An Evaluation of Methodologies for Uncertainty Analysis in Biological Waste Water Treatment.

Sebastiaan G.T. Kops, Peter A. Vanrolleghem
University of Ghent
Belgium

November 28, 1996

Keywords

Uncertainty analysis, Monod growth model, Monte Carlo simulation, Stochastic differential equations

Abstract

Because of the complexity of the combined system of sewers, wastewater treatment plants, and receiving waters it is necessary to include uncertainty analysis into the process of modeling this combined system. In this paper a start is made on the research of the uncertainty analysis in waste water treatment plants. Attention will focus on a crucial process in the activated sludge system. Assuming all parameters and uncertainties are known, the propagation of the model uncertainty into the model output is approximated. Results will be presented of an uncertainty analysis for a Monod growth model. Three methods approximating the propagation of model uncertainty are evaluated, namely Monte Carlo simulation, Monte Carlo simulation with stochastic parameters, and a method which uses the theory of stochastic differential equations.

1 Introduction

The past years the combined modeling of sewers, waste water treatment plants and receiving waters has become more and more a topic of discussion. Since this combined system is very complex, it will be impossible to model reality perfectly.

When predictions of such a complex model are to be matched with some observed behavior it is most likely that the model (for any parameter combination) will not give a perfect match. This will give rise to uncertainty in the predictions obtained from the model, called prediction uncertainty. Since prediction uncertainty is a measure indicating whether a model is capable of matching observed behavior, the analysis of this uncertainty must be incorporated in the process of modeling such a complex system as the combined environmental system of sewers, waste water treatment plants, and receiving waters.

In the past a lot of research has been done with regards to the uncertainty analysis of the sewer systems and receiving waters (*e.g.,* [Bec87]). However, there has been done very little research involving the waste water treatment plants. This might be considered a bit odd since uncertainty is most important in modeling waste water treatment plants. The input into the plants is always uncertain (the amount of input but also its characteristics), but also uncertainty occurs in parameters or even in the model structure itself. Introducing uncertainty in control strategies regarding waste water treatment plants might also result in more robust control strategies for the waste water treatment plants. In recent years, the focus of research regarding waste water treatment plants has been on the identification of parameters in models describing the activated sludge system [Van94, VKC96, VK96]. Using batch experiments a lot of measurements of the activated sludge process have been obtained. So in contrast with water quality modeling, there are a lot of measurements of the activated sludge process which are very useful in the research of uncertainties involved in the process.

In this paper some methodologies are being examined with respect to the Monod growth model used in modeling the activated sludge process[1]. The first one is a very often applied method, *i.e.,* Monte Carlo simulation. Two other not as well known methods are being examined. The two methods take into consideration that the parameters may vary in time. The first of these two is a modified Monte Carlo simulation where stochastic (time-varying) parameters are implemented and, second, a method will be evaluated which uses the theory of stochastic differential equations. In the next section a general introduction into uncertainty analysis is given.

[1] Note that all parameter values as well as uncertainties used in this paper are purely theoretical. Since here only the main characteristics of the models and their output is important this will not be a problem.

The three methods reviewed are explained in section 3 and results are given in section 4. In the last section conclusions are given.

2 Uncertainty Analysis

One of the goals of modeling is to *predict*. With this prediction and the precise knowledge of the system as translated in a mathematical model it is possible to control the modeled system. To predict it is necessary to have a *perfect* view of the true environmental process or system. However, the environmental system is very complex and infinitely large. Therefore, it is *impossible* to know all environmental processes and, hence, there will always be some uncertainty in the predictions we obtain from mathematical models. Therefore it is just as important not to give solely the predictions but also their uncertainty.

In theory, the uncertainty of a prediction obtained from a model, the *prediction uncertainty*, only depends on the *model uncertainty*. By model uncertainty is meant the *uncertainty caused by everything which is not modeled*, or in other words, the uncertainty caused by all processes which are not included in the model. It is possible that the model uncertainty changes with time. Let us consider, for example, in a model of the water quality of a river that only rainfall is not included. The model will show a low model uncertainty during a dry period. However if it suddenly starts to rain, a process occurs which is not included in the model and this will cause an increase in the model uncertainty.

So the real uncertainty of predictions obtained using a model depends on the model uncertainty of that model. Therefore it is important to quantify this model uncertainty. However, there is still very little known about this model uncertainty. In the following we will briefly discuss some uncertainties which will influence the model uncertainty and the determination of the prediction uncertainty. These uncertainties are:

1. Parameter uncertainty.
2. Measurement uncertainty.
3. Mathematical uncertainty.

An environmental process always depends on other processes. To define the dependency of the modeled process on other processes constant values are used. The constant values are called parameters[2]. In modeling an environmental process we will always look for a 'best' parameter set for a given model. Once this parameter set is found it will always corresponds with that particular model[3]. The 'best' parameter set may, for example, be given by the parameter set which physically makes the most sense or the parameter set with the lowest *parameter uncertainty*. Let us assume that the 'best' parameter set is given by the set of parameters with the lowest parameter uncertainty.

Suppose we have the perfect model, which means that the model uncertainty is zero. This also will imply that the parameter uncertainty is zero. Or, to put it in other words, it is possible to find the true parameter set since there will not be any uncertainty caused by other processes. So a model uncertainty of zero implies a parameter uncertainty of zero. This will again imply that if the parameter uncertainty is not zero the model uncertainty will also not be zero. In fact, if the parameter uncertainty increases, which means that it is more probable that the parameters which will predict the process perfectly using a particular model will not be found, the model uncertainty will also increase. Therefore it is possible to use the parameter uncertainty as an *indicator* for the model uncertainty.

To obtain the constant values of the parameters often parameter estimation methods are used. These methods make use of certain measurements of the modeled process to estimate the parameters in the model. However, it is never sure that all measurements of a certain process are perfect. In other words, there will always be some *measurement uncertainty*. The measurement uncertainty will influence the estimation of the parameters and their uncertainty. Measurements are also used to determine the prediction uncertainty, by comparing these with the simulation results of the model. Using this comparison the prediction uncertainty is determined and, therefore, the measurements will also influence the determination of the prediction uncertainty. Another problem involving measurements is that there will never be enough measurements to fully quantify the model uncertainty. This means that there will always be some uncertainty whether the measurements taken are enough to describe the whole process.

Another problem in estimating the parameters of a model are the mathematical limitations of the parameter estimation methods used. A lot of environmental processes are modeled using a nonlinear, continuous model. At this moment it is still impossible to estimate the parameters of a nonlinear model perfectly well. For the estimation of parameters computers are used. These computers work in discrete time. The measurements are also taken at discrete moments. Since the modeled process is continuous the model has to be discretized. This will also cause some uncertainty. Another uncertainty which may occur is caused by the finite precision of a com-

[2] Since the initial conditions are also part of the model and are considered to be constant, they can also be considered as parameters.

[3] If all parameters are identifiable.

puter. All the uncertainties involving our mathematical knowledge and the practical use of these mathematical knowledge are called *mathematical uncertainty*. Since a model has to be simulated to obtain predictions, the parameter uncertainty and the determination of the prediction uncertainty are also influenced by the mathematical uncertainty.

To summarize, the determined uncertainty of the predictions obtained from a certain model depends on the uncertainties caused by all processes which are not modeled, the model uncertainty, the uncertainties caused by the mathematical methods used to obtain these predictions from a certain model, the mathematical uncertainties, and the uncertainties involving the measurements of the process, the measurement uncertainty. The model uncertainty is very difficult to quantify but the parameter uncertainty of the parameter set with the lowest parameter uncertainty given a certain model can be used as an indicator for the model uncertainty. However, the parameter uncertainty found depends on the methods and the measurements used for estimating these parameters and their uncertainties. Each method has its own mathematical uncertainty and therefore the determined parameter uncertainty depends on the mathematical uncertainty. As mentioned before, the determination of the prediction uncertainty is also influenced by the mathematical uncertainty since a model has to be simulated for determining the prediction uncertainty. What remains is the measurement uncertainty. This uncertainty influences the parameter uncertainty since the parameters are estimated using measurements of a certain process. Since the determination of the prediction uncertainties always is done by using measurements of a given process, the measurement uncertainty also influences the determination of the prediction uncertainty. All these interactions are given in figure 1.

In the foregoing discussion it is said that the parameter uncertainty is an indicator for the model uncertainty and therefore for the determined prediction uncertainty. However, since it is said that model uncertainty is caused by everything which is not modeled yet, the parameter uncertainty is only *one* of the many sources which may cause model uncertainty. Theoretically it is very difficult to classify the other sources of uncertainty. However, practically it is possible to describe all these sources (except for the parameter uncertainty) by adding *one* noise term to the ordinary differential equation describing the system. Two ways to include model uncertainty in a model are:

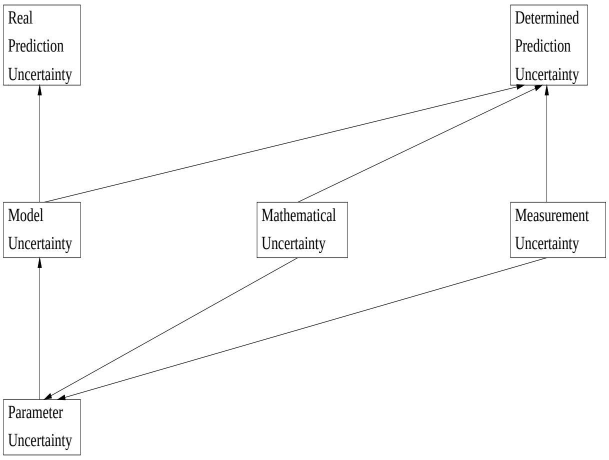

Figure 1: *A view on modeling prediction uncertainty*

- Account only for parameter uncertainty. This means that in a certain model the parameter p will become $p(t) = \bar{p} + N(t)$ where $\bar{p}$ is the parameter without noise and $N(t)$ is the noise term.
- Lump all model uncertainty in *one* noise term. The model $\frac{dx}{dt} = f(x,p,t)$ will then become $\frac{dx}{dt} = f(x,p,t)+N(t)$ where p is a parameter (set)and $N(t)$ is a noise term.

These two methods may be combined if there are more sources of model uncertainty than only the parameter uncertainty. The second method is physically not very plausible but must also be considered because of its technical simplicity.

Furthermore, one might have noticed that *input uncertainty* is not mentioned. Mostly the input of a model consists of measured data. Therefore the uncertainties involved in these inputs are scaled under measurement uncertainties.

In the following sections the propagation of the model uncertainty into the determined prediction uncertainty is examined. This is done by assuming that all parameters and model uncertainty (including parameter uncertainty) are known. Since nothing has to be estimated, no measurement uncertainty has to be included. Mathematical uncertainty has to be mentioned since the propagation of model uncertainty into the determined prediction uncertainty is only an approximation. However, it will not be determined or taken into account.

3 Three Methodologies

In this section three methodologies for approximating the prediction uncertainty, given some model uncertainty, are being reviewed. All these methods approximate the mean and variance of a model given the variance of a parameter (set) or noise term. The mean is the expected output of the model and the variance is a measure for the uncertainty. The first method examined is well known and frequently used, the others are not as well known. These two methods try to solve a big disadvantage of the first method. All three methods are also mentioned in [Kre83]. Note that these methods are, of course, not the only ones which may be used to approximate the prediction uncertainty. Some other methods are mentioned in [Kre83].

3.1 Monte Carlo Simulation

Monte Carlo Simulation is a well known and frequently used method to approximate some stochastic properties of a modeled system. A deterministic model is run repeatedly with every run a different set of parameter values. These parameter values are determined at the beginning of every run from specific probability distributions. Note that the parameter values do not change during one run. Monte Carlo simulation is frequently used since it is conceptually very simple and easy to use given some previously developed random number generators. It is, for example, often used in sensitivity analysis, *i.e.*, whether a model is sensitive to a change in the parameter values.

Another application is the approximation of the prediction uncertainty of the model given know probability distributions for the parameters. The mean and variance (uncertainty) of the model are approximated using

$$EX(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t) \tag{1}$$

$$var(X(t)) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i(t) - EX(t))^2 \tag{2}$$

where n is the number of total simulations (this number has to be sufficiently large) and i is the i-th simulation run.

To summarize, some advantages of Monte Carlo simulation are

- The simplicity of Monte Carlo simulation, both conceptually and technically.
- The freedom of choosing any probability distribution for the parameters.

One disadvantage of Monte Carlo simulation, apart from the relatively long simulation time needed, is that one has to assume that the parameter values have to remain constant during one run. This assumption, especially during long-term experiments, may not always be valid. The methods described in the next subsections try to solve this problem.

3.2 Monte Carlo Simulation with Stochastic Parameters

Despite the many advantages of Monte Carlo simulation with stochastic parameters this method has rarely been used. It is in concept the same as the previously mentioned Monte Carlo simulation. However, in this method the parameter values will, in contrast with Monte Carlo simulation, vary in time. The parameters will be determined *at each time instant* from given probability distributions. This method has the same advantages as the Monte Carlo simulation. Moreover, the disadvantage of Monte Carlo simulation, namely the assumption of the parameters being constant during one run, has been canceled. A disadvantage of this method is that it is relative slow.

Until now, only parameter uncertainty is mentioned. However, this method is not only capable of simulating a model with stochastic parameters but also of simulating a deterministic model with a noise term added to it.

3.3 Stochastic Differential Equations

The basis of the third method is the theory of the so called (Itô) Stochastic Differential Equations (SDE)(*e.g.* [Bag93]). Starting from some ordinary differential equation (ODE) a random differential equation may be obtained by assuming the parameters are stochastic processes or by adding a noise term to the ODE. If the stochastic parameters or the noise term has a Gaussian distribution it is possible to obtain an (Itô) SDE. For linear equations it is possible to determine the mean and variance of the model exactly using the Fokker-Planck equation. However, for non-linear equations the mean and variance have to be approximated. This may , relatively easy, also be done by using the Fokker-Planck equation to determine ODE's which approximate the mean and variance. These approximations will only hold if the model is not too complex. Since ODE's are used to approximate the mean and variance, this method is relatively fast. Other ways to approximate the mean and variance of SDE's are the stochastic numerical integration methods given in [KP92]. These methods will stay valid for complex systems but will take, relatively, much more simulation time.

A disadvantage of using SDE's for approximating prediction uncertainty is the assumption of Gaussian distribution for the parameters or noise terms. This assumption (especially in environmental systems) may not always be valid.

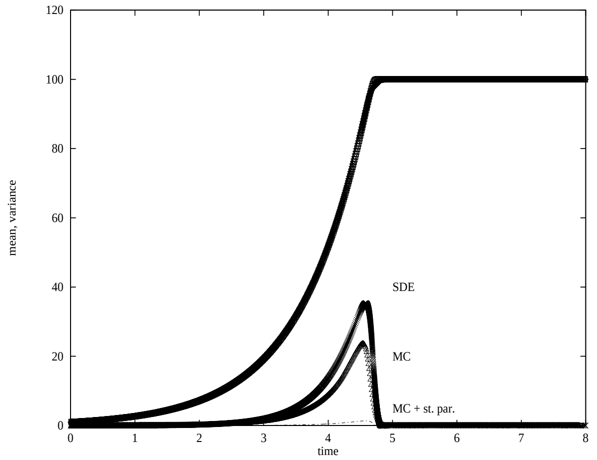

Figure 2: *Approximated means(high) and variances(low) (MC, MC with stoch. param., SDE approach; noise on whole ODE)*

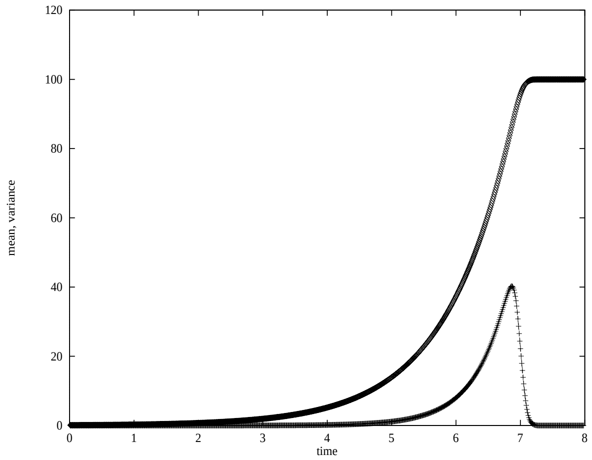

Figure 3: *Approximated mean(high) and variance(low) (MC with stoch. param.; noise on whole ODE)*

4 Results

The Monod growth model is given by

$$\frac{dx_t}{dt} = \frac{\mu_m s_t}{k_s + s_t} x_t \quad (3)$$

$$\frac{ds_t}{dt} = -\frac{1}{Y}\frac{dx_t}{dt} \quad (4)$$

in which x and s are respectively the biomass and substrate concentration, μ_m is the maximum growth rate, k_s is the half-saturation constant, and Y is the yield. Using $s_t = \frac{1}{Y}(x_m - x_t)$ where $x_m = x_{t_0} + Y s_{t_0}$, $\tilde{t} = \mu_m t$, and $\tilde{x}_t = \frac{100 x_t}{x_m}$ results in

$$\frac{d\tilde{x}_t}{d\tilde{t}} = \frac{100 - \tilde{x}_t}{K + 100 - \tilde{x}_t}\tilde{x}_t \quad (5)$$

This last equation is used to evaluate Monte Carlo simulation, Monte Carlo simulation with stochastic parameters, and a method using SDE's (SDE approach) for the approximation of prediction uncertainty in the Monod model. In this paper both Monte Carlo simulation methods are being evaluated using Gaussian distributions for the parameters and other noise terms. In [Ste83] was shown that if only parameter uncertainty was accounted for in the SDE approach (p becomes $p(t) = \bar{p} + N(t)$), it showed worse results than if a noise term was added to the whole equation 5. Since the second possibility is also much more simple we only use this one to evaluate the SDE approach. The number of runs for both Monte Carlo simulations was 100 (there was no significant difference if more runs were used).

In figure 2 the approximated means and variances of all methods are shown. The parameter values and initial conditions are: $K = 1$, $x(0) = 1$, $var(N(t)) = 0.1$, and $var(x(0)) = 0.01$. All approximated means are (approximately) the same. However, the variances are different. The SDE approach gives a higher variance than both the Monte Carlo simulations. It can be shown that the approximation of the mean and variance of the SDE obtained using equation 5 was mathematically quite correct. Therefore the comparison between the SDE approach and the Monte Carlo simulation with stochastic parameters (these are comparable since both use time varying parameters) becomes more a question of which result is more realistic: a variance which reaches a maximum of almost 40 percent or approximately 1 percent.

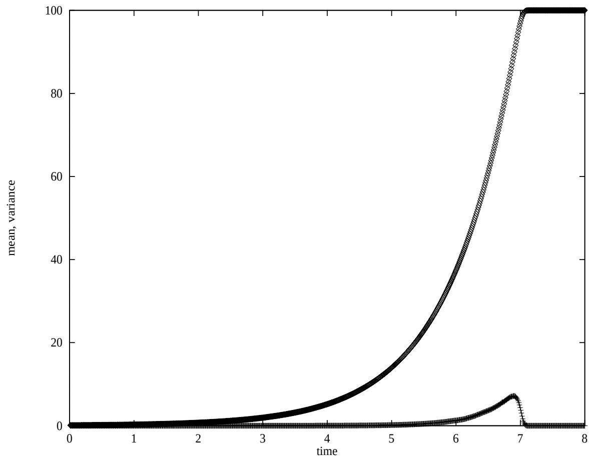

Figure 4: *Approximated mean(high) and variance(low) (MC with stoch. param.; noise on μ)*

Figure 3 gives the approximated mean and variance for the Monte Carlo Method with stochastic parameters, but this time with $x(0) = 0.1$ and $var(x(0)) = 0.001$ (So in both cases the variance of $x(0)$ is one percent of its actual value). In this case the determined ODE's to approximate the mean and variance proved to be unstable. It can be proven that the initial values and parameters have to be in a certain region for the ODE's to be stable. Therefore the SDE approach might not be very good whenever the true initial conditions and parameters are lying outside this interval. The normal Monte Carlo simulation also proved to be unstable but this might be caused by the integration algorithm itself (step size too large). A graph of Monte Carlo simulation with a stochastic μ ($var(\mu) = 0.1$) is given in figure 4. In this case the normal Monte Carlo simulation also proved to be unstable.

One may also conclude from these three figures that all methods show a maximum of the prediction uncertainty when the expected output almost reaches its maximum (at the 'second turn' in the S-curve). This was intuitively known by biologists but now it is 'mathematically' shown.

5 Conclusions

In this paper some methods for approximation the prediction uncertainty given a known model uncertainty, with respect to the Monod model, were being examined. All three methods showed a maximum of the prediction uncertainty when the expected output almost reached its maximum (at the 'second turn' in the S-curve). This was intuitively known by biologists but by using one of the methods examined in this paper it can be shown 'mathematically'. The example of Monod model is not a very difficult example which made it possible to test the results with the intuition of the biologists. However, if a model becomes more and more complex, it might become necessary to use one of these methods to determine the period(s) of high uncertainty in a model. The first method examined was Monte Carlo simulation. A disadvantage of this method is the assumption that parameters have to stay constant during one run. This problem was solved using the other two methods. The first one was Monte Carlo simulation with stochastic parameters and the second one a method which used stochastic differential equations (SDE approach). Both methods were able to incorporate time varying parameters. When examining the SDE approach it showed that all initial values and uncertainty values has to lie within a certain region for the approximations to be stable. This might become a problem if realistic values lie outside of this region[4]. Another disadvantage was the assumption of a Gaussian distribution for stochastic parameters or noise terms. The Monte Carlo simulation with stochastic parameters does not assume Gaussian distribution and might therefore be used where other probability distributions are involved. Another advantage is that this method will always be stable whenever the deterministic model is stable. When comparing the results of both method with the same initial values and parameters, the SDE approach showed a much higher maximum of the variance than the Monte Carlo simulation with stochastic parameters did. As mentioned previous, in this case the comparison between the SDE approach and the Monte Carlo simulation with stochastic parameters becomes more a question of which result is more realistic: a variance which reaches a maximum of almost 40 percent or approximately 1 percent.

[4] This might be solved by using stochastic numerical integration methods instead of an approximation using the Fokker-Planck equation

References

[Bag93] A. Bagchi. *Optimal Control of Stochastic Systems.* Series in Systems and Control Engineering. Prentice-Hall, London, UK, 1993.

[Bec87] M.B. Beck. Water quality modelling: a review of the analysis of uncertainty. *Water Resources Research*, 23(8):1393–1442, 1987.

[KP92] P.E. Kloeden and E. Platen. *Numerical Solutions of Stochastic Differential Equations.* Springer-Verlag, Heidelberg, 1992.

[Kre83] J.N. Kremer. Ecological implications of parameter uncertainty in stochastic simulation. *Ecological Modelling*, 18:187–207, 1983.

[Ste83] L. Steenhaut. Theoretische analyse van stochastische modellen voor microbiele groei (*in dutch*). Master's thesis, Universiteit Gent, Gent, België, 1983.

[Van94] P. Vanrolleghem. *On-line modelling of Activated Sludge Processes: Development of an Adaptive Sensor.* PhD thesis, Universiteit Gent, 1994.

[VK96] P.A. Vanrolleghem and K.J. Keesman. Identification of biodegradation models under model and data uncertainty. *Water Science Technology*, 33(2):91–105, 1996.

[VKC96] P.A. Vanrolleghem, Z. Kong, and F. Coen. Full-scale on-line asessment of toxic wastewaters causing change in biodegradation model strucure and parameters. *Water Science Technology*, 33(2):163–175, 1996.