ENVIRONMETRICS, VOL. 7, 109–121 (1996)

GREY BOX MODELLING OF OXYGEN LEVELS IN A SMALL STREAM

JUDITH L. JACOBSEN AND HENRIK MADSEN

Institute of Mathematical Modelling, Bldg. 321, DTH, DK-2800, Lyngby, Denmark

SUMMARY

Data from a stream is used to identify a dynamic model of the oxygen level as a function of solar radiation and precipitation. The time series of the oxygen level is occasionally corrupted by pumping of external water into the stream, which is of no interest to the biological processes in the stream. In this paper a grey box modelling approach, which is a statistical method taking the known physical relations into account, is used. Using this approach, an identification of a stochastic continuous time model for the oxygen level based on the discrete time data, where the corrupted data are considered as missing values, is outlined.

KEY WORDS oxygen dynamics; time series analysis; missing data; continuous time modelling; grey box models

1. INTRODUCTION

The purpose of this work is to establish a stochastic model of the oxygen level in a creek as a function of solar radiation and precipitation. Oxygen, measured as dissolved oxygen, is the most universal environmental factor when water quality in rivers and streams is evaluated. The level of oxygen determines the kind of aquatic fauna that may exist, and thereby the trophic level of the ecosystem. Harremoës and Malmgren-Andersen (1990) state that there is a pronounced need for stochastic descriptions of the different processes that influence water quality in order to obtain a better understanding of the involved phenomena.

This paper describes a grey box modelling approach to identify a continuous time linear stochastic dynamical model of the oxygen level in a small stream based on time series with missing values. The missing values are due to an occasional disturbance of the measurements, which is of no interest to the dynamics we wish to model. They are therefore identified, and considered as missing values.

Time series analysis traditionally deals with black box models, i.e. models which are identified purely on the basis of data. It may be questioned whether such models are only representative of the data themselves, and not of the general dynamics and physics behind the data.

Bohlin (1984), Madsen *et al.* (1985), Madsen and Melgaard (1991), Melgaard and Madsen (1993) and Melgaard (1994) describe a so-called grey box approach that combines available physical knowledge with statistical modelling tools. This is done by using well-known physical equations, and including them in the analysis, along with a set of data, which are then used to estimate the parameters, and validate or falsify the model structure.

No matter how well understood a system is, there are always uncertainties and phenomena that have not yet been detected or fully realized. Therefore, combining physical knowledge with information from data is an optimal approach towards understanding a physical system.

CCC 1180–4009/96/010109–13

Received December 1994
Revised June 1995

In Denmark, the Water Pollution Act, WPA, has set standards for stationary waste water discharge. A combined sewage system is designed with a specific hydraulic capacity, which includes an expected rate of overflow. The system has built-in storage basins to take some of the load when the sewage system cannot handle the extra during heavy rainfall. Overflows of polluted oxygen-free water will occasionally enter the environment, with negative consequences for the recipient quality.

The essence of the problem involves three different components: the drainage system, the sewage treatment plants and the recipient. Each of these components are equally important, but while the first two have been the subject of much research, the recipient has not yet been submitted to such thorough investigation. Rainfall is a powerful transient burden of the entire system, that may give rise to discharge of oxygen consuming components, resulting in lowered oxygen levels and dead fish as some of the repercussions. Thus the WPA-standards eventually may result in huge investments in renewal and restoration of the sewage system, as well as in the monitoring of the recipients.

Many of the dominant mechanisms which determine the variations of the oxygen level in small streams and rivers are quite well known and described deterministically by, for example, Simonsen (1994), Harremoës *et al.* (1989) and Harremoës and Malmgren-Andersen (1990). However, there are many approximations involved, as the oxygen dynamics involves several transient components: the natural diurnal and annual oxygen oscillation, caused by photosynthesis and respiration from plants, and the sudden external influence during rain. This causes so much high frequency variation (noise) that calibration of deterministic models is very difficult, and stochastic modelling has to be considered.

2. THE DATA AND THE VARIABLES

The dependent variable is the oxygen content of a small creek that runs along a recreational area in Copenhagen. The input variables are temperature, solar radiation, precipitation and water depth.

Oxygen (mg/l) and *temperature* (°C) data were measured online, in the middle of the stream, every 3–5 minutes, and show a marked diurnal variation.

Solar radiation, measured as global radiation, has a great influence on the oxygen content in a creek and therefore a clear diurnal (and annual) variation is found for the oxygen. The variation in amplitude is, however, according to Harremoës and Malmgren-Andersen (1990) and Madsen *et al.* (1985), of a random nature due to the variation of cloud cover. Solar radiation (W/m^2) is measured every 10 minutes at Højbakkegård, a nearby agricultural meteorological station.

Precipitation data, given as an intensity (μm/s), is collected and stored by the Danish Meteorological Institute. A thorough explanation of the collecting system and the measuring methods may be found in Bramslev (1989) and Jacobsen and Voss (1994).

Water depth is measured a bit down stream from the other measurements and is therefore only an approximation to the depth at the measuring station. However, since it is an important parameter, it has been included.

2.1. Sampling

Since the observations in the different series were not equally spaced, sampling was necessary. For the three variables of interest, the hourly measurements of the solar radiation is the longest time interval and the radiation at time t is given as the mean over the preceding hour. It was then chosen to sample all the data like the solar radiation. See Jacobsen and Voss (1994) for a more thorough description of the sampling techniques employed.

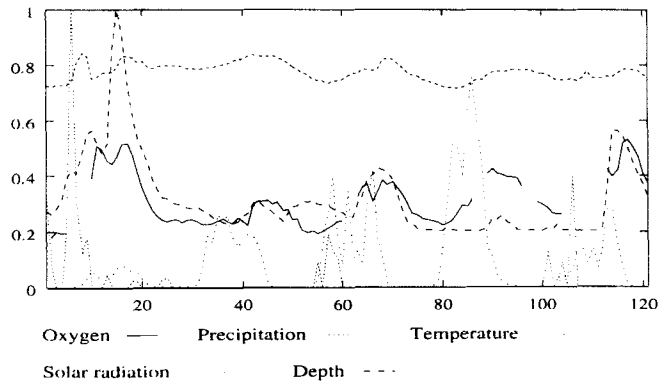

Figure 1. The scaled measurements of oxygen, solar radiation depth, temperature and precipitation. The oxygen measurements show the missing values

This sampling technique suppresses a lot of the high frequency variation, giving a smoother variation. A more correct treatment of filtering and sub-sampling, in order to avoid aliasing, was considered, but since the solar radiation series were available only as averaged hourly values, it was decided to treat the other series likewise for consistency. After sampling, each of the three time series contained 1460 hourly observations including 411 missing values.

3. A DYNAMICAL MODEL FOR THE VARIATION OF OXYGEN

The measurements of the input and output variables for five days are depicted together, scaled for comparison, in Figure 1. The figure shows that the oxygen level is affected by variations in both solar radiation and precipitation. There seems to be a delayed response in relation to the input, specifically from a shift in the solar radiation. The response to precipitation input is not as distinct as the response to the solar radiation. The depth of the water in the stream is seen to be highly correlated with the precipitation, although a little delayed. Whenever the water depth shows an increase, so does the oxygen level.

A dynamical system, such as a stream, is very complex. All the physical processes are not known and therefore difficult to describe in detail. In this section, only a brief introduction to the physical and chemical processes and the derivations of the governing differential equations will be shown for the oxygen dynamical system. For further details, see Jacobsen and Voss (1994).

Autocorrelations and partial autocorrelations revealed that a system of at least two state variables were to be expected (Jacobsen and Voss 1994). When the oxygen level of a small stream is to be evaluated, the concentration of organic matter is known to be the foremost important factor. Therefore, it seemed natural to have the concentrations of oxygen and organic matter as the two state variables.

The mass balance for oxygen and organic matter are thoroughly explained and derived in Harremoës and Malmgren-Andersen (1990). The mass balance for a water element is based on the assumption of total mixing and no longitudinal dispersion. The assumption of total mixing is usually fulfilled, except very close to the source of an organic matter discharge.

3.1. The differential equations

In the present complex dynamic system, where both the physical and chemical parameters behind the oxygen levels in a creek are interrelated, both non-linear and stochastic elements should be included in the description. A sketch of the processes involved is shown in Figure 2.

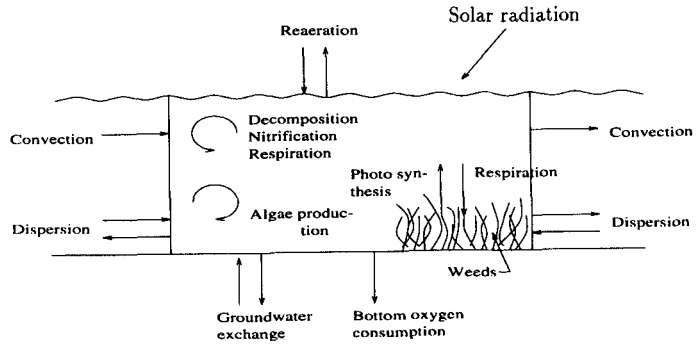

Figure 2. Overview of the processes involved with oxygen concentrations

The following two differential equations are suggested for the state variables oxygen, C, and organic matter, L. They are mainly inspired by Harremoës *et al.* (1989) and Harremoës and Malmgren-Andersen (1990), though the precipitation terms, involving P_r, are our own presumptions:

$$\frac{dC}{dt} = -\frac{K}{h\sqrt{h}}C - K_{1c}L + P(I) + \frac{K}{h\sqrt{h}}C_m(T) - R(T) - \frac{1}{h}\sqrt{CK_bP_r} \tag{1}$$

$$\frac{dL}{dt} = -K_{1l}L + K_3C + \kappa P_r \tag{2}$$

where

$$C_m(T) = 14{\cdot}5294 - 0{\cdot}3735771T + 0{\cdot}004973443T^2 \qquad (\text{mg/l})$$

$$R(T) = R_{15}\theta^{T-15} \qquad (\text{mg/l}) \tag{3}$$

$P(I)$ is some function of the solar radiation, I, e.g. $P(I) = \beta I$ (Cosby 1984), which is used in the following. This linear function does not incorporate possible saturated or even inhibiting conditions, which are assumed to be of limited importance. The coefficients in the empirical function for $C_m(T)$, the solubility of oxygen in water as a function of water temperature, are estimated via curve fitting to a table given in Harremoës *et al.* (1989). Respiration at temperature 15°C, R_{15}, is estimated along with the other parameters in the model. Otherwise, the function is empirical with $\theta = 1{\cdot}07$.

The physical explanation for the differential equations may be expressed as the change in oxygen is depleted by a contribution from re-aeration of the measured oxygen, $K/(h\sqrt{h})C$, but replenished from the saturated oxygen concentration, $K/(h\sqrt{h})\,C_m(T)$. Degradation of organic matter from various sources in the water volume also uses oxygen, $K_{1c}L$, while photosynthesis, $P(I)$, from solar radiation produces it. Depletion of oxygen is also caused by the respiration, $R(T)$, and from the sediment at the bottom of the creek via K_b. The change in organic matter is depleted by degradation, $K_{1l}L$, but replenished by oxygen involved in nitrification, K_3C, which provides nourishment in the form of nutrient salts.

Photosynthesis as a function of solar radiation has been the subject of much research for lakes, but not much for small streams and rivers. It obviously depends on plants, or rather the chlorophyll in the plants, which in turn depends on pH and nutrients in the stream.

3.1.1. Physical and chemical factors

The two main differential equations for the state variables, oxygen concentration and organic

matter (equations (1) and (2)), involve both physical and chemical factors. These are described in the following.

Solar radiation, I, may be influenced by solar elevation and cloud cover, possibly depending on water quality. It is assumed that there is no light saturation or inhibition. According to Harremoës and Malmgren-Andersen (1990), about 44 per cent of the total solar radiation lies in the photosynthetic active spectral area.

Re-aeration is an oxygen exchange from the air, striving to maintain equilibrium. Absorption of oxygen, in the water, is assumed to be a first-order process:

$$\frac{dC}{dt} = \frac{K}{h\sqrt{h}}(C_m(T) - C) \quad K = 2{\cdot}9 \times 10^{-4}(1 + \sqrt{\mathscr{F}_r})\sqrt{(gI_o)} \tag{4}$$

$$\sqrt{\mathscr{F}_r} = \frac{v}{\sqrt{(gh)}} \tag{5}$$

where K is a constant comprising Froudes number, $\mathscr{F}_r$, the absorption coefficient, the gradient of the stream, I_o and the gravitational acceleration, g. Equation (5) shows a dependence on the water depth, h, but, since it is negligible and to keep it simple, Froudes number is considered constant. $C_m(T)$ is the saturated oxygen concentration at temperature T, and C is the oxygen concentration. Harremoës *et al.* (1989) and Harremoës and Malmgren-Andersen (1990) give a detailed account for the re-aeration process. Stream velocity, slope of the stream bottom, i.e. water flow, are important factors. The dependence of temperature for the saturated oxygen concentration, $C_m(T)$, is tabulated in Harremoës *et al.* (1989). In this project a function is empirically fitted to the tabulated values. Higher temperature causes higher total respiration and low re-aeration due to high oxygen saturation.

Respiration, $R(T)$, from the plants in the creek uses oxygen. It is weakly dependent on the temperature, as described by equation (3). Respiration of organic matter, represented by CH_2O, is considered constant over the day. That makes it independent of the content of organic matter, plant types, the light intensity and of the oxygen-concentration itself, which may not be entirely likely, but commonly accepted as a first approximation (see Harremoës and Malmgren-Andersen (1990)). It utilizes oxygen, O_2, produces carbon dioxide, CO_2, and generates energy, which is stored by the opposite process, photosynthesis.

$$O_2 + CH_2O \xleftarrow{\text{energy}} CO_2 + H_2O. \tag{6}$$

Photosynthesis, $P(I)$, is a process where plants convert light into chemical energy. In a model for photosynthetic production of oxygen, light is assumed to be the only limiting factor, and, for simplicity, the production of oxygen is assumed proportional to the light intensity. With photosynthesis depending primarily on the light intensity and respiration being almost constant, the diurnal cycle from solar radiation is clearly recognized in the oxygen variation (see Figure 1).

Degradable organic matter in excess of what is needed to sustain life is not in itself harmful for the fish in the stream, but mineralization removes oxygen from the water. The cutting of weeds in the stream has a large, but unknown, effect in this context. The classical theory (as presented in Harremoës and Malmgren-Andersen (1990)), regarding decomposition of organic matter, is based on the assumption that the processing of organic matter, in a element of water following the stream, can be described as a first-order process:

$$\frac{dL}{dt} = M_L = -K_{1l}L \tag{7}$$

where M_L is the production of organic matter. The degradation constant, K_{1l}, is related to the organic matter, L, as indicated by the subscript.

Sedimentation, K_s, and extraction, K_e, is particular organic matter falling to the bottom of the creek by gravity, and dissolved organic matter being extracted from the water phase, respectively. A first-order removal is also assumed here, i.e. $dC/dt = (K_s + K_e)L$. These terms, as well as the degradation of organic matter, influencing the oxygen concentration, will be incorporated in one term only: $K_{1c}L$.

Nitrification uses oxygen in the biochemical oxidation of ammonia over nitrite to nitrate. Ammonia is present in waste water, run-off from farm land and decomposition of organic matter in the mineralization process. The stoichiometric coefficient is about 4·3 mg O_2 per mg $NH_4^+ - N$ ($NH_4^+ - N$ is ammonium bound as nitrogen). Since the organic matter is nourished by these salts, the production of organic matter depends on oxygen through the nitrification process. This term is then included as $dL/dt = K_3C$.

Precipitation, P_r, onto, and evaporation from, the surface of the stream itself will be neglected due to the very small surface. Precipitation run-off, though, may be of significance, thus accounting for some direct dependency and resulting in a delayed oxygen consumption. If this dependency is related to oxygen diffusion in the sediment, Harremoës *et al.* (1989) has shown that it is a half order process, thus resulting in the term with the square root of the oxygen concentration. The unknown contribution from precipitation (with an optimum delay, found to be 4 hours), in relation to the two state variables, will be incorporated with the term P_r.

4. A MAXIMUM LIKELIHOOD METHOD FOR PARAMETER ESTIMATION

In this section it is shown how the parameters in a linear stochastic differential equation are estimated using discrete time measurements and the maximum likelihood method.

4.1. The model in state space form

A model formulation based on physical laws will typically be a set of coupled differential equations. To account for the stochastic fluctuations, as well as the model discrepancies from the true system, stochastic terms are added. Thus the uncertainties and lack of knowledge about the system are included in the stochastic terms.

Including stochastic terms, the differential equation (equations (1) and (2)) can be written in state space form:

$$\begin{bmatrix} dC \\ dL \end{bmatrix} = \begin{bmatrix} -\frac{K}{h\sqrt{h}} & -K_{1c} \\ K_3 & -K_{1l} \end{bmatrix} \begin{bmatrix} C \\ L \end{bmatrix} dt + \begin{bmatrix} \beta & \frac{K}{h\sqrt{h}} C_m(T) + R(T) & \frac{1}{h}\sqrt{CK_b} \\ 0 & 0 & \kappa \end{bmatrix} \begin{bmatrix} I \\ 1 \\ P_r \end{bmatrix} dt + \begin{bmatrix} dw_1 \\ dw_2 \end{bmatrix} \quad (8)$$

where the last term is an additive noise term, which is introduced to describe the deviation between equations (1) and (2) and the true variation of the state variables C and L.

Only the oxygen concentration is measured. The recorded oxygen concentration, C_r, can be written as

$$C_r(t) = [1 \quad 0] \begin{bmatrix} C(t) \\ L(t) \end{bmatrix} + e(t) \quad (9)$$

where $e(t)$ is the measurement error.

In a short notation the equations are written:

$$\mathbf{dX} = \mathbf{AX}dt + \mathbf{BU}dt + \mathbf{dw}(t) \quad (10)$$

and

$$\mathbf{Y}(t) = \mathbf{C}\mathbf{X}(t) + \mathbf{e}(t) \tag{11}$$

where the matrix **A** characterizes the dynamical behaviour of the system and **B** is a matrix which specifies how the input signals enter the system. **X** is the state vector and **U** the input vector. Furthermore, $\mathbf{w}(t)$ is assumed to be a process with independent increments. With the purpose of calculating the likelihood function, $\mathbf{w}(t)$ is restricted to be a Wiener process with the incremental covariance $R_1^c(t)\mathrm{d}t$. $\mathbf{e}(t)$ is the measurement error, assumed normally distributed white noise with zero mean and variance $\mathbf{R}_2$. Furthermore, it is assumed that $\mathbf{w}(t)$ and $\mathbf{e}(t)$ are mutually independent.

It is a crucial question whether the parameters of a specified state space model can be identified. If a non-identifiable model is specified, the methods for estimation will not converge. See Madsen *et al.* (1990), Melgaard and Madsen (1993) and Melgaard (1994) for a further discussion on identifiability.

4.2. From continuous to discrete time

Since it is assumed that the system is described by the stochastic differential equation (10), it is possible analytically to perform an integration which under some assumptions exactly specifies the system evolution between discrete time instants.

The discrete time model corresponding to the continuous time model (10) is obtained by integrating the differential equation through the sample interval $[t, t+\tau]$. Thus the sampled version of (10) can be written as:

$$\mathbf{X}(t+\tau) = \mathrm{e}^{\mathbf{A}(t+\tau-t)}\mathbf{X}(t) + \int_t^{t+\tau} \mathrm{e}^{\mathbf{A}(t+\tau-s)}\mathbf{B}\mathbf{U}(s)\mathrm{d}s + \int_t^{t+\tau} \mathrm{e}^{\mathbf{A}(t+\tau-s)}\mathrm{d}\mathbf{w}(s). \tag{12}$$

Under the assumption that the input vector, $\mathbf{U}(t)$, is constant in the sample interval, the sampled version of (10) can be written as the following discrete time model in state space form:

$$\mathbf{X}(t+\tau) = \mathbf{\Phi}(\tau) + \mathbf{\Gamma}(\tau)\mathbf{U}(t) + \mathbf{v}(t;\tau) \tag{13}$$

where

$$\mathbf{\Phi}(\tau) = \mathrm{e}^{\mathbf{A}\tau} \tag{14}$$

$$\mathbf{\Gamma}(\tau) = \int_0^{\tau} \mathrm{e}^{\mathbf{A}s}\mathbf{B}\mathrm{d}s \tag{15}$$

$$\mathbf{v}(t;\tau) = \int_t^{t+\tau} \mathrm{e}^{\mathbf{A}(t+\tau-s)}\mathrm{d}\mathbf{w}(s). \tag{16}$$

Equation (13) is the discrete form of equation (10).

On the assumption that $\mathbf{w}(t)$ is a Wiener process, $\mathbf{v}(t;\tau)$ becomes normally distributed white noise with zero mean and covariance:

$$\mathbf{R}_1(\tau) = \mathbf{E}[\mathbf{v}(t;\tau)\mathbf{v}(t;\tau)'] = \int_0^{\tau} \mathbf{\Phi}(s)\mathbf{R}_1^c\mathbf{\Phi}(s)'\mathrm{d}s. \tag{17}$$

If the sampling time is constant (equally spaced observations), the stochastic difference equations can be written:

$$\mathbf{X}(t+1) = \mathbf{\Phi}\mathbf{X}(t) + \mathbf{\Gamma}\mathbf{U}(t) + \mathbf{v}(t) \tag{18}$$

where the time scale now is transformed such that the sampling time becomes equal to one time unit.

The calculations above were carried out assuming that $\mathbf{U}(t)$ was constant through the sampling interval $[t, t+\tau]$. In the following we shall, however, assume that $\mathbf{U}(t)$ is linear through the sampling interval. It is advantageous for cases where the measurements have been low pass filtered, e.g. by simple averaging, or generally when the input are not controlled.

The implementation of the linear interpolation is a modification to the derivations above. The modification is in the calculation of the term: $\mathbf{\Gamma}(\tau)\mathbf{U}(t)$ in equation (13).

Assuming that $\mathbf{U}(s)$ is linear through the points $(t, \mathbf{U}(t))$ and $(t+\tau, \mathbf{U}(t+\tau))$ results in the calculation:

$$\begin{aligned}\int_t^{t+\tau} \mathrm{e}^{\mathbf{A}(t+\tau-s)}\mathbf{B}\mathbf{U}(s)\mathrm{d}s &= \int_0^{\tau} \mathrm{e}^{\mathbf{A}s}\mathbf{B}\left(\frac{\mathbf{U}(t+\tau)-\mathbf{U}(t)}{\tau}(\tau-s)+\mathbf{U}(t)\right)\mathrm{d}s \\ &= \int_0^{\tau} \mathrm{e}^{\mathbf{A}s}\mathbf{B}(\tau-s)/\tau \mathrm{d}s(\mathbf{U}(t+\tau)-\mathbf{U}(t))+\mathbf{\Gamma}(\tau)\mathbf{U}(t) \\ &= \mathbf{\Gamma}_h(\tau)(\mathbf{U}(t+\tau)-\mathbf{U}(t))+\mathbf{\Gamma}(\tau)\mathbf{U}(t). \end{aligned} \tag{19}$$

Hence equation (18) becomes:

$$\mathbf{X}(t+1) = \mathbf{\Phi}\mathbf{X}(t)+\mathbf{\Gamma}_h(\tau)(\mathbf{U}(t+t)-\mathbf{U}(t))+\mathbf{\Gamma}\mathbf{U}(t)+\mathbf{v}(t). \tag{20}$$

The numerical details are described in Madsen *et al.* (1991).

4.3. The likelihood function

It is assumed that the observations are obtained at regularly spaced time intervals, and hence that the time index t belongs to the set $\{0, 1, 2, \ldots N\}$. N is the number of observations, or rather the length of the time series. Some of the observations may be missing. In order to obtain the likelihood function we further introduce:

$$\mathbf{Y}^*(t) = [\mathbf{Y}(t), \mathbf{Y}(t-1), \ldots, \mathbf{Y}(1), \mathbf{Y}(0)]' \tag{21}$$

i.e. $\mathbf{Y}^*(t)$ is a matrix containing all the observations up to and including time t. Finally, let $\boldsymbol{\theta}$ denote a vector of all the unknown parameters – including the unknown variance and covariance parameters in $\mathbf{R}_1^c$ and $\mathbf{R}_2$. The case of a missing observation is easily handed by assuming $\mathbf{R}_2 = \infty$ for that time instant, and taking $Y(t)$ to an arbitrary value.

The likelihood function is the joint probability density of all the observations assuming that the parameters are known, i.e.

$$\begin{aligned} L'(\boldsymbol{\theta}; \mathbf{Y}^*(N)) &= p(\mathbf{Y}^*(N)\,|\,\boldsymbol{\theta}) \\ &= p(\mathbf{Y}(N)\,|\,\mathbf{Y}^*(N-1), \boldsymbol{\theta}) p(\mathbf{Y}^*(N-1)\,|\,\boldsymbol{\theta}) \\ &= \left(\prod_{t=1}^{N} p(\mathbf{Y}(t)\,|\,\mathbf{Y}^*(t-1), \boldsymbol{\theta})\right) p(\mathbf{Y}(0)\,|\,\boldsymbol{\theta}) \end{aligned} \tag{22}$$

where successive applications of the rule $P(A \cap B) = P(A\,|\,B)P(B)$ is used to express the likelihood function as a product of conditional densities.

Since both $\mathbf{v}(t)$ and $\mathbf{e}(t)$ are normally distributed, the conditional density is also normal. The normal distribution is completely characterized by the mean and the variance. Hence, in order to parameterize the conditional distribution, we introduce the conditional mean and the conditional variance as

$$\hat{\mathbf{Y}}(t\,|\,t-1) = E[\mathbf{Y}(t)\,|\,\mathbf{Y}^*(t-1), \boldsymbol{\theta}] \tag{23}$$

and

$$\mathbf{R}(t|t-1) = V[\mathbf{Y}(t)|\mathbf{Y}^*(t-1), \boldsymbol{\theta}] \tag{24}$$

respectively. It is noted that (23) is the one-step prediction and (24) the associated variance. These can be calculated recursively by a Kalman filter.

4.4. The Kalman filter

A Kalman filter is a recursive, linear, real-time processing algorithm, which gives an optimal estimate of the states of a dynamic system in a noisy environment. It is used to recursively calculate the one-step prediction for the state of the system together with formulas for updating (or reconstructing) this estimate, based on a new observation.

In the present case where the transfer of the states of the system in discrete time is described by equation (2) and the observations by equation (11), the equations for *updating* the estimate of the state $\mathbf{X}$, according to Madsen (1989) and Melgaard and Madsen (1993), becomes

$$\hat{\mathbf{Y}}(t|t) = \hat{\mathbf{X}}(t|t-1) + \mathbf{K}_t(\mathbf{Y}(t) - \hat{\mathbf{Y}}(t|t-1)) \tag{25}$$

$$\mathbf{P}(t|t) = \mathbf{P}(t|t-1) - \mathbf{K}_t\mathbf{R}(t|t-1)\mathbf{K}_t' \tag{26}$$

where the Kalman gain, $\mathbf{K}_t$, is

$$\mathbf{K}_t = \mathbf{P}(t|t-1)\mathbf{C}'\mathbf{R}(t|t-1)^{-1}. \tag{27}$$

The formulas for *prediction* becomes

$$\hat{\mathbf{X}}(t+1|t) = \boldsymbol{\Phi}\hat{\mathbf{X}}(t|t) + \boldsymbol{\Gamma}_h(\mathbf{U}(t+1) - \mathbf{U}(t)) + \boldsymbol{\Gamma}\mathbf{U}(t) \tag{28}$$

$$\hat{\mathbf{Y}}(t+1|t) = \mathbf{C}\hat{\mathbf{X}}(t+1|t) \tag{29}$$

$$\mathbf{P}(t+1|t) = \boldsymbol{\Phi}\mathbf{P}(t|t)\boldsymbol{\Phi}' + \mathbf{R}_1 \tag{30}$$

$$\mathbf{R}(t+1|t) = \mathbf{C}\mathbf{P}(t+1|t)\mathbf{C}' + \mathbf{R}_2 \tag{31}$$

The formulas require some initial values which describe the prior knowledge about the states of the system in terms of the prior mean and variance:

$$\hat{\mathbf{X}}(1|0) = E[\mathbf{T}(1)] = \boldsymbol{\mu}_0 \tag{32}$$

$$\mathbf{P}(1|0) = V[\mathbf{T}(1)] = \mathbf{V}_0. \tag{33}$$

The matrix $\mathbf{P}(t+1|t)$ is the variance of the one-step prediction of the state, $\hat{\mathbf{X}}(t+1|t)$, of the system.

Using equations (21)–(24) the logarithm of the conditional likelihood function (conditioned on $\mathbf{Y}(0)$) is calculated:

$$\log L(\theta; \mathbf{Y}^*(N)) = -\frac{1}{2}\sum_{t=1}^{N}(\log\det\mathbf{R}(t|t-1) + \boldsymbol{\varepsilon}(t)'\mathbf{R}(t|t-1)^{-1}\boldsymbol{\varepsilon}(t)) + \text{constant} \tag{34}$$

where

$$\boldsymbol{\varepsilon}(t) = \mathbf{Y}(t) - \hat{\mathbf{Y}}(t|t-1) \tag{35}$$

is the one-step prediction error. It may be noted that in the case of missing observations we set $\mathbf{R}_2 = \infty$, which gives contribution to the likelihood at that time instant.

Since it has not been possible to determine an explicit expression for the ML-estimator, numerical methods have been used (see Madsen and Melgaard (1991)).

5. RESULTS AND DISCUSSION

Different models were tried and a satisfying model was found by means of varying the initial values, their range and checking the diagnostic information provided by the software used (see Melgaard and Madsen (1993)). Estimating parameters in a continuous time model has the advantage that the parameter estimates may be directly evaluated, physically. The found estimates and their standard deviations are shown in Table I.

All parameters, except organic matter L and respiration $R(T)$, are significantly different from zero. Organic matter, therefore, seems of little significance in itself in this creek, however, the parameters linking the two states were significant. It seems that the depth variations of the creek play a larger role than is commonly the case. This may be because the exchange with air turns out to be a dominant factor. In general, all the estimated parameters falls into ranges proposed in the literature (see Harremoës *et al.* (1989), Harremoës and Malmgren-Andersen (1990) and Simonsen (1994)).

Stochastically the model is evaluated by using tests for parameters equal to zero, as discussed above, and tests for white noise of the residuals. The tests have shown that an assumption of white noise is reasonable.

Figure 3 shows the measured and the simulated oxygen concentration, while Figure 4 shows the measured values and the one-step predictions for a part of the series. It is seen that the obtained model fits quite well, though of course not perfectly.

This model had no direct time delay between the solar radiation input and the oxygen output, and a four hour delay between precipitation input and the oxygen output. The parameter, κ, governing the precipitation is negative, suggesting that rain removes organic matter from the creek. However, this is possibly due to a dilution effect. Previously it had been expected that oxygen would be consumed when it rains, as a result of increasing levels of organic matter from sewage pipes and the precipitation itself. As seen in the simulation in Figure 3, there is some variation that is not accounted for by the model.

The finding of no direct time delay between the solar radiation input and the oxygen response may just be because the time resolution is too large compared to the one hour sampling time. Still the time constants of the system will describe a phase shift between input and output variables, and hence the time difference between, for example, the maximum radiation and the corresponding maximum oxygen (as seen in Figure 1). The rain has to be routed to the sewage system and then through that before it finds its way to the recipient, making the four hour delay seem possible.

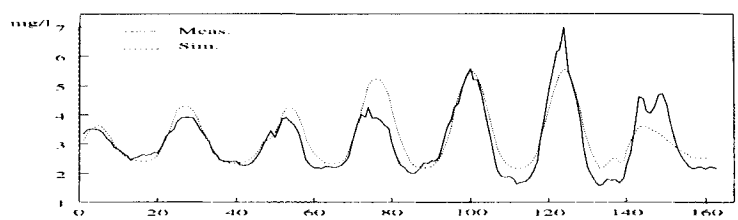

Figure 3. Measured and simulated oxygen for a part of the series

Table I. Maximum likelihood estimates for the oxygen system, per day, and their standard deviations

	Parameter							
	K_{1c}	K	K_3	K_{1l}	β	K_b	R_{15}	κ
Estimate	21·276	0·159	2·086	10·422	$0{\cdot}12 \times 10^{-2}$	14·574	$0{\cdot}27 \times 10^{-9}$	−110·42
Standard Deviation	(1·643)	(0·014)	(0·121)	(1·041)	$(0{\cdot}73 \times 10^{-4})$	(4·309)	$(0{\cdot}13 \times 10^{-7})$	(28·795)
Unit	$mgO_2/mgLh$	d^{-1}	mgO_2/mgL	mgL/mgO_2h	$mgO_2/W10^3l$	mgO_2/d	mgO_2/l	mgL/l

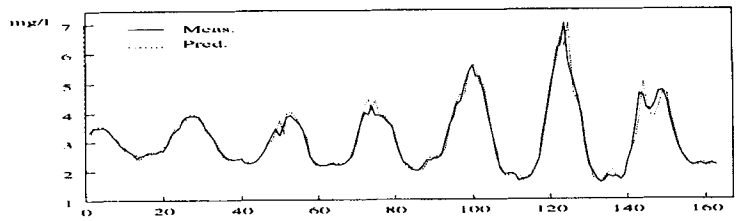

Figure 4. Measured and one-step predicted oxygen for a part of the series

More exhaustive research will have to be done in refining the equations. A more thorough investigation would include a non-stationary mass balance, such as differential mass balance over an infinitely small element of water. Also the analysis of values from two or more stations in the stream should prove more accurate. The estimated model in the present paper may be considered as a first step towards a better modelling and understanding of the oxygen dynamics in creeks.

6. CONCLUSION

A continuous time linear stochastic model for the oxygen dynamics of a small creek is formulated and estimated using discrete time data, with missing values, and a maximum likelihood method. The stochastic evaluation showed that the model was not contradicted by the statistical tests.

Purely statistical models, such as a transfer function model, are useful in classifying data and describing poorly understood systems, but they generally offer little physical insight; they are usually sufficient for prediction and control applications. The advantages of the so-called black box approach is that traditional time series methods can be used for identifying the structure of the transfer function. As an alternative, physical information may be used to establish the structure of the model. Here a continuous time model has been formulated based on the known physical equations for the system and estimated using discrete time series. Thus, the model is based on the physical laws for the system, as well as the actual data, and the resulting model is called a grey box model.

With the use of the grey box approach, it was possible to obtain a model for a rather complex dynamical system using the available physical knowledge combined with statistical modelling tools. The resulting parameter estimation was physically evaluated as describing the system in accordance to what is known and expected for this kind of system.

Keeping the model formulation and parameter estimation in continuous time has several advantages. Experiments with missing observations, which most traditional time series analysis approaches cannot deal with, are easily accommodated. The model can easily be improved by a combination of detailed comparison with data and the use of physical facts, and furthermore, a direct physical interpretation of the estimated parameters is possible.

The model obtained had no direct time delay from solar radiation to the oxygen concentration and a four hour time delay for the influence from precipitation. While organic matter usually is of great importance when evaluating the oxygen dynamics of a stream, it seemed to have little influence in this particular creek, while water depth had more impact.

ACKNOWLEDGEMENTS

The authors wish to thank Højbakkegård for providing the solar radiation data, the Department of Agricultural Sciences, Section of Soil and Water and Plant Nutrition, The Royal Veterinary and Agricultural University, Copenhagen, Denmark. Also gratitude is due to the Danish

Meteorological Institute, DMI, and Ballerup Municipality for providing precipitation, oxygen and temperature data, respectively.

REFERENCES

Bohlin, T. (1984). 'Computer-aided grey-box validation', Technical report TRITA-REG-8403, Department of Automatic Control, Royal Institute of Technology, Stockholm, Sweden.

Bramslev, J. P. (1989). *Treatment of Precipitation Data from a Precipitation Measuring System, Covering the Country*, Department of Technical Hygiene, Technical University of Denmark. (In Danish: *Bearbejdning of nedbørsdata fra et landsdækkende regnmålersystem*).

Cosby, B. (1984). 'Dissolved oxygen dynamics of a stream: model discrimination and estimation of parameter variability using an extended Kalman filter', *Water Science and Technology*, **16**,(5–7), 561–570.

Harremoës, P. and Malmgren-Andersen, A. (1990). *Textbook for Water Pollution*, Polyteknisk Forlag. (In Danish: *Lærebog i Vandforurening*).

Harremoës, P., Henze, H., Arvin, E. and Dahi, E. (1989). *Theoretical Water Sanitation*, Polyteknisk Forlag. (In Danish: *Teoretisk vand-hygiejne*).

Jacobsen, J. L. and Voss, L. (1994). 'Time series analysis and modelling of corrupted data', Master's report, no. 12/94. Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, Lyngby, Denmark.

Madsen, H. and Melgaard, H. (1991). 'The mathematical and numerical methods used in CTLSM – a program for ML-estimation in stochastic, continuous time dynamical models', Technical report 7/1991, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, Lyngby, Denmark.

Madsen, H., Spliid, H. and Thyregod, P. (1985). 'Markov models in discrete and continuous time for hourly observations of cloud cover', *Journal of Climate and Applied Meteorology*, **24**, 629–639.

Madsen, H., Melgaard, H. and Holst, J. (1990). 'Identification of building performance parameters', *in* Bloem, J. (ed), *Workshop on Advanced Identification Tools in Solar Energy Research*, Non Nuclear Energy, Commission of the European Communities, DG XII, pp. 37–60.

Madsen, H. (1989). *Time Series Analysis*, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark. (In Danish: *Tidsrækkeanalyse*).

Melgaard, H. and Madsen, H. (1993). 'CTLSM version 2.6 – a program for parameter estimation in stochastic differential equations, Technical report 1/1993, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, Lyngby, Denmark.

Melgaard, H. (1994). 'Identification of physical models', Ph.D. thesis, Institute of Mathematical Modelling, Technical University of Denmark.

Simonsen, J. (1994). 'Oxygen fluctuation in streams', Ph.D. thesis, Department of Technical Hygiene, Technical University of Denmark.