Supplemental Materials

1. Methods

1.1 Explanatory variable dataset

We assembled a set of explanatory variables previously identified as important in explaining variation in well contamination by nitrate (Table S1). Nitrate contamination of groundwater is related to factors that affect the source of nitrate inputs, factors that affect the transport or downward movement of nitrate into groundwater, and attenuation factors that relate to the potential for denitrification as nitrate moves through the subsurface into groundwater (Gurdak and Qi 2012). Transport factors include the presence of confining layers that prevent downward movement of water, and soil characteristics that can reduce permeability and subsurface drainage. Attenuation factors include the presence of organic matter, dissolved oxygen concentrations in aquifers, poorly drained soil, and other factors that would lead to low oxygen conditions that facilitate the denitrification of nitrate and hence reduce the quantity of leached N that reaches groundwater (Nolan and Hitt 2006, Burow et al. 2010, Kellogg et al. 2010, Wick et al. 2012, MPCA 2013). We did not have information on hydraulic gradients or sub-surface lateral flow pathways so the model parameters only address factors related to the vertical transport of nitrate.

Depth is also an important predictor of groundwater nitrate contamination. Shallow wells and wells tapping aquifers with fewer overlying layers of bedrock are more likely to be affected by changes in contaminant loading at the surface (Burow et al. 2010, Gurdak and Qi 2012, Lichtenberg and Shapiro 1997, Nolan and Hitt 2006, Gardner and Vogel 2005, Tesoriero et al. 2004, Runkel et al. 2013). For each well, we estimated the vertical rank as the number of geologic formations above the tapped aquifer, counting down from the uppermost bedrock unit (Figure S1). Aquifer rank below surficial bedrock was found to be significant in previous studies on groundwater contamination in the region (Mubarak 2003, Harkanpar 2008).
Figure S1: Stratigraphic column for bedrock of southeastern Minnesota. Numbers indicate the vertical rank of each formation. Numbers were assigned to each well based on the aquifer identified in the driller’s records. We then estimate the difference between the rank of the uppermost (surface) bedrock at each well site (from surficial geology maps) and the rank of the tapped aquifer (aquifer rank variable). Cross-section modified from Mossler (2008) and Runkel et al. (2013).

Similar to previous studies, we assigned soil characteristics and surface nitrate loading values to each well based on a circular buffer around each well point (Tesoriero and Voss 1997, Nolan and Hitt 2006, Nolan et al. 2002, Gardner and Vogel 2005, Burow et al. 2010, Frans et al. 2012, Gurdak and Qi 2012). We acknowledge that a buffer around each well only captures vertical flow of water from the region above each well and does not explicitly address lateral flow paths, residence time, or aquifer volume, all of which may be important in predicting groundwater well contamination. Estimating the true recharge zone for each well requires information about flow paths and subsurface transport that are unknown in most study regions, including even the well-characterized geology of SE MN. Previous studies have estimated soil characteristics and land use and management variables within a 400 m to 3200 m radius around each
well point (Tesoriero and Voss 1997, Nolan and Hitt 2006, Nolan et al. 2002, Gardner and Vogel 2005, Burow et al. 2010, Frans et al. 2012, Gurdak and Qi 2012). We experimented with different radii and while the direction of the relationship between nitrate load and well contamination remained positive, the most positive relationship between surface land use (and associated nitrate load) was found at 500m. A 500 m radius was also used in studies of groundwater nitrate contamination published by Nolan et al. (2002), Nolan and Hitt (2006), Burow et al. (2010), Frans et al. (2012), and Gurdak and Qi (2012). We generated 500 m buffers around each well in ArcGIS and then used spatial joins and zonal statistics to assign surface nitrate load, geologic attributes, and soil characteristics to each well (Table S1).

Table S1. Variables used in fitting the nitrate contamination model. In addition to the continuous variables listed here we also investigated the significance of nominal variables for surficial geology type, bedrock geology unit, presence of confining layers (aquitards), presence of grout in well construction, and aquifer type in predicting nitrate contamination. Soils data were obtained from the USDA Natural Resources Conservation Service (NRCS) Soil Survey Geography database (SSURGO). Unless otherwise noted, the soils data are based on the dominant component for all known layers (0-999 m).

| Explanatory variables considered in model selection | Max | Min | Mean | SD | Source |
|---------------------------------------------------|-----|-----|------|----|--------|
| Sum of the surface load of nitrate-N around each well based on 2007 land use land cover (kg nitrate-N/ha/yr) | 11,429 | 315 | 2,906 | 2,158 | See Table S2 |
| Annual average precipitation (mm/yr) based on 2007-2012 data. | 1001 | 797 | 868 | 41 | PRISM climate group |
| Aquifer rank below the uppermost bedrock unit | 6 | 0 | 1.7 | 1.3 | See Figure S1 |
| Well depth (m) | 850 | 11 | 283 | 144 | County well index |
| Drift thickness and permeability (scale 1-3, with 3 being most permeable) | 2.97 | 0.95 | 1.60 | 0.34 | Minnesota Geologic Survey (categorical variable representing relative permeability of unconsolidated sediments) |
| Average percent clay in well recharge zone | 38.40 | 0.02 | 20.13 | 4.65 | NRCS SSURGO |
| Average percent organic matter in well recharge zone | 21.38 | 0.01 | 1.27 | 0.95 | NRCS SSURGO |
| Average percent sand in well recharge zone | 90.50 | 0.11 | 28.72 | 15.99 | NRCS SSURGO |
| Average drainage class in well recharge zone (refers to the natural, undisturbed soil condition and describes how quickly water drains from the soil profile) | 6.61 | 1 | 4.40 | 0.81 | NRCS SSURGO |
Average soil water content in well recharge zone  433.3  24.1  238.8  57.4  NRCS SSURGO

Percent of pixels in well recharge zone classified as "well drained" (hydrologic soil groups A and B)

To estimate the annual average load of nitrate to the recharge zone around each well under each scenario, we used land use-specific nitrate export coefficients. For the baseline land-use scenario we used a reclassified version of the 2007 Cropland Data Layer produced by the U.S. Department of Agriculture to define the extent and spatial pattern of land use around each well. Export coefficients for each major land cover or crop type were estimated from a literature review and represent the nitrate available for leaching through surface or subsurface pathways due to inputs from crop residues, fertilizers, and atmospheric deposition (Reckhow and Simpson 1980). Where possible, nitrate export values were adapted from field studies in similar soil and climatic conditions. For example, values for corn represent annual export of nitrate observed in Minnesota watersheds with representative drainage, fertilizer use, and tillage (Table S2).

Table S2: Nitrate export coefficients assigned to land use codes in the 2007 and 2012 Cropland Data Layers. Export coefficients are widely available in the literature, with several meta-analyses (Reckhow et al. 1980, Frink 1991, Parn et al. 2012) providing average values for different vegetation classes. Where available, retention coefficients from regional (Southeastern MN) studies were used (Randall and Iragavarapu 1995, Randall and Mulla 2001, Randall et al. 1997, Weed and Kanwar 1996). In the absence of local data, values from national meta-analyses were used or averaged across multiple national and international studies matching flow paths (inclusive of surface and sub-surface transport), constituent type (values were used for Nitrate-N), and climate (Reckhow et al. 1980, Frink 1991, Parn et al. 2012). This approach differs from previous statistical models of groundwater contamination which use proxies for nutrient load such as the percentage of agricultural or developed land, or counts of the number of septic systems or number of confined livestock operations in a well recharge zone (Nolan et al. 2002, Gardner and Vogel 2005, Wick et al. 2012, Gurdak and Qi 2012, Liu et al. 2005). The advantage of using land-cover specific nitrate load values instead of a per area estimate of acreage of a given land cover type is that our approach better captures relative differences among different land cover types. Also our metric for nitrate load is likely to better approximate true nitrate load to the surface because values are adapted from field studies under similar conditions and reflect agricultural practices of the region.

| Land use Code | Land Use Description | Nitrate-N load in kg/ha/yr |
|---------------|----------------------|---------------------------|
| 1             | water                | 0.00                      |
| 2             | developed            | 9.60                      |
| 3             | barren land coniferous forest | 0.50                   |
| 4             | deciduous forest and mixed | 3.72                    |
| 5             | mixed                | 3.72                      |
We constructed multiple logistic regression models for each nitrate contamination threshold (4 ppm and 10 ppm). Logistic regression predicts the probability that a sample falls within a given response category and does not require the response variable to be normally distributed (well nitrate concentration is highly skewed due to the significant number of non-detects). Logistic regression has been applied in many previous studies on groundwater nitrate contamination, most commonly using a contamination threshold of 1-3 ppm to identify wells that exceed background or naturally-occurring levels of nitrate (Tesoriero et al. 2004, Gardner and Vogel 2005, Tesoriero and Voss 1997, Liu et al. 2005, Gurdak and Qi 2012).

In logistic regression, the mean response at any level of input is the probability of being in a category (above or below a threshold), where

\[ P = \frac{e^{(b_0 + bx)}}{1 + e^{(b_0 + bx)}} \]

and P is the probability of exceeding a given threshold, \( b_0 \) is a constant and bx is the vector of slope coefficients and explanatory variables. In order to transform the probability function so that a linear function can be fitted to the explanatory variables a logit transformation is applied. The logit function is

\[ \ln \left( \frac{P}{1 - P} \right) = b_0 + bx \]

With this transformation, the logit is linearly related to the model parameters and standard linear regression tools can be used to estimate values for \( b_0 \) and bx. Explanatory variables are fit to the logit function and then converted back into probability units.

### 1.3 Selecting variables for the logistic regression model
For the subset of wells with recent chemistry data we screened all candidate explanatory variables for significance in a logistic regression model based on the 4 ppm and 10 ppm thresholds. We first noted that the nominal variable representing the boundary of thick unconsolidated glacial deposits or drift was significant in classifying contaminated and un-contaminated wells. The drift boundary marks a region where the depth to bedrock is greater than fifty feet, compared to the bedrock-dominated eastern region where drift is thin and patchy (Figure 1). Only nine wells located within the boundary of glacial drift had nitrate concentrations greater than 4 ppm (or 2% of all wells in the drift zone), whereas 16% of all wells outside the zone of glacial drift had nitrate levels above 4 ppm. Due to the observation that wells within the drift boundary appear to be fairly well protected from nitrate contamination we removed these wells from the sample set based on the conservative assumption that these wells are less sensitive to changes in land use and corresponding surface loading of nitrate.

To screen the remaining candidate variables for inclusion in the final logistic regression models, we used backwards stepwise regression using the minimum Akaike information criterion (AIC) and minimum Bayesian information criterion (BIC) in JMP Pro 10 (SAS Institute Inc.). Additionally, we used the “best glm” (general linearized model) package in the statistical software package R which uses a cross-validation approach to identify significant explanatory variables and parameter estimates. In a cross-validation approach, subsets of data are excluded as test sets, models and significant parameters are estimated based on training sets, and then adjusted based on performance relative to the test sets.

1.4 Evaluation of model fit

We used two statistical tests to evaluate the predictive ability of the multivariate logistic regression model. The Lack of Fit or Goodness of Fit test addresses whether more complex terms are needed in the current model or if there appears to be enough information with the existing variables. The null hypothesis for this test is that the model fits the data, therefore a higher p-value indicates a well-calibrated model. For both the 4ppm and 10ppm logistic regression models, the lack of fit p-value was 1.0 indicating that there is little to be gained by introducing additional variables to the model.

We also used Receiver Operating Characteristic (ROC) curves to evaluate each model (Figure S2). Curves which are further to the left and higher on the vertical axis have greater predictive capacity. The accuracy of the model is represented as the tradeoff between specificity or the rate of false positives and sensitivity, which is the rate of true positives. Accuracy is measured by the area under the ROC curve (AUC), where an area of 1 represents a perfect model and an area of .5 represents zero predictability. The AUC estimates for the 4 ppm and 10 ppm models are 0.77 and 0.86, respectively, represent fair to good predictive power and meet or exceed AUC estimates from previous groundwater models (Gurdak and Qi 2012).
Figure S2. ROC curves for the 4ppm model (top) and 10ppm model (below). Axes represent the rate of false positives (specificity) and true positives (sensitivity). The greater the area under the curve, the better the predictive ability of the model (AUC 4ppm = 0.77, AUC for 10ppm = 0.86). The straight line is drawn at a 45 degree angle tangent to the ROC Curve and is useful in identifying the probability threshold which balances the frequency of false negatives and false positives. The 45 degree angle tangent to the ROC curve representing this point for the 4 ppm model corresponds with a 0.16 cutoff and a 0.08 cutoff for the 10 ppm model.

1.5 Adjusting the probability cutoff used to assign wells to categories

The output of the logistic regression model is a continuous probability value, therefore in order to translate probabilities into a nominal contamination prediction for each well, we needed to select the probability threshold to use in assigning wells to each response category (contaminated or uncontaminated). This threshold is related to the frequency of positive detections, such that low frequency events are associated with lower probability thresholds. The appropriate threshold is one that both minimizes total error and comes closest to estimating the correct frequency of positive response events without overestimating contamination.
Figure S3 plots a range of probability thresholds based on how these values influence the total error rate and rate of positive detection. For the 4 ppm model, a probability threshold of 0.27 predicted 98% of the total number of contaminated wells (true positives + false positives/total number of observed positives), with a 15% total error rate (number of false positives + the number of false negatives/total samples). The lowest total error rate was at a cutoff probability of 0.36, with 12% error. At this level, the model estimates only 52% of the total number of actual contaminated wells. For the 10 ppm model, a 0.23 cutoff predicted 100% of the total number of actual contaminated wells and yielded a 6% error rate. Minimizing total error rate to the lowest value of 5% selects a cutoff of 0.39 which predicts only 44% of total contamination. The penalty of raising the probability cutoff to positive rate detection was greater than the penalty to total error, therefore we selected probability cutoffs that minimized the total number of misclassifications (false positives and false negatives) while coming closest to a “true” prediction of the total number of contaminated wells without overestimating contamination. We selected probability cutoffs of 0.28 for the 4 ppm model and 0.24 for the 10 ppm model which slightly under-predicts the total number of contaminated wells. Note that these thresholds are greater (more conservative) than the probability cutoffs identified in the ROC curves (Figure S2) which represent an equal balance between the frequency of false negatives and false positives.
Figure S3: Plot of the total error rate and the rate of positive detection (correct number of contaminated wells) for each probability threshold. The vertical line marks the threshold where the model correctly predicts the total number of contaminated wells.

References:

Burow, K. R., B. T. Nolan, M. G. Rupert, and N. M. Dubrovsky. 2010. Nitrate in groundwater of the United States, 1991–2003. Environmental Science & Technology 44:4988-4997.

Frans, L.M., M.G. Rupert, C.D. Hunt Jr., K.D. Skinner. 2012. Groundwater quality in the Columbia Plateau, Snake River Plain, and Oahu basaltic-rock and basin-fill aquifers in the northwestern United States and Hawaii, 1992-2010: U.S. Geological Survey Scientific Investigations Report 2012-5123.

Frink, C. R. 1991. Estimating nutrient exports to estuaries. Journal of Environmental Quality 20:717-724.

Gardner, K. K., R. M. Vogel. 2005. Predicting ground water nitrate concentration from land use. Ground Water 43:343-352.

Gurdak, J. J., S. L. Qi. 2012. Vulnerability of recently recharged groundwater in principle aquifers of the United States to nitrate contamination. Environmental Science & Technology 46:6004-6012.

Harkanpar, K. 2008. Nitrate pollution externality associated with increased corn demand for ethanol: A case study of Olmsted County in Minnesota. The 82nd Annual Conference of the Agricultural Economics Society, Royal Agricultural College.

Kellogg, D., A. J. Gold, S. Cox, K. Addy, and P. V. August. 2010. A geospatial approach for assessing denitrification sinks within lower-order catchments. Ecological Engineering 36:1596-1606.

Lichtenberg, E., L. K. Shapiro. 1997. Agriculture and nitrate concentrations in Maryland community water system wells. Journal of Environmental Quality 26:145-153.
Liu, A., J. Ming, and R. O. Ankumah. 2005. Nitrate contamination in private wells in rural Alabama, United States. Science of the Total Environment 346:112-120.

MPCA, Minnesota Pollution Control Agency. 2013. The condition of Minnesota's groundwater, 2007 - 2011. August, 2013.

Mossler, J.H. 2008. Paleozoic stratigraphic nomenclature for Minnesota: Minnesota Geological Survey Report of Investigations 65, 76 p., 1 pl.

Mubarak K. 2003. Utilizing GIS to Estimate The Quantity and Distribution of Nitrate-Nitrogen And Chloride in Olmsted County Groundwater. St. Mary's University of Minnesota, Winona, MN.

Nolan, B. T., K. J. Hitt. 2006. Vulnerability of shallow groundwater and drinking-water wells to nitrate in the United States. Environmental Science & Technology 40:7834-7840.

Nolan, B. T., K. J. Hitt, and B. C. Ruddy. 2002. Probability of nitrate contamination of recently recharged groundwaters in the conterminous United States. Environmental Science & Technology 36:2138-2145.

Pärn J., G. Pinay, and Ü. Mander. 2012. Indicators of nutrients transport from agricultural catchments under temperate climate: A review. Ecological Indicators 22:4-15.

Randall G., T. Iragavarapu. 1995. Impact of long-term tillage systems for continuous corn on nitrate leaching to tile drainage. Journal of Environmental Quality 24:360-366.

Randall G., T. Iragavarapu, and B. Bock. 1997. Nitrogen application methods and timing for corn after soybean in a ridge-tillage system. Journal of Production Agriculture 10:300-307.

Randall G. W., D. J. Mulla. 2001. Nitrate nitrogen in surface waters as influenced by climatic conditions and agricultural practices. Journal of Environmental Quality 30:337.

Reckhow K. H., M. N. Beaulac, and J. T. Simpson. 1980. Modeling phosphorus loading and lake response under uncertainty: A manual and compilation of export coefficients. US EPA.

Reckhow, K., J. Simpson. 1980. A procedure using modeling and error analysis for the prediction of lake phosphorus concentration from land use information. Canadian Journal of Fisheries and Aquatic Sciences 37:1439-1448.

Runkel, A.C., J.R. Steenberg, R.G. Tipping, A.J. Retzler. 2013. Geologic controls on groundwater and surface water flow in southeastern Minnesota and its impact on nitrate concentrations in streams. Minnesota Geologic Survey.

Tesoriero, A. J., F. D. Voss. 1997. Predicting the probability of elevated nitrate concentrations in the Puget Sound Basin: implications for aquifer susceptibility and vulnerability. Ground Water 35:1029-1039.

Tesoriero, A. J., T. B. Spruill, and J. L. Eimers. 2004. Geochemistry of shallow ground water in coastal plain environments in the southeastern United States: Implications for aquifer susceptibility. Applied Geochemistry 19:1471-1482.
Weed D., R. Kanwar. 1996. Nitrate and water present in and flowing from root-zone soil. Journal of Environmental Quality 25:709-719.

Wick, K., C. Heumesser, and E. Schmid. 2012. Groundwater nitrate contamination: Factors and indicators. Journal of Environmental Management 111:178-186.