Structural Identification and Validation in Stochastic Differential Equation based Models - With application to a Marine Ecosystem NP-model.

Jan Kloppenborg Møller

DTU Informatics, Richard Pedersens Plads, Technical University of Denmark Building 321 DK-2800 Lyngby, Denmark.
National Environmental Research Institute, Fredriksborgvej 399, DK-4000 Roskilde, Denmark.

Jacob Carstensen

National Environmental Research Institute, Fredriksborgvej 399, DK-4000 Roskilde, Denmark.

Henrik Madsen

DTU Informatics, Richard Pedersens Plads, Technical University of Denmark Building 321 DK-2800 Lyngby, Denmark.

Summary. Stochastic differential equations (SDEs) for ecosystem modelling have attracted increasing attention during recent years. The modelling has mostly been through simulation based experiments. Estimation of parameters in SDEs is, however, possible by combining Kalman filter and likelihood techniques. The resulting filter equations handle additive diffusion effectively, while state dependent diffusion is difficult to handle. In many cases it is however possible to transform the state-space to avoid state dependent descriptions. It is demonstrated how pure random walk hidden state formulation and state estimation of key parameters can generate data driven model formulations. The resulting models are based on short-term predictions and it is demonstrated how considerations on stationarity of the distribution and inspection of probabilistic properties of simulation results can generate further model improvements of simulation models. The proposed methodology is demonstrated using phytoplankton and nitrogen data from a Danish estuary covering a 16 years period (1988-2003). It is demonstrated how non-linear relationships between states can be identified by plotting the (random) production parameter as a function of the state variables and global radiation. Further improvements of both the drift and the diffusion term are achieved by comparing simulated densities and data.

Keywords: Stochastic differential equations, Maximum likelihood, Extended Kalman filter, Structural identification, Validation, Lamperti transform, Simulation performance, NP-models.

1. Introduction

Stochastic differential equations (SDEs) are stochastic generalisations of ordinary differential equations (ODEs), where the differential increments are given a probabilistic interpretation (Øksendal (2003)). The theory of SDEs is in a mature state and the literature on theoretic properties of SDEs is bulk (e.g. Klebaner (2005), Karatzas and Shreve (1991)), and the use of SDEs is standard in mathematical finance.

SDEs has also proven useful in diverse fields such as pharmacokinetic (Tornøe et al. (2004)), engineering (Madsen at al. (1987)) and geolocation of fish (Pedersen et al. (2008)). These applications uses data to estimate parameters in a continuous-discrete time stochastic state space formulation, which allow a splitting of the noise processes into observation noise

and system noise. In SDEs system noise is often referred to as diffusion, which describes the stochastic part of the state-space formulation.

The estimation in the present work is based on an implementation of the Extended Kalman Filter (EKF) (e.g. Jazwinski (1970)) and approximate likelihood estimation as presented in Kristensen et al. (2004a). The EKF allows for optimal state estimation, and through modelling parameters in the model as pure random walk hidden states it is possible to formulate data-driven hypotheses based on the reconstructed or smoothed state of the system (Kristensen et al. (2004b)).

The EKF filter approach is effective in handling additive (state independent) diffusion, however, such an assumption is in many cases a strong simplification of real life systems, that will not fulfil basic requirements of the system, such as positive states. The assumption also exclude a large class of well known diffusion processes (such as "Black and Scholes" type models, (Øksendal (2003)) and the Feller diffusion (Iacus (2008))). For one-dimensional diffusion processes this is effectively handled by transformation of the state-space (Baadsgaard et al. (1997)), the transformation is often referred to as the Lamperti transform (Iacus (2008)). For multivariate processes this is a more delicate matter (Luschgy and Pagés (2006); Aït-Sahalia (2008)), but for a restricted class of diffusion processes it can be handled by Itô's lemma, and a general multivariate formulation that allow for a Lamperti type transformation is presented.

The present work is a further development of the methodology presented in Kristensen et al. (2004b), in addition to the consideration based on the likelihood and reconstruction of the random walk hidden states, the structural development is based on considerations about the stationary solution and simulations results. The proposed methodology is exemplified with a comprehensive study of a marine ecosystem.

Marine ecosystems represent very complex structures of coupled subprocesses of which each subprocess represent a detailed discipline in its own right, and the subprocesses interacts across a wide spectrum of space and time in a complicated manner, which ultimately determines the dynamics of the complete system.

Typically a model for a complex system is obtained by coupling deterministic sub-models together, where each sub-model describes a specific subprocess. The functional relations are therefore not based on the specific conditions observed at the study site. The output of the resulting model is compared to observations from the specific study-site and parameters are tuned to mimic observations in the ecosystem. An early and simple example of this approach is found in Fasham et al. (1990), a more resent and complex example is Bartell et al. (1999). The later example illustrates the profound complexity of ecosystem models.

The complexity of ecosystem models makes these especially useful for illustrating the methodology presented here, since the dynamics of the full system is not determined by the individual subprocess, but by the subprocesses *and* the way these are interconnected and working on different time scales.

Section 2 introduce the continuous-discrete time stochastic state-space formulation with emphasis on the transformations that enables estimation of system with state dependent noise. The proposed methodology is presented in Section 3, with the example constituting the main part of the article in Section 4. Finally the results from the example and some general implications of the proposed methodology are discussed in Section 5.

2. Continuous-discrete time stochastic state-space models

Stochastic differential equations are stochastic generalisations of ordinary differential equations in the sense that the deterministic skeleton of an SDE is an ODE. The continuous time state of the SDE is observed indirectly in discrete-time through the observation equation. This gives the continuous-discrete time stochastic state-space formulation

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)dt + \boldsymbol{\sigma}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)d\boldsymbol{w}_t \tag{1}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_{t_k}, \boldsymbol{u}_{t_k}, \boldsymbol{\theta}, \boldsymbol{e}_k, t_k), \tag{2}$$

where $t \in \mathbb{R}_0$ is time, t_k ($k \subset \mathbb{N}_0$) is the sample times, $\boldsymbol{x}_t \subseteq \chi \in \mathbb{R}^n$ is a vector of state variables belonging to the state-space (χ), $\boldsymbol{u}_t \in \mathbb{R}^r$ is a vector of inputs, $\boldsymbol{w}_t$ is the standard Brownian motion, $\boldsymbol{\theta} \in \mathbb{R}^p$ is a parameter vector, $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$ is a vector function referred to as the drift term, $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{m \times n}$ is a matrix function referred to as the diffusion term, $\boldsymbol{y}_k$ is the observation at time t_k, $\boldsymbol{h}(\cdot) \in \mathbb{R}^l$ is the observation function and $\boldsymbol{e}_k \in \mathbb{R}^l$ is a random observation error. Hence, the state-space formulation consist of the system equation (1), which describes the time-evolution of the states, and the observation equation (2), which describes the how the actual observations relates to the states.

The system equation (1) is a short-term notation for the integral interpretation, and in this context the Itô interpretation is used. Details on the formulation of SDEs and the general theory can be found in e.g. Øksendal (2003).

2.1. Parameter and state estimation

The estimation procedure employed here is based on the Extended Kalman Filter (EKF) and maximum likelihood estimation. A general account for the procedures can be found in Kristensen et al. (2004a), however the basic assumption is that the differential increments in Eq. (1) are Gaussian and that the observations are also Gaussian. For the filter equation to take on a sufficiently simple form to allow efficient implementation, the continuous-time stochastic state formulation and the discrete-time observation formulation is restricted to the form

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)dt + \boldsymbol{\sigma}(\boldsymbol{u}_t, \boldsymbol{\theta}, t)d\boldsymbol{w}_t \tag{3}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_{t_k}, \boldsymbol{u}_{t_k}, \boldsymbol{\theta}, t_k) + \boldsymbol{e}_k, \tag{4}$$

where $\boldsymbol{\sigma} \in \mathbb{R}^{n \times n}$ is a quadratic matrix function independent of the state, and $\boldsymbol{e}_k \in \mathbb{R}^l$ is a Gaussian random variable with zero mean and covariance $\boldsymbol{S}(\boldsymbol{u}_t, \boldsymbol{\theta}, t)$. All other terms are as explained above. The first restriction ($\boldsymbol{\sigma}$ quadratic) is not a real restriction since the estimation is based on the likelihood (weak solution) which, as a consequence of the fact that the Kolmogorov forward (Fokker-Planck) equation (Gard (1988)) only depends on $\boldsymbol{\sigma}\boldsymbol{\sigma}^T$. The independence between the diffusion matrix and the state can to some extent be dealt with by transformation of the state-space (see below) to obtain a formulation where the diffusion is independent of the state. The last restriction (observation noise additive and Gaussian), is crucial for the EKF, real life observations are, however, often not Gaussian, but it is often possible to deal with this by transformations of the observations before estimation (e.g. Box-Cox transformations).

The approximate likelihood estimation is based on the assumption that the conditional

density is Gaussian, and in this case the likelihood can be written as

$$L(\boldsymbol{\theta};\mathcal{Y}_N) = p(\boldsymbol{y}_0|\boldsymbol{\theta}) \prod_{k=1}^{N} \frac{\exp\left(-\frac{1}{2}\boldsymbol{\epsilon}_k^T \boldsymbol{R}_{k|k-1}^{-1}\boldsymbol{\epsilon}_k\right)}{\sqrt{(2\pi)^l \det(\boldsymbol{R}_{k|k-1})}}, \tag{5}$$

where $\mathcal{Y}_N = \{\boldsymbol{y}_0, ..., \boldsymbol{y}_N\}$ are all observations up to time $T = t_N$, $p(\boldsymbol{y}_0|\boldsymbol{\theta})$ is the conditional (on the parameters) density of the first observation, the innovation covariance matrix is given by $\boldsymbol{R}_{k|k-1} = V\{\boldsymbol{y}_k|\mathcal{Y}_{k-1};\boldsymbol{\theta}\}$ and the innovation is given by $\boldsymbol{\epsilon}_k = \boldsymbol{y}_k - E\{\boldsymbol{y}_k|\mathcal{Y}_{k-1};\boldsymbol{\theta}\}$. While the restriction of the observation equation (4) is necessary for (5) to form a reasonable approximation, it is not sufficient, since the observation equation (4) consist of a function of the state $h(\cdot)$ *and* an additive Gaussian error. The assumption is therefore that the conditional density of $\boldsymbol{x}_t$ is approximately Gaussian (possibly after a transformation h). This is likely to hold when the sampling frequencies are fast (compared to the dynamics of the system), while there exist methods to verify this (Bak et al. (1999)), we will not be concerned with this issue in the present study, since the evaluation of the final model is with respect to long term simulations not short term predictions (see Section 3).

In addition to the parameter estimation provided by the maximum likelihood procedure the filtering procedure allows for state estimation to obtain the state reconstruction ($E[\boldsymbol{x}_t|\mathcal{Y}_t]$), and the smoothed state ($E[\boldsymbol{x}_t|\mathcal{Y}_T]$), where $\mathcal{Y}_t$ is the information provided by observations up to time t. The estimation procedure is implemented in the open source software CTSM† (Kristensen et al. (2004a); Kristensen and Madsen (2003)).

2.2. Transformation of the state-space

As noted the diffusion matrix should be independent of the state in order to allow the filtering equation to be simple enough to allow efficient and numerically stable solutions. Transformations of SDEs is an application of Itô's Lemma (Øksendal (2003)), the special case where the transformed system has state independent diffusion is often referred to as the Lamperti transform (Iacus (2008); Luschgy and Pagés (2006)). For one dimensional processes this is well-known (Baadsgaard et al. (1997); Iacus (2008)) and the transformation is only limited by the ability to find an explicit expression for the inverse transformation (Iacus (2008)).

Unfortunately the generality of the Lamperti transform is restricted to one dimensional diffusion (Luschgy and Pagés (2006)). It is however possible to construct a Lamperti type transformation for a restricted class of diffusion processes (Luschgy and Pagés (2006)) given by

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)dt + \boldsymbol{\sigma}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)\boldsymbol{R}(\boldsymbol{u}_t, \boldsymbol{\theta}, t)d\boldsymbol{w}_t, \tag{6}$$

where $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n\times n}$ is a diagonal matrix, with diagonal elements $\sigma^{ii}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t) = \sigma^i(x_{i,t}, \boldsymbol{u}_t, \boldsymbol{\theta}, t)$ and $\boldsymbol{R}(\cdot) \in \mathbb{R}^{n\times n}$ is any matrix function (independent of $\boldsymbol{x}_t$). If z_i is chosen as

$$z_t^i = \psi^i(x_t^i, \boldsymbol{u}_t, \boldsymbol{\theta}, t) = \int \frac{d\xi}{\sigma^i(\xi, \boldsymbol{u}_t, \boldsymbol{\theta}, t)}\bigg|_{\xi=x_t^i}, \tag{7}$$

†Available at `www2.imm.dtu.dk/~ctsm`

then by Itôs lemma (e.g. Øksendal (2003)), z_t^i is also an Itô process given by

$$dz_t^i = \frac{\partial}{\partial t}\psi^i(\cdot,t)dt + \frac{\partial}{\partial \xi}\psi(\xi,\cdot)\bigg|_{\xi=x_t^i} dx_t^i + \frac{1}{2}\frac{\partial^2}{\partial \xi^2}\psi^i(\xi,\cdot)\bigg|_{\xi=x_t^i} (dx_t^i)^2 \tag{8}$$

$$= \left(\psi_t(\cdot,t) + \frac{f_i(\cdot)}{\sigma^i(\cdot)}dt - \frac{1}{2}\sigma_x^i(\cdot)\sum_{j=1}^{N}\left[\boldsymbol{R}(\cdot)\right]_{i,j}^2\right)dt + \sum_{j=1}^{N}\left[\boldsymbol{R}(\cdot)\right]_{i,j} dw_j, \tag{9}$$

where the diffusion term is independent of the state. The Lamperti transformation is essentially one dimensional (Luschgy and Pagés (2006)), which is also the construction applied here. The construction (9) involves the time derivative of $\boldsymbol{\sigma}(\cdot)$. In real life application such time dependence will often be through some observed input and the time-differentiation will involve numerical differentiation of the input. It is therefore recommended that time dependence in $\boldsymbol{\sigma}(\cdot)$ is avoided if possible. Aït-Sahalia (2008) provides a more general result than (6), but (6) is simpler to apply and will suffice for our purpose.

3. Description of methodology

The procedure proposed here is an iterative procedure, where each step is repeated until an acceptable model has been achieved. The procedure is divided into 4 steps (Figure 1), the aim of step 1-3 is to identify possible model improvement, this being either model extensions or model reductions. Traditionally model extensions is implemented by formulating a hypothesis based on mechanistic knowledge and hypothesis testing by e.g. likelihood ratio testing. However, inspection of pure random walk processes adapted to data by the EKF, for different parameters can also generate data driven hypotheses, which can be tested by conventional likelihood testing.

Maximum likelihood estimation is equivalent to optimisation of the one-step predictions (to the next available observation) error. However, if the model objective is different, such as a k-step prediction or simulation, then investigating model performance with respect to this objective may lead to model extensions, potentially different from those suggested from optimising the one-step prediction.

Step 0: Data considerations

In this initial step, before model development and estimation, the main question is, if the Gaussian assumption is fulfilled. If the Gaussian assumption of observation noise is not fulfilled observations should be transformed such that the observation equation has the required form (Eq. (4)). Other considerations could be outlier detection, data aggregation, etc. These considerations should be made a priori, since changes affect the iterations in step 1-3 and especially the classical statistical hypotheses testing depend critically on the transformation of data.

Step 1: Statistical inference

This is the classical statistical step where a candidate model is formulated and statistical testing is performed by comparing likelihoods or information criteria (e.g. AIC, BIC). Possible model reductions are considered in this step, even if a model reduction seems plausible

from a statistical point of view it might not be so from a modelling point of view, for instance a model reduction might lead to a model that does not fulfil basic model requirements (e.g. positive states). Additionally, some model reductions may prove unreasonable from a mechanistic understanding of the system in question, despite that statistical testing has rendered parameters equal to zero. In such cases it is preferable to maintain insignificant parameters in the model. Some of these steps are described in more details in Kristensen et al. (2004a).

If possible, model validation should also be performed by considering the autocorrelation function or generalisations like lag dependent functions (Nielsen and Madsen (2001)). These standard model validation tools are, however, not applicable for non-equidistant sampled data.

Since testing is based on one-step predictions only, rather than performance with respect to the required purpose, it might be appropriate to skip the testing part and go directly to the validation (Step 3), when model reformulation is based on considerations with respect to the required purpose.

Step 2: Structural identification

The initial model formulation will often be simpler than the complexity of the system suggests, and the challenge is to identify possible model extensions that lead to significant improvements of the model while avoiding over-parameterisation. One way to identify potential model deficiencies is to examine the diffusion term (Kristensen et al. (2004b)), because large diffusion coefficients indicate model deficiencies in the corresponding state. Examining the observation noise may similarly pinpoint model deficiencies. Although observation noise will always be present, large observation variance suggests that the observation bear no information or alternatively that the state equations do not sufficiently describe the dynamics of the observations. Thus, the diffusion and observation noise are both expected to be positive, but large parameter estimates give hints for model improvement.

The considerations above should lead to the selection of one parameter for further analysis. This parameter is formulated as a pure random walk hidden (unobservable) state denoted by $\theta_{i,t}$,

$$d\theta_{i,t} = \sigma_{\theta_i} dw_{\theta_i}. \tag{10}$$

The model is re-estimated with the random walk diffusion (σ_{θ_i}) added to the parameter vector, and θ_i replaced by the initial state of $\theta_{i,t}$ ($\theta_{i,0}$). Ideally σ_{θ_i} should be estimated, but for complicated models this might, as we will see in Section 4.2.3, lead to small estimates on the diffusion σ_{θ_i}. Although small estimates of the diffusion term indicates that the main interactions are captured by the model, it is advisable to fix the diffusion to a moderate value that allows regular and sufficient updating of the parameter to describe parameter variations over time. The time evolution of the random walk parameter should not show systematic variations with neither time nor other states, provided that the drift term is adequately modelled.

The smoothed state or state reconstruction for the random walk parameter is calculated using the estimation procedure described above (Section 2.1), and plotted against state variables and inputs to identify possible functional relationships. Non-parametric modelling tools such as generalised additive models (GAM) (Hastie and Tibshirani (1990)) can be employed as part of this identification approach, but the significance of these relationships has to be confirmed either by testing (Step 1) or validation (Step 3).

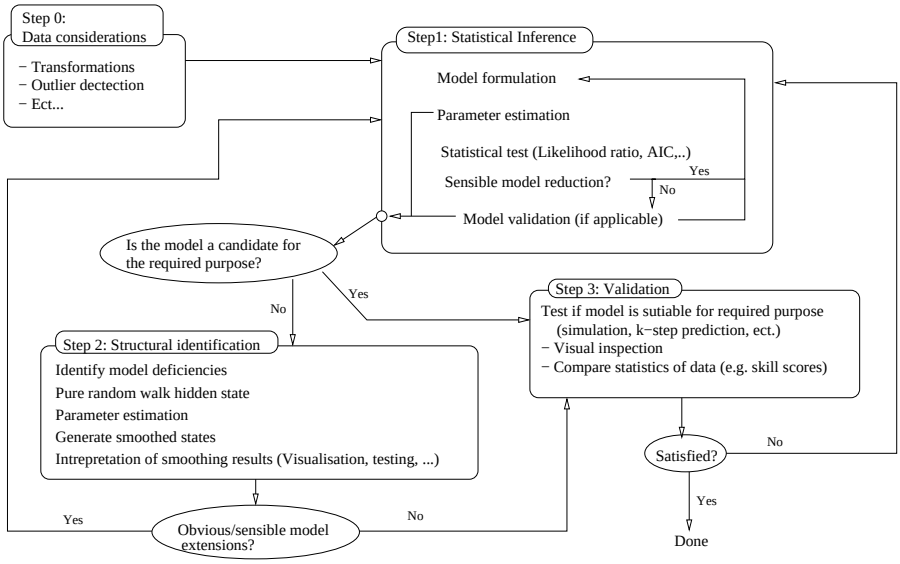

Fig. 1. Conceptual diagram of model development method.

Step 3: Validation

The model developed in Step 1 and 2 should be evaluated against its objective. However, the validation methodology depends on the objective, that is, if the objective is short-term prediction then the development tools presented in Step 1 and 2 are likely to produce satisfactory results, because the likelihood estimation procedure is based on one-step predictions. On the other hand, if the model objective is long-term prediction the model developed during Step 1-2 may prove inappropriate, since model components governing the long-term predictions may not significantly affect the short-term predictions and thus may not be included in the model formulation (e.g. parameters characterising higher order moments of the stationary distribution). The general applicability of a model developed with one specific objective in mind can be assessed by various methods of cross validation, see e.g. Hastie et al. (2001) for a general discussion or Madsen (2008) for a time series oriented discussion.

In the example presented below the objective is to develop a model suitable for simulations and in this context visual inspection of the state distributions is relevant and useful for the model development. Visual inspection is, however, subjective by nature and should be combined with some kind of objective skill score. In the presented example we employ a quantile skill score and interval skill score (Gneiting and Raftery (2007)) that assigns one number to the ability of the model to predict quantiles and confidence intervals.

4. Example: A multivariate Nitrogen-Phytoplankton model

The methodology presented above is applied to a simple two-state model for the interaction between water column nitrogen and phytoplankton, structural identification is used to generate hypothesis on primary production, these include both non-linear and multiplicative terms for nitrogen and light saturation. The model development is exemplified with

water quality data from the Danish estuary, Skive Fjord, and global radiation gauged in the vicinity of Skive Fjord.

4.1. Data sources and processing

Skive Fjord is a shallow estuary located in the northern part of Denmark, which has been extensively monitored during the Danish National Aquatic Monitoring and Assessment Program (DNAMAP), and from this monitoring program measurements of total nitrogen, chlorophyll as a proxy for phytoplankton biomass and primary production sampled weekly or biweekly were used in the present study. Freshwater discharge and nitrogen input from the entire Skive Fjord watershed, calculated as combination of measured and modelled inputs, were given with a monthly resolution. Finally, global radiation with an hourly resolution was provided by the Danish Meteorological Institute (DMI).

The temporal resolution of nitrogen and freshwater inputs was increased to daily values by means of piecewise linear functions to maintain total monthly inputs. Chlorophyll data (μg chla/l) was converted into kg N/m^3 using the standard carbon to chlorophyll weight ratio of 50:1, the Redfield ratio (C:N=106:16 (molar)), and primary production (mgC/m^2) was converted to kg N/m^3 by means of the Redfield ratio and the average depth of Skive Fjord (3.2 m).

The time series for global radiation had gaps and occasionally erroneous zero values during daytime, that were also treated as missing values. Missing values were filled by linear interpolation if the sequence of missing data was short (data available within two hours from the missing observation or available at the same time of day the day before and after). After this initial gap filling, longer sequences of missing observation were filled using a general harmonic function (including a diurnal and a daily seasonal component) fitted to data. The average daily global radiation, after completing all gaps, was used as input to the model. All data had pronounced seasonal variation, but also contained fast random variations, particularly evident for phytoplankton and primary production (Figure 2).

4.2. The multivariate stochastic differential equation model

The conceptual setting (Figure 3) is that Skive Fjord is enriched with nitrogen discharges from the surrounding watershed (N_{ex}), whereas atmospheric deposition is relatively smaller and neglected for this model development study. To maintain the water balance of the estuary nitrogen and phytoplankton are flushed out of the system depending on the freshwater inflow (Q). The estuarine circulation will lead to additional dilution that is contained in the general loss processes, that also include denitrification and burial in the sediment (a_{wl}). a_{wl} also describes other systematic effects such as the diffusive nitrogen exchange across the sediment-water interface. Measured global radiation is a proxy for the photo-synthetic active radiation (PAR) that sustains phytoplankton growth (a_{wp}), transforming inorganic nitrogen from the water column into organic biomass. Besides the flushing described above phytoplankton loss is assumed to be mediated through the water column nitrogen (a_{pw}). Clearly, this system is a rather coarse (lumped) simplification of the many complex processes taking place, but the idea is to focus on the primary production process and lump other processes to simple first-order approximations.

In the initial step the conceptual model (Figure 3) is transformed into the simplest possible mathematical formulation encapsulating the different mass flows, thereby minimising the risk of over-parameterisation and imposing false hypotheses based on the initial model.

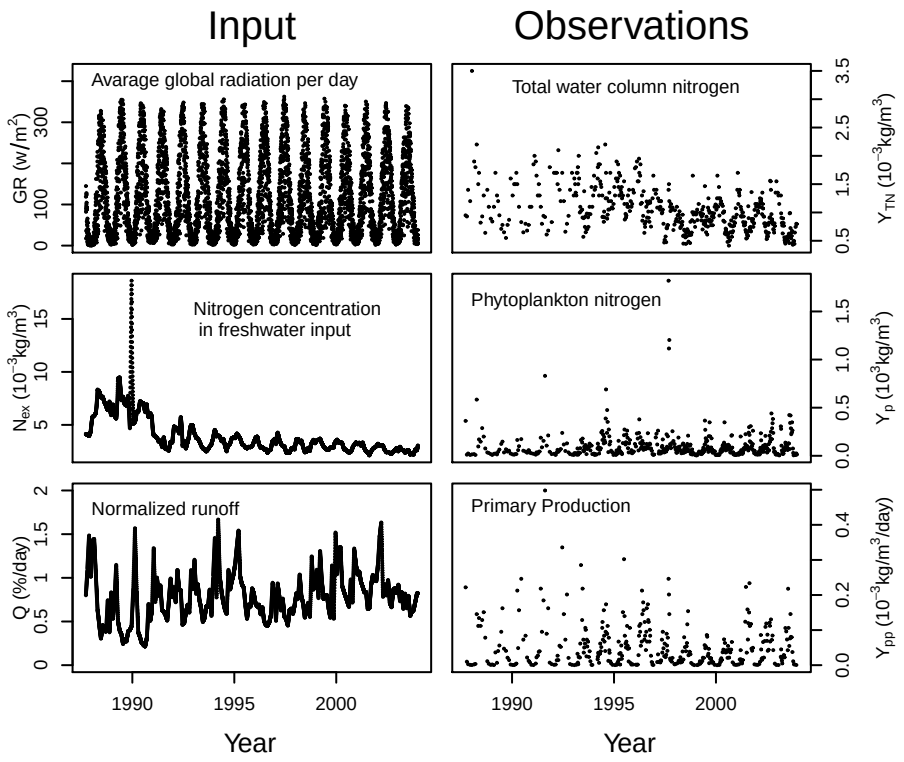

Fig. 2. Data used for modelling. Input and output variables in the left panel and right panels, respectively.

The minimal requirements of the mathematical formulation is that mass balance is maintained in the drift term, and that the state-space does not contain negative values. The simplest way to ensure this is by introducing noise proportional to the states, leading to the initial model formulation

$$d\begin{bmatrix} X_{w,t} \\ X_{p,t} \end{bmatrix} = \begin{bmatrix} N_{ex,t}Q_t \\ 0 \end{bmatrix} dt + \begin{bmatrix} -Q_t - a_{wp} - a_{wl} & a_{pw} \\ a_{wp} & -a_{pw} - Q_t \end{bmatrix} \begin{bmatrix} X_{w,t} \\ X_{p,t} \end{bmatrix} dt + \begin{bmatrix} \sigma_w X_{w,t} & 0 \\ 0 & \sigma_p X_{p,t} \end{bmatrix} \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} d\boldsymbol{w}_t, \tag{11}$$

where $X_{w,t}$ is the water column nitrogen not contained in phytoplankton, $X_{p,t}$ is phytoplankton nitrogen, $N_{ex,t}$ is the input of nitrogen from land, Q_t is the normalised (by the volume of Skive Fjord) freshwater input, $a_{wp}X_{w,t}$ ($a_{wp} > 0$ constant) is the primary production, $Q_t X_{i,t}$ is the flushing of nitrogen, $a_{wl}X_{w,t}$ (a_{wl} constant) is the loss/exchange of water column nitrogen through various processes besides primary production and phytoplankton mortality, $a_{pw}X_{p,t}$ ($a_{pw} > 0$ constant) is the phytoplankton mortality, $\sigma_i X_{i,t}$ ($\sigma_i > 0$) describes the system noise and r_{12} determines the noise correlation. The multiplicative noise ensures that the state-space of $\boldsymbol{X}_t = (X_{w,t}, X_{p,t})^T$ is strictly positive almost surely ($a.s.$) if $\boldsymbol{X}_0 > 0$.

Conservation of mass is not maintained in the diffusion term, and this seems reasonable

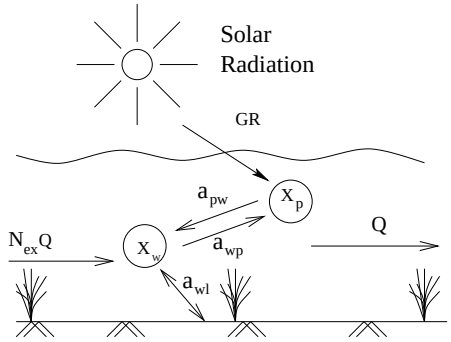

Fig. 3. Conceptual diagram of the model. The state variables are X_w water column nitrogen not contained in phytoplankton, and X_p is phytoplankton nitrogen. The forcing are nitrogen input $N_{ex}Q$, nitrogen loss QX_i and the parameters are phytoplankton mortality rate a_{pw}, phytoplankton birth process a_{wp} and interactions between water column nitrogen and other compartment e.g. the sediment.

because the lumped model for the loss/exchange processes in the drift term is too simple to describe in detail all the complex and interacting processes. Further, mass balance in the diffusion term would imply that random loss (gain) of phytoplankton biomass should appear in the water column, and would therefore be equivalent to only birth and death processes being stochastic. This is a quite strong assumption for a model that cannot be considered as a closed system. Additionally, mass balance in the diffusion term can to some extend be accounted for by the correlation r_{12}, and in this way it is tested by estimation.

As noted in Section 2.1, the estimation procedure does not allow state dependent diffusion, and therefore the system equation (11) is transformed according to the Lamperti transform (Eq. (7))

$$Z_{i,t} = \frac{\log(X_{i,t})}{\sigma_i} \quad \Rightarrow \quad X_{i,t} = e^{\sigma_i Z_{i,t}}, \tag{12}$$

the transformed system is given by

$$\begin{aligned} d\begin{bmatrix} Z_{w,t} \\ Z_{p,t} \end{bmatrix} = &\begin{bmatrix} \frac{1}{\sigma_w}\left(\frac{N_{ex,t}Q_t + a_{pw}X_{p,t}}{X_{w,t}} - (Q_t + a_{wl} + a_{wp})\right) & -\frac{1}{2}\sigma_w(1+r_{12}^2) \\ \frac{1}{\sigma_p}\left(a_{wp}\frac{X_{w,t}}{X_{p,t}} - (a_{pw} + Q_t)\right) & -\frac{1}{2}\sigma_p(1+r_{12}^2) \end{bmatrix} dt \\ &+ \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} d\boldsymbol{w}_t, \end{aligned} \tag{13}$$

where $X_{i,t}$ is a function of $Z_{i,t}$ defined by the inverse transformation in Eq. (12).

While the state-space of the original system (11) is $[0,\infty)\times[0,\infty)$ the state-space of the transformed system is $\mathbb{R}^2$, which is tractable from an estimation point of view. The system given in (13) is observed through a set of observation equations. Under the assumption that the observations are log-normal distributed around the expectation values, these are

$$\begin{bmatrix} Y_{TN,k} \\ Y_{p,k} \\ Y_{pp,k} \end{bmatrix} = \begin{bmatrix} \epsilon_{TN,k} & 0 & 0 \\ 0 & \epsilon_{p,k} & 0 \\ 0 & 0 & \epsilon_{pp,k} \end{bmatrix} \begin{bmatrix} X_{w,k} + X_{p,k} \\ X_{p,k} \\ a_{wp}X_{w,k} \end{bmatrix}, \tag{14}$$

where $Y_{TN,k}$ is observed total nitrogen in the water column, $Y_{p,k}$ is the observed phytoplankton nitrogen, and $Y_{pp,k}$ is the observed primary production, all at time $t = t_k$. In the log-domain we obtain the observation equation

$$\begin{bmatrix} \log(Y_{TN,k}) \\ \log(Y_{p,k}) \\ \log(Y_{pp,k}) \end{bmatrix} = \begin{bmatrix} \log(X_{w,k} + X_{p,t_k}) \\ \log(X_{p,k}) \\ \log(a_{wp} X_{w,k}) \end{bmatrix} + \begin{bmatrix} e_{TN,k} \\ e_{p,k} \\ e_{pp,k} \end{bmatrix}, \tag{15}$$

where $X_{w,k}$ and $X_{p,k}$ are described by the inverse transformation Eq. (12), and $e_{i,k} \sim N(0, s_i^2)$. In order to strengthen conclusions obtained in the structural identification step we have chosen to use all available data in for estimation and the validation described Section 4.3.1 is therefore based on the same data (but not on likelihood performance).

4.2.1. Model 1: The linear model

The parameters of the linear model Eqs. (13) and (15) were estimated. All parameters, except a_{pw}, display good estimation statistics (Table 1). Although a_{pw} is not statistically significant at this step of the modelling procedure, it is not advisable to remove it, since this would lead to a biologically meaningless model with a zero death rate. The estimated correlation coefficient is negative, implying that part of the randomness introduced by the diffusion is affecting the difference between primary production and mortality.

The model does not implicitly contain any other seasonal elements than the nitrogen input (N_{ex}) to describe the strong seasonality of the data displayed in Figure 2. The problem of how to identify possible improvements of the model Eqs. (11) and (14) in an objective way is the main theme of this paper. While the parameterisation of the diffusion term is not extremely important for the estimation problem as such, because the process is kept in place by the EKF, the diffusion term is a main driver for the distributional properties of the process and will be addressed in Section 4.4.1. The diffusion estimates (Table 1) highlights that the phytoplankton diffusion (σ_p) is approximately 10 times larger than the water column nitrogen diffusion (σ_w). Further, the observation noise of primary production (s_{pp}^2) is very large, which makes it reasonable to select the primary production parameter (a_{wp}) for a more thorough examination. This is also a plausible hypothesis from a biological point-of-view since primary production is traditionally modelled as a function of PAR and phytoplankton biomass as well as available nitrogen. For notational convenience, we will represent the primary production process by

$$a_{wp,t}^i = a_{wp}^0 X_{w,t} f_i(\boldsymbol{X}_t, \boldsymbol{u}_t), \tag{16}$$

where f_i is the functional expression to be identified (with $f_1(\boldsymbol{X}_t, \boldsymbol{u}_t) = 1$ in the linear model). In each step, i, a_{wp}^0 is replaced with $a_{wp,t}^0$ to identify a candidate model, $i + 1$ (replacing $f_i(\cdot)$ with $f_{i+1}(\cdot)$).

The primary production parameter ($a_{wp,t}^0$) is now modelled as a random walk process that will adapt to the data through the EKF. To ensure $a_{wp,t}^0 > 0$, $\forall t$ the random walk is introduced in the log domain, resulting in the following equation for the random walk parameter

$$d\log(a_{wp,t}^0) = \sigma_{a_{wp}} dw_{a_{wp},t}. \tag{17}$$

The parameters of this modified model description and the smoothened state $E(\boldsymbol{X}_t|\mathcal{Y}_T)$ (the mean of posterior distribution) from the EKF are estimated. A clear seasonal variation

Table 1. Estimation results for Model 1-6, bold face number refer to significant (on a 5% level) parameters, while number in parenthesis refer to standard deviation of parameter estimates.

	Unit	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
				Drift parameters			
a^0_{wp}	[a] d^{-1}	**0.016**	**0.159**	**0.361**	**0.376**	**0.362**	**0.405**
		(0.002)	(0.006)	(0.020)	(0.021)	(0.020)	(0.024)
a_{wl}	d^{-1}	**0.007**	**0.018**	**0.018**	**0.018**	**0.02**	**0.024**
		(0.003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
a_{pw}	d^{-1}	0.053	**0.124**	**0.319**	**0.310**	**0.322**	**0.379**
		(0.034)	(0.007)	(0.019)	(0.020)	(0.018)	(0.021)
k_w	$\frac{g}{m^{-3}}$			**0.067**	**0.106**	0.025	0.006
				(0.017)	(0.020)	(0.025)	(0.020)
k_{gr}	$\frac{W}{m^{-3}}$				**0.990**	**9.283**	**18.068**
					(0.244)	(3.247)	(3.360)
a_{p0}	$\frac{g}{m^3 \cdot d}$					**0.002**	**0.004**
						(0.001)	(0.001)
				Diffusion parameters			
σ_w	[b] $\frac{g}{m^3\sqrt{d}}$	**0.061**	**0.055**	**0.063**	**0.063**	**0.061**	**0.062**
		(0.004)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
σ_p	[b] $\frac{g}{m^3\sqrt{d}}$	**0.625**	**0.252**	**0.152**	**0.145**	**0.197**	**0.065**
		(0.085)	(0.015)	(0.008)	(0.009)	(0.017)	(0.012)
r_{12}		**−0.164**	**−0.135**	−0.057	−0.046	0.014	0.009
		(0.046)	(0.034)	(0.035)	(0.039)	(0.039)	(0.042)
γ_w							**0.662**
							(0.128)
γ_p							**0.546**
							(0.062)
				Variance of observation noise			
s^2_{TN}	$\frac{g^2}{m^6}$	**0.013**	**0.029**	**0.013**	**0.013**	**0.012**	**0.011**
		(0.002)	(0.004)	(0.002)	(0.002)	(0.002)	(0.002)
s^2_p	$\frac{g^2}{m^6}$	**0.304**	**0.141**	**0.172**	**0.185**	**0.198**	**0.205**
		(0.148)	(0.047)	(0.018)	(0.036)	(0.022)	(0.031)
s^2_{pp}	$\frac{g^2}{m^6 d^2}$	**3.546**	**2.193**	**1.174**	**1.126**	**0.851**	**0.715**
		(0.285)	(0.179)	(0.095)	(0.098)	(0.102)	(0.075)

[a] Unit does not apply to Model 2, where the unit is $\frac{m^3}{g \cdot d}$.

[b] Units does not apply to Model 6, where the unit of σ_i is $\left(\frac{g}{m^3\sqrt{d}}\right)^{-\gamma_i}$.

Table 2. Likelihood table or Model 1-6, column 1-4 refer to the model set up, column 5 report the log-likelihood, column 6 report the total number of degrees of freedom (including the initial state), while column 6-7 report AIC and BIC for all models and the last column reports the likelihood ratio test when applicable.

	f	a_{p0}	γ_i	$\log(L)$	DF	AIC	BIC	$P(x > -2Dl)$
Model 1	$f_1 = 1$	0	1	-1446	11	2914	2971	
Model 2	$f_2 = f_1 X_{p,t}$	0	1	-1374	11	2770	2828	
Model 3	$f_3 = \frac{f_2}{k_w + X_{w,t}}$	0	1	-1079	12	2182	2245	
Model 4	$f_4 = \frac{f_3 GR_t}{k_{gr} + GR_t}$	0	1	-1067	13	2159	2227	
Model 5	$f_5 = f_4$	free	1	-1007	14	2042	2115	0.0000
Model 6	$f_6 = f_4$	free	free	-986	16	2004	2088	0.0000

in $\hat{a}^0_{wp,t|T}$ (Figure 4A) as well as evident correlations with phytoplankton nitrogen ($X_{p,t}$), water-column nitrogen ($X_{w,t}$), and global radiation GR_t (Figure 5A-C) emerges, however most pronounced with phytoplankton. This is also confirmed by comparing AIC for linear models of the hypothesised relationships

$$\hat{a}^0_{wp,t|T} = \alpha_{gr,0} + \alpha_{gr,1} Gr_t + \epsilon_{Gr,t} \qquad AIC = -22928 \tag{18}$$

$$\hat{a}^0_{wp,t|T} = \alpha_{p,0} + \alpha_{p,1} \hat{X}_{p,t|T} + \epsilon_{p,t} \qquad AIC = -28410 \tag{19}$$

$$\hat{a}^0_{wp,t|T} = \alpha_{w,0} + \alpha_{w,1} \hat{X}_{w,t|T} + \epsilon_{w,t} \qquad AIC = -23998. \tag{20}$$

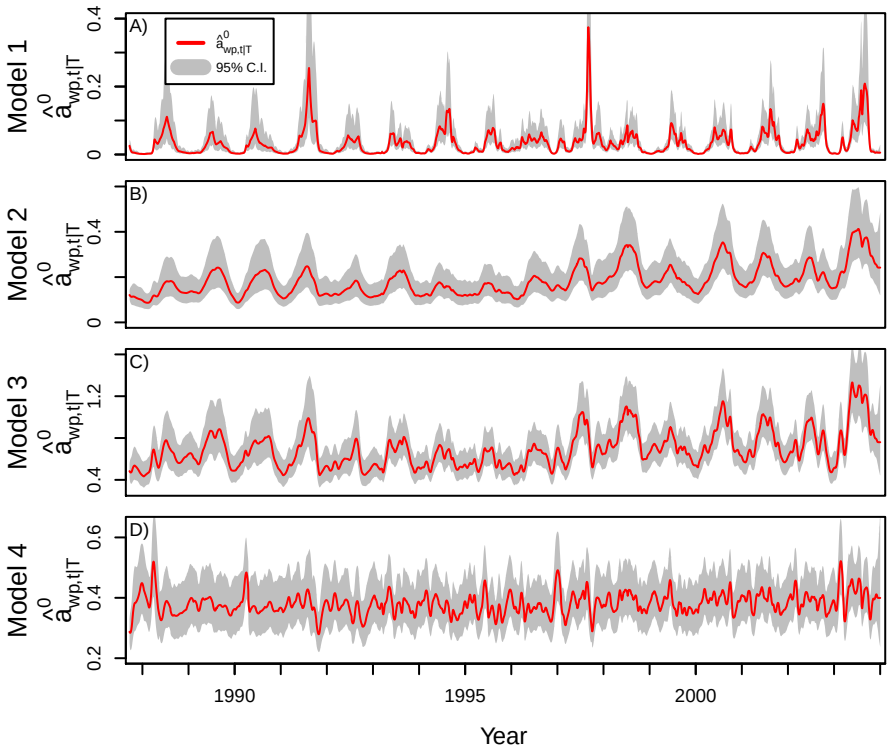

Fig. 4. Smoothened state of the random walk phytoplankton growth parameter ($\hat{a}^0_{wp,t|T}$) in Model 1-4 as function of time. 95% confidence interval (grey area) is base on a Gaussian assumption of $\log(\hat{a}^0_{wp,t|T})$. Red lines represent the median of $\hat{a}^0_{wp,t|T}$.

4.2.2. Model 2: Including phytoplankton

Based on the statistics in Eqs. (18)-(20) and Figure 5A-C phytoplankton nitrogen is included in the production process and hence $f_2(\cdot)$ (in Eq. (16)) is modelled as

$$f_2(\boldsymbol{X}_{t_k}, \boldsymbol{u}_t) = X_{p,t}. \tag{21}$$

Model 1 is not a proper subset of Model 2 implying that likelihood ratio testing is not valid. We therefore base the model evaluation on AIC and BIC, improvements in both criteria in the order of 140 (Table 2) are seen. All parameters in this model formulation, including the death rate, are well determined (Table 1), though the phytoplankton diffusion (σ_p) and primary production observation noise (s^2_{pp}) are still large compared to the diffusion of X_w, despite that both decreased.

To address these large errors the primary production parameter is again modelled as a random walk and plotted as a function of the potential explanatory variables and time (Figures 5D-F and 4B). A clear seasonal variation (Figure 4B) still remains as well as evident correlation between $\hat{a}^0_{wp,t|T}$ and the state variables and global radiation (Figure 5D-F). It is also seen that linear relations would be a poor fit (Figure 5D-F), and the following hypotheses are therefore considered

$$H_1 : a_{wp} = \frac{Gr_t}{k_{gr} + Gr_t} \tag{22}$$

$$H_2 : a_{wp} = \frac{X_{p,t}}{k_p + X_{p,t}} \tag{23}$$

$$H_3 : a_{wp} = \frac{1}{k_w + X_{w,t}}. \tag{24}$$

H_1 is based on Figure 5F and the well-known fact that light saturation occurs for primary production, although the parametric description here is simpler than the usual parameterisation of light saturation (e.g. Fasham et al. (1990)). H_2 is based on empirical evidence only (Figure 5D), while H_3 is based on Figure 5E and Michaelis-Menten kinetics for nitrogen. These hypotheses are transformed into linear relations and the best relationship based on AIC is chosen

$$\frac{1}{a_{wp,t}} = \alpha_{gr,1}\frac{1}{Gr_t} + \alpha_{gr,2}Gr_t + \epsilon_{Gr,t} \qquad AIC = 34812 \tag{25}$$

$$\frac{1}{a_{wp,t}} = \alpha_{p,1}\frac{1}{X_{p,t}} + \alpha_{p,1}X_{p,t} + \epsilon_{p,t} \qquad AIC = 30108 \tag{26}$$

$$\frac{1}{a_{wp,t}} = \alpha_{w,0} + \alpha_{w,1}X_{w,t} + \epsilon_{w,t} \qquad AIC = 15040. \tag{27}$$

The lowest AIC was the formulation with water column nitrogen Eq. (27), which is also most apparent from the less scatter in Figure 5E.

4.2.3. Model 3: Including nitrogen saturation

Based on the statistics in Eqs. (25)-(26) and Figure 5D-F, water column nitrogen is included in the model, and the primary production process is reformulated as

$$f_3(\boldsymbol{X}_{t_k}, \boldsymbol{u}_t) = \frac{X_{p,t}}{k_w + X_{w,t}}. \tag{28}$$

This model, now including nitrogen saturation of primary production, was first fitted without random walk parameters (Table 1). All parameters, except the correlation parameter, were significant, even though the statistics suggest to remove the correlation parameter, it is maintained in the model and the model reduction step is postponed to the final model.

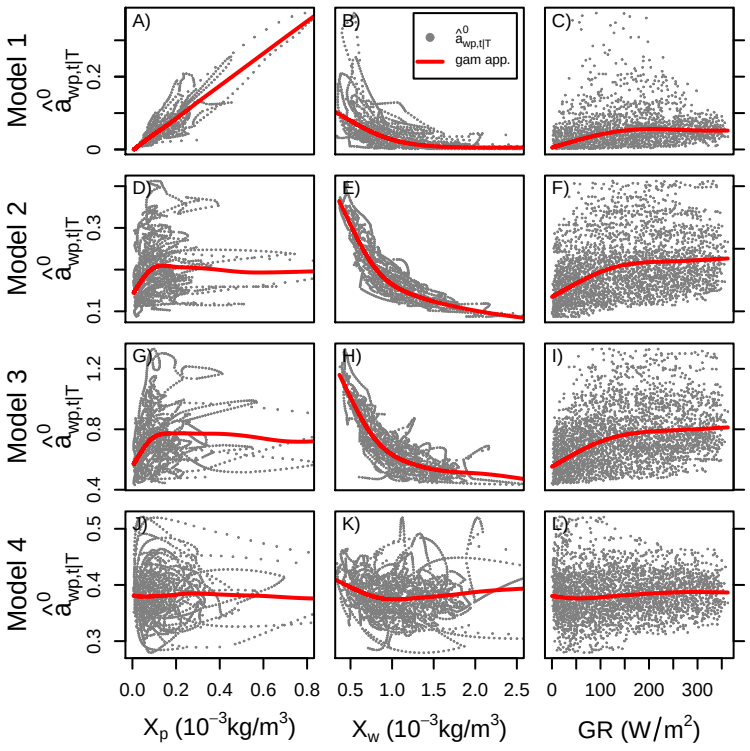

Fig. 5. Smoothened state of the random walk phytoplankton growth parameter ($\hat{a}^0_{wp,t|T}$) in Model 1-4 as a function of smoothen phytoplankton level (first column), smoothen water column nitrogen (second column), observed global radiation (third column). Grey dots represent point estimation of $\hat{a}^0_{wp,t|T}$, while red lines represent GAM (smoothing splines) fit to the points.

Since Model 2 is not a proper subset of Model 3 model evaluation is again based on AIC and BIC, improvements was about 600 for both criteria (Table 2). It should also be stressed that the diffusion of X_p and the observation noise of primary production (s^2_{pp}) both decreased.

The estimated random walk diffusion for the primary production parameter ($a^0_{wp,t}$) almost disappeared ($2.4 \cdot 10^{-4}$) after this model change, and there was no seasonal variation in the random walk, despite an anticipated seasonal pattern yet uncovered in the model formulation. To address this artefact from the estimation procedure and allow regular updating of $a^0_{pw|t}$, the diffusion for the random walk parameter was fixed to 0.05, which is comparable to the diffusion of the water column diffusion parameter. Following this modification the strongest relationship to the random walk parameter is still with water column nitrogen (Figure 5H), however, it is also evident that the random walk parameter was related to global radiation (Figure 5I). As the intention is to build a model that is well suited for simulation, it is necessary to include a seasonal input, but also cross correlation between X_w and GR might influence the results seen in Figure 5G-I. Therefore global radiation is included, acknowledging that the relationship between $a^0_{pw|t}$ and GR is clearly not linear, the simple light saturation function given in Eq. (22) is chosen.

4.2.4. Model 4: Including global radiation

Based on the reasoning above $f_4(\cdot)$ is chosen as

$$f_4(\boldsymbol{X}_{t_k}, \boldsymbol{u}_t) = \frac{Gr_t X_{p,t}}{(k_{gr} + Gr_t)(k_w + X_{w,t})}. \tag{29}$$

Again all parameters except for the correlation coefficient are clearly significant (Table 1). The phytoplankton diffusion (σ_p) and primary production observation noise (s_{pp}^2) did not decrease (Table 1), but the improvements of AIC and BIC is about 20 (Table 2).

The primary production parameter is again modelled as a random walk with fixed diffusion parameter equal 0.05. There is no strong correlation between the states and the input (Figure 5J-L), and the distinctive seasonal pattern observed in Figure 4A-C has disappeared (Figure 4D). As the developed model is a potential candidate for simulation studies, we continue to Step 3 in the model development procedure to validate if the model is suitable for simulation.

4.3. Simulations

The modelling so far has focused on formulation hypothesis based on pure random walk primary production parameter, likelihood testing and information criteria. Likelihood testing is equivalent to optimisation of the ability to predict the observation at time t_{k+1} given the information up to time t_k. The time between water quality samples vary from 4 to 94 days with an average sampling time ranging from 11.5 to 16 days for the different water quality variables. For simulation studies the objective of the model is to predict perhaps one or several years ahead. Further the aim of a simulation is not only to predict one value (like the mean) of future states, but to predict the distribution of the future states. This imply that we need to get both the drift and the diffusion term right. The simulation studies in the following is based on the Euler scheme (Kloeden and Platen (1999)) applied to the transformed process with $\Delta t = \frac{1}{96}d$. The time step Δt is decreased until the simulation plots (like Figure 6) do not change.

4.3.1. Simulation of Model 4

Model 4 was simulated over the entire span of the dataset, with the initial state drawn from a Gaussian distribution around the smoothened state, having a mean and variance equal to that of the smoothen state (in the transformed domain). The simulated water column nitrogen seems to be well captured (Figure 6A and D), while phytoplankton (Figure 6G and J) and primary production (not shown, but similar to the phytoplankton plot) perform poorly, with the simulated distribution drifting away from the observations. To address this problem, we reconsider the developed phytoplankton equation

$$\begin{aligned} dX_{p,t} = &\left(a_{wp} \frac{X_{w,t} Gr_t}{(k_w + X_{w,t})(k_{gr} + Gr_t)} - Q_t - a_{pw} \right) X_{p,t} dt \\ &+ \sigma_p X_{p,t} (r_{12} dw_{1,t} + dw_{2,t}). \end{aligned} \tag{30}$$

The stationary distribution (considering X_w, Q_t and GR_t as constants) for this equation is either 0 or ∞ depending on the factor in front of $X_{p,t}dt$. Such a behaviour is clearly not

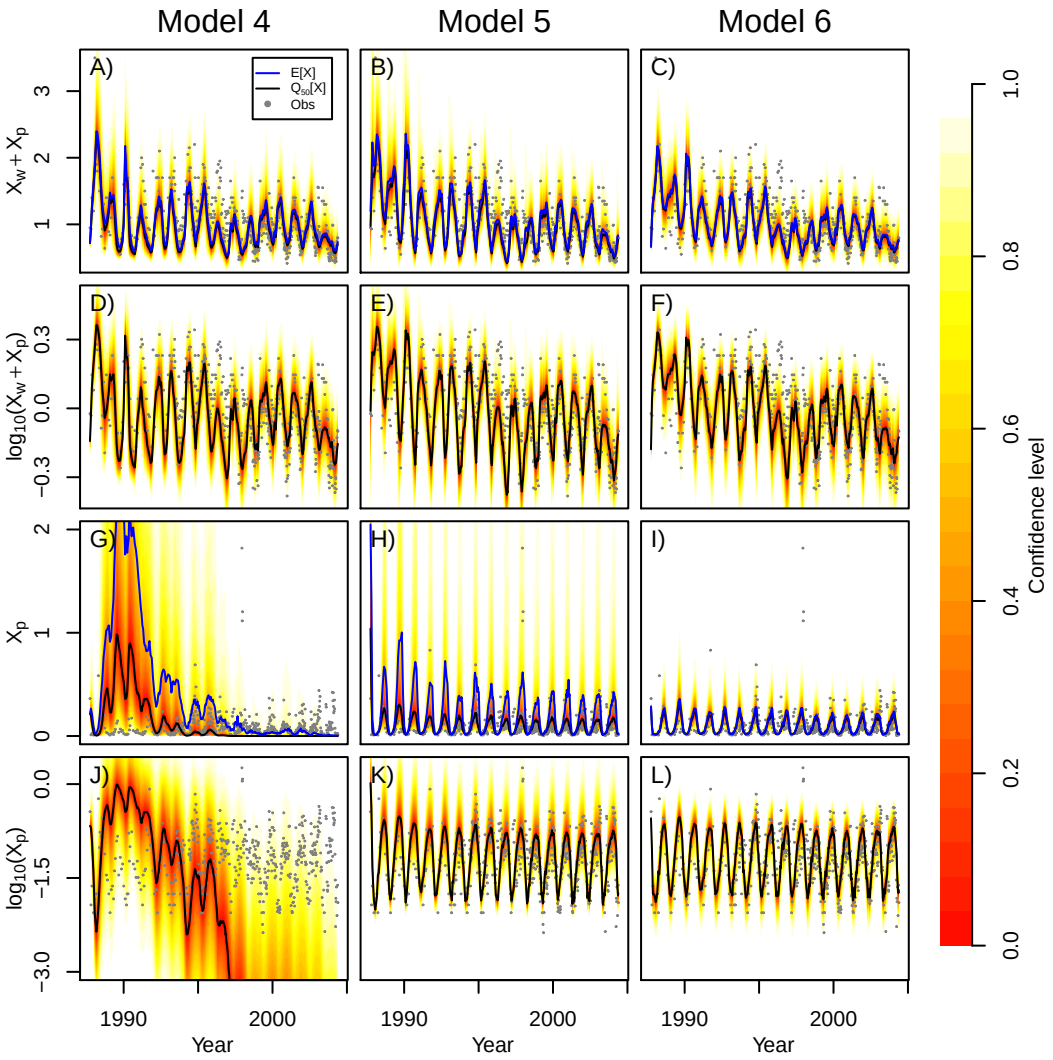

Fig. 6. Simulation for the time span of observations with Model 4-6, colour code refer to confidence intervals around the median of the distribution, while gray dots are the measurements.

desirable and is the main course of the drift seen in Figure 6G and J. In order to solve this problem we add a constant ($a_{p0} > 0$) to the equation to get

$$\begin{aligned} dX_{p,t} =& a_{p0}dt + \left(a_{wp}\frac{X_{w,t}Gr_t}{(k_w + X_{w,t})(k_{gr} + Gr_t)} - Q_t - a_{pw} \right) X_{p,t}dt \\ &+ \sigma_p X_{p,t}(r_{12}dw_{1,t} + dw_{2,t}). \end{aligned} \tag{31}$$

The argumentation for the constant a_{p0} is clearly mathematical convenience, but we can think of a_{p0} as an inoculum. Also if the constant is small it will not influence (local) phytoplankton growth greatly.

4.4. Model 5: Including a constant inoculum

The model with a constant inoculum factor is also well determined (Table 1) and the likelihood improved by 59.9 on 1 degree of freedom (Model 4 is a proper subset of Model 5), which is significant with $p << 0.0001$. To evaluate if a_{p0} is small, imagine that no water column nitrogen is present and that $Q_t = 0$, then the mean value (of the stationary distribution) of the phytoplankton would be $\frac{a_{p0}}{a_{pw}} = 6.8 \cdot 10^{-3} \frac{g}{m^3}$ (Iacus (2008); Forman and Sørensen (2008)), which is low compared to the observed values (about 0.7% of the observed phytoplankton nitrogen is below this value). The nitrogen saturation k_w constant is not significant in t-test (Table 1), and it is low compared to the observed values of water column nitrogen, it is however comparable to the saturation constant used by Fasham et al. (1990) (0.007 gNm^{-3}) and the reported saturation constant by Fisher et al. (1992) (0.028 gNm^{-3}). Furthermore if k_w is removed the model would be able to produce undesirable negative values of water column nitrogen, and k_w is therefore kept in the model.

An important test of Models 5 is its behaviour in a simulation study. The annual variations of the phytoplankton state is captured much better (Figure 6H and K), however the model does predict very large mean values for phytoplankton (Figure 6H) and primary production (not shown, but similar to the phytoplankton plot), and the confidence intervals are also large, with values exceeding the largest observations.

4.4.1. Model 6: Analysis of diffusion

These wide confidence intervals obtained with Model 5 could be due to the diffusion scaling too fast with the state. In order to analyse this question, we replace the diffusion term, in Eq. (11), with

$$\begin{bmatrix} \sigma_w X_{w,t}^{\gamma_w} & 0 \\ 0 & \sigma_p X_{p,t}^{\gamma_p} \end{bmatrix} \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}. \tag{32}$$

The coefficients γ_w and γ_p are commonly chosen as either 1 (Øksendal (2003)) or 0.5 (Klebaner (2005)) in biological models. For γ_i equal to 0.5 the linear one-dimensional process is known as Feller diffusion in biology (Klebaner (2005)) or CIR model in finance (Iacus (2008)). It has a positive probability of reaching zero if the loading parameter is small compared to the diffusion parameter (Iacus (2008)), while for γ_i larger than 1 the system does not fulfil the linear growth condition, and existence and uniqueness is not guaranteed (Øksendal (2003)).

The Lamperti transform presented in Eq. (12) needs to be reformulated as

$$Z_{i,t} = \frac{X_{i,t}^{1-\gamma_i}}{\sigma_i(1-\gamma_i)} \quad \Rightarrow \quad X_{i,t} = [\sigma_i(1-\gamma_i)Z_{i,t}]^{\frac{1}{1-\gamma_i}}. \tag{33}$$

For $\gamma_i \in (0,1)$ the state-space of $Z_{i,t}$ is equal to the state-space of $X_{i,t}$ ($[0,\infty)$). The

transformed system is given by

$$\begin{aligned} dZ_{w,t} =& \frac{X_{w,t}^{-\gamma_w}}{\sigma_w}\left(N_{ex,t}Q_t - \left(Q_t + a_{wl} + a_{wp}^0 f_4(\boldsymbol{X}_t, \boldsymbol{u}_t)\right) X_{w,t} + a_{pw}X_{p,t}\right) dt \\ & - \frac{1}{2}\sigma_w\gamma_w X_{w,t}^{\gamma_w - 1}(1 + r_{12}^2)dt + dw_1 + r_{12}dw_2 \end{aligned} \tag{34}$$

$$\begin{aligned} dZ_{p,t} =& \frac{X_{p,t}^{-\gamma_p}}{\sigma_p}\left(a_{p0} + a_{wp}^0 f_4(\boldsymbol{X}_t, \boldsymbol{u}_t)X_{w,t} - (a_{pw} + Q_t)X_{p,t}\right) dt \\ & - \frac{1}{2}\sigma_p\gamma_p X_{p,t}^{\gamma_p - 1}(1 + r_{12}^2)dt + r_{12}dw_1 + dw_2. \end{aligned} \tag{35}$$

Rearranging and using the short hand notation, $\tilde{\theta}_{w,0} = (N_{ex,t}Q_t + a_{pw}X_{p,t})/\sigma_w$, $\tilde{\theta}_{w,1} = (Q_t + a_{wl} + a_{wp}^0 f_4(\boldsymbol{X}_t, \boldsymbol{u}_t))/\sigma_w$, and $\tilde{\theta}_{w,2} = \frac{1}{2}\sigma_w\gamma_w(1 + r_{12}^2)$ gives the following SDE for the water column nitrogen

$$dZ_{w,t} = X_{w,t}^{-\gamma_w}\left(\tilde{\theta}_{w,0} - \tilde{\theta}_{w,1}X_{w,t} - \tilde{\theta}_{w,2}X_{w,t}^{2\gamma_w - 1}\right) dt + dw_1 + r_{12}dw_2. \tag{36}$$

Now consider the limit where $X_{w,t} \to 0$, for $\gamma_w \neq \frac{1}{2}$ we get

$$\lim_{X_w \to 0} dZ_{w,t} = \lim_{X_w \to 0} \begin{cases} -\tilde{\theta}_{w,2}X_{w,t}^{\gamma_w - 1} & = -\infty \quad a.s. \quad \gamma_w < \frac{1}{2} \\ \tilde{\theta}_{w,0}X_{w,t}^{-\gamma_w} & = +\infty \quad a.s. \quad \gamma_w > \frac{1}{2}. \end{cases} \tag{37}$$

For $\gamma_w = \frac{1}{2}$ the limit splits into three cases

$$\lim_{X_w \to 0} dZ_{w,t} = \lim_{X_w \to 0} \begin{cases} (\tilde{\theta}_{w,0} - \tilde{\theta}_{w,2})X_{w,t}^{-\frac{1}{2}} & = -\infty \quad a.s. \quad \tilde{\theta}_{w,0} < \tilde{\theta}_{w,2} \\ dw_1 + r_{12}dw_2 & \qquad\qquad\quad \tilde{\theta}_{w,0} = \tilde{\theta}_{w,2} \\ (\tilde{\theta}_{w,0} - \tilde{\theta}_{w,2})X_{w,t}^{-\frac{1}{2}} & = +\infty \quad a.s. \quad \tilde{\theta}_{w,0} > \tilde{\theta}_{w,2}. \end{cases} \tag{38}$$

These considerations show that the transformed system is not consistent with the topology of the state-space when $\gamma_w < \frac{1}{2}$, and for $\gamma_w = \frac{1}{2}$ the transformed system is only consistent with the topology of the state-space for a restricted set of parameter values ($\tilde{\theta}_{w,0} > \tilde{\theta}_{w,2}$). The reasons for these inconsistencies is that the transformation is only valid inside the state-space and not on the boundary. The parameter set should reflect that $P(X_{w,t} = 0) = 0$. Clearly the simplest way to ensure this is by restricting γ_w to the interval $(0.5, 1)$. Similar arguments apply to the phytoplankton equation, and γ_w is therefore also restricted to the interval $(0.5, 1)$.

The parameters were estimated with $\gamma_i \in (0.5, 1)$. Most parameters are well determined (Table 1) (with the exception of the correlation coefficient and nitrogen saturation), and further the diffusion coefficient of X_p and X_w have comparable sizes. The simulation study (Figure 6 third column) shows that the confidence intervals for phytoplankton nitrogen has narrowed (Figure 6I and L) and that phytoplankton mean is now close to the median of the distribution. The consequence is that more extreme values are not included in the distribution of the simulations.

4.5. Quantification of simulation performance

The purpose of simulation studies like the ones presented in Figure 6 is to predict the distribution of the future state of the system. One way of quantifying this analysis is to

compare simulation quantiles with observed data. The visual inspection of the simulation results (Figure 6) is clearly relevant and led to the introduction of a_{p0} and γ_i. However, to quantify the model skills we need to represent the performance by a single number, in the same way as the likelihood values represent the overall quality of short-term predictions.

A good model candidate should 1) be reliable meaning the quantiles of the estimated distribution should hold the right proportion of the data (often referred to as reliability) and 2) have narrow confidence regions (often referred to as sharpness (Gneiting and Raftery (2007)). In addition, the ability to adapt to different uncertainty regimes is sometimes considered (often referred to as resolution). Though misleading results may emerge if sharpness and resolution are considered only and reliability is not taken into account (Pinson et al. (2007); Møller et al. (2008)). A proper skill score for quantiles should therefore combine these objectives into one single number. One such skill score is (Gneiting and Raftery (2007))

$$S(r_1,..r_k;x) = \sum_{i=1}^{k} (\alpha_i s_i(r_i) + (s_i(x) - s_i(r_i))\mathbb{I}\{x \leq r_i\}) + h(x), \tag{39}$$

where x is the observation, r_i is the quantile predicted by the simulation, s_i is non-decreasing and h is arbitrary. Here we choose $s_i(x) = x$ and $h(x) = -\sum_i \alpha_i x$, and in this case Eq. (39) is the sum of loss functions for quantile regression as defined by Koenker and Basset (1978). Model 6 perform consistently better than Model 4 and 5 when comparing the skill score for all observations (Table 3). Based on these statistics Model 6 is the better choice of model.

To evaluate individual confidence intervals the (negatively oriented) interval score (Gneiting and Raftery (2007)) is calculated by

$$S_\alpha^{int}(l,u;x) = (u-l) + \frac{2}{\alpha}(l-x)\mathbb{I}\{x < l\} + \frac{2}{\alpha}(x-u)\mathbb{I}\{x > u\}, \tag{40}$$

where $[l,u]$ is the confidence interval and α is the nominal coverage. As noted by Gneiting and Raftery (2007) this is intuitively appealing, since sharpness is formulated directly $(u-l)$ and values outside the interval are penalised. Model 6 performs consistently better for total nitrogen and primary production (Figure 7), Model 5 performs slightly better for high confidence levels of phytoplankton (above 0.6) while Model 6 performs better for confidence levels below 0.6. The combined conclusion is therefore that Model 6 is the preferred candidate for the final model.

The better performance of Model 5 compared to Model 6, in terms of nominal coverage of phytoplankton nitrogen (Figure 7), reflects that extreme values are not well described in Model 6. This is not surprising as neither Model 6 nor any of the other models contain any mechanisms to specifically capture these extreme events. To capture these extremes, information on e.g. oxygen depletion would be needed, but oxygen depletion events are governed by local weather conditions such as wind, which cannot be predicted long time ahead. It is therefore not realistic to predict extremes with a simulation model that runs for several of years.

4.6. Model reductions

Model 6 is the candidate for the final model formulation, although the results in Table 1 suggest two potential model reductions. Nitrogen half saturation (k_w) is not significant, however, as discussed above, it is not reasonable to remove this parameter since it will lead

Table 3. Quantile skill scores and quantile skill score differences for the simulation models. Skill scores are calculated in the log-domain, and index refer to observations.

	S_{TN}	DS_{TN}	S_p	DS_p	S_{pp}	DS_{pp}
Model 4	-8.96		-214.32		-187.96	
Model 5	-8.98	-0.02	-24.00	190.32	-30.34	157.62
Model 6	-8.39	0.60	-23.65	0.35	-28.29	2.05

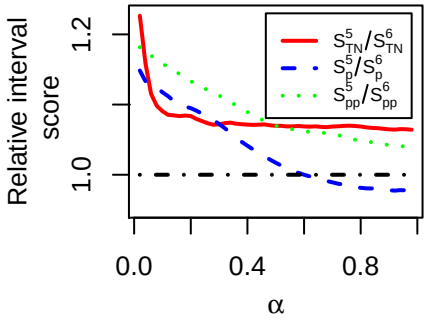

Fig. 7. Relative interval quantile skill score for simulation models 5 and 6; α refer to nominal coverage. Values above 1 indicate better performance of Model 6 while values below 1 indicate better performance of Model 5.

to a positive probability of reaching negative values for water column nitrogen. It should be noted that an attempt to estimate parameters in a Model 6 with $k_w = 0$ failed, because the optimiser tested values that led to negative values of water column nitrogen.

The correlation coefficient (r_{12}) is also not significant, and consequently this parameter is removed and the model re-estimated, yielding a log-likelihood decrease of 0.023, which corresponds to a p-value of 0.83. Furthermore, the relative (to the standard deviation) changes of the parameters are all less than 0.1. Thus the final model consists of the syatem equation

$$d\begin{bmatrix} X_{w,t} \\ X_{p,t} \end{bmatrix} = \begin{bmatrix} N_{ex,t}Q_t - (Q_t + a_{wl})X_{w,t} - \frac{a^0_{wp}X_{w,t}X_{p,t}GR_t}{(k_w+X_{w,t})(k_{gr}+GR_t)} + a_{pw}X_{p,t} \\ a_{p0} + \frac{a^0_{wp}X_{w,t}X_{p,t}GR_t}{(k_w+X_{w,t})(k_{gr}+GR_t)} - (a_{pw} + Q_t)X_{p,t} \end{bmatrix} dt + \begin{bmatrix} \sigma_w X^{\gamma_w}_{w,t} & 0 \\ 0 & \sigma_p X^{\gamma_p}_{p,t} \end{bmatrix} d\boldsymbol{w}_t, \tag{41}$$

and the observation equation

$$\begin{bmatrix} \log(Y_{TN,k}) \\ \log(Y_{p,k}) \\ \log(Y_{pp,k}) \end{bmatrix} = \begin{bmatrix} \log(X_{w,k} + X_{p,t_k}) \\ \log(X_{p,k}) \\ \log(a_{wp}X_{w,k}) \end{bmatrix} + \begin{bmatrix} e_{TN,k} \\ e_{p,k} \\ e_{pp,k} \end{bmatrix}. \tag{42}$$

4.7. Discussion of the biological model

The model development presented here is based primarily on mathematical and statistical reasoning, while the biological/physical reasoning is mostly used in the initial model

formulation. The model development based on visual inspection of the (smooth) path of the primary production parameter is, however, consistent with existing knowledge from N-P models. Identified model extensions based on the random walk hidden states are all significant, as are the extensions based on considerations of the simulated or stationary distribution.

Clearly, the presented model is very simple and represents a coarse simplification of the complex ecosystem, and a higher dimensional state-space would be needed to give a realistic description owing to all the known processes. The simplicity is a deliberate choice, since the focus is placed on demonstrating the model development and the significant role of the diffusion term. The structural identification and the considerations of the simulated distributions are, however, not limited to simple systems and the identification methodology can principally be applied directly to models with a higher number of state variables as well. Further, the probabilistic formulation of state transitions also lump weak interactions that may otherwise be formulated in a complicated ODE model, often with a large degree of uncertainty. The SDE setting provides a direct quantification of the uncertainty, which cannot be estimated directly from ODE models. The model development does not include explicit knowledge about the parameters, and prior knowledge is only included implicitly in the hypothesis formulations.

The key importance of the diffusion term is illustrated by the large improvements in the likelihood criteria, when introducing the exponents γ_i, stressing the importance of the specific parametric formulation of diffusion term. One of the main results of the model development is the large reductions of the confidence band width of the simulated densities obtained by introducing γ_i. Actually, a hypothesis of $\gamma_i = \frac{1}{2}$ cannot be rejected using unconditional t-tests for each of the exponents, and the results therefore support the hypothesis that the diffusion is of the Feller type. It is, however, argued that estimating γ_i in the open interval $\left(\frac{1}{2}, 1\right)$ is robust from a estimation point of view, but just as important is that strictly positive states included in the drift term are maintained in the stochastic formulation when $\gamma_i > \frac{1}{2}$.

The model focuses mainly on primary production, whereas loss processes are lumped together. The partitioning of the loss term in the model could be carried out following the same methodology as presented for the production process, provided that explanatory information or observations needed to describe the different terms (e.g. zooplankton and filter feeder biomass) is available. This will probably result in a more variable set of loss functions as opposed to the constant loss rate (a_{wl}) used in the present context. Another issue is that the model does not distinguish between labile and non-labile nitrogen for phytoplankton growth, and a better partitioning of the nitrogen pool may lead to further model improvements. All these potential improvements imply additional states in the system, which makes both estimation and inferences more complicated, and therefore render such model extensions less appropriate for introducing the methodology. More importantly, a more detailed process description requires additional information that is not available in standard monitoring programmes.

The light half-saturation parameter is well determined by the procedure. Comparing the half-saturation parameter (Figure 8) with the range of data shows that there are observations on both side of the constant. In contrast, the half-saturation parameter for water column nitrogen is far below any observed value, implying that phytoplankton growth is not severely limited by nitrogen. Skive Fjord is a eutrophic ecosystem with large land-based inputs of nitrogen, and ambient concentrations of dissolved inorganic nitrogen are mostly above the levels reported to limit phytoplankton growth in experiments (e.g. Fisher et al.

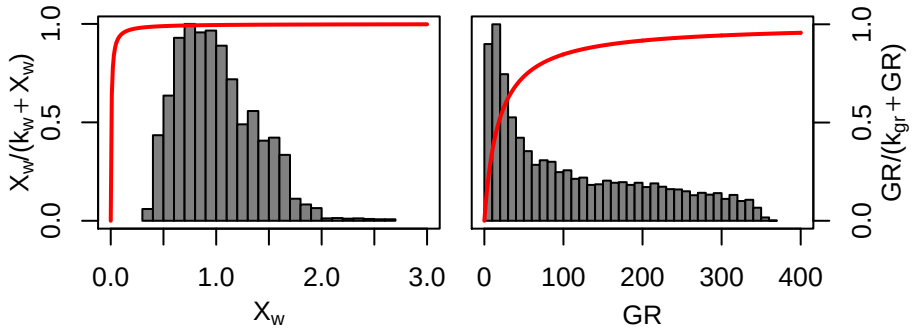

Fig. 8. Nonlinear multiplicative effects for nitrogen and light saturation. Red lines show the estimated relationships, and the histogram show the distributions of water column nitrogen (left) and observed global radiation (right).

(1992)).

5. Conclusion

The methodology presented in this study is based on likelihood estimation for identifying probabilistic models of physical/biological phenomena, formulated as a system of Itô stochastic differential equations partially observed in discrete time through a set of observation equations. By formulating parameters of the stochastic differential equation model as pure random walk hidden (unobserved) states, it is demonstrated how embedded structural information can be extracted by analysis of the smoothen state estimates. The selection of which parameter should be analysed to improve the model is based on considerations about the diffusion matrix *and* the observation covariance matrix. Large values of either or both of these terms for a particular state propose model deficiencies in the corresponding state and therefore possible model improvements. The diffusion of the random walk hidden state parameter should preferably be estimated to describe the dynamics of the parameter. It is demonstrated that even when the random walk parameter is estimated (close) to zero, it is still possible to identify model improvements by fixing the diffusion to a value that allowed regular and sufficient updating. In the presented example this is fixed such that the diffusion is comparable to the diffusion of other states of the system.

All suggested model improvements was tested by means of information criteria or likelihood ratio testing. The first part of the identification, based on random walk state estimation for identification and information criteria for selection, resulted in large reductions of the diffusion and the observation variance, implying that the deterministic skeleton dictated by the drift term, provides a better description of the model as the complexity of the model increase. The validation step is based on simulation studies, which also pinpointed model deficiencies and suggested further model extensions for both drift and diffusion. These model extensions lead to significant improvements of the likelihood, but more importantly to a stable simulation model and reductions of the simulation variance while improving the simulation performance (in terms of quantile skill score).

The diffusion term is shown to be important for both short-term (likelihood estimation)

and long-term (simulations) dynamics of the system. In the present context we are limited by the ability to find solutions for the Lamperti transform and an explicit inverse transformation. This excludes a large class of diffusion processes, where mass balance or partial mass balance is taken into account in the formulation of the diffusion matrix. The model development did however demonstrate that the correlation coefficient is not an important parameter in the model.

The simplicity of the model example implies that most parameters are well determined in a statistical sense. For more complicated models (or models with less information provided by the observations) it might be appropriate to include prior knowledge in the optimisation. The statistical software used here, does include the possibility of including prior knowledge in the estimation through the Maximum A Posterior (MAP) procedure described in Kristensen et al. (2004a). Such prior knowledge could be obtained from literature studies or from site specific experiments.

References

Aït-Sahalia, Y. (2008) Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, **32**, (2), 906-937

Baadsgaard, M., Nielsen, J. N., Spliid, H., Madsen, H., and Preisel, M. (1997) Estimation in stochastic differential equations with state dependent diffusion term. *SYSID '97 - 11th IFAC symposium on system identification, IFAC*

Bak, J., Nielsen, H. Aa., and Madsen, H. (1999) Goodness of fit od stochastic differential equations. *Symposium i Anvendt Statistik, Copenhagen*, 341-346

Bartell, S. M., Lefebvre, G., Kaminski, G., Carreau, M., and Campbell, K. R. (1999) An ecosystem model for assessing ecological risks in Québec river, lakes and reservoirs. *Ecological Modelling*, **124**, 43-67

Fasham, M. J. R, Ducklow, H. W., and McKelvie S. M. (1990) A nitrogen-based model of plankton dynamics in the oceanic mixed layer. *Journal of Marine Research*, **58**, 591-639

Fisher, T. R., Peele, E. R., Ammerman, J. W., and Harding, L.W. (1992) Nutrient limitation of phytoplankton in Chesapeake Bay. *Marine Ecology Progress Series* **82**, 51-63.

Forman, J., and Sørensen, M. (2008) The Pearson Diffusion: A Class of Statistically Tractable Diffusion Processes. *Scandinavian Journal of Statics*, **35**, 438-465

Gard, T. C. (1988) Introduction to stochastic differential equations. *Marcel Dekker, Inc.*, New York.

Gneiting, T., and Raftery, A. E. (2007) Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102**(477), 359-378

Hastie, T. J., and Tibshirani, R. J. (1990) Generalized additive models. *Monographs on Statistics and Applied Probability*, **43**, Chapman & Hall/CRC

Hastie, T., Tibshirani, R., and Friedman, J. (2001) The Elements of Statistical Learning - Data Mining, Inference and Prediction. *Springer Series in Statistics, Springer New York.*

Iacus, S. M. (2008) Simulation and Inference for Stochastic Differential Equations - With R Examples. *Springer Series in Statistics, Springer New York.*

Jazwinski, A. H. (1970) Stochastic Processes and Filtering Theory. *Academic Press, New York, ASU.*

Karatzas, I., and Shreve S. E. (1991) Brownian Motion and Stochastic Calculus - Second edition. *Springer New York.*

Klebaner, F. C. (2005) Introduction to Stochastic Calculus with Applications. *Imperial College Press, London.*

Kloden, P., and Platen, E. (1999) Numerical solutions of Stochastic Differential Equations. *Springer-Verlag, Berlin.*

Koenker, R., and Bassett, G. J. (1978) Regression Quantile. *Econometrica*, **46**(1), (33-50).

Kristensen, N. R., and Madsen, H. (2003) Continuous time stochastic modelling - CTSM 2.3 - Mathematics Guide. *Technical University of Denmark*

Kristensen, N. R., Madsen, H., and Jørgensen, S. B. (2004a) Parameter estimation in stochastic grey-box models. *Automatica*, **40**, 225-237

Kristensen, N. R., Madsen, H., and Jørgensen, S. B. (2004b) A method for systematic improvement of stochastic grey-box models. *Computers & Chemical Engineering*, **28**, 1431-1449

Luschgy, H., and Pagés, G. (2006) Functional quantization of a class of Brownian diffusions: A constructive approach. *Stochastic Processes and their Applications* **116***, 310-336*

Madsen, H. (2008) *Time Series Analysis.* Chapman & Hall/CRC

Madsen, H., Holst, J., and Thyregod P. (1987) A Continuous Time Model for the Variations of Air Temperature. *10. Conference on Probability and Statistics in Atmospheric Science, American Meteorological Society,* pp. 52-58. Edmonton

Møller, J. K., Nielsen, H. Aa., and Madsen, H. (2008). Time-adaptive quantile regression. *Computational Statistics & Data Analasis*, **52**, 1292-1303

Nielsen, H. Aa., and Madsen, H. (2001) A Generalization of some Classical Time Series Tools. *Computational Statistics & Data Analysis* **37**, 13-31

Pedersen, M. W., Righton, D., Thygesen, U.H., Andersen K., and Madsen, H. (2008). Geolocation of North Sea cod using Hidden Markov Models and behavioral switching. *Canadian Journal of Fisheries and Aquatic Sciences*, Vol. 65, pp. 2367-2377

Pinson, P., Nielsen, Aa. H., Møller, J. K., and Madsen, H. (2007) Non-parametric Probabilistic Forecast of Wind Power: Required Properties and Evaluation. *Wind Energy*, **10**, pp. 497-516

Tornøe, C. W., Jacobsen, J., Pedersen, O., Hansen, T., and Madsen, H. (2004) Grey-box Modelling of Pharmacokinetic/Pharmacodynamic Systems. *J. Pharmacokinet. Pharmacodyn*, Vol. 31, pp. 401-417

Øksendal, B. (2003) Stochastic Differential Equati¡ons - An Introduction with Applications, Sixth edition. *Springer-verlag, Berlin*