Prediction of phosphorus load from non-point sources to South African rivers

DH Meyer[1]* and J Harris[2]
[1]*Department of Statistics, University of the Witwatersrand, PO Wits 2050, South Africa*
[2]*WATERTEK, CSIR, PO Box 395, Pretoria 0001, South Africa*

Abstract

It has been found that the simple linear regression equation: $\ln Y_t = \ln\beta_0 + \beta_1 \ln X_t$ does not adequately describe the relationship between phosphorus export (Y_t) and runoff (X_t) for South African rivers. Serial correlation plays an important role in this relationship and must be incorporated into the model. In addition, for some rivers the above equation under-estimates phosphorus export for both low and high runoffs. This can be corrected by including a quadratic $(\ln X_t)^2$ term in the model. Finally, in order to eliminate the estimated bias in phosphorus prediction produced by anti-logging, it is necessary to apply a correction factor.

Introduction

Water eutrophication is caused by excessive fertilisation with nutrients such as phosphorus and nitrogen. Eutrophication leads to excessive growth of algae and aquatic plants. This interferes with the use of water for domestic and industrial water supply, irrigation, recreation and fisheries. Jones and Lee (1982) maintain that phosphorus is the nutrient which most commonly determines algal growth in a water body. Consequently eutrophication control usually relies on the control of phosphorus loading.

The most common sources of phosphorus in a water body are, according to Jones and Lee (1982), domestic waste-water treatment plant effluents, runoff from land, atmospheric precipitation and dry fall-out. The first of these sources is commonly referred to as a point source while the remaining sources are referred to as non-point sources. In models designed to test the effect of phosphorus control on eutrophication, models to simulate non-point source phosphorus export to impoundments are required. As illustrated in Fig. 1, for non-point source dominated rivers such as the Vaal River, monthly phosphorus loads are strongly influenced by monthly runoffs.

Grobler and Rossouw (1988) have modelled the non-point source phosphorus export for South African rivers in terms of runoff. They describe the phosphorus export for non-point source dominated rivers by the equation:

$$Y_t = b_0 X_t^{b_1} \qquad (1)$$

where:

- Y_t denotes the phosphorus export for month t
- X_t denotes the runoff for month t
- b_0 and b_1 are coefficients estimated for each river using nonlinear regression.

In this paper we consider the Grobler and Rossouw (1988) monthly data for 6 non-point source dominated South African rivers. For these rivers we find that an improved method for estimating and modelling phosphorus export can be developed.

Data

Grobler and Rossouw (1988) estimated monthly phosphorus loads from continuous flow measurements and periodic measurements of phosphorus concentration. Walker (1986) suggested numerous methods for achieving this. His most successful method, stratified regression (Cochran, 1977), was used by Grobler and Rossouw (1988) to compile their data.

In this method contemporaneous daily flows and phosphorus concentration data are divided into strata on the basis of daily flow. Frequencies, mean flows and mean concentrations are calculated for each stratum. A weighted regression analysis is then performed on the mean flow and concentration data using the strata frequencies as weights. This produces a regression line which can be used to predict daily phosphorus concentrations, allowing the calculation of daily phosphorus loads by taking the product of the estimated concentrations and the measured daily flows. Grobler and Rossouw (1988) calculated monthly flows (m^3) and monthly phosphorus loads (kg), in this manner. Their final figures were standardised for catchment size.

Ostensibly this method may produce spurious correlations between phosphorus load and flow, since the loads are calculated as a product of concentration and flow. But, since the influence of runoff on phosphorus concentration is a real phenomenon (Chesters *et al.*, 1980), we consider this argument to be invalid.

Methods

Several measures, namely bias (%), R^2, s_e and the Durbin-Watson statistic, have been used to assess the adequacy of the approaches considered for predicting phosphorus load from runoff.

Bias indicates the extent to which mean predicted values for phosphorus export exceed mean observed phosphorus export. Ideally there should be no bias. The formula used to estimate "Bias (%)" is given by:

$$\text{Bias (\%)} = \frac{100(\bar{y} - \bar{y})}{\bar{y}} \qquad (2)$$

where:

$\bar{y}$ = *mean for* $\hat{y}_t$
$\bar{y}$ = *mean for* y_t.

For want of a better name the adequacy of the fit is measured by the R^2, defined as follows for t = 1, 2,...n:

$$R^2 = 1 - \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} (y_t - \bar{y})^2}. \qquad (3)$$

The R^2 is used in linear regression to measure the proportion of

*To whom all correspondence should be addressed.
Received 19 October 1990; accepted in revised form 18 April 1991.

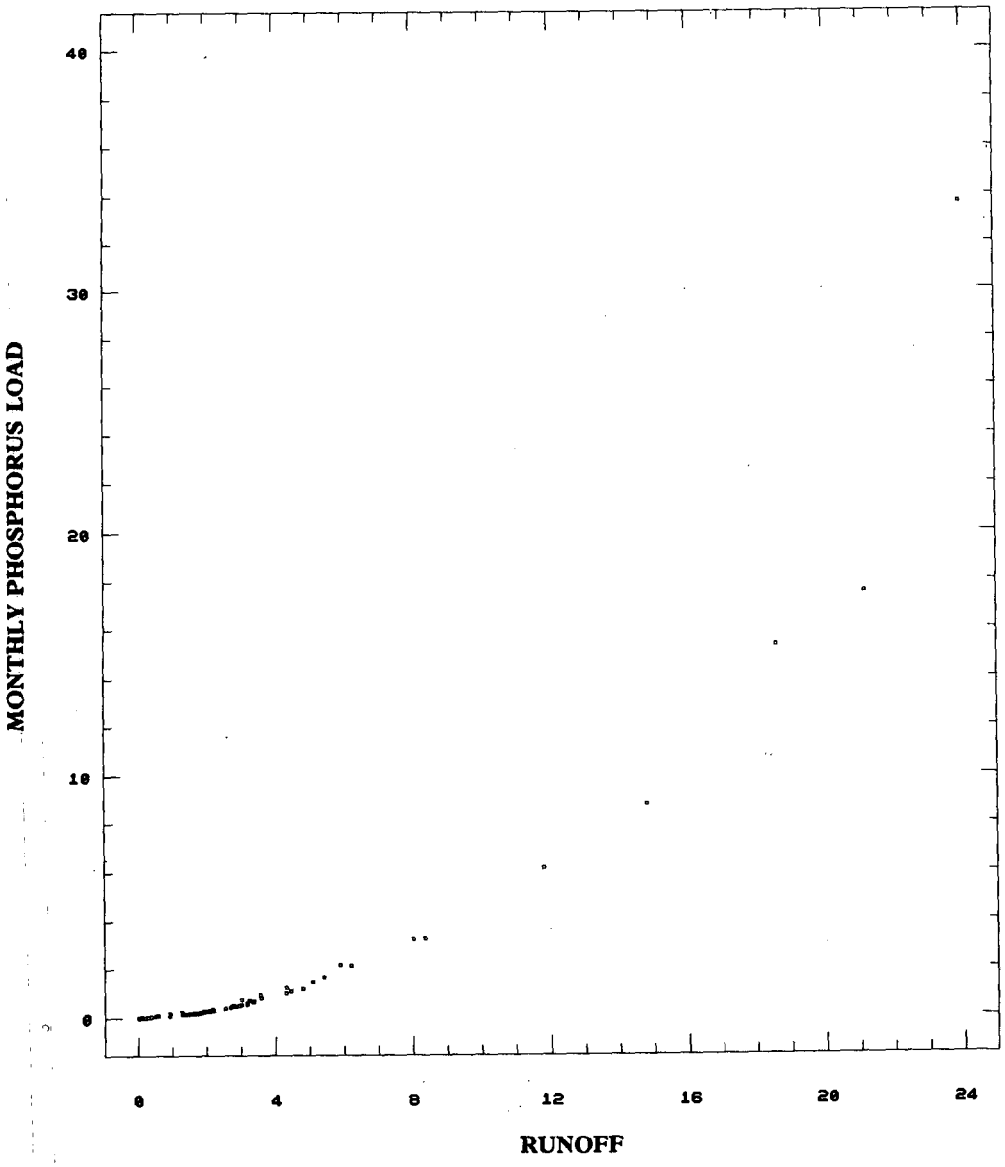

Figure 1
Vaal River: Monthly phosphorus load versus runoff

variability in y_t which is explained by the regression equation. In linear regression $0 \leqq R^2 \leqq 1$ with an R^2 of one indicating a perfect fit.

The root mean squared error, s_e, defined by the equation:

$$s_e = \sqrt{\frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n}} \quad (4)$$

is used to compare the relative accuracy of the various approaches. A low s_e is indicative of a good fit.

Finally the Durbin-Watson statistic is used as the measure of error independence. If errors cannot be considered as independent it means that time dependencies in the data are incorrectly modelled. The Durbin-Watson statistic is defined as follows:

$$DW = \frac{\sum_{t=1}^{n-1} (e_{t+1} - e_t)^2}{\sum_{t=1}^{n} e_t^2} \quad (5)$$

where:

$e_t = y_t - \hat{y}_t$.

A Durbin-Watson statistic value of close to 2 suggests independent consecutive errors. Values close to zero suggest positively correlated consecutive errors and values close to four suggest negatively correlated consecutive errors. Critical values for this statistic were originally produced by Durbin and Watson (1951) but are reproduced in many standard statistics texts, for example Neter *et al.* (1988).

Results

Regression

When Eq. (1) is log-transformed one obtains the regression line:

$$\ln(Y_t) = \ln(b_o) + b_1 \ln(X_t) \quad (6)$$

In an attempt to check the linearity of this relationship the logged phosphorus loads were smoothed using LOWESS smoothing (Cleveland, 1979) and plotted against log-transformed runoffs. In this smoothing procedure the smoothed points are obtained from weighted regression lines, a different regression line for each point. The weights used reduce as distance from the line increases

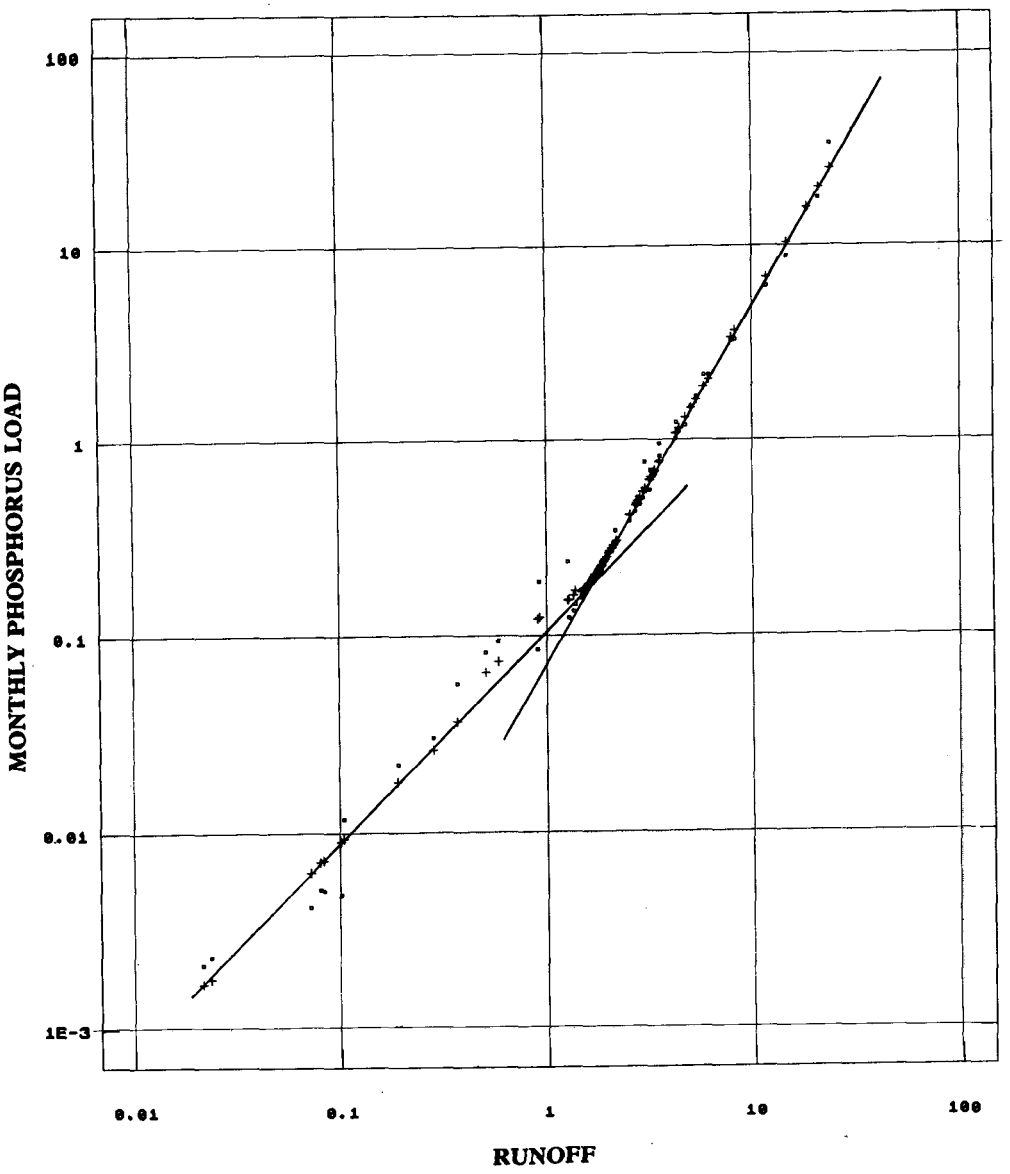

Figure 2:
Vaal River: Phosphorus load: raw and LOWESS smoothed

and as distance from the point in question increases. As illustrated in Fig. 2, for the Vaal River the LOWESS smoothed points do not suggest a linear relationship between ln(phosphorus load) and ln(runoff). Two lines with different slopes for low and high runoff levels are suggested. Such a relationship is better described by a quadratic equation with coefficients b_o, b_1 and b_2 of the form:

$$\ln Y_t = \ln b_o + b_1 \ln X_t + b_2 (\ln X_t)^2 \qquad (7)$$

than the simple linear regression Eq. (6).

Regression lines (6) and (7) were fitted to the raw (unsmoothed) data and predicted values for Y_t, $\hat{y}_t$ were obtained by anti-logging the predicted values obtained from these equations. The results shown in Table 1 indicate that the quadratic regression (7) produces less bias and better fits for four of the six rivers considered.

The values for the Durbin-Watson statistic(DW) in Table 1 indicate that the errors from Eqs. (6) and (7), Y_t-$\hat{y}_t$, exhibit significant serial correlation. Neter *et al.* (1988, p.880) explain that, when the errors exhibit significant serial correlation, the s_e in Table 1 may be seriously understated, causing the corresponding R^2 values to be seriously overstated. This can be avoided and the fit improved by incorporating time dependence in the phosphorus prediction equation, using a time series transfer model (Box and Jenkins, 1970).

Time series transfer model

The transfer model corresponding to Eq. (6) has the form:

$$\ln Y_t - \gamma \ln Y_{t-1} = \beta_1 (\ln X_t - \gamma \ln X_{t-1}) + \epsilon_t \qquad (8)$$

where γ and β_1 are parameters and ϵ_t is assumed to be a normally distributed random variable with mean μ and variance σ^2. Predicted values for Y_t, $\hat{y}_t$, are obtained from the equation:

$$\hat{y}_t = \exp(\hat{\mu} + \hat{\gamma} \ln y_{t-1} + b_1 [\ln X_t - \hat{\gamma} \ln X_{t-1}]) \qquad (9)$$

using least squares estimates for β_1, γ and μ (denoted by b_1, $\hat{\gamma}$ and $\hat{\mu}$). These parameter estimates are obtained by minimising:

$$\sum_{t=1}^{n} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{n} e_t^2 \qquad (10)$$

TABLE 1
LINEAR REGRESSION PHOSPHORUS PREDICTION

River	Simple linear regression				Quadratic regression			
	Bias%	R^2%	s_e	DW	Bias %	R^2%	s_e	DW
Magalies	7,4	94,7	0,10	1,18	2,5	98,5	0,05	1,37
Vaal	44,7	51,7	2,85	0,68	5,5	95,8	0,84	0,89
Vet	8,5	96,9	0,07	1,11	-7,9	79,3	0,17	1,25
Umgeni	21,5	77,1	1,12	1,17	5,3	88,3	0,80	1,33
Karkloof	2,9	99,7	0,07	1,04	-2,0	99,3	0,11	1,33
Sterk	15,8	91,7	0,29	0,84	-1,9	99,3	0,08	1,76

TABLE 2
TRANSFER MODEL PHOSPHORUS PREDICTION

River	Simple linear model			Quadratic model			
	Bias%	R^2%	s_e	Bias%	R^2%	s_e	P-values for b_2
Magalies	4,9	96,8	0,08	1,1	99,9	0,04	0
Vaal	25,0	77,0	1,96	4,3	95,1	0,91	0
Vet	-0,6	93,9	0,09	-9,1	80,8	0,16	0,00005
Umgeni	13,8	82,2	0,99	2,2	88,6	0,79	0
Karkloof	-7,7	99,4	0,10	-2,1	99,5	0,10	0,00012
Sterk	6,5	97,1	0,17	-2,3	99,2	0,09	0

The transfer model corresponding to Eq. (7) has the form:

$$\ln Y_t - \gamma \ln Y_{t-1} = \beta_1(\ln X_t - \gamma \ln X_{t-1}) + \beta_2([\ln X_t]^2 - \gamma[\ln X_{t-1}]^2) + \varepsilon_t \quad (11)$$

where β_1, β_2 and γ are parameters and it is assumed that ε_t has the same distribution as in (8). Predicted values for Y_t, $\hat{y}_t$, are obtained from the equation:

$$\hat{y}_t = \exp(\hat{\mu} + \hat{\gamma}\ln y_{t-1} + b_1[\ln X_t - \hat{\gamma}\ln X_{t-1}] + b_2[(\ln X_t)^2 - \hat{\gamma}(\ln X_t)^2]) \quad (12)$$

using least squares estimates for β_1, β_2, γ and μ (denoted by b_1, b_2, $\hat{\gamma}$ and $\hat{\mu}$). These estimates are also obtained by minimising Eq. (10).

The s_e and R^2 values calculated for models (8) and (11) in Table 2 are reliable because the errors from these models can be considered to be independent. Table 2 suggests that the quadratic model (11) produces better fits than the simple linear model (8) for five of the six rivers. In the case of the Vet River, model (8) gives better results than model (11), despite the very low P-value associated with b_2.

The levels of bias found in Table 2 tended to be lower than in Table 1. This strengthens the argument for transfer models as opposed to linear regression models for phosphorus load prediction.

Bias removal

In our final developmental phase we multiply the predictions, $\hat{y}_t$, by a suitable factor which produces a mean predicted phosphorus load equal to the mean observed phosphorus load. In this way we eliminate the estimated bias in our predictions of phosphorus load.

As suggested by Fig. 3 for the Vaal River, the errors obtained when Eqs. (11) (and (8)) are fitted to the data do tend to be normally distributed. In Fig. 3 the mean is estimated by the median and the standard deviation is estimated by 74% of the interquartile range. In this case the χ^2 goodness of fit statistic produces a P-value of 0,52 when we test for a normal distribution. This means that the hypothesis of a normal distribution for these errors cannot be rejected.

If the ε_t in Eqs. (8) and (11) are distributed $N(\mu, \sigma^2)$ then it is possible, theoretically speaking, to eliminate the estimated bias by multiplying the $\hat{y}_t$ predictions by:

$$\exp(\tfrac{1}{2}\sigma^2) \quad (13)$$

However, in practice we find that, if we substitute the maximum likelihood estimate for σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{n}(e_t - \bar{e})^2}{n} \quad (14)$$

in Eq. (13), the estimated bias is not equal to zero. Indeed, the estimated bias may actually increase.

Instead of using the factor given in (13) to correct for bias, it is suggested that we multiply y_t by the factor:

$$\exp((k_0 - 1)\hat{\mu} + \tfrac{1}{2}k_0^2\hat{\sigma}^2) \quad (15)$$

where:

$$k_0 = \frac{-\hat{\mu} - \sqrt{\hat{\mu}^2 + 2\hat{\sigma}^2(\ln\frac{\bar{y}}{\bar{y}} + \hat{\mu})}}{\hat{\sigma}^2} \quad (16)$$

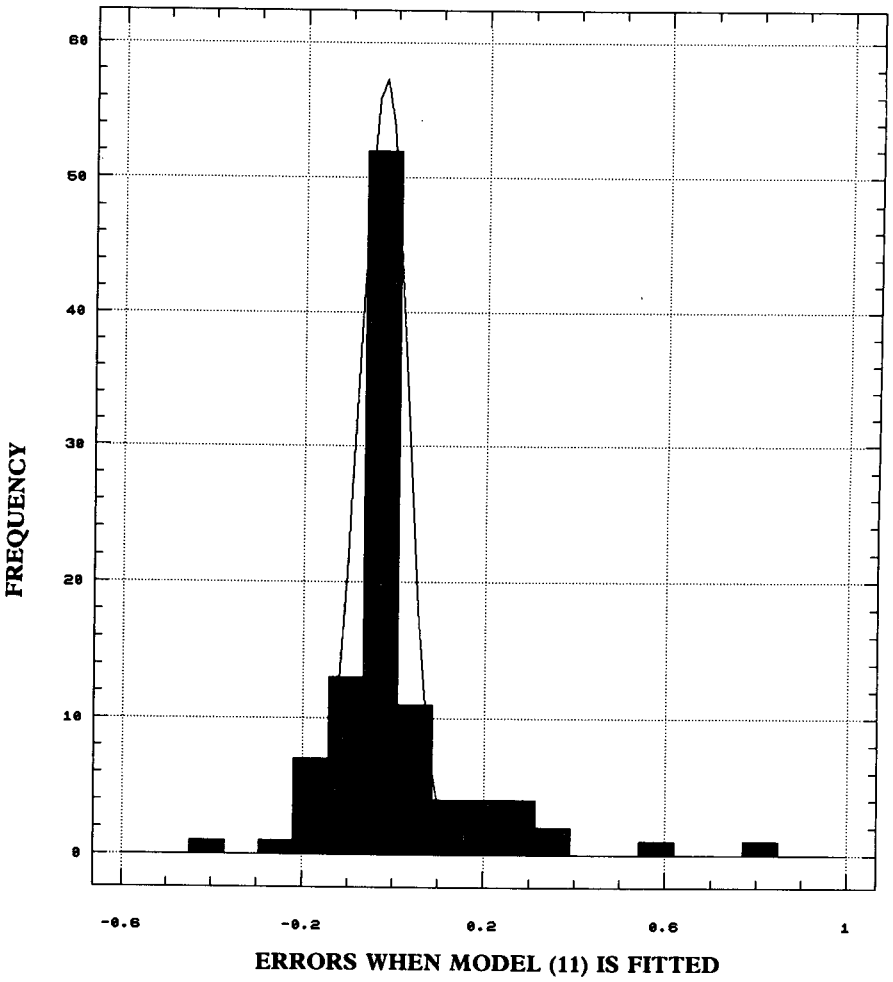

Figure 3
Frequency histogram: Vaal River errors N(-.0274,.002853)

In Eq. (15) k_o is an additional coefficient which must be estimated separately for each river using Eq. (16). As confirmed by the R^2 values for this and the other rivers in Table 3, bias correction has little, if any, detrimental effect on the R^2.

In Fig. 4 the final, bias corrected predictions for phosphorus load are compared to the observed loads in the case of the Vaal River.

Discussion

In Table 3 we find that for the Vet and Karkloof Rivers the R^2 for the simple linear transfer model is higher than for the quadratic transfer model. It appears that for the Vet and Karkloof Rivers a simple linear transfer model is more appropriate than a quadratic transfer model. This is a surprising result because the P-values associated with the Eq. (11) quadratic coefficients, b_2, in Table 2, are all highly significant. The effect of anti-logging is to amplify high predicted values. In the case of the Vet and Karkloof Rivers the effect of the quadratic model is to over-predict for high values of $\ln(Y_t)$. When we anti-log these over-predictions we inflate the error, hence reducing the R^2. This means that only for highly significant quadratic coefficients, b_2, (P-values < 0,00001), it is necessary to consider a quadratic rather than a simple linear transfer model.

Conclusions

It has been found that the equation:

$$\ln Y_t = \ln b_o + b_1 \ln X_t \qquad (17)$$

TABLE 3
ELIMINATION OF ESTIMATED BIAS

River	Simple linear transfer model		Quadratic transfer model	
	R^2	s_e	R^2	s_e
Magalies	97,7%	0,07	99,1%	0,04
Vaal	88,6%	1,38	95,7%	0,84
Vet	94,0%	0,09	87,4%	0,13
Umgeni	86,0%	0,88	88,6%	0,80
Karkloof	99,9%	0,04	99,7%	0,08
Sterk	98,1%	0,14	99,4%	0,08

does not adequately describe the relationship between phosphorus export and runoff for South African rivers.

For 4 of the 6 rivers studied it was found that a quadratic $(\ln(X_t))^2$ term was needed in the equation. In addition, for all 6 rivers serial correlation was found in the prediction errors obtained from this equation. This problem was circumvented by recognising the time dependence in the data, using a time series transfer model in place of a regression model to predict phosphorus loads. In order to eliminate the estimated bias entirely, the phosphorus predictions had to be multiplied by a suitable scaling factor, namely:

$$\exp\left((k_0-1)\hat{\mu} + \frac{1}{2}k_0^2\,\hat{\theta}^2\right) \qquad (18)$$

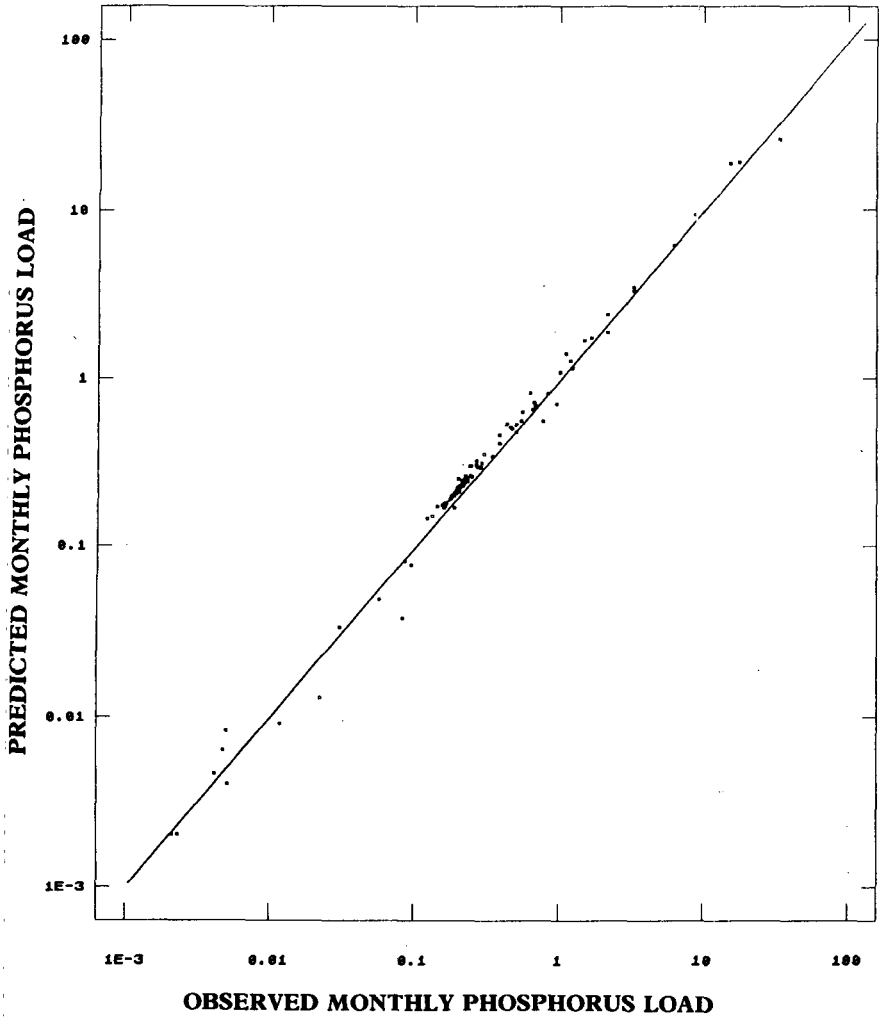

Figure 4
Vaal River: Observed v. predicted phosphorus load

where:

$\hat{\mu}$ and $\hat{\sigma}^2$ denote estimates for μ and σ^2

and k_o is optimised for each river as indicated in (16).

Acknowledgements

We gratefully acknowledge the assistance of Dr. Dirk Grobler of the Environmental Research Programme at the CSIR and Mr. Nico Rossouw of WATERTEK, CSIR. In addition we would like to thank the Water Research Commission for funding and for the provision of data.

References

BOX, GEP and JENKINS, GM (1970) *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco, California.

CHESTERS, G, COOTE, DR, JEFFS, DN, KONRAD, JC, OSTRY, RC and ROBINSON, JB (1980) Pollution from land runoff. *Environ. Sci. Technol.* **14** 148-153.

CLEVELAND, WS (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **74** 829-836.

COCHRAN, WG (1977) *Sampling Techniques.* New York Wiley & Sons.

DURBIN, J and WATSON, GS (1951) Testing for serial correlation in least squares regression 11. *Biometrika* **38** 173-175.

GROBLER, DC and ROSSOUW, JN (1988) Nonpoint source derived phosphorus export from sensitive catchments in South Africa. Confidential report to the Department of Water Affairs, Pretoria.

JONES, RA and LEE, GF (1982) Recent advances in assessing impact of phosphorus loads on eutrophication-related water quality. *Water Res.* **16** 503-515.

NETER, J, WASSERMAN, W and WHITMORE, GA (1988) *Appl. Stat.* Allyn and Bacon Inc.

WALKER, WW (1986) Empirical methods for predicting eutrophication in impoundments, Report 4, Phase 111: Applications Manual. Confidential report to the Department of the Army, US Army Corps of Engineers, Washington DC.