Predicting nitrate exposure from groundwater wells using machine learning and meteorological conditions

Randall Etheridge[1], Jacob Hochard[2], Ariane L. Peralta[3], Thomas J. Vogel[4]

[1]Department of Engineering, Center for Sustainable Energy and Environmental Engineering, East Carolina University

[2]Haub School of Environment and Natural Resources, University of Wyoming

[3]Department of Biology, East Carolina University

[4]Coastal Resources Management PhD Program, East Carolina University

Abstract

Private groundwater wells have the potential to be an unmonitored source of contaminants that can harm human health for millions of people throughout the United States. Developing models that predict potential exposure to contaminants, such as nitrate, could guide sampling efforts and allow the residents to take action to reduce their risk. Machine learning models have been successful in predicting nitrate contamination using geospatial information such as proximity to nitrate sources or soil type, but previous models have not considered meteorological factors that change temporally. In this study, we test random forest (regression and classification) and linear regression models to predict nitrate contamination of wells using rainfall and temperature records over the previous 180-days. We trained and tested models for (1) all of North Carolina, (2) each geographic region in North Carolina, (3) a three-county region with high density animal agriculture, and (4) a three-county region with a low density of animal agriculture. All regression models had poor predictive performance ($R^2 < 0.04$) for all areas tested. The random forest classification model for the coastal plain region showed fair agreement (Cohen's kappa = 0.23) when trying to predict whether contamination occurred. All other classification models had slight or poor predictive performance. Our results show that temporal changes in rainfall and temperature alone are not enough to predict nitrate contamination in most areas of North Carolina but show potential in the coastal plain region.

Introduction

Private groundwater wells are the source of drinking water for 45 million people in the United States of which 2.4 million are in North Carolina (NCDHHS, 2019). The regular testing required for public water distribution systems is not required for private wells, which makes these wells a potential unknown source of contaminants that can harm human health (Rogan & Brady, 2009). For example, in North Carolina private groundwater wells are only required to be tested when they are drilled (NCDHHS, 2019). Annual testing is recommended, but it is rarely carried out due to the cost or low perception of risk (Jones et al., 2005; Postma et al., 2011). Since a low-cost and regular testing program is not available to serve the millions of private wells, it would be valuable to develop low-cost alternate methods of determining risk of contamination that use readily available data to guide sampling efforts and protect human health.

Nitrate is a drinking water contaminant widely known for causing methemoglobinemia, also known as blue baby syndrome. The EPA recommended maximum concentration in drinking water of 10 mg L^{-1} of nitrate-nitrogen was set to reduce the occurrence of methemoglobinemia (Rogan & Brady, 2009; Ward et al., 2018). There is evidence of many additional health effects of nitrate in drinking water when concentrations are below the 10 mg L^{-1} limit. The risks include specific cancers and birth defects (Schullehner et al., 2018; Ward et al., 2018). These health effects show the importance of being able to determine the likelihood of nitrate concentrations being slightly elevated in drinking water. Public water systems are required to monitor nitrate concentrations in drinking water, whereas regular testing is not required for private wells. Monitoring of nitrate levels is also more economical in drinking water treatment plants that can measure the nitrate concentration at one point to protect the health of its hundreds to thousands of customers. Nitrate analysis of private drinking water wells usually occurs through collecting a sample and sending it to a laboratory (Rogan & Brady, 2009). This only allows the well owner to know the quality of their drinking water at the time the sample was collected. It does not account for how the concentrations of contaminants may change through time. There is a need for a method to determine risk of private well owners that does not require expensive sampling and can account for changing conditions through time.

Wheeler et al. (2015) and Messier et al. (2019) developed random forest models to predict nitrate contamination in private groundwater wells in Iowa and North Carolina, respectively. Their models were primarily based on geospatial characteristics such as proximity

to potential sources of nitrate, land use, aquifer characteristics, and soil types. The models they developed have few variables that account for changes through time. However, the potential importance of seasonal patterns and annual variations is illustrated by year and month being two of the top three most important variables for a model developed for North Carolina (Messier et al., 2019) and year being one of the top ten most important variables in Iowa (Wheeler et al., 2015).

Abiotic factors (e.g., temperature, precipitation) factors that change through time affect nitrate concentrations in surface water and groundwater (Petry et al., 2002; Rivett et al., 2008; Stuart et al., 2011). Temperature is a primary driver of the rates of biological processes that alter nitrate concentrations such as nitrification and denitrification (Kadlec, 2012; van Kessel, 1977; Stanford et al., 1975; Stark, 1996). The rates of these processes are typically at their highest during the growing season when temperatures are the highest. Rainfall is another meteorological factor that alters nitrate transport and transformation. Hydrological conditions in soils often drive the predominant nitrate transformation process with saturated conditions promoting nitrate loss to the atmosphere through denitrification and unsaturated conditions promoting nitrate accumulation if there is a source of organic nitrogen or ammonium (Foulquier et al., 2013; Kadlec, 2012; Peralta et al., 2013). Nitrate transport is also linked to rainfall as storm events can mobilize nitrate that has accumulated in the soil (Baker & Showers, 2019; Hinckley et al., 2019; Jordan et al., 2003). The seasonal and annual variations in rainfall and temperature promote determining whether these variables are major drivers of nitrate contamination in private groundwater wells. Wheeler et al. (2015) did take into account climate variables such as the mean annual precipitation, mean annual minimum temperature, and mean annual maximum temperature, but these do not show how the changes in these variables have the potential to alter concentrations at shorter time scales. They found that precipitation and mean annual maximum temperature were important variables for predicting potential contamination (Wheeler et al. 2015).

The results of previous work in predicting nitrate in private groundwaters wells and current knowledge of the factors that influence nitrate concentration dynamics point to the potential of temperature and rainfall affecting nitrate contamination and human health. The objectives of this study were to develop and test multiple models that predict nitrate contamination in private drinking water wells in the three geographic regions of North Carolina

based on temperature and precipitation records over the previous 180 days to determine whether these temporal variables can be used to assess risk of contamination. Our modeling approach used both linear regression and random forest models to predict nitrate contamination using only preceding rainfall and temperature. Our models exclude geospatial variables (e.g. area of agricultural land within 1 km radius or number of hog lagoons within 5 km radius) that have been used in other modeling efforts in an attempt to create a simpler model that only uses readily available meteorological parameters.

Methods

Study area and nitrate data

The dataset used in this project comes from private wells in North Carolina that were measured for nitrate by the North Carolina Department of Health and Human Services State Laboratory for Public Health (SLPH) from 2013 to 2018. Samples that are run by the SLPH are typically collected by county health officials before being sent to the laboratory. This dataset does not include samples processed by private laboratories; however, it contains samples from all 100 counties in North Carolina. To facilitate matching with meteorological data, samples were georeferenced based on the residential address reported for each sample. A custom fuzzy matching algorithm was used to join the addresses from our dataset with those in the North Carolina Master Address Dataset for georeferencing. Samples that were not matched with an address were not included in our models. Our statewide dataset included 12,140 georeferenced samples (Figure 1A). Geographic regions were used as a simplistic means of dividing the samples into regions of different geology. We expected the transport and transformation of nitrate to vary in each region based on the geological properties; therefore, the explanatory variables were expected to differ between each region. When divided into geographic regions, the mountains, piedmont, and coastal plain had 2,584, 6,813 and 2,779 samples, respectively (Figure 1A).

Recognizing our approach does not take into account sources of nitrate for each well, we developed models for two subsets of counties that have different types and density of sources of nitrate. These two subsets are similar geographically and geologically. The first subset includes Bladen, Duplin and Sampson counties (animal production – AP; Figure 1B). These counties were chosen due to the high density of animal production operations that could serve as a source of nitrate (Burkholder et al., 2007; Osterberg & Wallinga, 2004). The models developed by

Messier et al. (2019) showed location relative to a hog lagoon or a poultry farm as important variables for prediction of groundwater well nitrate. Duplin and Sampson counties are the top two pork producing counties in the United States and Bladen County ranks eleventh (USDA, 2019). The second subset includes Cumberland, Harnett, and Lee counties (control – C; Figure 1B). The density of animal production is much lower in these neighboring counties. Identifying different important variables in models developed for the two county subsets could show the importance of nitrate source in the factors that drive nitrate contamination. All modeling methods were tested for the whole state, each geographic region and the two county subsets so a total of six models were developed for each method.

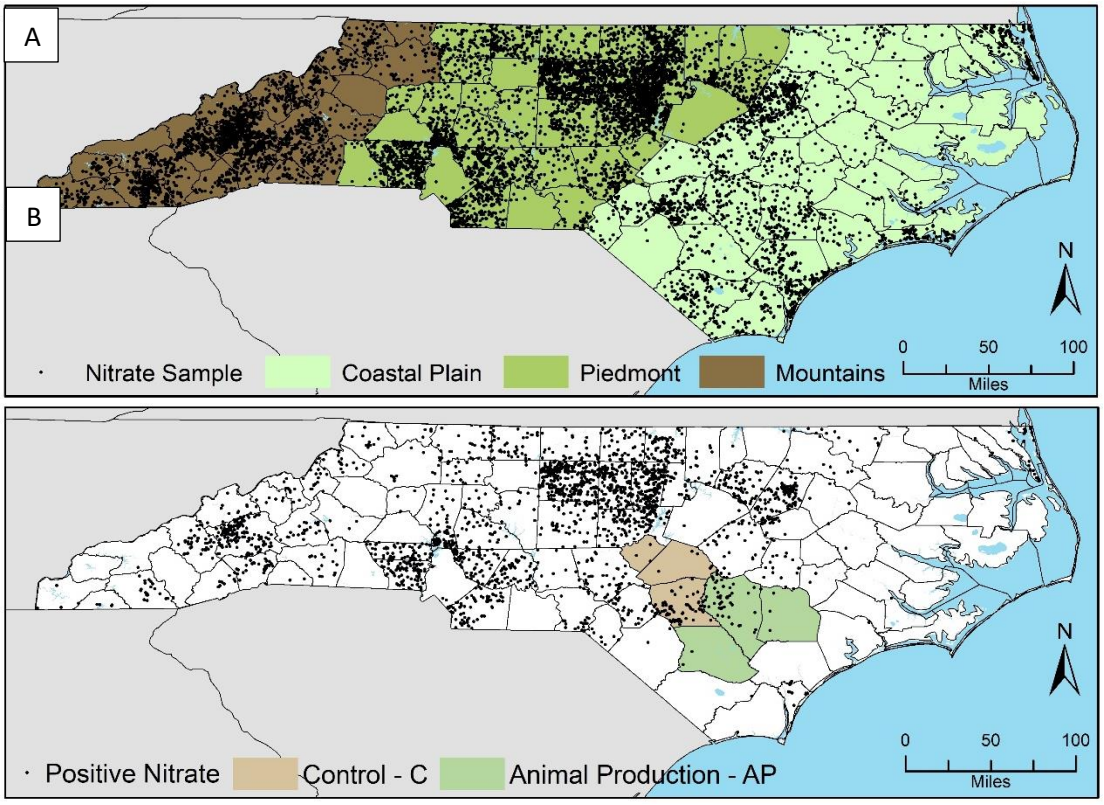

Figure 1: (A) Location of nitrate samples included in our analysis and North Carolina counties shaded based on region. (B) Samples with nitrate-nitrogen concentrations above 1 mg L^{-1} and locations of the two county subsets.

The minimum detection limit of nitrate for samples run by the SLPH is 1 mg L^{-1} of nitrate-nitrogen. As a result of this relatively high minimum detection limit, 76% of the samples in our dataset had no detected nitrate. For models that were developed to predict nitrate concentrations, a concentration of 0.5 mg L^{-1} was used for samples below the detection limit. Due to the evidence that slightly elevated nitrate concentrations can have health effects, detection

of nitrate at or above 1 mg L^{-1} of nitrate-nitrogen was used for models that were trying to predict whether a well was contaminated or not contaminated.

Temperature and precipitation data

The Oregon State University PRISM Climate Group (2004) historical data were the source of the temperature and precipitation data. The daily precipitation depth and temperature (°F) at each sample location was recorded for 180 days prior to sample collection through the day of sample collection. Records of temperature and precipitation depth were divided into periods of lag until sample collection of 0, 7, 14, 21, 30, 60, 90, 120, 150, and 180 days.

The daily sum, maximum, and mean precipitation depths were recorded for each lag period for each sample. Nitrate transport and transformation from surface sources are likely to be different for periods with high intensity rainfall that falls during a small portion of the lag period (e.g. during a hurricane) compared to low intensity rainfall that is distributed throughout the lag period. This likely difference is the reason that the sum, maximum, and mean precipitation depths are all included as variables for model testing. The daily minimum, maximum, mean minimum, mean maximum, mean average (i.e. mean of daily mean temperature over previous *n* days) temperatures were recorded for each lag period for each sample. All these variables were included in model testing due to the potential for short-term and seasonal trends in temperature to alter rates of nitrate transformation processes.

Model testing

Random forests continuous models and lasso regression models were utilized to predict nitrate concentration based on the temperature and precipitation prior to sample collection. Random forests were used because of their previous success in handling complex interactions and uncovering functions and relationships in environmental data (Nelson et al., 2018; Rahmati et al., 2019). Lasso regression was used because of its ability to choose the most important explanatory variables for inclusion in the model (Efron et al., 2004). In addition to the regression models, random forests classification models were utilized to predict whether a well was contaminated or not contaminated with nitrate. Although a regression model that predicts the nitrate concentration would be more useful for quantifying health risk, the work by Messier et al. (2019) showed the challenges of predicting nitrate concentrations in North Carolina due to the large number of samples where nitrate was below the detection limit. A classification model would help identify periods where well water may not be safe for drinking. This would allow targeted sampling to

determine whether a well is contaminated or allow the residents to use an alternate source of water during the period of time the well may be contaminated. The model development and testing was carried out in the R environment (R Version 4.0.3, R Core Team, 2019) using the tidymodels, tune, and workflow packages. The random forests models were fit using the ranger package and lasso regression models were fit using the glmnet package.

All the models were fit and tested with 5-fold cross validation following a split into a training and testing data sets. Root mean square error (RMSE) and the coefficient of determination (R^2) were used to assess model fit for the regression models. The model with the highest R^2 for the training set was applied to the testing data for further assessment. Accuracy and Cohen's kappa were used to assess model fit for the classification models. The model with the highest Cohen's kappa was also applied to the testing data. The variable importance scores and the significant variables were recorded for the random forests and lasso regression models, respectively, to determine the most important explanatory variables.

Results and Discussion

Regression model performance

The random forests regression models for the six different datasets had poor predictive performance (Table 1). The best training and test fits were for the coastal plain and control datasets. The five most important variables for each of the random forest regression models are listed in Table 2 and a full list of variable importance factors can be found in the supplemental information. Due to the poor performance of the models, variable importance is only discussed based on trends that hold for nearly all models. All the models had four or five of the top five most important variables related to temperature when attempting to predict nitrate concentrations using random forest regression models. The coastal plain and control datasets, which performed the best, have the variables with the greatest lag between the meteorological factor and sample collection.

Similarly, the lasso regression models performed poorly for each of the datasets (Table 1). The coastal plain dataset had the best performance for the lasso models as it did for the random forest regression models. The number of variables included in the models ranged from 38 for the coastal plain to zero for the mountain region. The variables with the coefficients of greatest magnitude are shown in Figure 2. A list of all variables included in the models can be found in the supplemental information. In contrast to the most important variables for the random forest

models, the variables with coefficients of greatest magnitude were all related to precipitation. Similar to the most important variables for random forest models, the variables with coefficients of greatest magnitude for the most accurate models had greater lag times. The 180-day mean rainfall was included in four out of six models and in each case has a negative coefficient with magnitude greater than 0.85. These models indicate that elevated long-term rainfall typically decreases nitrate concentrations in well water. These results could also indicate that meteorological records longer than 180 days are needed to accurately model the influence of rainfall on groundwater nitrate concentrations. This aligns with groundwater travel time research that has been conducted in North Carolina's coastal plain that show groundwater travel time from source to a stream can be years (Gilmore et al., 2016; Solomon et al., 2015).

Table 1: Summary of regression model performance for training and test splits for each dataset.

Dataset	R^2 – training	RMSE – training (mg L^{-1})	R^2 – test	RMSE – test (mg L^{-1})
Random Forests				
State	-0.005	2.36	-0.006	1.83
Coastal Plain	0.006	2.44	0.033	2.73
Piedmont	-0.01	1.94	-0.017	3.52
Mountain	-0.09	1.93	-0.021	1.39
AP	-0.15	4.62	-0.020	2.48
C	-0.016	2.35	0.013	4.84
Lasso Regression				
State	0.003	2.36	0	1.83
Coastal Plain	0.03	2.40	0.028	2.74
Piedmont	-0.001	1.93	0	3.49
Mountain	-0.011	1.89	0	1.37
AP	-0.03	4.38	-0.07	2.55
C	-0.015	2.38	-0.023	4.92

AP - Animal Production; C - Control; RMSE - Root Mean Square Error

Table 2: Five most important explanatory variables for the random forest regression models. The explanatory variables come from the Oregon State University PRISM Climate Group historical data.

Dataset	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
State	90-day T mean average	60-day T min	14-day T mean average	14-day T max	0-day T min
Coastal Plain	90-day T mean min	21-day T mean min	120-day P sum	60-day T mean min	90-day T maximum
Piedmont	14-day T mean average	14-day T mean max	60-day T min	60-day T mean average	21-day T mean average
Mountain	14-day T min	180-day P mean	7-day T mean average	30-day T mean average	7-day T mean max
AP	60-day T max	60-day T min	7-day T max	90-day T mean average	120-day P mean
C	60-day T min	90-day T min	120-day T mean min	120-day T min	90-day T mean min

AP - Animal Production; C - Control

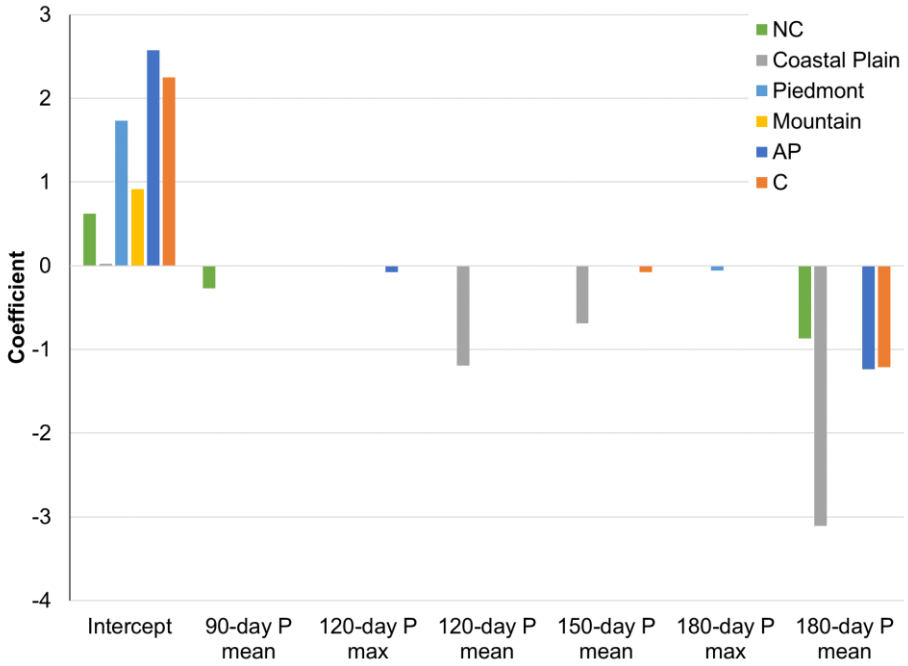

Figure 2: Coefficients of greatest magnitude for variables in the lasso regression models.

Classification model performance

The random forest classification models all had predictive accuracy at 0.68 or above; however, the strength of agreement based on Cohen's kappa ranged from poor (<0) to fair (0.2-0.4) (Landis & Koch, 1977). Most of the models fit into the slight (0-0.2) agreement category with the coastal plain model being the only model in the fair category. It is interesting to note that the

animal production dataset had the lowest Cohen's kappa of any dataset and it is located in the coastal plain. This indicates that rainfall and temperature are not the primary factors that influence nitrate contamination in this area of dense animal production. The models developed by Messier et al. (2019) that attempted to predict the class of concentration (<1, 1-5, ≥ 5 mg L^{-1}) of each well had kappa values with fair strength of agreement. Their models performed well in the animal production area. Our results combined with previous results promote the training and testing of models that take into account geospatial factors such as proximity to a nitrate source and temporal changes in rainfall and temperature.

Table 3: Summary of classification model performance for training and test splits for each dataset.

Dataset	Accuracy – training	Kappa – training	Accuracy – test	Kappa – test
State	0.74	0.052	0.75	0.074
Coastal Plain	0.80	0.150	0.81	0.232
Piedmont	0.68	0.036	0.68	0.081
Mountain	0.84	0.055	0.83	0.074
AP	0.76	-0.002	0.83	-0.060
C	0.68	0.140	0.68	0.050

AP - Animal Production; C - Control

The five most important explanatory variables for the random forest classification models are shown in Table 4. The full ranking of explanatory variables can be found in the supplemental information. All of the most important variables for the coastal plain and statewide models have lag times of 120 days or greater. The 180-day precipitation sum and mean were important variables in both models, which again shows that our work may not include a long enough record of rainfall. The important variables for all models indicate that near-term (less than one month) precipitation has little influence on the risk of nitrate contamination. The trends for temperature are not as clear as multiple models include temperature lag from zero to 30-days.

Table 4: Five most important explanatory variables for the classification models.

Dataset	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
State	180-day T mean max	180-day P sum	180-day P mean	180-day T mean average	150-day T mean max
Coastal Plain	150-day T max	180-day P sum	180-day P mean	180-day T max	120-day T max
Piedmont	180-day P max	180-day P mean	0-day T min	150-day P max	180-day P sum
Mountain	21-day T mean max	30-day T mean max	30-day T max	60-day P mean	30-day T mean min
AP	14-day T min	60-day T min	150-day T mean min	120-day T max	180-day T max
C	150-day P max	150-day T mean average	150-day P sum	60-day P mean	90-day T mean min

AP - Animal Production; C - Control

Area and Regional Differences

The models developed in this study typically had poor predictive performance; however, there are some common trends across the models that show differences between the regions and areas. The models for the coastal plain showed the best potential for predicting nitrate contamination based on temperature and rainfall. The lag duration of the most important variables typically decreased as you moved from the coastal plain to the piedmont and into the mountains. This may indicate a more direct linkage between the source of the contaminant and the contaminated well in the western part of the state. That is, the travel time from the source of nitrate to the groundwater well is much lower in the mountains than it is in the coastal plain.

An unexpected result from this study was the animal production (AP) area models had worse predictive power than the control area models. We expected the AP models to perform better because the high density of nitrate sources throughout the area would decrease the variability caused by having a lack of nitrate sources at the larger scales. The models for the control area were expected to have worse predictive power due to a more random placement of nitrate sources that the models would not have been able to take into account. Well depth was the most important variable for Wheeler et al. (2015) when predicting nitrate contamination in Iowa. Messier et al. (2019) and our models for North Carolina did not include well depth due to this variable not being readily available at the state or county level. Developing models that combine well depth, geospatial variables, and temporal variation in temperature and rainfall is the next step toward accurately assessing the risk of nitrate contamination in drinking water wells.

Conclusion

Assessing and reducing the risk of groundwater well nitrate contamination has the potential to improve human health in rural areas. Machine learning models have proven acceptable for predicting nitrate contamination risk in Iowa and North Carolina based on geospatial variable input. In this study, we developed and tested models for predicting nitrate contamination risks based on temperature and rainfall over the previous 180-days for different regions of North Carolina. The machine learning models (regression and classification) for the coastal plain performed the best of the tested models; however, their performance was only fair. This work underscores the need for variables in addition to rainfall and temperature to predict nitrate contamination risk accurately. Future work should test models that include temporal variables, well depth, and geospatial variable to predict contamination risk.

Acknowledgements

This work benefited from funding under National Science Foundation award #1902282 and U.S. Environmental Protection Agency (EPA) award #RD836942. The authors would also like to thank Glen Henry for his support in data acquisition.

References

Baker, E. B., & Showers, W. J. (2019). Hysteresis analysis of nitrate dynamics in the Neuse River, NC. *Science of The Total Environment*, *652*, 889–899. https://doi.org/10.1016/j.scitotenv.2018.10.254

Burkholder, J., Libra, B., Weyer, P., Heathcote, S., Kolpin, D., Thorne, P. S., & Wichman, M. (2007). Impacts of Waste from Concentrated Animal Feeding Operations on Water Quality. *Environmental Health Perspectives*, *115*(2), 308–312. https://doi.org/10.1289/ehp.8839

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499. https://doi.org/10.1214/009053604000000067

Foulquier, A., Volat, B., Neyra, M., Bornette, G., & Montuelle, B. (2013). Long-term impact of hydrological regime on structure and functions of microbial communities in riverine wetland sediments. *FEMS Microbiology Ecology*, *85*(2), 211–226. https://doi.org/10.1111/1574-6941.12112

Gilmore, T. E., Genereux, D. P., Solomon, D. K., & Solder, J. E. (2016). Groundwater transit time distribution and mean from streambed sampling in an agricultural coastal plain

watershed, North Carolina, USA: GROUNDWATER TRANSIT TIME. *Water Resources Research*, *52*(3), 2025–2044. https://doi.org/10.1002/2015WR017600

Hinckley, B. R., Etheridge, J. R., & Peralta, A. L. (2019). Storm Event Nitrogen Dynamics in Waterfowl Impoundments. *Water, Air, & Soil Pollution*, *230*(12), 294. https://doi.org/10.1007/s11270-019-4332-5

Jones, A. Q., Dewey, C. E., Doré, K., Majowicz, S. E., McEwen, S. A., Waltner-Toews, D., et al. (2005). Public perception of drinking water from private water supplies: focus group analyses. *BMC Public Health*, *5*(1), 129. https://doi.org/10.1186/1471-2458-5-129

Jordan, T. E., Whigham, D. F., Hofmockel, K. H., & Pittek, M. A. (2003). Nutrient and sediment removal by a restored wetland receiving agricultural runoff. *Journal of Environmental Quality*, *32*(4), 1534–1547.

Kadlec, R. H. (2012). Constructed marshes for nitrate removal. *Critical Reviews in Environmental Science and Technology*, *42*(9), 934–1005.

van Kessel, J. F. (1977). Factors affecting the denitrification rate in two water-sediment systems. *Water Research*, *11*(3), 259–267. https://doi.org/10.1016/0043-1354(77)90057-4

Landis, J. R., & Koch, G. G. (1977). An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics*, *33*(2), 363–374. https://doi.org/10.2307/2529786

Messier, K. P., Wheeler, D. C., Flory, A. R., Jones, R. R., Patel, D., Nolan, B. T., & Ward, M. H. (2019). Modeling groundwater nitrate exposure in private wells of North Carolina for the Agricultural Health Study. *Science of The Total Environment*, *655*, 512–519. https://doi.org/10.1016/j.scitotenv.2018.11.022

NCDHHS. (2019, December). Well Water and Health: Facts & Figures. Retrieved December 14, 2020, from https://epi.dph.ncdhhs.gov/oee/wellwater/figures.html

Nelson, N. G., Muñoz-Carpena, R., Phlips, E. J., Kaplan, D., Sucsy, P., & Hendrickson, J. (2018). Revealing Biotic and Abiotic Controls of Harmful Algal Blooms in a Shallow Subtropical Lake through Statistical Machine Learning. *Environmental Science & Technology*, *52*(6), 3527–3535. https://doi.org/10.1021/acs.est.7b05884

Osterberg, D., & Wallinga, D. (2004). Addressing Externalities From Swine Production to Reduce Public Health and Environmental Impacts. *American Journal of Public Health*, *94*(10), 1703–1708. https://doi.org/10.2105/AJPH.94.10.1703

Peralta, A. L., Ludmer, S., & Kent, A. D. (2013). Hydrologic history influences microbial community composition and nitrogen cycling under experimental drying/wetting treatments. *Soil Biology and Biochemistry*, *66*, 29–37. https://doi.org/10.1016/j.soilbio.2013.06.019

Petry, J., Soulsby, C., Malcolm, I. A., & Youngson, A. F. (2002). Hydrological controls on nutrient concentrations and fluxes in agricultural catchments. *Science of The Total Environment*, *294*(1), 95–110. https://doi.org/10.1016/S0048-9697(02)00058-X

Postma, J., Butterfield, P. W., Odom-Maryon, T., Hill, W., & Butterfield, P. G. (2011). Rural children's exposure to well water contaminants: Implications in light of the American Academy of Pediatrics' recent policy statement. *Journal of the American Academy of Nurse Practitioners*, *23*(5), 258–265. https://doi.org/10.1111/j.1745-7599.2011.00609.x

PRISM Climate Group. (2004, February 4). Oregon State University. Retrieved from http://prism.oregonstate.edu

R Version 4.0.3, R Core Team. (2019). *R: A language and environment for Statistical Computing* (Vol. http://www.R-project.org).

Rahmati, O., Choubin, B., Fathabadi, A., Coulon, F., Soltani, E., Shahabi, H., et al. (2019). Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods. *Science of The Total Environment*, *688*, 855–866. https://doi.org/10.1016/j.scitotenv.2019.06.320

Rivett, M. O., Buss, S. R., Morgan, P., Smith, J. W. N., & Bemment, C. D. (2008). Nitrate attenuation in groundwater: A review of biogeochemical controlling processes. *Water Research*, *42*(16), 4215–4232. https://doi.org/10.1016/j.watres.2008.07.020

Rogan, W. J., & Brady, M. T. (2009). Drinking Water From Private Wells and Risks to Children. *Pediatrics*, *123*(6), e1123–e1137. https://doi.org/10.1542/peds.2009-0752

Schullehner, J., Hansen, B., Thygesen, M., Pedersen, C. B., & Sigsgaard, T. (2018). Nitrate in drinking water and colorectal cancer risk: A nationwide population-based cohort study. *International Journal of Cancer*, *143*(1), 73–79. https://doi.org/10.1002/ijc.31306

Solomon, D. K., Gilmore, T. E., Solder, J. E., Kimball, B., & Genereux, D. P. (2015). Evaluating an unconfined aquifer by analysis of age-dating tracers in stream water. *Water Resources Research*, *51*(11), 8883–8899. https://doi.org/10.1002/2015WR017602

Stanford, G., Dzienia, S., & Pol, R. A. V. (1975). Effect of Temperature on Denitrification Rate in Soils. *Soil Science Society of America Journal*, *39*(5), 867–870. https://doi.org/10.2136/sssaj1975.03615995003900050024x

Stark, J. M. (1996). Modeling the temperature response of nitrification. *Biogeochemistry*, *35*(3), 433–445. https://doi.org/10.1007/BF02183035

Stuart, M. E., Gooddy, D. C., Bloomfield, J. P., & Williams, A. T. (2011). A review of the impact of climate change on future nitrate concentrations in groundwater of the UK. *Science of The Total Environment*, *409*(15), 2859–2873. https://doi.org/10.1016/j.scitotenv.2011.04.016

USDA. (2019). *2017 Census of Agriculture* (United States Department of Agriculture). Retrieved from https://www.nass.usda.gov/Publications/AgCensus/2017/index.php

Ward, M. H., Jones, R. R., Brender, J. D., De Kok, T. M., Weyer, P. J., Nolan, B. T., et al. (2018). Drinking Water Nitrate and Human Health: An Updated Review. *International Journal of Environmental Research and Public Health*, *15*(7), 1557. https://doi.org/10.3390/ijerph15071557

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

Supplemental Information for

Predicting nitrate exposure from groundwater wells using machine learning and meteorological conditions

Randall Etheridge[1], Jacob Hochard[2], Ariane L. Peralta[3], Thomas J. Vogel[4]

[1]Department of Engineering, Center for Sustainable Energy and Environmental Engineering, East Carolina University

[2]Haub School of Environment and Natural Resources, University of Wyoming

[3]Department of Biology, East Carolina University

[4]Coastal Resources Management PhD Program, East Carolina University

386 **Table S1:** Variable importance factors for the random forest regression models.
387

Variable	State	Coastal Plain	Piedmont	Mountain	AP	C
0-day P sum	0.124	0.104	0.452	0.306	0.082	0.054
0-day T min	2.443	0.275	2.827	0.161	0.392	0.102
0-day T mean min	0.959	0.475	1.056	0.666	0.372	0.124
0-day P max	0.292	0.084	0.517	0.169	0.103	0.089
0-day T mean max	0.752	0.353	1.263	0.485	0.526	0.099
0-day T max	0.458	0.383	0.451	0.582	0.227	0.127
0-day P mean	0.275	0.074	0.600	0.181	0.099	0.101
0-day T mean average	1.561	0.444	1.731	0.376	0.269	0.118
7-day P sum	0.614	0.225	0.653	0.159	0.164	0.070
7-day T min	1.064	0.486	0.930	0.411	0.499	0.270
7-day T mean min	0.614	0.406	0.761	0.689	0.464	0.301
7-day P max	0.620	0.266	0.395	0.213	0.281	0.135
7-day T mean max	0.928	0.549	0.785	0.915	0.610	0.262
7-day T max	1.060	0.414	0.634	0.685	0.778	0.312
7-day P mean	0.685	0.194	0.628	0.193	0.181	0.069
7-day T mean average	0.842	0.399	0.921	1.052	0.513	0.328
14-day P sum	0.849	0.241	0.696	0.273	0.175	0.153
14-day T min	0.969	0.509	1.670	1.114	0.501	0.248
14-day T mean min	1.295	0.504	1.671	0.476	0.443	0.396
14-day P max	0.398	0.290	0.423	0.168	0.164	0.197
14-day T mean max	1.221	0.343	3.537	0.435	0.593	0.140
14-day T max	2.546	0.445	1.386	0.439	0.518	0.211
14-day P mean	0.534	0.305	0.801	0.329	0.246	0.227
14-day T mean average	2.610	0.477	6.204	0.699	0.578	0.266
21-day P sum	0.383	0.286	0.377	0.301	0.336	0.289
21-day T min	0.871	0.628	2.039	0.747	0.410	0.312
21-day T mean min	1.083	0.750	1.250	0.570	0.569	0.322
21-day P max	0.255	0.239	1.017	0.134	0.200	0.115
21-day T mean max	1.865	0.349	1.473	0.846	0.440	0.129
21-day T max	0.972	0.395	0.892	0.352	0.583	0.131
21-day P mean	0.385	0.278	0.401	0.272	0.276	0.369
21-day T mean average	0.797	0.515	2.891	0.645	0.514	0.427
30-day P sum	0.462	0.427	0.379	0.449	0.360	0.277
30-day T min	0.705	0.487	0.806	0.760	0.559	0.308
30-day T mean min	1.190	0.490	1.576	0.467	0.441	0.327
30-day P max	0.366	0.248	0.409	0.209	0.355	0.084
30-day T mean max	1.167	0.340	1.629	0.325	0.498	0.310
30-day T max	0.763	0.328	1.849	0.347	0.718	0.171
30-day P mean	0.366	0.484	0.433	0.729	0.314	0.354

Variable	State	Coastal Plain	Piedmont	Mountain	AP	C
30-day T mean average	1.905	0.628	1.778	0.936	0.638	0.362
60-day P sum	0.462	0.392	0.386	0.316	0.446	0.307
60-day P max	0.324	0.250	0.329	0.186	0.115	0.209
60-day P mean	0.393	0.465	0.508	0.273	0.377	0.292
60-day T min	2.829	0.442	3.373	0.286	0.817	0.618
60-day T mean min	2.408	0.692	1.932	0.240	0.641	0.407
60-day T mean max	1.596	0.453	1.087	0.146	0.675	0.237
60-day T mean average	1.356	0.510	3.129	0.432	0.726	0.310
60-day T max	0.791	0.328	2.206	0.713	0.985	0.329
90-day P sum	0.792	0.498	0.429	0.265	0.496	0.332
90-day P max	0.376	0.253	0.462	0.296	0.271	0.086
90-day P mean	0.596	0.338	0.404	0.234	0.590	0.272
90-day T min	0.817	0.609	1.854	0.644	0.618	0.610
90-day T mean min	1.187	0.869	1.147	0.296	0.327	0.459
90-day T mean max	1.025	0.485	0.909	0.582	0.652	0.215
90-day T mean average	4.032	0.567	1.853	0.521	0.744	0.379
90-day T max	1.136	0.633	2.226	0.501	0.620	0.287
120-day P sum	0.467	0.742	0.830	0.225	0.543	0.345
120-day P max	0.308	0.386	0.451	0.138	0.129	0.116
120-day P mean	0.460	0.556	1.432	0.403	0.733	0.335
120-day T min	0.819	0.570	0.883	0.616	0.637	0.471
120-day T mean min	1.032	0.619	0.623	0.753	0.364	0.530
120-day T mean max	0.797	0.497	1.043	0.230	0.508	0.328
120-day T mean average	1.697	0.596	2.591	0.369	0.621	0.352
120-day T max	0.844	0.553	1.366	0.478	0.727	0.206
150-day P sum	0.639	0.481	0.747	0.371	0.589	0.378
150-day P max	0.178	0.264	0.899	0.109	0.111	0.163
150-day P mean	0.534	0.351	0.655	0.435	0.402	0.373
150-day T min	0.690	0.556	0.779	0.272	0.440	0.215
150-day T mean min	0.760	0.505	0.692	0.499	0.627	0.158
150-day T mean max	0.745	0.552	0.623	0.123	0.508	0.229
150-day T mean average	1.288	0.559	1.194	0.402	0.441	0.351
150-day T max	0.751	0.499	0.989	0.205	0.460	0.272
180-day P sum	0.330	0.479	0.705	0.912	0.619	0.358
180-day P max	0.351	0.225	0.503	0.221	0.190	0.118
180-day P mean	0.415	0.477	0.456	1.086	0.568	0.433
180-day T min	0.555	0.454	0.661	0.649	0.526	0.210
180-day T mean min	0.730	0.496	0.659	0.257	0.508	0.181
180-day T mean max	0.579	0.313	0.445	0.230	0.177	0.203
180-day T mean average	0.711	0.503	0.549	0.246	0.319	0.302
180-day T max	0.470	0.262	0.744	0.199	0.606	0.160

388
389 **Table S2:** A list of all variables and coefficients included in the lasso regression models. Blank
390 cells indicate the variable was not included in the model.
391

Variable	State	Coastal Plain	Piedmont	Mountain	AP	C
Intercept	0.6198	0.0228	1.7316	0.9144	2.5748	2.2493
0-day P sum		-0.0031				
0-day T min			0.0002			
0-day P max		-0.0034				
0-day T max		0.0021				
0-day P mean		-0.0024				
7-day P sum		0.0042				
7-day P max		0.0100	-0.0056			
7-day T mean max		-0.0001				
7-day T max		-0.0030				
7-day P mean		0.0299				
14-day P sum	-3.00E-05					
14-day T min					0.0013	
14-day P max					-0.0228	
14-day T mean max		-0.0125				
14-day T max		-0.0050				
14-day P mean	-2.13E-05					
14-day T mean average		-0.0018				
21-day T mean max		-0.0048				
21-day T mean average		-0.0004				
30-day P max		0.0101				
30-day T mean max		-0.0096				
60-day P sum	-4.73E-06					
60-day P max					-0.0195	
60-day P mean	-1.00E-05					
60-day T mean min	-0.0009	0.0008				
60-day T mean max		-0.0059				
90-day P sum	-0.0040					
90-day P max		0.0070	-0.0137			
90-day P mean	-0.2727					
90-day T min		0.0001			0.0037	
90-day T mean min		0.0013				
90-day T mean max	0.0033	-0.0028				
90-day T max		0.0149				
120-day P sum		-0.0100				
120-day P max	-0.0001				-0.0783	
120-day P mean		-1.1933				

Variable	State	Coastal Plain	Piedmont	Mountain	AP	C
120-day T mean min		0.0022				
120-day T mean max		-0.0072				
120-day T max		0.0253				
150-day P sum		-0.0045				-0.0003
150-day P mean		-0.6888				-0.0789
150-day T mean max		-0.0024				
150-day T max		0.0245				
180-day P sum	-0.0058	-0.0172			-0.0149	-0.0073
180-day P max			-0.0613			
180-day P mean	-0.8691	-3.1108			-1.2380	-1.2133
180-day T min						-0.0038
180-day T mean max		-0.0204				
180-day T mean average		-0.0005				
180-day T max	0.0078	0.0282	-0.0033			

392

Table S3: Variable importance factors for the random forest classification models.

Variable	State	Coastal Plain	Piedmont	Mountain	AP	C
0-day P sum	0.0006	0.0009	0.0005	0.0007	0.0001	0.0004
0-day T min	0.0059	0.0033	0.0063	0.0078	0.0008	0.0015
0-day T mean min	0.0034	0.0021	0.0028	0.0047	0.0007	0.0019
0-day P max	0.0007	0.0008	0.0006	0.0007	0.0001	0.0001
0-day T mean max	0.0032	0.0023	0.0028	0.0040	0.0011	0.0019
0-day T max	0.0048	0.0050	0.0042	0.0055	0.0018	0.0016
0-day P mean	0.0007	0.0009	0.0005	0.0005	0.0001	0.0003
0-day T mean average	0.0032	0.0018	0.0028	0.0051	0.0014	0.0023
7-day P sum	0.0030	0.0015	0.0025	0.0040	0.0012	0.0012
7-day T min	0.0053	0.0033	0.0049	0.0063	0.0027	0.0005
7-day T mean min	0.0052	0.0037	0.0045	0.0053	0.0017	0.0009
7-day P max	0.0037	0.0022	0.0044	0.0043	0.0010	0.0017
7-day T mean max	0.0046	0.0062	0.0041	0.0070	0.0021	0.0008
7-day T max	0.0058	0.0065	0.0051	0.0067	0.0029	0.0016
7-day P mean	0.0029	0.0018	0.0026	0.0043	0.0018	0.0014
7-day T mean average	0.0056	0.0038	0.0043	0.0082	0.0021	0.0006
14-day P sum	0.0045	0.0032	0.0036	0.0053	0.0017	0.0008
14-day T min	0.0055	0.0040	0.0058	0.0060	0.0101	0.0017
14-day T mean min	0.0050	0.0034	0.0046	0.0065	0.0020	0.0014
14-day P max	0.0051	0.0041	0.0045	0.0047	0.0017	0.0016
14-day T mean max	0.0041	0.0052	0.0039	0.0086	0.0015	0.0015
14-day T max	0.0039	0.0055	0.0034	0.0075	0.0013	0.0008
14-day P mean	0.0045	0.0033	0.0038	0.0056	0.0016	0.0014
14-day T mean average	0.0056	0.0024	0.0038	0.0081	0.0016	0.0004
21-day P sum	0.0034	0.0027	0.0032	0.0053	0.0024	0.0008
21-day T min	0.0046	0.0026	0.0044	0.0048	0.0032	0.0020
21-day T mean min	0.0040	0.0041	0.0033	0.0050	0.0015	0.0019
21-day P max	0.0035	0.0038	0.0030	0.0054	0.0019	0.0017
21-day T mean max	0.0043	0.0033	0.0039	0.0113	0.0019	0.0015
21-day T max	0.0040	0.0053	0.0034	0.0073	0.0010	0.0012
21-day P mean	0.0035	0.0026	0.0029	0.0042	0.0024	0.0008
21-day T mean average	0.0045	0.0030	0.0038	0.0086	0.0013	0.0024
30-day P sum	0.0040	0.0038	0.0035	0.0063	0.0008	0.0012
30-day T min	0.0051	0.0036	0.0040	0.0065	0.0013	0.0009
30-day T mean min	0.0048	0.0031	0.0035	0.0089	0.0010	0.0007
30-day P max	0.0046	0.0032	0.0041	0.0046	0.0013	0.0013
30-day T mean max	0.0047	0.0047	0.0030	0.0097	0.0040	0.0016
30-day T max	0.0043	0.0047	0.0032	0.0094	0.0026	0.0013
30-day P mean	0.0042	0.0041	0.0037	0.0055	0.0014	0.0011

Variable	State	Coastal Plain	Piedmont	Mountain	AP	C
30-day T mean average	0.0055	0.0030	0.0037	0.0072	0.0014	0.0022
60-day P sum	0.0048	0.0039	0.0036	0.0079	0.0011	0.0023
60-day P max	0.0041	0.0049	0.0032	0.0068	0.0020	0.0016
60-day P mean	0.0050	0.0038	0.0037	0.0091	0.0011	0.0027
60-day T min	0.0052	0.0019	0.0043	0.0055	0.0055	0.0010
60-day T mean min	0.0047	0.0029	0.0034	0.0055	0.0014	0.0011
60-day T mean max	0.0057	0.0033	0.0038	0.0052	0.0018	0.0014
60-day T mean average	0.0056	0.0028	0.0042	0.0066	0.0019	0.0016
60-day T max	0.0045	0.0056	0.0029	0.0069	0.0027	0.0008
90-day P sum	0.0056	0.0053	0.0038	0.0064	0.0033	0.0022
90-day P max	0.0046	0.0049	0.0044	0.0055	0.0026	0.0013
90-day P mean	0.0060	0.0058	0.0037	0.0072	0.0020	0.0015
90-day T min	0.0065	0.0025	0.0042	0.0085	0.0014	0.0012
90-day T mean min	0.0059	0.0042	0.0040	0.0065	0.0020	0.0026
90-day T mean max	0.0058	0.0066	0.0043	0.0053	0.0029	0.0019
90-day T mean average	0.0074	0.0039	0.0050	0.0074	0.0022	0.0021
90-day T max	0.0045	0.0066	0.0036	0.0054	0.0036	0.0019
120-day P sum	0.0071	0.0056	0.0041	0.0065	0.0030	0.0023
120-day P max	0.0048	0.0047	0.0051	0.0067	0.0033	0.0018
120-day P mean	0.0071	0.0052	0.0040	0.0064	0.0030	0.0022
120-day T min	0.0081	0.0053	0.0038	0.0070	0.0021	0.0015
120-day T mean min	0.0077	0.0057	0.0040	0.0049	0.0019	0.0014
120-day T mean max	0.0076	0.0059	0.0051	0.0049	0.0013	0.0015
120-day T mean average	0.0078	0.0068	0.0050	0.0049	0.0021	0.0018
120-day T max	0.0056	0.0147	0.0044	0.0056	0.0044	0.0001
150-day P sum	0.0086	0.0072	0.0038	0.0065	0.0038	0.0029
150-day P max	0.0057	0.0058	0.0060	0.0052	0.0021	0.0049
150-day P mean	0.0092	0.0069	0.0042	0.0072	0.0036	0.0026
150-day T min	0.0071	0.0048	0.0042	0.0052	0.0032	0.0015
150-day T mean min	0.0071	0.0062	0.0039	0.0087	0.0052	0.0004
150-day T mean max	0.0098	0.0065	0.0030	0.0052	0.0016	0.0016
150-day T mean average	0.0094	0.0057	0.0051	0.0068	0.0015	0.0029
150-day T max	0.0075	0.0203	0.0028	0.0045	0.0023	0.0004
180-day P sum	0.0167	0.0187	0.0059	0.0074	0.0037	0.0009
180-day P max	0.0064	0.0098	0.0079	0.0058	0.0033	0.0013
180-day P mean	0.0166	0.0186	0.0068	0.0083	0.0035	0.0023
180-day T min	0.0060	0.0063	0.0031	0.0056	0.0028	0.0014
180-day T mean min	0.0057	0.0060	0.0031	0.0052	0.0021	0.0016
180-day T mean max	0.0172	0.0074	0.0035	0.0081	0.0033	0.0009
180-day T mean average	0.0112	0.0091	0.0047	0.0078	0.0012	0.0024
180-day T max	0.0088	0.0185	0.0031	0.0063	0.0040	0.0006

395