Assessment of water quality in Little Vermillion River watershed using principal component and nearest neighbor analyses
Shiyang Li, Rabin Bhattarai, Li Wang, Richard A. Cooke, Fang Ma and Prasanta K. Kalita

ABSTRACT
Because of increased use of fertilizers to feed the increasing global population, the nutrient loads in surface and subsurface water have increased substantially in the last few decades. Many studies have been conducted to investigate the factors affecting nitrate load in surface and subsurface flow. The objective of this study is to investigate the relationship between the various factors affecting nitrate transport using principal component analysis (PCA) and nearest neighborhood analysis methods. Hydrological and biogeochemical data from a small (<500 km²) agricultural watershed in east central Illinois, USA for the duration of 10 years have been used in this study. The PCA approach divided various factors that influence nitrate transport into three principal components. The first component contained air temperature, cover phenotype, evapotranspiration, cover factor and dry mass factors. The second component contained precipitation and flow, which was defined as the hydrologic component. The third component included tillage practices and nitrogen application and was termed the anthropogenic component. The results from the PCA approach suggested all three components had significant influence on nitrate transportation and transformation. Among these three components, the hydrological components had the highest contribution on both surface and subsurface nitrate load. The nearest neighborhood analysis yielded a similar conclusion.

Key words | nitrate, non-point source pollution, PCA, subsurface drainage, watershed

INTRODUCTION
Agricultural and related activities are the leading cause of non-point pollution of water resources (Baker et al. 1975). Nitrogen (N) and phosphorus (P) are the main components of agricultural fertilizers needed for crop production. However, excess N and P can result in non-point sources pollution of water resources. Since excess nitrate and nitrite in water are associated with many health and environmental problems, the United States Environmental Protection Agency standard for NO₃ is 10 mg/L for drinking water (Kalita et al. 2006). The excess nitrate consumption can induce the stillbirths and infant deaths due to anencephalus (Fan & Steinberg 1996). Because of increased use of fertilizers and other agrochemicals in the Midwestern USA, the flushing of excess nutrients to the Mississippi River has led to the creation of a large hypoxic zone in the Gulf of Mexico (Liu et al. 2010). To understand nutrient transport through a heavily tile-drained watershed, a watershed monitoring program was conducted in Little Vermillion River (LVR) watershed from 1991 to 2004. The LVR watershed is located in east central Illinois and an example of a typical tile-drained agricultural area of the central USA.

Because of the presence of the subsurface tile drainage system, surface and subsurface transport of nutrients have become more complicated. The impact of the presence of tile drainage on surface and subsurface nutrient transport has been discussed in a series of earlier studies (Goswami

doi: 10.2166/ws.2014.117
The soluble agricultural chemicals like nitrate transport are enhanced because of the presence of the subsurface drainage system (Kladivko 2001). Gentry et al. (2000) demonstrated that the rainfall after fertilizer application could cause significant nitrate loss, and the melting snow could have a similar effect. Mitchell et al. (2000) and Kladivko et al. (2004) reported that nitrate concentrations followed a pronounced seasonal cycle, with maximum concentrations occurring in the spring and minimum concentrations occurring in the autumn. Kladivko et al. (2004) analyzed 15 year drainage agricultural area data from Indiana and reported that the N loss was a function of the drain spacing. They also observed that the narrower drain spacing resulted in more N losses per acre. Kalita et al. (2006) observed that the cropping management practices had a significant effect on nitrate concentration based on 10 years data from the LVR watershed. Another study conducted by Goswami et al. (2009) indicated that base flow played a more important role than tile flow in raising nitrate concentration in the Big Ditch watershed in Illinois, USA. Precipitation, antecedent moisture condition, fertilizer application time and evapotranspiration are the other contributing factors for nitrate load in streams and rivers (Goswami & Kalita 2009).

Principal component analysis (PCA) has been used to assess both surface (Parinet et al. 2004; Shrestha et al. 2008; Gvozdic et al. 2012) and groundwater (Zhang et al. 2011; Mahapatra et al. 2012) quality in earlier studies. The PCA analysis revealed that the dissolved mineral salt was the primary factor influencing water quality along the Mekong River (Shrestha et al. 2008). Gvozdic et al. (2012) reported that the water quality of River Drava could be divided into three zones using PCA analysis results and the level of pollution increased downstream. Arslan (2013) incorporated the spatial weight into the water quality analysis by using PCA analysis and observed that the PCA result was different from the spatial weighted approach compared to the classical method. The PCA analysis for surface water quality in the Indian River Lagoon, Florida revealed that land management was found to be the most influential anthropogenic factor (Wan et al. 2014).

K-nearest neighbor analysis (KNN) as a classification statistics method usually works together with PCA. Serrano & Gallego (2004) used the PCA method to identify the internal structure of data and the best discriminant variables and KNN predicted unknown water sample type. Deegalla & Bostrom (2006) used the PCA method to reduce the dimension of data sets and then classified the data by using the KNN method. St-Hilaire et al. (2012) developed the water temperature predication model by using KNN which could estimate water temperature. Evaluation of eutrophic lake water quality by using PCA and KNN showed that lake water quality related to biochemical processes, atmospheric deposition and runoff (Ikem & Adisa 2011). Modaresi & Araghinejad (2014) reported that KNN could be used to classify water quality. But these two methods have not been used for surface and subsurface water quality assessment in agricultural watershed to our knowledge.

Although previous studies were helpful to understand the mechanism and few specific factors affecting nitrogen transport, the level of impact of these factors on nitrogen transport has not been quantified clearly. A statistical method like PCA has not been used to study the factors that impact nitrate load in a subsurface drainage system. The objective of this study was to find the dominant components and most influential factors for nitrate load in surface and subsurface flow. In this study, 10 factors in surface and subsurface were monitored at three sites in the LVR watershed in east central Illinois, USA for 10 years. A multivariate methodology based on PCA and nearest neighbor analysis was conducted for the assessment. The relationship and interaction between the factors which impact the nitrate load in surface and subsurface flow were also investigated.

**METHODS**

**Site location and characteristics**

The LVR watershed area covers about 489 km² area and is located in sections of Champaign, Vermilion and Edgar counties of east central Illinois, USA (40°06’21.45″N, 87°41’34.12″W) (Figure 1). A long-term water quality-monitoring project was established at LVR watershed in 1991 (Mitchell et al. 2000). Data collected at three sampling locations (A, B and C in Figure 1) were used in this study. Surface and subsurface flow were monitored at all three stations and the samples were analyzed for nitrate-N.
concentration. Other parameters which can influence surface and subsurface flow like rainfall, temperature and farming operations were also monitored. Sites A, B and C drained agricultural fields with 6.68, 3.34 and 6.81 ha area, respectively.

Data collection and parameters selection

Surface and subsurface water samples were collected using an automatic sampler at 15 min intervals (Mitchell et al. 2000). Nitrate concentration was measured using the hydrazine sulfate reduction method and the daily load time series was calculated using linear interpolation.

In this study, nine influential parameters were selected as principal component factors to determine the impact of these factors on surface and subsurface nitrate load. The factors considered for PCA were:

1. temperature (monitored data in LVR project);
2. crop phenotype (soybean is 1 and corn is 2);
3. evapotranspiration (obtained from Illinois State Water Survey);
4. cover factor (soil loss ratio for crop stage period and canopy cover, estimated data);
5. dry mass of plant (estimated data according to crop stage and yield);
6. flow volume (monitored data in LVR project);
7. precipitation (monitored data in LVR project);
8. N application load (monitored data in LVR project);
9. land management (monitored management practice and transferred to numerical value);
10. nitrate-N load in flow subsurface and surface (monitored data in LVR project).

All nine factors selected were based on the previous works of researchers on dominant factors impacting nitrate load. Precipitation, temperature and fertilizer application were found to be important factors for nitrate transport in Gentry et al. (2000). Management and crop yield were the main factors reported in the study conducted by Kalita et al. (2006). Goswami et al. (2009) found that base flow is an important factor for nitrated load. Goswami & Kalita (2009) identified evapotranspiations and antecedent moisture condition could impact nitrate load in flow.

Data treatment and statistical method

Data treatment

One management practice can affect several aspects of water quality but the level of impact can be different. The effectiveness of management practices on water quality is presented in Table 1 (modified from Ohio State University 2014).

In this study, land management is not a numeric value. So, a weighting method was used to convert it to a
number following Equation (1). The equivalent factor is computed as

\[ M_j = \frac{A_4 + A_5}{\sum_{i=1}^{9} A_i} \]  

where \( i \) is a factor with ranges 1–9, \( j \) is management practices with ranges 1–5 and \( A_i \) is water quality factors.

Before standardizing the data, all the data for the three sites were tested for significant difference using t-test. The \( P \)-value showed that the temperature and precipitation data were not significantly different for all three sites.

### Principal component analysis and nearest neighbor analysis

The PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (Hotelling 1933).

The results of a PCA are usually discussed in terms of component scores, also called factor scores (the transformed variable values corresponding to a particular data point) and loadings (the weight by which each standardized original variable should be multiplied to get the component score). Based on component group and factor score, dominant factors can be identified.

The component score was computed by the equation below (Abdi & Williams 2010)

\[ F_j = \sum_{i=1}^{n} a_i \times \frac{c_i}{\sqrt{b_j}} \]  

where \( F \) is component score, \( i \) is factor number from 1–9, \( j \) is component number from 1–3, \( a \) is factors’ standardized data, \( b \) is the eigenvalue for each component and \( c \) is the factor loading.

After the extraction of principal component \( PC_j \), \( F_j \) is \( PC_j \)'s value, a model using \( PC_j \) as the independent variables was developed to predict nitrate load both in surface and subsurface. A multiple linear regression (MLR) analysis was used to relate nitrate load with all \( PC_j \). The general MLR equation is shown below

\[ y_i = c_1PC_1 + c_1PC_1 + \cdots + c_jPC_j + b \]  

where \( y_i \) is a dependent variable (nitrate load for surface or subsurface flow), \( c_j \) is a regression coefficient, \( PC \) is component score and \( j \) is component number from 1 to 3.

Because only three principal components were extracted in this study, the standard regression was applied by entering all independent variables together. The significance of each independent variable was tested and the analysis of variance (ANOVA) of each model was conducted to evaluate the validity to the dependent variable.

Nearest neighbor analysis can be used to find the spatial relationships among the multiple objects by calculating their distances. A graphical display of how these objects are clustered can also be obtained using this analysis (Nemes et al. 2006). A standardized \( n \)-space Euclidian distance is calculated to measure the similarity and relation which can be computed, as shown in Equation (4). Using nearest
neighbors, the unknown variable properties are predicted using Equation (5)

\[ d_{a,b} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \]  

\[ y = \frac{1}{k} \sum_{i=1}^{k} d_{a,b} \]  

where \( d \) is factor distance, \( i \) is factor number from 1 to 9, \( a \) and \( b \) are representing factors, \( Y \) is the prediction variable and \( k \) is the number of nearest neighbors.

In this study, the value of \( k \) is 3, based on the number of principal components extracted. The PCA was used to reduce the number of factor dimensions and obtain the independent principal components. Then nearest neighbor analysis was used to check and prove the results from PCA and identify the important factor(s) contributing to nitrate load in water. The PCA and nearest neighbor analysis were carried out using SPSS 19 software (IBM).

**RESULTS AND DISCUSSION**

**Six group data PCA results and analysis**

Before conducting PCA analysis, the correlation analysis among the variables was conducted and the results are presented in Table 2. It was observed that all nine factors had a strong correlation with one or several factors (Table 2). This indicated that the PCA process was appropriate for dimensionality reduction and components extraction.

The PCA analysis was carried out for A, As, B, Bs, C, Cs group data, respectively. The eigenvalues for the component number for each group are shown in Figure 2. All six groups’ eigenvalue for components showed a similar trend, all the eigenvalues for the first three components were greater than 1. This indicated that the six group data structures were homogeneous and it also revealed that all six groups’ data could be dimensionally reduced from nine to three.

As shown in Table 3, the total variance explained results indicated that those three leading components could explain more than 70% of variance for all six groups.

In addition, all six PCA loading patterns were similar to each other (Table 4). For example, the component 1 for Site A-surface data strongly correlated to temperature, cover phenotype, evapotranspiration, C factor and dry mass and the loading contribution of these factors to component 1 were 90.7, 89.5, 79.7, 60.7 and 49.8%, respectively. Similarly, the component 1 for Sites B and C surface data were also influenced strongly by temperature, cover phenotype, evapotranspiration, C factor and dry mass. Similarly, the loading contribution of these factors was 92.4, 89.5, 85.4, 69.7 and 55.3% for Site B; and 90.3, 89.1, 85.8, 79.8 and 52.4% for Site C, respectively. It is noticeable that the loading contribution of these factors for Sites B and C surface data was very close to those for Site A. Hence, \( t \)-test was used to check whether there was significant difference in factor loadings for Sites A, B and C. The analysis showed that

| Table 2 | Upper triangular correlation significance matrix of nine factors used in the principal component analysis |
|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| **C factor**    | **Cover phenotype** | **Tillage practice** | **Flow** | **N application** | **Precipitation** | **Dry mass** | **Evapotranspiration** | **Temperature** |
| Sig. (1-tailed)  |                  |                 |               |                  |                   |            |                          |                  |
| C factor         | <0.001           | 0.034           | <0.001        | 0.284            | <0.001           | 0.053        | <0.001                     | <0.001           |
| Cover phenotype  | 0.182            | 0.123           | 0.274         | <0.001           | <0.001           | <0.001       | <0.001                     | <0.001           |
| Tillage practice | 0.475            | 0.040           | 0.214         | 0.490            | 0.470            |              | 0.248                      |                  |
| Flow             | 0.225            | <0.001          | <0.001        | 0.011            | 0.454            |              |                            |                  |
| N application    | 0.020            | 0.112           | 0.003         | 0.100            |                  |              |                            |                  |
| Precipitation    | 0.411            | <0.001          | <0.001        | <0.001           |                  |              |                            |                  |
| Dry mass         | <0.001           |                  |               | <0.001           |                  |              |                            |                  |
| Evapotranspiration|                  |                  |               |                  |                  |              |                            | <0.001           |
| Temperature      |                  |                  |               |                  |                  |              |                            |                  |

Downloaded from https://iwaponline.com/ws/article-pdf/15/2/327/414956/ws015020327.pdf by guest
this meant there was no significant difference ($P > 0.05$) among factor loadings for Sites A, B and C. For components 2 and 3, the factor loadings for each site were not significantly different, where $P$-values were higher than 0.05 for both components. The results were similar for subsurface factor loading.

From the analysis, few inferences can be made. Whereas all the factor data were standardized, the higher factor loading meant factor contributed more to the component. Hence, the temperature was the dominant factor in component 1. It is evident that all other four factors in this component were impacted by temperature. Evapotranspiration and plant growth are driven by temperature, and cover factor and phenotype are indirectly influenced by it. Parinet et al. (2004) reported that nitrate concentration in lake water was related to temperature and biomass by using PCA. A similar finding was also reported by Ikem & Adisa (2014). For component 2, flow was found to be the dominant factor as it plays a vital role in water quality status. For component 3, nitrogen application rate was found to be more influential than management practice. Based on the characteristics of the factors in each component, component 1 in which all factors were related to vegetation could be interpreted as the botanical component. Since component 2 included precipitation and flow, it can be assigned as the hydrologic component. The factors in component 3 (N application and tillage practice) were related to anthropogenic activities. Hence, it can be termed an anthropogenic component. Wan et al. (2014) reached a similar conclusion that anthropogenic factors played an important role in water quality assessment.

To show the relationship between different factors, the component spatial distribution charts were created (Figure 3). It was observed that three orthogonal components were distinctly independent of each other (Figure 3 and Table 4). Component 1 contained five factors and explained more than 35% of variance of all factors. The second component, which was located at the second quadrant, contained two factors and explained more than 20% variance. Component 3 contained two factors and explained nearly 15% variance. All six charts revealed a homologous distribution pattern. The difference in flow mechanism (surface vs. subsurface) and site locations did not change the principal components. This indicated that nitrate load might be influenced by three independent components in the same manner rather than a random one, in this study. To determine which component influenced the nitrate load the most, multiple linear

![Figure 2](https://iwaponline.com/ws/article-pdf/15/2/327/414956/ws015020327.pdf)

**Figure 2** | The scree plot for PCA results of A, As, B, Bs, C and Cs.

<table>
<thead>
<tr>
<th>Component</th>
<th>Eigenvector value</th>
<th>% of variance</th>
<th>Cumulative %</th>
<th>Eigenvector value</th>
<th>% of variance</th>
<th>Cumulative %</th>
<th>Eigenvector value</th>
<th>% of variance</th>
<th>Cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>B</td>
<td></td>
<td></td>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.363</td>
<td>15.144</td>
<td>70.962</td>
<td>1.458</td>
<td>16.204</td>
<td>74.399</td>
<td>1.342</td>
<td>14.906</td>
<td>73.475</td>
</tr>
<tr>
<td>As</td>
<td></td>
<td></td>
<td></td>
<td>Bs</td>
<td></td>
<td></td>
<td>Cs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3.182</td>
<td>35.358</td>
<td>35.358</td>
<td>3.455</td>
<td>38.390</td>
<td>38.390</td>
<td>3.658</td>
<td>40.643</td>
<td>40.643</td>
</tr>
<tr>
<td>2</td>
<td>1.841</td>
<td>20.459</td>
<td>55.818</td>
<td>1.786</td>
<td>19.847</td>
<td>58.237</td>
<td>1.438</td>
<td>15.982</td>
<td>56.625</td>
</tr>
<tr>
<td>3</td>
<td>1.363</td>
<td>15.144</td>
<td>70.962</td>
<td>1.456</td>
<td>16.180</td>
<td>74.417</td>
<td>1.322</td>
<td>14.690</td>
<td>71.316</td>
</tr>
</tbody>
</table>

**Table 3** | Components variance explained for each group’s data in the principal component analysis

To show the relationship between different factors, the component spatial distribution charts were created (Figure 3). It was observed that three orthogonal components were distinctly independent of each other (Figure 3 and Table 4). Component 1 contained five factors and explained more than 35% of variance of all factors. The second component, which was located at the second quadrant, contained two factors and explained more than 20% variance. Component 3 contained two factors and explained nearly 15% variance. All six charts revealed a homologous distribution pattern. The difference in flow mechanism (surface vs. subsurface) and site locations did not change the principal components. This indicated that nitrate load might be influenced by three independent components in the same manner rather than a random one, in this study. To determine which component influenced the nitrate load the most, multiple linear
Regression models were built based on component score and standardized nitrate load data.

Regression analysis

The regression model results and model coefficients are presented in Figure 4 and Table 5, respectively. The ANOVA results showed all six models developed were statistically significant \((P < 0.001)\). The lowest coefficient in model As was 0.034 for botanical component and this coefficient was not significant in this model. It was also observed that the contribution of anthropogenic component to subsurface nitrate was negligible because the coefficient of significance was negative. Although the hydrologic component was significant, it was weakly related to the nitrate load \((R^2 = 0.17)\). This implied the model could only explain 17% nitrate load source, even though it was statistically significant \((P < 0.001)\). All the subsurface models (As, Bs and Cs) weakly corresponded with nitrate load.

The comparison of predicted nitrate load (RegZload) using the regression model and observed nitrate load (Zload-inwater) in water is shown in Figure 4. All the subsurface models showed a weak correlation to nitrate load. This meant the three components obtained from PCA could weakly explain nitrate load in subsurface water. In the case of the botanical component, the sinks for nitrate are plant uptake and denitrification. These nitrate sink processes are relatively slow compared to surface runoff generation and infiltration. Similarly, the impact of botanical component on nitrate loading in subsurface water is

### Table 4 Component loading matrices for six groups’ data in the principal component analysis

<table>
<thead>
<tr>
<th></th>
<th>Component</th>
<th></th>
<th>Component</th>
<th></th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Temperature</td>
<td>0.907</td>
<td>-0.127</td>
<td>-0.093</td>
<td>0.924</td>
<td>-0.100</td>
</tr>
<tr>
<td>Cover phenotype</td>
<td>0.895</td>
<td>-0.193</td>
<td>0.115</td>
<td>0.895</td>
<td>-0.072</td>
</tr>
<tr>
<td>Evapotranspiration</td>
<td>0.797</td>
<td>-0.283</td>
<td>-0.159</td>
<td>0.854</td>
<td>-0.179</td>
</tr>
<tr>
<td>C factor</td>
<td>0.607</td>
<td>0.328</td>
<td>0.430</td>
<td>0.697</td>
<td>0.280</td>
</tr>
<tr>
<td>Dry mass</td>
<td>0.498</td>
<td>-0.476</td>
<td>-0.156</td>
<td>0.553</td>
<td>-0.465</td>
</tr>
<tr>
<td>Flow</td>
<td>0.143</td>
<td>0.852</td>
<td>-0.288</td>
<td>0.139</td>
<td>0.839</td>
</tr>
<tr>
<td>Precipitation</td>
<td>0.471</td>
<td>0.711</td>
<td>-0.357</td>
<td>0.449</td>
<td>0.746</td>
</tr>
<tr>
<td>N application</td>
<td>0.018</td>
<td>0.092</td>
<td>0.819</td>
<td>-0.060</td>
<td>-0.193</td>
</tr>
<tr>
<td>Tillage practice</td>
<td>0.252</td>
<td>0.367</td>
<td>0.474</td>
<td>0.244</td>
<td>-0.373</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Component</th>
<th></th>
<th>Component</th>
<th></th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Temperature</td>
<td>0.907</td>
<td>-0.127</td>
<td>-0.093</td>
<td>0.923</td>
<td>-0.097</td>
</tr>
<tr>
<td>Cover phenotype</td>
<td>0.895</td>
<td>-0.193</td>
<td>0.115</td>
<td>0.898</td>
<td>-0.072</td>
</tr>
<tr>
<td>Evapotranspiration</td>
<td>0.797</td>
<td>-0.283</td>
<td>-0.159</td>
<td>0.862</td>
<td>-0.172</td>
</tr>
<tr>
<td>C factor</td>
<td>0.607</td>
<td>0.328</td>
<td>0.430</td>
<td>0.695</td>
<td>0.285</td>
</tr>
<tr>
<td>Dry mass</td>
<td>0.498</td>
<td>-0.476</td>
<td>-0.156</td>
<td>0.552</td>
<td>-0.463</td>
</tr>
<tr>
<td>Flow</td>
<td>0.143</td>
<td>0.852</td>
<td>-0.288</td>
<td>0.125</td>
<td>0.839</td>
</tr>
<tr>
<td>Precipitation</td>
<td>0.471</td>
<td>0.711</td>
<td>-0.357</td>
<td>0.440</td>
<td>0.749</td>
</tr>
<tr>
<td>N application</td>
<td>0.018</td>
<td>0.092</td>
<td>0.819</td>
<td>-0.069</td>
<td>-0.198</td>
</tr>
<tr>
<td>Tillage practice</td>
<td>0.252</td>
<td>0.367</td>
<td>0.474</td>
<td>0.234</td>
<td>-0.378</td>
</tr>
</tbody>
</table>
less significant compared to flow and infiltration. Although most nitrate loss on surface occurs through runoff, uptake and denitrification are other pathways for nitrate consumption. When there is no rainfall, the botanical component can contribute to nitrate load reduction in surface water.

The results from the As (nitrate load in subsurface flow) model suggested that botanical and anthropogenic components were not important contributors for subsurface load variation. The hydrologic component could impact the load to a certain degree and there could be few additional factors that could impact nitrate load. For example, soil type can have a significant impact on subsurface nitrate load. But the soil types in all three sites were similar in this study. Hence, the impact of variation in soil type on subsurface nitrate load could not be studied. It was observed that the result from regression models for surface nitrate load had a strong correlation with observed nitrate load for all three sites. The $R^2$ values for all three surface models were high. For Sites B and C, $R^2$ values were higher than 0.8 which indicated that all three components played important roles for the variation in nitrate load in surface flow. Most coefficients were significant ($P < 0.01$) for A, B and C models, indicating the high influence of components. Consequently, the hydrologic component was the more dominant component for surface load change compared to anthropogenic and botanical components. It is evident that rainfall and flow are the main causes of nitrate loss. The anthropogenic component will impact the nitrate source and runoff volume and pattern, while the botanical component can impact the nitrate load during non-rainfall time at a slow rate.

**Nearest neighbor analysis**

The three nearest neighbor factors results are shown in Figure 5 for nitrate load in each site. The basic function of the nearest neighbor analysis is to classify new factors based on the property of nearest old factors. It was observed that flow, N application and temperature were the nearest factors to nitrate load in every site. This implied that flow, N application and temperature could explain the variation in the nitrate load in water compared to the other six factors. Although this analysis is similar to Serrano & Gallego's (2004) studies, our approach is different from theirs. They predicted unknown samples based on nearest known samples.

![Figure 5](https://iwaponline.com/ws/article-pdf/15/2/327/414956/ws015020327.pdf)
In our case, the target factor and the factors affecting it were identified first. The factor, which is nearer to the target factor, can have a higher impact on it. From the value distribution of the three nearest neighbors in Figure 5, the surface nitrate loads were closer to flow than temperature and N application. This suggested the flow played the most important

Table 5  Multiple linear regression model coefficients and significance level table (the equation is $y_i = a_1F_1 + a_2F_2 + a_3F_3 + b$, the significance level of independent variables are labeled by *)

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Model</th>
<th>$a_1$ (C1)</th>
<th>$a_2$ (C2)</th>
<th>$a_3$ (C3)</th>
<th>ANOVA</th>
<th>b</th>
<th>Mean square</th>
<th>Sig.</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1*</td>
<td>0.542**</td>
<td>−0.135</td>
<td></td>
<td></td>
<td>−0.001</td>
<td>21.780</td>
<td>&lt;0.001</td>
<td>0.594</td>
</tr>
<tr>
<td>As</td>
<td>0.034</td>
<td>0.295**</td>
<td>−0.086</td>
<td></td>
<td></td>
<td>0.001</td>
<td>6.409</td>
<td>&lt;0.001</td>
<td>0.171</td>
</tr>
<tr>
<td>B</td>
<td>0.115**</td>
<td>0.622**</td>
<td>0.275**</td>
<td></td>
<td></td>
<td>−0.001</td>
<td>31.204</td>
<td>&lt;0.001</td>
<td>0.838</td>
</tr>
<tr>
<td>Bs</td>
<td>0.112</td>
<td>0.39**</td>
<td>0.075</td>
<td></td>
<td></td>
<td>−0.003</td>
<td>11.177</td>
<td>&lt;0.001</td>
<td>0.300</td>
</tr>
<tr>
<td>C</td>
<td>0.08**</td>
<td>0.645**</td>
<td>−0.231**</td>
<td></td>
<td></td>
<td>0.001</td>
<td>31.218</td>
<td>&lt;0.001</td>
<td>0.814</td>
</tr>
<tr>
<td>Cs</td>
<td>−0.006</td>
<td>0.707**</td>
<td>−0.012</td>
<td></td>
<td></td>
<td>0.001</td>
<td>16.057</td>
<td>&lt;0.001</td>
<td>0.429</td>
</tr>
</tbody>
</table>

*Significant ($P < 0.05$).  
**Extra significant ($P < 0.01$).  

Figure 4  The comparison of regression model results with standardized nitrate load (RegZload represents regression result and Zloadinwater represents standardized nitrate load).
role in nitrate load variation, which ratifies the results from PCA analysis. However, the subsurface results showed a different picture compared to PCA analysis. Temperature was found to be the closest factor to nitrate load. One plausible reason for this is there was no one factor that had significant influence on nitrate load without considering other related factors in subsurface sites. The regression analysis result also supported this, as the levels of significance for all the subsurface regression model coefficients were less than 0.5.

**CONCLUSION**

Based on PCA analysis result, nitrate load in both the surface and subsurface flow in three different sites seemed to share the same pattern of component structure. Nine factors could be grouped to three components according to their interrelationships. These components could be categorized as botanical, hydrologic and anthropogenic components.
Using the component score, regression models were developed for the nitrate load in both the surface and subsurface. The results showed that the hydrologic component was the most influential component in both surface and subsurface models. Botanical and anthropogenic components could also impact surface load but were weakly associated with subsurface load. The regression models created for surface load were more significant (high $R^2$ value) compared to the subsurface load one.

And flow, N application and temperature were considered as the most influential factors on nitrate load in both surface and subsurface sites by nearest neighbor analysis. This study helps to understand how the complicated multiple factors play a complex role in governing water quality status in surface and subsurface flow.

This study outlines a simple and readily transferable method to find dominant factors for water quality impairment in the agricultural watershed. Based on the findings in this study, it can be suggested that it is necessary to control or slow down the surface runoff and tile flow from farmland in order to improve the watershed water quality. The grass waterway, drop spillway and vegetative filter strip are examples of few BMPs for this purpose. Drainage water management devices and denitrification bioreactor at tile outlets can also help in reducing nitrate load in subsurface flow.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (51179041), the Major Science and Technology Program for Water Pollution Control and Treatment (2013ZX07201007), and the State Key Laboratory of Urban Water Resource and Environment (Harbin Institute of Technology) (No. 2014TS05).

REFERENCES


First received 17 July 2014; accepted in revised form 31 October 2014. Available online 14 November 2014