solids is related to rock-water interaction as residence time increases. Upward groundwater flux through confining units and older groundwater in shallow aquifers indicated that deep sources (at least deeper than the modeled hydrogeologic units) of saline water contributed to observed salinity in the MRVA, MCAQ, and LCAQ. Surficial sources of salinity, especially Cl, were indicated by the importance of surficial explanatory variables that describe soil texture and recharge rates. BRT models support hypotheses of both surficial and deep sources of salinity and were able to map areas of high salinity, despite the heterogeneity of salinity sources throughout the aquifers. The BRT salinity predictions for the Mississippi embayment provide the first spatially continuous ML modeling predictions of groundwater quality over the aquifer extent in the region. Furthermore, the models link physical flow models (via output from the MODFLOW groundwater-flow model) to water-quality predictions, providing an integrated assessment of the water resources. Although the specific BRT models presented here (in terms of hyperparameters and explanatory variables) are limited to the Mississippi embayment, the modeling approach can be applied in other aquifer systems. Additionally, continuous salinity predictions are valuable as input into other modeling efforts, such as other water-quality constituent predictions from ML methods, freshwater volume estimates, and

FIGURE 9. Observed TDS concentrations at (a) all groundwater wells compared to predicted concentrations at raster cells and (b) at all groundwater wells with observed concentrations less than or equal to the secondary maximum contaminant level of 500 mg/L. Red line is the 1:1 line.
Using Boosted Regression Tree Models to Predict Salinity in Mississippi Embayment Aquifers, Central United States

hydrogeologic framework characterization from bulk resistivity measurements.

**DATA AVAILABILITY**

Model predictions and explanatory variables are available as a U.S. Geological Survey data release (Knierim et al. 2020).

**SUPPORTING INFORMATION**

Additional supporting information may be found online under the Supporting Information tab for this article: Table of explanatory variables and figures.

**ACKNOWLEDGMENTS**

The authors thank the contribution from B. Tom Nolan, retired U.S. Geological Survey, whose mentorship and leadership in machine learning modeling were critical to the success of the research. Thanks also to the colleagues who reviewed the manuscript for the U.S. Geological Survey and the Journal of American Water Resources Association. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

**AUTHORS’ CONTRIBUTIONS**

**Katherine J. Knierim:** Conceptualization; data curation; formal analysis; investigation; methodology; project administration; software; validation; visualization; writing-original draft; writing-review & editing.

**James A. Kingsbury:** Conceptualization; data curation; formal analysis; investigation; validation; visualization; writing-original draft; writing-review & editing.

**Connor J. Haugh:** Conceptualization; data curation; investigation; resources; writing-original draft.

**Katherine M. Ransom:** Conceptualization; investigation; methodology; resources; software; writing-original draft; writing-review & editing.

**LITERATURE CITED**

Alhassan, M., C.B. Lawrence, S. Richardson, and E.J. Pindilli. 2019. “The Mississippi Alluvial Plain Aquifers — An Engine for Economic Activity.” U.S. Geological Survey Fact Sheet 2019-3003. https://doi.org/10.3133/fs20193003.

Arslan, H. 2012. “Spatial and Temporal Mapping of Groundwater Salinity Using Ordinary Kriging and Indicator Kriging: The Case of Bafra Plain, Turkey.” Agricultural Water Management 113: 57–63. https://doi.org/10.1016/j.agwat.2012.06.015.

Ayotte, J.D., B.T. Nolan, and J.M. Gronberg. 2016. “Predicting Arsenic in Drinking Water Wells of the Central Valley, California.” Environmental Science & Technology 50 (14): 7555–63. https://doi.org/10.1021/acs.est.6b01914.

Belitz, K., R.B. Moore, T.L. Arnold, J.B. Sharpe, and J.J. Starn. 2019. “Multi-Order Hydrologic Position in the Conterminous United States: A Set of Metrics in Support of Groundwater Mapping at Regional and National Scales.” Water Resources Research 55. https://doi.org/10.1029/2019WR025908.

Belitz, K., T.L. Arnold, J.B. Sharpe, and J.J. Starn. 2019. “National Multi Order Hydrologic Position (MOHP) Predictor Data for Groundwater and Groundwater-Quality Modeling.” U.S. Geological Survey Data Release. https://doi.org/10.5066/P9HILU4YY.

Biecek, P. 2018. “DALEX: Explainers for Complex Predictive Models in R.” Journal of Machine Learning Research 19 (84): 1–5.

Broom, M.E., T.F. Kraemer, and W.V. Bush. 1984. “A Reconnaissance Study of Saltwater Contamination in the El Dorado Aquifer, Union County, Arkansas.” U.S. Geological Survey Water Resources Investigations Report 84–4012. https://doi.org/10.3133/wri844012.

Clark, B.R., P.M. Barlow, S.M. Peterson, J.D. Hughes, H.W. Reeves, and R. Viger. 2018. “National-Scale Grid to Support Regional Groundwater Availability Studies and a National Hydrogeologic Framework.” U.S. Geological Survey Data Release. https://doi.org/10.5066/P7PS4B24.

Clark, B.R., and R.M. Hart. 2009. “The Mississippi Embayment Regional Aquifer Study (MERAS): Documentation of a Groundwater-Flow Model Constructed to Assess Water Availability in the Mississippi Embayment.” U.S. Geological Survey Scientific Investigation Report 2009–5172. https://doi.org/10.3133/sir20095172.

Clark, B.R., R.M. Hart, and J.J. Gurdak. 2011. “Groundwater Availability of the Mississippi Embayment.” U.S. Geological Survey Professional Paper 1785. https://doi.org/10.3133/pp1785.

Clark, B.R., K.J. Knierim, and L. Duncan. 2019. “Zonepy: Zonal Analysis Tools for Raster and Vector Datasets.” https://github.com/brl Clark-usgs/zonepy.

DeSimone, L.A., P.B. McMahon, and M.R. Rosen. 2015. “The Quality of Our Nation’s Waters: Water Quality in Principal Aquifers of the United States, 1991–2010.” U.S. Geological Survey Circular 1360. https://doi.org/10.3133/cir1360.

DeSimone, L.A., J.P. Pope, and K.M. Ransom. 2020. “Machine-Learning Models to Map PH and Redox Conditions in Groundwater in a Layered Aquifer System, Northern Atlantic Coastal Plain, Eastern USA.” Journal of Hydrology: Regional Studies 30: 100697. https://doi.org/10.1016/j.ejrh.2020.100697.

Dieter, C.A., K.S. Linsey, R.R. Caldwell, M.A. Harris, T.I. Ivanenko, J.K. Lovelace, M.A. Maupin, and N.L. Barber. 2017. “Estimated Use of Water in the United States County-Level Data for 2015.” U.S. Geological Survey Data Release. https://doi.org/10.5066/f7thb15v5.

Elith, J., J.R. Leathwick, and T. Hastie. 2008. “A Working Guide to Boosted Regression Trees.” Journal of Animal Ecology 77 (4): 802–13. https://doi.org/10.1111/j.1365-2656.2008.01390.x.

Ericsson, M.L., S.M. Elliott, C.A. Christenson, and A.L. Krall. 2018. “Predicting Geogenic Arsenic in Drinking Water Wells in Glacial Aquifers, North-Central USA: Accounting for Depth-Dependent Features.” Water Resources Research 54 (12): 10172–87. https://doi.org/10.1029/2018WR023106.
States.” U.S. Geological Survey Hydrologic Atlas 706. https://doi.org/10.3133/ha706.
Pettijohn, R.A.1996. “Geochemistry of Ground Water in the Gulf Coast Aquifer Systems, South-Central United States.” U.S. Geological Survey Water-Resources Investigation Report 96-4107. https://doi.org/10.3133/wri964107.
Pollock, D.W.2012. “User Guide for MODPATH Version 6 — A Particle-Tracking Model for MODFLOW.” U.S. Geological Survey Techniques and Methods 6-A41. https://pubs.usgs.gov/tm/6a41/
Python Software Foundation. 2019. “Python Language Reference, v 3.6. Python.org.” https://www.python.org.
R Core Team. 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; https://www.r-project.org/.
Ransom, K.M., B.T. Nolan, J.A. Traum, C.C. Faunt, A.M. Bell, J.A.M. Gronberg, D.C. Wheeler et al. 2017. “A Hybrid Machine Learning Model to Predict and Visualize Nitrate Concentration throughout the Central Valley Aquifer, California, USA.” Science of the Total Environment 601-602: 1160-72. https://doi.org/10.1016/j.scitotenv.2017.05.192.
Hardke, J. T. 2018. “Rice Production Handbook”, University of Arkansas Division of Agriculture Cooperative Extension Service, University of Arkansas System, Little Rock, Arkansas. https://www.uaex.edu/publications/pdf/mp192/mp192.pdf.
Ridgway, G. 2019. “Gbm: Generalized Boosted Regression Models.” https://cran.r-project.org/web/packages/gbm/gbm.pdf.
Rosecrans, C.Z., B.T. Nolan, and J. Gronberg. 2017. “Prediction and Visualization of Redox Conditions in the Groundwater of Central Valley, California.” Journal of Hydrology 546: 341-56. https://doi.org/10.1016/j.jhydrol.2017.01.014.
Sahoo, S., T.A. Russo, J. Elliott, and I. Foster. 2017. “Machine Learning Algorithms for Modeling Groundwater Level Changes in Agricultural Regions of the U.S.” Water Resources Research 53 (5): 3878-95. https://doi.org/10.1002/2016WR019933.
Saucier, R.T. 1994. Geomorphology and Quaternary Geologic History of the Lower Mississippi Valley. Vicksburg, MS: U.S. Army Corps of Engineers.
Schwarz, G., A.B. Hoos, R.B. Alexander, and R.A. Smith. 2006. “Section 3. The SPARROW Surface Water-Quality Model — Theory, Application and User Documentation.” U.S. Geological Survey Techniques and Methods 6–B3. https://doi.org/10.3133/tm6B3.
Stanton, J.S., D.W. Anning, C.J. Brown, R.B. Moore, V.L. McGuire, S.L. Qi, A.C. Harris, K.F. Dennehy, P.B. McMahon, J.R. Degnan, and J.K. Böhlke.2017. “Brackish Groundwater in the United States.” U.S. Geological Survey Professional Paper 1833. https://doi.org/10.3133/pp1833.
Starr, J.J., and K. Belitz. 2018. “Regionalization of Groundwater Residence Time Using Metamodeling.” Water Resources Research 54 (9): 6357–73. https://doi.org/10.1029/2017WR021531.
Tesoriero, A.J., J.A. Gronberg, P.F. Juckem, M.P. Miller, and B.P. Austin. 2017. “Predicting Redox-Sensitive Contaminant Concentrations in Groundwater Using Random Forest Classification: Predicting Contaminant Classifications.” Water Resources Research 53: 7316–31. https://doi.org/10.1002/2016WR020197.
U.S. Environmental Protection Agency. 2013. “Safe Drinking Water Information System (SDWIS).” https://www.epa.gov/epawaters/sdwis-search.
U.S. Environmental Protection Agency. 2020. “Secondary Drinking Water Standards: Guidance for Nuisance Chemicals.” https://www.epa.gov/sdwa/secondary-drinking-water-standards-guidance-nuisance-chemicals#table.
U.S. Geological Survey. 2016. “National Map 3DEP Products and Services, 1/3 Arc-Second Digital Elevation Model.” https://viewer.nationalmap.gov/basic/?basemap=b1&category=ned,ncsrs&title=3DEP%20View.
U.S. Geological Survey. 2017. “USGS Water Data for the Nation: U.S. Geological Survey National Water Information System Database.” https://doi.org/10.5066/F7P55KJN.
USGS Advanced Research Computing. 2019. USGS Yeti Supercomputer. U.S. Geological Survey. https://doi.org/10.5066/P7D79MJ.
Wacaster, S.R., J.M. Clark, D.A. Westerman, and W. Kress. 2018. “Digital Dataset for the Geomorphology of the Lower Mississippi River Valley in Missouri, Kentucky, Arkansas, Tennessee, Louisiana, and Mississippi.” U.S. Geological Survey Data Release.https://doi.org/10.5066/F7N878QN.
Waldron, B., D. Larsen, R. Hannigan, R. Csontos, J. Anderson, C. Dowling, and J. Bouldin. 2011. “Mississippi Embayment Regional Ground Water Study.” U.S. Environmental Protection Agency Report 600/R-10/130. https://cfpub.epa.gov/si/si_public_record.cfm?Lab=NRMRL&dirEntryId=230439.
Waskom, M.2019. “Seaborn. V 0.10.” https://github.com/mwaskom/seaborn.
Whitfield, M.S.1975. “Geohydrology and Water Quality of the Mississippi River Alluvial Aquifer, Northeastern Louisiana.” U.S. Geological Survey Water Resources Technical Report 10.
Worland, S.C., W.H. Farmer, and J.E. Kiang. 2018. “Improving Predictions of Hydrological Low-Flow Indices in Ungaged Basins Using Machine Learning.” Environmental Modelling & Software 101: 169–82. https://doi.org/10.1016/j.envsoft.2017.12.021.
Worland, S.C., S. Steinschneider, W. Asquith, R. Knight, and M. Wieczorek. 2019. “Prediction and Inference of Flow-Duration Curves Using Multi-Output Neural Networks.” Water Resources Research 55 (8): 6650–686. https://doi.org/10.1029/2018WR024463.
Zimmerman, R.K. 1992. “Fractured Smackover Limestone in Northeast Louisiana; Implications for Hydrocarbon Exploitation.” Gulf Coast Association of Geological Societies Transactions 42: 401–12.