Structural Equation Model of Total Phosphorus Loads in the Red River of the North Basin, USA and Canada

Karen R. Ryberg*

Abstract

Attribution of the causes of trends in nutrient loading is often limited to correlation, qualitative reasoning, or references to the work of others. This paper represents efforts to improve causal attribution of water-quality changes. The Red River of the North basin provides a regional test case because of international interest in the reduction of total phosphorus loads and the availability of long-term total phosphorus data and ancillary geospatial data with the potential to explain changes in water quality over time. The objectives of the study are to investigate structural equation modeling methods for application to water-quality problems and to test causal hypotheses related to the drivers of total phosphorus loads over the period 1970 to 2012. Multiple working hypotheses that explain total phosphorus loads and methods for estimating missing ancillary data were developed, and water-quality related challenges to structural equation modeling (including skewed data and scaling issues) were addressed. The model indicates that increased precipitation in season 1 (November–February) or season 2 (March–June) would increase total phosphorus loads in the basin. The effect of agricultural practices on total phosphorus loads was significant, although the effect is about one-third of the effect of season 1 precipitation. The structural equation model representing loads at six sites in the basin shows that climate and agricultural practices explain almost 60% of the annual total phosphorus load in the Red River of the North basin. The modeling process and the unexplained variance highlight the need for better ancillary long-term data for causal assessments.

Core Ideas

• Multiple working hypotheses for the causes of total phosphorus loads were developed.
• Challenges associated with water-quality and related ancillary data are addressed.
• Hypotheses with appropriate data were tested with structural equation modeling.
• Climate and agricultural practices explain 57% of the variation in the load.
• Climate has a greater effect on P loads than do agricultural practices.

This open access article distributed under the terms of the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Supplemental material is available online for this article. Published online August 17, 2017.

Abbreviations: CFI, Comparative Fit Index; CRP, Conservation Reserve Program; EGRET, Exploration and Graphics for RivEr Trends; LVM, latent variable model; MWHs, multiple working hypotheses; PCA, principal components analysis; RMSEA, root mean square error of approximation; SEM, structural equation model; SRMR, standardized root mean-square residual; TP, total phosphorus; WRTDS, weighted regressions on time, discharge, and season; WWTP, wastewater treatment plant.
Phosphorus is a naturally occurring, widespread nutrient essential for plant growth; however, there are also anthropogenic sources, and phosphorus is often the nutrient responsible for accelerated eutrophication (Schindler, 1974; Mueller and Helsel, 1996). Nonpoint sources of phosphorus include minerals, rocks, soil, fertilizer, and dead biomass, all of which can contribute excess phosphorus to streams through natural runoff and soil erosion (natural or as the result of tillage practices). The most common point source is sewage effluent (Hem, 1985; Mueller and Helsel, 1996). Minimizing the generation of waste containing phosphorus and losses of phosphorus in the environment is a global concern for both food production and adequate water quality (Withers et al., 2015). The Federal Water Pollution Control Act, as amended in 1972 (now known as the Clean Water Act), resulted in numerous efforts to improve water quality (USEPA, 2016). From 1970 to 1992, urban streams experienced a decline in phosphorus following mandated phosphorus controls in wastewater treatment plants and limits on the phosphate content of detergent (Mueller and Helsel, 1996). Prompted by concerns about phosphorus in lakes, Minnesota banned phosphorus in lawn fertilizers in 2005 (State of Minnesota, 2005). Major manufacturers in the United States removed phosphorus from dishwater detergent in 2010 (McCoy, 2011).

In SEM, theory is the basis of the “causal proposition,” not the data or statistics (Liu et al., 1997). Structural equation models test hypotheses about relations among observed and latent (unobserved) variables and hypotheses about patterns of relations between variables (Kaplan, 2000; Baggozzi and Yi, 2012; Hoyle, 2012). Structural equation models have been applied to ecology (Johnson et al., 1991; Mitchell, 1992; Pugesek and Tomer, 1996; MacCallum and Austin, 2000; Malae et al., 2000; Iriondo et al., 2003; Pugesek et al., 2003; Grace and Bollen, 2005; Arhonditis et al., 2006; Grace, 2006, 2008; Grace et al., 2010, 2012, 2015; Riseng et al., 2011) and have had more limited application to water quality (Reckhow et al., 2005; Chen and Lin, 2010; Pollman, 2014), particularly for groundwater (Liu et al., 1997; Chenini and Khemiri, 2009; Belkhirri and Narany, 2015; Viswanath et al., 2015) and lakes and reservoirs (Liu et al., 2010; Hu and Ou, 2013; Wu et al., 2014). One reason SEM has not been applied widely to water quality is the limited availability of ancillary data for potential causal factors, as observations are often sporadic or the period of record is short. The many challenges to better causal analysis and the use of SEMs include missing data, differing periods of record in water-quality observations and ancillary data, differing spatial scales represented in the data, and non-normal distributions common in water-quality and associated ancillary data.

Given the total phosphorus concerns in the Red River, the potential of SEM, and the pitfalls of SEM, this study has two purposes. The first is to investigate SEM methods for application to water-quality problems and document model specification decisions that can apply to other water-quality problems. The second is to test causal hypotheses related to the drivers of total phosphorus (TP) loads in the Red River over the period 1970 to 2012.

Materials and Methods

Total phosphorus is the sum of all forms of phosphorus, including dissolved (that portion that can pass through a filter at the time of water-quality sampling, including phosphate) and particulate (that portion adsorbed to sediment and in plant and animal tissue). Four sources of data for TP concentrations in the Red River were identified: Environment Canada (data provided in electronic format to the USGS), USGS data available from the Water Data for the Nation database (USGS, 2015), data from the Minnesota Pollution Control Agency Lake and Stream Information database (Minnesota Pollution Control Agency, 2015), and data from the North Dakota Department of Health Surface Water Quality Data for North Dakota database (North Dakota Department of Health, 2015). Supplemental Table S1 lists the data sources. Multiple sources of data at sites were statistically compared, where they overlapped, using the Kolmogorov–Smirnov test (a nonparametric test for whether two populations differ in distribution; Higgins, 2004) to determine whether the data could be combined. The comparison process was previously described for the sites on the Red River at Fargo, ND, and at Emerson, MB (Ryberg, 2015; Ryberg et al., 2015), and that process was followed for the additional sites included in this study.

Discharge data were obtained from the USGS Water Data for the Nation database (USGS, 2015). Exploratory analyses were performed using Exploration and Graphics for RivEr Trends (EGRET; Hirsch and De Cicco, 2014), a package for the statistical computing software R (R Core Team, 2016), and results were published in Ryberg (2015). A major finding, based on decadal boxplots, was that median daily discharge in the 1970s through the early 1990s was approximately one-half of discharge for the early 1990s through 2012 for the Red River at both Emerson, MB, and Fargo, ND.

Total phosphorus load was modeled using weighted regressions on time, discharge, and season (WRTDS), a method for analysis that can be used to characterize trends in concentration and flux (Hirsch et al., 2010; Hirsch and De Cicco, 2014). The WRTDS method combines water-quality and discharge data into an analysis that decomposes the record into four parts: (i) a trend that is similar to a time series moving average over several years, (ii) a seasonal component that can change gradually over time, (iii) a discharge relation that is relatively smooth with the potential for gradual changes over time, and (iv) a random part that remains after the removal of the trend, seasonal, and discharge components (Hirsch et al., 2010). The WRTDS method assumes changes in the discharge regime can happen gradually; however, the discharge analysis showed an abrupt change to much wetter conditions in 1993, which was also documented elsewhere (Williams-Sether, 1999). Because of the very strong nonstationariness evident in the discharge dataset, the WRTDS analyses were divided into two slightly overlapping periods, 1970 to 1993 and 1993 to 2012. The year 1993 is included in both periods because the change happened mid-year 1993 and, for some locations, the increase in precipitation was seen more in 1994 than 1993 (Williams-Sether, 1999). Therefore, the estimates of daily and annual concentration and load can differ for 1993 because the first period (1970–1993) experienced comparatively drier conditions, whereas the second period (1993–2012) was wetter with higher runoff. In the subsequent SEM analysis,
the load for 1993 was the average of the two estimates. The WRTDS estimates were limited to the period of record, which varied from site to site (shown in Supplemental Table S1; start dates varied from 1970 to 1994, and end dates varied from 2010 to 2012). An associated USGS data release (Ryberg et al., 2016) provides the annual load estimates.

**Multiple Working Hypotheses Framework**

Merz et al. (2012, p. 1379) argued that flood trend attribution (which has many similarities to water-quality trend attribution) is generally “based on qualitative reasoning or even speculation” and often is simply a listing of references to related works that support the authors’ conclusions. Harrigan et al. (2013) cited the work of Merz and used Chamberlin’s (1890) method of multiple working hypotheses (MWHs) for attribution of detected changes in streamflow by identifying a wider set of potential drivers of hydrological change in a basin. They described each hypothesized driver’s potential influence on the basin, made a judgment as to whether the driver affected the basin (or acknowledged lack of current information), and identified which drivers were appropriate for further statistical analysis (Harrigan et al., 2013). Clark et al. (2011) advocated for the MWHs framework in hydrologic modeling; their work is relevant to SEM as well in that they encouraged the use of MWHs to describe the overall system as well as processes within it. Following these ideas, MWHs potentially describing the causes of annual changes in total phosphorus load in the Red River basin were considered, and available data to test the hypotheses were examined. Causal mechanisms and potential datasets were systematically considered and are described in Supplemental Table S2. In some cases, sufficient data were not available to test potential causal mechanisms; those decisions are documented in Supplemental Table S2.

Data representing potential causal mechanisms for total phosphorus loads in the basin were compiled for the six watersheds represented by the six streamgages (Supplemental Table S1). These data include basin monthly mean precipitation; estimates of percentage of cropland in the Conservation Reserve Program; estimates of phosphorus from fertilizer and manure; percentage of land in categories developed, semideveloped, and agricultural; and estimates of total phosphorus load from wastewater treatment plants (WWTPs). The data, as well as shape files with the six streamgage locations and subbasins, are available in an associated USGS data release (Ryberg et al., 2016), which includes metadata. More information about these data is available in the Supplemental Material.

**Estimates of Missing Values: Interpolation and Imputation of Missing Values with Principal Components Analysis**

For the ancillary data described above, the study period includes many years for which one or more of the datasets do not have estimates; however, annual estimates were needed for the subsequent SEM. One option is to exclude years with missing values; however, the subsequent SEM would not be feasible with a period reduced to those years with complete data. In some cases, there was little change from observation to observation; in those instances, missing values were linearly interpolated. In other datasets, linear interpolation was not appropriate because it greatly underestimates the variability in the underlying processes; therefore, missing values were imputed using principal components analysis (PCA) via the function imputePCA in the R (R Core Team, 2016) package missMDA (Husson and Josse, 2015). Codes for each observation in the datasets are defined in the USGS data release metadata and indicate which estimation method was used for each observation (Ryberg et al., 2016). Additional details are included in the Supplemental Material.

**Structural Equation Modeling**

Potential SEM models for the factors that influence total phosphorus load in the Red River basin were developed using a priori knowledge of the basin and of the phosphorus cycle in the basin. These models were based on the multiple hypotheses framework and data availability highlighted in Supplemental Table S2.

The limit of the number of conceivable models approaches infinity, constrained only by the science of the phosphorus cycle and the number of variables available to represent causal factors (this number can be augmented by the use of latent variables, explained below). These variables can be combined together as one or more regression equations and one or more latent variables that then combine to model TP load. In addition, variables can have a direct effect on the TP load or a mediated effect through another variable. However, sample size considerations limit the complexity of realistic models. Structural equation modeling is a large sample method and, generally, one should have a ratio of 10 to 20 samples per parameter estimate in the final model (sample size considerations and rules of thumb are discussed further in the Supplemental Material).

Four types of models were considered. The first was a separate SEM model at each site, which would allow the variables used in the model to differ from site to site and the parameter estimates to vary for the same variables. The second consisted of individual models at each site that were then connected together in a hierarchical manner. In this case, the TP load for an upstream site would become a variable in the model for the next downstream site. The third type of model considered was a grouped model in which the same model was applied to all the sites, with site identification number as a grouping variable. The same parameters would be estimated at each site but allowed to vary among the sites. The fourth type was a model in which all the sites were combined. This would increase the sample size but necessitate scaling and standardization of variables so that they were on the same scale across sites.

The SEM was fitted using 223 observations and the software package lavaan, latent variable analysis (Rosseel, 2012, 2016), for the statistical software R (R Core Team, 2016). Wastewater treatment plant estimates were not imputed for the years 1970 to 1977 because no satisfactory method could be found for imputation (see Supplemental Material) and robust maximum likelihood methods were used that can estimate incomplete observations. In the sem function of lavaan, the estimator was indicated as “MLR” (robust maximum likelihood estimation for both complete and incomplete data, with a scaled test statistic) and the missing argument was set equal to “fiml” (full information maximum likelihood in which years with partial data can contribute to estimation of all model parameters) (Rosseel, 2012; Beaujean, 2015). Model revisions were based on modification
indices (indices available as output in lavaan that report the change in chi-square that results from adding a path to a model) and knowledge of the phosphorus cycle. Specific modeling concerns are described in the following sections.

**Transformation to Multivariate Normality**

Maximum likelihood methods require multivariate normality of endogenous variables, which are internal or dependent variables (variables with an arrow pointing to them in the graphical depiction of an SEM model; Fig. 1); therefore, endogenous variables were transformed to approximate multivariate normality, by basin (Kline, 2012; Lei and Wu, 2012). The process followed that of an example in the *Handbook of Structural Equation Modeling* (Hoyle, 2012) in which the endogenous variables were first rescaled to a range of 1 to 100, then transformed to approximate multivariate normality, and finally standardized to 0 means and unit standard deviations (Fox et al., 2012). Additional details are available in the Supplemental Material.

Univariate tests for normality (Shapiro et al., 1968; Shapiro and Francia, 1972; R Core Team, 2016) failed for the Conservation Reserve Program (CRP) data, as did multivariate tests (Mardia, 1974; Korkmaz et al., 2014) when it was included with the other endogenous variables. Structural equation modeling failed the overall test of fit when CRP data were included. This occurred for several reasons, including that many zeroes were in the data series (because CRP started in 1986) and many extremely small values. Even when only the nonzero values were examined, CRP data still violated assumptions of normality; therefore, CRP was dropped from consideration in models for all basins. Non-normalities were also found in the WWTP data; however, WWTP was used as an exogenous variable (a variable that is not dependent on an outcome of other variables) in the

![Fig. 1. Graphical depiction of the structural equation model for annual total P loads in the Red River of the North basin. LoadKgPerYr, annual total P load; AgPract, latent variable representing agricultural management practices; PfromFert, P from fertilizer; PfromManure, P from manure; wheat, percentage cropland harvest wheat; soy, percentage cropland harvested soybeans; s1precip, total precipitation for season 1 (November–February); s2precip, total precipitation for season 2 (March–June); wwtpTPLoad, annual total P load from wastewater treatment plants. Green arrows are positive coefficients; red arrows are negative coefficients. The darker the color, the larger the coefficient. Double arrows are variances estimated by the model. wwtpTPLoad was not statistically significant (p value > 0.01) and is therefore represented by a faint line. The numeric values are the standardized path coefficients and variance estimates.](image-url)
model, and SEM does not require distributional assumptions for exogenous variables (Eliason, 1993).

**Latent Variable Model**

Structural equation models can contain one or more latent variables that are in themselves a model and represent unobserved constructs known to exist. Latent variables allow one to discuss those constructs, potentially in comparison to other latent variables. A reflective latent variable model (LVM) is indicated by observed, or manifest, variables that in some cases may not be significant on their own but contribute to a latent construct. One could construct an LVM representing urban influences, for example, in which multiple causal factors indicate an overall urban influence on water quality. The LVM represents the underlying structure of the relations between the observed variables. In the urban example, the variation in urbanization might cause variation in indicator variables such as WWTP effluent, percentage impervious surface, population, and storm water runoff. An LVM needs enough nonredundant information to generate unique parameter estimates (Beaujean, 2015), and four indicator variables may be sufficient. In the urban example, population and wastewater effluent may be highly correlated; therefore, some exploratory analysis is required. In some cases, LVMs can be estimated with fewer indicators, but with conditions imposed on the model (Beaujean, 2015) that can reduce its utility.

In this case, an LVM representing agricultural practices was considered because of the availability of agricultural-related data. The agricultural practices LVM represents a set of economic, societal, and climatic conditions that affect how farmers till their soil, what crops they grow, and what amendments, such as fertilizer, they add to the soil.

**Measures of Model Fit**

Four measures assessed model fit. The first was an overall test of fit, a robust chisquare test determined using robust maximum likelihood methods and a Yuan–Bentler scaled test statistic (Yuan and Bentler, 1998) from the output of the sem function in lavaan (Rosseel, 2012). The scaled test statistic reflects the multivariate kurtosis in the observations (Beaujean, 2015). A nonsignificant result (p value > 0.01) indicates the model fits the data relatively well. The chi-square test statistic can be biased and indicate rejections of satisfactory models when there are distributional violations. Bootstrapping can correct for this bias (Grace, 2006; Hancock and Liu, 2012). Because nonnormalities are common in water-quality data and evident in some of the univariate and multivariate tests of the data, the overall test of fit and additional fit measures were bootstrapped using the Bollen–Stine bootstrapping method (Bollen and Stine, 1992; Hancock and Liu, 2012; Rosseel, 2012).

Additional measures of model fit included an absolute (or standalone) index, standardized root mean square residual (SRMR), which does not compare models or account for complexity but simply measures absolute fit; an incremental fit index (also called relative or comparative indexes), Comparative Fit Index (CFI; Bentler, 1990; Beaujean, 2015); and a parsimony index, root mean square error of approximation (RMSEA), which takes into account model complexity and penalizes models with more parameters (fewer residual degrees of freedom). These indices were recommended for use with maximum likelihood methods because of their generally accepted cut-off values for model assessment; the provision for penalizing model complexity, in the case of RMSEA; and relative independence of sample size, in the case of CFI (Hu and Bentler, 1998, 1999; Bagozzi and Yi, 2012). The SRMR ranges from 0 to 1, with values closer to 0 indicative of a better fit (Rosseel, 2012). The CFI ranges from 0 to 1, values closer 1 indicating a better fit (Ullman, 2006; Bagozzi and Yi, 2012; Rosseel, 2012). The RMSEA ranges from 0 to 1 (but can go higher than 1), with values closer to 0 being better (Beaujean, 2015). The following cutoff values were used to indicate good model fit: SRMR ≤ 0.08; CFI ≥ 0.95; and RMSEA ≤ 0.06 indicative of close model fit, with RMSEA > 0.10 indicative of a poor-fitting model (Hu and Bentler, 1998, 1999; Ullman, 2006; Bagozzi and Yi, 2012; Rosseel, 2012; Beaujean, 2015). Additional details about these measures of model fit are provided in the Supplemental Material.

**Results**

Sample size limitations quickly became evident. The sample size limitation removed the first three types of models (separate model at each site, a hierarchical SEM, or a grouped SEM) from consideration. The data used represent up to 43 yr of water-quality sampling, which is a long-term sampling program in the realm of water quality. However, when summarized to annual load estimates, a sample size of 43 is too small for SEM. In the hierarchical or grouped model scenarios, the number of parameters estimated greatly increases, and the 223 estimates of annual TP load in the basin do not support that level of model complexity.

Modeling urban and agricultural practices LVMs was considered; however, only the agricultural practices LVM was feasible based on the available data and the desire for four (or more) indicator variables for each model. The agricultural practices LVM was indicated by phosphorus from fertilizer, phosphorus from manure, percentage of cropland harvested wheat (*Triticum aestivum* L.), and percentage of cropland harvested soybeans (*Glycine max* (L.) Merr.). Corn (*Zea mays* L.) was also a potential indicator variable; however, soybeans and corn were strongly positively correlated (Kendall’s tau ranged from 0.81 to 0.98 for the six basins), and soybeans and corn were both strongly negatively correlated with wheat (Kendall’s tau ranged from −0.57 to −0.96 for correlation between soybeans and wheat and −0.46 to −0.97 for correlation between corn and wheat). Therefore, only two of the three crop-type variables were needed for nonredundant information. The corn data violated assumptions of univariate and multivariate normality, despite transformation attempts; therefore, corn was the crop-type variable dropped. In the final SEM, the agricultural practices LVM represented the underlying structure (covariance) that produced relations among the indicator variables (Beaujean, 2015).

The SEM model (Fig. 1, Table 1) indicated that agricultural practices have a direct influence on the annual total phosphorus load. Season 1 (November–February) and season 2 (March–June) precipitation also directly influenced the total phosphorus load, whereas season 2 influenced agricultural practices, thereby having another, indirect, influence on total phosphorus load. Phosphorus from fertilizer, phosphorus from manure, percentage cropland harvested wheat, and percentage cropland harvested soybeans were indicative of the latent construct,
agricultural practices. This model has a parameter estimate to sample size ratio of approximately 12:1.

Measures of model fit indicated that the model fit the data reasonably well ($p$ value for robust chi-square test = 0.028). The overall fit test and additional measures of model quality were bootstrapped and the empirical distribution of the overall fit test was compared to the original test statistic as described in Hancock and Liu (2012). The bootstrap $p$ value was 0.260, indicating the model fits well (a bootstrap $p$ value < 0.01 would indicate the original model was inadequate). The mean of the bootstrapped fit measures indicated excellent model quality (average SRMR = 0.06; average CFI = 0.99; and average RMSEA = 0.05).

Season 1 precipitation (November–February) had the largest effect on TP load, on the basis of the magnitude of the standardized coefficients. In terms of standard deviation units, if season 1 precipitation had increased by one standard deviation, then (scaled, transformed, standardized) load would have increased by 0.661 standard deviation units. The combined direct and indirect effect of season 2 precipitation was 0.337 (direct effect, 0.268, plus the indirect effect 0.323 × 0.215). The direct effect of agricultural practices on the load was 0.215 standard deviation units. The wastewater contribution (wwtp TPLoad) to TP load was not statistically significant in the model ($p$ value = 0.310).

The latent variable, agricultural practices (AgPract in tables and figure), was used to represent the underlying covariance structure among the observed indicator variables phosphorus from fertilizer (PfromFert), phosphorus from manure (PfromManure), percentage cropland harvested wheat (wheat), and percentage cropland harvested soybeans (soy). Agricultural practices explained a great deal for the variance in phosphorus from manure (communality = 0.87, Table 2, where communality is the proportion of an indicator's variance explained by the latent variable influencing it, the squared standardized loading) and percentage cropland harvested as soybeans (communality = 0.95).

The SEM explained 57% of the annual TP load in the Red River basin ($R^2$, Table 3). Season 2 precipitation explained 10% of the year-to-year changes in agricultural practices (Table 3).

### Table 1. Standardized parameter estimates for structural equation model of annual total phosphorus in the Red River of the North basin.

<table>
<thead>
<tr>
<th>Left-hand side†</th>
<th>Operation relating variables‡</th>
<th>Right-hand side†</th>
<th>Standardized coefficient or variance</th>
<th>Standard error</th>
<th>$p$ value for $t$ test of significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>LoadKgPerYr</td>
<td>~</td>
<td>s1precip</td>
<td>0.661</td>
<td>0.032</td>
<td>0.000</td>
</tr>
<tr>
<td>LoadKgPerYr</td>
<td>~</td>
<td>s2precip</td>
<td>0.268</td>
<td>0.049</td>
<td>0.000</td>
</tr>
<tr>
<td>LoadKgPerYr</td>
<td>~</td>
<td>AgPract</td>
<td>0.215</td>
<td>0.049</td>
<td>0.007</td>
</tr>
<tr>
<td>LoadKgPerYr</td>
<td>~</td>
<td>wwtpTPLoad</td>
<td>0.047</td>
<td>0.046</td>
<td>0.310</td>
</tr>
<tr>
<td>AgPract</td>
<td>~</td>
<td>s2precip</td>
<td>0.323</td>
<td>0.059</td>
<td>0.000</td>
</tr>
<tr>
<td>AgPract</td>
<td>=~</td>
<td>PfromFert</td>
<td>0.284</td>
<td>0.089</td>
<td>NA§</td>
</tr>
<tr>
<td>AgPract</td>
<td>=~</td>
<td>PfromManure</td>
<td>−0.933</td>
<td>0.019</td>
<td>0.000</td>
</tr>
<tr>
<td>AgPract</td>
<td>=~</td>
<td>wheat</td>
<td>−0.766</td>
<td>0.038</td>
<td>0.000</td>
</tr>
<tr>
<td>AgPract</td>
<td>=~</td>
<td>soy</td>
<td>0.977</td>
<td>0.013</td>
<td>0.000</td>
</tr>
<tr>
<td>PfromFert</td>
<td>~~~</td>
<td>PfromFert</td>
<td>0.919</td>
<td>0.051</td>
<td>0.000</td>
</tr>
<tr>
<td>PfromManure</td>
<td>~~~</td>
<td>PfromManure</td>
<td>0.130</td>
<td>0.036</td>
<td>0.000</td>
</tr>
<tr>
<td>wheat</td>
<td>~~~</td>
<td>wheat</td>
<td>0.414</td>
<td>0.059</td>
<td>0.000</td>
</tr>
<tr>
<td>soy</td>
<td>~~~</td>
<td>soy</td>
<td>0.046</td>
<td>0.025</td>
<td>0.062</td>
</tr>
<tr>
<td>LoadKgPerYr</td>
<td>~~~</td>
<td>LoadKgPerYr</td>
<td>0.427</td>
<td>0.038</td>
<td>0.000</td>
</tr>
<tr>
<td>AgPract</td>
<td>~1</td>
<td>AgPract</td>
<td>0.896</td>
<td>0.038</td>
<td>0.000</td>
</tr>
<tr>
<td>PfromFert</td>
<td>~1</td>
<td></td>
<td>−0.196</td>
<td>0.144</td>
<td>0.175</td>
</tr>
<tr>
<td>PfromManure</td>
<td>~1</td>
<td></td>
<td>1.051</td>
<td>0.247</td>
<td>0.000</td>
</tr>
<tr>
<td>wheat</td>
<td>~1</td>
<td></td>
<td>0.929</td>
<td>0.205</td>
<td>0.000</td>
</tr>
<tr>
<td>soy</td>
<td>~1</td>
<td></td>
<td>−1.072</td>
<td>0.259</td>
<td>0.000</td>
</tr>
<tr>
<td>LoadKgPerYr</td>
<td>~1</td>
<td></td>
<td>−2.969</td>
<td>0.194</td>
<td>0.000</td>
</tr>
</tbody>
</table>

† LoadKgPerYr, annual total P load; AgPract, latent variable representing agricultural practices; PfromFert, P from fertilizer; PfromManure, P from manure; wheat, percentage cropland harvest wheat; soy, percentage cropland harvested soybeans; s1 precip, total precipitation for season 1 (November–February); s2 precip, total precipitation for season 2 (March–June); wwtpTPLoad, annual total phosphorus load from wastewater treatment plants.

‡ ~ indicates that the right-hand side variable is a regression predictor of the left-hand side; =~ indicates a reflective latent variable (left-hand side) indicated by the right-hand side variable; ~~, variance; ~1, mean or intercept term.

§ PfromFert is a "marker variable" that serves to define the latent variable model's variance; therefore, a $p$-value is not calculated.

### Table 2. Communality of the indicator variables.

<table>
<thead>
<tr>
<th>Indicator variable†</th>
<th>Communality</th>
</tr>
</thead>
<tbody>
<tr>
<td>PfromFert</td>
<td>0.08</td>
</tr>
<tr>
<td>PfromManure</td>
<td>0.87</td>
</tr>
<tr>
<td>wheat</td>
<td>0.59</td>
</tr>
<tr>
<td>soy</td>
<td>0.95</td>
</tr>
</tbody>
</table>

† PfromFert, P from fertilizer; PfromManure, phosphorus from manure; wheat, percentage cropland harvest wheat; soy, percentage cropland harvested soybeans.

### Table 3. $R^2$ for the regression relations.

<table>
<thead>
<tr>
<th>Regression relation†</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgPract ~ s2precip</td>
<td>0.10</td>
</tr>
<tr>
<td>LoadKgPerYr ~ s1precip + s2precip + AgPract + wwtpTPLoad</td>
<td>0.57</td>
</tr>
</tbody>
</table>

† AgPract, latent variable representing agricultural practices; LoadKgPerYr, total P load in kilograms per year; s1 precip, season 1 (November–February) total precipitation; s2 precip, season 2 (March–June) total precipitation; wwtpTPLoad, total P load from wastewater treatment plants.
This indicated that agricultural practices are responsive to climatic conditions.

**Discussion**

Multiple working hypotheses as a precursor to SEM are a useful tool for summarizing possible causal factors of the water-quality process being described and for summarizing the availability of data. Even with careful consideration of potential causal factors and data, SEM is challenging for water-quality issues because of the need for ancillary data regularly measured over the entire period of record. Ancillary data often have varying periods of record and sporadic measurements. These challenges may be addressed using a variety of techniques that were incorporated in this paper, including linear interpolation of values that change little over the period of record (such as percentage agricultural land use in this study), imputation of missing values using PCA for variable data that require a more sophisticated model for estimation (such as some of the data found in Ryberg et al., 2016), or the use of robust maximum likelihood methods for incomplete observations (such as is the case with the WWTP data). In addition, water-quality data and the associated ancillary data are on widely varying scales and often are skewed. These issues were successfully addressed by rescaling to a common scale and transformation to approximate normality. While useful, rescaling and transformation processes are not a cure all as some variables defy transformation to univariate or multivariate normality (such as the CRP data).

Having a large enough sample size is another challenge in applying SEM to water-quality problems. The rescaling and standardization process supports combining multiple sites together to effectively increase the sample size and support more parameter estimates than would a single site model.

In this case study, almost 60% of the year-to-year variation in TP load in the basin is explained explicitly in the model, rather than through simple correlation relations or qualitative statements. This is an important step in better supporting the attribution of the causes of changes in water quality. Conversely, 40% of the variance remains unexplained. Several reasons are possible for this, including unquantified errors in the WWTP data, lack of data related to in-stream processes, generalizations made by the model, lagged factors, and unquantified errors in the TP load estimates.

It is possible that the WWTP data were not very effective in the model because of unquantified errors in the estimates based on treatment method and discharge. Examination of the WWTP data (Ryberg et al., 2016) shows little change from year to year for some basins and potential outliers in others. In addition, the decrease in phosphorus in urban streams from 1970 to 1992 in the United States is not seen in streams in the Red River basin. These data provide a dataset to support better attribution of the causes of changes in water quality; however, they will require further refinement and are likely affected by source water quality, individual differences between WWTPs, unique characteristics of each city served by a WWTP, and unquantified reporting errors in the Clean Watersheds Needs Assessment. It is also possible that WWTP effluent contributes a base level of TP but does not contribute much to the year-to-year differences in TP load in the basin.

Some of the remaining variability may be caused by in-stream processes for which there are insufficient long-term data. The model combined TP loads at six sites to find a generalized model for TP. Individual sites may vary as to the most important factors in load production; however, treating the sites as individuals would not work with SEM, which is a large sample method. Finally, estimating a structural model based on annual observations smooths over some timing issues, such as the timing of fertilizer application, which may be important for explaining the variability. Attempts were made to adjust for this by including seasonal precipitation. Lagged phosphorus from fertilizer was also attempted early in exploratory analysis (not shown) but did not contribute to a model that fit well. A model of seasonal TP load was considered; however, many of the potential causal factors are estimated only on an annual scale.

The interpretation of model coefficients is complex. Because the coefficients are standardized, a one standard deviation increase in season 1 (November–February) precipitation increases the variable representing annual TP load by 0.661 standard deviations; however, that variable is standardized, transformed, and scaled TP load and therefore difficult to interpret in the original units. However, the model does show us that season 1 precipitation is the most important driver of TP loads in this basin. It also indicates that agricultural practices are important, and if fertilizer usage, cropping choices, and other considerations for which there is not sufficient data, such as tillage practices, can be adapted to reduce losses of fertilizer to the streams, there could be a significant reduction in phosphorus loads.

The model still provides significant improvement over causal statements made on the basis of correlations, qualitative information, and references to the work of others. Season 1 precipitation had a large contribution to annual TP load (Fig. 1). High season 1 precipitation in the Red River basin generally indicates a large snowpack, which when it melts in the spring can cause overland flooding (Ryberg et al., 2007), bringing streams into contact with plant material and soils. When the water returns to the stream it brings phosphorus with it. Season 2 precipitation (March–June) also has an important influence on annual TP load. It directly affects the load by contributing to snowpack or spring rains that can wash TP sources into the stream before plants are mature (there is more bare ground at this time than later in the season) and occurs before evapotranspiration becomes a major factor in the hydrologic cycle; therefore, season 2 precipitation is more likely to runoff than season 3 (July–October). Season 2 precipitation also had an indirect influence on TP load in that it affects agricultural practices that in turn affect TP load. Agricultural practices, a combination of crop types and fertilizer methods, also influence TP load.

It is not surprising that climate and agricultural practices affect TP load. The benefit of the model is that it shows the relative importance of these factors and quantifies how much of the variability that they explain in TP load (almost 60%). The model also shows the power of latent variables that can be used to discuss and quantify constructs like agricultural practices. Further refinement of such a model might include an urban practices LVM, in which case additional or improved data are needed for urban factors.
Conclusions

Multiple working hypotheses and SEM are useful tools for framing and testing causal hypotheses about water quality, providing the benefit of modeling the causes and thereby moving the science of attribution beyond simple correlation relations, qualitative comments, and references to the work of others. The model presented here shows that climate and agricultural practices are the major drivers of TP load in the Red River basin, explaining almost 60% of the year-to-year variance in TP loads. Climate has the largest effect on the load, which is not a surprise; however, it important to understand this contribution to water quality, especially with interest in developing nutrient criteria for the Red River at the US–Canadian border. With approximately 72% of the land area being in agricultural production (Ryberg et al., 2016), agricultural practices are an important driver of water quality across the basin, and the model shows that fertilizer and crops contribute to variation in TP loads. The modeling process and the unexplained variance highlight the need for better ancillary data for causal factors. Not only do water-quality researchers need long-term water-quality monitoring, but they also need long-term monitoring of causal factors.

Acknowledgments

This project was funded by the US Geological Survey National Water Quality Program’s National Water-Quality Assessment Project, the Minnesota Pollution Control Agency, and the North Dakota Department of Health. Thanks to Aldo V. Vecchia, US Geological Survey, for an early review of this manuscript, and thanks to two anonymous reviewers for their constructive reviews. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US Government.

References


