Appendix C – Predictive Model

Summary

This document provides the technical underpinnings of the modeling approach taken to predict areas likely to have high nitrate concentrations across the State of Nebraska for the 2023-2024 Nebraska Department of Environment and Energy (NDEE) Water Quality Study. The model predictions represent the probability that nitrate concentrations will exceed certain threshold values in private domestic wells based solely on the model inputs listed in Table 1. In this study, threshold values of 3 mg/L, 5 mg/L, and 10 mg/L were modeled as representative of the background, elevated, and maximum contaminant level for nitrate in groundwater, respectively. Ultimately, this estimate is just one factor used in the web-based Geographic Information System (GIS) risk assessment tool for use by NDEE and agency partners. Regardless of the predicted risk, private domestic well owners are strongly encouraged to sample their well annually to properly assess their specific risk. Model construction and results offer valuable insights into the relationship between nitrate concentrations in Nebraska, common sources, hydrogeological factors, and land-use trends. Exploratory analyses and literature review were first conducted to identify potentially influential factors, then Boosted Regression Trees (BRTs) were trained to classify wells likely to exceed each threshold value. Finally, the BRTs were generalized for the internal NDEE GIS tool and evaluated against private domestic well samples from the free NDEE sampling effort. Model performance was strong for the testing and training data, and the model surfaces had acceptable performance compared to the fully independent private domestic well samples. However, additional work on the model is recommended to incorporate additional variables known to impact nitrate concentrations and reduce the false negative predictions (under prediction of nitrate concentration). A Model Card (based on the one proposed by Mitchell et al, 2019) is provided in Model Card

Table 1.

Outline

List of Figures

List of Tables

Model Card

Table 1.Model Card.

Model Owner	Nebraska Department of Environment and Energy (NDEE)
Model Date	November 2024
Model Version	1.0
Model Type	Boosted Regression Trees (Classification)
Spatial References	Datum: North American Datum 1983 State Plane Projection: Lambert Conformal Conic
Model Goals	• Evaluate the relationship between nitrate concentrations, common point sources (limited to registered onsite wastewater treatment (OWT) facilities and animal feeding operations (AFO)), contributing land-use patterns, well construction, and hydrogeological factors. • Predict the probability that nitrate in a domestic well will exceed three values: a background concentration, an elevated concentration, and the Maximum Contaminant Level (MCL). Identify high-risk areas of nitrate in groundwater. • Incorporate model results into a private domestic well risk assessment GIS tool to be used by NDEE and select agency partners.
Model Inputs	
Nitrate Well Samples	The median nitrate sample from the Nebraska Groundwater Clearinghouse database for the period 2003-2019 was calculated at each well modeled. These median values were converted to binary variables at three threshold values: 3 mg/L, 5 mg/L, and 10 mg/L. Concentrations above each threshold were assigned 1 and below were assigned 0. Around each well, a 1500-meter buffer was generated and used to aggregate predictor variables that were not defined at the well level (such as nitrate concentration or well construction details).
Well Construction	Well construction variables were derived from the Nebraska Department of Natural Resources (NDNR) registered wells database. One-half screened interval depth, pumping water level, static water level, the presence or absence of a well seal, and well depth variables were included in the model. Location for each well was represented in the model using latitude and longitude as numeric variables.
Land Use	USDA Cropland Data Layer (CDL). Pixel counts and class percentages were calculated for each well buffer. The percentage of cultivated soybeans and corn were included in the model. USGS 30-meter irrigated acres (LGRIP30) 2023 Release. The percentage of irrigated and rainfed crops in each well buffer were included in the model. Historic fertilizer data assembled by the USGS, derived from USDA National Agricultural census, were used to estimate application rates at the county-scale and joined to wells included in the model. Municipal boundary information from Nebraska Map, a census derived product, was used to represent the potential impacts of municipal wastewater collection systems and other potential urban sources of nitrate, such as lawn fertilizer. Nebraska Map is managed by the Nebraska Geographic Information Office (GIO).

Hydrogeological	Soil infiltration data from Soil Survey Geographic Database (SSURGO) was sampled at 30-meter resolution and aggregated by mean value inside each well buffer. The mean vertical soil infiltration (ksat) for each well was used in the analysis. Streams have an impact on nitrate concentrations where surface and groundwater are interconnected. The distance to the nearest stream was calculated for each well buffer. Stream data came from the NDEE Title 117 waterbodies database. Reservoirs and lakes can also impact nitrate concentrations much like streams. Similarly, the distance to the nearest Title 117 lake was calculated for each well buffer.
Model Inputs	
Point Sources	Registered OWT facilities from the NDEE integrated information system (IIS) were aggregated into well buffers as a per square mile density and as a distance measured from the buffer edge for each well. Animal Feeding Operation (AFO) facilities from the NDEE IIS were aggregated into well buffers in the same manner as OWT facilities, with the addition of a facility count metric for AFOs in each well buffer. Animal facilities were also represented by livestock watering wells data from the NDNR registered wells database. Watering wells may capture areas where animals graze and smaller operations not permitted under Title 130.
Model Outputs	Probability that the median nitrate concentration will exceed a background concentration, elevated concentration, and the MCL (based on model inputs within a 1500-meter radius of each well), confusion matrix, variable influence, partial dependence, variable interaction, evaluation statistics, and associated plots. Predictor variables were aggregated to a half-mile grid surface across Nebraska and passed to the trained models to generalize the predictions for use in a GIS tool for use by NDEE and key partners.
Model Evaluation	Models were optimized to maximize Matthew's Correlation Coefficient (MCC), Sensitivity, Specificity, and Overall Accuracy calculated from the Confusion Matrix for each model. Values are reported for testing data. MCC values were between 0.5 – 0.51. Sensitivity was from 55 – 88%. Specificity was from 59 – 92%. Overall Accuracy was from 78 – 81%. Model surfaces were compared to an independent set of domestic well samples collected in 2023-2024. Evaluation metrics were lower across the board for the model surfaces, but MCC (0.20 – 0.28), sensitivity (34 – 60%), specificity (68 – 87%), and overall accuracy were acceptable (65 – 79%) to recommend model results for incorporation into an internal tool for NDEE and key partners.
Credits	Author: Bridger Corkill Year: 2024 Affiliation: Nebraska Department of Environment and Energy
Intended Use	This model is intended to supplement a risk assessment tool for private domestic wells. The model considers many factors that may influence the nitrate level around a private domestic well. Estimating the probability that nitrate will exceed the modeled threshold concentrations can help assess risk for private domestic wells located in areas where nitrate samples are unavailable. Nitrate concentrations were modeled based on a range of threshold concentrations that reflect a low, medium, or high-risk potential to private domestic well owners. The model is not intended to predict the exact concentration at any one well location. Rather, the goal is to provide a reasonable baseline assessment of risk potential given the available data and model inputs. Additional risk factors will be included in the GIS tool. This model is not intended as a primary decision-making tool, and will be used exclusively by NDEE and select agency partners.

Introduction

One objective of the water quality study, conducted by the NDEE, was to develop a model identifying high-risk areas of nitrate in groundwater. Results of this modeling effort are intended to supplement a risk assessment Geographic Information System (GIS) tool that was developed during the water quality study. This tool assists NDEE and select agency partners in evaluating the potential risk of elevated nitrate in a domestic well. In the GIS tool, the user will enter a well location, and the tool queries information for that location to calculate a risk index and create a report for the user. Predictive model results are one part of this risk index and are intended to provide an estimate of how likely a private domestic well owner is to find elevated nitrate concentrations in their well, based on contributing factors and existing nitrate sample data. Ultimately, the only way to ensure a safe supply of drinking water is to have it tested.

Previous studies conducted in coordination with the U.S. Geological Survey (USGS), such as Nolan et al. (2014) and Wheeler et al. (2015), have employed machine learning (ML) methods to predict nitrate concentrations, including the probability that N as Nitrate will exceed several thresholds, in private domestic wells. Similar studies conducted by USGS in Wisconsin (Wellman and Rupert, 2016; Borchardt et al., 2021) use logistic regression analysis to predict the risk of domestic well contamination by several contaminants, including nitrate. Traditional regression methods were not used in this study because the nitrate data used to train the models does not meet many of the underlying assumptions for a regression model, such as Gaussian distribution of model residuals and a homogenous relationship between nitrate and predictor variables across the model space, i.e., the state of Nebraska. ML algorithms do not require a particular distribution or assume the data has a homogenous relationship across the model space. They also benefit from large, multi-dimensional datasets (Breiman et al., 1984). Because of these advantages, this study uses a forest-based classification algorithm, BRTs, to predict whether a well is likely to exceed several threshold values for nitrate concentration based on well characteristics, geologic conditions, land-use, and some common potential sources of nitrate. Predictions were made for wells considered representative of domestic well construction in Nebraska. Figure 1 shows the nitrogen cycle.

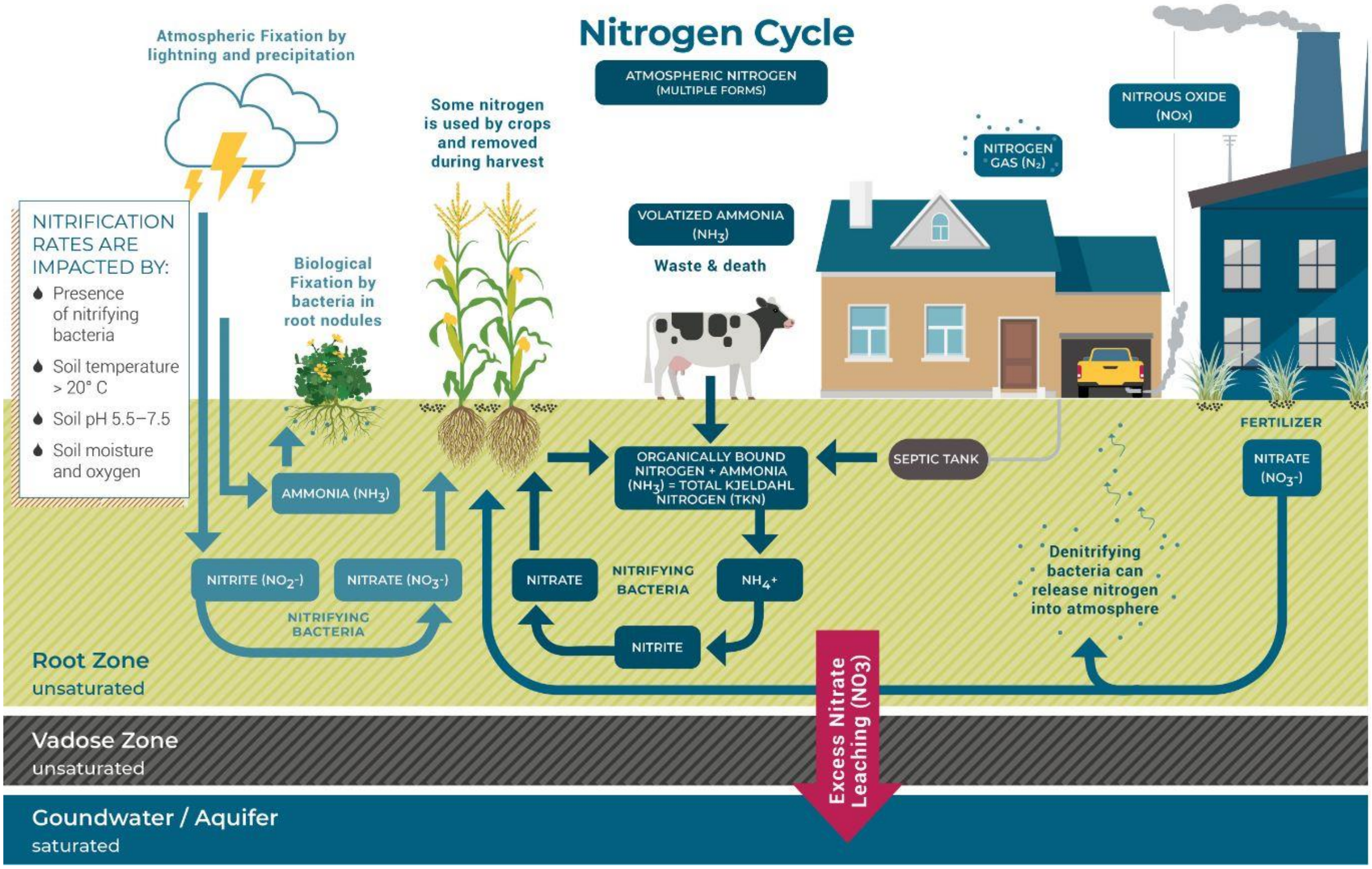

Nitrogen Cycle
ATMOSPHERIC NITROGEN (MULTIPLE FORMS)
Atmospheric Fixation by lightning and precipitation
Some nitrogen is used by crops and removed during harvest
NITROUS OXIDE (NOx)
NITROGEN GAS (N_2)
VOLATIZED AMMONIA (NH_3)
Waste & death
NITRIFICATION RATES ARE IMPACTED BY:
Presence of nitrifying bacteria
Soil temperature > 20° C
Soil pH 5.5–7.5
Soil moisture and oxygen
Biological Fixation by bacteria in root nodules
AMMONIA (NH_3)
ORGANICALLY BOUND NITROGEN + AMMONIA (NH_3) = TOTAL KJELDAHL NITROGEN (TKN)
SEPTIC TANK
FERTILIZER
NITRATE (NO_3-)
NITRITE (NO_2-)
NITRATE (NO_3-)
NITRIFYING BACTERIA
NITRATE
NITRIFYING BACTERIA
NH_4^+
NITRITE
Denitrifying bacteria can release nitrogen into atmosphere
Root Zone unsaturated
Excess Nitrate Leaching (NO3)
Vadose Zone unsaturated
Goundwater / Aquifer saturated

Figure 1. Nitrogen Cycle Conceptual Diagram.

Nitrogen takes multiple forms in the environment and comes from both organic and inorganic sources. Nitrogen typically enters the soil as ammonia where it is nitrified to nitrate under oxic conditions. Nolan and Hitt (2002) report that background concentrations in undeveloped forested areas of the United States are around 1 mg/L. Levels measured slightly higher in rangeland and grassland, between 2 and 3 mg/L. Further studies of the High Plains Aquifer (McMahon, 2007) have generally agreed that 4 mg/L is the highest observed “relative” background concentration in the system. Background and relative background concentrations of nitrate are an area of debate in literature. Nitrate concentrations between 0.5 and 3 mg/L are considered a transitional range between natural background and anthropogenic contamination (McMahon et al., 2007). For this study, a conservative background level of 3 mg/L was assumed based on Nebraska land-use trends. Anhydrous fertilizer and livestock manure application to cropland are two primary sources of nitrogen in the soil and streams (Spalding and Exner, 1993). Additional sources include human and livestock waste, certain industrial facilities, and wastewater treatment facilities (ATSDR, 2017).

Inorganic and organic nitrogen (as ammonia) are nitrified in the soil to nitrite and then nitrate under oxic conditions. In the High Plains system, dissolved oxygen levels are such that nitrate can persist for decades (Spalding and Exner,1993; McMahon et al., 2007). Nitrate management practices can reduce levels over time, but in Nebraska levels may still be rising (Exner, 2014). When nitrate is not biologically fixed—by plants or microorganisms—it leaches through the unsaturated root and vadose zone eventually reaching groundwater (Malakar et al., 2023). The time it takes for nitrate to reach groundwater is related to the thickness of the vadose zone, the depth to groundwater, soil characteristics, precipitation, and irrigation (Wells et al., 2018; Malakar et al., 2023). In areas where groundwater and surface water are interconnected, groundwater can be a source of nitrate in streams or vice versa (Green et al. 2018). Domestic wells are more likely to tap shallower formations and are often constructed near onsite wastewater treatment systems, cropland, and animal feed operations which can all contribute to contamination (Wheeler et al., 2015; Wellman et al., 2016; Borchardt et al., 2021). In addition to permitted animal feed operations and onsite facilities, livestock watering wells registered with the NDNR were incorporated into the model to represent areas where animals may graze that are not captured by a single facility location.

The water table varies throughout the year and upper levels of an aquifer may have different nitrate concentrations than deeper, older groundwater. In areas where groundwater and surface water are interconnected, flows from groundwater to surface water may act to ‘flush out’ excess nitrate into streams and rivers (Snow and Miller, 2018; Malakar et al., 2023). Seepage from losing reaches and reservoirs may have the opposite impact. Additionally, the varying geology across Nebraska dramatically changes the rate at which nitrate reaches groundwater (Spalding, 2001; Wells, 2018; Cherry, 2019).

Nitrate is more rapidly transported to groundwater under irrigated lands than non-irrigated lands. Irrigated crops typically receive more fertilizer application than non-irrigated crops and therefore have a higher nitrate soil concentration contributing to nitrate leaching (Spalding 2001; Exner 2014; Malakar et al., 2023). Excess water from irrigation not taken up by crops pushes nitrate through the unsaturated vadose zone. Irrigation wells, which may be constructed with gravel pack along their entire casing, can act as conduits for water high in nitrate to move rapidly into lower levels of the aquifer. Wells that are screened or gravel packed through multiple formations can cause aquifer comingling (Driscoll, 1986). The impact of agriculture was captured in this study using percentage of irrigated cropland, crop-percentages, cumulative nitrogen application estimates, and livestock facility data. The 30-meter irrigated acres (LGRIP30) product produced by the USGS was used to estimate the percentage of irrigated area around each well. It is nominally a 2015 product (Teluguntla, 2023); however, because of the permitting requirements and water management by the Nebraska Natural Resources Districts, the total irrigated acres over the study period should be relatively constant. Additionally, investigation of the cropland data layer (CDL) in Nebraska showed little change over time in the dominant crop classes.

Well samples for 281 water quality indicators, including nitrate, are available to the public in Nebraska via the Nebraska Quality Assessed Agrichemical Clearinghouse (the Clearinghouse). The Clearinghouse is a collaborative effort between the NDEE, the University of Nebraska-Lincoln Conservation Survey Division (UNL CSD), and the Natural Resources Districts of Nebraska (NRDs). Nitrate samples in this study were all sourced from and are publicly available on the Clearinghouse. Samples have been collected from monitoring, irrigation, domestic, public water supply, commercial/industrial, livestock, and groundwater source heat pump wells since mid-1974 to present. Each sample is given a quality flag based on the methodologies used for sampling and the laboratory method. The flag depends on the amount and type of quality assurance/quality control that was identified in obtaining each sample. At the time of the study, data for the years 2020 to present is incomplete. No data quality filter was applied to nitrate samples used to train and test the models.

Point sources of nitrate, such as failing onsite treatment systems, are an important source to consider for estimating the nitrate risk in a domestic well (Nolan et al., 2014; Wheeler et al., 2015). OWT facility data from Title 124 permit records were used to calculate the impact of OWTs on nitrate. There are important limits to this record. Title 124 Onsite Wastewater Systems requires registration of any OWT constructed, reconstructed, altered, modified, or otherwise changed by a certified professional, professional engineer, or registered environmental health specialist since January 1, 2004. There are currently approximately 29,600 registered OWT, but many OWTs are not registered, and some OWTs are exempt from registration. Data considered for inclusion in the predictive model are summarized in Table 2. Other point sources, such as those regulated by NDEE’s release assessment program, did not have the data quality needed for inclusion in the model.

Table 2. Datasets Considered for the Predictive Nitrate Model.

Dataset	Agency & Year	Description
Clearinghouse Well Samples	NDEE, UNL CSD, 2024	Nitrate samples from the Clearinghouse from Non-Public Water Supply wells for the years 2003 to 2023 were used as model inputs. Because of data gaps in the Clearinghouse, this is nominally a 2003-2019 product. The median nitrate concentration from all samples taken over the study period was calculated at each well.
Domestic Well Samples from the Free Sampling Effort	NDEE, 2024	Results from the NDEE free sampling effort were used as an independent testing set to evaluate model performance. These results are from samples collected by private well owners, per instructions they received with their nitrate test kit from the Nebraska Department of Health and Human Services (NDHHS) Public Health Environmental Lab. Some of these samples may have been collected following reverse osmosis or other treatment units and they may not all be representative of raw well water.
National Land Cover Dataset (NLCD)	USGS, 2022	Land-use trends and data were analyzed for relationships to nitrate levels in Nebraska. Data were aggregated to well buffers by pixel counts, and percentages for each land use type were compared to nitrate levels. In the models, LGRIP30 and the CDL were used instead of NLCD data.
Depth to Groundwater	NDEE, 2024	Groundwater elevations, based on the regional Nebraska hydrologic models, were calculated for the spring season and generalized across the state. These elevations were not incorporated into the modeling but may benefit future efforts.
Well Construction Information	NDNR, 2024	Well construction information (e.g., well depth and construction year) for wells in the Clearinghouse, provided by NDNR, was evaluated for relationships to nitrate levels. Well construction variables were derived from the NDNR registered wells database. Location, one-half-screened interval depth, pumping water level, static water level, presence or absence of a well seal, and well depth variables were included in the model. Location for each well was represented by latitude and longitude as numeric variables
Soil Survey Geographic Database (SSURGO) Soil Properties	NRCS, 2023	The SSURGO database was used to generate representative saturated soil infiltration rates (ksat) for each well buffer distance. Other SSURGO variables recommended for future modeling are discussed in the conclusions and recommendations section.
Cropland Data Layer (CDL)	USDA, 2022	The CDL was analyzed for relationships to nitrate levels and change over time. The percentage of corn and percentage of soybeans was calculated inside each well buffer and included in the modeling. Other notable classes, such as alfalfa and winter wheat were considered but ultimately excluded and covered by LGRIP30 data.

Dataset	Agency & Year	Description
LGRIP30	USGS, 2023	USGS 30-meter irrigated acres (LGRIP30) product was used as a model input. The percentage of irrigated and rainfed crops in each well buffer were included in the model.
Nebraska Permitted Irrigated Acres	NDNR, 2023	The irrigated acres from groundwater were queried from the Permitted Irrigated Acres data layer maintained by the NDNR. LGRIP30 was selected to represent irrigated acres in the dataset instead of this product.
Registered Onsite Wastewater Treatment (OWT) Systems	NDEE, 2023	Title 124 registered OWT facilities were aggregated by well buffer as count, distance, and density values. Domestic, industrial, and commercial facilities were included.
Registered Animal Feed Operations (AFOs)	NDEE, 2023	Animal feed operations (AFOs), as defined by Title 130, were aggregated into well buffers by facility count, distance, and density values. Facility data were retrieved from the NDEE IIS.
Livestock Watering Wells	NDNR, 2023	Stock wells were aggregated by count into well buffers and as a per square mile density value inside each buffer.
Historic Fertilizer Application Rates	USGS, 2006	County level fertilizer application data from USGS for the years 1987 to 2006 was normalized over the land area in each county and then joined to wells as a kg/land-acre rate value.
Groundwater Release Assessments	NDEE, 2024	Release assessment data is collected by NDEE but was not in a form that could be reliably included in the modeling.
Permitted Nitrate Precursor Storage Facilities	NDEE, EPA CAMEO, 2024	Tier two storage facilities are required to report through NDEE to the EPA on chemical storage facilities. These data were ultimately excluded from the model.
Nebraska Municipal Boundary Data	Census 2020; NE Geographic Information Office (GIO), 2024	Municipal boundaries in Nebraska are derived from the 2020 census and updated by NGIO using state data from the Department of Revenue and annexation ordinances from cities. Municipal boundaries were used in the models to represent urban sources of nitrate, such as fertilizer runoff and wastewater collection systems.
Title 117 Waterbodies Database	NDEE, 2024	NDEE maintains a database of regulated surface waters under Title 117. Streams, lakes, and reservoirs can all impact nitrate in groundwater when they are hydrologically connected. Data on Title 117 defined streams and lakes that were incorporated into the models as distance variables.
Title 123 Wastewater Treatment Facilities	NDEE, 2024	Permitted wastewater facilities defined by Title 123 were considered for inclusion in the model but were ultimately excluded and the impact of municipal treatment and collection systems was represented using the municipal boundary data.

Methods

Variable Aggregation

Nitrate well sample data from the Clearinghouse for the period 2003 to 2019 from non-PWS wells 300 feet or less in depth were used in the analysis. Because some wells have been sampled multiple times, the median concentration at each well was calculated prior to analysis. 1500-meter radius buffers around each well were created using ArcGIS Pro (Arc version 3.1) and used to aggregate variables. Buffer distances in this study were comparable to those used in previous studies (Tesoriero and Voss, 1997; Nolan et al., 2014; Borchardt et al., 2021). Variables can be broadly categorized as either aggregated at the well level or buffer level. Well construction information and nitrate sample data were joined to each well, while land-use variables, potential sources, distance features, and hydrogeologic features were aggregated in each well buffer.

Preliminary variables were assembled based on related modeling studies (Nolan et al., 2014; Wheeler et al., 2014; Wellman et al., 2016), potential sources of nitrate (ATSDR, 2015), historic information on nitrate in Nebraska (Spalding and Exner, 1993; Litke, 2001; McMahon, 2007), data availability, and consultation with modeling, hydrology, and engineering experts. A list of all datasets considered for inclusion is presented in Table 2. Notable exclusions from the model are discussed in the recommendations section for future work.

Well construction information, including well depth, static water level (SWL), pumping water level (PWL), drawdown (the difference between SWL and PWL), depth to the mid-point of the screened interval, length of gravel pack, presence of a seal, and pump rate were derived from the NDNR Registered Wells Database. Construction data are collected when the well is registered and may not reflect changes to water level, well depth, or pump level. Where information on well construction was unavailable, the variable was set to Null. Except for drawdown, gravel pack length, and pump rate, all available construction data was used in the modeling.

Hydrogeologic variables including vertical soil infiltration rate (Ksat), aquifer boundary data, stream location, and depth to bedrock geology were considered for inclusion in the model. Ksat was calculated from the USGS SSURGO dataset by first resampling the 10-m product to 30-m resolution and then zonal statistics were calculated inside each well buffer. The mean Ksat value was selected as the representative statistic and included in the model. Additional variables from the SSURGO database, such as hydric rating, drainage class, and soil geochemical properties were considered for inclusion. However, these data were not in a format that was usable in the modeling effort at time of writing. A discussion of additional SSURGO factors for the model is presented in the conclusions and recommendations. The distance to the nearest stream and nearest lake, as defined by Title 117, was calculated for each well buffer and included in the model. Aquifer boundary data were ultimately excluded from the model but may be a good option to divide the state into regions for future groundwater modeling efforts.

Distances between each well buffer and potential point-source datasets were calculated and used as explanatory variables. Models include only onsite wastewater treatment facilities, livestock watering wells, and permitted livestock facilities as listed in Model Card

Table 1. A discussion of missing facilities data for potential nitrate sources is provided in the conclusions and recommendations section. No maximum distance was established. Facilities inside the buffer had distance equal to zero. Facility counts by type inside each buffer were also calculated for livestock facilities and livestock watering wells. Density for these point facilities was calculated as a facility per square mile value using the focal statistics tool in ArcGIS Pro (Version 3.1) using a 6-mile moving window. Mean facility density values were aggregated into each well buffer distance using the Zonal Statistics tool in ArcGIS Pro 3.1. Livestock watering well data from the NDNR registered wells database was also used to calculate a well per square mile value across the state to represent areas where livestock may be moved to that are not captured by permitted facility data.

Land-use data from the National Land Cover Dataset (NLCD), an irrigated acres product derived from NLCD called LGRIP30 (Teluguntla et al., 2023), and the CDL (USDA NASS, 2023) were evaluated for inclusion in the model. The NLCD land-cover dataset did not have adequate variance to include as a model input, a vast majority of the land in Nebraska is either grassland or cropland. For each land use dataset, the 30-m products were aggregated into the 1500-meter well buffer and pixel statistics were calculated summarizing the land use percentages. LGRIP30, including irrigated and rainfed cropland data, and the CDL, including only the two largest classes, corn, and soybeans, were included in the models. The CDL from 2008 was used in the modeling. Analysis of the CDL in Nebraska showed little change in major crop classes over time.

Nitrogen application rates were estimated by USGS at the county level for the years 1988 to 2006 (Spahr et al., 2010). Previous studies discussed the impact of legacy fertilizer application on present-day nitrate levels (Exner, 2014). This study seeks to empirically account for this legacy nitrogen input based on the 2006 USGS county level estimates. A cumulative nitrogen application rate was calculated as follows: farm and non-farm tonnage was summed across years, then the sum of nitrogen applied in kilograms (kg) was divided by the total land area in each county (in acres) to estimate the cumulative application per acre. These county level values were joined to each well. As with the CDL, using values from 2006 is reflective of the lag-time between nitrogen application at the surface and elevated groundwater nitrate levels (Wheeler et al., 2015; Cherry et al., 2019).

Prior to modeling the wells with nitrate sample data from the Clearinghouse, data were randomly divided into testing and training groups. Training data are used in the model training. Testing data are not used in model training and are instead used to evaluate model performance (Breiman et al., 1984). Two-thirds of the sample data were set aside for training and one-third for testing. To ensure a repeatable split, wells were sampled in R (Kuhn, 2020) using a fixed seed. The same seed was used across models.

During the 2023-2024 water quality study, NDEE offered free nitrate test kits to private domestic well owners. Results from the NDEE free sampling effort were used as an independent testing set to evaluate model performance. These results are from samples collected by private domestic well owners per the instructions they received with their nitrate test kit from the Nebraska Department of Health and Human Services (NDHHS) Public Health Environmental Lab. Some of these samples may have been collected following reverse osmosis or other treatment units and they may not all be representative of raw well water. Because not all construction variables were known for these wells, they were used to evaluate the performance of generalized model results discussed later in this section. Samples were geocoded using ArcGIS Professional (Version 3.1, 2023) To alleviate data quality issues, duplicate addresses, P.O. boxes, points of interest, and street centerlines were removed from the set of geocoded points to calculate model evaluation metrics.

Boosted Regression Trees (BRTs)

Previous water quality investigations have used regression (Hirsch et al., 2010; Garcia et al., 2017), logistic regression (Black et al., 2023; Wellman and Rupert, 2016; Gross and Low, 2013; Lombard, et al. 2021), and machine learning methods like those employed in this study (Nolan et al., 2014; Nolan, 2015 et al.; Lombard et al., 2021; Knierim et al., 2022) to predict water quality in surface and groundwater. Logistic regression and regression were explored for predicting nitrate concentrations in this study, but the nitrate data available violate several important assumptions of traditional regression methods such as a Gaussian distribution of the model residuals and a uniform relationship between predictor variables and response variable across the model space. Additionally, machine learning methods had stronger predictive power during testing.

Random forest models use a set of tree predictors to classify data or fit regression coefficients to predict a continuous variable (Breiman et al., 1984). Forest-based models have been applied to water quality predictions in nitrate investigations (Nolan et al., 2014; Wheeler et al., 2015), and to predict other regulated contaminants like arsenic and manganese (Lombard et al., 2021; Knierim et al., 2022). Variables were aggregated in ArcGIS Professional (Version 3.1, 2023) and models were tuned in R using the dismo and gbm packages (Friedman, 2002; Hijmans, 2023). In this study, classification was chosen over continuous prediction. For the purposes of this investigation priority was placed on predicting whether a private domestic well is likely to exceed threshold concentrations and pose a health risk rather than predicting specific concentrations at a given well.

Forest-based classification uses combinations of input variables and an element of randomness to predict class membership (Breiman et al., 1984). Decision 'trees' based on a random sampling of predictor variables, vote on the most popular class for a given input vector. BRTs are a type of forest-based regression model that has been employed in species distribution modeling (Elith et al., 2008; Yu et al., 2020) and water quality analysis. Two studies with similar hydrology and investigation goals looked at relatively shallow, unconfined aquifers in the California Central Valley (Nolan et al., 2014) and the State of Iowa (Nolan et al., 2015) using forest-based classification and/or BRTs. Contaminants like arsenic (Lombard, et al. 2021) have also been modeled using BRTs. In this study, nitrate well samples were classified into binary variables at three concentration thresholds: 3 mg/L, 5 mg/L, and 10 mg/L. These values represent the upper-end background concentration in unpopulated grassland areas (Nolan and Hitt, 2003), an elevated level of nitrate, and the Safe Drinking Water Act (SDWA) Maximum Contaminant Level (MCL), respectively (US EPA, 1991). The BRT models in this study were trained to predict the probability that nitrate would exceed each concentration threshold.

BRTs are made up of many simple tree-predictors, which in aggregate, are optimized for predictive accuracy. This can be analogously thought of as many rules of thumb may be more practical than a single, complex rule to describe every situation (Elith et al, 2008). In the classification case, trees predict the most likely class instead of fitting a continuous response. Boosting, in BRT, is the combination of tree-predictor models which 'boosts' the strength of the constituent trees (Friedman, 2003). BRTs are well suited to modeling various predictor variables (continuous, categorical) and are robust to missing data (Breiman et al., 1984).

Each tree's contribution to the overall model is governed by the learning (shrinkage) rate. Generally, model performance is more robust using a low (slow) value, because of the optimization procedure. "Boosting is a form of functional gradient descent," where the unexplained deviance in the model is minimized at each stepwise addition of trees to the forest (Elith et al., 2008). A smooth descent along the curve leads to more stable model behavior (Friedman, 2003). Variable influence is calculated for the predictors in BRT models in the gbm package (Friedman, 2002) and is a measure of how frequently a variable is selected for splitting. Variables that contribute to a greater reduction in error are weighted more heavily by the measure. Relative variable influence for the model sums to 100%. Higher variable influence indicates that the variable is strongly influential to model predictions (Friedman, 2002). Percentage of relative influence does not equal percentage contribution to response variable. That is to say, the percentage influences reported by the BRT models do not correspond to percent contribution to nitrate levels in groundwater. Rather, they indicate how strongly each contributing variable is related to *predicting* the nitrate risk. Collinear factors, while largely unproblematic for BRT efficacy, do impact the calculations for variable influence and should be considered when interpreting the results (Dormann et al., 2013; Belitz and Stackelberg, 2021).

Tree complexity refers to the number of variable interactions possible in each decision tree constituent of the model. A complexity of 1 would be a "stump" with one variable and two terminal nodes. The addition of all these stumps would make up the BRT, where each stump casts its vote for the most likely class. A complexity of two allows for two-way interaction, and so forth (Elith et al., 2008; Breiman et al., 1984). Variable interaction can be tabulated and plotted from BRT models because interactions are inherent to the structures of decision trees. As splits in the tree progress, later predictor variables are dependent on the branches of earlier predictors. In this way, variable interaction is a part of the method (Breiman et al., 1984). By holding other predictors to mean values, partial variable influence plots can be developed for the response variable in the gbm package (Friedman, 2002). These partial influence plots offer insight into the shape and relative relationships between the predictor variables and the response. Because of the method for their creation, partial dependence plots should not be interpreted as individual models or used to interpolate specific values (Friedman, 2002).

Interaction can also be calculated and visualized between variables in the BRT ensemble. Variable interaction plots, created using the same principle as partial influence plots, can be created to show the interactions between influential factors in the model (Friedman, 2002). Like partial dependence, these plots are not intended to perfectly represent the relationship between nitrate and each predictor variable, but they do offer insights into how model variables interact with each other. For instance, it is expected that irrigation and soil infiltration rate will impact the rate at which nitrate reaches groundwater (Exner, 2014; Wells et al., 2018; Malakar et al., 2023), and the interaction between these factors in the model may shed additional light on that relationship.

Evaluation Metrics

Models were evaluated using Matthew's Correlation Coefficient (MCC), sensitivity, specificity, and total accuracy. A confusion matrix, with associated statistics, was calculated for each classification model using the R package caret (Kuhn, 2023). MCC was the primary evaluation metric, and all measures used to evaluate model performance are summarized in Table 3. There are four possible outcomes for binary classification in confusion matrix calculations: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). True positives refer to the samples that were above the threshold concentration accurately classified by the model. True negatives are the samples that were below the threshold concentration accurately classified by the model. False positives indicate a model prediction of above the threshold, but an actual value below. False negatives are the samples that were above the threshold concentration incorrectly classified as below. False negatives are more problematic to this study than false positives because a false positive may encourage someone to test their well, while a false negative may engender a false sense of safety.

In binary classification, MCC provides a measure of how model predictions compare to the performance of random predictions (Matthews, 1975, Chicco, 2021). MCC ranges between -1 and 1 where -1 indicates discord between predictions and actual values, 0 indicates predictions no better than random, and 1 indicates perfect agreement between model and observation. Positive MCC values can be interpreted on the same scale as Pearson's R (Chicco, 2021, Sokal et al., 1969). Sensitivity is the percentage of samples above the threshold concentration correctly classified by the model. Specificity is the percentage of true negatives predicted by the model out of total negative samples (Sokal et al., 1969). Accuracy, sometimes called overall accuracy, is a measure of sensitivity and specificity. Evaluation metrics were calculated using the following equations:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$Sensitivity = \frac{TP}{P}$$

$$Specificity = \frac{TN}{N}$$

$$Accuracy = \frac{TP + TN}{P + N}$$

Where TP stands for true positive, TN stands for true negative, FP stands for false positive, and FN stands for false negative. MCC was selected as the primary evaluation metric because it is robust to lopsided datasets and appropriate for binary classification problems (Chicco, 2021).

Table 3. Evaluation Metrics Used to Evaluate Models.

Diagnostic	Type of Model	Description
Accuracy	Binary Classification	Accuracy is a measure of Sensitivity for all classes, in the case of binary classification, accuracy is the same across classes (Sokal et al., 1969).
Sensitivity	Binary Classification	Sensitivity is the percentage of samples that fall above the threshold value correctly classified by the model (Sokal et al., 1969).
Specificity	Binary Classification	Specificity is the percentage of samples that fall below the threshold value correctly classified by the model (Sokal et al., 1969).
Matthews Correlation Coefficient (MCC)	Binary Classification	Also known as the mean square contingency value or phi statistic, the MCC is a measure of agreement between predicted and actual values. In binary classification, it is akin to comparing the model to a coin flip. Interpretation of MCC on the order of Pearson's R correlation, where 1 indicates perfect agreement between model and observation, -1 is disagreement, and 0 is no better than a random prediction (Matthews, 1975 and Chicco, 2021).

Generalizing Model Results for the GIS Tool

A key goal of this modeling investigation is to supplement a web-based GIS tool for NDEE and select agency partners to help evaluate the risk of elevated nitrate concentrations. In the ideal case, this tool would host the model weights and predictor datasets and predict to the user-entered well location based on the local factors. Because of technical limitations, this is not possible in the near term and an alternate product covering all possible input locations for the tool, i.e., a statewide product, is highly desirable.

Model results from each BRT were generalized across the state to form a smooth prediction surface by first aggregating the predictor variables to an arbitrary half-mile grid surface in the ArcGIS Professional (Version 3.1) software suite, then importing the data into R where trained model files predicted to the surface, and finally mapping the results. Variable aggregation followed the same procedure as the wells data with one notable exception. All available well construction data from the NDNR for active, registered wells was used to create the grid surface, including wells that were not sampled for nitrate or included in the model data.

Data Exploration

Nitrate samples from the Clearinghouse and variables summarized in Table 2 were explored prior to final aggregation strategy and modeling. This section summarizes the important elements of the data exploration. Because some wells have been sampled multiple times during the period 2003-2019, median nitrate concentration was calculated for each well and is mapped in Figure 2. Observed concentrations were converted to binary responses as described in the methods section. Training wells are shown in Figure 3, symbolized based on the 10 mg/L MCL threshold. GBM-MCL stands for Gradient Boosted Model – Maximum Contaminant Level. Each model is named following this convention which is used in figures throughout the text. Wells below that value are symbolized in navy and wells above the MCL are symbolized in yellow. Figure 4 shows the wells used to test the model symbolized in the same fashion.

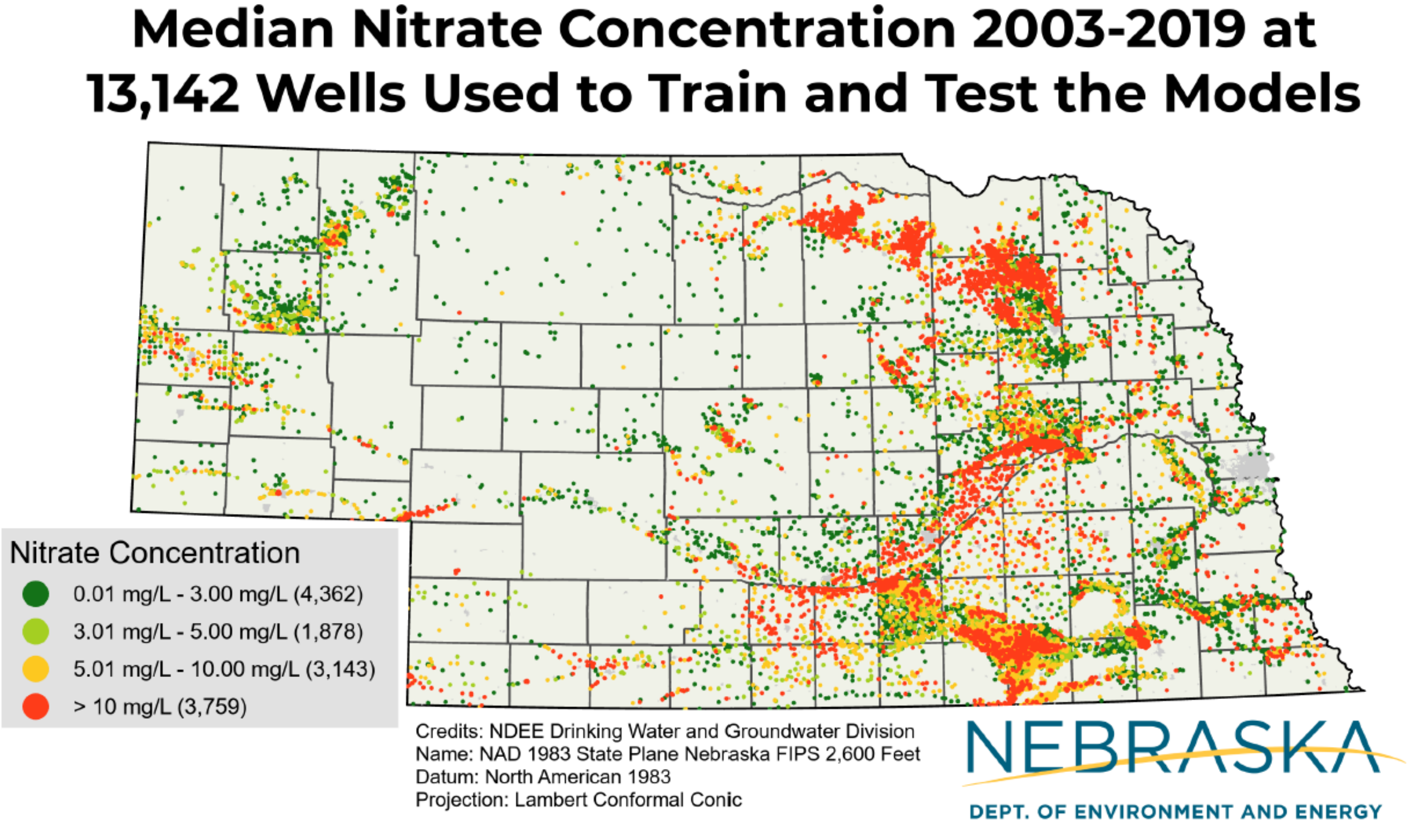
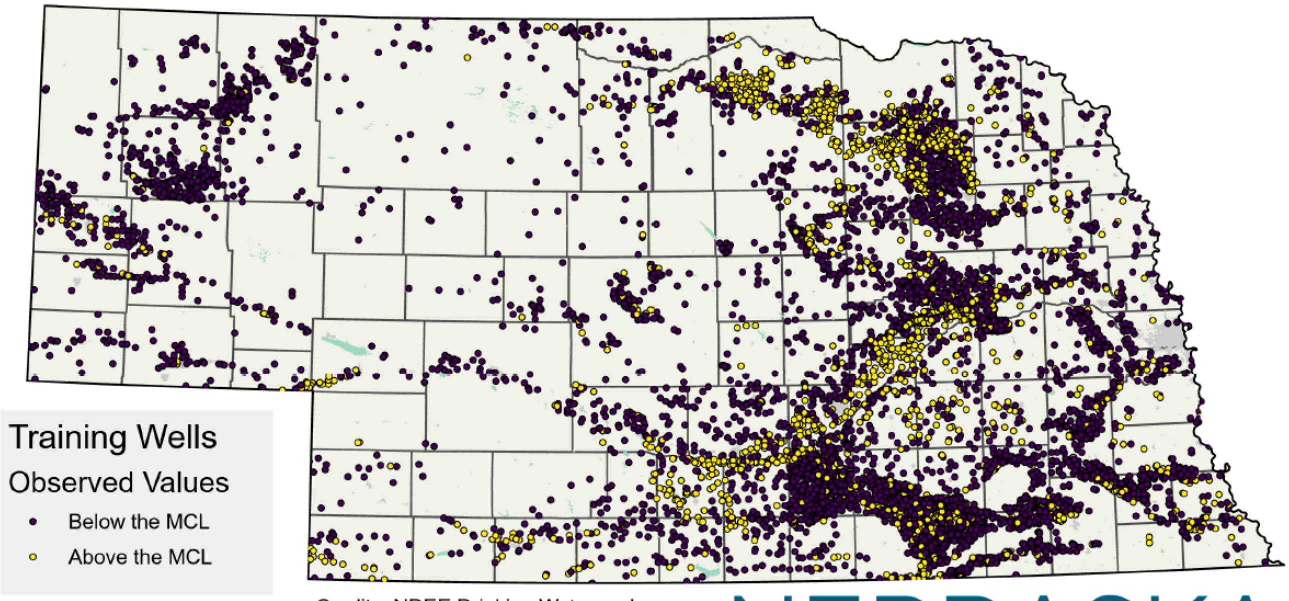

Median Nitrate Concentration 2003-2019 at
13,142 Wells Used to Train and Test the Models
Nitrate Concentration
● 0.01 mg/L - 3.00 mg/L (4,362)
● 3.01 mg/L - 5.00 mg/L (1,878)
● 5.01 mg/L - 10.00 mg/L (3,143)
● > 10 mg/L (3,759)
Credits: NDEE Drinking Water and Groundwater Division
Name: NAD 1983 State Plane Nebraska FIPS 2,600 Feet
Datum: North American 1983
Projection: Lambert Conformal Conic
NEBRASKA
DEPT. OF ENVIRONMENT AND ENERGY

Figure 2. Predictive Nitrate Model Input: All Well Locations Used to Train and Test Each Model by Median Nitrate Concentration.

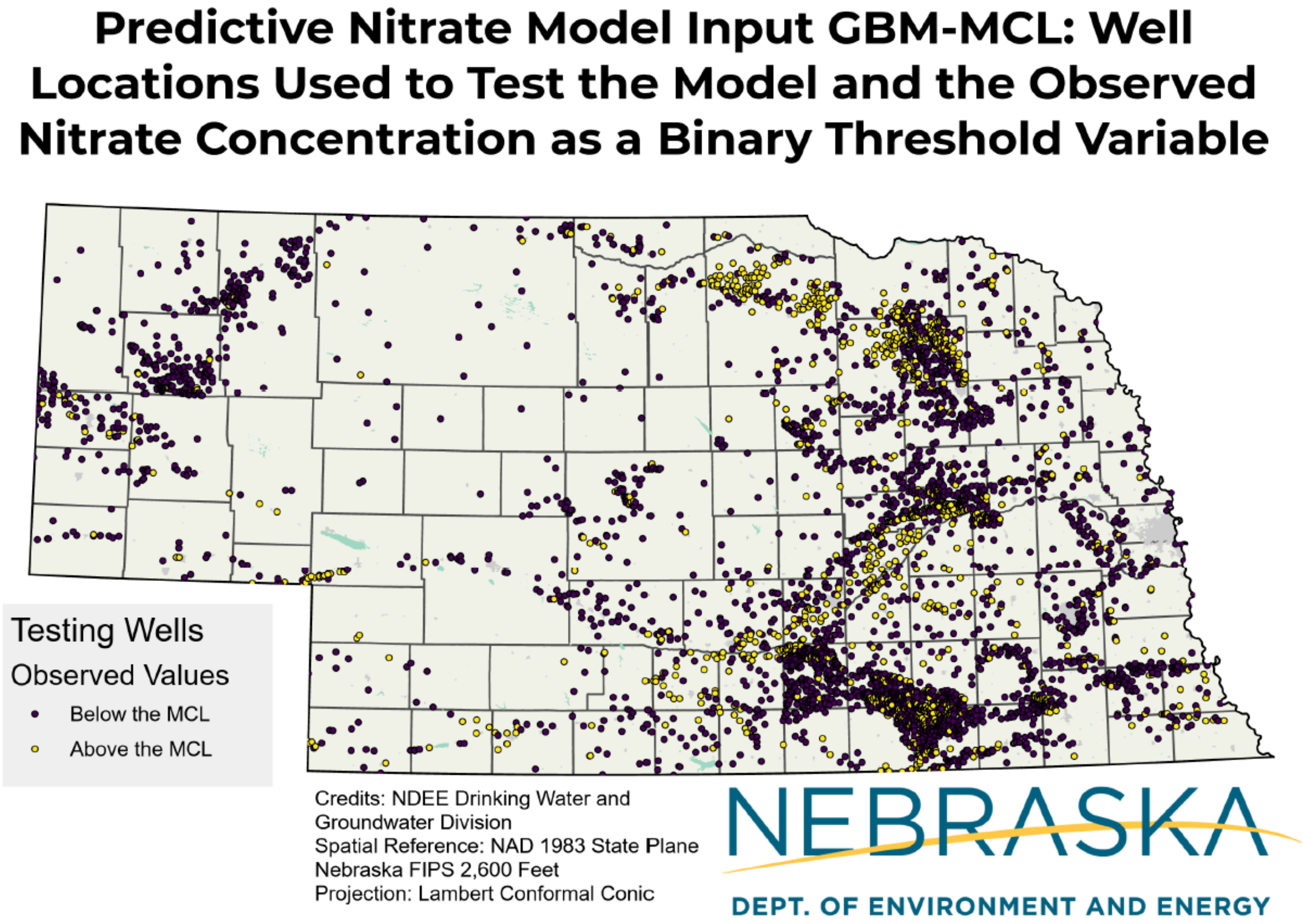


Predictive Nitrate Model Input GBM-MCL: Well Locations Used to Train the Model and the Observed Nitrate Concentration as a Binary Threshold Variable

Credits: NDEE Drinking Water and Groundwater Division
Spatial Reference: NAD 1983 State Plane Nebraska FIPS 2,600 Feet
Projection: Lambert Conformal Conic

NEBRASKA
DEPT. OF ENVIRONMENT AND ENERGY

Figure 3. Predictive Nitrate Model Input GBM-MCL Well Locations Used to Train the Model and the Observed Nitrate Concentration as a Binary Threshold Variable.



Predictive Nitrate Model Input GBM-MCL: Well Locations Used to Test the Model and the Observed Nitrate Concentration as a Binary Threshold Variable

Credits: NDEE Drinking Water and Groundwater Division
Spatial Reference: NAD 1983 State Plane Nebraska FIPS 2,600 Feet
Projection: Lambert Conformal Conic

Figure 4. Predictive Nitrate Model Input GBM-MCL Well Locations Used to Test the Model and the Observed Nitrate Concentration as a Binary Threshold Variable.

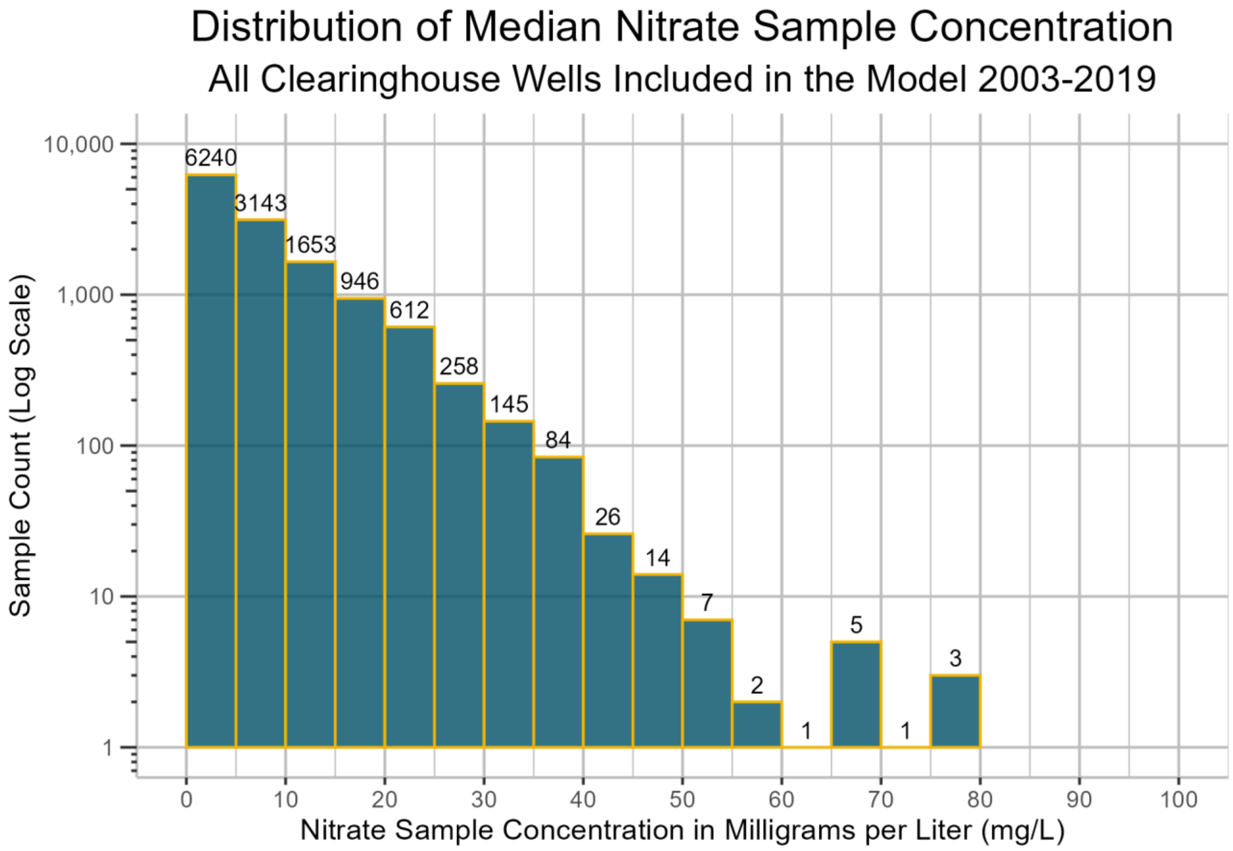


Figure 5. Median Nitrate Concentration Distribution at Wells Included in the Model.

Because of siting, design, and sampling requirements, Public Water Supply (PWS) wells are biased to lower nitrate levels and were excluded from modeling. See Table 4 for a summary of the well sample data in the Clearinghouse organized by well type. It shows the wide range in sampling patterns and concentrations between well classes. This range in concentrations between well types shown in Table 4 can be explained in part by the more stringent construction standards for PWS wells than other types of wells. Additionally, PWS wells must meet SDWA standards and those that do not are typically decommissioned, blended, or treated. This biases the PWS data toward lower nitrate concentrations overall. Sample data from PWS wells were excluded from training and testing data because it is not representative of nitrate levels in private domestic wells, the target of this modeling effort.

Table 4. Summary statistics for nitrate samples in the clearinghouse by well type from 2003 to 2019.

Clearinghouse Well Type	Mean Nitrate Concentration (mg/L)	Median Nitrate Concentration (mg/L)	Sample Count	Wells Sampled
Livestock Watering	12.43	8.40	522	105
Domestic	7.21	2.50	5,676	1,423
Irrigation	9.35	6.50	51,969	13,504
Monitoring	7.33	4.10	19,021	1,697
Public Water System	4.04	2.94	42,631	3,064
All Wells	7.05	4.5	119,992	19,768

Well Construction

Well depth, pumping water level (PWL), static water level (SWL), half the distance to the screened interval, the presence or absence of a surface seal, the length of gravel pack, and construction year were factors evaluated against the nitrate concentration in all sampled wells. Construction year was assigned a binary variable corresponding to wells built before or after state construction standards were established in 1988 and was ultimately insignificant in modeling. Only wells 300 feet or shallower were modeled. Domestic wells in Nebraska do not typically exceed 300 feet in depth. Figure 6 shows the depth of active, registered domestic wells in the state. All Clearinghouse samples included in the model are plotted against half the depth to the screened interval in Figure 7. Well depth, depth of the screened interval, PWL, and SWL, are all proxy measures for an important nitrate predictor: groundwater age (Nolan et al., 2015; Wells et al., 2018; Malakar et al., 2023). Depth to the screened interval shows a negative relationship with nitrate concentration.

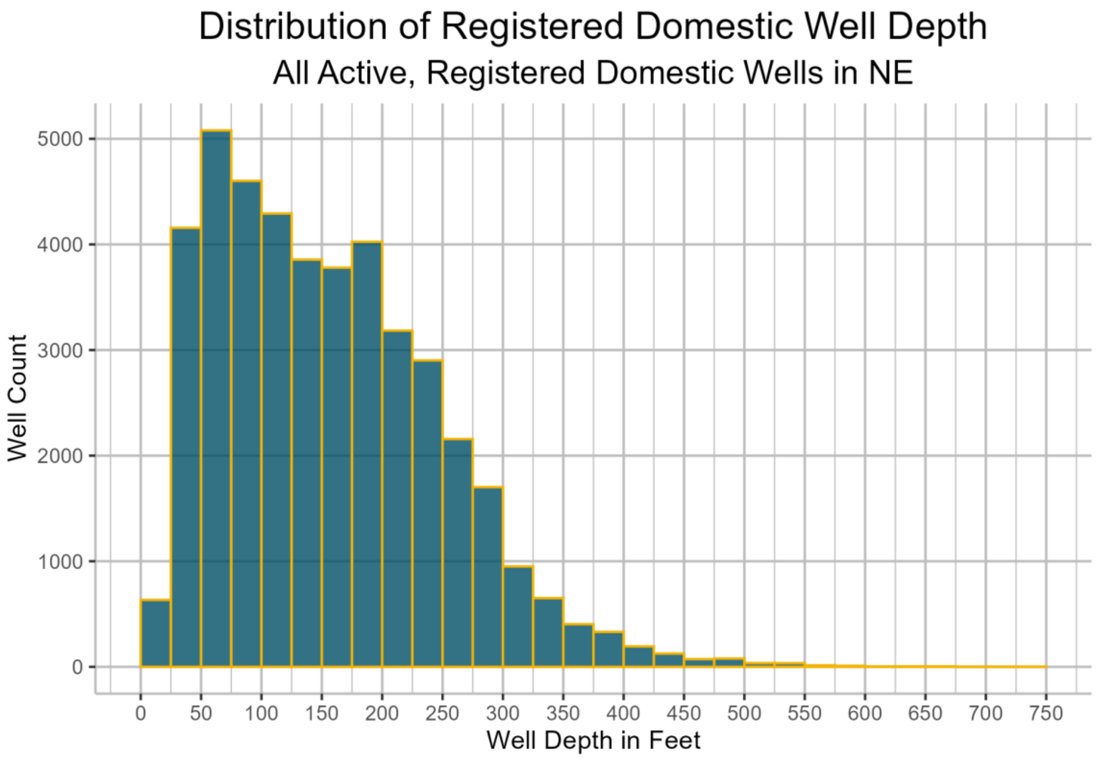


Figure 6. Distribution of Domestic Well Depth Among Active, Registered Wells in Nebraska.

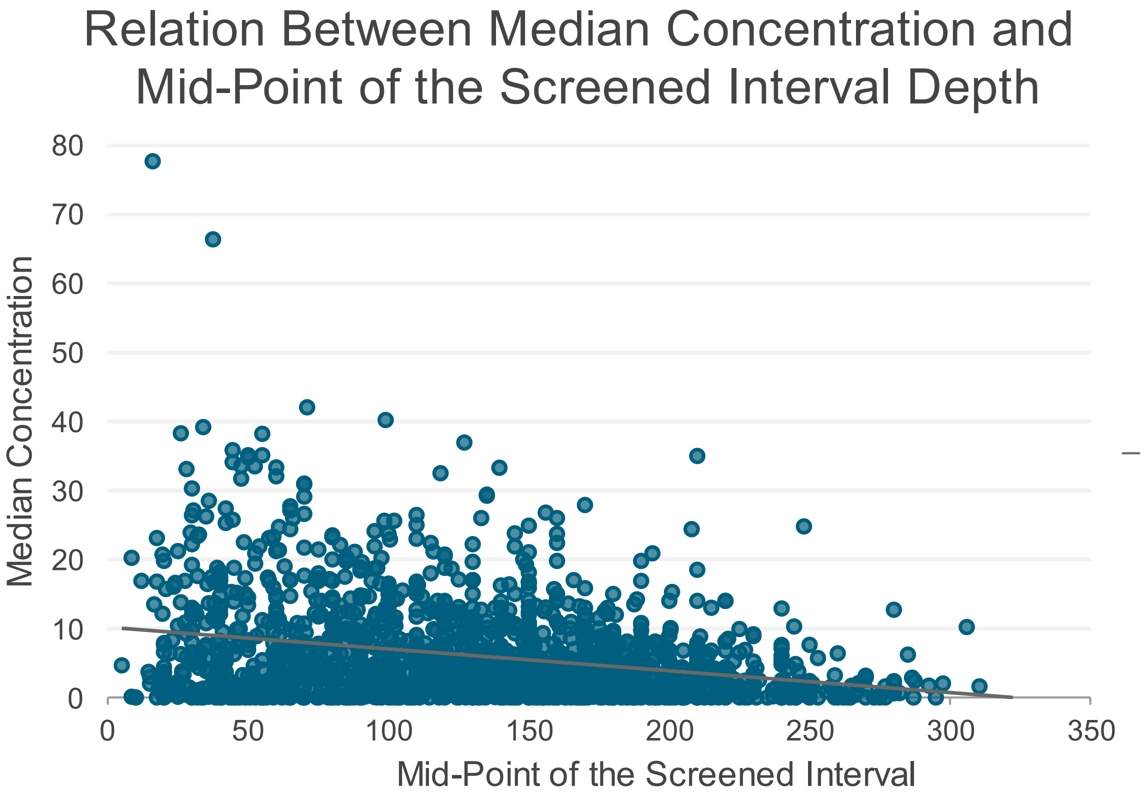


Figure 7. Nitrate Concentration in Milligrams per Liter (mg/L) and Depth to the Mid-Point of the Screened Interval in Feet (ft).

Land-Use

The CDL was evaluated for changes over time at the Township scale across the state (Figure 8) using the R package ggplot2 (Wickham, 2016). Significant changes were not seen between the major classes (grassland, corn, soy) over the study area since the initial release of the 2008 product. Additionally, around 80% of corn and soy acres appear to be in rotation with each other over the study period and these two should be considered a linked class. Corn shows similar correlation with the percentage of irrigated land and the historic fertilizer application rate. See a correlation matrix of the land use inputs in Figure 9 where Pearson's R values are plotted on the right diagonal (Sokal et al., 1969). A combined CornSoy class is shown illustrating the close relationship between the two factors. While factor independence is not a required assumption of BRTs (Elith et al., 2008), collinear factors do influence the variable influence and interactions in the model (Dormann et al., 2013).

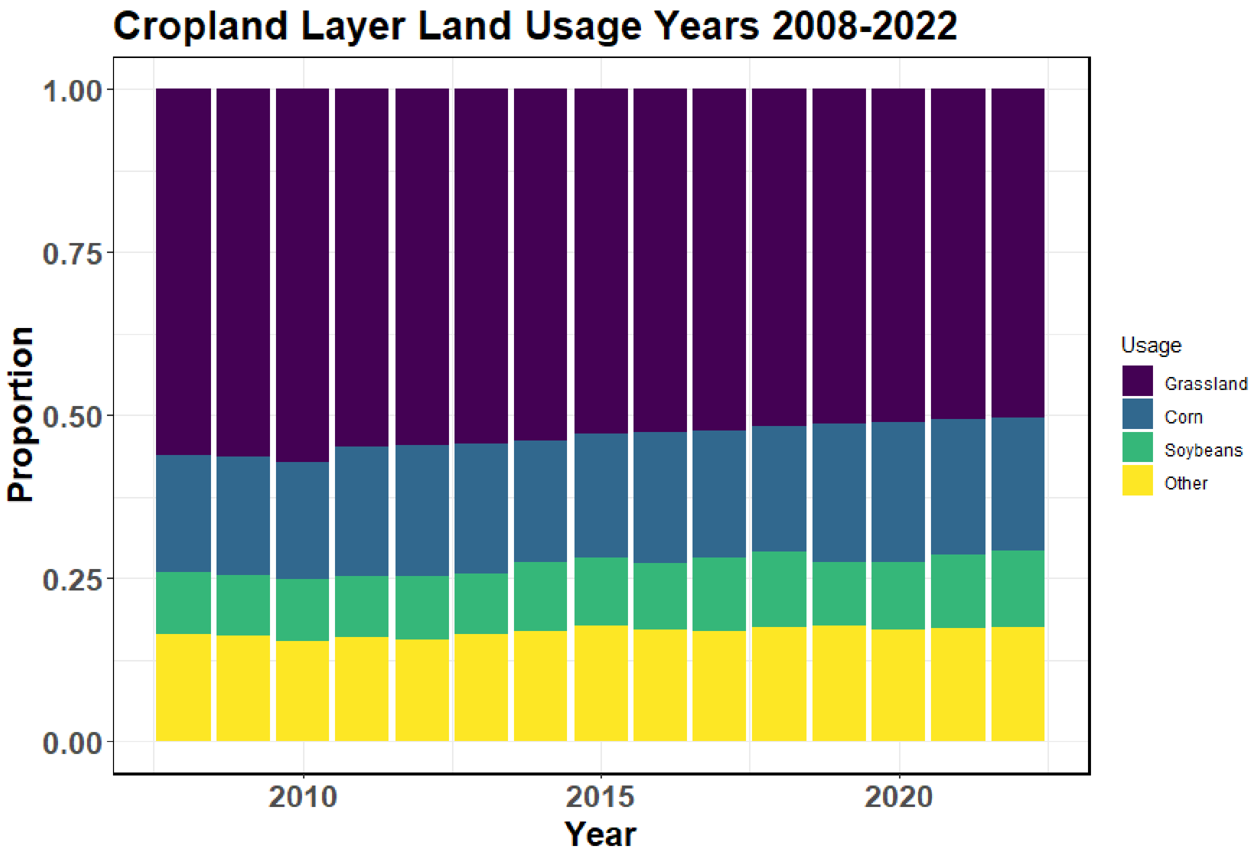


Figure 8. Cropland Data Layer Largest Land-Use Classes.

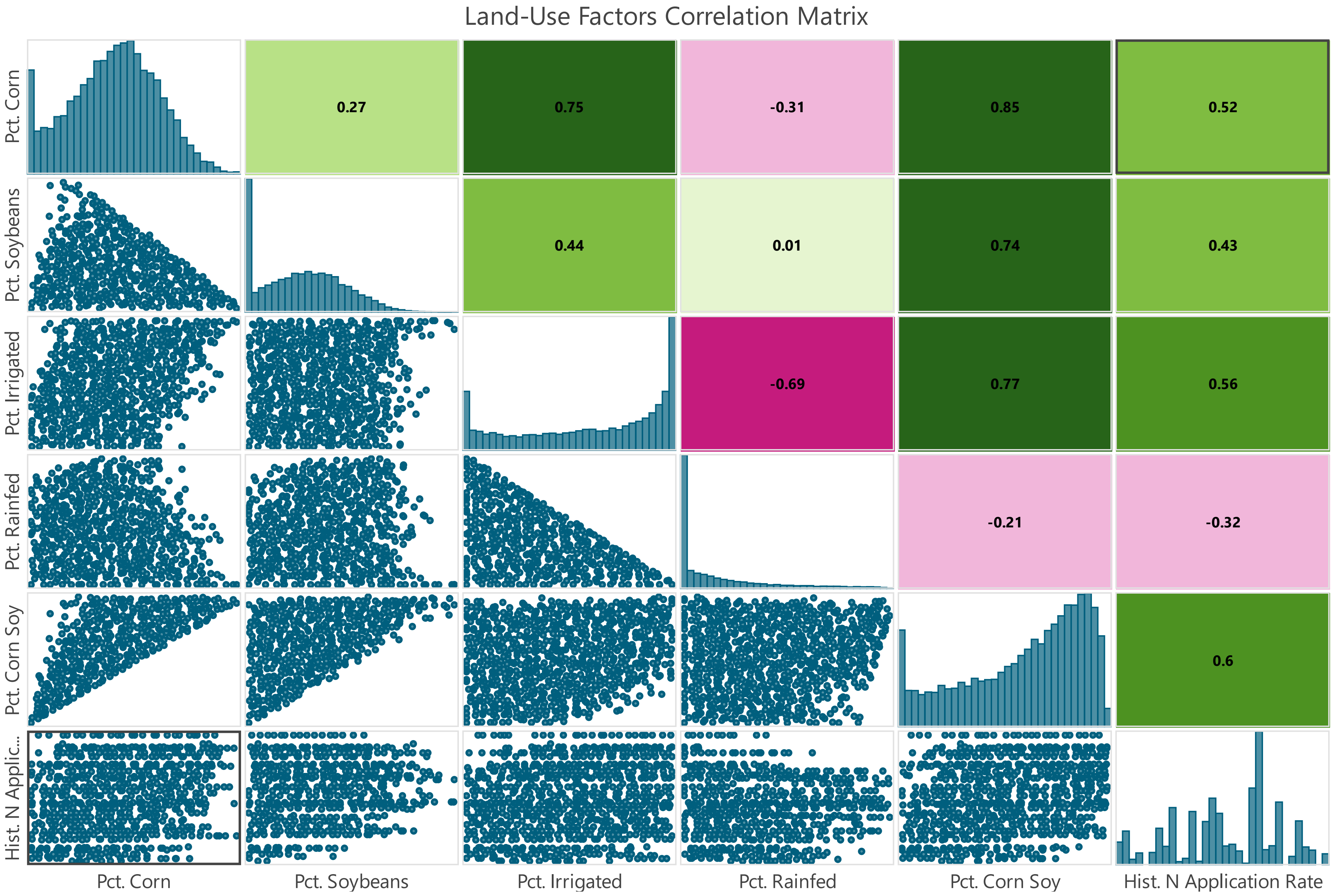


Figure 9. Correlation Matrix for Land Use Variables Considered for or Included in the Model.

Results & Discussion

Model predictions in probability terms were converted to binary values and compared with the training and testing data. Probability values of above 0.5 were treated as a prediction of one and equal to or below 0.5 as zero for mapping and confusion matrix calculations.

Models were tuned by varying the learning rate (from 0.001 to 0.1), tree complexity (from 3-5), and number of trees (from 100-15,000). Once a rough number of trees was established, values of complexity and learning rate were optimized. As a rule of thumb, when the complexity was increased by one, the learning rate was reduced by approximately one-half. Complexity was varied between three and five. A complexity of five was found to optimize all models. Unsurprisingly, optimal complexity was near the square root of the number of predictor variables (22). See table 5 for a summary of the model parameters after tuning.

Table 5. Model Parameters Used in each Tuned BRT Model.

Model	Number of Trees	Tree Complexity	Learning Rate
GBM-MCL (MCL)	8600	5	0.008
GBM10.18 (Elevated)	12100	5	0.005
GBM10.19 (Background)	13600	5	0.007

Table 6 summarizes diagnostic statistics for the models separated by training and testing data. Model sensitivity for the training data ranged between 76-97% with the highest sensitivity for classifying wells above the background level. For the testing data, sensitivity ranged from 55-88%. Again, the highest sensitivity was achieved classifying wells above the 3 mg/L background. Overall accuracy was high across the board, ranging from 75-91% in the testing and training data.

MCC values evaluated from the training data ranged between 0.74 and 0.82 indicating very strong (0.7 – 1.0) agreement between predictions and observations. In the testing data, MCC values ranged from 0.50 – 0.51 indicating strong agreement (0.4 – 0.69) between testing data and model predictions.

Table 6. Model Diagnostic Statistics.

Model Diagnostic Statistics Model	Specificity (0)	Sensitivity (1)	Accuracy	MCC
	Training Data			
GBM-MCL (MCL)	97%	76%	91%	0.78
GBM10.18 (Elevated)	86%	88%	87%	0.74
GBM10.19 (Background)	82%	97%	92%	0.82
	Testing Data			
GBM-MCL (MCL)	92%	55%	81%	0.51
GBM10.18 (Elevated)	72%	78%	75%	0.51
GBM10.19 (Background)	59%	88%	78%	0.50

Variable influence is plotted in Figure 10 for the GBM-MCL model. Variable influence, partial dependence, and variable interaction plots were roughly equivalent across models, results are reported for GBM-MCL and are representative of the other models. Across models the most influential factors were well location (lat/long) and soil infiltration rate (ksat). Variables with many null values – like depth to the midpoint of the screened interval (Half_ScreenDepth) – have lower influence. Partial dependence plots for the eight most influential factors in the GBM-MCL model are shown in Figure 11.

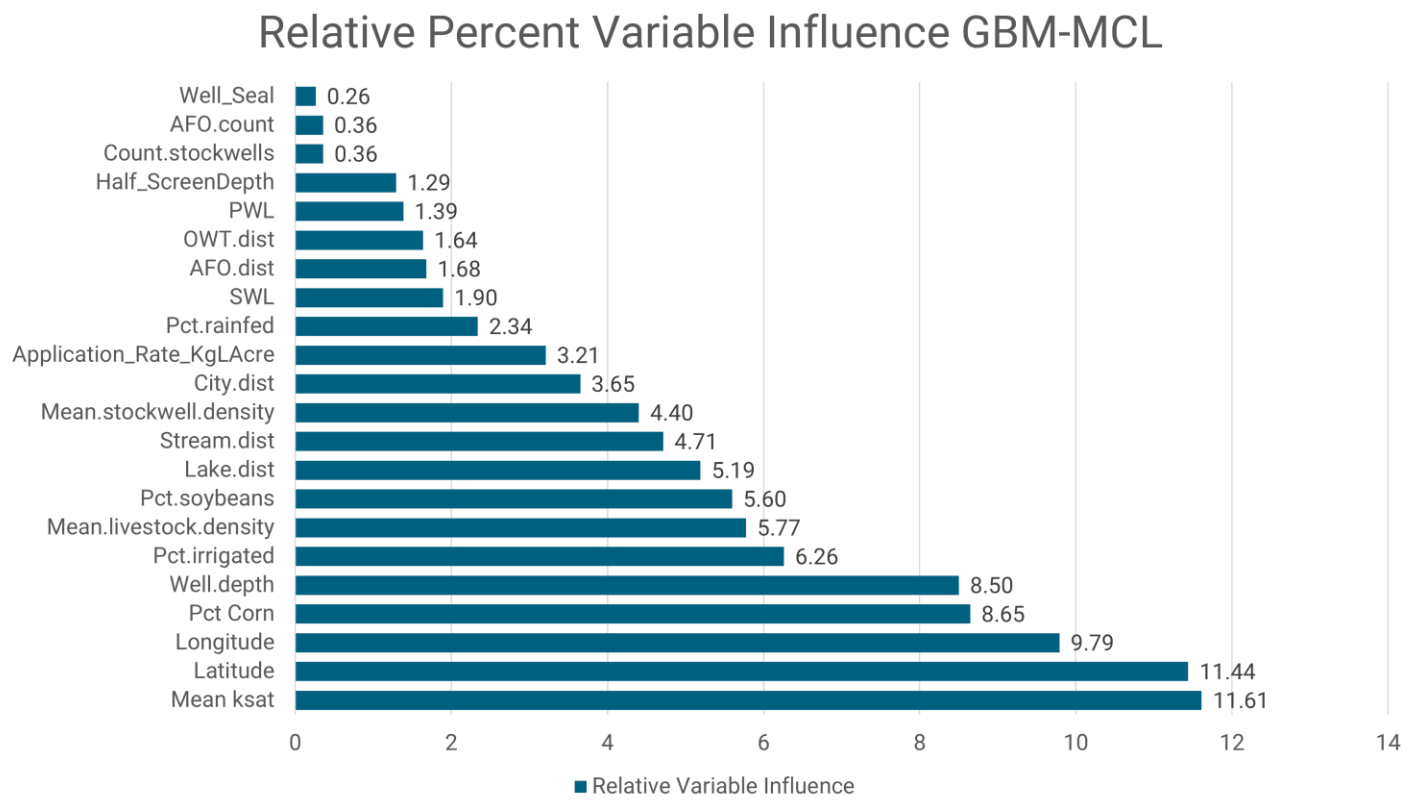


Figure 10. Relative Percent Variable Influence for the 22 Predictor Variables in the GBM-MCL Model.

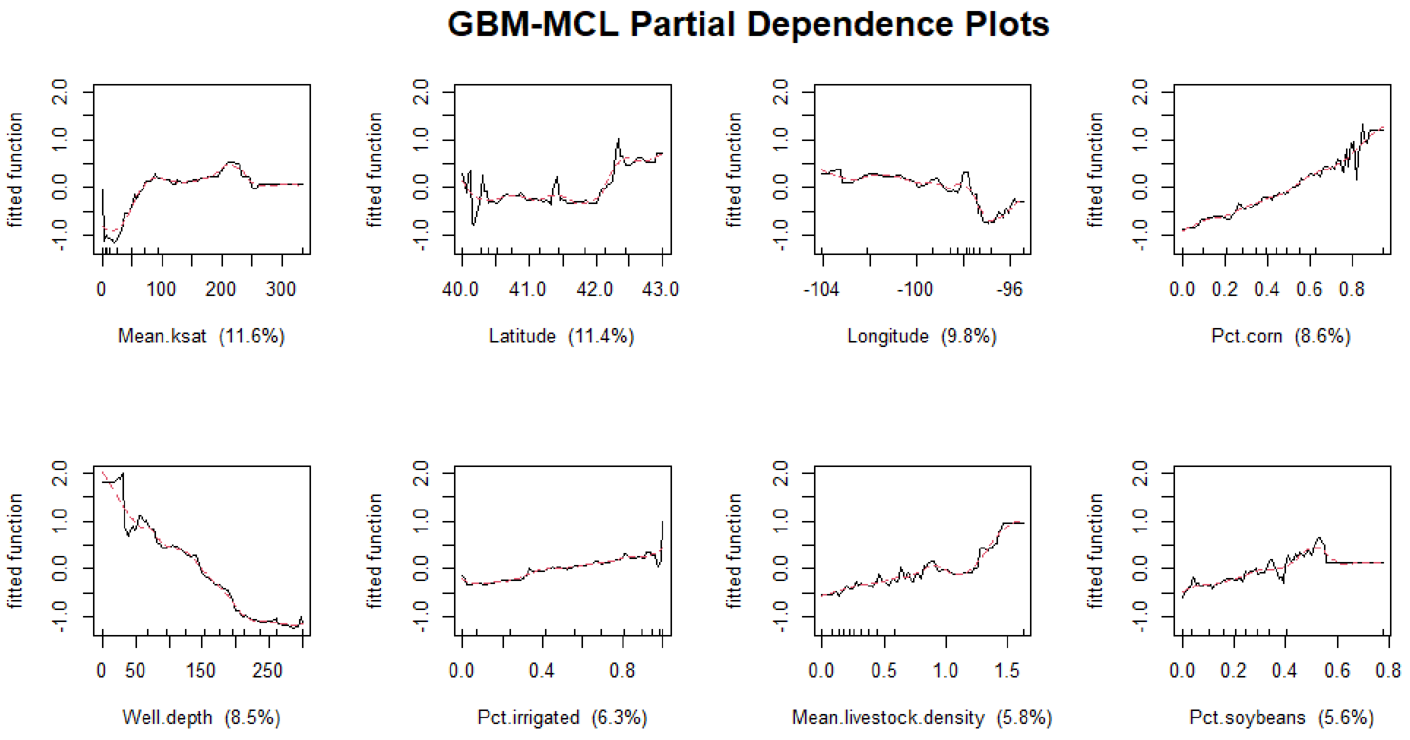


Figure 11. Partial Dependence Plots for the Eight Most Influential Variables in GBM-MCL.

Based on the variable influence and partial dependence plots, intensive agricultural land-use is a strong predictor of nitrate risk. Corn and soy should be considered a linked class based on crop rotation patterns in

Nebraska and the shape of their response supports this. Location, as measured by latitude and longitude, was consistently identified across models as a strong predictor with about 20% variable influence. This is unsurprising, given historic land-use patterns are strongly linked to nitrate levels in groundwater. Soil infiltration was another factor closely linked to estimated nitrate risk as expected. In general, nitrate risk increases with increasing ksat, but plateaus after a steep increase between 50 – 100 micrometers/second. Well depth had a negative relationship with nitrate risk. Shallower wells were more likely to be classified as high-risk, with the highest risk estimated in wells 50 feet or less in depth.

Variable interactions between land use trends and soil infiltration rate appear strong in the GBM-MCL model predicting against the MCL. Figure 12 shows the surface plot of predicted probability (z-axis) based on ksat and the percentage of irrigated land around each well. Figure 13 shows this interaction in two-dimensions, where more intense color indicates a higher predicted probability.

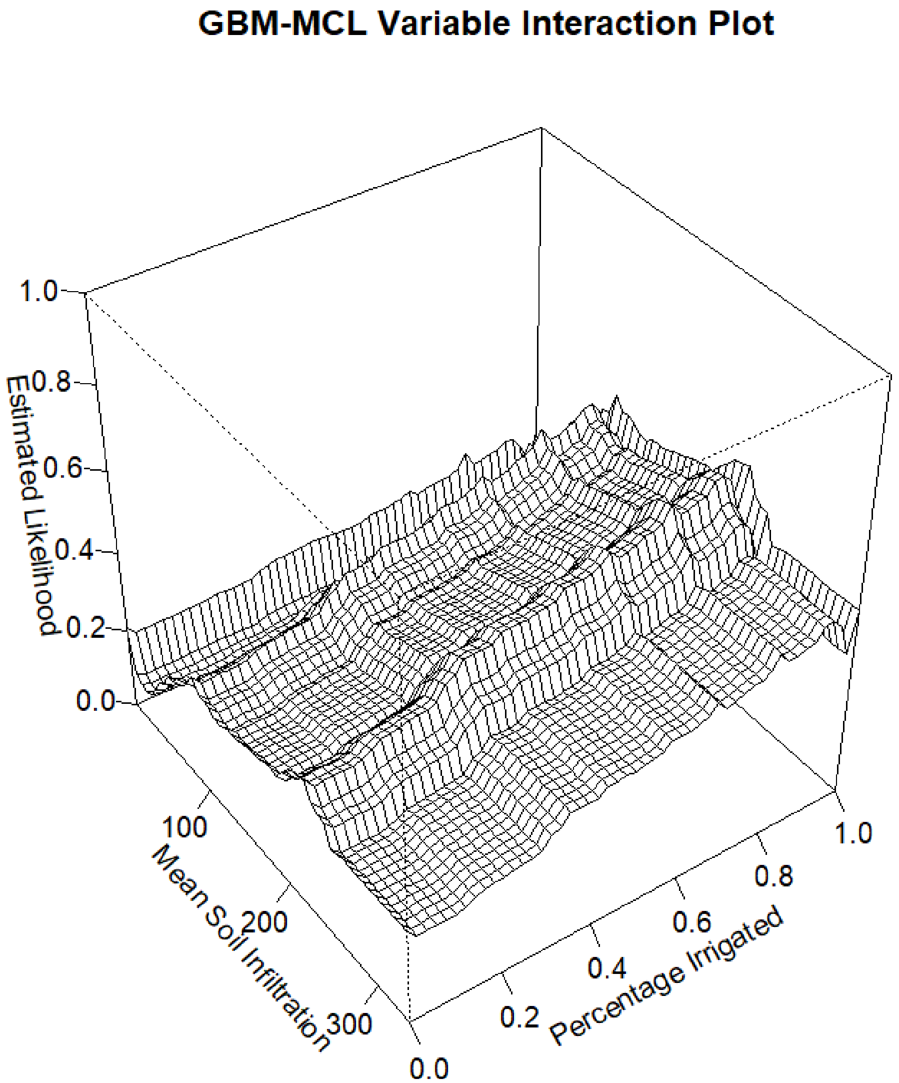


Figure 12. GBM-MCL Variable Interaction Surface Plot for Mean Soil Infiltration Rate and Percentage Irrigated Land.

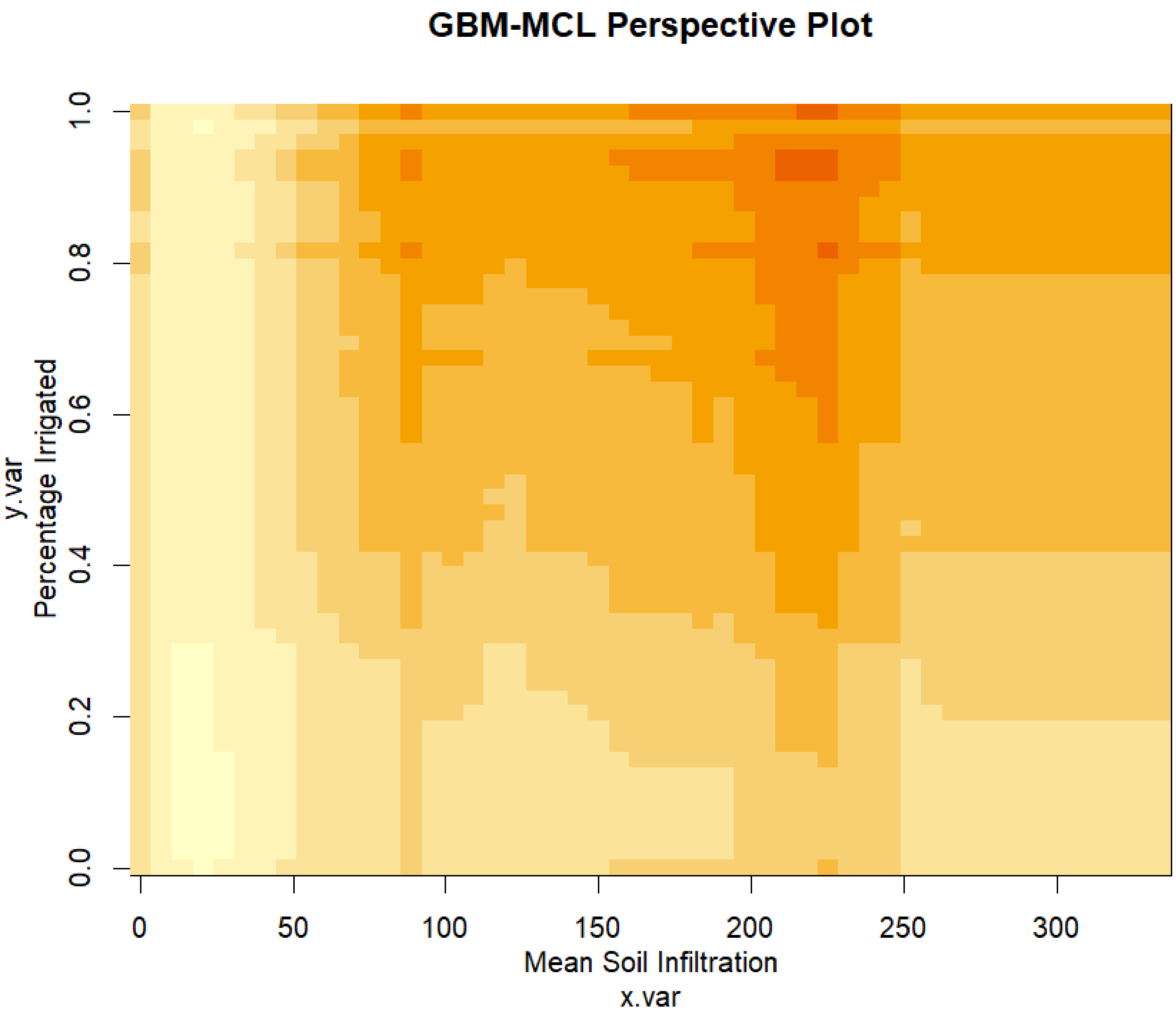


Figure 13. GBM-MCL Perspective Plot.

Physically, this interaction suggests that wells in heavily irrigated areas are at a higher risk of elevated nitrate levels if the soil infiltration rate is also above 50 micrometers per second. This relationship dips at high soil infiltration rates, suggesting soils that are unproductive for farming. It may also suggest that marginal soils on either end of the drainage spectrum receive comparably more inputs. Figure 14 plots the interaction between historic fertilizer rate and soil infiltration which supports this assertion. Correlation between other crop predictors such as corn, soy, historic fertilizer application, and irrigated land may be muting this relationship.

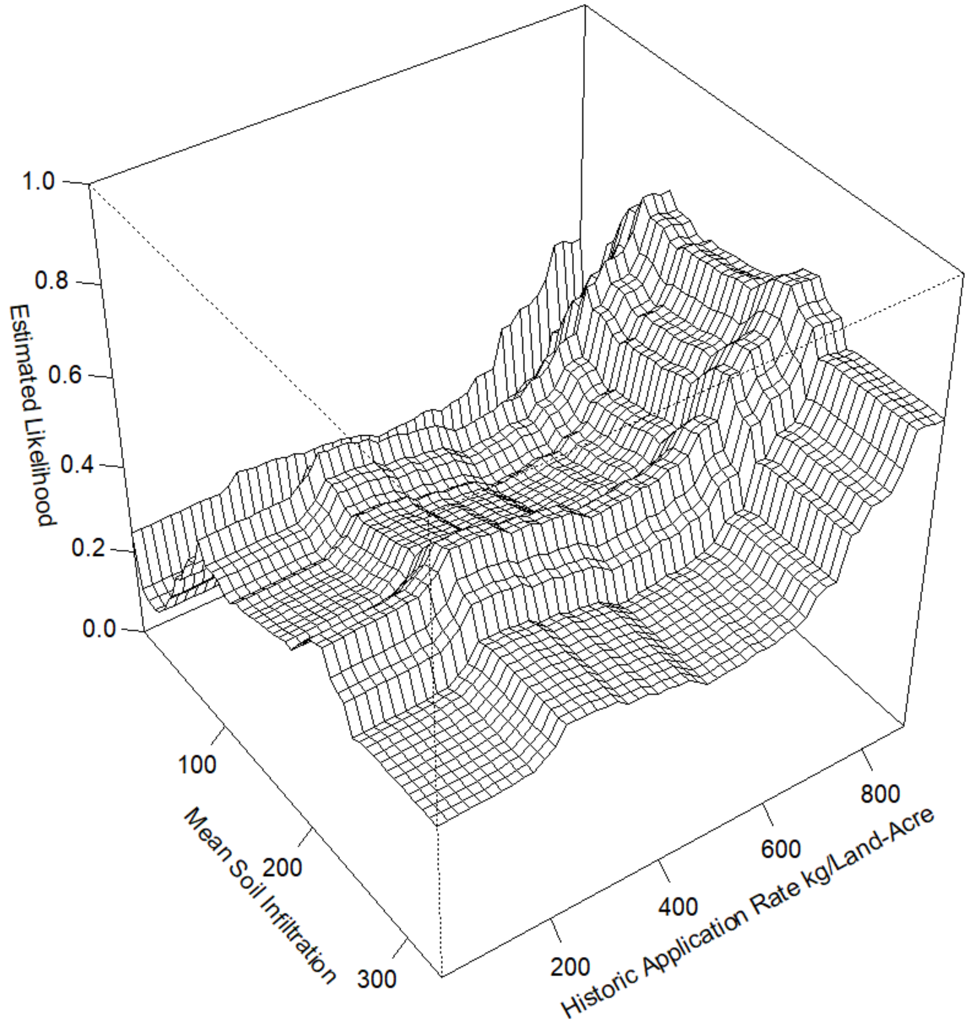


Figure 14. GBM-MCL Variable Interaction Plot Between Estimated Likelihood Soil Infiltration Rate and Historic Application Rate.

Model predictions were exported from R and mapped using ArcGIS Pro. Figure 15 plots the predicted and observed results from the GBM-MCL testing data. Light blue triangles are true positives, dark blue triangles are true negatives, orange x’s are false positives, and red x’s are false negatives. False negatives are rendered first in the figure, and it should be noted that at this scale the wells appear much closer together than they are. Mapping the results reveals that the model generalizes wells across Nebraska with the apparent exception of the Paleo Valley Aquifer systems.

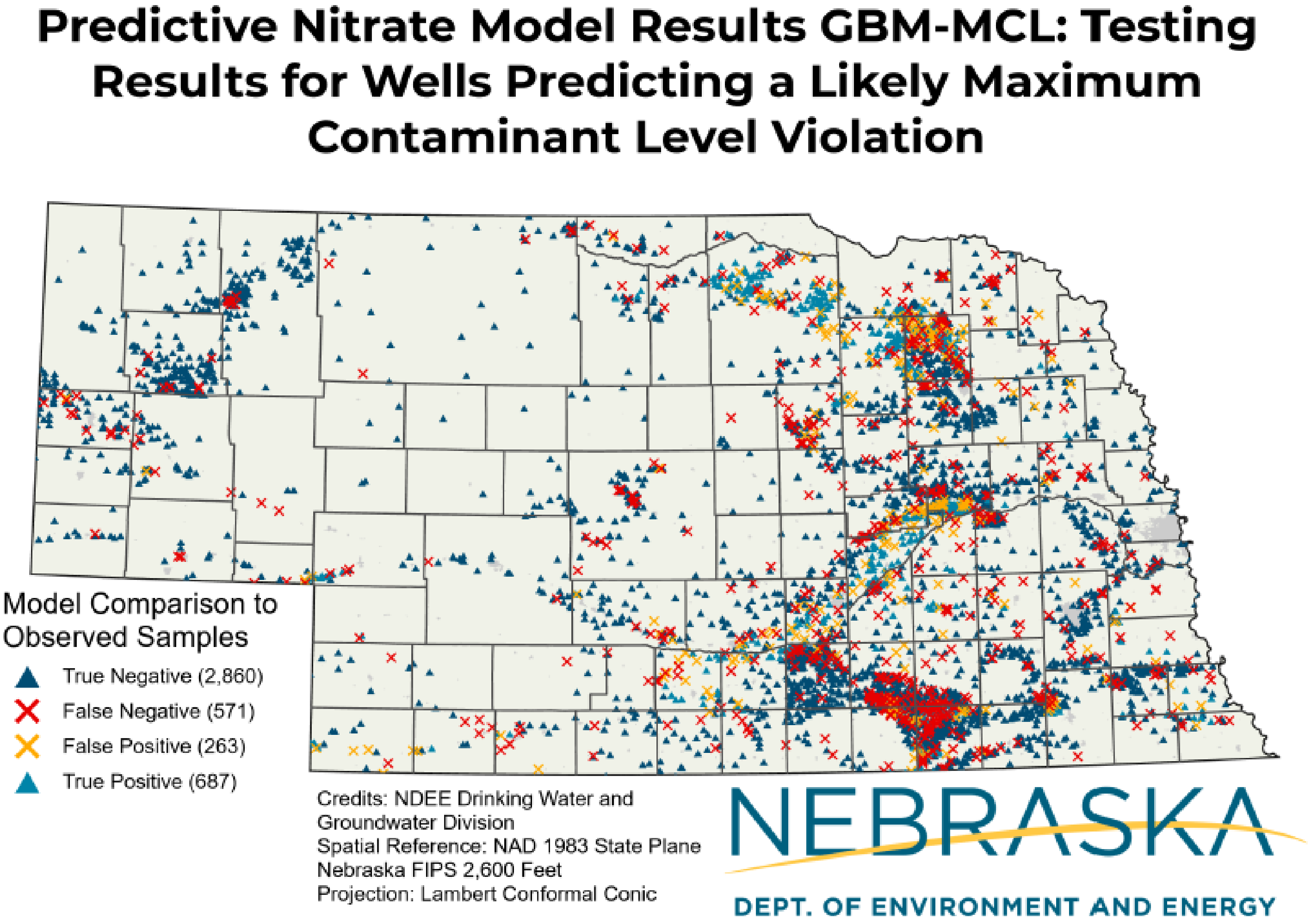
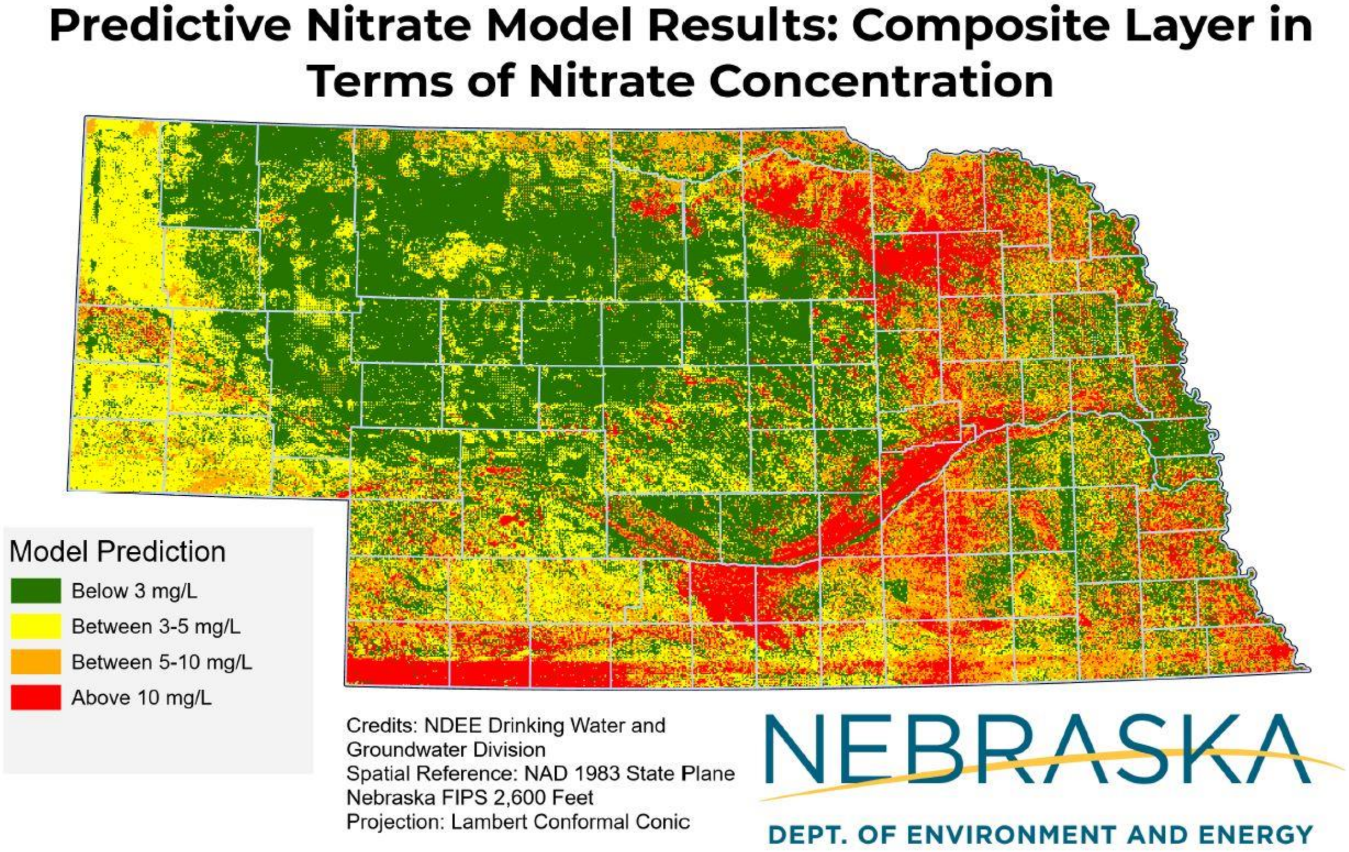

Predictive Nitrate Model Results GBM-MCL: Testing Results for Wells Predicting a Likely Maximum Contaminant Level Violation
Model Comparison to Observed Samples
▲ True Negative (2,860)
✕ False Negative (571)
✕ False Positive (263)
▲ True Positive (687)
Credits: NDEE Drinking Water and Groundwater Division
Spatial Reference: NAD 1983 State Plane Nebraska FIPS 2,600 Feet
Projection: Lambert Conformal Conic
NEBRASKA
DEPT. OF ENVIRONMENT AND ENERGY

Figure 15. GBM-MCL Testing Results.

Generalizing Model Results for the NDEE Domestic Well Risk Assessment Tool

Results from the three trained models were generalized across the state. First, predictor variables were aggregated into a half-mile grid surface covering the State of Nebraska, then the trained models were used to predict to that surface, and finally the predictions were mapped for incorporation into the tool. Figure 16 shows the composite model results as they are queried by the GIS tool. Areas in red are more likely than not to exceed the MCL, areas in orange are more likely than not to exceed the elevated concentration, areas in yellow are more likely than not to exceed the background concentration, and areas in green are more likely than not to fall below the background concentration. A fully independent set of testing data from the NDEE free private domestic well sampling effort was used to evaluate the performance of the gridded model predictions.


Predictive Nitrate Model Results: Composite Layer in Terms of Nitrate Concentration
Model Prediction
🟩 Below 3 mg/L
🟨 Between 3-5 mg/L
🟧 Between 5-10 mg/L
🟥 Above 10 mg/L
NEBRASKA
DEPT. OF ENVIRONMENT AND ENERGY

Credits: NDEE Drinking Water and Groundwater Division
Spatial Reference: NAD 1983 State Plane Nebraska FIPS 2,600 Feet
Projection: Lambert Conformal Conic

Figure 16. Predictive Nitrate Model Results: Composite Layer in Terms of Nitrate Concentration.

Model Predictions Compared to NDEE Private Domestic Sampling Effort Results

In addition to testing data (data not used in model training), model performance was evaluated against samples collected by private domestic well owners as a part of the 2023-2024 NDEE water quality study free private domestic well nitrate sampling effort. Because these were not used to train the models, they represent a fully independent testing set. This test estimates how well the model predictions generalized to the half-mile grid surface will perform in the GIS tool. Table 7 summarizes the metrics for each surface. Results are also plotted in Figure 17 to visualize how the model results compare to the independent dataset.

Based on testing, it was found that aggregating the variables into an arbitrary grid 'weakened' model efficacy by generally under-estimating probability of exceeding each threshold concentration when compared to the fully independent private domestic well data (MCC=0.13 – 0.28). Classification thresholds were systematically reduced (in increments of 0.05) on the grid surface to optimize predictive accuracy and provide a conservative estimate of risk. By adjusting the cutoff value from 0.5 to 0.25 for the MCL model, from 0.5 to 0.35 for the elevated model, and from 0.5 to 0.45 for the background model, comparison to the private domestic well samples were acceptable (MCC=0.20 – 0.28) and more in line with a comparison between the gridded surface and the testing data (MCC=0.40 – 0.44).

Figure 17 shows a comparison between the MCL predictions and the private domestic samples by outcome. Location information for these wells is based on the street address provided by well owners and then geocoded using ArcGIS Pro (Version 3.1). Addresses matching P.O. boxes, Points of Interest, and Street centerlines were removed for evaluation metric calculations. Addresses that requested more than one sample kit were also removed as some owners tested before and after treatment units or for multiple properties. Areas in darker blue on Figure 17 correspond to higher probability of exceeding the MCL. Figure 18 summarizes the results of the MCL comparison with the domestic samples by classification category. Figures for the elevated and background surfaces are available in the supplemental model material.

Table 7. Diagnostic Statistics based on Comparing the Gridded Model Predictions to the Domestic Samples from the 2023-2024 Domestic Well Sampling Effort.

Metric → **Model-Surface ↓**	**Specificity (0)**	**Sensitivity (1)**	**Accuracy**	**MCC**
	2024 Domestic Well Testing Data			
GBM-MCL (MCL)	87%	34%	79%	0.20
GBM10.18 (Elevated)	75%	52%	68%	0.26
GBM10.19 (Background)	68%	60%	65%	0.28

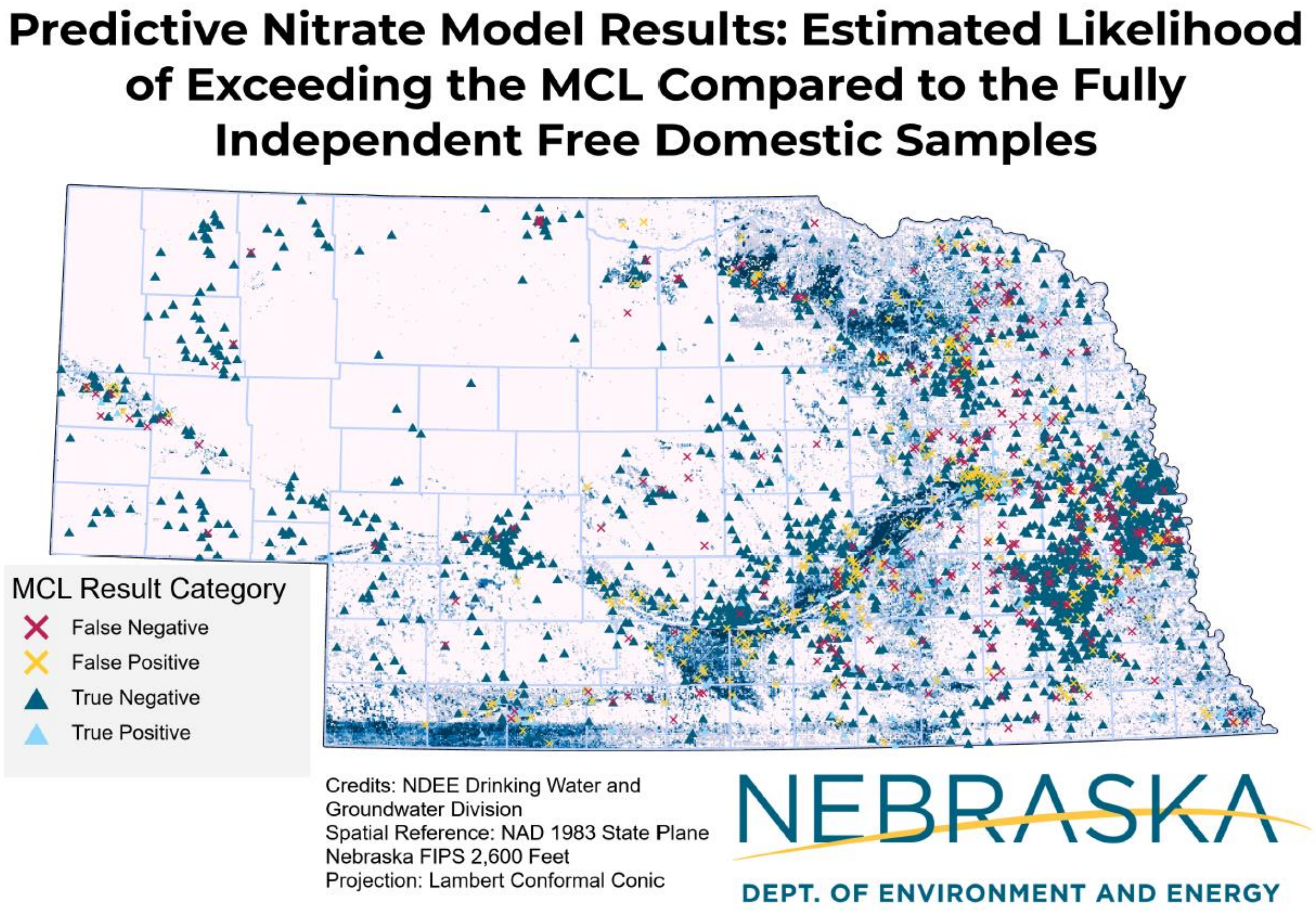

Predictive Nitrate Model Results: Estimated Likelihood of Exceeding the MCL Compared to the Fully Independent Free Domestic Samples
MCL Result Category
✕ False Negative
✕ False Positive
▲ True Negative
▲ True Positive
Credits: NDEE Drinking Water and Groundwater Division
Spatial Reference: NAD 1983 State Plane Nebraska FIPS 2,600 Feet
Projection: Lambert Conformal Conic
NEBRASKA
DEPT. OF ENVIRONMENT AND ENERGY

Figure 17. Predictive Nitrate Model Results: Half-Mile Grid Surface for GIS tool.

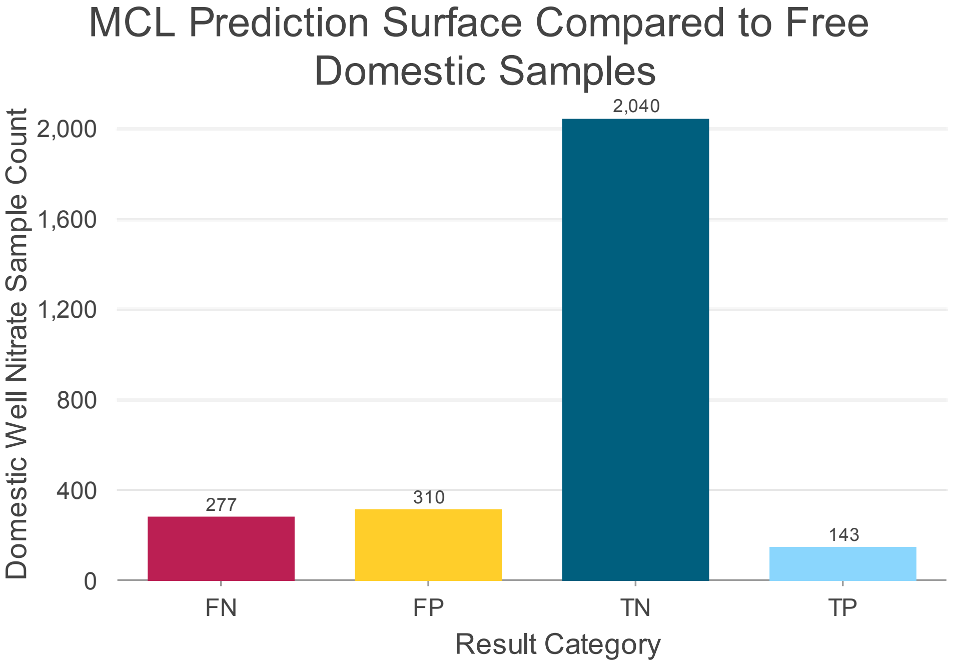


Figure 18. MCL Prediction Surface Compared to Domestic Well Samples by Result Classification.

Limiting Training Data

While BRTs are robust to null values for predictor variables and strongly related predictors, the influence of variables with missing values is systemically lowered. Preliminary testing that limited the dataset to only complete records and a limited number of collinear factors yielded interesting results (Figure 19). Depth to the midpoint of the screened interval became much more influential to the model predictions. Land-use influence was distributed mostly to the percentage of irrigated lands and the historic application rate. Ksat and well location were still key factors.

Modeling was repeated on a subset of the training data (n=1,559) where no factor contained null values. The results were overall in line with the models, with some differences in variable influence and interactions. For instance, in GBM-MCL no-NULL, the depth to the mid-point of the screened interval increased 900% in variable influence (Figure 19). This would indicate that screen depth is a stronger predictor than well depth, as expected, because it better estimates groundwater age. Figure 19 also illustrates the way that null values mute the variable influence measure. Low influence does not equal low impact. It may just indicate low coverage of a given predictor. Static water level and pumping water level increased by 200-250% by removing wells with null values from the training data.

Reducing the training set reduced the ability of the models to generalize, which was reflected in lower MCC values (0.41-0.45) when compared with the BRTs (0.5-0.51). Accuracy and sensitivity were also around 5% lower across the board for the testing data when using the limited training set. This is not particularly surprising given the data-hungry nature of ML methods and the steep reduction in class examples (from more than 13,000 to 1,559).

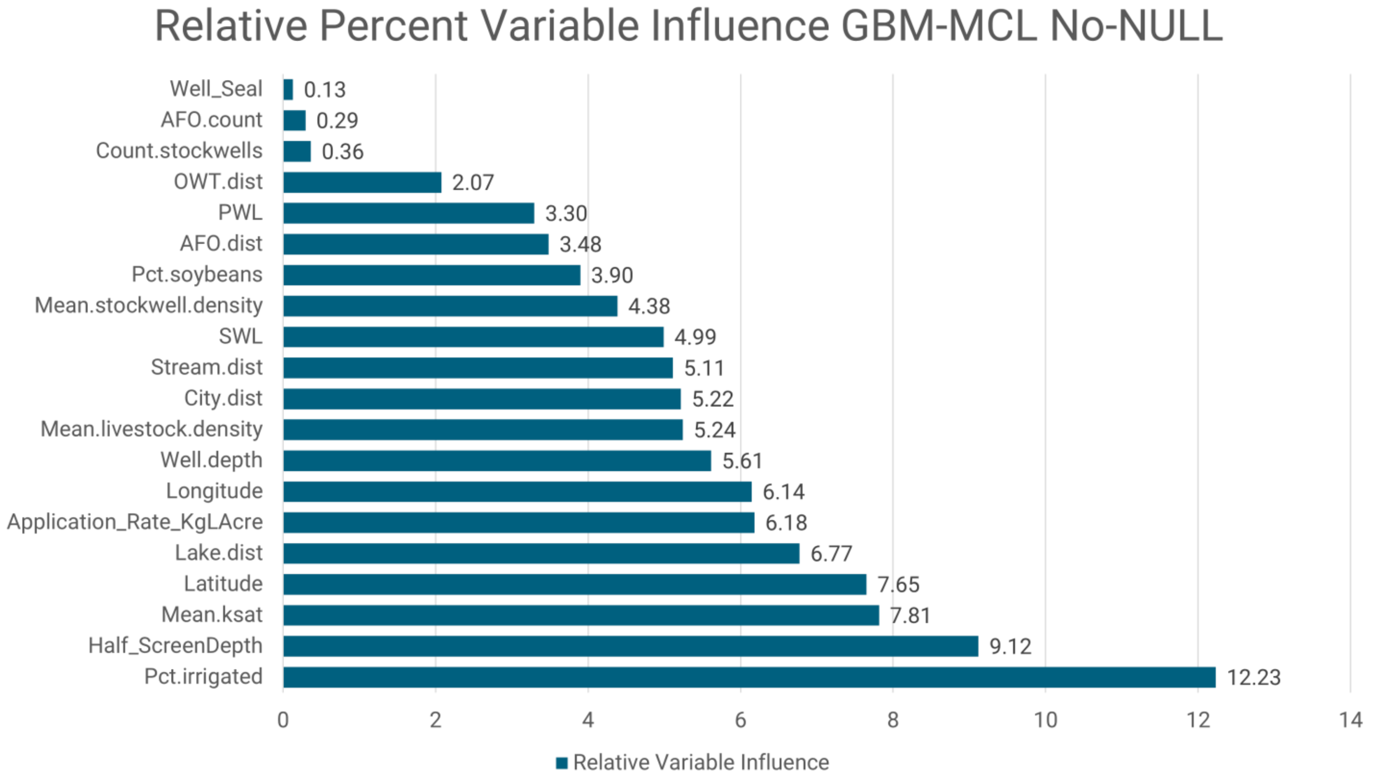


Figure 19. Relative Percent Variable Influence in the GBM-MCL Model with no NULL Factors.

Conclusions & Recommendations

In this study, three BRT models were trained to predict the probability that median nitrate concentration in a private domestic well would exceed water quality thresholds based on the available model inputs aggregated within 1500-meters of each well modeled. Input variables representing land-use, hydrogeologic factors, point sources, well construction details, and nitrate samples collected from 2003-2019 were used to train the model. Threshold values of 3 mg/L, 5 mg/L, and 10 mg/L were modeled to represent the background, elevated, and the SDWA MCL for nitrate, respectively.

Predictor variables used to train the BRTs were aggregated into a half-mile grid surface across Nebraska and model predictions were made for each grid cell to generalize models for the internal NDEE GIS tool. Predictions were only based on the input data listed in the model card, and it is important to note that this is not a complete accounting of the variables that could impact nitrate concentrations in a given well. Overall, the evaluation statistics for the models were strong, with MCC values between 0.5 and 0.51 for the testing data, and acceptable (MCC=0.20 – 0.28) when the gridded predictions were compared to the fully independent private domestic well sample set for use in the GIS tool. Overall model accuracy was high in the training and testing data (71 – 91%), and the estimates generalized acceptably to the private domestic well data collected during the 2023-2024 NDEE free private domestic well sampling effort (68 – 87% overall accuracy). Evaluation metrics from the training data indicate that models were well optimized, though additional predictor data, more complete predictor data coverage, and additional nitrate samples will almost certainly improve future modeling efforts.

False negatives (underprediction of nitrate concentration) are of greater concern than false positives for this study. NDEE intends to use the model internally and with key partners. Decision makers should note that a high false negative rate suggests the model generally underestimates risk of exceeding the MCL. GBM-MCL model sensitivity was 55%, with an overall accuracy of 81%. The testing specificity was very high – 92%, however the false negative rate of 45% could be improved with future work. Values compare favorably to those reported by Nolan (2014) and Wheeler and Nolan (2015) for nitrate prediction. GBM-MCL performance was comparable to Lombard (2021) in a similar study exploring arsenic MCL exceedances. From a regulatory perspective, it can be argued that predicting where nitrate is not likely to exceed the MCL is just as important as predicting where it is likely to exceed the MCL. Ultimately, testing the water is the only way to know the concentration for certain. When the predicted surface for the GBM-MCL model was compared to the private domestic well data collected during the water quality study (a fully independent test set) the results were acceptable, with 47% sensitivity and an overall accuracy of 73%.

Several insights about the relationship between nitrate and contributing factors can be taken away from the model. Variable influence and partial dependence plots show that well location, intensive agricultural land-use, irrigation, and high soil infiltration rates are strongly related to the level of nitrate in groundwater. This is consistent with other research demonstrating that shallow wells in high infiltration soils sited in areas where there is significant nitrogen surface loading are at the highest risk of nitrate contamination (Spalding and Exner, 1993; Spalding, 2001; Litke, 2001; Exner, 2014; Nolan et al., 2014; Davis et al., 2015; Wheeler et al., 2015; Garcia et al., 2017). At the scale of this modeling, point source impacts from nitrate were less influential to the nitrate risk of a given private domestic well than expected. However, it is nearly certain that variable aggregation, missing point source data, and data coverage issues play a large role in muting the signal from these sources. This modeling does not support the assertion that point sources do not impact nitrate levels in private domestic wells. Until recent decades, OWT systems like septic tanks were not permitted by the state or tracked. Like private domestic wells, there are likely thousands of unregistered OWT systems. Livestock operations data used to train the model only covers facilities regulated by Title 130, and as such is not a complete record of animal operations in the state. Smaller facilities, which are not permitted, are not included in the AFO data used to train the model. Livestock watering-well data may partially capture these facilities but is a proxy measure. Animal units, a common measure to generalize livestock counts across species, could be incorporated into future modeling to better reflect the size differences between operations, which directly impacts the loading rate.

Livestock operation density, nearby onsite treatment systems, and municipal boundaries were all significant factors, but were not as strongly predictive as crop variables or well location. The significance of well location to nitrate predictions is unsurprising, considering areas like the Upper Elkhorn River Basin have reported elevated nitrate concentrations since at least the 1970s (Spalding and Engberg, 1978), and increasing concentrations have been reported across the state as far back as 1930 (Litke, 2001; McMahon et al., 2007). Location reflects history, and the legacy nature of the nitrate issue in Nebraska. Location is also reflective of local geology and site conditions that likely boost its significance as a factor.

Future modeling efforts could better incorporate additional point source datasets which may more fully capture these impacts. Limited data on unregistered facilities discussed above, or facilities that do not require permitting, are likely reducing the impact of point source data in the model. This data gap could also be a contributor to the relatively high false negative rate (45%) of the GBM-MCL model. Data on release assessments and storage facilities for nitrogen precursors such as ammonia and fertilizer tier two facilities regulated by NDEE could be incorporated into future modeling efforts and may more completely capture the potential sources around each well.

Since 2010, an additional million acres of corn have been cultivated in the state and there is evidence that trends in fertilizer efficiency have plateaued (Ferguson, 2024). Future modeling could incorporate more classes from the CDL and explore other years to see what potential differences may arise in the data. Fertilizer and irrigation vary based on crop, and it is possible additional classes would improve the modeling. Across the state, the time it takes for this surface loading to reach groundwater varies from years to decades (Wells et al., 2018; Malakar et al., 2023). Future modeling efforts could better address the differences in transport time by broadly grouping wells based on soil characteristics as in Exner 2014. Another option would be to model wells in distinct groundwater regions, such as major aquifer units delineated by USGS or UNL CSD.

At time of the analysis – the public Clearinghouse for nitrate sample data does not have a complete sample record for the years 2020 to 2023. These models should be re-evaluated and trained when that data becomes available. BRTs, like other machine learning methods, benefit from large datasets (Breiman et al., 1984). While it is not expected that these data would change the variable influence or conclusions of this report, they would likely improve accuracy by virtue of providing more class examples to train models.

Well construction information was not available for all nitrate samples included in the analysis. While BRTs are robust to missing data, the variable influence is strongly impacted by missing values as seen in the discussion section. Future modeling could incorporate more of this well data. Variables representing meteorological factors, such as average annual precipitation, soil geochemical variables, and vadose zone transport rates have been valuable to other studies predicting water quality in domestic wells (Wheeler et al., 2015, Lombard et al., 2021). Additional SSURGO variables that could benefit the models include available water capacity, clay, sand, and silt content, soil organic matter, hydrologic group, drainage class, depth to bedrock geology, water table depths, conductivity, and pH value. Further work should incorporate this data into predicting nitrate levels in Nebraska. Annual precipitation data could come from the High Plains Regional Climate Center or NOAA. These data would improve model results and in turn the efficacy of the internal NDEE GIS tool.

Groundwater elevation was another variable not directly included in the model. However, variables such as well depth, static water level, pumping water level, and ½ the screened interval depth, indirectly capture the groundwater elevation. A groundwater elevation product, based on the regional groundwater models managed by NDNR, was developed as a part of the water quality study. Future modeling could incorporate this product directly. A more up-to-date accounting of fertilizer application rates and land-application sites could improve future modeling results. The Lower Loup Natural Resources District produced a GIS product that linked

Confined Animal Feed Operations (CAFOs) to their land application sites as permitted by NDEE. NDEE could explore developing a statewide product which could improve modeling efforts and management.

Vertical flow rates within each aquifer are influential to the stratification of nitrate concentrations in groundwater (Snow and Miller, 2018; Malakar et al., 2023). Transport rates could be derived from each of the regional groundwater models (as they are updated and improved) and have been valuable to other modeling efforts like Nolan 2015. As a model input, vertical transport rate through the aquifer would more directly capture the varying timescales it takes for nitrate to reach deeper groundwater across the state. Another option to reflect the varying transport rates would be group wells by hydrologic region or by aquifer and train regional models. This should be explored when the BRTs from this study are updated as additional data becomes available.

The state should consider developing regional groundwater models that incorporate a nitrogen cycle balance such as the one proposed by Garcia et al. (2019), using the coupled Community Multiscale Air Quality Bidirectional modeling system developed by USEPA and USDA Environmental Policy Integrated Climate (EPIC) agroecosystem model (Pleim and Ran, 2023). Such a system would allow for a much more detailed accounting of nitrogen at the surface for management efforts. This could help target efforts to lower concentrations in areas where groundwater is a primary source of drinking water.

Model accuracy is generally highest where there is more available data, such as in the river valleys and northeastern portions of the state. In areas with fewer nitrate samples, there are relatively fewer class examples to train the model on those locally relevant variables. The next section discusses recommendations for how model results should be incorporated into the internal NDEE GIS tool.

Determining Threshold Values and Risk Level

Based on the model performance, it is recommended to incorporate the results into the internal NDEE risk assessment GIS tool, with important caveats about the limits of these predictions. Each factor in the GIS tool is assigned points which are added together to determine an overall risk index. More points correspond with a higher risk. Recommended threshold values for incorporating the model into the tool are show in in Table 8. Points were assigned to the threshold values shown in Table 9 based on how model predictions relate to potential nitrate risk. Each model predicts where nitrate concentrations are likely to exceed the values in the table, and points are assigned based on each prediction. Threshold values reported in table 9 were determined based on: literature, model performance review, data from the free private domestic well sampling effort, and quality assurance procedures conducted by NDEE and project partners. As more data become available, the BRTs in this report should be updated and using the recommendations provided in the previous section. Updated models could improve the gridded surfaces in the GIS tool and reduce false negative rates.

Table 8. Threshold Values for the Predictive Nitrate Model Results Incorporated into the GIS Risk-Assessment Tool.

Predicted Nitrate Concentration Range (mg/L)	Points Assigned	Minimum Probability Predicted	Description
<3 mg/L	0	0.00	If the model predicts the input location has a probability of 0.25 or less of exceeding the MCL, a probability of 0.35 or less of exceeding 5 mg/L, and a probability of 0.45 or less of exceeding 3 mg/L, then the tool assigns zero points for this indicator. Language is provided to the user based on the model; it is likely that the nitrate level in their well is below background (less than 3 mg/L).
>3 mg/L	1	0.45	If the model predicts the input location has a probability of 0.25 or less of exceeding the MCL, a probability of 0.35 or less of exceeding 5 mg/L, but a probability greater than 0.45 of exceeding 3 mg/L, then the tool assigns one point for this

Predicted Nitrate Concentration Range (mg/L)	Points Assigned	Minimum Probability Predicted	Description
			indicator. Language is provided to the user that based on the model, it is likely that the nitrate level in their well is above background but below elevated (between 3 and 5 mg/L).
>5 mg/L	2	0.35	If the model predicts the input location has a probability of 0.25 or less of exceeding the MCL, but a probability greater than 0.35 of exceeding 5 mg/L, then the tool assigns two points for this indicator. Language is provided to the user that based on the model, it is likely that the nitrate level in their well is elevated (between 5 and 10 mg/L).
> 10 mg/L	3	0.25	If the model predicts the input location has a probability greater than 0.25 of exceeding the MCL, then the tool assigns the maximum number of points (three) for this indicator. In the GIS tool, language is provided to the user that based on the model, it is likely that nitrate level in their well exceeds the MCL (10 mg/L).

References

Agency for Toxic Substances and Disease Registry (ASTDR) (2017). Toxicological Profile for Nitrate and Nitrite, U.S. Department of Health and Human Services

Belitz, K., & Stackelberg, P. E. (2021). Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models. *Environmental Modelling and Software*, *139* (February), 105006. https://doi.org/10.1016/j.envsoft.2021.105006

Black, R. W., Wright, E. E., Bright, V. A. L., & Headman, A. O. (2023). Prediction of the Probability of Elevated Nitrate Concentrations at Groundwater Depths Used for Drinking-Water Supply in the Puget Sound Basin, Scientific Investigations Report 2023 – 5117. *USGS Scientific Investigations Report*.

Borchardt, M. A., Stokdyk, J. P., Kieke, B. A., Muldoon, M. A., Spencer, S. K., Firnstahl, A. D., Bonness, D. E., Hunt, R. J., & Burch, T. R. (2021). Sources and risk factors for nitrate and microbial contamination of private household wells in the fractured dolomite aquifer of northeastern Wisconsin. *Environmental Health Perspectives*, *129*(6), 1–18. https://doi.org/10.1289/EHP7813

Breiman, L., Friedman, J., Olshen, R.A., & Stone, C.J. (1984). Classification and Regression Trees (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470

Cherry, M., Gilmore, T., Mittelstet, A., Gastmans, D., & Gates, J. (2019). *Isotopic Composition of Groundwater and Precipitation in Nebraska, USA*. 1–4. https://digitalcommons.unl.edu/

Chicco, D., Warrens, M., Jurman, G. (2021) The Matthews Correlation Coefficient (MCC) is More Informative than Cohen's Kappa and Brier Score in Binary Classification Assessment, *IEEE Access, 9,* 78368-78361.

Davis, L. C., Bartholomay, R. C., Fisher, J. C., & Maimer, N. V. (2015). Water-Quality Characteristics and Trends for Selected Wells Possibly Influenced by Wastewater Disposal at the Idaho National Laboratory, Idaho, 1981 – 2012. *Scientific Investigations Report*, 106. http://dx.doi.org.10.3133/ sir20155003

Dormann C. F., Elith J., Bacher S., Buchmann C., Gudrun C., Carré G., Jaime R. Marquéz G., et al. (2013) "Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance." Ecography 36, no. 1: 27–46. https://www.jstor.org/stable/ecography.36.1.27.

Driscoll, F. G. (1986). Groundwater and Wells, *Johnson Filtration Systems, Inc.* ISBN: 0961645601

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. The Journal of animal ecology, 77(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.xExner, M. E., Hirsh, A. J., & Spalding, R. F. (2014). Nebraska's groundwater legacy: Nitrate contamination beneath irrigated cropland. *Water Resources Research*, *50*(5), 4474–4489. https://doi.org/10.1002/2013WR015073

Exner, M. E., Hirsh, A. J., & Spalding, R. F. (2014). Nebraska's groundwater legacy: Nitrate contamination beneath irrigated cropland*. Water Resources Research*, 50(5), 4474–4489. https://doi.org/10.1002/2013WR015073

Ferguson, R, Maharjan, B., Iqbal, J. (2024) Nitrogen Fertilizer Trends in Nebraska from 1955-2023, *UNL CropWatch Blog,* retrieved April 2024. https://cropwatch.unl.edu/2024/nitrogen-fertilizer-trends-nebraska-1955-2023

Friedman, J. H. (2002). Stochastic gradient boosting, Computational Statistics & Data Analysis, Volume 38, Issue 4, Pages 367-378, ISSN 0167-9473, https://doi.org/10.1016/S0167-9473(01)00065-2. (https://www.sciencedirect.com/science/article/pii/S0167947301000652)

Friedman, J.H. and Meulman, J.J. (2003), Multiple additive regression trees with application in epidemiology. Statist. Med., 22: 1365-1381. https://doi.org/10.1002/sim.1501

Garcia, V., Cooter, E., Crooks, J., Hinckley, B., Murphy, M., & Xing, X. (2017). Examining the impacts of increased corn production on groundwater quality using a coupled modeling system. *Science of The Total Environment*, *586*, 16–24. https://doi.org/10.1016/j.scitotenv.2017.02.009

Green, C. T., Liao, L., Nolan, B. T., Juckem, P. F., Shope, C. L., Tesoriero, A. J., & Jurgens, B. C. (2018). Regional Variability of Nitrate Fluxes in the Unsaturated Zone and Groundwater, Wisconsin, USA. *Water Resources Research*, *54*(1), 301–322. https://doi.org/10.1002/2017WR022012

Gross, E. L., & Low, D. J. (2013). Arsenic concentrations, related environmental factors, and the predicted probability of elevated arsenic in groundwater in Pennsylvania. *USGS Scientific Investigations Report*, 56. http://pubs.er.usgs.gov/publication/sir20125257

Hijmans, R. J., Phillips, S., Leathwick, J.; Elith, J. (2023). dismo Species Distribution Modeling Version 1.3-14. https://rspatial.org/raster/sdm/

Hirsch, R. M., Moyer, D. L., & Archfield, S. A. (2010). Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay River inputs. *Journal of the American Water Resources Association*, *46*(5), 857–880. https://doi.org/10.1111/j.1752-1688.2010.00482.x

Knierim, K. J., Kingsbury, J. A., Belitz, K., Stackelberg, P. E., Minsley, B. J., & Rigby, J. R. (2022). Mapped Predictions of Manganese and Arsenic in an Alluvial Aquifer Using Boosted Regression Trees. *Groundwater*, *60*(3), 362–376. https://doi.org/10.1111/gwat.13164

Kuhn, M., Wickham, H. (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. Version 1.2. https://www.tidymodels.org

Kuhn, M. (2023). caret Version 6.0-94 Classification and Regression Training, https://github.com/topepo/caret/

Litke, D. W. (2001). Historical Water-Quality data for the High Plains Regional Ground-Water Study Area in Colorado, Kansas, Nebraska, New Mexico, Oklahoma, South Dakota, Texas, and Wyoming, 1930-1998, *U.S. Geological Survey Water Resources Investigation Report 00-4254*

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

Malakar, A., Ray, C., D’Alessio, M., Shields, J., Adams, C., Stange, M., Weber, K. A., & Snow, D. D. (2023). Interplay of legacy irrigation and nitrogen fertilizer inputs to spatial variability of arsenic and uranium within the deep vadose zone. *Science of The Total Environment*, *897*, 165299. https://doi.org/10.1016/j.scitotenv.2023.165299

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure,* Volume 405, Issue 2, https://doi.org/10.1016/0005-2795(75)90109-9.

McMahon, Peter B., Dennehy, Kevin F., Bruce, Breton W., Gurdak, Jason J., and Qi, Sharon L., 2007, Water-quality assessment of the High Plains Aquifer, 1999–2004: U.S. Geological Survey Professional Paper 1749, 136 p

Mitchell, M., Zaldivar, A., Hutchinson, B., Spitzer, E., Raji, I. D., Vasserman, L., Barnes, P., & Gebru, T. (2019). Model Cards for Model Reporting., https://doi.org/10.1145/3287560.3287596

Nolan, B. T., Fienen, M. N., & Lorenz, D. L. (2015). A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. *Journal of Hydrology*, *531*, 902–911. https://doi.org/10.1016/j.jhydrol.2015.10.025

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. *Environmental Science and Technology*, *48*(10), 5643–5651. https://doi.org/10.1021/es405452q

Nolan, B.T., Hitt, K.J., (2003), Nutrients in Shallow Ground Waters Beneath Relatively Undeveloped Areas in the Conterminous United States, *U.S. Geological Survey Scientific Investigations Report 02-4289. https://pubs.usgs.gov/wri/wri024289/*

Pleim, J. AND L. Ran. (2023). Bidirectional ammonia flux modeling in the CMAQ-EPIC system. Dry Deposition and Surface Chemical Reactivity, Presented in Harwell, N/A, UK, October 04 - 05, 2023.

Snow, D. D., & Miller, D. M. (2018). *Bazile Ground Water Management Area Isotope and Recharge Study* (Vol. 372, Issue 2).

Sokal, R.R., Rohlf, F.J., Freeman, & Co. (1969). Biometry: The Principles and Practice of Statistics in Biological Research.

Spahr, N. E., Dubrovsky, N. M., Gronberg, J. M., Franke, O., & Wolock, D. M. (2010). Nitrate loads and concentrations in surface-water base flow and shallow groundwater for selected basins in the United States, water years 1990 - 2006. *U.S. Geological Survey Scientific Investigations Report 2010–5098*, 174. http://pubs.usgs.gov/ circ/1350/

Spalding, R. F., and Engberg, R. A. (1978). Groundwater Quality Atlas of Nebraska, *Conservation and Survey Division, Institute of Agriculture and Natural Resources, University of Nebraska-Lincoln.*

Spalding, R. F., and Exner, M. E., (1993), Occurrence of Nitrate in Groundwater – A Review, *Journal of Environmental Quality; 22:392-402*

Spalding, R.F., Watts, D.G., Schepers, J.S., Burbach, M.E., Exner, M.E., Poreda, R.J. and Martin, G.E. (2001), Controlling Nitrate Leaching in Irrigated Agriculture. *J. Environ. Qual.,* 30: 1184-1194. https://doi.org/10.2134/jeq2001.3041184x

Teluguntla, P., Thenkabail, P., Oliphant, A., Gumma, M., Aneece, I., Foley, D., McCormick, R. (2023). *Landsat-Derived Global Rainfed and Irrigated-Cropland Product 30 m V001* [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. Accessed 2024-04-08 from https://doi.org/10.5067/Community/LGRIP/LGRIP30.001

Tesoriero, A.J. and Voss, F.D. (1997), Predicting the Probability of Elevated Nitrate Concentrations in the Puget Sound Basin: Implications for Aquifer Susceptibility and Vulnerability. Groundwater, 35: 1029-1039. https://doi.org/10.1111/j.1745-6584.1997.tb00175.x

U.S. Environmental Protection Agency (1991). Integrated Risk Information System (IRIS) Chemical Assessment Summary (https://iris.epa.gov/static/pdfs/0076_summary.pdf). Accessed 2024.

U.S. Department of Agriculture (USDA) National Agricultural Statistics Service (NASS), (2023). Cropland Data Layer: USDA NASS, USDA NASS Marketing and Information Services Office, Washington, D.C. https://croplandcros.scinet.usda.gov/

Wellman, T.P., and Rupert, M.G., (2016), Groundwater quality, age, and susceptibility and vulnerability to nitrate contamination with linkages to land use and groundwater flow, Upper Black Squirrel Creek Basin, Colorado, 2013: U.S. Geological Survey Scientific Investigations Report, 2016–5020, 78 p., http://dx.doi.org/10.3133/sir20165020.

Wells, M.J.; Gilmore, T.E.; Mittelstet, A.R.; Snow, D.; Sibray, S.S. (2018) Assessing Decadal Trends of a Nitrate-Contaminated Shallow Aquifer in Western Nebraska Using Groundwater Isotopes, Age-Dating, and Monitoring. Water, 10, 1047. https://doi.org/10.3390/w10081047

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. *Science of the Total Environment*, *536*, 481–488. https://doi.org/10.1016/j.scitotenv.2015.07.080

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Yu, H., Cooper, A.R., & Infante, D.M. (2020). Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees. *Ecological Modelling, 432*, 109202.