Supporting Information for

Nitrate Variability in Groundwater of North Carolina using Monitoring and Private Well Data Models

Kyle P. Messier[†], *Evan Kane*[‡], *Rick Bolich*[‡], *Marc L. Serre*[†]*

Authors' Affiliation:

† Department of Environmental Science and Engineering, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC

‡ North Carolina Department of Environment and Natural Resources, Division of Water

Resources

***Corresponding Author:**

Marc L. Serre

Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina, 1303 Michael Hooker Research Center, Chapel Hill, NC 27599

Phone: (919) 966-7014 Fax: (919) 966-7911

The Supporting Information includes 22 pages, 9 tables, 9 figures, and 2 movie links.

Spatial Explanatory Variables

1) Nitrate Mass in Fertilizer, Manure, and Atmospheric Deposition. Estimates of nitrate were based on USGS estimates of nitrate mass in farm fertilizer, non-farm fertilizer, manure, and atmospheric deposition. The estimates are based on county-level estimates compiled from fertilizer sales, census of agriculture, and population estimates following the methods outlined in Ruddy et al.[1], and employed by Hoos and McMahon[2] for the analysis of nitrogen loads in streams using spatially referenced regression on watershed attributes (SPARROW).

Nitrate mass estimates in kilograms per year per county was obtained from Ruddy et al[1] and averaged over all of the available years to obtain an average mass per year per county estimate. Similar to Hoos and McMahon[2], in order to more accurately represent the spatial distribution of the county-level data, nitrate farm fertilizer and manure estimates were distributed to only agricultural land according to the 2006 National Land Cover Database [3]. The non-farm fertilizer was distributed to the developed, forest, shrub, and grassland land cover classes. The atmospheric deposition was distributed evenly across each county. The total amount of nitrate mass was divided by the number of 30-meter cells within each county that was portioned mass estimates resulting in variables that represent the average amount of nitrate mass input (from the respective source) per year per 900 square-meters, which is then multiplied by 900 square-meters to obtain nitrate mass per year. Following the creation of nitrate mass variables, we calculate the mean nitrate mass per year per hectare from each source (*l=Farm Mass, Non-Farm Mass, Manure, or Atmospheric deposition)* as:

$$NM_i^{(l)}(\lambda_l) = \frac{1}{\pi\lambda^2}\sum_{j=1}^{n_i(\lambda_l)} M_j^{(l)} \qquad (1)$$

where $NM_i^{(l)}(\lambda_l)$ is the mean nitrate mass per year per hectare of type (l) within a radius λ_l of nitrate point *i,* $M_j^{(l)}$ is the estimated nitrate mass (kg/year) of type l for the j^{th} pixel described above surrounding nitrate point i , $\pi\lambda^2$ is the area of the circular buffer, and $n_i(\lambda_l)$ is the number of pixels within the circular buffer of radius λ_l around nitrate point i. Area units are converted from square meters to hectares, which is more common in the agricultural field.

2) Point Source Variables. Following Messier et al.[4], we calculate the sum of exponentially decaying contribution from various potential nitrate point sources including wastewater treatment residuals (WTR) application fields[5], swine farms, swine waste lagoons, cattle farms, chicken farms, and wastewater treatment plants (WWTP). Equation 2 shows the general form of the point source variables,

$$PS_i^{(l)}(\lambda_l) = \sum_{j=1}^{n_l} C_{0j}^{(l)} \exp\left(-3 * \frac{D_{ij}}{\lambda_l}\right), \qquad (2)$$

where $PS_i^{(l)}(\lambda_l)$ is the sum of exponentially decaying contribution from point sources type (l) at nitrate point i, n_l is the total number of point sources of type (l), D_{ij} is the distance between the j-th point source of type (l) and the nitrate point i, C_{0j} is a proxy for the initial nitrate concentration at the point source if available, or equal to 1 otherwise, and λ_l is the exponential decay range corresponding to the distance it takes for nitrate released by source of type (l) to be reduced by 95%. WWTP initial values are based on the design capacity of the plant; cattle, chicken, and swine farms are weighted based on the number of animals; and the other point source variables do not have information available to provide reasonable estimates of the initial concentration.

3) On-Site Wastewater Treatment. On-site wastewater treatment, or septic tanks, variables are created following the methods of Pradhan et al[6] with adjustments for our variables' circular buffers as opposed to watershed polygons. The 1990 US census was the last census to collect information on the method of wastewater treatment used in residential homes, which was obtained at the census block group level as the number of septic or other on-site wastewater treatment systems (i.e. latrine, straight pipe) per census block group. We calculated the estimated septic system density as follows:

$$SD_i(\lambda) = \frac{\sum_{j=1}^{n_i(\lambda)} \xi_j^{(\lambda)}}{\pi\lambda^2} \tag{3}$$

where $SD_i(\lambda)$ is the septic system density (#/mi^2) around nitrate point i within circular buffer λ, $n_i(\lambda)$ is the total number of census block groups within circular buffer λ, $\xi_j^{(\lambda)}$ is the number of septic systems in the overlapping area between census block j and the circle created by radius λ assuming a constant density of septic tanks in each census block, and $\pi\lambda^2$ equals the area of the circular buffer created with radius λ.

The average nitrate loading from septic system is

$$SN_i(\lambda) = \sum_{j=1}^{n_i(\lambda)} PD_j * a_{j\lambda} * p_j * 10 \tag{4}$$

where $SN_i(\lambda)$ is the septic nitrate (lb/yr) around nitrate point i circular buffer λ, $n_i(\lambda)$ is the total number of census block groups within circular buffer λ, PD_j is the population density (people/mi^2) in census block group j, $a_{j\lambda}$ is the area of overlap between census block group j and λ, p_j is the proportion of people (dimensionless) in census block j that are on septic systems, and the result is multiplied by 10 lb/person-year based on the worst case-scenario that the amount of nitrate septic influent is estimated at 10 pounds per person per year [6].

4) Population density. Population density represents a surrogate variable associated with non-farm nitrate inputs and is calculated for each circular buffer using the 2000 census population data at the block level and assumes population is evenly distributed over each block.

5) National Land Cover Database. We construct explanatory variables based on the National Land Cover Database (NLCD) satellite imagery file that characterizes land cover types at 30 meter resolution. We create variables for every NLCD land cover type and aggregated land cover type that represent attenuation variables including deciduous forest, evergreen forest, mixed forest, herbaceous wetlands, and woody wetlands . For a NLCD variable (l) of interest we calculate

$$LC_i^{(l)}(\lambda_l) = \frac{1}{n_i(\lambda_l)} \sum_{j=1}^{n_i(\lambda_l)} I_j^{(l)} \qquad (5)$$

where $LC_i^{(l)}(\lambda_l)$ is the percent of land cover of type (l) within a radius λ_l of nitrate point i, $I_j^{(l)}$ is an indicator variable equal to 1 if the j^{th} pixel surrounding nitrate point i is of type l, and zero otherwise, and $n_i(\lambda_l)$ is the number of pixels within the circular buffer of radius λ_l around nitrate point i.

6) *Slope and Topographic Wetness Index.* Slope and Topographic Wetness Index (TWI) [7] are variables that represent possible attenuation and transport variables and are calculated from a digital elevation raster. Slope is calculated as the average gradient between adjacent cells within a circular buffer centered on each well. TWI expresses the potential wetness in soils due to topography and is commonly used in watershed scale hydrological models [7,8] and as a predictor variable for groundwater contaminants [9]. The mean TWI within a circular buffer is calculated as

$$TWI_i(\lambda) = \frac{1}{n_i(\lambda)} \sum_{j=1}^{n_i(\lambda)} \ln\left(\frac{F_{Aj}}{\tan(\beta_j)}\right) \qquad (6)$$

where F_{Aj} is the j-th flow accumulation calculated from a D8 flow algorithm, and β_j is the j-th pixel slope, and $n_i\ (\lambda)$is the number of pixels that are within radius λ around nitrate point i.

7) *Soil variables.* Soil based variables are calculated as the average of the given soil characteristic within a circular buffer. We use the multilayer soil characteristics dataset for the conterminous United States (CONUS-SOIL), which contains soil estimates of pH, permeability, hydrologic soil groups, available water capacity, and depth to bedrock [10]. Data on histosol soil type, a soil group that contains large amounts of organic matter in the upper profile, was obtained directly from the supporting information of Nolan and Hitt[11].

8) *USGS withdrawals.* Similar to Nolan and Hitt[11], we calculate the average water withdrawals from groundwater, surface water, and the sum of groundwater and surface water. Water withdrawal rates per county [12] are distributed evenly over each county, which is then used to calculate the average water withdrawal within a circular buffer.

Model Coefficient Interpretations

Interpretations of regression sources parameters are based on the nonlinear model formulation: Since nitrate was log-transformed and the nonlinear model has multiplicative interaction, the percent increase of the geometric mean of nitrate is the exponential of the source coefficient multiplied by the result of the attenuation and transport terms held to their mean value. Below is the derivation of this interpretation:

In matrix format, let us write an equation for the log of the nitrate with the equation form in this paper, with the attenuation and transport term simplified into one exponential term.

$$Ln\ (N) = X\beta \exp(Z\gamma)$$

For simplicity, let's reduce it to one source and one attenuation/transport variable.

$$Ln\ (N) = \beta_1 X_1 \exp(\gamma_1 Z_1)$$

Let us write another equation that represents a one unit increase in source X_1.

$$Ln\ (N_2) = \beta_1 (X_1 + 1)\exp(\gamma_1 Z_1)$$

For clarity, rename $N = N_1$ and evaluate the attenuation/transport term at the mean values, leading to a constant value. We have two equations:

$$\begin{cases} Ln\ (N_1) = \ \beta_1 X_1 K \\ Ln\ (N_2) = \beta_1 (X_1 + 1) K \end{cases}$$

Subtract the equations and simplify

$$Ln(N_1) - Ln(N_2) = -\beta_1 K$$
$$-B_1 K = Ln\left(\frac{N_1}{N_2}\right)$$
$$\beta_1 K = Ln\left(\frac{N_2}{N_1}\right)$$
$$\exp(\beta_1 K) = N_2/N_1$$

Using the derived formula the model source interpretations for the monitoring well model are as follows:

1) The percent increase in the geometric mean of nitrate in mg/L for every 1 kg/yr/ha of farm manure while other sources and attenuation/transport is constant is $\exp(0.0759 * 0.456) = 1.04 = 4\%$.

2) The percent increase in the geometric mean of nitrate in mg/L for every 1 unit of wastewater treatment residuals while other sources and attenuation/transport is constant is $\exp(0.245 * 0.456) = 1.12 = 12\%$.

3) The percent increase in the geometric mean of nitrate in mg/L for every 1 kg/yr/ha of farm fertilizer while other sources and attenuation/transport is constant is $\exp(0.132 * 0.456) = 1.06 = 6\%$.

4) The percent increase in the geometric mean of nitrate in mg/L for every 100 pigs in swine CAFO's while other sources and attenuation/transport is constant is $\exp(0.117 * 0.456) = 1.06 = 6\%$.

5) The percent increase in the geometric mean of nitrate in mg/L for every 1 percent increase in developed low land while other sources and attenuation/transport is constant is $\exp(0.112 * 0.456) = 1.05 = 5\%$.

6) The percent increase in the geometric mean of nitrate in mg/L for every 1 kg/yr/ha of nitrate in atmospheric deposition while other sources and attenuation/transport is constant is $\exp(0.447 * 0.456) = 1.23 = 23\%$.

For private wells:

1) The percent increase in the geometric mean of nitrate in mg/L for every 1 kg/yr/ha of farm fertilizer is while other sources and attenuation/transport is constant $\exp(0.0432 * 0.4636) = 1.02 = 2\%$.

2) The percent increase in the geometric mean of nitrate in mg/L for every 10 percent increase in developed land while other sources and attenuation/transport is constant is $\exp(0.0112 * 0.4636 * 10) = 1.05 = 5\%$.

3) The percent increase in the geometric mean of nitrate in mg/L for every 1 unit of swine lagoons while other sources and attenuation/transport is constant is $\exp(0.1079 * 0.4636) = 1.05 = 5\%$.

4) The percent increase in the geometric mean of nitrate in mg/L for every 100 kg/yr/ha of nitrate in atmospheric deposition while other sources and attenuation/transport is constant is $\exp(2.9e - 11 * 0.4636 * 100) = 1.02 = 0.0000000014\%$. This seemingly negligible increase is due to the fact that the hyperparameter is 25km, thus the increase in atmospheric deposition in widely distributed.

Tables

Table S1. Groundwater Nitrate Data Source Basic Information.

Data Source	Median (mg/L)	Mean (mg/L)	Unique Wells	Space/Time Samples	Year Range	Percent Detected
NC-DWR	1.30	4.61	366	11,004	1980-2011	79.7
USGS	0.10	6.14	585	1,318	1990-2012	61.4
Private Well	0.62	1.66	18,664	22,067	1990-2011	30.6

Table S2. Spatial explanatory variable model category. The candidate variables are listed according to their category in the groundwater NO_3^- model. Details on how each variable calculated is presented in the previous section of the supporting information.

	Sources	Attenuation	Transport
Variable Names	Farm Fertilizer; Non-Farm Fertilizer; Manure; Nitrate Atmospheric Deposition; Points Source: WWTP, Cattle Farms, Poultry Farms, Swine Farms, Swine Lagoons, Waste Treatment Residuals (WTR); On-Site Wastewater Treatment input; On-Site Wastewater treatment density; National Landcover Database: Developed Open, Developed Low, Developed Medium, Developed High, Developed All, Pasture/Hay, Crops, Agriculture combined	National Landcover Database: Deciduous, Evergreen, Mixed Forest, Forest All, Grassland, Woody Wetlands, Herbaceous Wetlands, Wetlands All; Histosol Soils	Soil Permeability; Depth to Bedrock; pH; Hydrologic Soil Groups: A,B,C,D; Available Water Capacity; Water Withdrawals: Groundwater, Surface Water, Total; Topographic Wetness Index; Mean Slope

Table S3. Nonlinear regression model variables selected via CFN-RHO and parameter estimates for spatially-smoothed/time-averaged NO_3^- monitoring (left) and private well (right) models. All variables are significant with *p-value* < 0.025. Variables units: **a-** Kg- NO_3^-/yr/ha, **b-** Dimensionless, **c-** 100 pigs, **d-** percent, **e-**cubic meters per second. (-) Not a variable in the model.

	25 KM Spatially Smoothed/Temporally Averaged Nitrate					
	Monitoring Well			Private Well		
Variable	Variable Range	Coefficient Estimate	Standard Error	Variable Range	Coefficient Estimate	Standard Error
Constant	**n/a**	-3.71	0.191	n/a	-1.570	0.0382
	Source Variables					
Wastewater Treatement Residuals (WTR)[b]	40 km	0.0235	0.0056	-	-	-
Farm Fertilizer[a]	25 km	4.67e-9	8.0e-10	25 km	7.2e-10	3.5e-11
Swine Lagoons[b]	-	-	-	35 km	0.0385	0.0016
Atmospheric Deposition[a]	25 km	3.07e-8	4.8e-9	25 km	8.49e-9	1.4e-10
Wastewater Treatment Plant	25 km	0.0132	0.0003	-	-	-
	Attenuation and Transport Variables					
Deciduous Forest[d]	25 km	-0.0416	0.0026	25 km	-0.0312	5.5e-4
Mixed Forest	-	-	-	25 km	-0.0395	0.0021
Herbaceous Wetlands[d]	25 km	-0.7042	0.0649	25 km	-0.1757	0.0112
Histosol[d]	25 km	-0.0482	0.0076	25 km	-0.0924	0.0037
Hydrologic Soil Group D[d]	25 km	-0.013	0.0019	25 km	-0.0271	5.7e-4
Hydrologic Soil Group C[d]	25 km	-0.0123	0.0027	-	-	-
GWW[e]	-	-	-	25 km	-1.8014	0.0448

Table S4. The number of times each variable in the full spatially-smoothed/time-averaged LUR model for monitoring wells was selected in the ten-fold cross-validation runs.

Variable	Number out of 10 the variable was picked in 10 fold cross-validation
Farm Mass	10
NADP	7
WWTP	9
WTR	10
Deciduous	10
Herbaceous Wetlands	10
HSG-C	7
HSG-D	8
Histosols	10

Table S5. The number of times each variable in the full spatially-smoothed/time-averaged LUR model for private wells was selected in the ten-fold cross-validation runs.

Variable	Number out of 10 the variable was picked in 10 fold cross-validation
Farm Mass	10
Atmospheric Deposition	10
Swine Lagoons	10

HSG D	10
Deciduous	10
Herbaceous Wetlands	10
GWW	10
Histosol	10

Table S6. The number of times each variable in the full time-averaged LUR model for monitoring wells was selected in the ten-fold cross-validation runs.

Variable	**Number out of 10 the variable was picked in 10 fold cross-validation**
Manure	**6**
WTR	**10**
Farm Fertilizer	**10**
Swine CAFO’s	**10**
Developed Low	**7**
Atmospheric Deposition	**7**
Forest	**7**
Herbaceous Wetlands	**10**
Histosol	**8**
Slope	**7**

Table S7. The number of times each variable in the full time-averaged LUR model for private wells was selected in the ten-fold cross-validation runs.

Variable	Number out of 10 the variable was picked in 10 fold cross-validation
Farm Fertilizer	**10**
Developed	**10**
Swine Lagoons	**2**
Atmospheric Deposition	**7**
Histosol	**7**
HSG D	**10**
Deciduous	**0**

Table S8. 2 x 2 table showing the percent of area in North Carolina as predicted by this study's LUR-BME model to be (I) below 0.25 mg/L for both monitoring and private wells, (II) above 0.25 mg/L for monitoring wells and below 0.25 for private wells, (III) below 0.25 mg/L for monitoring wells and above 0.25 mg/L for private wells, and (IV) above 0.25 mg/L for both monitoring and private wells.

		Monitoring Well	
		<0.25 mg/L	**>=0.25 mg/L**
Private Well	**<0.25mg/L**	***I*** **43.2**	***II*** **1.4**
	>=0.25mg/L	***III*** **30.6**	***IV*** **24.8**

Table S9. 2 x 2 table showing the percent of area in North Carolina as predicted by this GWAVA models (Nolan and Hitt, 2006) to be (I) below 0.25 mg/L for both monitoring and private wells, (II) above 0.25 mg/L for monitoring wells and below 0.25 for private wells, (III) below 0.25 mg/L for monitoring wells and above 0.25 mg/L for private wells, and (IV) above 0.25 mg/L for both monitoring and private wells.

		Shallow Groundwater	
		<0.25 mg/L	**>=0.25 mg/L**
Drinking Water	**<0.25mg/L**	***I*** **25.4**	***II*** **6.0**
	>=0.25mg/L	***III*** **2.6**	***IV*** **66.0**

Figures

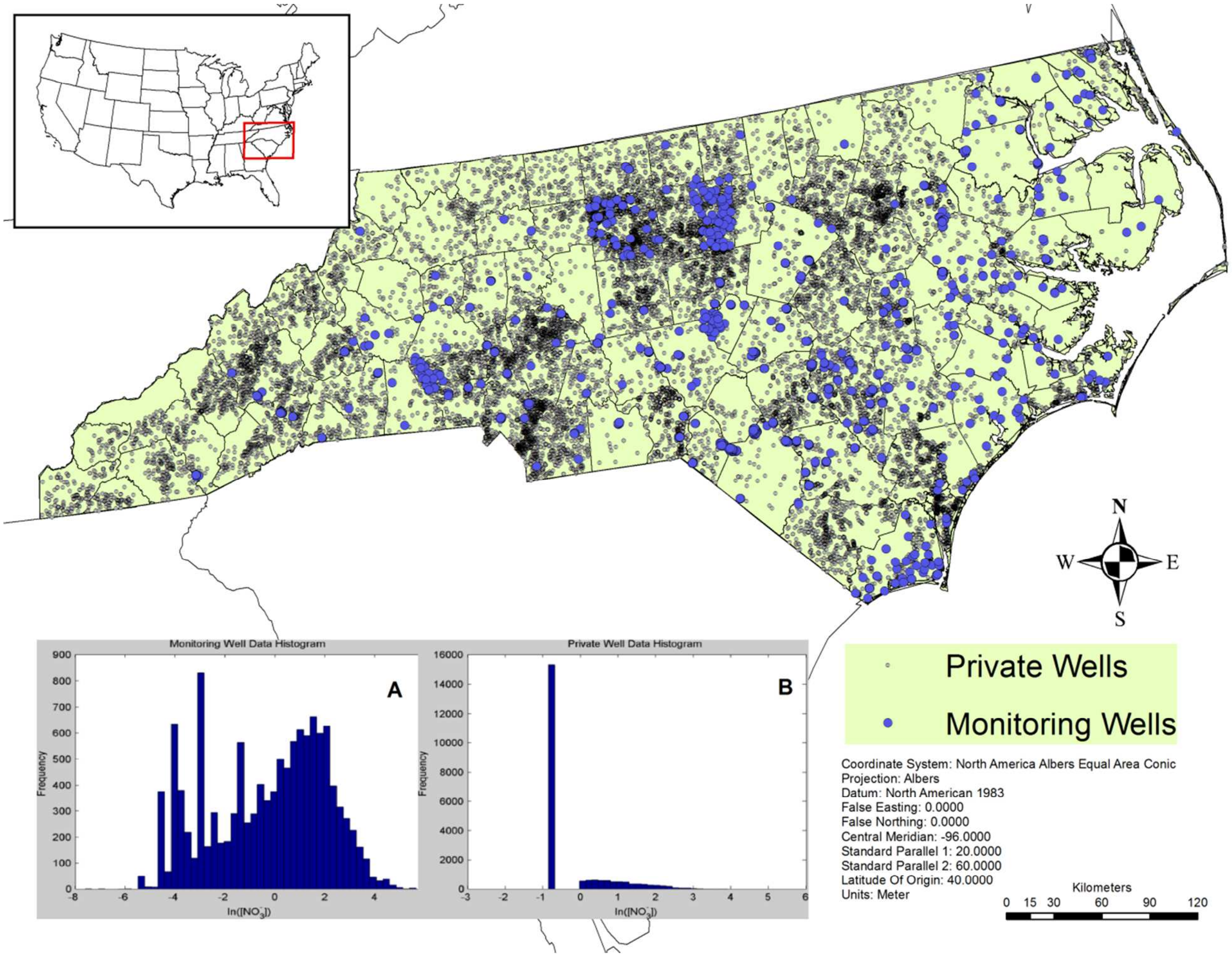


Figure S1. North Carolina study area with private well and monitoring well nitrate databases. The convex hull of monitoring and private wells covers 88 and 99.5 percent of North Carolina, respectively. A) Frequency histogram of the log-nitrate concentration for monitoring well data. B) Frequency histogram of the log-nitrate concentration for private well data.

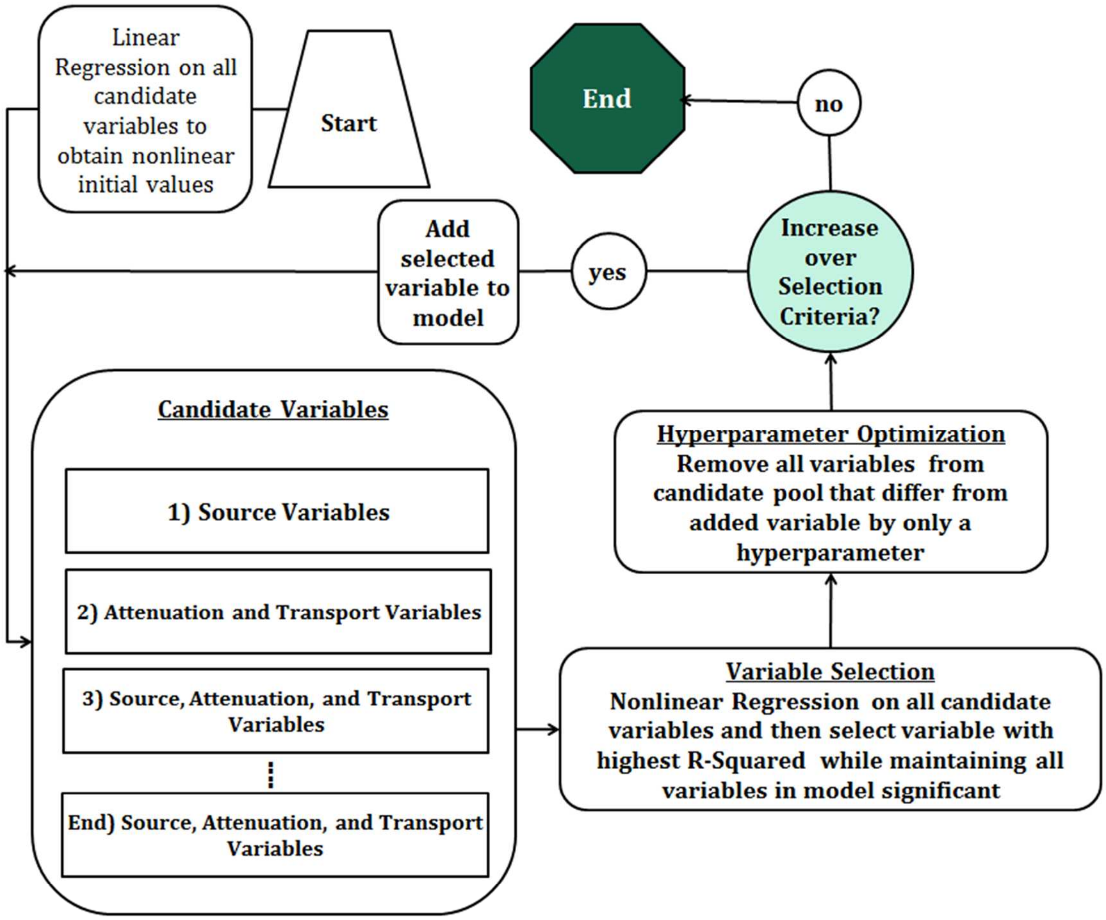


Figure S2. Flow diagram of the constrained forward nonlinear and hyperparameter optimization model selection procedure.

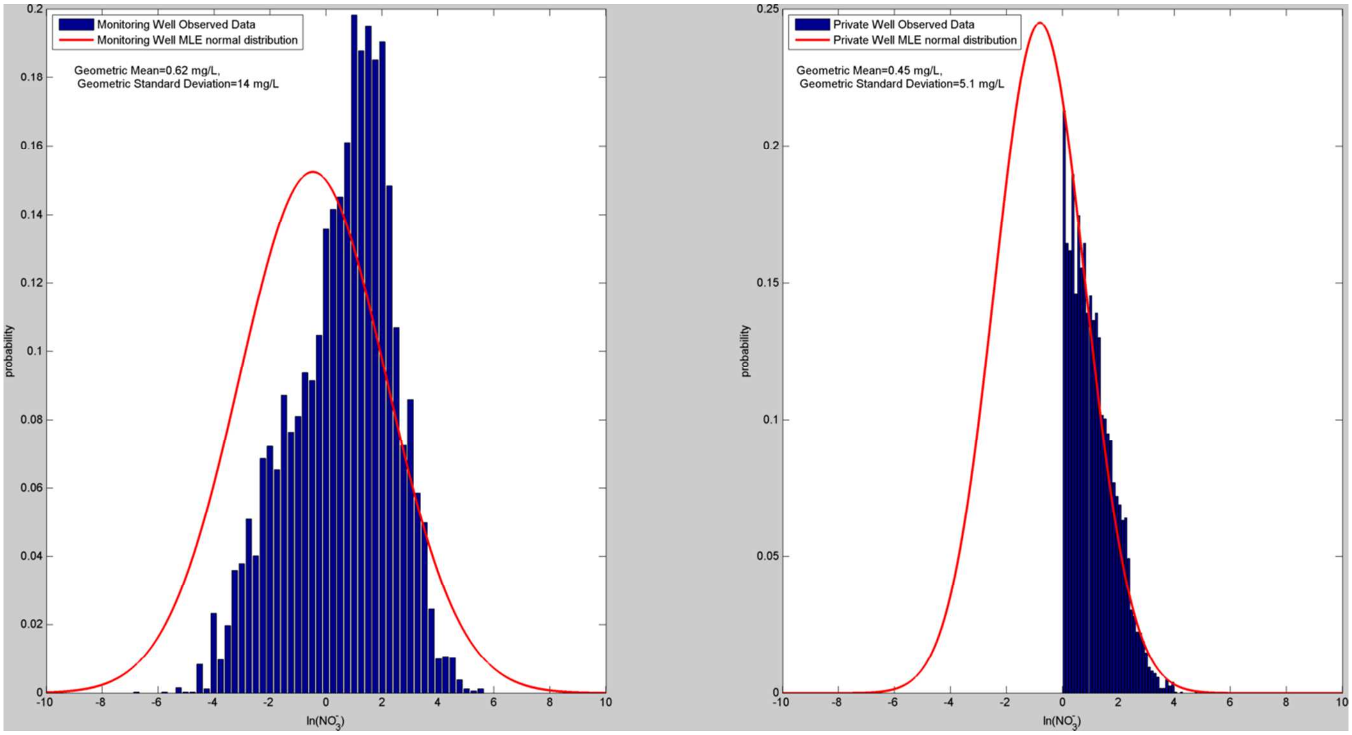


Figure S3. Left) Histogram (blue) of monitoring well data only observed above the detection limit, log-transformed. The fitted normal distribution (red) based on the maximum likelihood estimation method accounting for nondetects and their detection limits. Right) Histogram (blue) of private well data only observed above the detection limit, log-transformed. The fitted normal distribution accounting for nondetects (red).

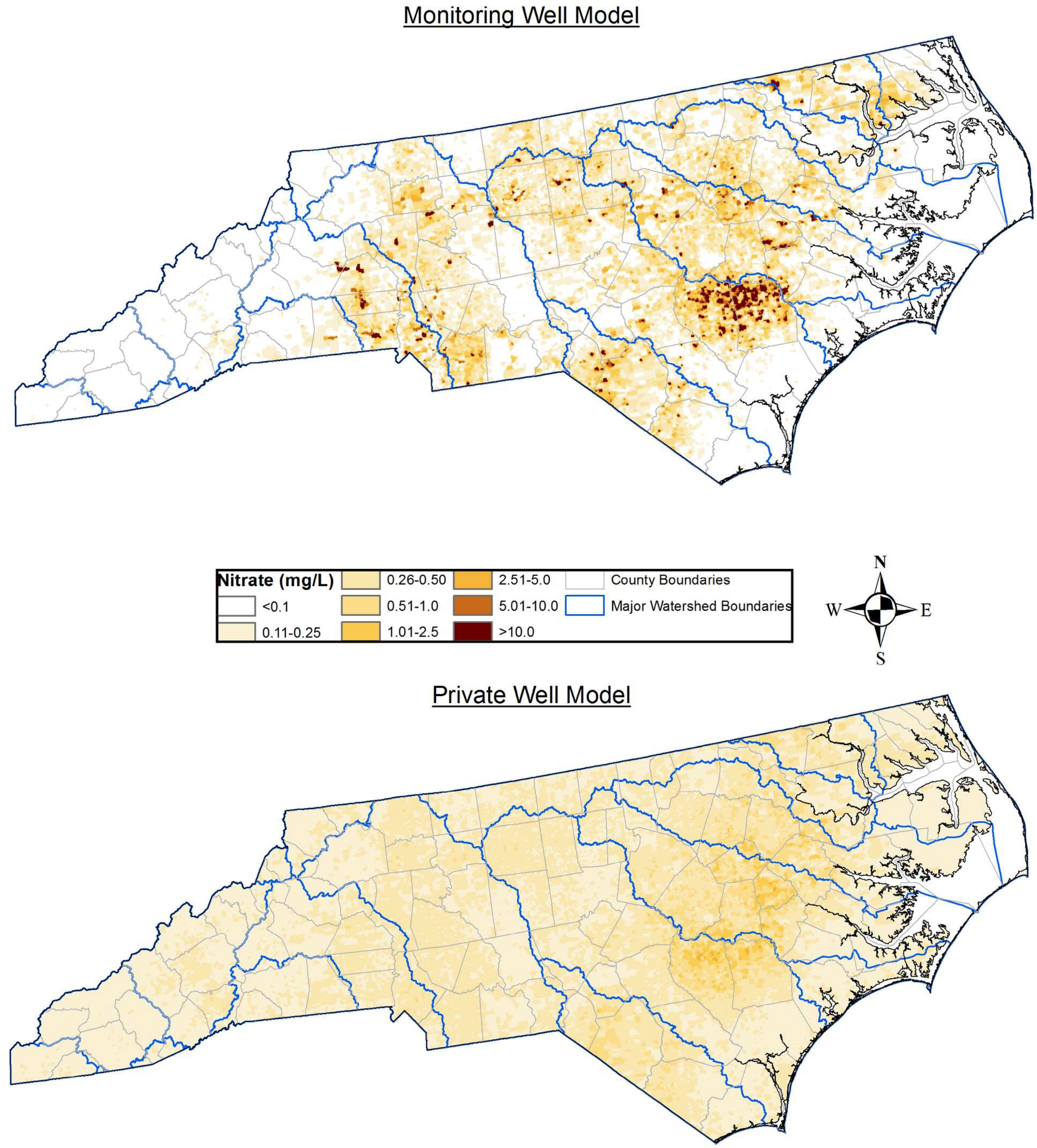


Figure S4. Land Use Regression results from the Constrained Forward Nonlinear Regression and Hyperparameter Optimization procedure for the monitoring and private well models. There are significant areas of predicted nitrate above 10 mg/L in the southeastern plains region for the monitoring wells. This area also has relatively widespread contamination above 1 mg/L in the private wells. Prediction variance should be used in conjunction with results at unmonitored locations.

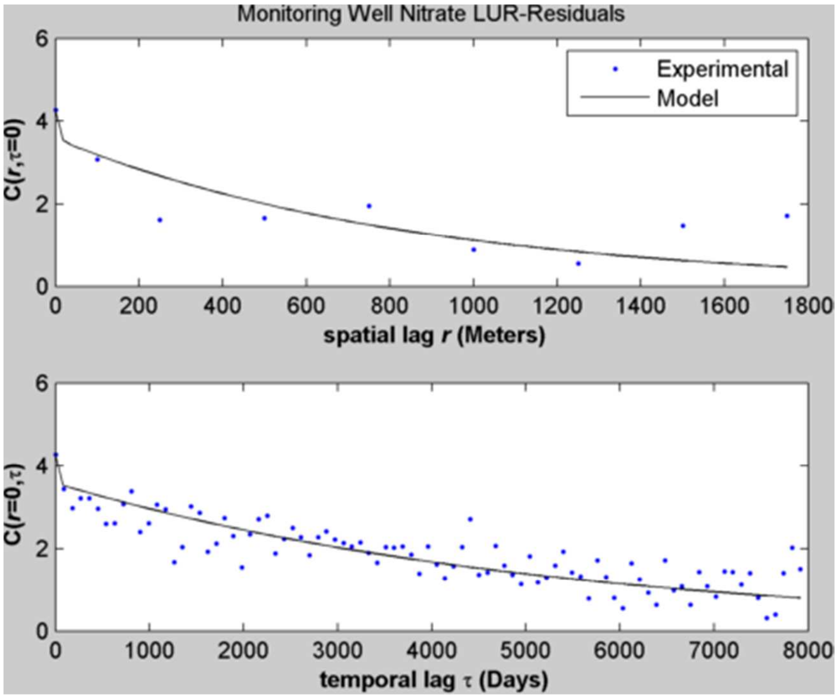


Figure S5. Monitoring well nitrate LUR residual experimental and modeled spatial (top) and temporal (bottom) covariance. The model is fit based on a least-squared fit with weights equal to the experimental covariance at the lag times the square root of the number of pairs used to calculate the covariance.

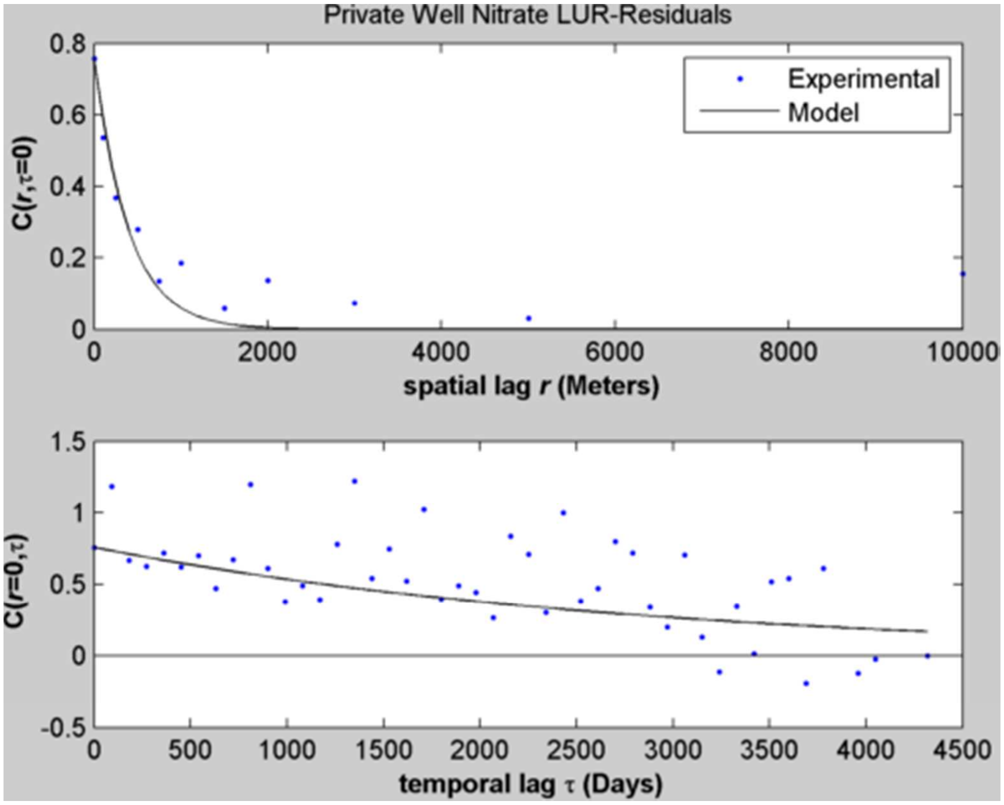


Figure S6. Private well nitrate LUR residual experimental and modeled covariance.

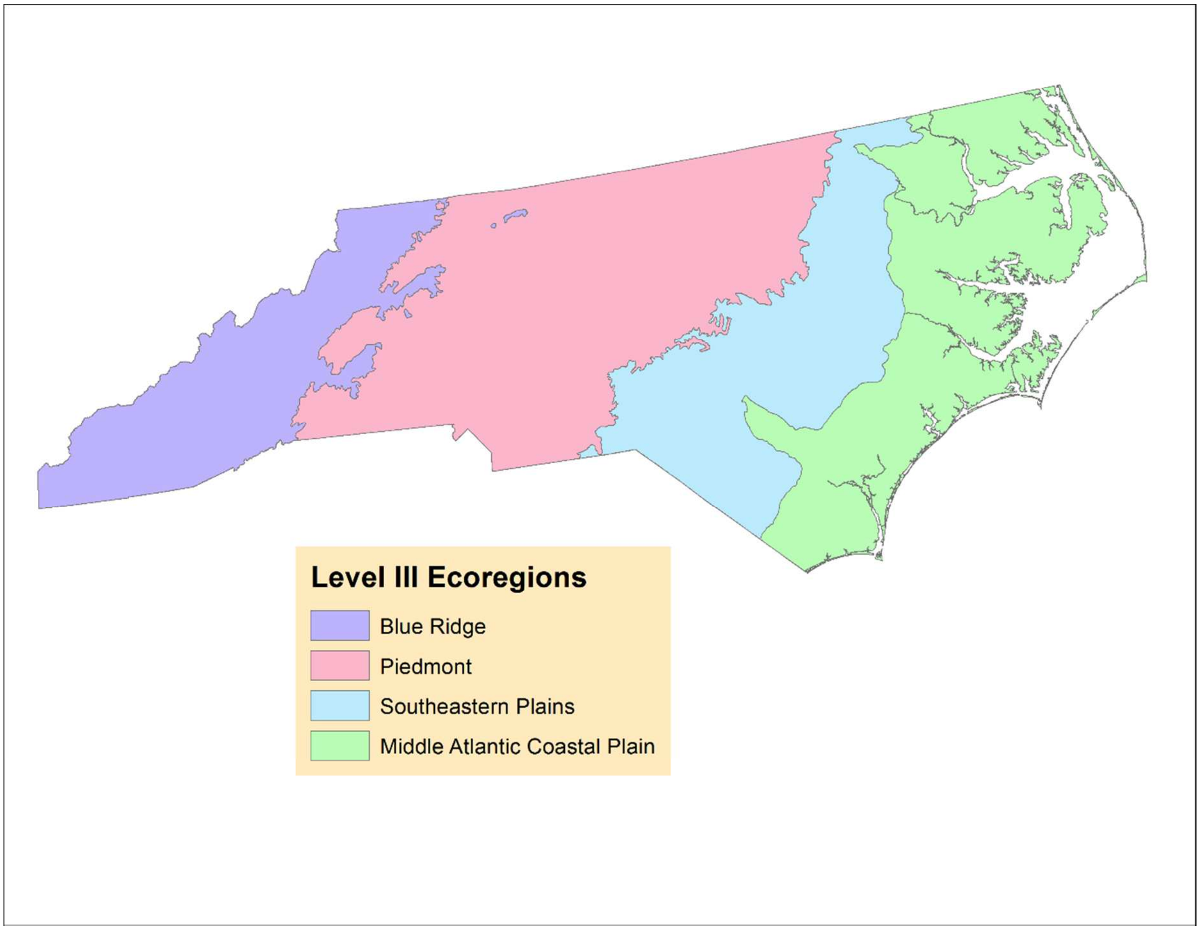
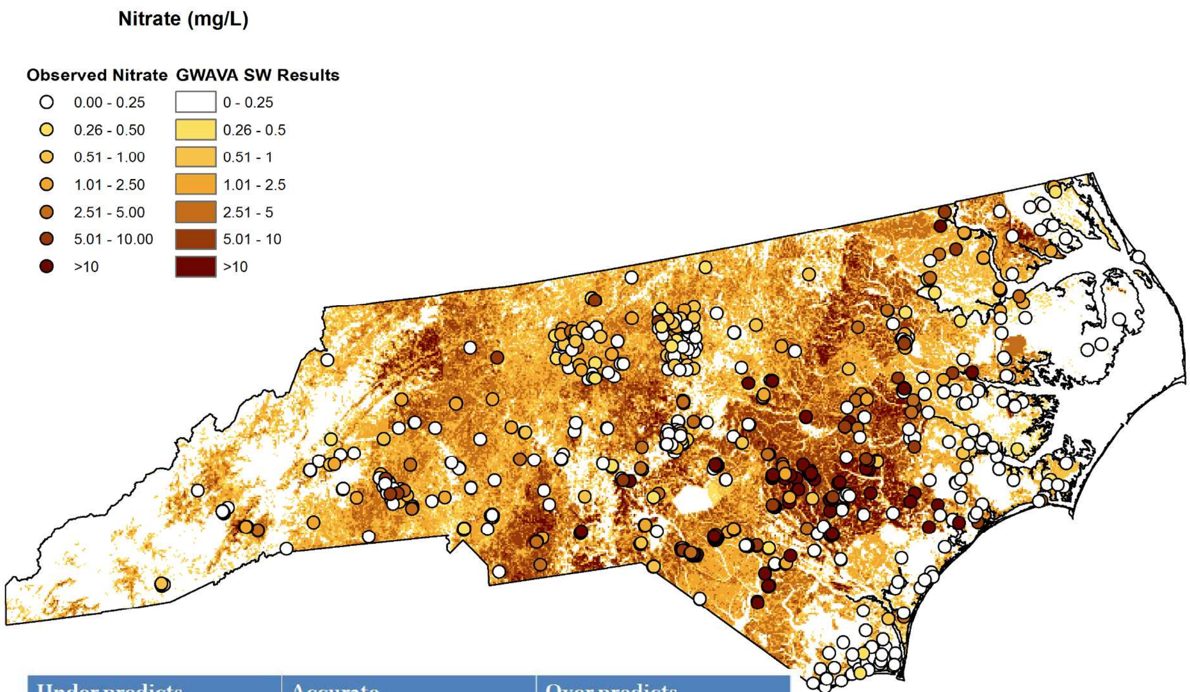


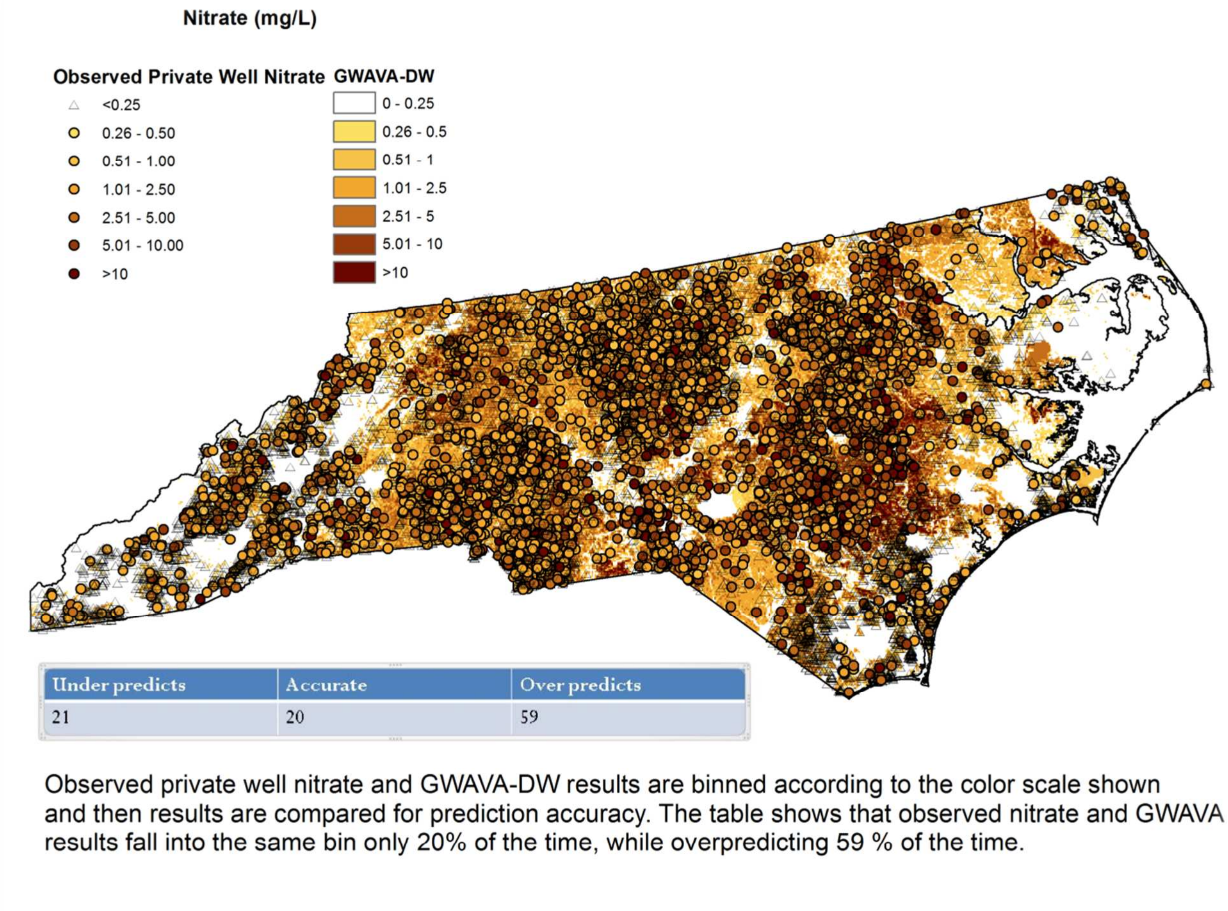
Figure S7. Level III Ecoregions in North Carolina defined by the US Environmental Protection Agency.



Under predicts	Accurate	Over predicts
33	19	48

Observed nitrate and GWAVA-SW results are binned according to the color scale shown and then results are compared for prediction accuracy. The table shows that observed nitrate and GWAVA results fall into the same bin only 19% of the time, while overpredicting almost half the time.

Figure S8. Observed monitoring well nitrate from this study overlaid with the GWAVA-SW model results.



Under predicts	Accurate	Over predicts
21	20	59

Figure S9. Observed private well nitrate from this study overlaid with the GWAVA-DW model results.

Movies

Movie S1: A movie showing the LUR-BME estimates for multiple days across the study time period is available for viewing and download at http://www.unc.edu/depts/case/BMElab/studies/KM_NO3_NC/

Movie S2: A movie showing the explanatory variables for the monitoring well LUR model is available for viewing and download at http://www.unc.edu/depts/case/BMElab/studies/KM_NO3_NC/

References for SI

(1) Ruddy, B. C.; Lorenz, D. L.; Mueller, D. K. County-Level Estimates of Nutrient Inputs to the Land Surface of the Conterminous United States , 1982 – 2001 Scientific Investigations Report 2006 – 5012 County-Level Estimates of Nutrient Inputs to the Land Surface of the Conterminous United States , 19. **2006**, 1982–2001.

(2) Hoos, A. B.; Mcmahon, G. Spatial analysis of instream nitrogen loads and factors controlling nitrogen delivery to streams in the southeastern United States using spatially referenced regression on watershed attributes (SPARROW) and regional classification frameworks. *Hydrol. Process.* **2009**, *23*, 2275–2294.

(3) Fry, J.; Xian, G.; Jin, S.; Dewitz, J.; Homer, C.; Yang, L.; Barnes, C.; Herold, N.; Wickham, J. Completion of the 2006 National Land Cover Database for the Conterminous United States. *Photogramm. Eng. Remote Sensing* **2011**, *77*, 858–864.

(4) Messier, K. P.; Akita, Y.; Serre, M. L. Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. *Environ. Sci. Technol.* **2012**, *46*, 2772–2780.

(5) Keil, A.; Wing, S.; Lowman, A. Suitability of Public Records for Evaludating Health Effects of Treated Sewage Sludge in North Carolina. *NC Med J.* **2011**, *72*, 98–104.

(6) Pradhan, S. S.; Hoover, M. T.; Austin, R. E.; Devine, H. A. *Potential Nitrogen Contributions from On-site Wastewater Treatment Systems to North Carolina ' s River Basins and Sub-basins*; Raleigh, North Carolina, 2007.

(7) Beven, K. J.; Kirkby, M. J. A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci.* **1979**, *24*, 43–69.

(8) Tague, C. L.; Band, L. E. RHESSys: Regional Hydro-Ecologic Simulation System—An Object-Oriented Approach to Spatially Distributed Modeling of Carbon, Water, and Nutrient Cycling. *Earth Interact.* **2004**, *8*, 1–42.

(9) Rodríguez-Lado, L.; Sun, G.; Berg, M.; Zhang, Q.; Xue, H.; Zheng, Q.; Johnson, C. A. Groundwater arsenic contamination throughout China. *Science* **2013**, *341*, 866–868.

(10) Miller, D.; White, R. A Conterminous United States Multi-Layer Soil Characteristics Data Set for Regional Climate and Hydrology Modeling. *Earth Interact.* **1998**, *2*.

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

(12) Kenny, J. F.; Barber, N. L.; Hutson, S. S.; Linsey, K. S.; Lovelace, J. K.; Maupin, M. A. Estimated Use of Water in the United States in 2005. *USGS Circ. 1344* **2005**.