Deterministic and stochastic models in the natural sciences

IMS meeting, Banff, July 31, 2002

Hans R. Künsch

Seminar für Statistik, ETH Zürich

Roland Brun and Peter Reichert

Institute for Environmental Science and Technology, EAWAG

1. Introduction

Often, one reads or hears statements like this: “In all fields of science, more and more data are collected, frequently in an automated way. In order to analyze these data, more statisticians are needed.”

In this talk, I try to give some explanations why this is not what happens. It is based on my experience in building up a collaboration with environmental scienctists at ETH.

My main point is that the current state of knowledge about environmental systems makes it far more difficult for statistics to add something meaningful to the analysis of the available data than one might naively think.

I will explain how I came to this conclusion. I will also try to show some statistical problems and tools that I consider to be worthwile in this context.

2. Data and models for aquatic systems

Data from environmental systems usually come as a multivariate space-time series. I consider here examples of aquatic systems, but the situation is similar for soil, atmospheric and oceanic systems.

A sewage treatment plant:

During two months, data were collected on water flows and on phosphate, ammonia, oxygen and suspended solids in some of the reactors of one plant at half-hourly intervals.

Monitoring of Lake Zurich:

Lake Zurich is a long narrow lake with a length of 30 km, a width between 0.5 and 2.5 km and a maximal depth of 137 m. The city of Zurich is at the lower end where the only outflow of the lake is. The shores are densely populated, and the lake provides drinking water for up to a million inhabitants.

A large number of physical, chemical and biological variables are measured monthly since 1972 at six stations. Data on inflow from rivers and sewage plants and atmospheric deposition are also available.

What would a statistician do with such data ?

Presumably, we would try to decompose the series into trends, seasonal effects and irregular part and then model the covariance structure of the irregular part. Interactions between temporal and spatial components might be complicated to handle. Finally, one might try to fit a (nonlinear) vector autoregression.

However, environmental scientists and engineers have little interest in the outcome of such a purely data-driven exercise. They are interested in models that

- take as much knowledge about the underlying processes into account as possible,
- contribute to the understanding of these processes,
- are transferable to similar systems,
- allow prediction of the same system under different driving conditions than those observed,
- have parameters with a clear subject matter interpretation.

Why can't we take knowledge about the underlying processes into account in our statistical models ?

These processes can be described mathematically only in continuous time and space (with ordinary or partial differential equations) and they involve many variables that are not observed. Subject matter knowledge cannot be formulated in a way that involves only the observed variables at the given discrete points in time and space.

In the following, I give a brief introduction into the kind of models that are used by environmental scientists. Then I will try to describe some of the problems that occur when one tries to fit such models to data.

2.1 A simple biochemical process

Consider three variables X_t, S_t, O_t denoting biomass, substrate and oxygen which evolve in time according to the differential equation

$$\begin{aligned}
\frac{\mathrm{d}X_t}{\mathrm{d}t} &= \mu\frac{S_t}{K_S + S_t}X_t - bX_t \\
\frac{\mathrm{d}S_t}{\mathrm{d}t} &= -\frac{1}{Y}\mu\frac{S_t}{K_S + S_t}X_t \\
\frac{\mathrm{d}O_t}{\mathrm{d}t} &= K_{ex}(O^{sat} - O_t) - \frac{1-Y}{Y}\mu\frac{S_t}{K_S + S_t}X_t - (1-f_I)bX_t.
\end{aligned}$$

This takes the 3 processes growth of biomass, decay of biomass and reaeration into account. For instance, the growth rate is $\mu\frac{S_t}{K_S+S_t}X_t$ (linear in X_t, but limited in a nonlinear way by the available substrate). This rate adds to the rate of change for each variable with different factors.

The unknown parameters are μ, K_S, b, Y, K_{ex}, O^{sat}, f_I and the three initial conditions.

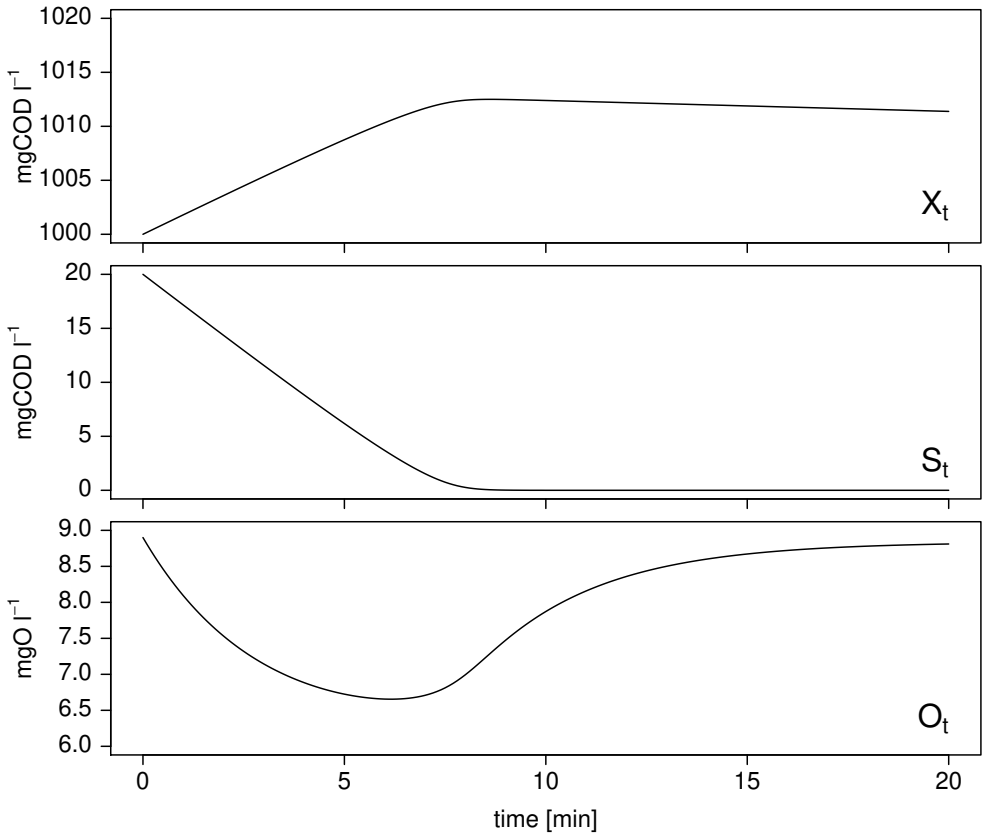

Figure 1: Solution of the differential equation for some initial condition. In a first phase, the biomass increases, using up the substrate and reducing oxygen. In a second phase, biomass decreases and oxygen moves back to the saturation level.

2.2 Biochemical processes in larger systems

Many variables, many processes, more complicated rates. In addition, transport processes have to be taken into account.

For our sewage plant:

Variables: phosphate, oxygen, nitrate, nitrogen, autotrophic and heterotrophic bacteria concentrations etc. in different reactors. (17 in total).

Processes: aerobic/anoxic/anaerobic growth of bacteria, hydrolysis etc. (19 in total).

For the lake Zurich:

Variables: phosphate, ammonia, nitrate, oxygen, organic phosphorus and dry mass without P of two groups of algae, dry mass of zooplankton, etc. (14 in total).

Processes: growth, respiration and death of two groups of algae and of zooplankton, aerobic and anoxic mineralization etc. (13 in total).

Differential equations for these processes in case of perfect mixing (no spatial dependence)

$$\frac{\mathrm{d}X_i}{\mathrm{d}t} = r_i(X) = \sum_j \nu_{ij}\rho_j(X).$$

Here ρ_j is the rate of process j, and ν_{ij} is the factor by which this rate contributes to the production rate r_i for variable i.

For example growth of one group of algae depends linearly on the concentration of this group and nonlinearly on the concentrations of nitrate and phosphate and on temperature and light intensity.

2.2 Spatial dependence, transport processes

Concentrations of variables and rates depend not only on time, but also on spatial locations. Changes in space and time of concentrations are determined by physical transport processes (advection, diffusion, sedimentation).

In sewage plants, reactors are usually well stirred and can thus be considered to a good approximation as homogeneous. One obtains a system of ODE's for each reactor, coupled by the flow between reactors.

In lakes, horizontal mixing is typically fast, but vertical mixing is slow. Hence one considers horizontally averaged variables, leaving depth z as the only spatial dimension. The biochemical processes described above are then combined with the physical processes in a mass balance equation

$$\frac{\partial X}{\partial t} = -\frac{\partial}{\partial z} q + r$$

where r is the production rate as above and q is the flow.

For instance, for diffusive mixing, $q = -K\partial X/\partial z$ where the mixing coefficient K depends on depth and time.

For the lake of Zurich, there are in addition two compartments corresponding to two sediment layers, and there is exchange between these layers and the water column.

As a result, one has a complicated system of ordinary or partial differential equations that reflect laws of nature and the current understanding of the processes involved. On an “absolute scale” these models are fairly large and complicated, but relative to the complexity of the system they must still be considered to be primitive. In particular, there are very few possibilities for simplification.

These models contain a large number of parameters (75 in the case of the sewage plant, 52 in the case of the lake Zurich). In principle one would like to estimate them by weighted nonlinear least squares, that is comparing the solution of the differential equations with the data. Here the data are some of the variables at a discrete set of time points and spatial locations.

Peter Reichert has developed a simulation and analysis program called AQUASIM. It allows to define models of the kind described above, solve the differential equations, fit parameters by least squares and perform sensitivity and uncertainty analysis.

3. Parameter estimation in the non-identifiable case

Problems arise because all these models have many more parameters than can be estimated from the available data. The algorithms for computing the optimal fit do not converge, and there is a whole complicated surface in parameter space where the sum of squares is minimal for all practical purposes.

This sounds like a wonderful occasion for Bayesian statistics. Indeed, there is prior knowledge from laboratory experiments and earlier studies about most of the parameters, and this can be quantified. There is also some feeling about which parameters are connected, but it is an impossible task to specify a joint prior for 20 or 50 parameters.

In addition, computing one solution of the system of differential equations is time consuming, and therefore simulating from the posterior is out of reach currently.

3.1 Sensitivity and collinearity measures

Our approach tries to find subsets of the parameters which can be estimated reasonably from the data when other parameters are fixed at some prior value.

Some notation: Let $\mathbf{y}$ be the available data put into a $n \times 1$ vector in some arbitrary order and let γ_i be the scale of variable y_i. Let $\boldsymbol{\theta}$ be the $m \times 1$ vector of all parameters and let $\boldsymbol{\eta}(\boldsymbol{\theta})$ be the output vector of the model (same variables, times and locations as for $\mathbf{y}$, in the same order). The least squares criterion is then

$$(\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}))^T W (\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}))$$

where W is $\mathrm{diag}(\gamma_i^{-2})$.

Finding a subset of parameters which can be estimated from the available data involves the following steps

- Select a reasonable parameter value $\boldsymbol{\theta}_0$.
- Define an uncertainty range $\Delta\theta_j$ for each component θ_j.
- Compute the sensitivity matrix $S = (s_{ij})$,

$$s_{ij} = \frac{1}{\gamma_i} \frac{\partial \eta_i(\boldsymbol{\theta})}{\partial \theta_j} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \Delta\theta_j \,,$$

  and compute from it importance indices for each component θ_j and collinearity indices for subsets K of components.
- Look for subsets of important parameters with low collinearity indices.
- Minimize the least squares criterion with respect to the components of the chosen subset to obtain a new $\boldsymbol{\theta}_0$ and iterate.

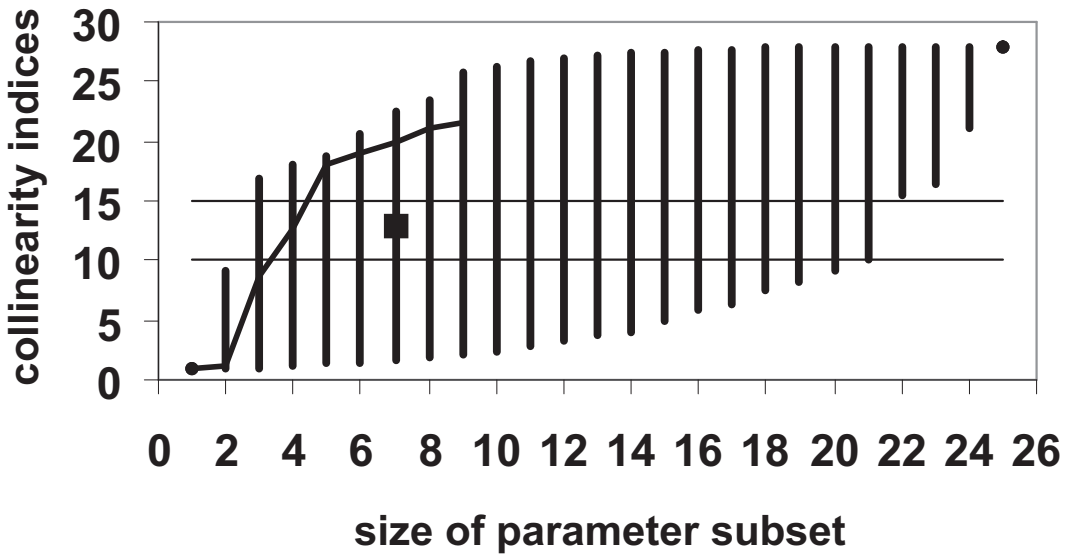

Figure 2: Fitting the model for Lake Zurich: Collinearity indices for different subsets

At the end, one should look again at the collinearity indices in order to explore the influence of the fixed components on the estimated components.

How does this method work for the sewage plant and the Lake Zurich ?

The starting value $\boldsymbol{\theta}_0$ is typically good, and only a few components undergo substantial changes in the process.

Studying the influence of fixed values on the estimated parameters reveals that these changes reflect not so much an error in the current opinion on the value of these components, but rather a lack of knowledge about some crucial variables (e.g. the precise composition of inflows) or a model deficit.

From an engineering point of view, the fit of the models is surprisingly good. However, time series plots of residuals reveal that systematic errors often dominate random errors.

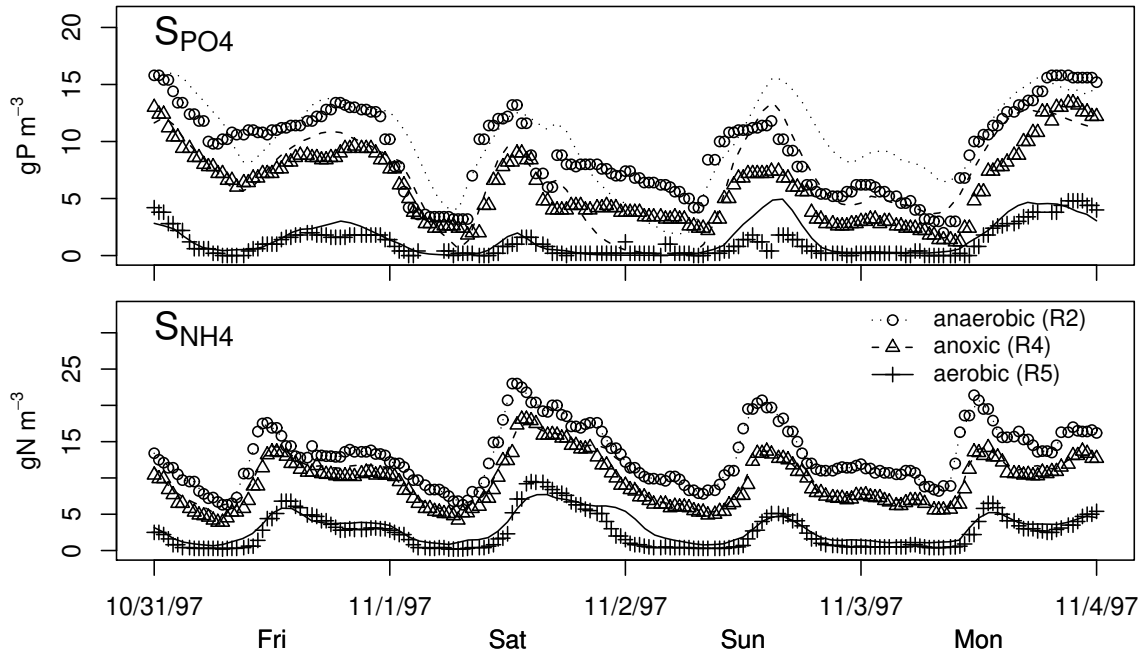

Figure 3: Comparison of observations and fitted model in the sewage plant example. The variables are phosphate and ammonia in 3 reactors.

4. Stochastic differential equations and time-varying parameters

If we want to include stochastic components in models that use the current biological, chemical and physical knowledge, we are led naturally to stochastic differential equations.

There is currently much activity in statistics for solutions of SDE's where partial and noisy observations are available at discrete time points. The main motivation for this work comes from financial mathematics, but I think it has also potential applications in natural sciences.

Recent work in this area has been done by Durham, Gallant, Shephard, Pitt and Roberts.

The most obvious reason for introducing randomness into a differential equation is that many parameters are not really constant, but vary in time in a way that is not predictable. This can account for instance for the systematic deviations that occur with deterministic models.

If a parameter θ enters linearly in a differential equation, $\frac{\mathrm{d}X_t}{\mathrm{d}t} = \theta f(X_t)$, then by putting $\theta = \overline{\theta}$ plus "noise", one obtains the SDE

$$\mathrm{d}X_t = \overline{\theta} f(X_t)\mathrm{d}t + \sigma f(X_t)\mathrm{d}W_t.$$

(this is for instance the first example in Oksendal's book).

However, this implies extremely fast fluctuations in the parameter which is is not reasonable for natural systems. We find it much more realistic to work with a model

$$\mathrm{d}\theta_t = \mu(\theta_t)\mathrm{d}t + \sigma(\theta_t)\mathrm{d}W_t, \quad \mathrm{d}X_t = f(X_t, \theta_t)\mathrm{d}t.$$

Simple models like Ornstein-Uhlenbeck or Cox-Ingersoll-Ross for θ_t are presumably sufficient.

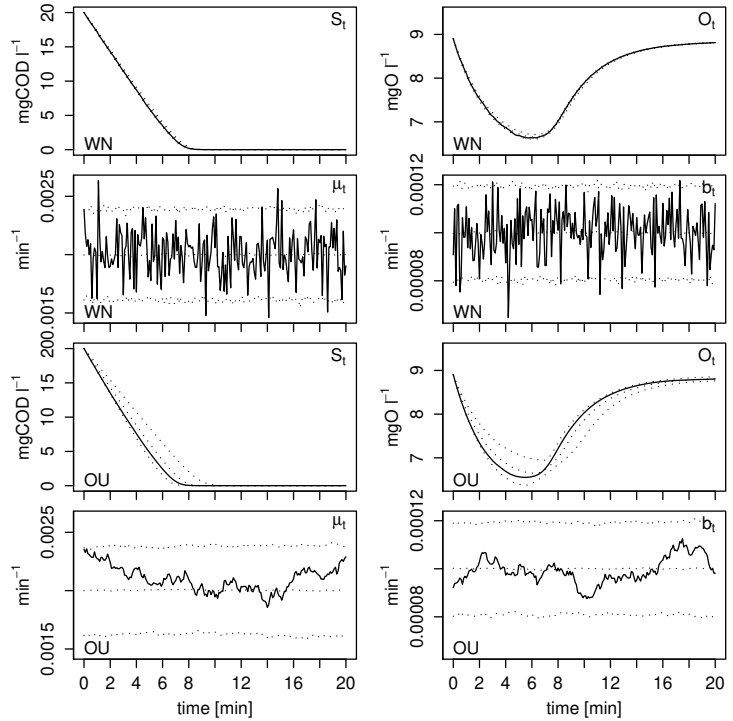

Figure 4: Model for biomass, substrate and oxygen with time varying parameters μ (maximal growth rate) and b (death rate) according to white noise and an OU-processes respectively. Shown are the 2.5%, 50% and 97.5% quantiles based on 1000 simulations, together with one arbitrary simulation.

4.1 MCMC for partially observed SDE's

Consider an unobserved SDE with unknown parameters β

$$\mathrm{d}X_t = \mu(X_t, \beta)\mathrm{d}t + \sigma(X_t, \beta)\mathrm{d}W_t, \quad (0 \leq t \leq T).$$

For implementations, we will always take the Euler approximation with small step size δ. What we observe is

$$Y_i = HX_{t_i} + \epsilon_i.$$

In the Bayesian approach, we want to simulate from the joint distribution of $((X_t), \beta)$ given (Y_i) with a Metropolis-Hastings algorithm. We will update either β, keeping the whole path fixed, or update the part on (r, s) of the path, keeping β and the parts on $[0, r]$ and $[s, T]$ fixed.

The key task is then to find a good proposal for a diffusion with given initial and end points and partial information for a few points in between.

However, as Roberts and Stramer have pointed out, there is an additional difficulty: If the volatility depends on β, the convergence is slow for small step size δ. The volatility is then almost completely determined by the path and it changes very little during the modification of the path.

4.2 The particle filter for partially observed SDE's

Partially observed SDE's are special cases of **state space** models. State space models consist of an unobservable time-discrete Markov chain of state variables Z_j and a sequence of observations Y_j such that Y_j depends only on Z_j and different Y_j's are conditionally independent. We denote the transition density of the states by $a(z_j, z_{j+1})$ and the conditional obervation density by $b(z_j, y_j)$.

state process	$\dots$	$\to$	Z_{j-1}	$\to$	Z_j	$\to$	Z_{j+1}	$\to$ $\dots$
			$\downarrow$		$\downarrow$		$\downarrow$	
observation	$\dots$		Y_{j-1}		Y_j		Y_{j+1}	$\dots$

In the case of partially observed SDE's, the state Z_j is the path of X on $(t_{j-1}, t_j]$. After a discretization of time, we take a to be the product of normal densities from the Euler approximation. b depends only on X_{t_j}.

The **particle filter** is a recursive Monte Carlo approximation to the conditional density of Z_j given $(Y_1, \ldots Y_j)$, the so-called filter density f_j. In principle, f_j can be computed with the recursion

$$f_{j+1}(z_{j+1}) \propto b(z_{j+1}, y_{j+1}) \int a(z_j, z_{j+1}) f_j(z_j) dz_j.$$

The problem is the computation of the integral (and to a lesser extent, the normalization).

The particle filter uses approximations $f_j(z_j) \approx \sum_{k=1}^{K} \lambda_{j,k} \Delta(z_{j,k})$ where $\lambda_{j,k}$ are weights and $\Delta(z)$ is a point mass at z. Inserting this into the recursion, we obtain

$$f_{j+1}(z_{j+1}) \propto b(z_{j+1}, y_{j+1}) \sum_{k=1}^{K} \lambda_{j,k} a(z_{j,k}, z_{j+1}).$$

In order to complete the recursion, we have to generate a weighted sample from this density. Due to lack of time, I skip the discussion how to do this.

What can we do with the filter density ?

It allows us to approximate the likelihood since

$$p(y_j \mid y_{j-1}, \ldots y_1) = \int f_{j-1}(z_{j-1}) a(z_{j-1}, z_j) b(z_j, y_j) dz_{j-1} dz_j.$$

There are some problems however, because we obtain discontinuous approximations of log likelihood, and for each value of the likelihood function we need a different particle filter.

Information about the unobserved states is also of considerable interest, but usually one wants the conditional distribution of Z_j given all observations $(y_1, \ldots y_n)$ rather than the filter distribution. The two are related by

$$p(z_j \mid z_{j+1}, y_1, \ldots y_n) \propto a(z_j, z_{j+1}) f_j(z_j),$$

but in the context of SDE's it is difficult to exploit this relation.

4.3 ODE's with random coefficients

$$\mathrm{d}\theta_t = \mu(\theta_t)\mathrm{d}t + \sigma(\theta_t)\mathrm{d}W_t, \quad \mathrm{d}X_t = f(X_t, \theta_t)\mathrm{d}t.$$

Since (X_t) is a deterministic function of θ_t, the MCMC algorithms need to be modified. Fixing the values of (X_t) outside (r, s) does not make sense. The particle filter can be used.

Consider the time evolution for biomass, substrate and oxygen from the beginning with maximal growth rate μ_t and death rate b_t following OU-processes. Observed are only oxygen values.

In the first phase, b_t is not identifiable. Therefore the filter values for b_t can be slightly off, leading to errors in biomass and substrate. These errors become visible at the beginning of the second phase also in the oxygen, and the filter has then to make large adjustments quickly. This can lead to a breakdown of the algorithm.

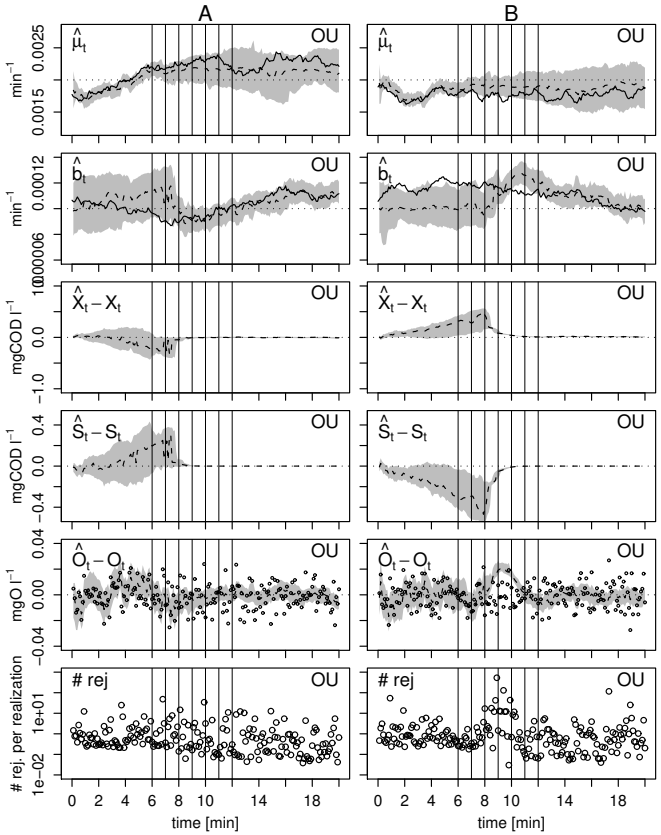

Figure 5: Particle filter for two realizations of the biomass, oxygen and substrate model. Solid: True value, Dashed: Filter median. Shaded: 95% filter interval.

5. Discussion

- The leading paradigm in many fields of science is the use of determinstic continuous models formulated as PDE's.
- These models are remarkably successful, and prior information allows even to use nonidentifiable models.
- There is uncertainty about processes that cannot be resolved in the current models, and quantifying this uncertainty is a challenge.
- Our students should get some exposure to PDE's (physical background, analytic and numerical methods).