JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

AMERICAN WATER RESOURCES ASSOCIATION

APPLYING STATISTICAL CAUSAL ANALYSES TO AGRICULTURAL CONSERVATION: A CASE STUDY EXAMINING P LOSS IMPACTS[1]

Song S. Qian and R. Daren Harmel[2]

ABSTRACT: Estimating the effect of agricultural conservation practices on reducing nutrient loss using observational data can be confounded by factors such as differing crop types and management practices. As we may not have the full knowledge of these confounding factors, conventional statistical meta-analysis methods can be misleading. We discuss the use of two statistical causal analysis methods for quantifying the effects of water and soil conservation practices in reducing P loss from agricultural fields. With the propensity score method, a subset of data was used to form a treatment group and a control group with similar distributions of confounding factors. With the multilevel modeling method, data were stratified based on important confounding factors, and the conservation practice effect was evaluated for each stratum. Both methods resulted in similar estimates of the conservation practice effect (total P load reduction avg. ~70%). In addition, both methods show evidence of conservation practices reducing the incremental increase in total P export per unit increase in fertilizer application. These results are presented as examples of the types of outcomes provided by statistical causal analyses, not to provide definitive estimates of P loss reduction. The enhanced meta-analysis methods presented within are applicable for improved assessment of agricultural practices and their effects and can be used for providing realistic parameter values for watershed-scale modeling.

(KEY TERMS: causal inference; conservation practices; multilevel modeling; observational data; propensity score.)

Qian, Song S. and R. Daren Harmel, 2015. Applying Statistical Causal Analyses to Agricultural Conservation: A Case Study Examining P Loss Impacts. *Journal of the American Water Resources Association* (JAWRA) 1-11. DOI: 10.1111/1752-1688.12377

INTRODUCTION

Implementing soil and water conservation practices is an important part of sustainable agriculture. In response to increased funding of conservation programs in the Farm Security and Rural Investment Act of 2002, the U.S. Department of Agriculture initiated the Conservation Effects Assessment Program (CEAP) to provide scientific understanding of the impacts and benefits of conservation practices (Duriancik *et al.*, 2008). CEAP produced a comprehensive bibliography on available literature, a suite of mechanistic models, and a series of model simulations on the effects of conservation practices at a watershed level for many regions of the United States. Many parameters used in watershed models are based on studies in the 1980s (e.g., Beaulac and Reckhow, 1982). Consequently, CEAP also facilitated collection of field-scale nutrient loading data for assessing vari-

[1]Paper No. JAWRA-14-0144-P of the *Journal of the American Water Resources Association* (JAWRA). Received July 1, 2014; accepted October 17, 2015. **Discussions are open until six months from issue publication**.

[2]Assistant Professor (Qian), Department of Environmental Sciences, The University of Toledo, 2801 West Bancroft Street, MS604, Toledo, Ohio 43606; and Supervisory Agricultural Engineer (Harmel), Agriculture Research Service, U.S. Department of Agriculture, Temple, Texas 76502 (E-Mail/Qian: song.qian@utoledo.edu).

TABLE 1. Variables (data types) in the MANAGE Database.

Variables	Definition
Watershed ID	Name of the watershed
Location (city, state)	City and state/province of the study (occasionally only a county or region was specified)
State	US state (or Canadian province) included to aid state-specific queries
Location (Lat, Lon)	Latitude and longitude of the study
Date	Beginning and end of period with annual nutrient load data (not necessarily the entire study duration)
Watershed years	Product of the number of monitored watersheds and the number of years with annual nutrient load data (ws-yr)
Land use	Identification of crop or vegetation type(s) and crop rotation
Tillage	Description of the tillage management divided into four options: no-till, conservation, conventional, or pasture
Conservation practice	Five options: waterway, terrace, filter strip, riparian buffer, or contour farming
Dominant soil type	Soil textural class and soil series
Hydrologic soil group	NRCS hydrologic soil group (HSG) classification (A, B, C, or D)
Soil test P	Maximum and minimum soil test P values for records with multiple watersheds or multiple years (ppm)
Soil test P extractant	Extractant used to determine soil test P
Land slope	Maximum and minimum land surface slopes for records with multiple watersheds (%)
Watershed size (ha)	Maximum and minimum watershed sizes for records with multiple watersheds
Type of fertilizer applied	Macro-nutrient composition (N-P-K)
Fertilizer application method	Fertilizer application method divided into four options: surface, injected, incorporated, or other
Annual maximum, minimum, and average values are provided for the following categories when specified	
N applied	The total annual amount of N applied to watershed(s) from all fertilizer sources (kg/ha/yr)
P applied	The total annual amount of P applied to watershed(s) from all fertilizer sources (kg/ha/yr)
Precipitation	Annual precipitation (mm/yr)
Runoff	Annual runoff (mm/yr)
Soil loss	Annual soil loss (kg/ha/yr)
Dissolved N	The total amount of dissolved N lost from the watershed(s) (kg/ha/yr)
Particulate N	The total amount of N lost from the watershed(s) in a particulate form (kg/ha/yr)
Total N	Total N load was specified in a number of the publications. If the total N load was not specified, it was determined as the sum of dissolved and particulate N loads, when both were specified (kg/ha/yr)
Dissolved P	The total amount of dissolved P lost from the watershed(s) (kg/ha/yr)
Particulate P	The total amount of P lost from the watershed(s) in a particulate form (associated with sediment) (kg/ha/yr)
Total P	Total P load was specified in a number of the publications. If the total P load was not specified, it was determined as the sum of dissolved and particulate P loads, when both were specified (kg/ha/yr)
Analysis technique	Techniques used to determine dissolved, particulate, and total N or P composition in runoff
Flow indication	Indication of the flow transport mechanisms (total, surface, base) addressed

ous agricultural conservation practices. To support CEAP and other national modeling and assessment efforts, the Measured Annual Nutrient loads from AGricultural Environments (MANAGE) database was created, summarizing 40 studies reported in peer-reviewed publications (Harmel *et al.*, 2006). The database provides readily accessible, easily queried watershed characteristic, nutrient export, and concentration data to guide policy and management decisions based on comparative nutrient load information from various land management alternatives. Harmel *et al.* (2008) expanded the database and added runoff concentration data. The data set includes studies conducted in 18 states, mainly in the Midwest states along the Mississippi River. It contains 1,677 watershed-years of data for various agricultural land uses, including pasture/rangeland, corn, various crop rotations, wheat/oats, barley, citrus, vegetables, sorghum, soy beans, cotton, fallow, and peanuts. Harmel *et al.* (2008) provides detailed information on spatial distribution of study sites and summarizes the variables (data types) included (Table 1). MANAGE compiles and summarizes measured annual nitrogen (N) and phosphorus (P) load and concentration data representing field-scale transport from agricultural and forest land, and drainage studies were recently added to further expand the database (Christianson and Harmel, 2015).

In the Farm Bill, federal funds are allocated for USDA conservation programs to provide financial assistance (i.e., for implementing practices) or technical assistance (i.e., for designing practices). Practices designed to control sediment loss were previously assumed to also control P loss from agricultural fields, and they do control particulate P losses; however, their effects on soluble P fate and transport has received intensive research attention in recent years. No-till has been promoted since the 1980s to decrease sediment loss from fields. Similarly, terraces and/or grassed waterways are installed in areas with concentrated flow and likely prone to gully erosion. These practices also reduce particulate P loading. Crop rotation and cover crops are increasingly used to control nutrient losses and improve soil quality. Other practices to limit nutrient loss include: contour tillage, reduced tillage, applying P based on soil test P recommendations, and avoiding application on frozen soils. See Sharpley *et al.* (1994, 2000), Sharpley and Withers (1994), and Smith *et al.* (2015) for more discussion of the impacts of conservation practices on P loss.

Initial analysis by Harmel *et al.* (2006) showed that annual particulate and total N loads can be reduced by using conservation tillage or no-till methods, but conservation effects were less evident for dissolved N and all P forms. In addition, some results were seemingly counterintuitive. For example, average annual N and P loads from fields with one or more conservation practices were often larger than from fields without conservation practices. The present analysis focused on the explanation of the apparent counterintuitive impacts of conservation practices on P loads.

Many previous studies on the effectiveness of conservation practices were focused on model simulations (Cho *et al.*, 2010; USDA-NRCS, 2011a, b, 2012a, b, c; White *et al.*, 2014). These simulation studies focused on watershed scale effects and were often based on an assumed level of field-scale reduction rate. As a result, field-scale nutrient reduction is often the basis of watershed-scale modeling studies. As randomized experiments for estimating field-scale nutrient reduction rates are rarely available, we must often use either results from field experiments limited to specific conditions or cross-sectional data such as MANAGE. As in almost all cross-sectional data, MANAGE represents a form of observational data and can be unsuitable for causal inference using conventional statistical approaches. In this article, we discuss the use of two statistical causal analysis methods (propensity score matching and multilevel modeling) designed to quantify causal effects, thereby expanding their applicability.

Statistical Causal Analyses

With databases such as MANAGE, fields with and without conservation practices can be compared to derive the practice effect. However, conventional statistical meta-analysis methods can be misleading because of confounding factors. For example, conservation practices are not applied randomly but are typically applied to fields that are more prone to sediment and nutrient loss. Similarly, data in MANAGE indicate that fields with conservation practices are more likely to receive higher fertilizer applications (Figure 1), which can mask the effect of the variable of interest (conservation practice). As a result, directly comparing nutrient loads between fields with and fields without conservation practices using observational data can be misleading. Fertilizer application is only one of many potential factors that can confound the result of comparing nutrient loads. R.A. Fisher recognized this problem when studying data from agricultural experiments, and he proposed the use of randomized experiments for causal inference (Fisher, 1971). However, randomized experiments are not always feasible. Observational data, such as MANAGE, are often the main source of information in many fields (e.g., social sciences and economics). Causal inference using observational data requires a set of specialized tools to select subsets of data to form a treatment subset and a control subset.

In our case, the treatment subset is a subset of all fields with one or more conservation practices, and the control subset is a subset of all fields without any conservation practices. These two subsets should differ only with respect to the conservation practices. As a result, difference in mean P loads from these two subsets can be attributed to the implementation of conservation practice(s). To ensure that the two subsets are "similar," we can use the propensity score matching method (Rubin, 2006). However, the propensity score method requires that we include all confounding factors. In many cases, researchers did not know or did not record all confounding factors when collecting data. As a result, we cannot definitely show that the basic assumption of the propensity score matching method is met in any specific study.

Propensity Score

The propensity score method is an approach for subsetting and matching (Rosenbaum and Rubin, 1983). That is, in the absence of randomized experimental data, we may identify two subsets of fields that have the same distributions on all observed covariates (confounding factors) but differ only in treatment assignment. If there are no other unobserved covariates that can predict treatment assignment (hidden confounding factors), we can consider that the treatment assignment is effectively random and can be used for causal analysis. The propensity score method has a long history in social science research, and some important results are included in Rubin (2006).

A propensity score is the conditional probability of a subject (a field) receiving the treatment (conserva-

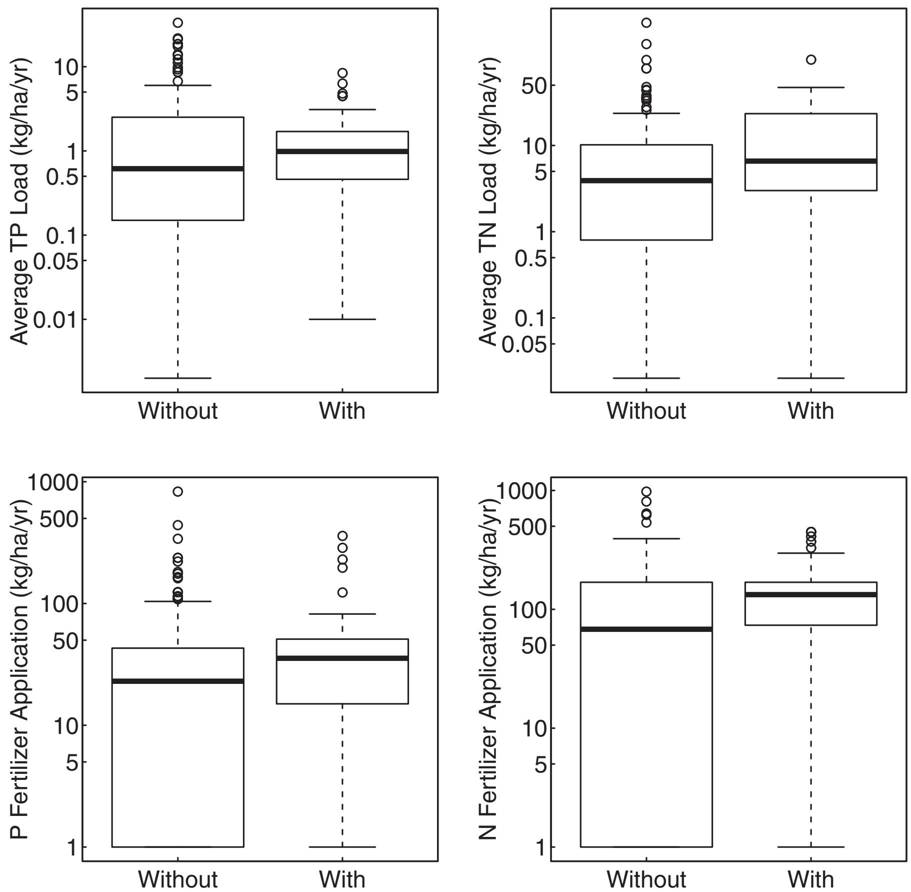

FIGURE 1. Box-Whisker Plots Comparing Fields with and without Conservation Practices; the Higher Mean Total P and Total N Loads from Fields with Conservation Practices (top row) Can Be Partially Explained by the Larger Fertilizer Applications to Fields with Conservation Practices.

tion practice) given all observed covariates. In a randomized experiment, the probability of a field receiving a conservation practice is determined by a flip of a coin, or 0.5; however, with observational data, some fields are more likely to have conservation practices than others (e.g., fields with higher fertilizer applications are more likely to have conservation practices, Figure 1), thereby the propensity score is not 0.5. Statistically, we can use covariates to "predict" the likelihood whether conservation practices are applied to a field. This likelihood is the propensity score. Using the propensity score, we match a field with conservation practices to a field without conservation practices if the two fields have similar propensity scores. By itself, a propensity score is meaningless. However, mathematical theory and experience have shown that when grouping subjects with similar propensity scores, the treatment and control groups will have similar covariate distributions (Gelman and Hill, 2007). In other words, propensity score matching retrospectively creates a control and a treatment group that can be considered as "randomized."

In practice, the propensity score is estimated by fitting a logistic regression model (Qian, 2010; Chapter 8) using all available covariates to predict the treatment assignment (Gelman and Hill, 2007, Section 10.3). In this case, we create a binary variable T to represent whether a field received conservation practices ($T = 1$) or not ($T = 0$). A logistic regression model is fit using T as the binary response and all available covariates as predictors. The model predicted probability of $T = 1$ for an observation is the propensity score for this observation. For each observed field in the treatment group, we match one or more observed fields from the control group which have a similar predicted probability of $T = 1$. Not all observed fields have matches. Consequently, the propensity score matched new dataset is a subset of the original data.

The resulting data have balanced covariates in the treatment and control groups; therefore, we can compare the mean P loads of the two groups either using a t-test or using model-based adjustments (regression). When using a t-test, we are fitting a regression model with the treatment indicator variable T as the only covariate:

$$y_i = \beta_0 + \theta T_i + \varepsilon_i \tag{1}$$

where $i = 1, \ldots n$, $T_i = 0$ or 1 represents the ith field is in the control or treatment group, β_0 is the mean of the response for fields without conservation practices (control) and θ is the estimated treatment effect, and ε_i is the residual term, assumed to follow a normal distribution with mean 0 and a constant variance. The treatment effect θ is the difference in mean P loads between fields with and without conservation practices. Because the P load is typically (natural) log transformed to conform to the normality assumption (van Belle, 2008), the difference represents a multiplicative factor in the original scale. Expressed in original scale, the effect θ is a multiplicative factor of e^{θ}. For example, $\theta = -0.8$ represents a factor of 0.45. This means that P loads from fields with conservation practices is 0.45 of the loads from fields without conservation practices (or 55% reduction).

When important covariates are available, we can control their effects on P loads by adding them to the t-test of Equation (1):

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \theta T_i + \varepsilon_i \tag{2}$$

where $x_1, \cdots, x_k$ are covariates, and θ is known as the causal effect after confounding factors are "controlled." Controlling for specific confounding factors can often change the estimated θ, as the meaning of θ is now the difference between treatment and control for fields with identical controlled factors (e.g., planted with the same crop and with the same fertilizer application rate). When a confounding factor (e.g., x_k) affects the treatment effect, interactions between T and x_k can be considered:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{k,i} + \alpha x_{ki} T_i + \theta T_i + \varepsilon_i \tag{3}$$

The treatment effect is now $\alpha x_{ki} + \theta$, varying as a function of the confounding factor x_k. The effect of x_k (slope) is now $\beta_k + \alpha T_i$, a function of the treatment. In other words, the interaction effect not only changes the treatment effect, but also the effect of the confounding factor. With the interaction term, θ is now the effect when $x_k = 0$. To ease interpretation, we often center predictors by subtracting their respective means (see Gelman and Hill, 2007, Chapter 4). For example, $x_k^c = x_k - \bar{x}_k$. When a centered predictor is 0 ($x_k^c = 0$), the original predictor is at its mean ($x_k = \bar{x}_k$.). As a result, the treatment effect θ is the effect when covariate x_k is at its mean when the regression model in Equation (3) was fit by using centered predictors. For this reason, we center all predictors before fitting a regression model.

Multilevel Modeling

Multilevel models are also known as random effects models or mixed effects models (Gelman and Hill, 2007; Qian, 2010). When using multilevel models for causal inference, we rely on the proper stratification to group data with similar attributes into strata. Within each stratum, a regression model is used:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \cdots + \beta_{kj} x_{kij} + \theta_j T_{ij} + \varepsilon_{ij} \tag{4}$$

where the subscript ij represents the ith observation in the jth stratum. The multilevel model imposes a common prior distribution on the regression model coefficients:

$$\begin{pmatrix} \beta_{0j} \\ \vdots \\ \beta_{kj} \\ \theta_j \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{\beta_0} \\ \vdots \\ \mu_{\beta_k} \\ \mu_\theta \end{pmatrix}, \quad \Sigma \right] \tag{5}$$

The common prior distribution creates a shrinkage effect; therefore, it improves the overall model performance (Qian *et al.*, 2015).

Whereas the propensity score method produces an estimate of the average effect θ, the multilevel model produces estimates for both the average effect μ_θ and effects for each stratum θ_j. In the propensity score method, the confounding factors are balanced through the matching process; therefore, the estimated average effect is reliable, as long as there are no hidden confounding factors. In the multilevel model, the estimated stratum-specific and overall average effects are contingent on the method of stratification. As a result, the multilevel model relies on subject matter knowledge to ensure that the stratification method used is scientifically meaningful.

Objectives

Using the propensity score and the multilevel modeling methods, we evaluated the effects of conservation practices represented in the MANAGE database. The specific objectives were to: (1) demonstrate the utility of applying the propensity score and multilevel modeling methods to the assessment of agricultural conservation practice effects, and (2) estimate the average and crop-specific effects of agricultural conservation practices on runoff P losses; both in an effort to demonstrate the applicability of these meta-analysis methods outside the social science realm.

MATERIAL AND METHODS

Data

To apply these statistical causal analysis methods, meta-data were extracted from the MANAGE database version 3 (Harmel *et al.*, 2008), which summarizes nutrient load and concentration data from 55 publications. Because very few studies in the database included fields with a filter strip or riparian buffer (1.5% of fields), we combined these two practices into one category. As a result, we evaluated the effects of four conservation practices [grassed waterways (USDA-NRCS, 2010a), contour farming (USDA-NRCS, 2007a), terraces (USDA-NRCS, 2010b), riparian forest buffer/filter strip (USDA-NRCS, 2007b, 2008)]. Our analyses (see Results and Discussion) suggest that the effects of individual conservation practices or their combinations are similar. As a result, we grouped the data into two groups, fields with at least one conservation practice and fields without any practice. The effect of conservation practices is defined as the amount of P loading reduction due to one or more conservation practices while all other factors are held constant.

Initial Analysis

We first grouped fields with nonzero P Applications into seven groups based on their conservation practice combinations. Each field was assigned a binary four digit code in following order: waterway, buffer, contour farming, and terrace. Each digit indicates whether that practice was implemented (i.e., 1 present, 0 absent). For example, 0100 represents a field with a buffer (filter strip or forested buffer), and 1001 represents a field with both a grassed waterway and terraces. After controlling the effects of soil loss, runoff, land use, tillage, and fertilizer application method, the estimated intercept and slope (P Applied random effects) were not different from 0 for fields with 1+ conservation practice, but fields without conservation practices had significantly higher intercept and slope values (Figure 2). As a result, we combined data from fields with 1+ conservation practices and therefore, estimated the effects of implementing 1+ conservation practices in fields with various crops and different fertilizer application methods.

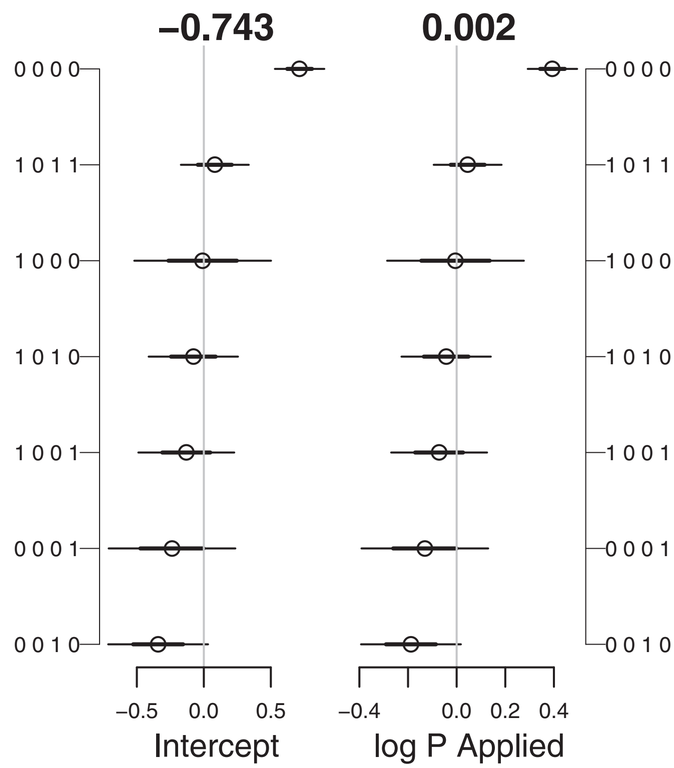

FIGURE 2. Estimated Multilevel Model Intercepts and Slopes for Fields with Various Conservation Practices. The *y*-axis label indicates whether any of the four conservation practices (waterway, buffer, contour farming, and terrace) are implemented (1) or not (0). The open circles are the estimated mean random effects and the thick and thin lines are the mean ±2 times standard error. The bold numbers are the fixed effects [average intercept (−0.743) and slope (0.002)]. The intercept represents mean of natural log of P loss (log kg/ha/yr) and the slope is unit-less because the units of P loss and fertilizer application are the same (both in log kg of P/ha/yr). Because both the response and predictor variables are (natural) log transformed, the slope is the % change in P loss per 1% increase in fertilizer application (see Qian, 2010, p. 157).

Causal Analyses

Given that propensity score and multilevel modeling have their appeals and potential problems, we applied both causal analysis methods to provide a better understanding of the potential variation in the effect of conservation practices. The propensity score method estimates the average effect based on two groups with similar distributions of covariates that represent a subset of the original data. The multilevel model method estimates the average effect dependent on the adequacy of data stratification. When the stratification method is justified, the multilevel model also yields stratum-specific effects. Therefore, we used both causal analysis methods to estimate the average effect of one or more conservation practices and thus increase the likelihood of detecting mistakes (i.e., hidden confounding factors or inappropriate stratification).

More importantly, we used the comparison of θ and μ_θ to serve as a "method checking" step. If the two estimates are similar, we are more confident on the estimated overall effect. In other words, if the two causal analysis methods produce the same conclusion, we can conclude that the methods are likely correct.

All analyses were carried out in R (R Core Team, 2015). We followed the steps outlined in Gelman and Hill (2007) for the propensity score method and used the R function "lmer()" from package "lme4" (Bates, 2010) to fit the multilevel model. The MANAGE database and our R code are in the online Supporting Information.

Propensity Score. For the propensity score method, we used seven potential covariates (P applied, runoff, soil loss, land use, fertilizer application method, tillage, and a term representing the land use/fertilizer application method interaction). We chose these covariates to represent factors important in determining P loads. Some obvious covariates (e.g., land slope, soil test P) were not included because they were reported by only a small number of studies in MANAGE.

The response variable (annual average total P load) was log-transformed. In MANAGE two annual total P loads were reported as 0, and they were replaced by 0.002 (half of the smallest nonzero values) in the analysis. Numerical covariates (P applied, runoff, and soil loss) were transformed using log $(x + 1)$ because of numerous 0 values in these covariates. When observations with missing values were removed, the dataset with these covariates had 135 observations, among which 29 used 1+ conservation practices. Using the nearest neighbor matching method, each of the 29 fields was matched with a field without conservation practices. The resulting subset had a total of 58 observations.

Multilevel Modeling

After trying many different multilevel model forms, a final model was selected based on both the relevancy of the predictors and AIC. We started with the most comprehensive model, which included all relevant numeric predictors and allowed varying intercept and slopes for all stratifications from relevant categorical predictors. The relevant numerical predictors were P applied, runoff, and soil loss. The relevant factor predictors were land use, tillage, and fertilizer application method. We then systematically eliminated terms that were statistically insignificant and that increased the AIC value. The resulting best model was:

$$\log(TP\text{Load}_{ijk}) = \beta_{0j} + \beta_1 x_{1ijk} + \beta_2 x_{2ijk} + \beta_{3jk} x_{3ijk} + \alpha_{jk} x_{3ijk} T_{ijk} + \theta_{jk} T_{ijk} + \varepsilon \quad (6)$$

where subscript ijk represents the ith observation in the jth land use type, and using kth fertilizer application method. Subscripts for model coefficients denote whether the respective coefficient values vary by groups. For example, β_{0j} indicates that the intercept β_0 is allowed to vary by land use type, and θ_{jk} represents that the conservation practice effect varies both by land use type and by fertilizer application methods. The varying coefficients were presented in terms of an overall mean (the fixed effect) and a group adjustment (random effects) (Table 2). The estimated model coefficients for a specific land use and fertilizer application method were the sum of the fixed effect and the appropriate random effects. For example, the intercept for corn field was 0.213 (0.133 + 0.080), and the conservation practice effect for a corn field using fertilizer application method "Incorporated" was −0.881 (−1.147−0.161 + 0.427).

Because both the fixed effects and random effects were estimated with uncertainty, the standard deviation of the estimated coefficients for a specific land use was calculated as the square root of the sum of squared standard deviations. For example, the corn field intercept of 0.213 was the sum of two terms, and

TABLE 2. Multilevel Model (Equation 6) Coefficients and Standard Error (SE). Units of model response and predictor variables are defined in Table 1.

Fixed Effects			
Coefficient	**Estimates**	**SE**	
β_0	0.133	0.199	
β_1	0.416	0.070	
β_2	0.513	0.048	
β_3	0.246	0.052	
θ	−1.147	0.465	
α	−0.351	0.191	
Random effects (land use)			
Category	β_0 (SE)	θ (SE)	α (SE)
Alfalfa	0.596 (0.302)	−1.199 (0.607)	−0.155 (0.079)
Corn	0.080 (0.139)	−0.161 (0.280)	−0.021 (0.036)
Cotton	0.010 (0.249)	−0.021 (0.501)	−0.003 (0.065)
Fallow	0.602 (0.402)	−1.211 (0.810)	−0.157 (0.105)
Oats/wheat	0.112 (0.236)	−0.224 (0.474)	−0.029 (0.065)
Pasture	−0.305 (0.108)	0.614 (0.218)	0.080 (0.028)
Peanuts	−0.293 (0.302)	0.589 (0.607)	0.076 (0.079)
Rotation	−0.660 (0.178)	1.327 (0.357)	0.172 (0.046)
Sorghum	−0.142 (0.241)	0.286 (0.484)	0.037 (0.063)
Random effects (fertilizer application method)			
Category		θ (SE)	α (SE)
Incorporated		0.427 (0.129)	0.286 (0.087)
Injected		−0.272 (0.170)	−0.182 (0.114)
Surface applied		0.091 (0.188)	0.061 (0.126)
Unknown		−0.246 (0.256)	−0.165 (0.172)

their respective standard errors were 0.199 and 0.139. The estimated standard error for the corn field intercept was then 0.243. Likewise, the estimated conservation practice effect for a corn field using fertilizer application method of "Incorporated" was a sum of three terms (each with a standard error), and the standard error of the estimated effect was the squared root of the sum of squares of the three respective standard errors ($\sqrt{0.465^2 + 0.28^2 + 0.129^2}$) or 0.56. We note that this approach can overestimate the uncertainty because the correlations among model coefficients were not considered.

RESULTS AND DISCUSSION

Propensity Score

A comparison of the distributions of the amount of the covariate P Applied before and after matching (Figure 3) illustrates results typical of the propensity score method. Before matching, the median annual total P loads were 0.63 and 0.99 kg/ha/yr for fields without and with any conservation practices, respectively. After matching, total P loads were 2.94 kg/ha/yr without any conservation practice and 0.97 kg/ha/yr with one or more.

Without considering covariates, the estimated treatment effect (logarithmic scale) can be estimated by using a *t*-test. The estimated effect was $\theta = -0.9563$ ($p = 0.01$). The standard deviation of θ was 0.36. This value represents a multiplicative factor of $e^{-0.9563} = 0.3843$, which is a 62% reduction in total P loads [with a 95% confidence interval of (21%, 81%)]. This value is the conservation practice effect averaged over all confounding factors represented in the propensity score matched dataset.

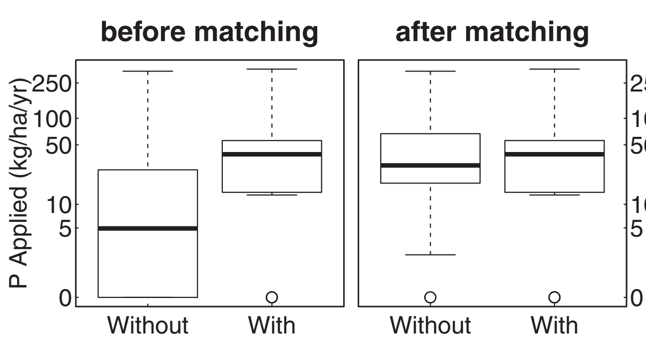

FIGURE 3. Distributions of the Centered log P Applied Are Compared between Fields with and without Conservation Practices, for before (left panel) and after (right panel) Propensity Score Matching.

Using relevant covariates, we further adjusted the estimated effect using regression. The best model [based on Akaike information criterion or AIC (Akaike, 1974)] uses runoff, soil loss, and P applied as covariate and the slope of P applied varies between the treatment and control groups. The fitted model is of the form:

$$\log(TP\text{Load}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha x_3 T + \theta T + \varepsilon \tag{7}$$

where x_1, x_2, and x_3 are centered log covariates runoff, soil loss, and P applied, respectively (e.g., $x_1 = \log(\text{Runoff}) - \overline{\log(\text{Runoff})}$). With this formulation, β_0 is the mean log total P load for fields without conservation practices ($T = 0$) when runoff, soil loss, and P applied are close to their respective (geometric) means. Because both the response and the predictors were log-transformed, the slopes β_1, β_2, β_3 represent the percent changes in total P load for every 1% change in the respective predictor (Qian, 2010, page 157). For example, $\widehat{\beta_1} = 0.77$ indicates approximately 0.77% (95% CI: 0.77 ± 0.18) increase in total P load for every 1% increase in runoff. θ is the average conservation practice effect. α is the conservation practice effect on the slope of covariate P applied [i.e., slope of x_3 is $\beta_3 = 0.14$ (se = 0.1) for fields without any conservation practice ($T = 0$), and the slope for x_3 is $\beta_3 + \alpha = -0.05$ (se = 0.16) for fields with conservation practice ($T = 1$)]. The estimated $\hat{\theta} = -1.06$ (se = 0.2) represents a multiplicative factor of $e^{-1.06} = 0.35$, or a 65% reduction in total P loads [95% CI: (48%, 77%)] for fields with average P applied (12.8 kg/ha/yr), as well as average runoff and soil loss. Using the estimated model coefficients (Table 3), we see that effect of P applied for control (β_3) is 0.14 (0.14% increase in P loads for every 1% increase in P applied), and the same effect for treatment ($\beta_3 + \alpha$) is -0.05 (not different from 0). We note that the estimated β_3 is statistically not different from 0. This is likely a result of the reduced sample size due to propensity score matching. We fit the same model using data before matching, and the estimated β_3 for fields without any conservation practices was positive (0.33 ± 2 × 0.05), while the slope for fields with conservation practices was not different from 0 (Table 1). Because the slope represents the increase in log P loads per unit increase of log P applied, a negative interaction term (α) suggests that conservation practices also reduced the incremental contribution to P loads from applied fertilizer. In other words, conservation practices are more effective for fields with a higher fertilizer application rate; the estimated 65% reduction is applicable to a field with the mean value of P applied (12.8 kg/ha/yr). Because of the high uncertainty for α, we reestimated the effect by remov-

TABLE 3. Regression Model (Equation 7) Coefficients (and standard error, SE) before and after Propensity Score Matching. The unit of the regression model response variable is (natural) log kg/ha/yr, and the units of predictor variables are log mm/yr for x_1 and log kg/ha/yr for x_2 and x_3.

Coefficient	After Matching (SE)	Before Matching (SE)
β_0	0.30 (0.16)	0.03 (0.16)
β_1	0.77 (0.09)	0.50 (0.07)
β_2	0.38 (0.06)	0.47 (0.05)
β_3	0.14 (0.10)	0.33 (0.05)
θ	−1.06 (0.20)	−0.75 (0.20)
α	−0.19 (0.13)	−0.36 (0.11)

ing the interaction term and resulting θ to be −1.19 (SE = 0.18) [or 70% (56%, 79%) reduction]. This value represents the conservation practice effect when P applied, runoff, and soil loss are the same for both treatment and control fields.

Multilevel Modeling

We interpreted the three numbers used to calculate the effect of conservation practices applied to corn fields using fertilizer application method "Incorporated" (−1.147−0.161 + 0.427) as follows. The average conservation practice effect across all crops was $\mu_\theta = -1.147$ (SE = 0.465), which is a factor of $e^{-1.147} = 0.318$ or a 68.2% reduction [95% confidence interval: (20%, 87%)]. For an average corn field, the effect is −1.147−0.161 (=−1.308 and SE = 0.54) or 73% with a 95% confidence interval of (20%, 91%). That is, conservation practices are more effective for corn fields than for the average across all field types. If applied fertilizer was incorporated on the corn field, the conservation practice effect was then −0.881 (59% reduction). The change suggests that the "Incorporated" method is itself an effective means for reducing fertilizer losses through surface runoff, and the incremental effect of conservation practices is thereby reduced. We note that the random effect for fertilizer application method of "Injected" is negative, which implies that this method may result in increased P load. This result is counterintuitive, and it is likely a result of imbalance in the data. There were 24 fields in the MANAGE database that used the injection method. Seventeen of them were corn fields, two each were cotton and soybeans, and three were rotations. Likewise, the method "Surface Applied" was used mostly on pasture. Given that these effects were not statistically different from 0, we could only conclude that the current database may be inadequate for assessing the effects of fertilizer application methods.

Because the methods produced similar estimates of the average effect, we used multilevel modeling to further explore how conservation practices change the effect of P applied (the slope of P applied). For example, we used the fitted model to understand the relationship between total P load and P applied, assuming effects of other factors are accounted for. In Figure 4, we graphed the relationship of $\log(TP\text{Load}_{ij}) = \beta_{0j} + \beta_{3j}x_3 + \theta_j T_{ij} + \alpha_j x_{3ij} T_{ij}$, which represents a comparison between fields with conservation practices and fields without using average runoff and soil loss and without considering the effects of fertilizer application methods. In all crops, the effect of conservation practices (distance between the dashed and solid lines) increased when P applied increased. In most cases, the slope for fields with conservation practices ($\beta_{3j} + \alpha_{2j}$) were not statistically different from 0, indicating no significant increase in P load when fertilizer application increased in fields with conservation practices.

CONCLUSIONS

Using observational data is often necessary when studying effects of natural resource management practices because comprehensive randomized experiments are cumbersome and expensive. In the present study, we demonstrated the applicability of two causal statistical methods (propensity score and multilevel modeling) to quantify the effects of water and soil conservation practices in reducing P loss from agricultural fields. These methods lead to more conclusive results than conventional statistical meta-analysis methods, which do not address confounding factors; therefore, application of these methods that are commonly applied in the social science realm can improve meta-analyses related to agricultural conservation.

As in all statistical methods, the validity depends on underlying assumptions. The propensity score method assumes no hidden confounding factors (known as the strong ignorable treatment assignment assumption), while the multilevel modeling method assumes adequacy of data stratification. Because these conditions are impossible to verify, we used both methods and compared the results. The average effects (after controlling the effects of confounding factors) from the two methods were similar even with large differences in sample sizes (58 for propensity score and 135 for multilevel modeling). The similarity of the 70% reduction from the propensity score and the 68% reduction from the multilevel model suggests that both methods are valid, in the case for estimat-

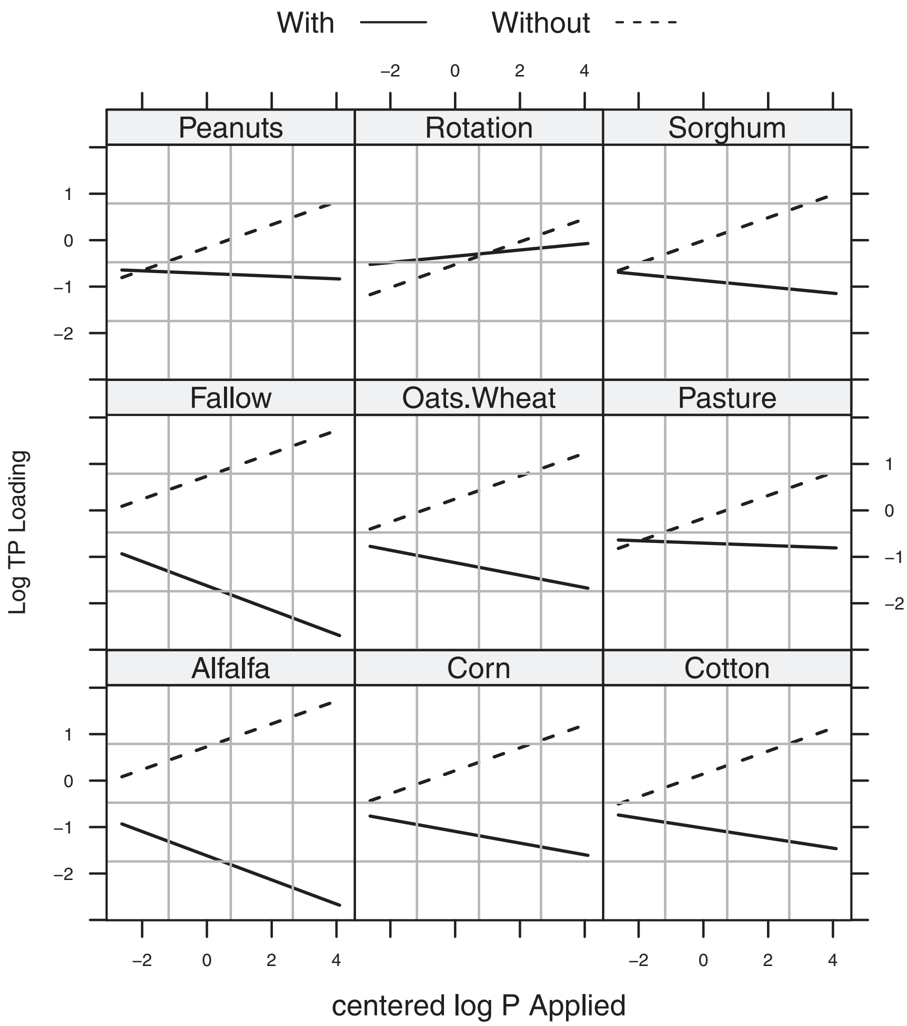

FIGURE 4. Effects of P Applied (slopes) Are Reduced When Using One or More Conservation Practices and the Effects Vary by Crop Type. Both the *x*- and *y*-axes are in natural logarithm scale (log kg/ha/yr). The values on the *x*-axis are centered log loading (log kg of P/ha/yr—mean of log P applied, for example, 0 represents the geometric mean of P applied).

ing the P loss reduction due to implementation of one or more conservation practice. In spite of these favorable results, the average rates of reduction should not be universally applied because the analyses were limited somewhat by limited data availability. Furthermore, conservation practices were shown to reduce incremental P loads per unit increase in fertilizer application rate; however, the P load reduction due to individual practices was not quantified because of the limited sample size for each conservation practice.

The value of our study can be realized in two ways. First, our estimates of conservation practices can be used in watershed modeling for scenario simulations to study the general effect on a watershed scale. Second, our estimates can be used to develop a prior distribution in a subsequent study of effects of individual conservation practices. A Bayesian estimation approach is most effective when informative prior distributions are available. Our study provides the basis for developing such priors.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article: the MANAGE dataset and R scripts for (1) processing the data, (2) calculating propensity score and matching, (3) multilevel modeling, and (4) producing figures used in the manuscript.

ACKNOWLEDGMENTS

We thank Pete Richards, Ibrahim Alameddine, and Kenneth H. Reckhow for constructive comments and suggestions. SSQ was partially supported by grants from the University of Toledo Research Council and the 4R Nutrient Stewardship.

LITERATURE CITED

Akaike, H., 1974. A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control 19(6):716-723.

Bates, D., 2010. Lme4: Mixed-Effects Modeling with R. Springer, http://lme4.r-forge.r-project.org/book/, *accessed* November 2015.

Beaulac, M.N. and K.H. Reckhow, 1982. An Examination of Land Use—Nutrient Export Relationships. Journal of the American Water Resources Association 18(6):1013-1024.

Cho, J., G. Vellidis, D. Bosch, R. Lowrance, and T. Strickland, 2010. Water Quality Effects of Simulated Conservation Practice Scenarios in the Little River Experimental Watershed. Journal of Soil and Water Conservation 65(6):463-473.

Christianson, L.E. and R.D. Harmel, 2015. The MANAGE Drain Load Database: Review and Compilation of More Than Fifty Years of North American Drainage Nutrient Studies. Agricultural Water Management 159:277-289.

Duriancik, L.F., D. Bucks, J.P. Dobrowolski, T. Drewes, S.D. Eckles, L. Jolley, R.L. Kellogg, D. Lund, J.R. Makuch, M.P. O'Neill, C.A. Rewa, M.R. Walbridge, R. Parry, and M.A. Weltz, 2008. The First Five Years of the Conservation Effects Assessment Project. Journal of Soil and Water Conservation 63(6):185A-197A.

Fisher, R.A., 1971. The Design of Experiments (Ninth Edition). Macmillan, New York, New York.

Gelman, A. and J. Hill, 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York, New York.

Harmel, D., S. Potter, P. Casebolt, K. Reckhow, C. Green, and R. Haney, 2006. Compilation of Measured Nutrient Load Data for Agricultural Land Uses in the United States. Journal of the American Water Resources Association 42(5):1163-1178.

Harmel, D., S.S. Qian, K. Reckhow, and P. Casebolt, 2008. The MANAGE Database: Nutrient Load and Site Characteristic Updates and Runoff Concentration Data. Journal of Environmental Quality 37:2403-2406.

Qian, S.S., 2010. Environmental and Ecological Statistics with R. Chapman and Hall/CRC Press, Boca Raton, Florida.

Qian, S.S., C.A. Stow, and Y.K. Cha, 2015. Implications of Stein's Paradox for Environmental Standard Compliance Assessment. Environmental Science and Technology 49(10):5913-5920.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/, *accessed* November 2015.

Rosenbaum, P. and D. Rubin, 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika 70(1):41-55.

Rubin, D., 2006. Matched Sampling for Causal Effects. Cambridge University Press, Cambridge, UK.

Sharpley, A., B. Foy, and P. Withers, 2000. Practical and Innovative Measures for the Control of Agricultural Phosphorus Losses to Water: An Overview. Journal of Environmental Quality 29 (1):1-9.

Sharpley, A.N., S.C. Chapra, R. Wedepohl, J.T. Sims, T.C. Daniel, and K.R. Reddy, 1994. Managing Agricultural Phosphorus for Protection of Surface Waters: Issues and Options. Journal of Environmental Quality 23:437-451.

Sharpley, A.N. and P.J.A. Withers, 1994. The Environmentally-Sound Management of Agricultural Phosphorus. Fertilizer Research 39:133-146.

Smith, D.R., W. Francesconi, S.J. Livingston, and C. Huang, 2015. Phosphorus Losses from Monitored Fields with Conservation Practices in the Lake Erie Basin, USA. Ambio 44(Suppl 2):S319-331.

USDA-NRCS, 2007a. Conservation Practice Standard: Contour Farming (Ac.) CODE 330. http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs143_026017.pdf, *accessed* November 2015.

USDA-NRCS, 2007b. Conservation Practice Standard: Riparian Forest Buffer (Acre) CODE 391. http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs143_026098.pdf, *accessed* November 2015.

USDA-NRCS, 2008. Conservation Practice Standard: Filter Strip (Ac.) CODE 393. http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs143_014868.pdf, *accessed* November 2015.

USDA-NRCS, 2010a. Conservation Practice Standard: Grassed Waterway (Ac.) CODE 412. http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs143_026051.pdf, *accessed* November 2015.

USDA-NRCS, 2010b. Conservation Practice Standard: Terrace (Ft.) CODE 600. http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1046935.pdf, *accessed* November 2015.

USDA-NRCS, 2011a. Assessment of the Effects of Conservation Practices on Cultivated Cropland in the Chesapeake Bay Region. Technical Report, Conservation Effects Assessment Project (CEAP). http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcsdev11_023934.pdf, *accessed* November 2015.

USDA-NRCS, 2011b. Assessment of the Effects of Conservation Practices on Cultivated Cropland in the Great Lakes Region. Technical Report, Conservation Effects Assessment Project (CEAP). http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1045480.pdf, *accessed* November 2015.

USDA-NRCS, 2012a. Assessment of the Effects of Conservation Practices on Cultivated Cropland in the Missouri River Basin. Technical Report, Conservation Effects Assessment Project (CEAP). http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1048710.pdf, *accessed* November 2015.

USDA-NRCS, 2012b. Assessment of the Effects of Conservation Practices on Cultivated Cropland in the Ohio-Tennessee River Basin. Technical Report, Conservation Effects Assessment Project (CEAP). http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1046342.pdf, *accessed* November 2015.

USDA-NRCS, 2012c. Assessment of the Effects of Conservation Practices on Cultivated Cropland in the Upper Mississippi River Basin. Technical Report, Conservation Effects Assessment Project (CEAP). http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1042093.pdf, *accessed* November 2015.

van Belle, G., 2008. Statistical Rules of Thumb (Second Edition). Wiley, Hoboken, New Jersey.

White, M.J., C. Santhi, N. Kannan, J.G. Arnold, R.D. Harmel, L. Norfleet, P. Allen, M. DiLuzio, X. Wang, J. Atwood, E. Haney, and M. Johnson, 2014. Nutrient Delivery from the Mississippi River to the Gulf of Mexico and Effects of Cropland Conservation. Journal of Soil and Water Conservation 69(1):26-40.