Water Quality Analysis of Rivers and Streams in the Continental United States.

Group: Hot Dawg
Alejandra Monsivais Ibarra, Jeonghoon Lee, Adeena Zahid

Abstract

A multitude of factors have been known to affect water quality of rivers and streams. This study sought to investigate two questions. Firstly, it looked at whether nitrogen, phosphorus, potassium, dissolved carbon and pH had an effect on water quality as measured by the EPT index. Secondly, it sought to investigate whether a correlation existed between land use and water quality as measured by the EPT index. Looking at data sourced from the United States Environmental Protection Agency's National Rivers and Streams Assessment for 2008/2009, this study found that none of the aforementioned chemicals or their interactions had any significant effect on water quality as measured by the EPT index. While correlations were found between land use and both EPT indices and chemical concentrations, these correlations were not statistically significant. To further investigate the correlation between land use on EPT index values, a case study was conducted on 6 sample sites near the Chambersburg Wastewater Treatment Plant along the Potomac river. Sites that were located further downstream from the facility had progressively lower EPT index values. However, this relationship was not significant due to lack of control for confounding variables affecting EPT at all 6 of the sample sites.

Introduction

Freshwater is an important resource. Beyond supporting human life, freshwater is an essential resource used in a variety of other industries from agriculture to mining. Most of the freshwater used on earth comes from readily accessible surface water sources such as rivers and lakes (Steffen et al., 2015). Despite their use as the main sources of freshwater on earth, resources from rivers and lakes are relatively scarce (Steffen et al., 2015). In fact, freshwater from rivers and lakes represent less than 0.0001% of all the water on earth (Steffen et al., 2015).

Given this fact and the undeniable importance of freshwater resources for all life on earth, it is clear that preserving water quality is of the utmost importance. Failing to do so can lead to negative human health effects, negative economic effects as well as wide-spread environmental degradation (Kerski, 2017). Yet preserving water quality is a difficult task. A wide variety of physical, environmental and chemical pressures are known to affect water quality (Kerski, 2017). Therefore, in order to enforce the necessary precautions to prevent and remediate water quality across different localities, a preliminary investigation into the biggest pressures most negatively affecting water quality in those localities must be done.

Although a variety of techniques for measuring water quality exist, one of the most effective and widely used methods is the EPT index (Jones et al., 2005). The EPT index, named for the three orders of benthic macroinvertebrates it monitors; Ephemeroptera, Plecoptera and Trichoptera, is a method of water quality analysis that uses the relative species abundances of these three orders of benthic macroinvertebrates as indicators of water quality (Narangarvuu et al., 2015). Past research has demonstrated that these three orders of benthic macroinvertebrates are highly abundant and that, more importantly, they are acutely sensitive to any sort of chemical or physical disturbance (Jones et al., 2005). Because of this, changes in their relative species abundances are reflective of physical habitat alteration, point and non-point contaminants and the cumulative effects of pollutants over their life cycle (Jones et al., 2005). In light of this, their relative abundances have been shown to be good measures of water quality in any given body of water (Reid, Somers & David, 1995). Higher values of the EPT index represent relatively good water quality, whereas lower values of the EPT index represent relatively poor water quality (Jones et al., 2005).

Using the EPT index as a metric of water quality, this study will seek to answer the following questions. Firstly, how do nitrogen, phosphorus, potassium, dissolved carbon content and pH affect water quality as measured by the EPT index across all rivers and streams in the continental United States? These five water chemistry variables were chosen because of their previously studied associations with phenomena leading to bad water quality. Nitrogen, phosphorus and potassium have been shown to contribute to the eutrophication of water bodies (Zeng et al., 2016). Dissolved carbon and pH are indicators of algal blooms, which negatively affect water quality by leading to anoxia or the release of cyanotoxins (Li et al., 2018). Secondly, this study sought to answer the question of whether a correlation existed between water quality and land use. We expect that higher concentrations of any or all chemical variables will lead to bad water quality, whereas lower concentrations will lead to relatively better water quality. Similarly, we expect that a correlation between land use and water quality will be observed.

Methods

Data sourcing and wrangling

The data used for this project were sourced from the United States Environmental Protection Agency (EPA) from their National Rivers and Streams Assessment (NRSA) for 2008-09 (United States Environmental Protection Agency [EPA], 2009). Three datasets were sourced from this assessment – benthic macroinvertebrate survey, water chemistry survey and site information data (EPA, 2009). Methods of collection were as per described in the EPA field operations manual (EPA, 2019).

The three datasets were sequentially revised for this project by filtering them on two main levels, where applicable. They were first filtered by the month of data collection, where only the data from sampling done in the summer months (i.e. June, July and August) were kept. This is because benthic community composition has been shown to vary seasonally, where some animals are more abundant at different times of the year (Jones et al., 2005). To control for this seasonal variation, only data collected in the summer months, the months in which it has been found that the community composition of benthics and environmental conditions are most sensitive to disturbance, were chosen (Jones et al., 2005). Furthermore, most of the sampling done by the EPA for NRSA were done in the summer months. Therefore, by filtering for survey results from the summer months, seasonal variations and temporal sampling bias in benthic macroinvertebrate abundances were controlled. The datasets were further filtered by sample category, where only primary observations were kept in the data. This was done for the purpose of removing duplicate observations from the dataset.

Following this, each of the data sources were filtered in a dataset-specific manner for the measurements of interest. The water chemistry dataset was filtered for five target observations – pH, nitrogen, phosphorus, potassium and dissolved carbon – the latter four of which were converted to micrograms/L in order to standardize the measurement scale. The site dataset was used to extract latitude and longitude data corresponding to each site.

Observation averaging was then performed on the benthic macroinvertebrate and water chemistry datasets; values for water chemicals for each of the five variables were averaged for each site per year (2008 and 2009). Following this, each site was graded based on a two-point Likert-scale and assigned a quality of either “good” or “poor” for each target chemistry variable based on literature threshold values determined by the EPA (EPA, 2000). For chemical variables that the EPA did not have threshold values, thresholds determined by the United Nations Institute for Environment and Human Security (UNU-EHS) (2016) were used. Ultimately, thresholds for nitrogen, phosphorus and pH were obtained from the UNU-EHS (UNU-EHS, 2016). Thresholds for phosphorus and potassium were obtained from the EPA (EPA, 2000). Benthic macroinvertebrates were also averaged for each site per year and the EPT index was then calculated for each site per year using the following equation:

(Plecoptera abundance + Trichoptera abundance + Ephemeroptera abundance) /
(Total macroinvertebrate abundance)

Most of the referenced EPT studies calculate the EPT Index as the sum of the absolute number of species of each of these three orders (Jones et al., 2005). The EPT calculation for this paper, however, was different for two primary reasons. Firstly, the NRSA survey did not include data on the number of species sampled for each order. Secondly, this study aimed to account for spatial variation in the distribution of all benthic macroinvertebrate abundances. To standardize the index, EPT was calculated as a relative measure of the abundances of each of the three target orders compared to the abundances of all benthic macroinvertebrates within each site.

All three datasets were then conglomerated to obtain our working dataframe which contained concentration values for nitrogen, phosphorus, potassium, dissolved carbon and pH, as well as the EPT index and site location for each site per year.

Shapiro-Wilk and Wilcoxon rank-sum tests

Shapiro-Wilk tests were performed to check for normality of distribution of the EPT indices from each of the two sampling years. These tests concluded that both of these distributions were not normally distributed, justifying the subsequent use of a Wilcoxon rank-sum test for testing the difference in medians between the distributions from the two years. A Shapiro-Wilk test was also performed on the working dataframe with both years combined, which concluded that this data was not normally distributed either.

The NRSA data showed complete separation, where sites were only sampled in 2008 or 2009, but never both. To accomodate for any variations due to sampling year, a Wilcoxon rank-sum test was performed on the data from 2008 and 2009 combined. The p-value of this test was 0.9047. Therefore, the null hypothesis could not be rejected and the medians for both years were not significantly different from each other. As a result, the data was treated as one continuous sampling season for further analyses.

Modelling

Two main models were created to analyze the effects of water chemical content on water quality assessed via the EPT index. First, a linear model was created (sample size = 132) with the response variable being the EPT index values and the predictor variables being site qualities for each of the five chemicals of interest; nitrogen, phosphorus, potassium, pH and dissolved carbon, as well as all combinations of their interactions. A linear mixed effects model (sample size = 132) was then created following the same format as the linear model, with year added as a random effect. Upon preliminary analyses, the data showed nesting by year only and not by site ID because each site only had one observation in the summarized dataset. This justifies the use of only year as a random effect within the linear mixed effects model. AIC values were used to compare model structures for the two models. The output showed that the linear model without any random effects had a lower AIC score than the linear mixed effects model. As a result, the linear model was used to interpret the data. An effort was made to construct multiple, different linear regression models in order to produce an optimized model via model averaging (using the dredge function). However, there was an unresolvable error where the C wrapper and its associated sprintf function could not appropriately handle +/- Inf values as well as NaN values. Thus, this method was abandoned because a literature review produced no documented solution. Finally, a residuals vs. fitted plot was constructed in order to check the linear regression model for homoscedasticity.

Scatterplot, Distribution Histogram and Violin Plot

Based on the output of the linear model, the interaction between nitrogen site quality and phosphorus site quality had the highest absolute estimate value for the EPT response variable. A violin plot was plotted to show the density of EPT index values across the interaction combinations between nitrogen and phosphorus site qualities. A dotplot was layered on top to show the distribution of EPT index values across these interaction combinations. Mean and significant error terms were also plotted. The y-axis was log-transformed to elucidate the EPT distribution.

A faceted scatterplot was then created to show the distribution of EPT indices across each of the five target chemistry variables. Both axes of the scatterplot were log-transformed to visualize this distribution. A distribution histogram was also created to show the frequency distribution of EPT indices for the data.

Raster plots and Potomac River Investigation

Raster plots were created using ggmap to show the distribution of EPT indices across the United States for all sites and for both years. Most of the sampling was found to be concentrated around the north-eastern United States. To visualize this densely clustered area, a raster plot of the area between -79 to -76.5 degrees of longitude and 38.75-39.75 degrees of latitude was produced. In this area, a single site with a relatively high EPT index value was found. Given the peculiarity of this site, its location was investigated further. This anomalous site was found to be close to a water purification facility – the Chambersburg Wastewater Treatment Plant. This site lay along a tributary of the Potomac river which emptied into the main river. Given that multiple sampling sites existed both upstream and downstream of the site of interest, an investigation as to how water quality changed in respect to distance from the water treatment plant was conducted. To do this, six sites that were in relatively rural areas and along the Potomac in proximity to the water treatment plant were chosen. Two of the sites were located upstream of the water treatment facility and three of the sites were located downstream of the water treatment facility. Despite more sites existing further downstream of the water treatment facility, these sites were found along highly urbanized areas and were thus not included in the investigation to account for the potentially confounding effects of urban development on water quality.

Following this, the absolute distance of each site was calculated from the Chambersburg Wastewater Treatment Plant in kilometers using the google maps 'measure distance' tool. Two plots were then created, plotting the distribution of EPT indices against distance from the plant, as well as the distribution of each target chemistry variable over increasing distance from the plant. Additional linear models were constructed to fit regression lines onto the line plots. Each linear regression model investigated if distance (the predictor variable) had a significant effect on EPT index, nitrogen content, phosphorus content, potassium content, dissolved carbon content or pH. Each linear model had a sample size of 6.

Results

Distribution Histogram and Scatterplot

Shapiro-Wilk tests performed on the data filtered for each year produced non-significant results (p-values = 6.229*10^-13 for 2008 and 2.066*10^-15 for 2009), indicating non-normality of distribution. Figure 1 shows the log-transformed distribution of EPT indices by their frequency of occurrence. Data was not normally distributed (Shapiro-Wilk's test p-value = 2.2*10^-16). Low EPT indices occurred more frequently and the distribution was visibly right-skewed. This indicates relatively low EPT indices all across the sites sampled by the NRSA.

A plot of the distribution of EPT indices across concentrations of the five target chemistry variables (Figure 2) showed no clear trend in EPT distribution. This is further supported by the linear model output (Table 1) which concluded that none of the target chemistry variables were significant predictors of EPT index on their own (p-values were all greater than 0.05).

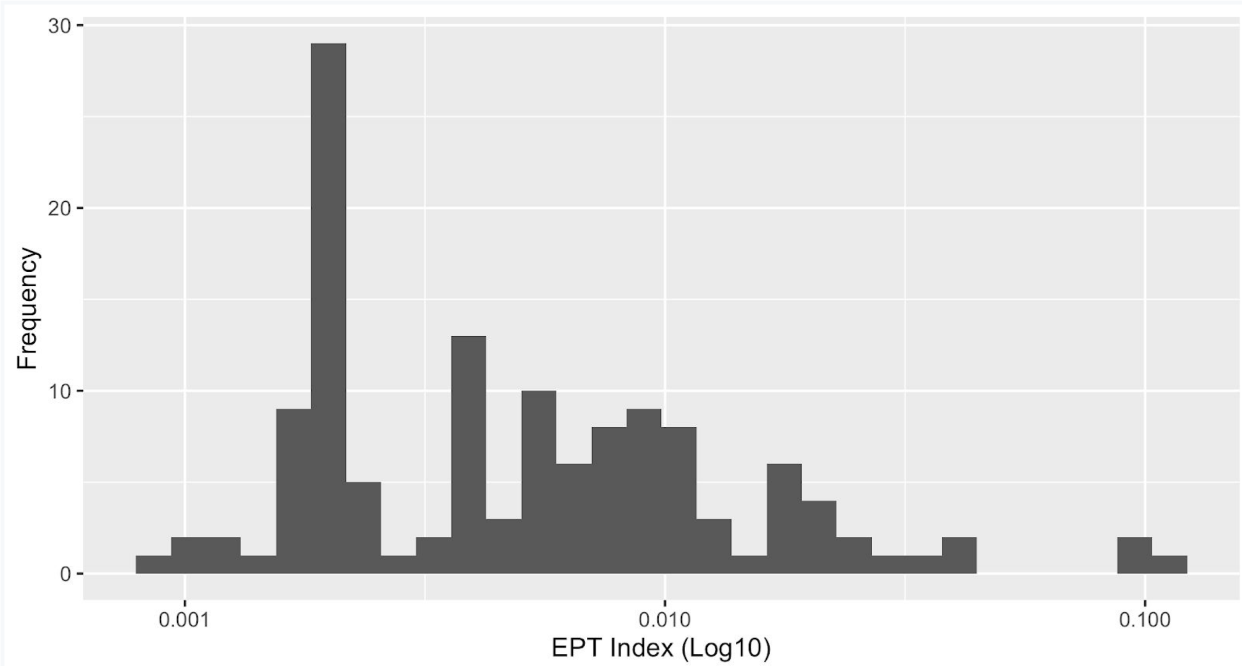

Figure 1. Histogram depicting the frequency distributions of EPT indices across all sites sampled during the years 2008 and 2009. Distribution is not normal (Shapiro-Wilks test p-value = 2.2*10^-16). EPT index values have been log-transformed.

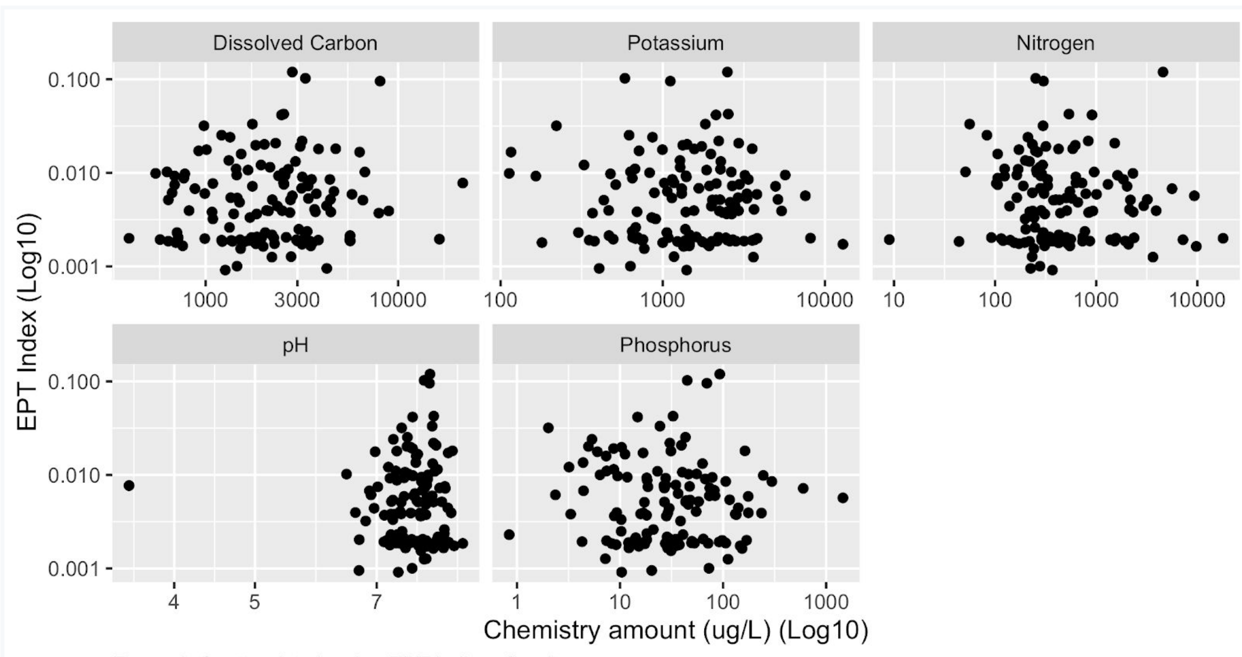

Figure 2. Scatterplot showing EPT indices for sites with concentrations of the target variables; dissolved carbon, potassium, nitrogen, pH and phosphorus, as the independent variables in both 2008 and 2009 combined. Each point represents one sample site. Note that chemical concentrations, as well as EPT index values, have been log transformed.

Model Output and Violin Plot

Figure 3 tests the assumption of the linear model to check whether all predictor variables and their interactions had constant variation. Here, a fanning effect was clearly distinguished, heavily implying heteroscedasticity of the predictor variables and their interactions.

Table 1 summarises the output of the main linear regression model (without year as a random effect) and the supplementary linear mixed effects regression model (with year as a random effect). Table 1 shows how none of the predictor variables and their interactions had a significant effect on the response variable (EPT index). However, the interaction between nitrogen site quality and phosphorus site quality had the largest, absolute estimator value (-0.03). Additionally, Table 1 shows how the linear regression model without year as a random effect had a lower AIC value than the linear mixed-effects regression model with year as a random effect.

The results of the linear model which accounted for the interaction between nitrogen and phosphorus show that EPT index values in order of lowest to highest occur in the interaction between good nitrogen and good phosphorus site quality (EPT = 0.008110306), poor nitrogen and poor phosphorus site quality (EPT = 0.008992542), good nitrogen and poor phosphorus site quality (EPT = 0.01057073) and poor nitrogen and good phosphorus site quality (EPT = 0.01125076). Differences between each mean EPT values were not statistically significant (p-value = 0.8416)

Table 1. Table summarizing the output of the two linear models run on the data for this study. For both models, EPT index was the response variable and the fixed effect predictors were the same. The first three Estimates, CI and p columns represent the estimated predictor effect, 95% confidence interval values and p-values for each predictor effect in the linear model with no random effects. The next three Estimates, CI and p columns represent the linear mixed effects model, with year as a random effect. The last column represents the degrees of freedom corresponding to each fixed effect for both models. The output for the random effect of year in the linear mixed effects model is included below the fixed effects. Number of observations for both studies were 132. R2 and adjusted R2 values for both models are also included, as well as the AIC values. The linear model with no random effects had the lower AIC score. All values have been rounded to three decimal places.

	EPT Index						
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*df*
Intercept	0.01	-0.03 – 0.05	0.694	0.01	-0.03 – 0.05	0.694	116.34
Nitrogen (Poor)	0.02	-0.03 – 0.07	0.456	0.02	-0.03 – 0.07	0.456	116.99
Phosphorus (Poor)	0.00	-0.03 – 0.04	0.927	0.00	-0.03 – 0.04	0.808	116.63
Potassium (Poor)	-0.01	-0.04 – 0.03	0.759	-0.01	-0.04 – 0.03	0.810	116.26
Dissolved Carbon (Poor)	-0.00	-0.04 – 0.04	0.975	-0.00	-0.04 – 0.04	0.975	116.93
pH (Poor)	-0.00	-0.04 – 0.03	0.930	-0.00	-0.04 – 0.03	0.759	116.21
Nitrogen (Poor) * Phosphorus (Poor)	-0.03	-0.06 – 0.01	0.201	-0.03	-0.06 – 0.01	0.930	116.50
Nitrogen (Poor) * Potassium (Poor)	-0.02	-0.08 – 0.04	0.543	-0.02	-0.08 – 0.04	0.377	116.95
Phosphorus (Poor) * Potassium (Poor)	0.00	-0.02 – 0.02	0.808	0.00	-0.02 – 0.02	0.543	116.12
Nitrogen (Poor) * Dissolved Carbon (Poor)	0.02	-0.02 – 0.05	0.377	0.02	-0.02 – 0.05	0.697	116.04
Phosphorus (Poor) * Dissolved Carbon (Poor)	-0.00	-0.03 – 0.03	0.996	-0.00	-0.03 – 0.03	0.996	114.57
Potassium (Poor) * Dissolved Carbon (Poor)	0.01	-0.02 – 0.04	0.697	0.01	-0.02 – 0.04	0.927	116.48
Nitrogen (Poor) * pH (Poor)	-0.01	-0.06 – 0.04	0.810	-0.01	-0.06 – 0.04	0.333	116.01
Nitrogen (Poor) * Phosphorus (Poor) * Potassium (Poor)	0.02	-0.02 – 0.07	0.291	0.02	-0.02 – 0.07	0.291	116.49
Nitrogen (Poor) * Potassium (Poor) * Dissolved Carbon (Poor)	-0.02	-0.06 – 0.02	0.333	-0.02	-0.06 – 0.02	0.201	116.99
Random Effects							
σ^2					0.00		
τ_{00}					0.00 $_{YEAR}$		
N					2 $_{YEAR}$		
Observations		132			132		
R^2 / R^2 adjusted		0.069 / -0.042			0.062 / NA		
AIC		-684.246			-563.473		

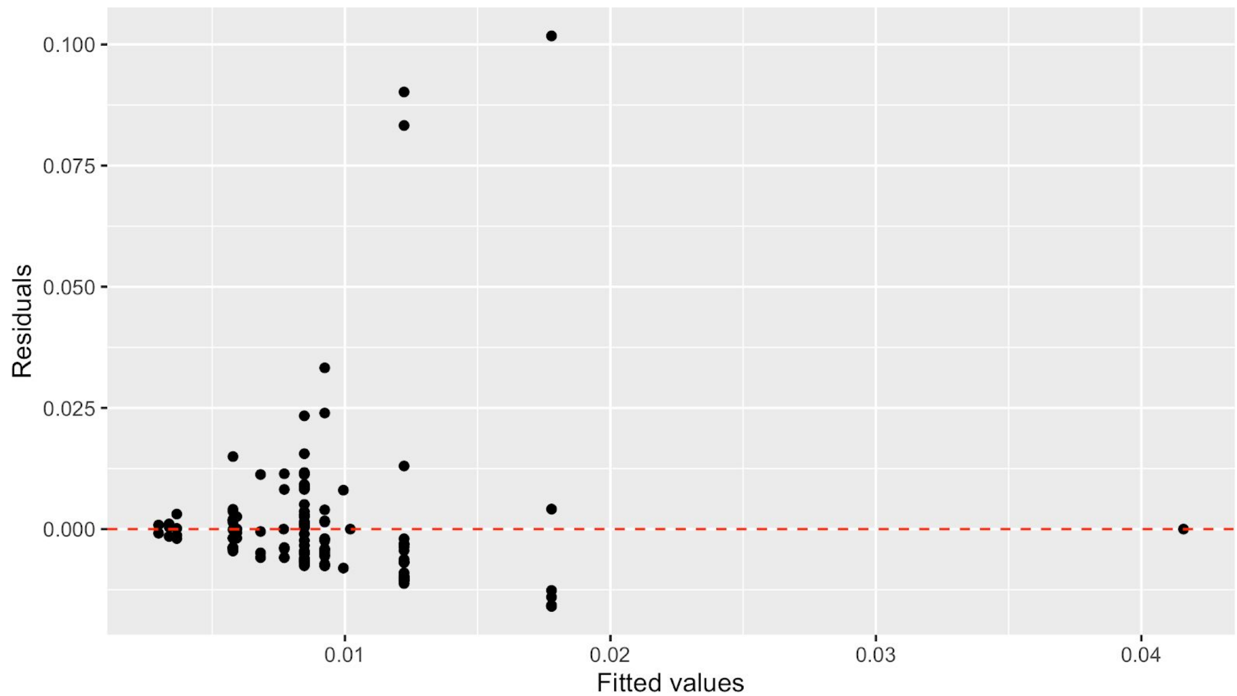

Figure 3. Plot of the residual values against fitted values for the linear model
with the equation (EPT Index ~ nitrogen * phosphorus * potassium * dissolved carbon * pH).

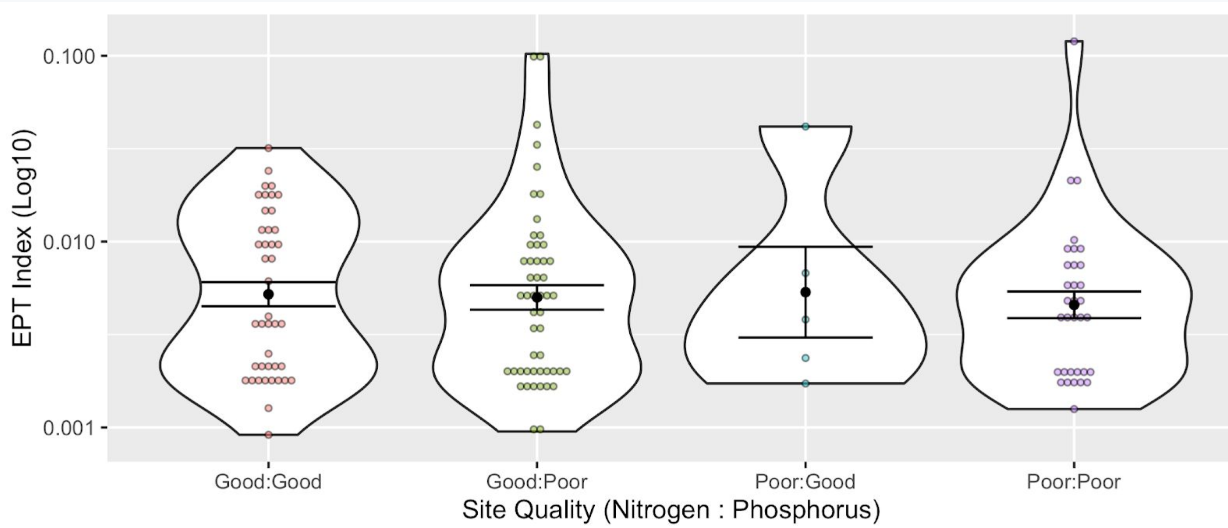

Figure 4. Violin plot depicting the results of a linear model
with the effects of the interaction between nitrogen site quality and phosphorus site quality.
EPT index values are shown for the different combinations of nitrogen-phosphorus site quality interactions.
Sites were given a site quality ranking of ‘good’ if they were below the maximum threshold deemed safe by the USEPA
If they were above this threshold, a site quality ranking of ‘poor’ was given.
Threshold value for nitrogen concentrations was 700 micrograms per litre.
Threshold value for phosphorus was 20 micrograms per litre.
EPT index value for good:good was 0.008110306. EPT index value for good:poor was 0.01057073.
EPT index value for poor:good was 0.01125076. EPT index value for poor:poor was 0.008992542.
P value = 0.8416. EPT Index values have been log transformed.

Raster Plots and Potomac River Investigation

As represented in Figure 5, EPT indices were generally relatively low for most sites across the US. Three sites stand out in this figure as having relatively high EPT indices. Upon further investigation, these three sites were found to have some level of correlation with surrounding land use. The site in Wyoming was an isolated site in Yellowstone National Park; the one in New Mexico was located by a relatively isolated freshwater spring. The site in Maryland appeared to be an anomaly since it was located in the middle of a dense urban area and was surrounded by sites with much lower EPT indices. Further investigation showed that this site was present just downstream of the Chambersburg Water Treatment Plant. While all anomalous sites were found to be in close proximity to land-use types that could explain their values, no tests were conducted to prove their association, therefore their correlation should be taken as suggestive, not definitive.

Upon further investigation of the sites both upstream and downstream of the Chambersburg water treatment plant (Figure 6), it was found that all five of these sites had relatively poor water quality as represented by their relatively low EPT index values. Looking at Figure 7, it is visible that the sample site located closest in proximity to and downstream of the water treatment had the highest EPT index value, and consequently, best water quality. Progressively lower EPT index values were found with increasing distance from the Chambersburg Water Treatment Plant. Figure 8 plots the distribution of the five target chemistry variables along the same axis of distance from the Wastewater Treatment Plant. Most of these variables showed no significant trend as seen in the p-values reported in Table 2. One observation stood out in this figure - the nitrogen spike at the site with the best water quality. Possible reasons for this spike are discussed in the Discussion section of this paper.

As seen in Table 2, distance was not a significant predictor of EPT Index or of any of the five target variables in the Potomac river investigation (p-values > 0.05). The model with EPT as the response variable was the best fit for this data, since it had the lowest AICc value. Even though Figure 7 may not show a significant trend, there are some clear patterns in this distribution and the land use corresponding to these sites may serve as one possible explanation. Table 2 also concludes that distance from the water purification facility did not significantly predict distribution of the target chemicals. This provides further support to the output of the main linear model (Table 1) where none of the target chemicals significantly predicted EPT distribution individually.

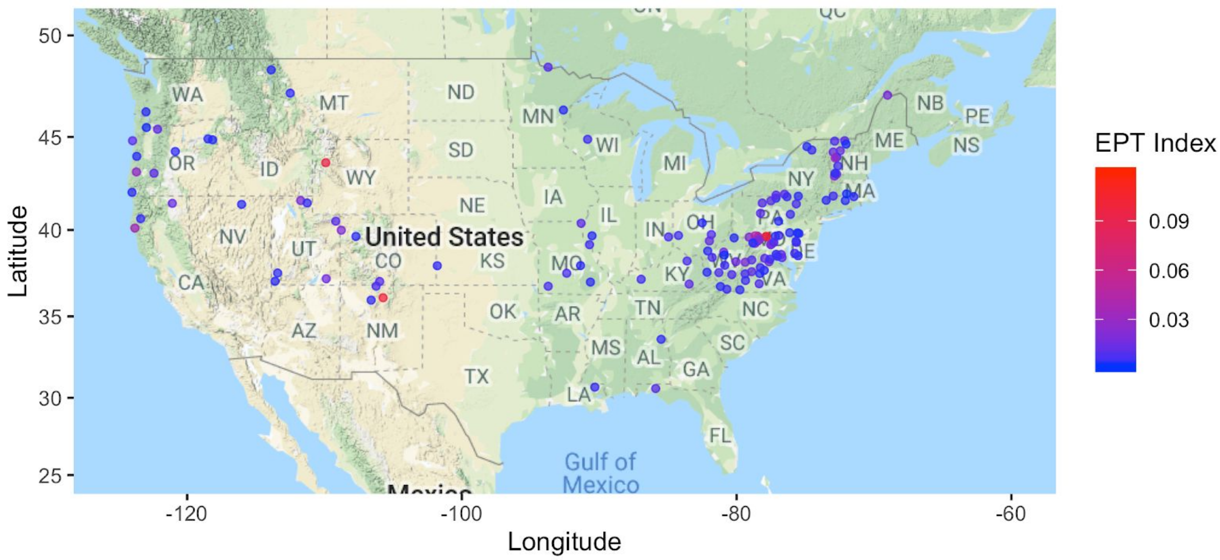

Figure 5. Raster plot depicting EPT index values and location for all sites sampled in both 2008 and 2009 throughout the continental United States.

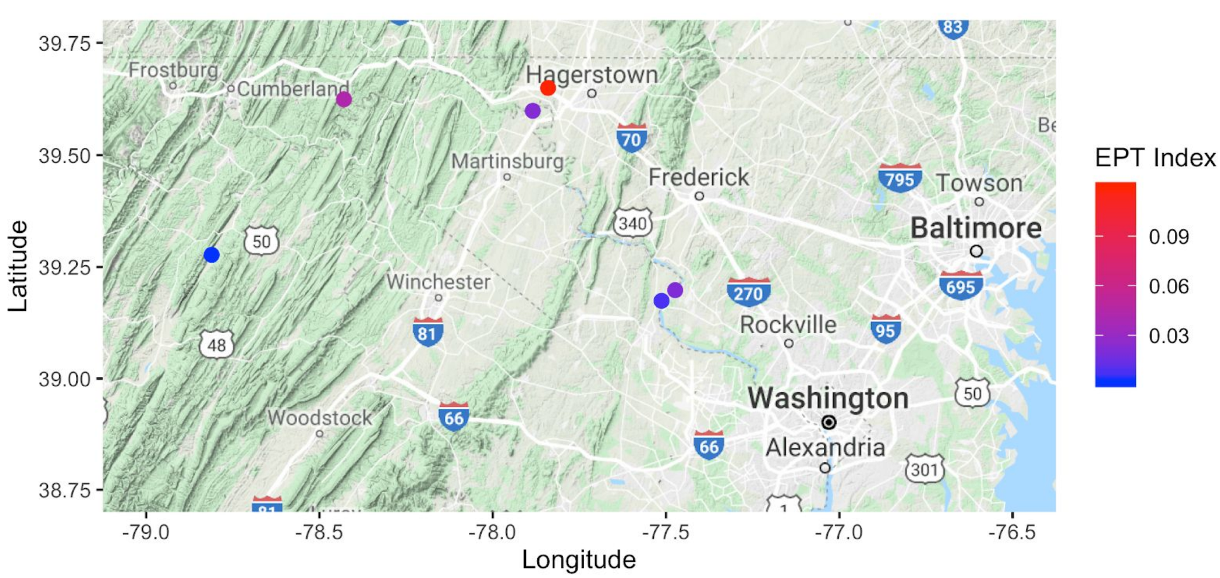

Figure 6. Raster plot depicting EPT index values and location for sites along the Potomac river and in close proximity to the Chambersburg Water Treatment Facility. Facility is located at 39°88'69.53"N 77°69'75.46"E.

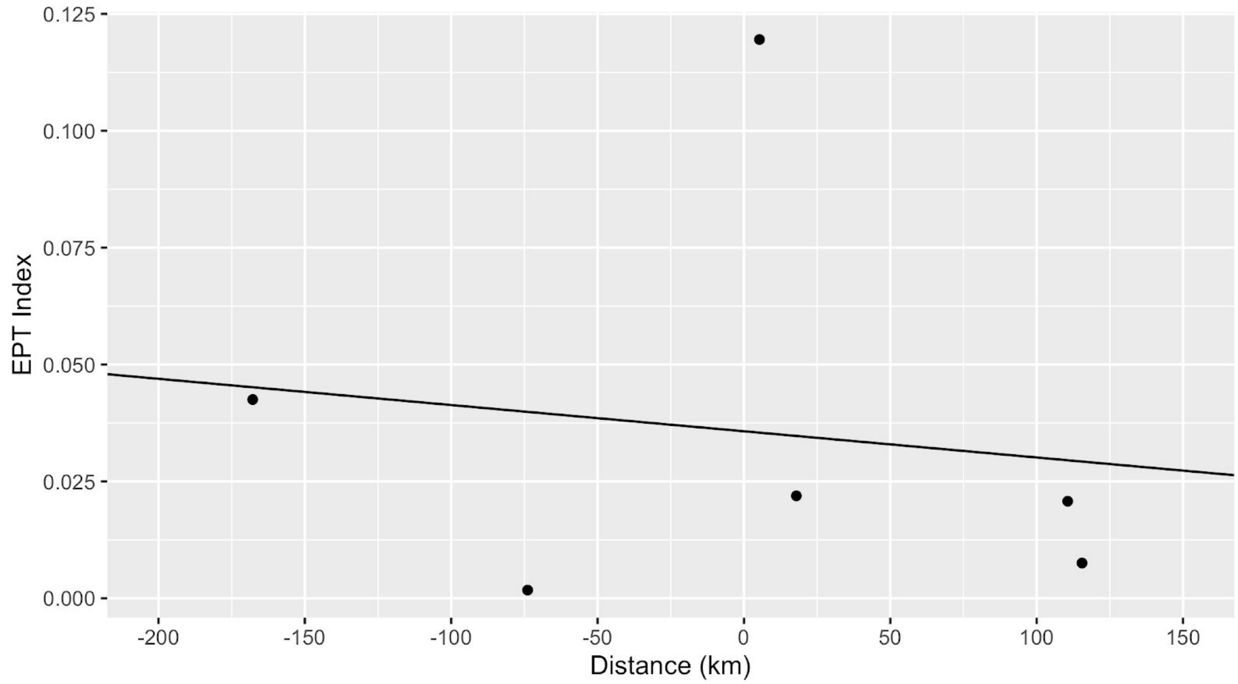

Figure 7. Distribution of EPT indices across sites corresponding to their distance from the Chambersburg Wastewater Treatment Plant along the Potomac river, both upstream and downstream. Upstream sites represent negative distance and downstream sites represent positive distance.

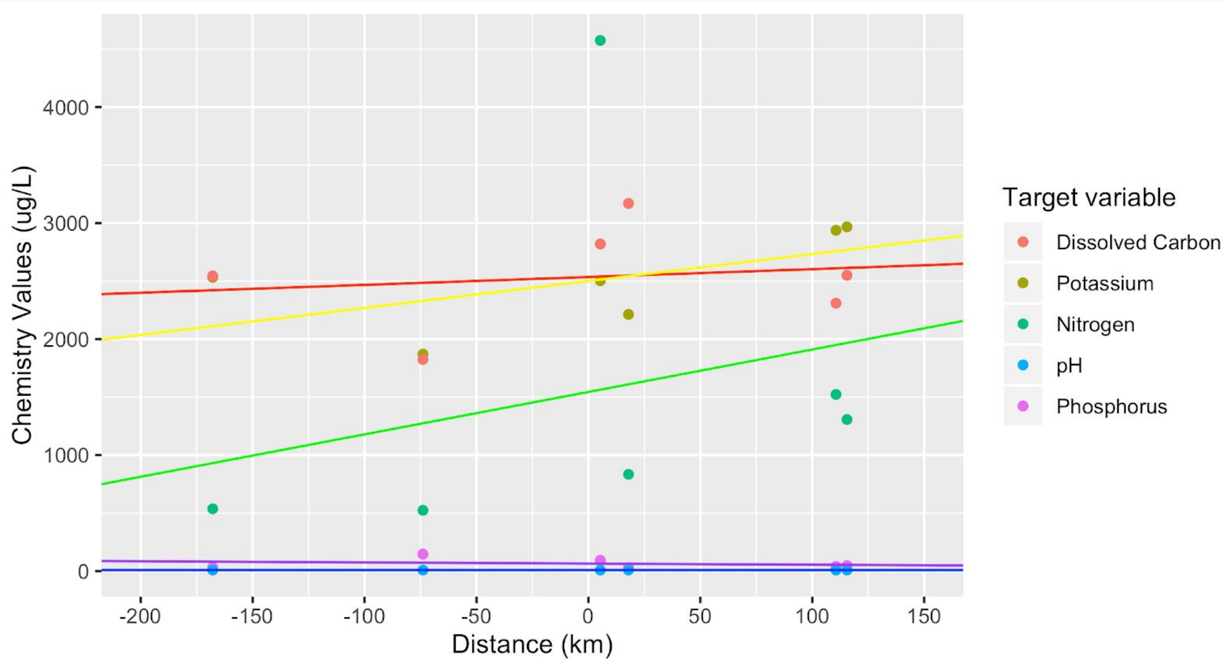

Figure 8. Distribution of the 5 target chemistry values across sites corresponding to their distance from the Chambersburg Wastewater Treatment Plant along the Potomac river, both upstream and downstream. Upstream sites represent negative distance and downstream sites represent positive distance.

Table 2. Summary of the output of linear models in the Potomac river investigation. A total of six linear models were run, with distance from the Chambersburg Wastewater Treatment Plant as the predictor variable for all of them. EPT index (Figure 7) and the five target chemistry variables (Figure 8) served as the respective response variables. For each model, the estimate value of the predictor and the intercept, 95% confidence interval values and p-values have been included. Total number of observations for all models were 6. R2 and adjusted R2 values for all models are also included, as well as the AIC~C values. The model with EPT Index as the response variable had the lowest AIC~C score. All values have been rounded to three decimal places.

	EPT Index			*Nitrogen*			*Phosphorus*			*Potassium*			*Dissolved Carbon*			*pH*		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
Intercept	0.04	-0.02 – 0.09	0.143	1545.39	-334.74 – 3425.51	0.085	64.97	8.23 – 121.72	**0.034**	2501.24	2073.82 – 2928.66	**<0.001**	2535.81	1965.66 – 3105.96	**<0.001**	8.31	8.04 – 8.58	**<0.001**
Distance	-0.00	-0.00 – 0.00	0.790	3.66	-15.21 – 22.53	0.619	-0.10	-0.67 – 0.47	0.652	2.32	-1.97 – 6.62	0.207	0.68	-5.04 – 6.40	0.758	-0.00	-0.00 – 0.00	0.987
Observations	6			6			6			6			6			6		
R^2 / R^2 adjusted	0.020 / -0.225			0.068 / -0.165			0.056 / -0.180			0.361 / 0.202			0.027 / -0.217			0.000 / -0.250		
AIC	-3.825			121.559			79.553			103.783			107.241			15.203		

Discussion

The linear regression model with EPT index as the response variable and the site qualities for each of the five target chemistry variables (nitrogen, phosphorus, potassium, dissolved carbon, pH) as the predictors did not produce any significant results. It was found that none of the predictors had a significant effect on EPT. Similarly, none of the interactions between any of the target chemical variables were found to be significant. That being said, the interaction between nitrogen site quality and phosphorus site quality produced the highest, absolute estimate value (-0.03) and the smallest p value (0.21). Previous studies have found that for a variety of algal species and bacteria, nitrogen and phosphorus have synergistic effects leading to their bloom (Zeng et al., 2016). Therefore, this synergistic effect could be responsible for the relatively high impact of this interaction on EPT index values relative to other chemicals or their interactions. There are a variety of factors that could account for the non-significance of the results. To begin with, the models are highly dependent on water chemistry threshold values. This study used the threshold values presented by the EPA and the UNU-EHS. However, when comparing these threshold values against concentrations tested in the literature, it is found that the values outlined by both of these institutions are generally higher than those found to have negative effects in scientific studies. For example, while the EPA threshold value for nitrogen is 1mg/L, a study conducted by Zeng et al., (2016) found that concentrations of nitrogen equal to or greater than 0.3mg/L led to eutrophication in the natural river systems tested (EPA, 2000). Therefore, it could be that the threshold value utilized in the model could be obscuring significant results.

The main linear regression model which did not include the random effect of year was used for the aforementioned analyses, yet using this model presents a couple key issues. Firstly, it is possible that the linear model which did not include the random effect of year might have had a lower AIC value than the model which did, not because it was indeed a better model, but rather because it was biased due to sampling design. None of the sites included in our study were sampled in both years that the study covered. Rather, all sites were only sampled in either 2008 or 2009 exclusively. Therefore, data suffered from complete separation, producing the Hauck-Donner effect. The Hauck-Donner effect would have caused abnormally large Wald standard errors and inflated p-values. Another issue associated with using this model comes from the variance of the data. One of the major assumptions of linear regression models is that variance is constant across all range of values. However when plotting residual values against fitted values of the model, a fanning effect is clearly distinguishable. Points are not evenly distributed across the y intercept. Therefore, it is clear that our data violates the assumption of constant variance. Instead, our data is heteroscedastic. Future research in this topic should take care to construct a regression model which accounts for the heteroscedasticity of the data. Additionally, the fact that the EPT index values were not normally distributed presents another issue with the use of the model in the study. Again, an assumption of linear regression models is that data is normally distributed. However, as seen in Figure 1, EPT index values are right-skewed. Using skewed data in our model produces estimate values that do not represent maximum likelihood values.

Although outliers are not as visible in the log-transformed scatterplot of EPT index values against concentrations of the five chemical variables (Figure 2), when the axes were not log-transformed, a few sites stood out as clear outliers. To answer the question of whether a relationship existed between water quality as measured by the EPT index and land use, the sites which represented outliers amongst the data were investigated further. For nitrogen, the site which had an anomalously high concentration value was found to be located in the midst of cropland. It is possible that runoff from agricultural fertilizers could account for the increased concentration of nitrogen found in this site (Sharip et al., 2012). For pH, the outlier which was found to have acidic water was proven to be located in close proximity to a hydraulic fracturing facility. Given that hydraulic fracturing involves the use of a plethora of chemicals to facilitate gas extraction, accidental leakage of these chemicals into the watershed could account for the abnormally low pH value observed at this site (Jackson et al., 2014). For potassium, the site presenting an outlier was found to occur in close proximity to a granite mining facility. Given that potassium represents one of the elements that constitute granite, it is possible that flow of dust or the runoff of rainwater from this facility into the watershed could have been responsible for the anomalous concentration of potassium in the water at this sample site (Hinchey et al., 2011). For phosphorus, the site that produced an outlier was found next to a mechanic workshop. Phosphorus is an additive in car engine oil; therefore, it is possible that oil spills from cars at the mechanic and their subsequent runoff into the riverbed could be responsible for the increased concentrations of phosphorus observed at this site (Spikes, 2008). For dissolved carbon, there were two sites that produced outliers, both with extremely high dissolved carbon concentrations. One of these sites was downstream of the Lake of the Woods, which has been documented to undergo periodic algal blooms (Sellers, 2018). Dissolved carbon content in water is positively correlated to eutrophication and as such, these algal blooms may explain the abnormally high carbon content of the water (Sellers, 2018). The second of these two sites was located downstream of Oak Lake, where the nutrient concentrations are documented to be driven by shoreline development practices, septic systems, agricultural inputs and internal loading (Hutchinson Environmental Sciences Ltd. [HES], 2019). Assessment of the water quality in Oak Lake showed abnormally high concentrations of dissolved carbon as well, and as such, any combination of the aforementioned anthropogenic factors influencing Oak Lake's water quality could be driving the high dissolved carbon concentration in this outlying site (HES, 2019). While possible reasons for the anomalies were found, no statistical tests or further studies were conducted to prove their correlation, therefore their relationship should be interpreted as suggestive not correlative. Possibilities for expansion of this study to prove correlation should include tests which control for confounding variables and that focus on each site mentioned.

When comparing EPT index values along a gradient away from the Chambersburg Water Treatment Plant located along the Potomac river, it was found that sample sites in close proximity and downstream of the water treatment plant had the highest EPT index values relative to those further downstream and upstream of the water treatment plant. While the relationship between distance from water treatment plant and EPT index values was not statistically significant, some clear patterns between distance and EPT index values could be seen. Upstream of the water treatment plant, EPT indices were low. Downstream of the water treatment plant, with increasing distance, lower EPT index values were observed. The site with the highest EPT index value was the one in closest proximity and downstream of the water treatment plant. Despite these patterns, the relationship cannot be attributed completely to distance from the water treatment plant. To begin with, this study did not control for other variables that could confound the results of the EPT index calculations. Research shows that factors such as leaf litter concentrations, water temperature, and dissolved oxygen content can affect EPT index values (Narangarvuu et al., 2015). Therefore, while a gradient seems to show decreasing EPT index values with increasing distance, other variables could be contributing to this result as well.

When investigating the relationship between chemistry values and distance to the water treatment plant, no significant trend was seen for any of the five target variables of our investigation (Figure 8). This further supports the results of our main linear model which suggests that none of these target variables have an effect on EPT index values on their own, rather, it is their interactions that possibly produce an effect. Another interesting result of this investigation showed that nitrogen concentration was highest at the site with the highest EPT index value. Based on nitrogen thresholds in the literature, one would expect that these concentrations would lead to bad water quality (UNU-EHS, 2016). However, as seen in our results, this is not the case. A few factors could be responsible for this outcome. To begin with, concentrations could have inaccurately been measured due to sampling or technological error. Additionally, it could be that as the literature suggests, worst water quality effects are seen with a synergistic combination of chemicals, not chemicals on their own. Therefore, it could be that at this site, this combination of concentrations of chemicals in the water did not negatively affect water quality.

Conclusion

The results of our study showed that overwhelmingly, EPT indices in rivers and streams across the continental United States were relatively low, suggesting poor water quality. However, given the limitations of our study, namely to do with sampling design and the consequent effect this had on data distribution, these conclusions cannot be reliably counted on.

Further studies on the topic should seek to improve sampling design by firstly, including species level data for benthic macroinvertebrates which would facilitate the calculation of EPT indices as in other similar studies. Secondly, to limit complete separation within data, all sample sites should be sampled annually. Despite these limitations, this study elucidated some key points, namely that land use can have an effect on water quality and that therefore, precautionary measures should be taken to minimize negative anthropogenic disturbances. This study further showed how water quality changes with concentrations of some key chemistry variables in the continental United States, namely the effect of the interaction between nitrogen and phosphorus. In doing so, it supported previous research which shows that fertilizer runoff poses serious threats to water quality. Therefore, reducing fertilizer runoff could be a useful strategy for water conservation.

Works Cited

Hinchey, A.M., Davis, W., Ryan, J., & Nadeau, L. (2011). Neoarchean high-potassium granites of the Boothia Mainland Area, Rae domain, Churchill Province: U-Pb zircon and Sm-Nd whole rock. *Canadian Journal of Earth Sciences,* 48(2), 247.

Hrovat, M., Urbanic, G., & Sivec, I. (2014). Aquatic insects along environmental gradients in a karst river system: A comparative analysis of EPT larvae assemblage components. *International Review of Hydrobiology*, 99(3), 222-235. doi:10.1002/iroh.201301641

Hutchinson Environmental Sciences Ltd. (2019). Oak Lake Water Quality Monitoring Report and Lake Management Plan. Retrieved from https://www.quintewest.ca/en/your-city-hall/resources/Public_Works/190723_J180050-Oak-Lake-Water-Quality-Report.pdf

Kerski, J. J. (2017). Water quality. *The Geography Teacher, 14*(1), 36-41. doi:10.1080/19338341.2016.1260623

Jackson, R.B., Vengosh, A., Carey, J.W., Davies, R.J., Darrah, T.H., O'Sullivan, F. & Pétron, G. (2014). The Environmental Costs and Benefits of Fracking. *Annual Review of Environment and Resources,* vol. 39, pp. 327-362.

Jones, C., Somers, K. M., Craig, B., Reynoldson, T.B. (2005). Ontario Benthos Biomonitoring Network Protocol Manual.

Li, Z., Zhao, Y., Xu, X., Han, R., Wang, M., & Wang, G. (2018). Migration and transformation of dissolved carbon during accumulated cyanobacteria decomposition in shallow eutrophic lakes: a simulated microcosm study. *PeerJ.* doi:10.7717/peerj.5922

Narangarvuu, D., Oyunbileg, J., Yang, P., & Boldgiv, B. (2015). Distribution of ephemeroptera, plecoptera, and trichoptera assemblages in relation to environmental variables in headwater streams of mongolia. *Environmental Earth Sciences*, 73(2), 835-847. doi:10.1007/s12665-013-2968-9

Reid, R., Somers, K., & David, S. (1995). Spatial and temporal variation in littoral-zone benthic invertebrates from 3 south-central Ontario lakes. *Canadian Journal of Fisheries and Aquatic Sciences,* 52(7), 1406-1420. doi:10.1139/f95-136

Sellers, T. (2018). Lake of the Woods Algae and Nutrients Highlighted in Aquatic Ecosystem Health Report. *International Joint Comission.* Retrieved from https://www.ijc.org/en/lake-woods-algae-and-nutrients-highlighted-aquatic-ecosystem-health-report

Sharip, Z., Schooler, S., Hipsey, M., & Hobbs, R. (2012). Eutrophication, agriculture and water level control shift aquatic plant communities from floating-leaved to submerged macrophytes in lake chini, malaysia. *Biological Invasions, 14*(5), 1029-1044. doi:10.1007/s10530-011-0137-1

Spikes, H. (2008). Low- and zero-sulphated ash, phosphorus and sulphur anti-wear additives. *Lubrication Science, 20*(2), 103-136.

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., . . . Sörlin, S. (2015). Planetary boundaries: Guiding human development on a changing planet. *Science, 347*(6223), 1259855. doi:10.1126/science.1259855

United Nations Institute for Environment and Human Security. (2016). International Water Quality Guidelines for Ecosystems (IWQGES). HOw to develop guidelines for healthy freshwater ecosystems. A policy oriented approach. *United Nations Environmental Programme.* Retrieved from http://web.unep.org/sites/default/files/Documents/20160315_iwqges_pd_final.pdf

United States Environmental Protection Agency. (2000). Ambient Water Quality Critera Recommendations. Information Supporting the Development of State and Tribal Nutrient Criteria. Rivers and Streams in Nutrient Ecoregion XIV. *United States Environmental Protection Agency.* Retrieved from https://www.epa.gov/sites/production/files/2018-10/documents/nutrient-criteria-manual-rivers-streams.pdf

United States Environmental Protection Agency. (2009). National Rivers and Streams Assessment (NRSA) 2008-2009 report. *United States Environmental Protection Agency.* Retrieved from https://www.epa.gov/national-aquatic-resource-surveys/nrsa

United States Environmental Protection Agency. (2009). National Rivers and Streams Assessment (NRSA) 0809 Benthic Macroinvertebrate Count - Data . *United States Environmental Protection Agency.* Retrieved from https://www.epa.gov/sites/production/files/2016-11/nrsa0809bentcts.csv

United States Environmental Protection Agency. (2009). National Rivers and Streams Assessment (NRSA) 0809 Site Information - Data. *United States Environmental Protection Agency.* Retrieved from https://www.epa.gov/sites/production/files/2015-09/siteinfo_0.csv

United States Environmental Protection Agency. (2009). National Rivers and Streams Assessment (NRSA) 0809 Water Chemistry - Data. *United States Environmental Protection Agency.* Retrieved from https://www.epa.gov/sites/production/files/2015-09/chem.csv

United States Environmental Protection Agency. (2019). National Rivers and Streams

Assessment 2018/2019 Field Operations Manual Non-Wadeable. Version 1.2

Zeng, Q., Qin, L., Bao, L., Li, Y., & Li, X. (2016). Critical nutrient thresholds needed to control eutrophication and synergistic interactions between phosphorus and different nitrogen sources. *Environmental Science and Pollution Research, 23*(20), 21008-21019. doi:10.1007/s11356-016-7321-x