Jul 1st, 12:00 AM

Water Quality Modelling in Rivers with Limited Observational Data: River Elbe Case Study

Renata J. Romanowicz

Ulrich Callies

Peter C. Young

Water Quality Modelling in Rivers with Limited Observational Data: River Elbe Case Study

Renata J. Romanowicz[1], Ulrich Callies[2] and Peter C. Young[1]

[1] *Lancaster University, Lancaster, LA1 3EZ*

[2] *GKSS, Research Centre, Institute for Coastal Research, Max Planck-Str. 1, D-21502, Geesthacht, Germany*

Abstract: Water quality predictions in an ungauged catchment require the development of a model that is able to capture the basic physical features of the process and depends only on variables that are easily available. From this point of view, the model has similar requirements to those used in future climate scenario analysis. The mechanistic water quality model, developed in GKSS, Germany, for the purpose of climate change analysis, uses only climatic variables, such as temperature, radiation and discharge, to predict the time variability of algae concentrations. This paper presents the development of a statistical analogue to this mechanistic model. The goal of this research is the derivation of a data-based model that has the minimum number of parameters required to explain the data and, at the same time, is able to represent the physical features of the process (a Data-Based Mechanistic or DBM model). The approximation of the mechanistic model is obtained by a statistical analysis of the relations between the model input and output variables, as well as the linearisation of the mechanistic algae equations, leading to the development of a statistically tractable model. The result of this analysis is a nonlinear, Multi Input Single Output (MISO) transfer function model that provides a statistical counterpart of the mechanistic algae model. The model is used to reconstruct hourly chlorophyll-*a* concentrations (a measure of algae concentrations) during the "pre-unification of Germany" period (before 1990) in the River Elbe, Germany. The uncertainty of the predictions is assessed and the results are validated against available monthly chlorophyll-*a* measurements.

Keywords: Data Based Mechanistic models; algae concentrations; mechanistic model emulation; Stochastic Transfer Function; uncertainty analysis.

1. INTRODUCTION

It is believed that mechanistic models, incorporating most aspects of a real system's behaviour, are suitable for predictions in ungauged catchments. However, ecosystems are complex and operate under random environmental conditions. There is no model able to predict environmental variables without error, even in fully instrumented catchments. We argue that complex, over-parameterised mechanistic models introduce a large amount of uncertainty related to parameter ambiguity and we propose here the application of *Data Based Mechanistic* (DBM) [Young, 2001 and references therein] modelling as a counterpart to mechanistic modelling. The advantage of DBM models lies in their parametric efficiency. The main problem with them lies in difficulties related to the physical interpretation of statistically derived relations between process variables. However, this process might be easier when a statistical equivalent to a mechanistic model is used as an intermediate stage.

The problem of identifying a statistical equivalent to a mechanistic model can be approached from at least two different directions. In one approach, we can start the analysis from the statistical properties of the available observations and derive the minimal order dynamical relations between the process variables [Young, 1999]. The other approach starts from the mechanistic description of the process and transforms the process equations into a state space description, which can then be linearised, so enabling the subsequent use of statistical tools.

The first approach has the advantage of providing, straight away, a representation of the data that is optimal from the statistical point of view but it might prove difficult in the case when the mechanistic model structure is complex. The difficulty of interpreting the parameters in physically meaningful terms arose from the rather special form of the mechanistic water quality model [Schroeder, 1997], which is based on approximations. These difficulties led us to use the alternative approach and start from mechanistic

model transformation, rather than the statistical model, and then utilize the statistical modelling methods after this first stage. This is a more conventional method, which normally leads to the linearisation of physical model equations. In this way, we are able to identify a consistent description of the process combining the optimal statistical representation of the process, from the point of view of the available observations, with a physical interpretation of the statistical model parameters.

The mechanistic model for algae and its statistical DBM equivalent use only external driving forces in the form of temperature, radiation and discharge. We want to investigate if the information contained in these data sets is sufficient to provide an adequate description of biological processes in river under totally different conditions to those in which the models were calibrated. In order to achieve this goal the chlorophyll-*a* concentrations in the Elbe in the period before German unification are used for DBM model validation.

In what follows, we describe the physical process that we address, available data and applied methodology (Section 2). In Section 3 we present a short description of the mechanistic model and a derived statistical emulator of its output. Two different DBM models are derived. First a DBM model for chlorophyll-*a* estimates in the Elbe has a structure partially derived from the mechanistic model linearised equations and thus its parameters can be related to physically meaningful parameters. The second model uses only the nonlinear transformation of temperature from the algae model equation. In the last section, we present the validation of both models on the pre-unification years.

2. DATA AND METHODOLOGY

2.1 Time series observations

The observation sets used in this study were collected by GKSS, at station Geesthacht (Elbe 586 km from the source) and include hourly observations of water quality, chlorophyll-*a*, silica, water temperature, radiation and discharge, starting from the year 1997. Chlorophyll-*a* is used as a surrogate measure of algae concentrations. After the unification of Germany (1990), water quality in the Elbe changed dramatically, following the closure of chemical factories in the upper reaches of the river. Thus the Elbe in the time before unification may be used to illustrate the ability of the models to work in very different conditions than the calibration conditions. Model validation is performed using monthly instantaneous observations of chlorophyll-*a* from Schnackenburg, 100 km up the river from Geesthacht, provided by the Elbe Water Authorities (WGE) [Romanowicz and Petersen 2004].

2.2 Methodology

In this study we apply Data Based Mechanistic methods introduced by Young [1999, 2001]. This approach tries to avoid theoretical preconceptions in the initial stage of the analysis, but wherever possible, the structure of the model is inferred directly from observations. Only then is the model interpreted in a physically meaningful manner. This physical interpretation is an essential element in all DBM modelling. No matter how well it matched the data, it is only considered truly credible if it can be interpreted in a physically meaningful way.

The Multiple Input, Single Output (MISO) Stochastic Transfer Function (STF) model used for off-line predictions has the form:

$$y_t = \sum_{i=1}^{i=M} \frac{B_i(z^{-1})}{A_i(z^{-1})} u_{i,t-\delta_i} + \xi_t \qquad (1)$$

where y_t is the algae concentration prediction at the end of sample time t; $u_{i,t-\delta_i}$ is the vector of input variables, i=1,..,M (e.g. temperature, radiation, or discharge) at the same sample time t; δ_i denotes any pure, 'advective' time delay for the ith input; and ξ_t represents the noise (which in some cases can be considered as zero mean, serially uncorrelated Gaussian white noise).

The polynomials $A_i(z^{-1})$ and $B_i(z^{-1})$ are defined as:

$$\begin{aligned} A_i(z^{-1}) &= 1 + a_{i,1}z^{-1} + \ldots + a_{i,n}z^{-n} \\ B_i(z^{-1}) &= b_{i,0} + b_{i,1}z^{-1} + \ldots + b_{i,m_i}z^{-m_i} \end{aligned} \qquad (2)$$

where a_i, b_j, $i=1,..,n$; $j=0,...,m_i$., are model parameters and the operator z^{-i} denotes a backward shift in time of i time steps, i.e. $z^{-i}u(t) = u(t-i)$. The value of the STF method depends on the amount of information available to statistically estimate the model parameters. Here, the SRIV algorithm in the CAPTAIN Matlab toolbox and associated Data Based Mechanistic (DBM) modelling concepts are used to identify the order of the STF model (the values of n, m_i and δ_i) and to estimate the associated parameters [e.g. Young, 1999, 2000, 2001].

The STF-based approach used here chooses, from among linear STF model structures, only those that have an inverse solution, i.e., the parameters of the model can be estimated uniquely. Moreover, this

technique estimates the covariance structure of the parameters together with the estimation errors, under Gaussian assumptions. The CAPTAIN toolbox methods use all the available information to derive the best estimates of the parameters using recursive time series analysis [Ljung and Soderstrom, 1983, Young, 1984]. Due to the uncertainty and simplifications involved in the description of environmental systems, the method will normally yield different estimated parameter values when calibrated over different sets of observations. However, in cases where the data sets have a sufficiently rich information content, the parameter values will not be significantly different in a statistical sense [Young, 1984].

3. COMPARISON OF MECHANISTIC AND DBM BASED APPROACHES

3.1 Mechanistic model approach

The main objective of the mechanistic model used in this study is the simulation of meteorologically induced variations of algal biomass. For this purpose a zero-dimensional model algae developed by Schroeder [1997] is applied. This model uses only external meteorological variables: temperature, radiation and discharge as the only time dependent model input variables. The model is set up in a Lagrangian framework, with a series of water packages, each assumed to be well mixed in volume, travelling downstream towards the Geesthacht Weir. The initial state variables are set constant and the modelled bio-chemical processes in each water body evolve independently, with negligible horizontal dispersion.

3.2 Data Based Mechanistic model

3.2.1 Statistical emulation of mechanistic model output

The STF model was based on temperature, radiation and discharge being treated as input variables and chlorophyll-*a* concentrations modelled by mechanistic model, treated as an observation variable. The model structure and its parameters were estimated using observations from the year 2000 and subsequently validated on the other observation sets (years 1997, 1998 and 1999).

The estimated first order Stochastic Transfer Function (STF) model has the following form:

$$y_t = 0.9824\,y_{t-1} - 0.0719\,temp_{t-27} + 0.0081\,rad_t - 0.0008\,dis_{t-1} + \varepsilon_t \quad (3)$$

where y_t denotes chlorophyll-*a* concentrations (μg/L), ε_t is the prediction error, *temp* denotes temperature in degree Celsius, *rad* is radiation in (W) and *dis* denotes discharge in (m^3/h).

The identified time constant is 56h, and the STF model explained R_T^2 =0.84 of the variance of the mechanistic model output.

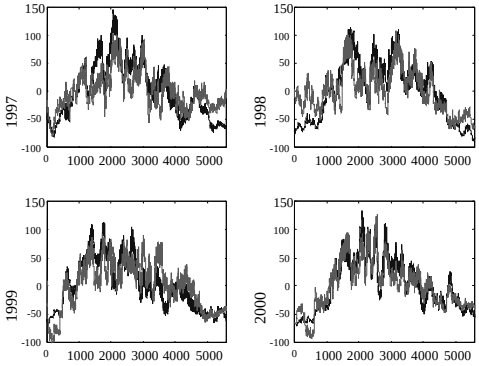

Figure 1. Validation stage for the 1st order STF model, years 1997, 1998, 1999. Darker lines denote algae model predictions. Calibration stage on 2000 year model output (algae) is shown in the lower right panel.

The validation stage shows that the model dynamics are well represented by this 1st order model and all the years reproduce well the trend of the chlorophyll-*a* concentrations. The decomposition of the model response into three components suggests that temperature is mainly responsible for the model trend. Discharge has a negative correlation with the modelled algal concentrations, which corresponds to the negative correlation found in the observation sets and to the dillution effect in the river. Radiation is mainly driving the rapid changes of model output.

3.2.2 STF analogue to mechanistic model

We shall transform the basic equation of algae model in order to linearise it:

$$Alg_t = Alg_{t-1} + \Delta t * Alg_{t-1} * (G - L - R) \quad (4)$$

where Alg_t denotes the algae concentration in time t, Δt denotes time period, G (Growth), L (Loss) and R (Respiration) describe rate constants of the basic biological processes.

Loss and Respiration depend on temperature only while Growth depends also on radiation and algal concentration. In order to transform this equation to one equivalent to STF, it is necessary to introduce logarithms on both sides of (4).

We obtain the following transformed equation for log(Alg):

$$y_t = y_{t-1} + \log(1 + G - L - \mathrm{R}) + \log(\Delta t) \tag{5}$$

where $y_t = \log(A\mathrm{lg}_t)$

After further approximations and introducing the relations for G, L and R [Romanowicz et al., 2002], we obtain a first order approximation to the mechanistic model equation (4):

$$y_t = y_{t-1} + a_1 k^{T-T_r} f_1(rad) f_2(Si) f_3(y_{t-1}) + a_2 k^{T-T_r} + \zeta_t \tag{6}$$

where $f_1(.)$; $f_2(.)$; $f_3(.)$ denote the nonlinear relations for radiation, silica and nonlinear feedback respectively. The term with temperature only corresponds to the first order approximation of loss and respiration functions, whilst the term with radiation and temperature corresponds to the growth function. There is also a non-linear feedback dependence on algae mass present in this term. The error ζ_t represents all the omitted higher order non-linear dependencies on radiation and temperature present in (5).

In order to get full equivalence between the stochastic and mechanistic model, we need to introduce the transport process of algae in the river. Mechanistic model uses a Lagrangian approximation to the transport problem. The reaction equation (4) is run for a finite period of time "backwards", thus giving an approximation of water body movement along the river. The time period is specified for each running time of the model, using an empirically derived relation combining the time of passage of water body and the observed amount of discharge. This procedure is in some sense representing the enhanced stream water quality model, Qual2 [Callies et al., 2000]. In order to obtain a fully equivalent stochastic model we should incorporate the reaction equation (4) into the partial differential equation describing the transport of the water body. Instead, in what follows we shall introduce discharge as an additional input to the DBM model in a similar fashion to the STF model (3).

The STF analysis performed using the observed chlorophyll-*a* concentrations (year 2000), and with temperature transformed according to equation (6), resulted in the best 1st order model, which explained 75% of the data:

$$y_t = 0.9775 y_{t-1} - 6.97 u_{t-3} + 6.95 u_{t-4} + 0.000004 rad_{t-1} - 0.000002 dis_{t-3} \tag{7}$$

where

$$x_t = 0.1 * 1.1^{T_t - Tref}$$
$$u_t = x_t - \frac{x_t^2}{2} + \frac{x_t^3}{3} \tag{8}$$

$y_t = \log(A\mathrm{lg}_t)$, *rad* denotes radiation [W] and *dis* denotes discharge in [m^3/h]; $T_{ref} = 20^o C$ and T_t is the temperature.

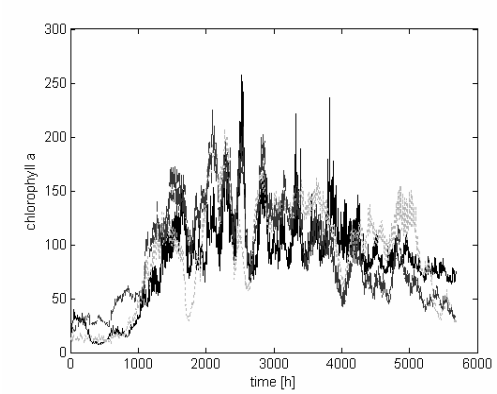

Figure 2. Comparison of STF model simulations (dark solid line) with observations of chloropyll a concentrations (gray dashed line) and mechanistic model results (black dot-dashed line).

Comparison of the simulations of mechanistic model and DBM models with observations (Figure 2) shows that both models have still much more in common with each other than with the real observations, in particular in the range of high values of chlorophyll-*a* concentrations. This confirms our previous investigations regarding the linear nature of mechanistic model realisations. This result shows also that mechanistic model uses the data efficiently, apart from the periods of lower concentrations where its performance could be improved. In particular, the analysis showed that the influence of discharge on mechanistic model performance should be improved.

As an alternative approach, we shall derive the STF model without logarithmic transformation of chlorophyll-*a* concentrations, while still applying an empirical nonlinear relation between algal concentrations and temperature (Eq. 8). Introducing this relation to the linear transfer function model (1), we get the second order STF model, which also explains 75% of the chlorophyll-*a* variations and has the following form:

$$y_t = -1.847 y_{t-1} + 0.848 y_{t-2} + 2.056 u_{t-3} + 0.0095 rad_{t-2} - 0.009 rad_{t-3} + 0.0001 dis_{t-2} + \varepsilon_t \tag{9}$$

The notation is the same as in Eq. 3 and the year 2000 is used for calibration of this model, shown in Figure 3.

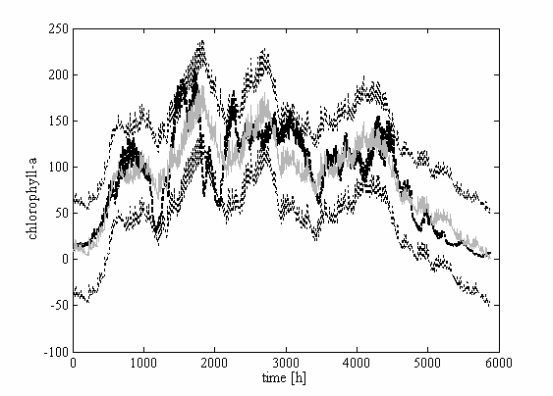

Figure 3. DBM model (9) without log-transformation: calibration on the year 2000; observations are shown as black continuous line, grey line denotes model predictions; dotted lines denote the 95% confidence bands.

This model has two real roots, corresponding to time constants 126 and 6 hours. This points to the existence of slow and fast responses of the model. Compared with our previous modelling experience [Romanowicz et al., 2002], this model gives a much better explanation of the data than the STF model with linear dependence on temperature.

4. VALIDATION OF DBM MODELS ON DATA FROM GERMAN PRE-UNIFICATION PERIOD

In this section we present the application of both derived DBM models to reconstruction of algal concentrations in the River Elbe during the German pre-unification period. The available observations of chlorophyll-*a* have the form of monthly instantaneous measurements taken during various parts of the day [Romanowicz and Petersen 2004]. DBM models provide the variance of the one step ahead predictions, which can be used to derive the confidence limits of the predictions. Figure 4 shows the validation results of the DBM model with logarithmic transformation of the state variables obtained for the year 1987. Due to the logarithmic transformation of variables, variance is heteroscedastic, i.e. it increases with the increase of the predicted chlorophyll-*a* concentrations. As a result, confidence limits have very wide bands for high values of chlorophyll-*a*, indicating large uncertainty of these predictions. Nevertheless, the observed values lie within the lower part of the confidence limits, indicating that the model overestimates the chlorophyll-*a* concentrations. This should have been expected, since before German unification the chlorophyll-*a* levels were very low, due to chemical pollution of the river.

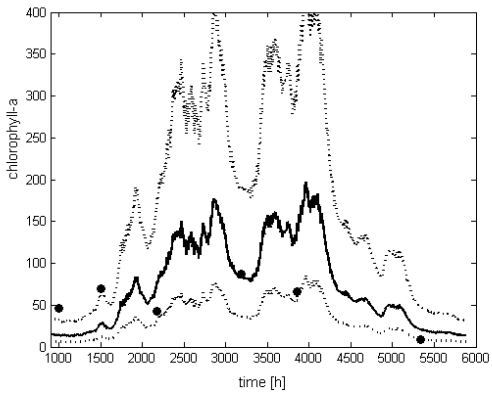

Figure 4. DBM model with log-transformation: validation on the year 1987; observations are shown as black dots, continuous line denotes model predictions; dotted lines denote the 95% confidence bands.

The same validation data were subsequently applied to the DBM model without log-normal transformation of chlorophyll-*a* concentrations (9). The results of the predictions, with 95% confidence bands, are shown in Figure 5.

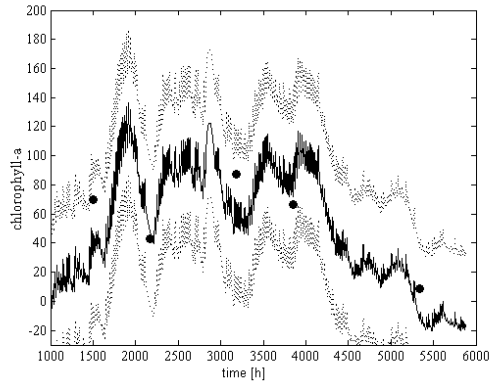

Figure 5. DBM model without log-transformation: validation on the year 1987; observations are shown as black dots, continuous line denotes model predictions; dotted lines denote the 95% confidence bands.

These model predictions have much narrower confidence bands (additive errors) and better explain the data. This result indicates that the assumption of additive errors rather than multiplicative is more suitable for chlorophyll-*a* predictions in the pre-unification period. However, this may result from very limited algal concentration variations in this particular case.

5. CONCLUSIONS

In this paper we have presented the application of DBM methodology to the prediction of chlorophyll-*a* concentrations in a river, working as

a surrogate for algae, in the River Elbe in the pre-unification period. In order to obtain a model suitable for application to an ungauged catchment, we followed the idea behind the development of the mechanistic algae model [Callies et al., 2002]. Namely, we applied only temperature, radiation and discharge as external driving factors (inputs) to the model. Firstly the STF analogue to mechanistic model was developed, which explains about 80% of the model performance, (i.e. most of its performance can be expalined by a linear model).

In the next stage, the mechanistic model equations were linearised and two DBM models were developed. One model, taking the closest resemblance with mechanistic description, uses the logarithm of chlorophyll-*a* concentration as a state variable and non-linear power transformation for the temperature. The second model uses the same transformation for the temperature without log transformed state variables. Both models were calibrated on data for the year 2000 and validated for the year 1987. The results indicate that the model without logarithmic transformation has much smaller confidence limits than the other model and gives a better fit to the data.

In this application we used hourly data, but daily data can also be used as the driving force for the DBM model predictions of algae. This apparent flexibility is very important where there is very poor instrumentation of the catchment. This work also shows that DBM models can be successfully used in modelling future climate scenarios.

6. ACKNOWLEDGMENTS

This paper describes work performed in collaboration with our colleagues from GKSS (G. Blöcker, W. Petersen, M. Scharfe, F. Schroeder) on the derivation of a statistical analogue to the mechanistic model and it constitutes part of our work on the river Elbe case study undertaken during the IMPACT EU project (2000-2003).

7. REFERENCES

Callies U., M. Scharfe, G. Blöcker,, R. Romanowicz, and P. C. Young,, Normalisation involving mechanistic models: benefits compared to purely statistical approaches, SCA Project IST 1999-11313, Deliverable 7, 2002.

Ljung L. and T. Soderstrom, *Theory and Practice of Recursive Estimation*, Cambridge Mass: MIT Press, 1983.

Romanowicz R. and W. Petersen, Modelling algae concentrations in the River Elbe in the years 1985-2001 using observations of daily oxygen concentrations, temperature and pH, *Acta Hydrochimica et Hydrobiologica*, in print, 2004.

Romanowicz R., P. C. Young,, U. Callies and M. Scharfe, Development of statistical TF based analogue for mechanistic model of water quality data, SCA Project IST 1999-11313, Report, 2002.

Schroeder, F., Water quality in the Elbe estuary: significance of different processes for the oxygen deficit at Hamburg. *Environmental Modelling and Assessment* **2**, 73-82, 1997.

Young, P. C., *Recursive Estimation and Time-Series Analysis*, Berlin: Springer-Verlag, 1984.

Young P.C., Data-based mechanistic modelling, generalised sensitivity and dominant mode analysis. *Computer Physics Communications*; 115: 1-17, 1999.

Young, P. C., Stochastic, dynamic modelling and signal processing: Time variable and state dependent parameter estimation. In W. J. Fitzgerald, A. Walden, R. Smith, & P. C. Young (Eds.), *Nonstationary and Nonlinear Signal Processing.* Cambridge University Press, 74-114, 2000.

Young, P. C., The identification and estimation of nonlinear stochastic systems. In A. I. Mees (Ed.), *Nonlinear Dynamics and Statistics*. Boston: Birkhauser, 127-166, 2001.