Austrian Journal of Statistics
Volume 27 (1998), Numbers 1&2, 39-48

Estimating Trends in Stream Water Quality with a Time-varying Flow Relationship

David Hirst
Biomathematics and Statistics Scotland

Abstract: The concentration of many elements in stream water depends in some way on the discharge. When looking for trends in these concentrations it is helpful to allow for this dependency. There are two advantages to this: the noise due to varying flow can be removed and thus the trend more clearly seen, especially if there are large variations in the flow over time (e.g. wet and dry years), and if the changes in the flow relationship are correctly modelled trends in high and low flow water can be investigated separately. This paper proposes an exploratory model in which the logarithm of the concentration is linearly related to that of the flow, but in which the slope of this relationship is allowed to vary smoothly through time, both with a long term trend and seasonally. The model also fits a trend in the intercept, which is also allowed to vary seasonally. The seasonal pattern is fixed, but the amplitude of the seasonal variations is allowed to vary smoothly through time. Thus the model is very flexible, and allows more aspects of the changes in water quality to be investigated than is the case with simpler models. For example, by predicting the concentrations at 95 and 5 percentile flow, the trends in high and low flow water quality can be investigated, even though these flows are rarely achieved in most data sets.

Keywords: Water Quality, Trend, Spline, Regression, Seasonality.

1 Introduction

Consider the data in Figure 1a. This shows the log concentration of sulphate measured in a stream at Glensaugh, Aberdeenshire, UK over a period of eight years. The measurements are monthly for the first four years and weekly thereafter. There is considerable temporal variation, but there is a clear seasonal pattern, and some evidence of longer term changes. There is also variation that cannot be ascribed to seasonality or trend, some of which may be due to variation in flow (Figure 1b). This data is typical of much water quality data, and the techniques developed in this paper are intended to be of use in a wide range of similar situations.

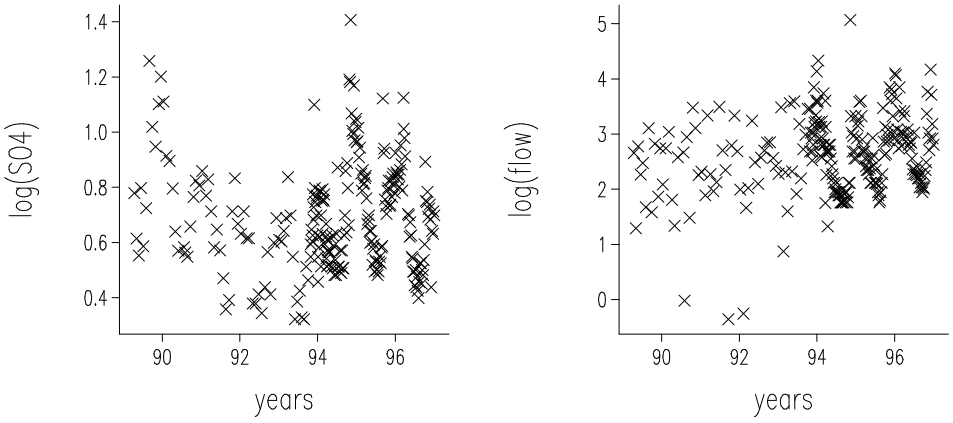

Figure 1: Time series of a) sulphate and b) flow in a stream in Glensaugh, Aberdeenshire

Many techniques have been suggested for exploring this kind of data. Esterby (1996) reviews much of the current methodology. The techniques can broadly be divided into two kinds: significance tests for trends, and descriptive methods. The significance tests, such Mann-Kendall, seasonal Kendall and Senn's t test, test for a trend defined in a very specific way, for example a monotonic decrease over time. While this may be appropriate in some situations, for example if a pollution source has recently been removed, in others there is no preconceived idea of what a trend may look like. This is the case at Glensaugh. Here we can only define trend as changes in the mean concentration after other variables such as flow and season have been accounted for. In this situation descriptive or exploratory methods are more useful. The simplest descriptive model simply fits a curve using a smoothing technique, such as a smoothing spline (Hastie and Tibshirani, 1990). This is clearly inadequate here as it ignores the marked seasonal pattern. Several authors have suggested ways of including seasonal effects, either by treating each month as a fixed effect estimated over the whole data set, or by fitting a continuous pattern such as a sine curve or a Loess estimator (Cleveland 1979). Robson and Neal (1996) used STL (Cleveland et al. 1990), a seasonal version of Loess, to

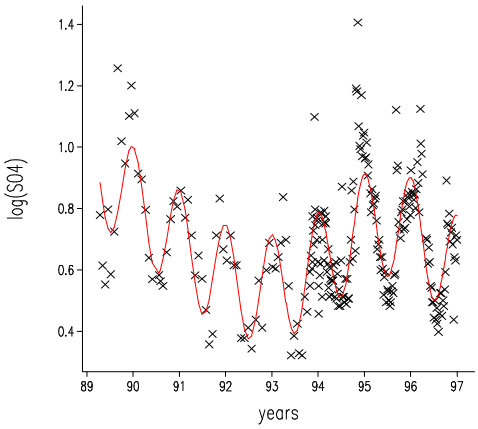

Figure 2: Results from fitting a constant seasonal pattern plus a smooth trend

analyse data from a Welsh catchment. Here a fixed non-parametric seasonal pattern is added to the smooth trend.

A similar model to that of Robson and Neal (1996) is fitted to the Glensaugh data in Figure 2. A trend estimated by a smoothing spline has been modified by including a seasonal pattern estimated by a sine curve, rather than a loess curve. The fit is fairly good, but the fixed amplitude of the seasonal pattern is not adequate, and there is still unexplained variation. Adding flow as a covariate has been suggested by many authors, but in this case it makes virtually no difference to the fit. The reason for this is clear from Figure 3 which shows the relationship between log(flow) and log(concentration). There is apparently no correlation. This plot however hides the fact that the relationship varies over time. This is shown in Figure 4 which plots the relationship separately for each year. The slope is positive when the concentrations are greatest, i.e. in 1989 and in 1995/96. When the concentration is lowest, in 1992, the slope is negative. This could be accounted for by estimating the slope separately each year, but it is better to allow it to vary smoothly through time.

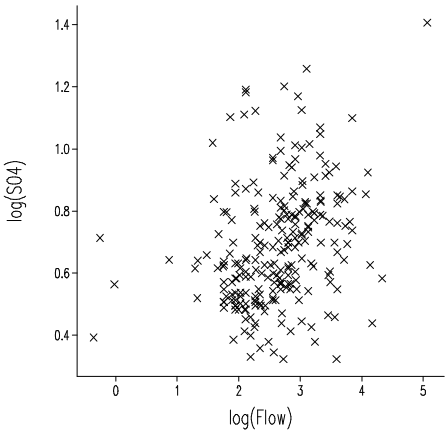

Figure 3: Log(concentration) versus log(flow) for the whole data set

This was the approach taken by Stalnacke and Grimvall (1996). Here the slope of a regression of concentration on flow is allowed to vary through time, with a roughness penalty to ensure it changes smoothly. Their approach is sufficiently flexible to allow for fixed seasonal effects but it is not clear that a variable seasonal pattern could be included. In this paper a different approach to allowing the flow-concentration relationship to vary smoothly through time is taken. This approach is easily extended to include a variable amplitude seasonal pattern.

The model is:

$$log(concentration(t)) = a_1(t) + a_2(t).log(flow(t)) + a_3(t).season_1(d) + a_4(t).season_2(d).log(flow(t)) + \varepsilon(t)$$

$a_1(t)$-$a_4(t)$ are smooth functions of time t

$season_1(d)$ and $season_2(d)$ are seasonal patterns with fixed amplitude

e.g. $season_1(d) = b_1\sin(d)+b_2\cos(d)$,

where b_1 and b_2 are constants such that $b_1^2+b_2^2=1$

d is time within the year in radians

$\varepsilon(t)$ is an error term.

The fourth term allows a seasonal variation in the flow-concentration relationship, and is included for completeness as there is no evidence in the example data that such variation occurs. In other data however this can be a significant part of the model.

In all subsequent sections, the notation for dependence on t or d is dropped, and any vectors are printed in bold, i.e. $a_1(t)$ becomes $\boldsymbol{a}_i$.

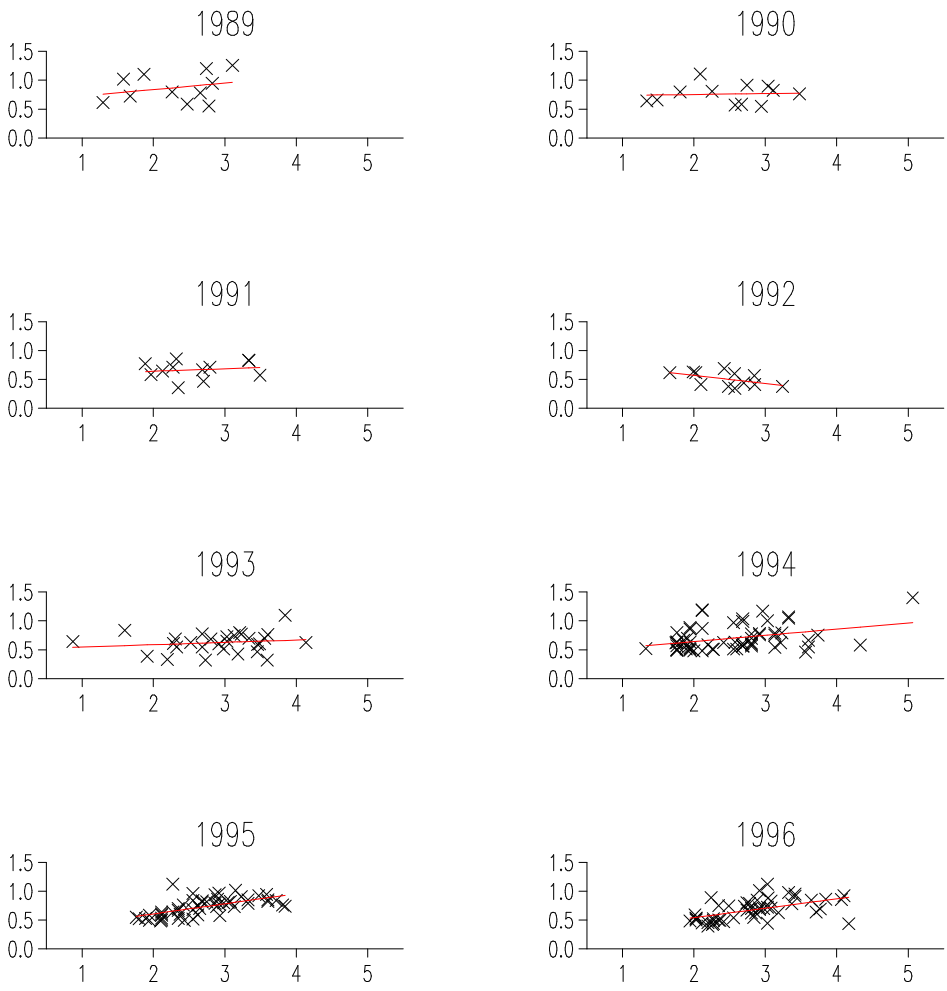

Figure 4: Log(concentration) vs. log(flow) for each year separately showing how the fitted linear regression varies over time

2 Fitting the Model

The model can be fitted using a backfitting algorithm, similar to that of Hastie and Tibshirani (1990). Any model of the form $\boldsymbol{y}=\Sigma\ \boldsymbol{a_i x_i} + \boldsymbol{e}$, i=1...*n*, where $\boldsymbol{a_i}$ is a smoothing spline in some variable such as time, $\boldsymbol{x_i}$ is a covariate (all except $\boldsymbol{x_1}$ adjusted to have zero mean) and $\boldsymbol{e}$ an error term can be fitted in the following way:

1) choose some starting values for the $\boldsymbol{a_i}$

2) let $\boldsymbol{y_{fit}}=\Sigma\ \boldsymbol{a_i x_i}$

3) for each *j* in turn, calculate $\boldsymbol{z_j} =(\boldsymbol{y}-\Sigma\ \boldsymbol{a_i x_i})/\ \boldsymbol{x_j}$ where the sum is over all $i \neq j$

4) re-estimate $\boldsymbol{a_j}$ so as to minimise the sum of squared differences between $\boldsymbol{a_j}$ and $\boldsymbol{z_j}$, weighted by $\boldsymbol{x_j}$, i.e. $\boldsymbol{a_j}$ minimises $\|(\boldsymbol{a_j} - \boldsymbol{z_j})\, \boldsymbol{x_j}\|^2$, subject to $\boldsymbol{a_j}$ having a chosen number of degrees of freedom. The weighting by $\boldsymbol{x_j}$, is necessary so that the same function is minimised in each iteration: $\|(\boldsymbol{a_j} - \boldsymbol{z_j})\, \boldsymbol{x_j}\|^2 = \|\boldsymbol{y}-\boldsymbol{y_{fi}}\|^2$
5) repeat steps 2 to 4 until convergence.

This algorithm therefore minimises the sum of squared residuals, subject to the $\boldsymbol{a_j}$ being smoothing splines with specified degrees of freedom. One or more of the $\boldsymbol{x_j}$ could themselves be regressions or other modelled terms, e.g. a seasonal effect $\boldsymbol{x_j}=b_1\sin(\boldsymbol{d})+b_2\cos(\boldsymbol{d})$ where $\boldsymbol{d}$ is time of year. To include this effect all that is necessary is to add an extra loop after estimating the $\boldsymbol{a_j}$, to estimate b_1 and b_2.

I.e. let $\boldsymbol{q_j}=(\boldsymbol{y}- \boldsymbol{y_{fit}} + \boldsymbol{a_j}\, \boldsymbol{x_j})/\boldsymbol{a_j}$
find b_1 and b_2 by a weighted regression of $\boldsymbol{q_j}$ on $\sin(\boldsymbol{d})$ and $\cos(\boldsymbol{d})$, where the weights are $\boldsymbol{a_j}$.

Note that it is necessary for convergence for all the terms bar one to have zero mean, and that the seasonal terms should be standardised to a constant amplitude (e.g. $b_1^2+b_2^2=1$) if the spline is to be interpreted as the amplitude of the effect. In the analysis of water quality data the x terms could be
$\boldsymbol{x_1}=1$
$\boldsymbol{x_2}=\log(\mathbf{flow})$
$\boldsymbol{x_3}=\mathbf{season}$ ($=b_1\sin(\boldsymbol{d})+b_2\cos(\boldsymbol{d})$+further terms if desired)
and sometimes
$\boldsymbol{x_4}=\mathbf{flow\ seasonality}$ ($=c_1\ \log(\boldsymbol{flow})\ \sin(\boldsymbol{d})+c_2\ \log(\boldsymbol{flow})\ \cos(\boldsymbol{d})$)
This would enable $\boldsymbol{a_1}$-$\boldsymbol{a_4}$ to be interpreted as:
$\boldsymbol{a_1}$ =trend
$\boldsymbol{a_2}$ =log(flow) - log(concentration) slope
$\boldsymbol{a_3}$ =amplitude of seasonal pattern in trend
$\boldsymbol{a_4}$ =amplitude of seasonal pattern in flow-concentration slope.

3 The Model Fitted to the Glensaugh Data

The model was fitted with 8 degrees of freedom for each of $\boldsymbol{a_1}$ to $\boldsymbol{a_3}$. The term $\boldsymbol{a_4}$ was initially fitted but then dropped as there was no evidence for seasonal change in the flow-concentration relationship. The choice of the degrees of freedom is clearly very important to the interpretation and fit of the model. There are unfortunately no obvious diagnostic plots available to help with the choice. Experience suggests one per year is a sensible choice. The requirements of a good model are that the residuals should be uncorrelated in time, and that the smooth terms should not be unrealistically rough. The fitted model is shown in Figure 5. Although it does not explain all the variation, the fit is clearly much better than the previous model. The three components, trend ($\boldsymbol{a_1}$), flow ($\boldsymbol{a_2x_2}$) and seasonality ($\boldsymbol{a_3x_3}$) are plotted separately in Figure 6. They are added together to get the final fit. Their relative magnitude can be compared either visually from the plot,

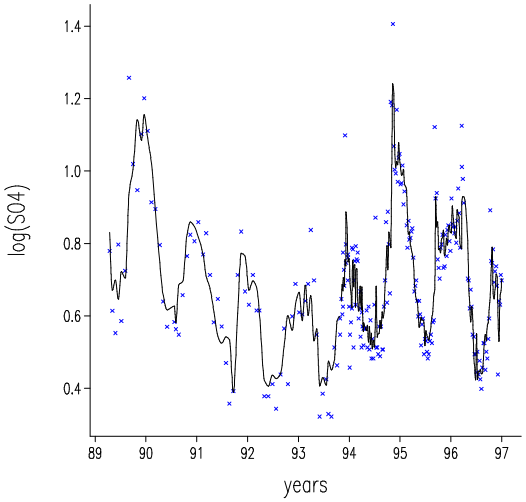

Figure 5: Fitted values from the final model, fitting a variable flow effect, variable amplitude seasonality and a smooth trend

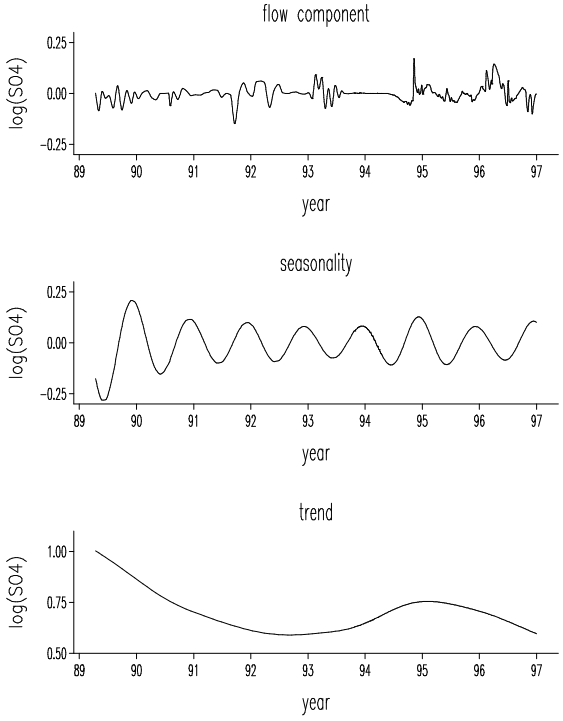

Figure 6: The three components of the model displayed separately. They are added together to get the final fit

or from their variances. In this case the percentage variances accounted for by each component are: Trend 25.8, seasonality 29.6, flow 5.8, residual 38.9. These variances can also be used for significance testing (see Section 4). They are given as percentages since their absolute values are of no interest.

The trend can be interpreted as the concentration that would be achieved at any time, if all other effects were at their mean values. It is therefore a 'flow adjusted concentration', which removes the effect of wet and dry periods, and seasonality. One advantage of this model is that trends at flows other than the mean can be investigated. Because the flow relationship varies, trends at high and low flow can be very different. This can be seen by predicting concentrations at 5 and 95 percentile flow. This is done in Figure 7. It can be seen that the low flow trend is less variable than the high flow trend.

4 Significance Testing and Confidence Intervals

An approximate analysis of variance table for each term in the model can be constructed for significance tests of the individual terms. For example, for the above data the table is the following (The sums of squares

have been rescaled to total 100 here):

component	df	SS	MS	F	p
trend	8	25.8	3.23	17.2	<0.001
season	8	29.6	3.70	19.7	<0.001
flow	8	5.8	0.73	3.9	<0.001
residual	207	38.8	0.19		

This is only approximate because the degrees of freedom are not known exactly. The components are all correlated and so the true degrees of freedom will be less than the nominal 8 for each term. Similarly the residual degrees of freedom will be larger than 207. This means that the significance tests are conservative, but often, as in this case, this is not a problem. The tests also make the assumption that the residuals are independent. In this case this appears to be justified, but removing all autocorrelation may be difficult. Choosing the correct degrees of freedom for each term is important here.

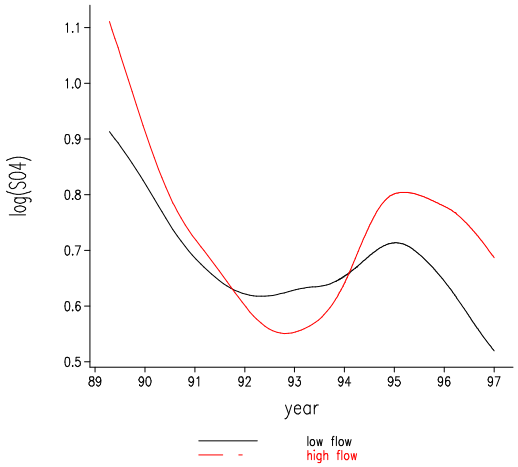

Figure 7: Predicted concentrations at low (5 percentile) and high (95 percentile) flow

The significance tests from the above analysis of variance test the hypothesis that there is no effect of the component, against the alternative that there is some effect. Although a common test to make, it could be argued that this is merely a measure of how much data is available. Probably more useful than significance tests are confidence intervals for each component. These can be calculated using the following bootstrapping technique:

i) Fit the full model and store the fitted values and residuals.
ii) Resample from the residuals (with replacement) and add back to the fitted values.
iii) Refit the model to the new data, and store the new fitted values and components.
iv) Repeat (ii) and (iii) 100 times. Pointwise confidence intervals can then be formed from the percentiles of the new components at each sample point.

It is an assumption of the bootstrapping technique that the residuals are exchangeable. This is similar to the assumption of independence in the ANOVA, and depends on choosing the correct degrees of freedom for the terms. Too few and the residuals will be autocorrelated.

90% confidence intervals for the main components of the model are shown in figure 8. Using this technique confidence intervals for any function of the fitted components can be constructed. In Figure 9 confidence intervals for the predicted concentrations at 5 and 95 percentile flow are shown. As would be expected, these are rather wide, reflecting the lack of information about concentrations at extreme flows. It would be

possible to construct intervals for the difference between these two predicted trends, if desired.

5 Alternative Ways of Fitting the Model

The algorithm in the previous section is only one way of fitting this type of model. In this paper the smoothed effects are fitted using smoothing splines, but this was merely for convenience given the software being used. Any smoothing technique could be used. Also there is no necessity to fit a variable linear effect of flow, it is reasonable to assume that the flow effect will be non-linear, even over short periods of time. Therefore it would make sense to fit this part of the model as a bivariate smooth in time and flow. This is perhaps best fitted using Loess (Cleveland, 1979). Preliminary results suggest that this will have the additional advantage of faster convergence.

The seasonal terms in the examples are assumed to be sine curves, with a constant location but variable amplitude. This can be made more sophisticated in two ways: Firstly the number of terms can be increased, i.e. to include sin(2t), cos(2t) and so on. There are no obvious diagnostic plots to help choose how many terms are necessary, and it is not always easy to choose between increasing the number of seasonal terms, and increasing the degrees of freedom for the trend, which will ultimately remove the seasonal pattern altogether. Secondly the phase as well as the amplitude can be allowed to vary. This is very easy to do, by setting two of the $\boldsymbol{x_j}$ to be *sin(t)* and *cos(t)*. If the seasonal pattern is not very clear however, this can produce some rather odd results, as there is no longer any restriction on when seasonal peaks occur. For example there could be two peaks in one year, and none in the next.

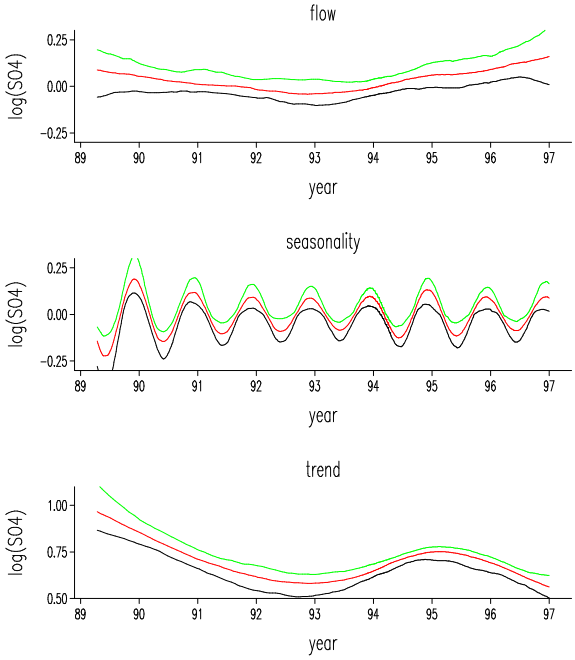

Figure 8: 90% confidence intervals for the three components of the model

6 Discussion

Although the model described is not based on a physical model of the catchment or environment, it is very flexible and can provide a lot of information about trends in water quality. These trends can then be related to data on physical processes, possibly with the aim of developing a physical model. If the trend term is all that is of interest, there is no real reason to fit such a sophisticated model, as the effect on the trend of adding seasonality and flow effects is usually fairly small. This would only not be the case if there was a long period of extreme flow, or very irregular sampling, or if the flow distribution changed over time. The advantage of the model is that it allows seasonal patterns and changes in the effect of flow to be examined in some detail. Changes in seasonality may relate to global warming, for example it has been suggested that the seasonal peak in nitrate concentration may be gradually moving. Also trends can be investigated at extreme flows, which will not have been sampled very frequently. It is to be expected that long term changes in precipitation will have different effects on low flow and high flow water. High flow water tends to react more quickly to changes in precipitation, and so shows more short term variation, which may be of immediate consequence to aquatic life. Low flow water will tend to change more slowly, and may be a better indicator of long term environmental changes. The model proposed allows some investigation of these effects, although it must be borne in mind that any estimation at extreme flows will be less reliable than estimation at the mean. Nevertheless being able to separate the two trends is a useful tool.

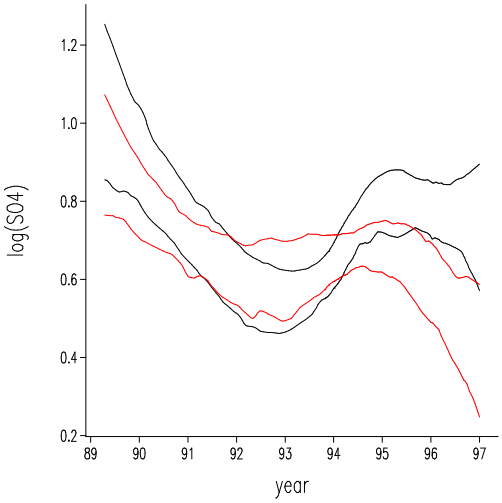

Figure 9: 90% confidence intervals for the predicted concentrations at low and high flow

There are still some problems with the model. It is difficult to decide how much to smooth each term, and there are no readily available diagnostic plots to help this process. In principle this could be decided by cross-validation, but the present implementation of the algorithm is too slow to do this in practise. The bootstrap confidence interval procedure is also very slow, and it could not reasonably be used if a large number of variables were being analysed. These problems may be resolved by using a different algorithm. Fitting a bivariate smooth in time and flow using Loess would improve the speed, and also allow a non-linear effect of flow to be fitted.

Acknowledgements

The data was collected by John Miller of the Macaulay Land Use Research Institute and the UK Environmental Change Network. The author is funded by the Scottish Office Agriculture, Environment and Fisheries Department.

References

S.R. Esterby. Review of methods for the detection and estimation of trends with emphasis on water quality applications. *Hydrological Processes*, 10, 127-129, 1996.

T.J. Hastie and R.J. Tibshirani. *Generalised Additive Models.* Chapman and Hall, London, 1990.

W.S. Cleveland. Robust locally weighted regression and smoothing scatter plots. *J. American Statistical Society*, 74, 386:829-836, 1979.

A.J. Robson and C. Neal. Water quality trends at an upland site in Wales, UK, 1983-1993. *Hydrological Processes*, 10, 183-203, 1996.

R.B. Cleveland, W.S. Cleveland, J.E. McRae, and I. Terpening. STL: A seasonal-trend decomposition based on loess. *Journal of Official Statistics*, 6, 3-73, 1990.

P. Stalnacke and A. Grimvall. A semiparametric method for estimation of time-varying relationships between runoff and riverine loads of substances. *Environmetrics*, 7,201-213, 1996.

Author's address:

David Hirst, Ph.D.
Biomathematics and Statistics Scotland
Macaulay Land Use Research Institute
Aberdeen AB15 8QH, UK
Electronic Mail: biossh@mluri.sari.ac.uk