STATISTICAL WATER QUALITY MODELLING (WITH SPECIAL EMPHASIS TO PHOSPHORUS MODELLING)

One of the most important aspects of environmental deterioration of inland waters is the progressive enrichment of waters with nutrients resulting in the mass production of algae, and other undesirable biotic changes. Many factors influence the level of plant biomass in an aquatic ecosystem including light, temperature, mixing, grazing, carbon dioxide gas, and nutrients, but, only some are controllable by man as a practical and economical method for decreasing plant productivity.

Although estimates have been made of the variety of nutrients, namely, nitrogen, phosphorus, iron, manganese, molybdenum and other trace elements, that may become limiting to algal communities (Goldman, 1965), it is probably only the geochemically rare (in relation to algal growth and plant growth requirements) macronutrients, namely nitrogen and phosphorus, which control the development of aquatic blooms (Goldman, 1965; Hasler, 1947; Hutchinson, 1957, 1973; Sawyer, 1960). Phosphates are frequently a limiting factor for both aquatic and terrestrial natural ecosystems. The addition of phosphorus to such ecosystems from anthropogenic sources frequently thus leads to increased algal productivity. The fertilization and resulting increased productivity in aquatic ecosystems leads to conditions which decrease the beneficial uses of water. The consequences to water uses include:

1. Aesthetic and recreational interferences-algal mats, decaying algal clumps, odours, and discoloration may occur;

2. Large diurnal variations in dissolved oxygen (DO) can result in low levels of DO at night as a result of algal respiration, which in turn may result in the death of desirable fish species;

3. Phytoplankton and weeds settle to the bottom of the water system and create a sediment oxygen demand (SOD);

4. Large diatoms (phytoplankton that require silica) and filamentous algae may clog water treatment plant filters thereby reducing the time between backwashing;

5. Extensive growth of rooted aquatic macrophytes (larger plant forms) interfere with navigation, aeration, and channel-carrying capacity; and,

6. Toxic algae have been associated with eutrophication in coastal regions and have been implicated in the occurrence of 'red tide' which may result in paralytic shellfish poisoning.

The consequences obviously warrant considerable attention to phosphorus levels and the opportunities for abatement and control.

SOURCES OF PHOSPHORUS AND MODELLING NEEDS

The various uses of phosphorus by humans include as fertilizers, in animal feed lots, detergents, and industrial uses. Phosphorus inputs to surface waters vary with discharge regulations and geography

for the diffuse sources. On the basis of the present technical literature the natural and cultural activities which introduce phosphorus to surface waters may be classified into 1) nonbasin activities - transportation by precipitation; 2) agriculture - use of fertilizers and crop harvesting procedures; 3) domestic - human wastes and washing products; 4) industrial - metal finishing and detergent industries; 6) mining - mining and refining of phosphorus; and 7) animal waste - use of animal manure as fertilizers.

Environmental management of particular resources requires the system within which a resource is distributed, be well understood. Possible suggestions toward the elimination of some of the phosphours inputs to surface waters include phosphorus removal from detergent formulations, removal of phosphorus of domestic wastes, and the application of the concept of zero discharge from point sources. These manipulations of phosphorus input are considered rather simplistic solutions of complex problems and immediately the question arises, whether these manipulations will produce improvements in the conditions, and second, whether in fact these solutions may be counter-productive in terms of economy of the region.

To identify potential control tactics for a particular surface water body, the behaviour of phosphorus concentrations must be understood. Investigators have utilized empirical and theoretical relationships between the phosphorus inputs from point and nonpoint sources and the typical phosphorus concentrations in stream (Eisenreich et al. 1977; Gakstatter et al. 1978; Hydroscience, 1977a, 1977b; Rast and Lee, 1978; Thomann, 1972; IJC, 1978). Omernik (1977) used a simple regression model to relate runoff concentrations to the percent of land in agriculture use and the percent of land in urban use. However, phosphorus concentrations in surface water cannot be predicted merely from pollutant source loadings and hydrographic conditions. Major portions of the phosphorus are regenerated due to limnological transformations (Stumm et al., 1972). The release of phosphorus from sediments may also be important, where release depends upon mixing and other physical conditions at the observation site/reach. Of interest herein is the degree to which such phenomena can be studied in terms of other parameters e.g., suspended solids, turbidity, etc. It is evident from the above discussion that the ultimate fate processes of phosphorus are complex. The development of a mechanistic model is certainly challenging given the complexity of the various pathways of phosphorus; therefore, phosphorus modelling efforts described in this paper are directed towards development of statistical models specific to a particular seasons and river stretch using measured data. Specifically, statistical relationships are developed between total phosphorus and commonly monitored water quality parameters. Two regression methodologies, namely best subset regression and stepwise regression have been used for model discrimination. The established relationships are useful in the infilling of missing data and also, in basin management activities to overcome the eutrophication problem. The regression models for various seasons have been developed for the Grand River basin.

STUDY BASIN

Fig. 1 presents a location map of the Grand River with pertinent sampling locations identified. The Ontario Ministry of Environment has accumulated a lengthy historical water quality data (WQD) (WQD, 1972-1979). The data from February, 1975 to March, 1977 are available on a weekly basis and the rest of the data are available on monthly basis. The details of sampling locations are available PLUARG Study (1979).

The Grand River, located in southwestern Ontario, originates near the village of Dundalk and picks up its major tributaries, the Conestogo, Nith, and Speed rivers as it winds its way over 300 kilometres

southeast to Lake Erie collecting water from a drainage area of 6965 square kilometres. The average annual flow of the Grand River is 55 m^3/s at the mouth. The flow may range from a maximum of 1800 m^3/s in the spring to a minimum of 6 m^3/s in the winter (GRBWMS, 1982). The Grand, and its tributary rivers the Nith, Speed, and the Conestogo, flow through one of the most important socio-economic regions in Ontario.

Land use within the basin is varied, with agricultural and rural land uses dominant in the northern and southern portions and urban land uses concentrated in the central portion. Agricultural and urban and land uses respectively comprise 78 percent and 3 percent of the basin area. Wooded and/or idle areas account for approximately 19 percent of the basin area, while less than 1 percent lies in the other uses.

THE PROBLEM OF AQUATIC PLANTS IN THE STUDY AREA

The most serious water quality impairment problems are found in the middle portion of the basin (GRBWMS, 1982). Excessive plant growths were observed particularly in the Speed River and in the Grand River between Kitchener and Paris (GRBWMS, 1982).

Water quality in the upper Grand and Conestogo Rivers was also a matter of concern with respect to phosphorus concentrations (GRBWMS, 1982). Since 1970, improvements in phosphorus levels have been demonstrated but were still sufficiently high for nuisance algae growths. In 1978, large growths of algae were observed at Grand Valley and very large growths were reported downstream between Paris and Brantford (GRBWMS, 1982).

Although all the major municipalities had secondary treatment facilities installed between 1970 and 1975 under the phosphorus removal program of the Ontario Ministry of the Environment (GRCA 1979) removing over 80% of the phosphorus and over 90% of the BOD_5 and suspended solids in the incoming raw sewage since 1976, visual observation of the river below these points indicated turbidity and excessive algae growths of an unappealing nature.

The available data included streamflows (Q) (cfs), suspended solids (SS) (mg/l), total nitrogen (TN) (mg/l), specific conductivity (CON) (micromhos/cm, 25°C), turbidity (TUR) (formazin turbidity units), total coliform (T-coli) (MPN), filtered chloride (Cl) (mg/l), and total phosphorus (TP) (mg/l).

MODEL BUILDING

The general representation of statistical models is given by

$$Y_i = \sum_{j=0}^{k} \beta_j x_{ij} + e \qquad \ldots . (1)$$

with $x_{j0} \equiv 1$

where, x_{ij} are the independent variables for the ith observation (various water quality constituents in present study), Y_i is the dependent variable for the ith observation (total phosphorus concentration), β_j are the unknown coefficients to be estimated, $k+1$ are the number of coefficients (to be estimated)

in the model and e is the error in the determination of Y_i which is generally assumed as having zero mean and constant standard deviation (σ).

The method of ordinary least squares is the most widely used method for assigning β_j because of its simple concept and no assumption is necessary on the probability distribution of data. This method will be used for estimating the coefficients associated with the various water quality constituents for prediction of phosphorus concentrations. The detailed description of the least squares method is available in various textbooks on statistics (e.g., Draper and Smith, 1981; Weisberg, 1980).

Development of Regression Models for Total Phosphorus

Regression analyses were performed in each case on full data sets and their seasonal subsets. Each data set is segmented into four seasonal subsets as listed below.

1.	Spring season	:	March 21 to June 20.
2.	Summer season	:	June 21 to September 20.
3.	Fall/Autumn season	:	September 21 to December 20.
4.	Winter season	:	December 21 to March 20.

Preliminary Analysis

Prior to a statistical regression analysis of a data set, an initial filtering of the data which consisted of a statistical analysis, a preliminary regression analysis, partial visual inspection of the data files, and the creation of numerous scatter plots revealed obvious data input errors. Once the identified input errors were removed, a general regression analysis assuming all water quality parameters as independent variables and total phosphorus as dependent variables, was made to identify any outliers on the basis of leverage value, and studentized residual statistics (Wilkinson 1990).

Using the filtered data, a correlation (of each water quality constituent with total phosphorus) matrix is obtained considering two sets of water quality parameters for overall data and various seasons at all locations. The first set (I) includes all the water quality variables whereas in the second set (II) T-coli and TUR were excluded. The reason for considering two sets is that T-coli and TUR had a large number of missing values in the data sets and hence excluding them increased the number of data points available for regression analyses.

To enhance the visualization of the correlation matrix, Table 1 presents the square of correlation coefficient to indicate the contribution of individual water quality parameters in explaining the variation in the dependent variable for the fall season as an example. Since T-coli had no significant correlation (Table 1) with total phosphorus, this water quality parameter was not considered any further for model formulation.

Selection of Independent Variables in Regression Analysis

To make the model useful for predictive purposes, one wants to include as many independent variables as possible so that reliable fitted values can be determined. Furthermore, since R^2 gives the proportion of the variation in the dependent variables that is explained by the fitted regression model, one obviously desires R^2 to be large. On the other hand, because of the effort involved in

the monitoring of a large number of independent variables, there is interest in including as few independent variables as possible. The compromise between these extremes is what is usually called selecting the best regression variables and consequently the best model. There is no unique statistical procedure for doing this (Draper and Smith, 1981). However, there are many statistical procedures such as all possible regression, backward elimination, forward elimination in stepwise regression, ridge regression, principal component regression, and stagewise regression which may help in optimum model formulation (Draper and Smith, 1981; Montgomery and Peck, 1982, and Weisberg, 1980).

In the present study, the two procedures namely the best subset regression, and stepwise regression procedures are used to select the best set of independent variables.

Best Subset Regression

Using the R^2 information (e.g., Table 1), various best subsets of independent variables can be selected on the basis of proportion of variation explained in the dependent variable. For each subset the regression was assessed according to 1) the value of R^2 achieved, 2) the F value (defined in Equation 3) and 3) the number of observations used in developing the model. The model obtained from the larger data set and achieving higher values of R^2 and F value will always be preferred. The above two criteria (R^2 and F-values) which will be used in model selection are briefly described below.

R^2 Criterion

R^2 value is used as a criterion for comparing models. A computing formula for R^2 is

$$R_2 = 1 - \frac{SSE}{SS_y} = \frac{SSR}{SS_y} \qquad (2)$$

with $SS_y = \Sigma (Y_i - \bar{Y})^2$; $SSE = \Sigma (Y_i - \hat{Y}_i)^2$; $SSR = \Sigma (\hat{Y}_i - \bar{Y})^2$,

where, $\bar{Y}$ is the average value of dependent variable and $\hat{Y}_i$ are the model-computed values of the dependent variables.

A strong linear association between Y_i and $\hat{Y}_i$ yields a large value of R^2 and vice versa. Unfortunately, R^2 provides an inadequate criterion for subset model selection since, whenever comparing a subset model to a large model including the subset, the larger model will always have an R^2 value as large, or larger, than R^2 for the subset model. Thus, the full model will always have the largest possible value of R^2. However, for a fixed number of independent variables (equal to k), R^2 can be used to compare different models with a large value of R^2 indicating the preferred model.

F value Criterion

The F value is mathematically described as (Draper and Smith 1981):

$$F = \frac{N-k-1}{k} \frac{R^2}{1-R^2} \quad (3)$$

From the above expression, it is clear that apart from the constant multiple [(N-k-1)/k], the F-statistic is the ratio of the explained to the unexplained variation in Y_i. Therefore, it is natural to say that the regression is significant only when the proportion of explained variation is large. This occurs when the F-value is large.

The F-statistic can also be used to compare any two models as long as all the independent variables in the smaller model are also included in the larger model, i.e., the small model is a subset model of the larger model.

As defined earlier, the residual sum of squares reflects the variation in the dependent variable that is not explained by the model. If the predictor variables which are not included in the subset model are important, then deleting them from the subset model should result in a significant increase in unexplained variation of Y_i. That is, SSE_r should become considerably larger than SSE_f. A convenient test-statistic (Weisberg, 1980) using this idea is:

$$F_{k-m,N-k-1} = \frac{(SSE_r - SSF_f)/(k-m)}{SSE_f/(N-k-1)} \quad (4)$$

where, SSE_f (defined in Equation 2) and SSE_r are the residual error sum of squares of the full model (containing k independent variables) and the subset model containing k-m independent variables, (where m is the number of independent variables dropped from the full model) respectively. The larger model will be preferred when the $F_{k-m,N-k-1}$ statistic is sufficiently large. One reasonable rule would be to prefer the full model if $F_{k-m,N-k-1} > F^*$, where F* is the α x 100% point of the $F_{(k-m,N-k-1)}$ distribution. The choice of $\alpha = 0.05$ is typical (Weisberg, 1980).

Stepwise Regression

The stepwise procedure provides a systematic technique for arriving at a satisfactory regression equation with a smallest subset of independent variables (Draper and Smith, 1981). In this method, each time a new independent variable is entered into the model, all the variables in the previous model are checked for their continued importance. The main advantages of the stepwise procedure is that the procedure is fast, easy to compute, relatively inexpensive, and available on virtually all computer software. Unfortunately, there are important drawbacks to the use of stepwise procedures. Firstly, the model chosen by stepwise regression need not be the best for any criterion of interest and there is no guarantee that the model chosen will in fact include any of the variables that would be the best subset. The stepwise method is best when the independent variables are nearly uncorrelated, the condition under which finding a subset model is least likely to be relevant. It is true that the best single variable is entered as the first in a stepwise algorithm; however, there is no guarantee that the

best pair is entered as the first pair of variables (Weisberg, 1980). The ordering of the variables given by stepwise regression is an artifact of the algorithm used and need not reflect relationships of substantive interest.

As pointed out above, the stepwise regression procedure is best suited when independent variables are almost uncorrelated, the condition which is rarely met in the case of water quality parameters. As an example, it may be noted from Table 2 that water quality parameters SS-TUR, TN-TUR, CON-Cl and SS-TN are highly correlated (correlation coefficient in the range of 0.714-0.879) at St 56 for the fall season. Of interest is how to select independent variables to build a model to reflect relationship of substantive interest.

To demonstrate the methods of selection of independent variables as described above (e.g., best subset procedure involving R^2/F-value and stepwise regression), an example is given below using the data-set of fall season at St56.

SELECTION OF INDEPENDENT VARIABLES FOR FALL-SEASON DATA SET FOR ST56

Best Subset Procedure

It is clear from Table 1 that at station St56 for the fall season, TN is the best single variable explaining more than 90% variation in the TP (total phosphorus) levels. The other water quality parameters namely SS, TUR, CON, Cl, and Q if taken alone as independent variable explain approximately 40%, 30%, 8%, 8%, and 0.3% variation the TP levels respectively. Now, to increase the R^2, the various pairs of water quality parameters with TN are attempted, some of them showing similar R^2 are summarized in Table 3. It may be noted that the R^2 values in Table 2 are less than that obtained in Table 1 considering only TN; it is because the number of data points available for regression are higher for the subset models as indicated in Table 3.

In Table 3, the R^2 values are more or less the same but there is a large variation between the F-values and hence the independent variables pair having largest F-value will be the obvious choice. Here, TN and TUR are the preferred independent variables having the largest F-values (150.18). From Table 3 it is also clear that the SSR is the maximum and SSE is the minimum for the preferred variable subset which is the basic objective of the regression modelling.

To further increase the R^2 value, now various combinations of three water quality parameters are attempted. The summary of some of the all attempted combinations is presented in Table 4.

Again, the combination of independent variables having largest R^2-value and F-value will be the obvious choice as the number of variables are fixed. Here, the combination consisting of TN, SS, and CON is the preferred subset of independent variables as it has the largest R^2 and F-values. It is to be noted that TUR is not included in the selected subset of three variables while it was included in the selected subset of two independent variables.

To further increase the proportion of explained variation in the total phosphorus concentrations, the various combinations consisting of four independent variables are attempted. Some of the attempted subsets are summarized in Table 5.

From Table 5, the subset (TN+SS+CON+Cl) has the largest R^2-value but it is selected using 24 observations while the other combination (TN+SS+CON+Q) is selected using 30 observations and

have the equivalent R^2 and F-value and hence would be the preferred choice. However, both the subsets may be selected and left for further filtering, as will become clear in the succeeding analyses.

By examining the Table 4 and 5, there is no significant increase in the R^2 - value by adding the new variables in the regression set, as becomes further clear from Table 6. On the basis of the above analyses (Table 1, 3 through 6), the selected best subsets models are summarized in Table 7.

An examination of Table 7 raises the question that which set/subset is the best model. As quoted earlier that when comparing a subset model to a larger model including the subset, the larger model will always have a larger value of R-square than the subset model for the same N) and hence the R^2 and individual F value criteria are not adequate for subset model selection. In such circumstances the $F_{k-m,N-k-1}$ statistics as given by (4) can provide the answer for model selection. The statistic considers whether the reduction caused in the explained variation in the dependent variable by using the subset model (as compared to the full model) is statistically significant. If the reduction in explanation is not significant, one should choose a subset model as opposed to a full model. This examination is carried out and explained in Table 8.

Before a final model is recommended from Table 8, the following two points may recognized:

1. All the best models selected for the filtering from Tables 1, 3 through 6 are the best possible model in the specified category of fixed independent variables. For example, in Table 4, the model TN+SS+CON is the best model for any combination of three independent variables.

2. All the best selected models (Table 8) are the subset of a full model containing all the possible independent variables.

The recommended final model would be that model which will contain just the sufficient independent variables so that there is no significant (in a statistical sense) drop in percent explanation of the variation in the dependent variable as compared to the full model containing all possible independent variables. Table 8 examines this criteria and the model TN+SS+CON can be recommended as the best subset model.

The other procedure to select the model as described earlier is the stepwise regression. Finally, there are two models for a particular data set selected by the two different procedures. These two models may or may not be identical. If they are different then the selection will be made on the basis of R^2, F-value, and number of independent variables. These two models (one from best subset procedure and the other from stepwise regression) for St56 are compared in Table 9.

Observing the selected models by two different approaches (Table 9), it is noted that both models have equivalent R-square values. However, there is a significant difference between their F-values. Further, the best subset model uses more observations and hence is more representative of the data. Observing the t-values of the individual coefficients one may conclude that conductivity (CON), selected as third variable by the best subset procedure, is a better explanatory variable rather than chloride (Cl), selected by stepwise procedure as the third explanatory variable. Considering all these aspects one may conclude that in the present situation the best subset model is better than the stepwise model.

Final Model Selection for Other Locations and Seasons

It is to be noted that the above discussions on the model development and selection of final model using two different methods were directed to the sampling station St56 for the fall season. The same approach was applied to all the sampling locations and seasons considered with the results summarized in Tables 10 through 14.

Model Performance and Discussions of Results

It may be mentioned here that out of 30 comparisons between models suggested by stepwise regression and the best subset procedures, on 27 occasions, the best subset procedure provided a better model. It may be seen that the statistical models developed in this research perform very well in computing the total phosphorus levels at various locations and seasons (R^2 in the range of 0.675-0.990) except at location St75 for summer season (Table 13). The high F-values for all the regressions indicate statistically significant regressions. Figure 2 presents the comparison of observed and model-computed TP levels for the overall data set (data of all seasons combined) at St56 for more than 150 data points spaced over 8 years (1972-79) which suggests good agreement in observed and model-computed TP levels. Plots (Figures 3 through 6) developed for fall, summer, winter and spring seasons for location St56 also indicate a good fit between the observed and model-computed total phosphorous levels.

The model-computed and observed TP levels were also compared for annual and seasonal models at all other locations of the basin (not shown here) and a good agreement was found in computed and observed TP levels.

In Table 10 through 14, the first few parameters in each model are, by far, of greatest significance. The water quality parameters such as suspended solids (SS), turbidity (TUR), and total nitrogen (TN) play major roles in the prediction of total phosphorus (TP) levels.

In some cases SS levels alone explain more than 90% of the variation in TP levels. Similarly, turbidity and total nitrogen are also found to explain significant portions of phosphorus concentration level variation, if taken alone as the independent variable. The other water quality parameters such as conductivity (CON) and chloride (Cl) play a minor role in the prediction of total phosphorus levels as their addition as independent variables in the model, the R-square value in most of the cases improves marginally.

The eutrophication problem might be tackled in accordance with the location needs and the regression models could be used to provide the information regarding the sources of phosphorus whether it is surface water and/or ground water, and location characteristics such as mixing, etc. In 1976 and onwards, most of the domestic wastewater was being treated by secondary treatment, removing over 80% of the phosphorus and over 90% of the BOD_5 and suspended solids in the basin. The phosphorus levels were reduced to a significant extent, but still above the critical level 0.100 mg/L (GRCA, 1979). As phosphates are tenaciously adsorbed by the soil colloids and move from farmlands into streams through erosion of top soil particles on which it is adsorbed, this study strongly suggests that good soil conservation practices which prevent erosion might be the most effective means of controlling the eutrophication problems in the Grand River basin.

CONCLUSIONS

Useful regression models for predicting phosphorus concentrations from other constituents were developed for selected locations for both annual and seasonal concentrations in the Grand River basin. As most of the regression models are successful in explaining more than 90% of the variation in the total phosphorus levels, the developed models may be used for the prediction of missing observed values. However, the variability of the results from one location to another indicates that a general model was not obtained to predict the total phosphorus concentration levels at any location, given levels at another location. Furthermore, the independent variables for total phosphorus prediction change seasonally; this finding is consistent with the knowledge that the major portions of phosphorus are influenced by the prevailing migration pathways at the time and phosphorus portions are regenerated due to the limnological transformations which depend upon mixing and other physical conditions at the observation location.

The study findings strongly suggest that suspended solids play an important role in prediction of phosphorus and consequently, control problems associated with the growth of aquatic plants in the basin. The strong relationship of phosphorus with suspended solids and turbidity indicate the source of phosphorus from surface water runoff while total nitrogen indicates the source from ground water.

About the regression modelling, it is noted that when the data set contains a number of missing values and independent variables are strongly correlated, it is not necessary that the model selected by stepwise regression procedure will be the best model. This fact has been highlighted in the study and it is found that the best subset procedure as described in the text evolves a better model.

** **** **

Table 1. R-square of water quality parameters with TP for fall-season

Site	Set	Number of observations	Q	SS	TN	CON	TUR	Cl	T-coli
St37	I	33	0.029	0.284	.0001	0.343	0.633	0.026	0.043
	II	42	0.019	0.394	.0008	0.314	------	0.016	------
St56	I	19	0.010	0.401	0.927	0.083	0.310	0.085	0.047
	II	24	0.003	0.410	0.917	0.070	------	0.083	------
St80	I	13	0.558	0.561	0.528	0.192	0.698	0.001	0.005
	II	39	0.720	0.218	0.188	0.560	------	0.036	-------
St78	I	17	.0008	0.720	0.770	0.020	0.786	0.070	0.134
	II	26	0.027	0.450	0.746	0.010	------	0.051	-------
St75	I	40	0.042	0.750	0.180	0.001	0.675	.0001	0.001
	II	43	0.050	0.750	0.180	0.003	------	.0001	------
St76	I	21	0.001	0.011	0.040	0.001	0.118	.0007	0.044
	II	34	0.012	.0007	0.019	0.007	------	.0021	-------

Table 2. Correlation coefficient between water quality parameters at St56 for fall-season

	Q	SS	TN	CON	TUR	Cl
Q	1.000					
SS	0.187	1.000				
TN	-0.064	0.754	1.00			
CON	-0.553	-0.127	0.312	1.000		
TUR	0.028	0.879	0.714	0.175	1.000	
Cl	0.329	0.036	0.335	0.848	0.364	1.000

Table 3. Model statistics with various pairs as independent variables (St 56 fall season)

Subset variables	N	R^2	SSR	SSE	F-value
TN+TUR*	31	0.915	1.449	0.135	150.18
TN+SS	31	0.910	1.442	0.142	142.21
TN+CON	30	0.910	1.438	0.143	135.98
TN+Cl	25	0.908	1.381	0.141	108.12
TN+Q	31	0.903	1.431	0.154	129.66

*Best subset model for further filtering

Table 4: Model statistics with various combinations of three independent variables (St56, fall-season)

Subset variables	N	R^2	SSR	SSE	F-value
TN+SS+Q	31	0.912	1.445	0.140	92.87
TN+SS+TUR	31	0.916	1.452	0.133	97.93
TN+SS+CON*	30	0.941	1.487	0.094	137.05
TN+SS+Cl	25	0.931	1.417	0.105	94.92
TN+TUR+CON	30	0.929	1.469	0.112	113.63
TN+TUR+Cl	25	0.925	1.408	0.114	86.72
TN+TUR+Q	31	0.916	1.452	0.113	98.21

*Best subset model for further filtering

Table 5. Model statistics with various combinations of four independent variables (St56, Fall-season)

Subset of variables	N	R^2	SSR	SSE	F-value
TN+SS+CON+Cl*	24	0.949	1.438	0.077	88.94
TN+SS+CON+TUR	30	0.941	1.488	0.093	99.45
TN+SS+CON+Q*	30	0.945	1.493	0.087	107.02
TN+SS+TUR+Cl	25	0.932	1.418	0.103	68.56
TN+SS+TUR+Q	31	0.917	1.453	0.132	71.63

*Best subset model for further filtering

Table 6: Model statistics of other possible combinations independent variables (St56, Fall-season)

Set of independent variables	N	R^2	SSR	SSE	F-value
TN+SS+CON+TUR+Q	30	0.945	1.493	0.0872	82.19
TN+SS+CON+TUR+Cl*	24	0.950	1.438	0.0763	67.83
TN+SS+CON+TUR+Cl+Q*	24	0.955	1.446	0.0682	60.09

*Best subset model for further filtering

Table 7. Selected sets/subsets, candidate for possible model independent variables (St56, fall-season)

Set of independent variables	N	R^2	SSR	SSE	F-value
TN+SS+CON+TUR+Cl+Q	24	0.955	1.446	0.0682	60.09
TN+SS+CON+TUR+Cl	24	0.950	1.438	0.0763	67.83
TN+SS+CON+Q	30	0.945	1.493	0.0870	107.02
TN+SS+CON+Cl	24	0.949	1.438	0.0770	88.94
TN+SS+CON	30	0.941	1.487	0.0940	137.05
TN+TUR	31	0.915	1.449	0.1350	150.18
TN	31	0.899	1.424	0.1608	256.73

Table 8: Selection of model variables on the basis of F-statistic (St56, fall-season)

Full model with k parameters			Reduced model with (k-m) coefficients		k-m	N-k-1	$F_{k-m,N-k-1}$	F^* (α =0.05)	Preferred model
Model	N	SSE	Model	SSE					
TN+SS+CON+TUR+Cl+Q	24	0.0682	TN+SS+CON+TUR +Cl	0.0763	5	17	0.404	2.81	Reduced
TN+SS+CON+TUR+Cl+Q	24	0.0682	TN+SS+CON+Q	0.0870	4	17	1.170	2.96	Reduced
TN+SS+CON+TUR+Cl	24	0.0763	TN+SS+CON+Cl	0.0770	4	18	0.040	2.93	Reduced
TN+SS+CON+Q	30	0.0870	TN+SS+CON	0.0940	3	25	0.067	2.99	Reduced
TN+SS+CON+Cl	24	0.0770	TN+SS+CON	0.0940	3	19	1.390	3.13	Reduced
TN+SS+CON*	30	0.0940	TN+TUR	0.0135	2	26	5.670	3.37	Full
TN+SS+CON	30	0.0940	TN	0.1608	1	26	18.47	4.23	Full

*Best subset model for further filtering

Table 9. Comparison of best subset and Stepwise methods for St56 fall season

Method of variable selection	N	Selected variables	R^2	F-value
Best subset*	30	SS+TN+CON	0.941	137.05
Stepwise	25	SS+TN+Cl	0.931	94.92

*Best subset model

Table 10. Final models for overall data (all seasons)

Site	Model	R^2
St37	TP= 0.04742+0.00074 SS+0.00146 TUR+0.00001 Q-0.00007 CON	0.809
St56	TP= 0.01489+0.00097 SS+ 0.00363 TUR	0.753
St75	TP= 0.05087+0.00087 SS+0.01606 TN+0.00117 TUR-0.00012 CON	0.939
St76	TP= 0.17345+0.00034 SS+0.00662 TUR-0.00003 Q-0.0019 Cl	0.812
St78	TP=-0.01358+0.00134 SS+0.02488 TN+0.00001 Q	0.746
St80	TP=-0.04087+0.00024 SS+0.00012 Q +0.0027 Cl	0.897

Table 11. Final models for fall-season

Site	Model	R^2
St37	TP= 0.04843+0.00272 TUR-0.00266 Cl	0.778
St56	TP= 0.07905-0.00089 SS+0.25490 TN-0.00072 CON	0.941
St75	TP= 0.01723+0.00079 SS+0.00939 TN+0.00083 TUR	0.861
St76	TP= 0.03127+0.00199 SS	0.923
St78	TP=-0.03768+0.00599 SS	0.842
St80	TP=-0.01543+0.00008 SS+0.0089 TN+0.00009 Q	0.833

Table 12. Final models for spring-season

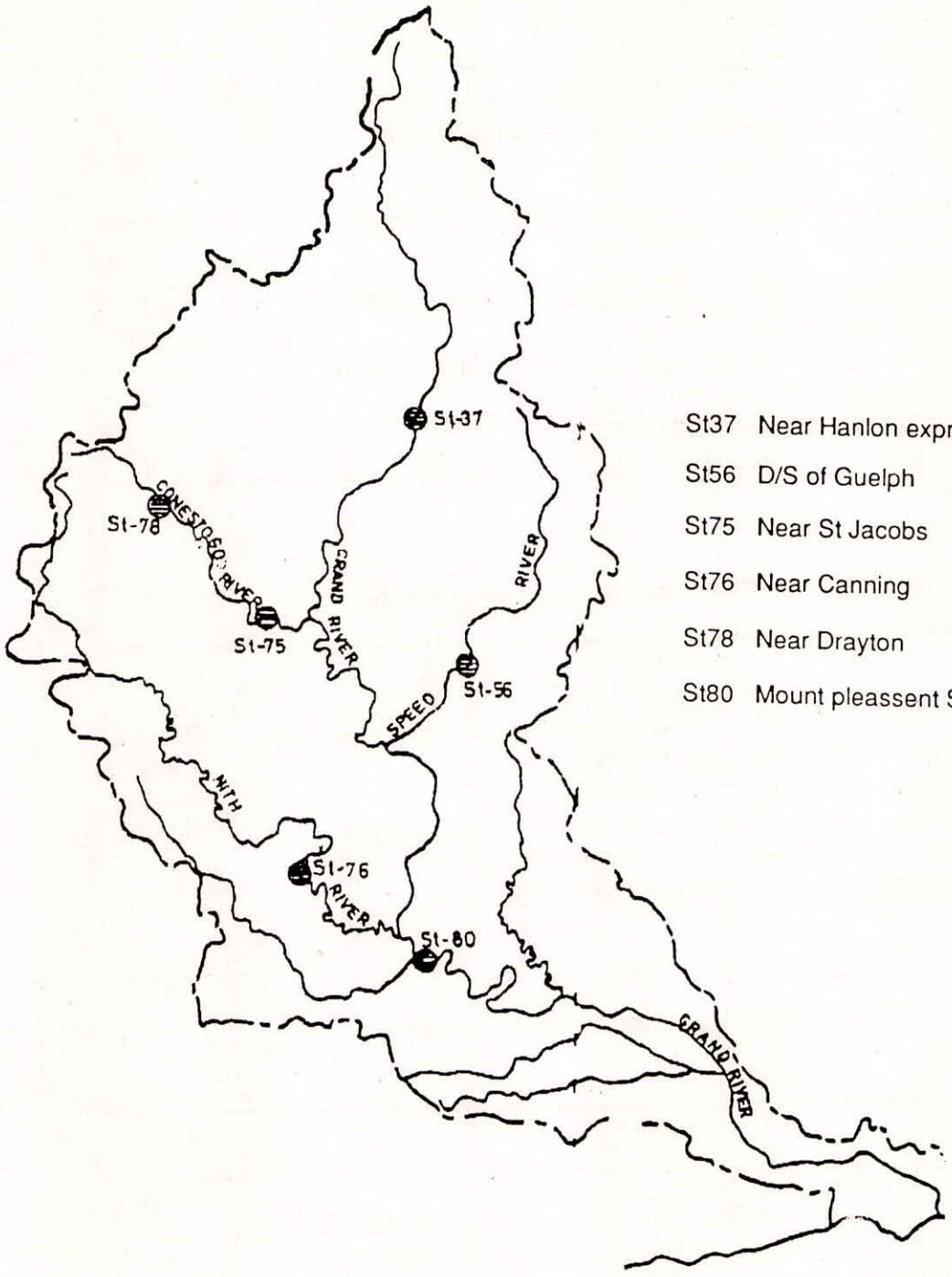
Site	Model	R^2
St37	TP= 0.03331+0.00087 SS+0.01714 TN+0.00159 TUR-0.00011 CON	0.907
St56	TP= 0.027+0.002 SS	0.978
St75	TP= 0.06373+0.00069 SS+0.01562 TN+0.00131TUR-0.00016 CON	0.983
St76	TP= 0.03107+0.00057 SS+0.00001 Q	0.990
St78	TP=-0.07397+0.00132 SS+0.04970 TN	0.797
St80	TP= 0.01119+0.00020 SS+0.00014 Q	0.954

Table 13: Final models for summer-season

Site	Model	R^2
St37	TP= 0.07098+0.00104 SS-0.01191 TN-0.00011 CON+0.00129 TUR	0.704
St56	TP=-0.20760+0.22501TN-0.01819 TUR	0.874
St75	TP= 0.02345+0.00101 SS+0.00104 TUR	0.454
St76	TP= 0.35043+0.00135 SS+0.01002 TUR -0.00016 Q-0.00558 Cl	0.698
St78	TP= 0.00251+0.00053 SS+0.00315 TUR+0.00004 Q	0.675
St80	TP=-0.10517+0.00024 SS+0.00016 Q+0.00678 Cl	0.883

Table 14: Final models for winter-season

Site	Model	R^2
St37	TP= 0.13070+0.00163 SS-0.00024 CON	0.955
St56	TP=-0.62471+0.00048 SS+0.29304 TN-0.00668 TUR+0.00011 Q	0.816
St75	TP= 0.12131+0.00107 SS+0.01837 TN+0.00040 TUR-0.00022 CON	0.985
St76	TP= 0.10409+0.00132 SS	0.934
St78	TP=-0.14668+0.00114 SS+0.06445 TN	0.786
St80	TP= 0.24712+0.00690TUR-0.00012 Q -0.00029 CON	0.992



St37 Near Hanlon expressway bridge

St56 D/S of Guelph

St75 Near St Jacobs

St76 Near Canning

St78 Near Drayton

St80 Mount pleassent Street, Brantford

Figure 1 Grand river basin and sampling locations

Figure 2 Comparison of Observed and Computed

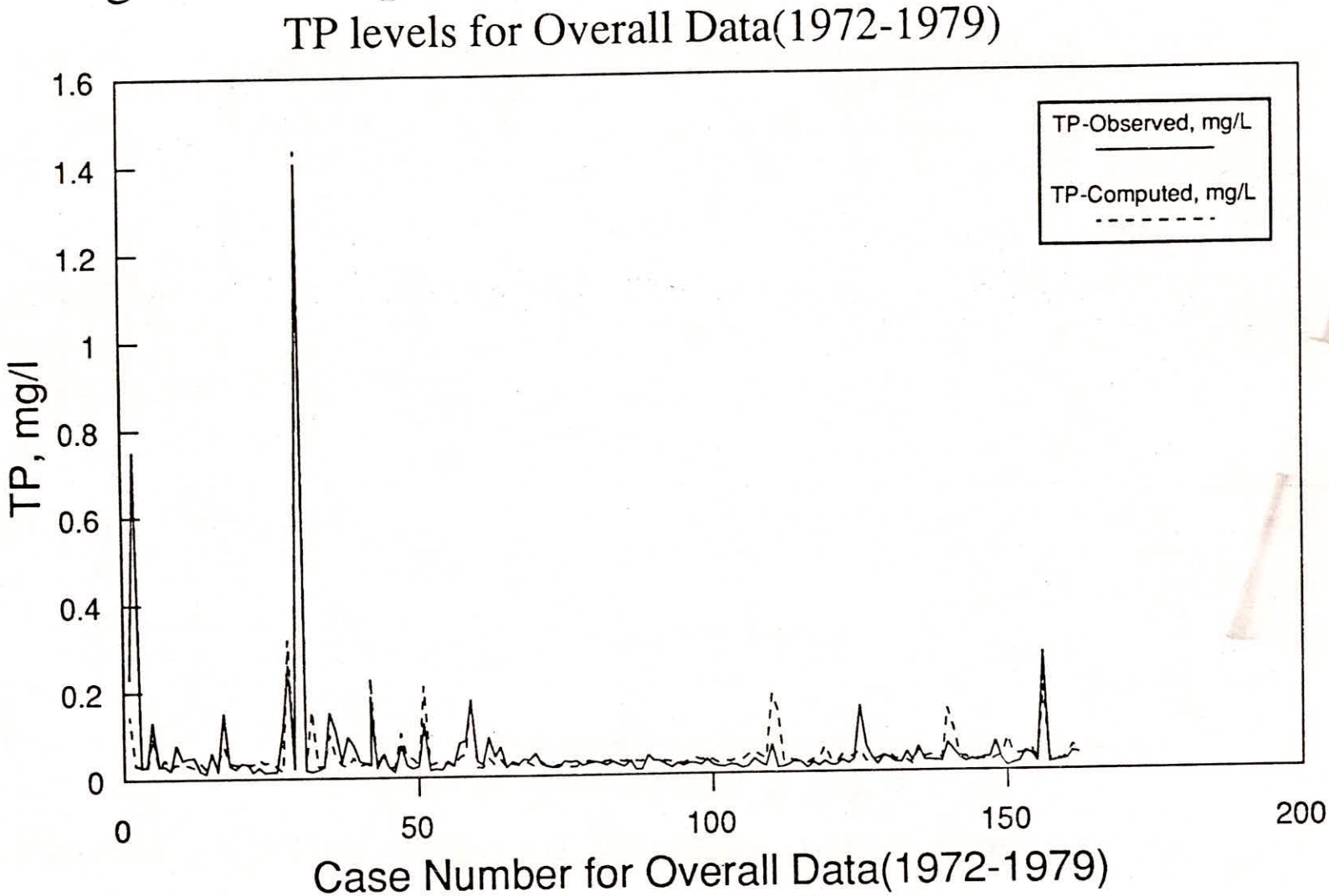


Figure 3 Comparison of Observed and Computed TP levels for Fall-season at St56.

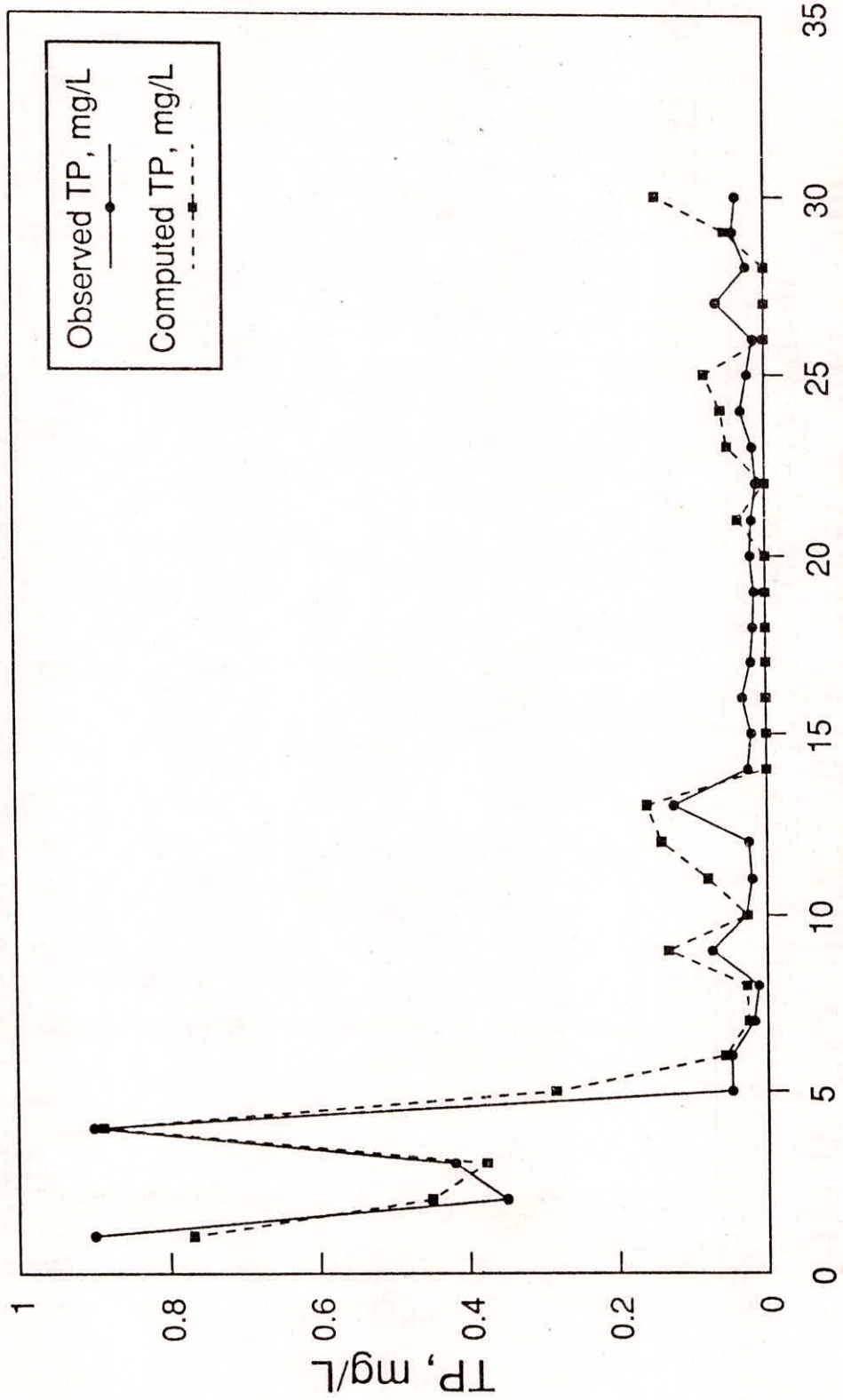


Figure 4 Comparison of Observed and Computed TP levels for Summer-season at St56.

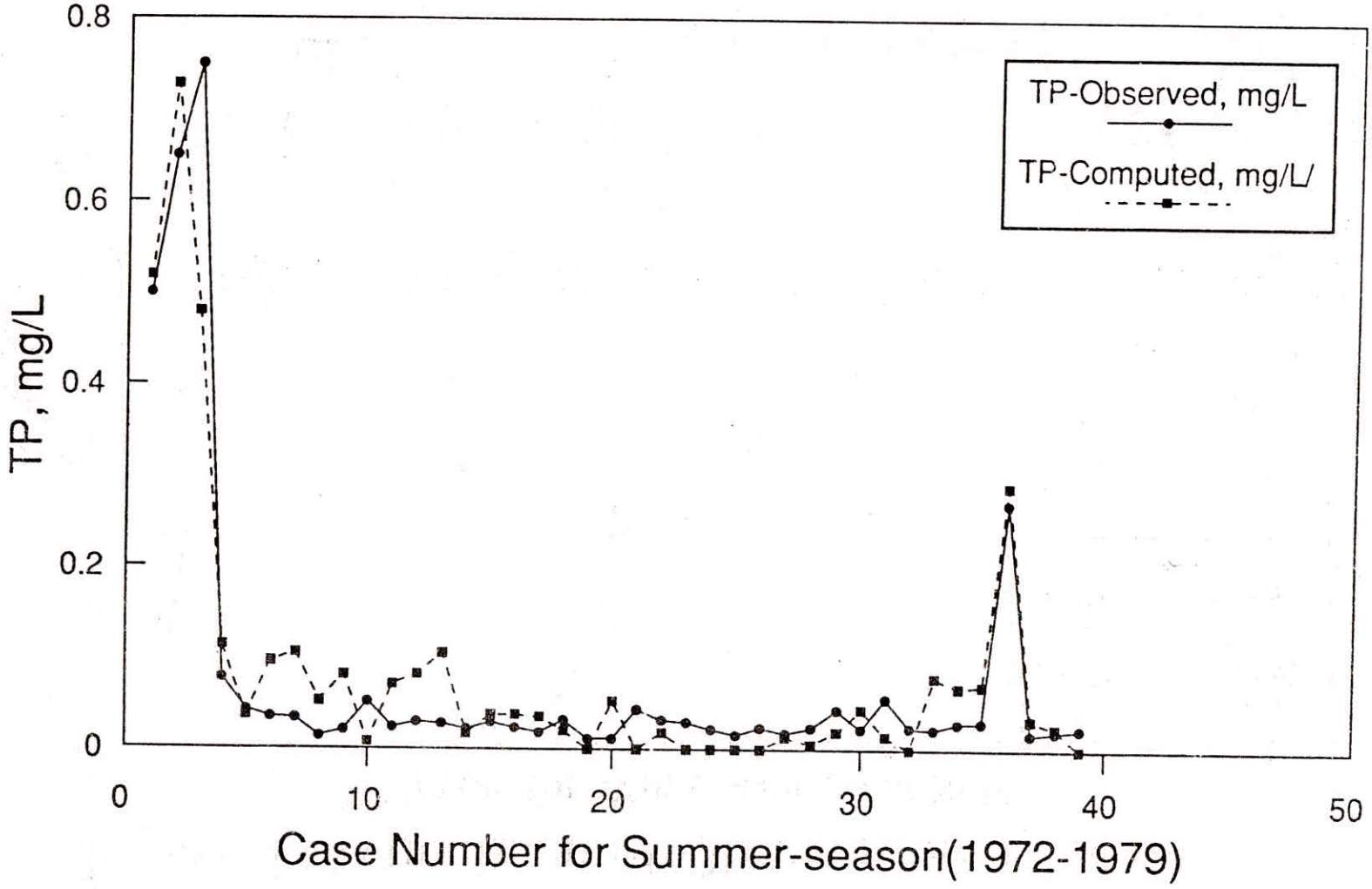


Figure 5 Comparison of Observed and Computed TP levels for Winter-season at St56.

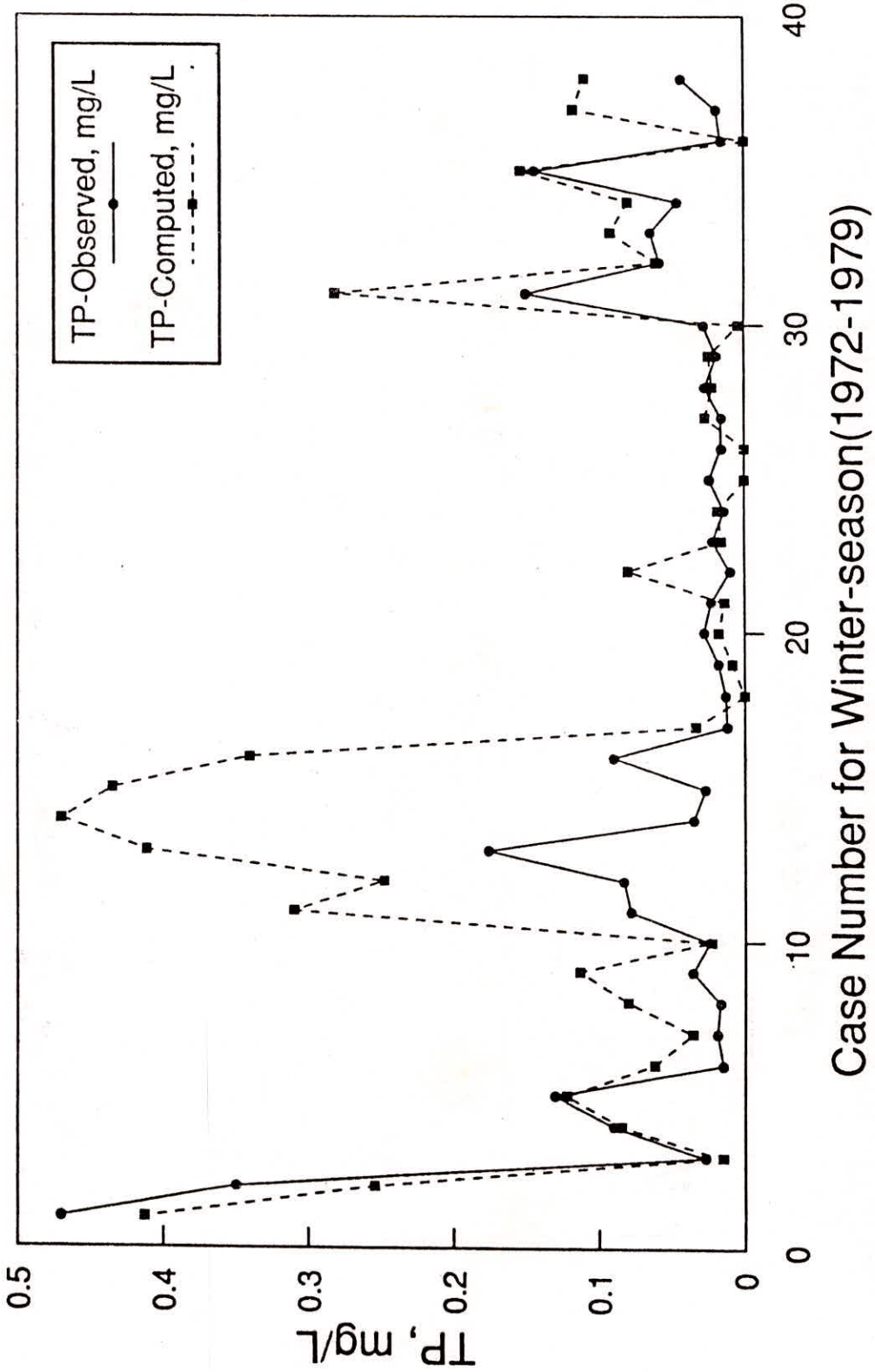


Figure 6 Comparison of Observed and Computed TP Levels for Spring-season at St56.