Chapter 4

Predicting In-stream Water Quality from Watershed Characteristics

Richard S. Huebner and Douglas G. Soutter

The Pennsylvania State University at Harrisburg
777 West Harrisburg Pike
Middletown, Pennsylvania 17057-4898

In-stream water quality studies are labour intensive and consume a significant amount of time and money. They are usually conducted on larger streams and rivers or those water courses where severe quality problems have been identified. In addition, current modelling efforts are at such a large scale that they do not provide information necessary for analyzing the contribution of local watersheds to in-stream contaminant loadings. Non-point source (NPS) pollution originates in smaller watersheds along tributaries to larger streams. If NPS pollution problems are to be substantially mitigated, it is critical to identify problem watersheds and the management practices needed to reduce their impact on receiving water quality. Identification and control of NPS pollution must be accomplished in these smaller watersheds without the resource intensive investigations that are currently the state-of-the-art.

This chapter presents a technique applied to watersheds in the Ridge and Valley Physiographic Province of central Pennsylvania. Although the expressions shown may not be directly applicable to watersheds outside this region, the methodology used should be transferable.

Two types of multiple linear regression expressions are shown. The first uses watershed properties, such as area, slope, U. S. Soil Conservation Service (SCS) curve number (CN) (Soil Conservation Service, 1986), hydrologic soil

Huebner, R.S. and D.G. Soutter. 1994. "Predicting In-stream Water Quality from Watershed Characteristics." *Journal of Water Management Modeling* R176-04. doi: 10.14796/JWMM.R176-04.

© CHI 1994 www.chijournal.org ISSN: 2292-6062 (Formerly in Current Practices in Modelling the Management of Stormwater Impacts. ISBN: 1-56670-052-3)

group, time of concentration, and percent of watershed covered by forest, agriculture, or urban area, to estimate the concentrations of water quality measures like pH, alkalinity, conductivity, nitrate-nitrogen, and water temperature. The second uses several of these water quality estimates to predict measures such as the concentration of ammonia-nitrogen and orthophosphate. Some of the expressions, however, represent weak causal relationships, for example, the expression for dissolved oxygen concentrations.

4.1 Introduction

The state of the art in nonpoint source pollution monitoring is to collect stream water samples, analyze them in the field and/or laboratory, and statistically summarize the results. The summarized results are used to determine problem watersheds and to identify water quality control measures to minimize the contamination from these watersheds. Modelling of larger systems like the entire Chesapeake watershed is being undertaken (Blankenship, 1991). This modelling effort provides contaminant loadings at the points where major rivers join the Chesapeake Bay but does not provide the information necessary for analyzing the contribution of local watersheds to those contaminant loadings. This information is critical to identifying problem watersheds and appropriate management practices.

There are several hundred streams that are tributary to the Susquehanna river (Higbie, 1965). Most are in Pennsylvania. Water quality in the Chesapeake Bay is directly related to the contribution of contaminants from the watersheds of these streams. It would be cost prohibitive to sample all of these streams as described above, even on a sporadic basis. An alternative is to develop a model that would predict contaminant concentrations in streams based upon a set of watershed characteristics. Problem watersheds could be identified by using this predictive model and the effectiveness of various management practices assessed.

The water quality study that provided the data base for this chapter was undertaken in order to provide county-wide baseline information on stream quality within Dauphin County, Pennsylvania. The data was collected according to its potential to 1) help determine the effectiveness of erosion and sediment pollution controls, 2) provide the basis for more specific future testing, and 3) show the effects of urbanization. It was also intended to aid in the targeting of Conservation District resources. Forty sites throughout Dauphin County, Pennsylvania were sampled, including a site on the Susquehanna River (for comparison purposes). Sampling and testing were conducted once every two weeks at each site for fifteen months. The study was originally planned to last twelve months from June, 1991 to June, 1992. However, due to the unusually dry summer of 1991, the period of study was extended from one year to fifteen months to include the growing season of 1992.

The following quality parameters were measured at all sites during the period from June, 1991 through September, 1992:

- 1) pH,
- 2) alkalinity,
- 3) turbidity,
- 4) conductivity,
- 5) nitrate-nitrogen,
- 6) ammonia-nitrogen,
- 7) orthophosphate,
- 8) dissolved oxygen,
- 9) water temperature,
- 10) air temperature, and
- 11) depth.

A combination of colorimetric methods and field testing instruments were used to make water quality measurements. Split sample testing by an independent laboratory and blind sample tests were also conducted during this period. For the work described in this chapter, only the first nine water quality measures were examined since air temperature and depth of flow are not water quality measures. Of particular concern in the list of parameters above are the concentrations of nutrients (Hirsch, Alley, and Wilber, 1988), specifically nitrate-nitrogen, ammonia-nitrogen, and orthophosphate. Since the watersheds in the study are tributary to the Susquehanna River and the Chesapeake Bay, these constitute contaminants that pose a significant threat to the bay and the river from the process of eutrophication (Thomann and Mueller, 1987; Tchobanoglous and Schroeder, 1987).

Of the 40 sites for which data was collected, only 34 were used to develop the regression equations. The six sites that were omitted represented very large watersheds and streams. Some of the 34 watersheds and streams used to develop the regression expressions are located in these six larger watersheds.

Multiple linear regression analysis (Poister, 1978; Draper and Smith, 1981) was used to examine causal relationships between watershed properties and the conventional water quality parameters, listed above. The water quality data consisted of averages of 31 observations from the fifteen month sampling period for each of the sites. Some of the averages were based on less than 31 observations due to missing values. The watershed properties that were used in the regression analysis were:

- 1) SCS curve number,
- 2) area,
- 3) stream slope,
- 4) hydrologic soil group,

- 5) time of concentration,
- 6) percentage of watershed covered by forest,
- 7) percentage of watershed used by agriculture, and
- 8) percentage of watershed urbanized or developed.

Watershed area, slope, hydrologic soil group, and ground cover or land use were available from maps (Pennsylvania Topographic and Geologic Survey, 1975) and other references (Soil Conservation Service, 1972). SCS curve number and time of concentration were computed using standard methods (Soil Conservation Service, 1986).

Sabater et al., established the relationship between watershed characteristics and the physio-chemical composition of the Ter river that flows into the Mediterranean sea (Sabater, 1990). The Dauphin County Conservation District project and the data base developed from it offered the opportunity to develop a similar, predictive model of water quality for watersheds that make up the Susquehanna basin.

4.2 Methodology

The data base for the regression analysis constituted a two-dimensional array with 34 rows, one for each sampling site included in the study, and 17 columns for the values of the watershed characteristics, eight columns, and the water quality parameters, nine columns. The watershed characteristics were independent variables and the water quality parameters dependent variables in the regression analysis. Table 4.1 shows the ranges of independent and dependent variables used in the development of the regression expressions.

A regression analysis was conducted for each of the nine water quality measures listed above. The analysis was conducted in three stages using Minitab Statistical Software (Minitab, Inc., 1991) on a personal computer. The first stage consisted of a stepwise linear regression to aid in the selection of the watershed characteristics that best predicted the variation in a water quality parameter. Next, a regression equation was developed using the watershed characteristics chosen by the stepwise regression. Finally, outliers identified after the regression equation was developed were removed from the data set and the regression equation developed again without them. Each regression equation included from three to five watershed characteristics as predictors of water quality. The watershed characteristics included in the equations changed among the different water quality parameters.

There were instances where watershed characteristics did not predict water quality parameters well. In these instances, an intermediate predictor of a specific water quality measure was investigated. If successful, a regression equation for the water quality parameter was developed based upon the value of another water quality parameter.

Table 4.1: Variable Data Ranges**Independent Variables - Watershed Characteristics**

<u>Variable</u>	<u>Range</u>	
	Low	High
SCS Curve Number	64	83
Area (sq. mi.)	1.4	86.7
Stream Slope	0.001	0.05
Hydrologic Soil Group	2	4
Time of Concentration (hr.)	1.2	20.1
Percentage of Watershed Covered by Forest	0	100
Percentage of Watershed used by Agriculture	0	100
Percentage of Watershed Urbanized or Developed	0	100

Dependent Variables - Water Quality Parameters

<u>Variable</u>	<u>Range</u>	
	Low	High
pH	4.7	8.1
Alkalinity (mg/l)	15	201
Turbidity (NTU)	0.69	130
Conductivity (mohms)	29	721
Nitrate-nitrogen (mg/l)	0.08	3.83
Ammonia-nitrogen (mg/l)	0.17	1.24
Orthophosphate (mg/l)	0.02	0.54
Dissolved Oxygen (mg/l)	9.0	12.3
Water Temperature (°C)	7.4	18.3

Finally, there were some water quality parameters for which a statistically significant and practically significant predictor equation could not be developed. These are discussed in the following section.

4.3 Results

The regression equations developed from the Dauphin County data are presented in Table 4.2 with the corresponding value of r^2 and the level of statistical significance of the regression expression, p . All of the expressions were statistically significant at the 1% level or less. The values of r^2 ranged from 0.394 to 0.899. The value of r^2 was adjusted for the number of degrees of freedom in each of the regression equations. Values for pH, alkalinity, conductivity, and nitrate-nitrogen concentration appeared reasonably well represented by the regression expressions developed from the watershed

Table 4.2: Water Quality Regression Equations

<u>pH</u>	
pH = 8.09 - 4.84 Slope - 0.225 Ag - 1.05 For	$r^2 = 0.724$ $p = 0.000$
<u>Alkalinity</u>	
Alk = - 93.3 + 4.33 CN - 39.6 Soil Type - 20.9 Ag - 74.8 For	$r^2 = 0.872$ $p = 0.000$
<u>Turbidity</u>	
Turb = - 4.05 - 25.3 Slope + 2.12 Soil Type + 4.85 Ag	$r^2 = 0.394$ $p = 0.001$
<u>Conductivity</u>	
Cond = - 219 + 14.4 CN - 165 Soil Type + 16.6 Ag - 242 For	$r^2 = 0.857$ $p = 0.000$
<u>Nitrate-nitrogen</u>	
NO ₃ -N = - 0.495 - 0.00687 Area + 0.230 Soil Type + 0.0637 Tc + 1.48 Ag	$r^2 = 0.709$ $p = 0.000$
<u>Ammonia-nitrogen</u>	
NH ₃ -N = 0.154 + 1.83 PO ₄	$r^2 = 0.899$ $p = 0.000$
<u>Orthophosphate</u>	
PO ₄ = 0.0316 + 0.0256 NO ₃ -N	$r^2 = 0.750$ $p = 0.000$
<u>Dissolved Oxygen</u>	
DO = 11.5 - 0.00100 CN + 0.0246 AREA - 0.214 Soil Type - 0.133 Tc + 0.138 Dev	$r^2 = 0.503$ $p = 0.005$
<u>Water Temperature</u>	
Temp = 12.5 - 0.146 Area + 0.725 Tc - 2.56 For	$r^2 = 0.623$ $p = 0.000$

characteristics, (r^2 greater than 70%). A significant amount of the variation in observed values of orthophosphate and ammonia-nitrogen concentrations were represented by the variation in other water quality parameters, specifically nitrate-nitrogen concentration, (r^2 values of 89.9% and 75%). However, there

appeared to be no significant predictor of turbidity (r^2 equal to 0.394) and the regression equations for dissolved oxygen concentration and water temperature were weak, (r^2 equal to 0.503 for dissolved oxygen and 0.623 for water temperature).

The r^2 values indicated reasonably strong causal relationships for most of the quality parameters. All of the equations were statistically significant at the 0.01 level or better. It was necessary, however, to examine the equations in light of the stated objective: to identify streams or stream reaches that exhibited water quality problems. This was done by using the appropriate watershed characteristics in each regression equation to estimate the corresponding value of a water quality parameter for each of the 34 sites. The results are described below.

For instance, the values of pH for each site tended to be over-estimated by the regression expression. All of the estimates, though, were within one pH unit of the observed values except for one location with a pH of 4.73. This location was influenced by acid mine drainage, a condition that would not be accounted for by the watershed characteristics used to develop the regression equation for pH. The equation for alkalinity also tended to over-estimate alkalinity concentrations. In some cases, the value was estimated to be as much as 50 mg/l higher than the observed values.

Conductivity is a measure somewhat related to alkalinity. Again the equation over-estimated observed conductivities but would have correctly identified five of the stream segments with the highest conductivities (those with conductivities greater than 340 mohms).

The three nutrient water quality measures, nitrate-nitrogen, ammonia-nitrogen, and orthophosphate, were important water quality measures with somewhat weak regression relationships. The nitrate-nitrogen equation had the highest r^2 value and was based solely upon watershed characteristics. When the concentrations of nitrate-nitrogen were predicted using the regression equation, relative error at one site exceeded 400%. However, out of eighteen sites with concentrations greater than 1 mg/l, only two would not have been identified using the regression expression. Two other sites with observed concentrations less than 1 mg/l would have mistakenly been identified as watersheds with nitrate-nitrogen concentrations over the threshold value, in this case 1 mg/l.

Orthophosphate concentrations were seriously under-predicted by the regression equation shown in Table 4.2. Relative errors ranged from 0.3% to 224%. The range of errors and the inability of the regression expression to identify problem streams was due to two factors. First, all of the streams sampled had relatively low orthophosphate concentrations (all less than 0.5 mg/l, most below 0.2 mg/l). Also, the regression equation for orthophosphate concentration was the second type of regression equation developed, based upon an intermediate variable, in this case, nitrate-nitrogen concentration. As described above, the estimation of nitrate-nitrogen concentrations exhibited significant error quantitatively. The two factors combined to negate the predictive value of the

regression expression for orthophosphate concentration.

Ammonia-nitrogen concentration was another water quality variable that was best predicted by a regression equation based upon an intermediate value, in this instance, orthophosphate concentration. As such, nitrate-nitrogen concentrations were predicted using watershed properties and the equation in Table 4.2. Orthophosphate concentrations were estimated using the predicted values of nitrate-nitrogen concentrations and finally ammonia-nitrogen concentrations calculated using the estimated concentrations of orthophosphate at each of the thirty-four sites. These were compared to the observed values of ammonia-nitrogen. Relative errors ranged from 0.3% to 78%. However, of the three sites with ammonia-nitrogen concentrations greater than 1.0 mg/l, two would have been identified by the predicted values. The values estimated by the regression equation ranged from 0.221 mg/l to 0.295 mg/l. This did not provide enough resolution to conclude that the regression equation would properly identify stream segments with high ammonia-nitrogen concentrations.

The comparison between predicted and observed values of dissolved oxygen concentrations and water temperature were not made because of the weak associations reflected by the r^2 values of their respective regression equations.

4.4 Conclusions

A technique for developing predictive equations of water quality from watershed characteristics has been presented in this chapter. The technique has several weaknesses. First, this study was based upon fifteen months of data at sites within a single county in the Ridge and Valley Physiographic Province of Pennsylvania. The equations developed are, therefore, limited in application. The water quality averages used were developed from a period that covered several significant seasonal changes. Also, the sample size, although adequate, was limited.

In the development of regression equations, there is a tendency to allow the equations to represent unrealistic relationships that, from a physical standpoint, make no sense. The equations shown in Table 4.2 were examined for this and were found to be conceptually correct. For instance, the equation for pH in Table 4.2 implies an inverse relationship between percentage of watershed covered by forest and pH. Conceptually, this relationship is consistent since one would expect lower pH in streams fed by forested areas due to the presence of humic acids from decaying organic matter.

Most of the causal relationships found in the regression equations in Table 4.2 have relatively strong associations with 70% or more of the variation in the dependent variable being explained by the regression expression. The relationship between orthophosphate and nitrate-nitrogen concentrations was particularly strong. This may, however, be a characteristic that reflects the

properties of local watersheds.

The failure to identify an equation for turbidity was an unexpected result. Conventional wisdom suggests that turbidity is caused by watershed properties like the extent of urbanization and agricultural land uses. However, the relative strength of the other relationships was also unexpected since factors other than watershed properties, such as biological processes, significantly affect water quality.

Additional work is required to develop water quality predictive equations beyond the results presented here. As noted above, seasonal aspects of water quality should be accounted for as well as wet weather and dry weather events. The data used in this study, in addition to representing fifteen continuous months of sampling also included a winter season and two summer seasons, one of which was a drought period for the state.

Multiple linear regression equations were developed from the data, but non-linear relationships may be more appropriate for some of the water quality parameters presented.

Nutrient quality equations need to be improved beyond those presented. Reliably estimating the concentrations of nutrients in smaller watersheds is critical to the control of non-point source pollution. The equations in Table 4.2 represent only a starting point for this work. Either water quality sampling on smaller watersheds or equations of the type presented here will be necessary to identify optimum land management practices in order to mitigate non-point sources of pollution.

Acknowledgements

The authors would like to thank the Dauphin County Conservation District, the Pennsylvania State University, and the Pennsylvania Department of Environmental Resources for their support during the conduct of the research associated with this chapter. The authors also wish to acknowledge the efforts of the Environmental Engineering Technology Class of 1992 at Penn State Harrisburg for their efforts in data collection in conjunction with the Dauphin County study.

References

- Blankenship, K. (1991), "Modelling the Bay's Watershed", Bay Journal, November, 1:8:1-5
- Commonwealth of Pennsylvania (1975), 7.5 Minute Series (Topographic), (various Quadrangle maps), Department of Environmental Resources,

Topographic and Geologic Survey, Harrisburg, PA, and U.S. Geologic Survey, Denver, CO

Draper, N. R. and Smith, H. (1981), *Applied Regression Analysis*, 2nd ed., John Wiley and Sons, Inc., New York

Higbie, H. W. (1965), *Stream Map of Pennsylvania*, The Pennsylvania State University, University Park, PA

Hirsch, R. M., W. M. Alley, and W. G. Wilber (1988), "Concepts for a National Water-Quality Assessment Program", U.S. Geological Survey Circular 1021, Books and Open-File Reports Section, U.S. Geological Survey, Denver, CO

Minitab Inc. (1991), *Minitab Reference Manual*, Release 8 PC Version, State College, PA, November

Poister, T. H. (1978), *Public Program Analysis: Applied Research Methods*, University Park Press, Baltimore, MD

Sabater, F., S. Sabater, and J. Armengol (1990), "Chemical Characteristics of a Mediterranean River as Influenced by Land Uses in the Watershed", *Water Research*, v 24, n 2, February, 24:2:143-155

Soil Conservation Service (1972), *Soil Survey - Dauphin County, Pennsylvania*, U. S. Dept. of Agriculture, Superintendent of Documents, Washington, D.C.

Soil Conservation Service (1986), *Urban Hydrology for Small Watersheds*, 2nd edition, U.S. Department of Agriculture, Superintendent of Documents, Washington, D.C., June

Tchobanoglous, G. and E. D. Schroeder (1987), *Water Quality*, Addison-Wesley Publishing Company, Reading, MA

Thomann, R. V. and J. A. Mueller (1987), *Principles of Surface Water Quality Modelling and Control*, Harper and Row Publishers, New York