University of North Carolina at Asheville
Journal of Undergraduate Research
Asheville, North Carolina
May 2013

Statistical Analysis of Water Quality in Western North Carolina Rivers

John Lombardi
Mathematics
The University of North Carolina at Asheville
One University Heights
Asheville, North Carolina 28804 USA

Faculty Advisor: Dr. Steve Patch

Abstract

Clean water and healthy watersheds are necessary for keeping water safe for human consumption and minimizing harmful changes to habitats for animals and plants by humans. Volunteers from the Volunteer Water Information Network collect samples from western North Carolina streams and rivers monthly. Samples are sent to the Environmental Quality Institute and analyzed for key indicators of water quality such as pH, total suspended solids, and Turbidity. Linear regression models are implemented to explore the effects of flow, season and time on water quality parameters. Least squares, auto-regressive, and moving-average models are compared and contrasted in order to assess which are best at predicting trends in the quantity of pollutants in western North Carolina Rivers. For five of the eight parameters analyzed, a least squares approach was best based on the criterion used. Significant trends were found for seasonality and flow where as a smaller amount of significant trends for time were found.

1. Introduction

Maintaining healthy water and watersheds is crucial for agriculture, keeping ecosystems healthy, and ensuring drinking water is safe for humans and animals. The Volunteer Water Information Network (VWIN) collects water samples from various rivers, streams, and lakes in Western North Carolina which are then analyzed by the UNC-Asheville Environmental Quality Institute (EQI). VWIN provides a database of information which EQI analyzes in order to assess water quality, what factors influence water quality, and provide quantitative information for environmental groups and local governments, in order to find areas that need the most work to improve water conditions.

VWIN data and EQI analysis of VWIN data were used from the French Broad Watershed in North Carolina. This includes Buncombe county, Haywood county, and Henderson county (Fig. 1). The French Broad watershed is a sizeable watershed; it is used as drinking water for 1 million people and is composed of 2,830 square miles of land in N.C.[1] The watershed is valuable not only for drinking water but also because it provides water to an extensive plant and animal network.

Determining the factors that influence water quality is greatly influenced by the type of model that is being used to predict changes for a water quality parameter. Temporal trends of water quality can be predicted well with linear models[2,3,4]. In order to select the best model two tests for the goodness of fit of a linear model were used to assess whether or not temporal trends of water quality parameters from month to month are necessary to account for when choosing a model. What factors have significant influences on water quality and what model best predicted changes in water quality were explored through statistical analysis.

2. Data

VWIN volunteers are trained by EQI, or partner organizations, about sample collection procedures by a VWIN coordinator as well as a training manual[6]. In order to reduce meteorological variability, samples are collected as close to noon as possible; however, samples are collected on separate days for the three counties in this study[6]. Buncombe and Haywood county samples are collected every second Saturday of each month; Henderson is sampled every third Saturday of each month[6]. Samples that were analyzed at other laboratories, when EQI closed temporarily, were also removed[3]. EQI analyzes water samples for 8

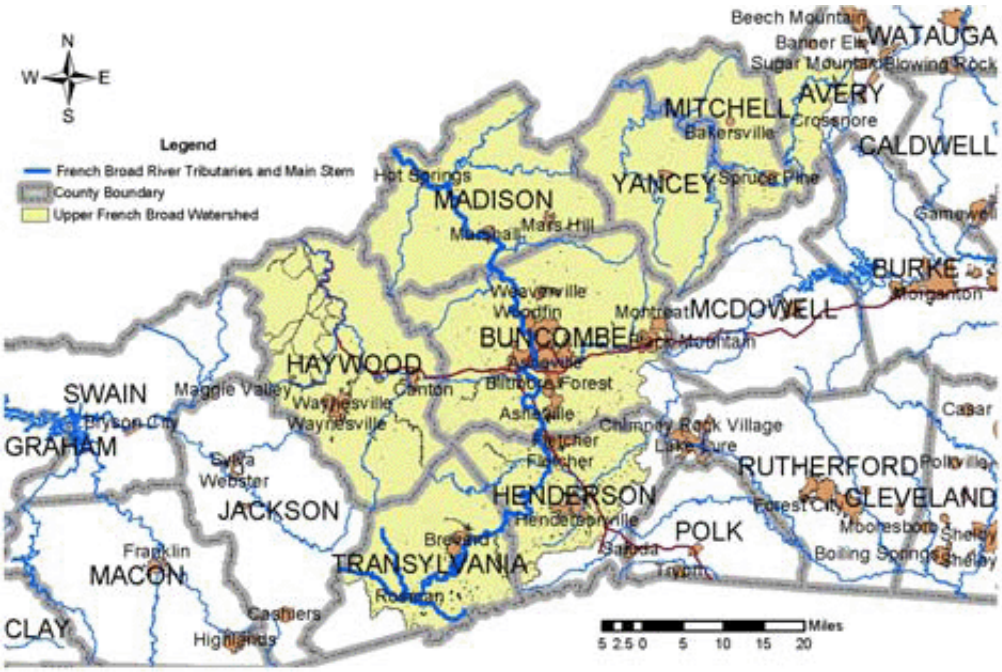

FIGURE 1. watershed separated by county[5]

different parameters: acidity, alkalinity, turbidity, total suspended solids (TSS), conductivity, ammonia-nitrogen (NH3), orthophosphate (PO4), and nitrate (NO3).

Data exists as early as 1990 but varies from site to site as well as from county to county. Sites were included only if they have ten years or more of data and have data for the past five years. After removing sites with less than ten years of data the total number of sites in the study was 114. Observations with a z-score> 3, after log transformation, were removed from each site[2]. Average proportions of outliers removed per site by parameter, as well as total amount of data for each parameter, are shown in table 1. When outliers were removed from the number of observations, as well as values that do not have a full data set for regression, i.e. missing a value for flow, then the total number of observations for each parameter in the regression was roughly 17,000. Outliers were removed because of interest in long-term trends of water quality; thus, removing instances where concentrations may be extremely high or low was necessary to remove their influence on predicting long-term changes to water quality. Flow was incorporated into the regression model and was obtained from USGS gages online[7]. Gages were selected for a site by similarity to stream/river conditions as well as minimizing geographic distance from the VWIN site.

Table 1. outliers and total number of observations

	Turb.	**TSS**	**NO3**	**NH3**	**PO4**	**Cond.**	**pH**	**Alk.**
Outlier proportion (Mean)	1.12%	0.81%	2.65%	1.60%	0.37%	0.92%	0.10%	0.50%
Total Observations	17884	17909	17538	17761	17993	17911	17987	17905

3. Methods

The statistical program R was used to aggregate the data and perform all statistical analyses. Three different linear models were tested in the study: least squares (LS), generalized least squares (GLS) with an auto-regressive error term, and GLS with a moving average error term. An auto-regressive model was implemented that was order one (AR1) which makes the assumption that the random error term in the regression, ϵ_t, was influenced by the previous error term. Auto-correlated regression error was defined to be: $\epsilon_t = \phi\epsilon_{t-1} + v_t$, where v_t is Gaussian white noise and are NID$(0, \sigma_v^2)$ and ϕ is the AR1 parameter[4]. The moving average

(MA) model was order one as well and the error was defined to be $\epsilon_t = v_t + \psi v_{t-1}$ [4]. Thus, the regression error is dependent on the Gaussian white noise from the current and previous error term.

The water quality parameters were log transformed in the form Log(parameter+ .5*Detection limit), except pH since it was already on a log scale. Detection limit was added before transforming parameters to eliminate instances in which a value was recorded as zero and would give negative infinity after being transformed. Statistical analysis of stream water level, temporal changes, and seasonality were explored through regression. Each regression had the following form

$$\text{Log}(\text{Parameter} + .5 * \text{Detection Limit}) = \text{Flow} + \text{Time} + \text{Fall.Vs.Spring} + \text{Winter.Vs.Summer} + \epsilon_i$$

Interlandi and Crockett observed strong influences on river discharge on water quality variables[3]. In order to adjust for the effect of instances where a big storm or heavy rain might affect parameters like TSS, flow was defined as follows:

$$\text{Log}(\frac{\text{Flow}}{\text{Mean Daily Flow}})$$

where mean daily flow is the long-term average flow for each date. Time was defined as

$$12 * (\text{Year} - 1990) + \text{Month}$$

or months since 1990. Seasonality was measured in the Spring vs Fall and Winter vs Summer term. These two variables set January 15th as the coldest day of the year, which is the coldest day of the year on average for western North Carolina[8]. This incorporates the effect of temperature on water quality. Seasonality was defined to be

$$\text{Fall.Vs.Spring} = \text{Sin}(\frac{2\pi x}{365}),$$

$$\text{Winter.Vs.Summer} = \text{Cos}(\frac{2\pi x}{365}),$$

where $x = (\text{Month} - 1 * 30.5 + (\text{Day}) - 15$.

For example, January 15th becomes: $(1-1)30.5 + (15) - 15 = 0$, and $Sin(0) = 0$, $Cos(0) = 1$. Thus making January 15th the coldest day for the term Winter.Vs.Summer. This method was chosen for its easy mathematical interpretation in lieu of treating season as a categorical variable.

In order to pick the best model for a parameter the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) was analyzed for the three models. The AIC and BIC are both measures of the goodness of fit for a linear model. Increasing the number of independent variables in a model will increase the goodness of fit for a model; thus, the AIC and BIC penalize for adding unnecessary independent variables. The AIC and BIC was computed for each site and then the median was computed for the AIC and BIC by parameter for every site. The model with the average of the AIC and BIC which was closest to zero was selected as the best model. Differences in the sign of the AIC and BIC for different parameters are caused by the maximized value of the likelihood function. The differences in sign are not significant; the minimum value for the AIC and BIC between models is the indicator for goodness of fit for the models tested. Once the model was selected, the Shapiro-Wilk test was used to determine the normality of the residuals[3]. The skewness and kurtosis of the residuals were also computed. Furthermore, the number of significant p-values at the $\alpha = .05$ level for each independent variable was calculated. The residuals and fitted values were graphed as well as the auto-correlation function for each parameter. For parameters that are best modelled by AR1 or MA1 an ANOVA test comparing the AR1 or MA1 to the LS model was used to compute the number of correlated error estimates that are significant ($\alpha = .05$).

4. Results

Results illustrate that there was no model that was consistently best for predicting trends for a water quality parameter (See Table 2) based on the criterion used. For five of the eight parameters a model that does not account for temporal correlation in the errors was best to detect trends to changes in water quality. For Turbidity, 38/114 of the parameter estimates for the moving average correlation were significant by the ANOVA model comparison to LS. NO3 had 72/114 significant phi estimates for the auto correlation and PO4 had 79/114. NO3 and PO4 had an auto regressive pattern that has statistical evidence which suggests

a necessity for its use instead of a regular LS model. Although the AIC and BIC was closest to zero for the MA1 model for Turbidity, the low number of significant estimates of the parameter ψ, 38/114, indicates that there was not strong evidence necessitating its use for turbidity.

TABLE 1. model results

Parameter	AIC	BIC	Average
Turbidity			
AR1	301.493	323.097	312.295
MA1	300.354	321.958	**311.156**
GLS	304.413	323.404	313.909
TSS			
AR1	357.798	379.937	368.8670
MA1	357.803	379.879	368.841
GLS	356.074	375.167	**365.620**
NH3			
AR1	-154.9	-132.16	-143.53
MA1	-145.42	-123.232	-134.326
GLS	-140.61	-122.935	**-131.773**
NO3			
AR1	90.287	109.876	**100.082**
MA1	90.52	110.059	100.290
GLS	97.755	116.258	107.006

Parameter	AIC	BIC	Average
PO4			
AR1	305.689	326.52	**316.105**
MA1	308.37	329.205	318.789
GLS	317.69	335.75	326.72
Conductivity			
AR1	-116.195	-96.6	-106.395
MA1	-112.48	-92.537	-102.509
GLS	-100.871	-85.406	**-93.138**
pH			
AR1	-37.84	-16.364	-27.103
MA1	-36.56	-14.872	-25.716
GLS	-34.318	-13.841	**-24.08**
Alkalinity			
AR1	11.742	32.92	22.333
MA1	11.07	33.429	22.24
GLS	12.868	29.55	**21.209**

After computing the best model, the number of significant p-values for the four independent variables was also computed (See Table 3). The regressions had very high numbers of significant p-values, indicating that there is strong evidence that water quality parameters are affected by the variables for season, time, and year.

TABLE 2. p-values for parameters for 114 sites

Parameter	Pos. Slope	Neg. Slope
Turb. (MA1)		
Flow	67	1
Time	11	15
Spring.Vs.Fall	61	2
Winter.Vs.Summer	4	68
TSS (LS)		
Flow	72	1
Time	10	14
Spring.Vs.Fall	86	1
Winter.Vs.Summer	1	85
NH3 (LS)		
Flow	20	11
Time	41	18
Spring.Vs.Fall	27	9
Winter.Vs.Summer	6	56
NO3 (AR1)		
Flow	26	9
Time	6	25
Spring.Vs.Fall	48	3
Winter.Vs.Summer	54	8

Parameter	Pos. Slope	Neg. Slope
PO4 (AR1)		
Flow	2	31
Time	26	14
Spring.Vs.Fall	1	8
Winter.Vs.Summer	1	47
Cond. (LS)		
Flow	4	62
Time	41	20
Spring.Vs.Fall	4	89
Winter.Vs.Summer	10	37
pH (LS)		
Flow	0	53
Time	24	9
Spring.Vs.Fall	1	51
Winter.Vs.Summer	3	95
Alk. (LS)		
Flow	2	56
Time	24	23
Spring.Vs.Fall	2	77
Winter.Vs.Summer	1	80

The lowest number of significant p-values for each parameter was often time. This indicates that, generally, since VWIN began sampling sites the month and year sampled was not a significant influence on water quality. For NH3 and Conductivity, time has 41 positive slopes suggesting that for a little less than half of all sites sampled, NH3 and Conductivity has increased over time. Alkalinity and NO3 have negative slopes for time for roughly 20% of sites, which shows that these two parameters have decreased for only a couple of sites since samples were first collected. Robinson et al. recorded decreases in nitrate over time as well[2].

Flow has significant p-values for over 50% of sites for 5/8 of the parameters. Flow was positive for most sites for Turbidity and TSS, which is to be expected since increased flow would cause more suspended solids and particulate matter to be present in the water. Alkalinity, pH, and Conductivity have strong statistical evidence suggesting they decrease with flow, for over 50% of sites, and have very few sites that increase with flow. NO3, NH3, and PO4 have a low number of sites with significant evidence for a positive or negative trend with flow.

Seasonality, as it was defined, does appear to be a good predictor on water quality; many sites have significant evidence for over 50% of sites. A positive slope for Spring.Vs.Fall would translate to a higher presence during Fall. Similarly, a positive slope for Winter.Vs.Summer indicates a higher presence of a parameter during winter. Turbidity and TSS both have statistical evidence that they are high in Fall and Winter which is also expected due to precipitation not varying drastically in western North Carolina which results in flow typically being higher in winter, thus increasing turbidity and TSS [4,7]. NH3 exhibits a similar pattern to turbidity and TSS, although there was a smaller number of sites that had strong statistical evidence that supports NH3 increasing in Fall. PO4, Conductivity, pH, and Alkalinity all exhibit a similar pattern: statistical evidence suggests that they all are higher in Fall and Summer. NO3 was the only parameter with a unique pattern for seasonality; results show that there is significant evidence that NO3 is highest in Spring and Winter.

Plots of residuals, on the y-axis, versus fitted values, on the x-axis, for all sites, separated by parameter, are shown in Figure 2. The residuals and fitted values are for the best model for each parameter. The graphs illustrate that no violations of homoscedasicity occur. Diagonal bands are present in the graphs for turbidity, TSS, NH3, NO3, PO4, and pH. These bands occur in the residuals when lower values of the parameter are rounded off to the same amount but have different predicted values.

Independence of errors are not expected to appear in graphs for AR1 and MA1 models since they were already assumed to exist and corrected for. Graphs for the AR1 and MA1 models verify that the models accounted for serial correlation. The GLS models also show that there was independence of the error terms.

The auto-correlation function for every site (ACF) was graphed, again separated by parameter, in figure 3. The graphs are the calculated auto-correlation function for every site after the best model to use for each parameter was chosen. The lags between residuals was graphed on the x-axis and the calculated ACF value on the y-axis. The dotted line is the confidence interval (at a confidence level of 0.95) for all sites which was computed after calculating the confidence interval for every site and taking the mean of all confidence intervals for every site, by parameter. The graphs illustrate that a significant portion of the ACF for each parameter was within the confidence interval.

Finally, the skewness, kurtosis, and normality of the residuals were computed and are shown in table 4. The Shapiro-Wilk test tests if the data came from a normally distributed population. The p-values for every parameter was extremely small thus giving sufficient statistical evidence to fail to reject the null hypothesis that the data came from a normal population. The data sets are extremely large, so it is no surprise that the Shapiro-Wilk test indicates that none of the parameters have residuals that are normally distributed. Skewness and kurtosis of residuals are generally acceptable; ideally a skewness level between $-1/2$ and $1/2$ is needed for a distribution to be approximately symmetric and a kurtosis level around 3 for a peak that is close to a normal distribution. Kurtosis levels are typically close to 3, except for Conductivity which has an extremely high kurtosis of 20. Skewness was in the ideal range for most parameters except NH3, which suggests that the residuals are highly positively or negatively skewed.

5. Conclusion

Detecting trends in water quality, as the evidence illustrates, was very dependent on the parameter being measured. Temporal influences may have much greater influences on water quality but does not always need to be accounted for. Monthly sampling at each site may explain the lack of temporal influence when regression was performed on the parameters, although results in [3] indicate there is little difference in varying sample

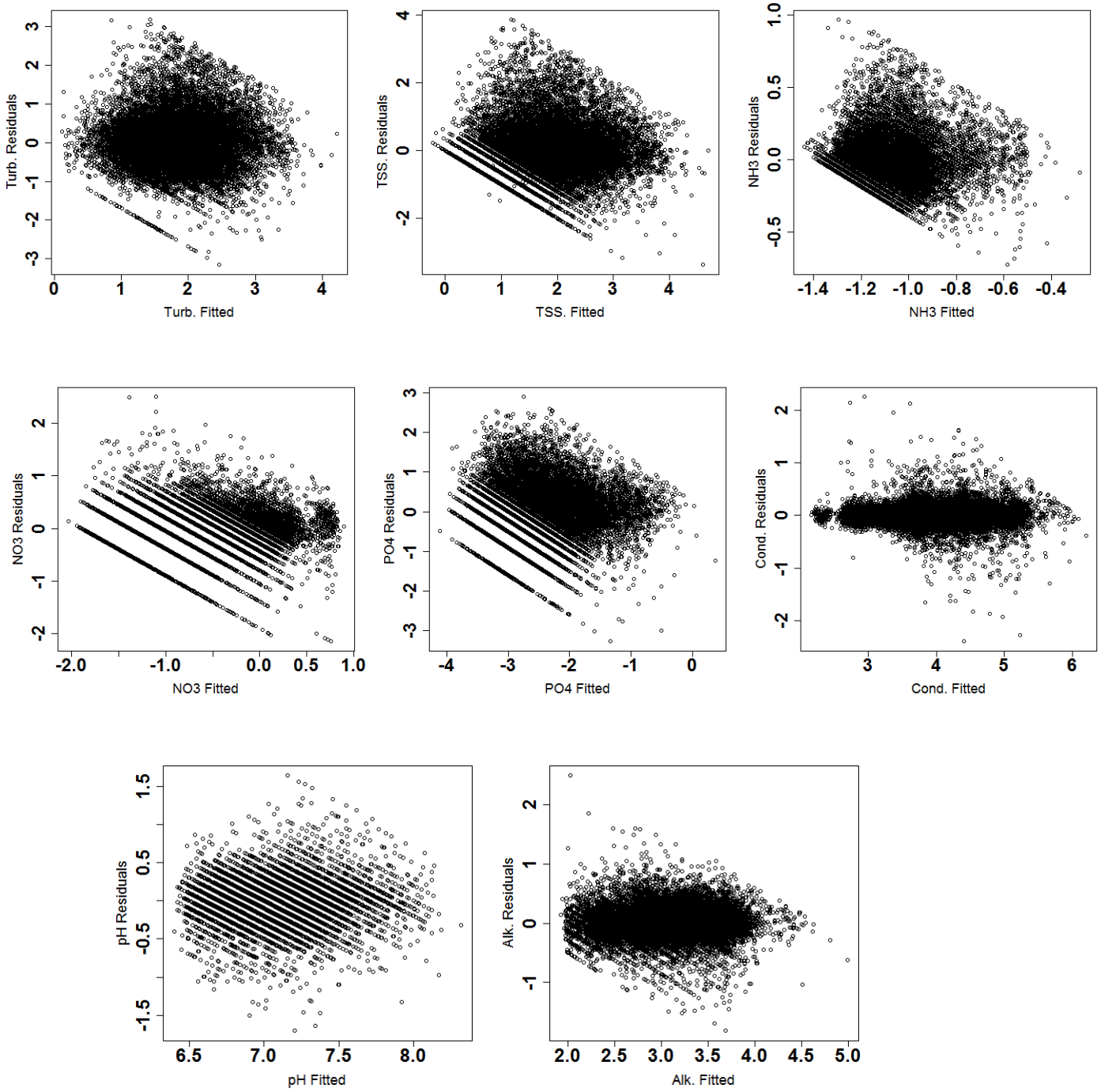

FIGURE 2. residuals for Turb., TSS, NH3, NO3, PO4, Cond., pH, Alk.

strategies. Regardless of the model used, significant trends were detected for all parameters. Evaluating trends site by site was the best approach to improve water quality in order by selecting what areas have significant changes over time or in certain seasons.

Both NO3 and Conductivity are increasing over time for a significant portion of sites, indicating that since samples were first collected in 1990 there is evidence that supports the quality of water is worsening for these

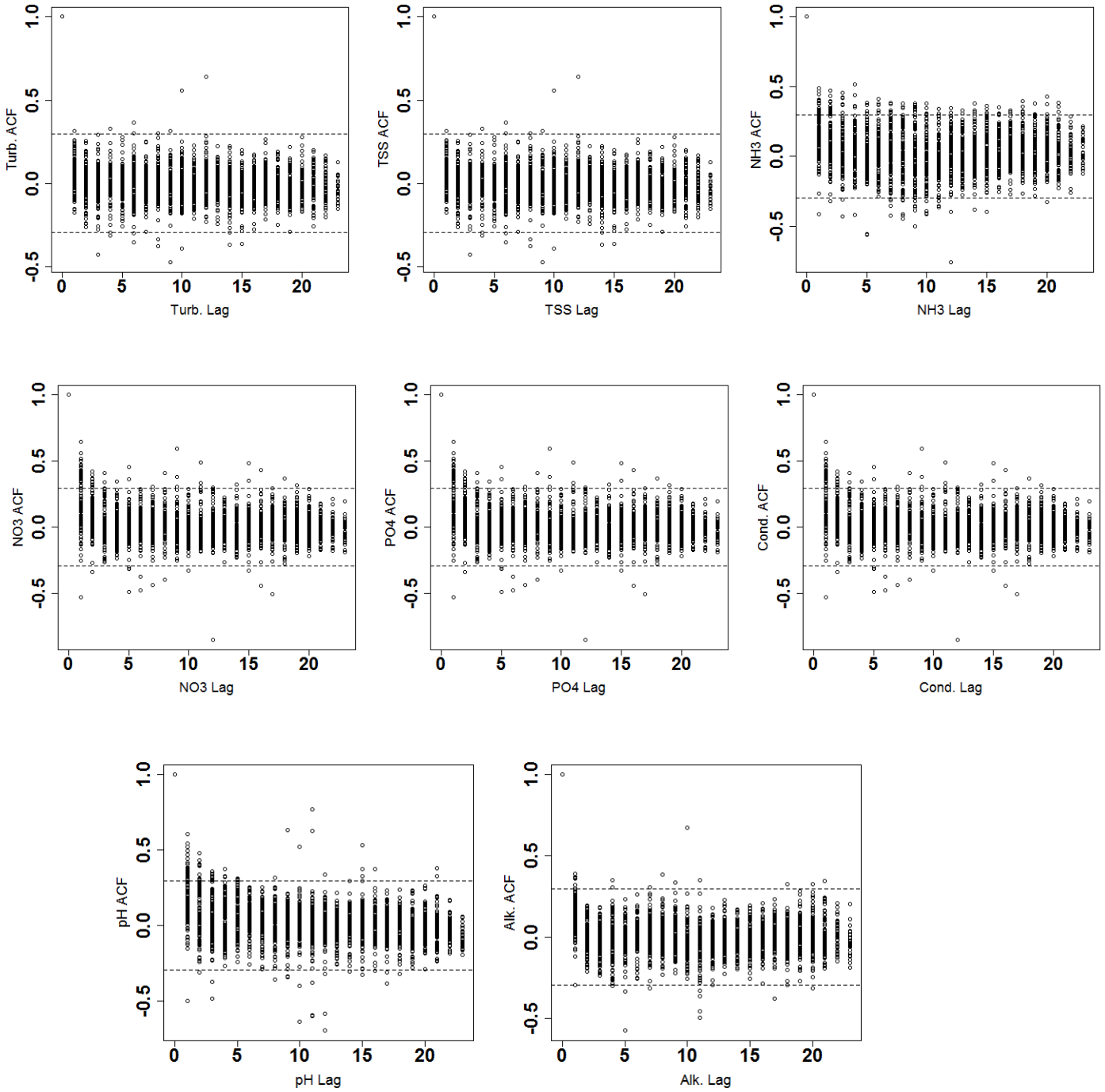

FIGURE 3. ACF for Turb., TSS, NH3, NO3, PO4, Cond., pH, Alk.

sites. Checking the relative geographic areas for sites would be essential in order to investigate if clusters of sites are worsening or if these are isolated VWIN sites. Flow was a strong factor for influencing both positive and negative trends in the water quality parameters analyzed. Results by Interlandi and Crockett show similar strong influences on flow for water quality [3]. Seasonality was also a strong factor that has positive

TABLE 3. Shapiro-Wilk, skewness, and kurtosis

Parameter	Shapiro-Wilk	Skewness	Kurtosis
Turb.	$W = 0.969$, p-val $< 2.2e-16$	0.6079	4.755
TSS	$W = 0.9661$, p-val $< 2.2e-16$	0.776	4.899
NH3	$W = 0.9392$, p-val $< 2.2e-16$	1.067	6.11
NO3	$W = 0.9716$, p-val $< 2.2e-16$	-0.17	6.124
PO4	$W = 0.9939$, p-val $< 2.2e-16$	0.063	3.703
Cond.	$W = 0.8756$, p-val $< 2.2e-16$	0.042	20.73
pH	$W = 0.957$, p-val $< 2.2e-16$	-0.256	7.54
Alk.	$W = 0.9467$, p-val $< 2.2e-16$	0.057	7.122

and negative trends for all eight water quality parameters in the study. Interlandi and Crockett also report similar influences of seasonality on water quality parameters [3].

Further work could be done by applying a stronger transform on variables with a high kurtosis or a negative skewness and high kurtosis (i.e. Conductivity or NO3 and pH). A Box-Cox transformation might significantly change the residuals from the GLS fit. Non-linear terms could be used in the regression that might capture the effect of time more accurately on certain parameters because time had low significant p-values for all parameters. Accounting for spatial correlation between parameters is another method that would improve the ability to detect trends for water quality data. Using geographic information systems would enable trends to be detected that accounts for spatial effects on parameters.

6. Acknowledgements

The author would like to thank Dr. Steve Patch for help throughout the research project and providing an endless amount of insight, assistance, and comedic relief throughout the research process. Ann Marie Traylor, director of EQI, was also instrumental in providing great feedback, support, and help. Staff at EQI, Chloe, Scott, and Hannah, to name a few, devoted time to doing chemical analyses of data collected by VWIN volunteers which was used in the statistical analysis. No statistical analysis can be done without data; a big thank you for all volunteers involved with VWIN.

7. References

1. RiverLink, 2009: FrenchBroadFacts. [http://www.riverlink.org/FrenchBroadFacts.asp].

2. Robinson, R. Bruce, Wood, Molly S., Smoot, James L., and Stephen E. Moore, 2004: Parametric modeling of water quality and sampling strategy in a high-altitude appalachian stream. *Journal of Hydrology*, **287**, 62-73.

3. Sebastian J. Interlandi, and Christopher S. Crockett, 2003: Recent water quality trends in the Schuylkill River, Pennsylvania, USA: a preliminary assessment of the relative influences of climate, river discharge and suburban development. *Water Research*, **37**, 1737-1748.

4. John Fox, 2002: Time-Series Regression and Generalized Least Squares.

5. Image taken from: http://www.bae.ncsu.edu/programs/extension/wqg/frenchbroad/.

6. Patch, Steven C., Westphal, Marilyn J., Pandolfo, Tamara, Fishburn, Jillian, and Elizabeth Wilcox, 2006: Water Quality in the Mountains: Henderson County Volunteer Water Information Network. Year-Thirteen Report. Technical Report 06-159, 55 pages.

7. United States Geological Survey, cited 2013: National Water Information System: Web Interface. [http://waterdata.usgs.gov/nwis/rt]

8. National Climatic Data Center, cited 2013: Global Historical Climatology Network [http://www.ncdc.noaa.gov/oa/climate/ghcn-daily/]