Identification, verification and validation of process models in wastewater engineering: a critical review

A. Aziz Guergachi and Gilles G. Patry

ABSTRACT

This article presents a critical review of the existing methodologies for process mathematical modelling in the area of wastewater engineering. It is argued that model identifiability is not a major issue in mathematical modelling. Model verifiability is a very demanding criterion that can be replaced by a less stringent one: model observability. The issue of 'complex models versus reduced-order models' is to be resolved by introducing a new concept: optimal model complexity. The traditional procedures of model validation are not adequate and a mathematical framework for model quality evaluation is needed.

Key words | identifiability, mathematical modelling, observability, validation, verifiability, wastewater engineering

A. Aziz Guergachi (corresponding author)
School of Information Technology Management,
Ryerson University, 350 Victoria Street,
Toronto, Ontario,
Canada M5B 2K3
E-mail: *a2guerga@ryerson.ca*

Gilles G. Patry
Department of Civil Engineering,
University of Ottawa,
350 Cumberland,
PO Box 450, Station A,
Ottawa, Ontario,
Canada K1N 6N5
E-mail: *patry@uottawa.ca*

INTRODUCTION

Since the development of the IAWPRC (International Association on Water Pollution Research and Control) model (Henze *et al.*, 1987), research in the area of biological wastewater treatment (WWT) process mathematical modelling has focused on the following three subjects:

(a) Model identification and identifiability

Rationale. The design of process control strategies requires models that are uniquely identifiable, i.e. models for which a unique set of parameters can be determined through the model identification procedure (Jeppsson, 1996). Most WWT process models do not meet the identifiability criterion.

(b) Model verifiability

Rationale. For a model to be truly verifiable, all its state variables have to be directly measurable (Jeppsson, 1996). This is not the case for most existing models which are then considered to be only partly verifiable.

(c) Model reduction

Rationale. Complex models with many parameters are generally difficult to identify uniquely, hence the need for reduced-order models that may not describe the full dynamics of the system (Jeppsson, 1996).

In this paper, it is argued that model identifiability is not a major issue in mathematical modelling. Model verifiability is a very demanding criterion and it is suggested to replace it by a less stringent one: model observability. The issue of 'complex models versus reduced-order models' is to be resolved by introducing a new concept: optimal model complexity. The traditional procedures of model validation are not adequate and a mathematical framework for model quality evaluation is needed.

The next section presents some general definitions of the objects that will be used in this paper. Then, the four sections that follow discuss the issues of 'model identifiability', 'verifiability versus observability', 'complex models or reduced-order models' and 'model validation', respectively.

DEFINITIONS

Consider a system S whose state space X is a finite-dimensional one and assume that this system is described by a mathematical model M of the form:

$$\dot{\mathbf{x}} = \mathbf{f}(t,\mathbf{x},\mathbf{p}) \quad (1)$$

where $\mathbf{x}$ is the system state vector, $\mathbf{p}$ is the parameter vector, t is the time and $\mathbf{f}$ is a function that is nonlinear. Such a model is generally developed by processing and combining information (usually obtained from three possible sources: first principles, empirical data and empirical knowledge) about the dynamics of the system. Although several types of models can be developed for the same system (fuzzy logic, neural network, time series, etc), we will limit our attention in this paper to mathematical models of the type of Equation (1).

A fundamental problem in system modelling is the determination of the values of model parameters such that the corresponding response of the model equation approximates as closely as possible the actual response of the physical system. Assume that the response of the physical system is given in the form of a set of real data:

$$Y_N: \mathbf{x}^{\text{data}}(t_1), \mathbf{x}^{\text{data}}(t_2), \ldots, \mathbf{x}^{\text{data}}(t_N) \quad (2)$$

The mathematical procedure for determining the model parameters on the basis of a set Y_N of data is called *model identification* (or calibration). Traditionally, it consists in minimizing an objective function $J(\mathbf{p})$ such that

$$J(\mathbf{p}) = \sum_{k=1}^{N} \left\| \mathbf{x}(\mathbf{p},t_k) - \mathbf{x}^{\text{data}}(t_k) \right\|^2 \quad (3)$$

where $\mathbf{x}(\mathbf{p},t)$ represents the solution to the model Equation (1). If there is only one unique minimum for J then the system is defined as *identifiable* (Jeppsson, 1996). A system model is said to be *verifiable* if all its state variables are directly measurable (Jeppsson, 1996). After a model is developed and identified, we need to know how well it mimics the true system behaviour. The procedure of verifying this property is called *model validation*.

THE MODEL IDENTIFIABILITY ISSUE

The lack of identifiability has been considered a handicap for process models in the area of wastewater engineering. The most recent work on WWT process model identifiability is probably that of Jeppsson (1996). Using a simple example, Jeppsson showed that models that use the Monod equation are not identifiable. He then developed a set of reduced models for which he investigated the identifiability using computer simulations. He did not provide, however, a formal proof of the identifiability of these models.

In this paper, it is argued that model identifiability is not a major problem in mathematical modelling: a model that is not identifiable can still be useful if it produces a good performance. The arguments in favour of this view are two-fold.

Models of complex systems are practically impossible to uniquely identify

This is a fact. When Beck (1986) pointed out the lack of identifiability of the IAWPRC model, he immediately added that 'there is nothing unusual in this, for the same problem is widespread in the environmental sciences and in the adjacent disciplines of pharmacokinetics and biomedical system analysis. . . . It is well known that there are difficulties with structural identifiability of biochemical process models, specifically in association with the use of the Monod expression.' The lack of model identifiability is due to the fact that available models explain just a portion of the behaviour of highly complex systems. The other portion which is *not* accounted for by those models shows itself through the variability of model parameters. Lack of identifiability is then an inherent feature of complex systems.

Systems that are identifiable are usually associated with some unique values of parameters called *universal constants*. Other identifiable systems do not give rise to universal constants, but their model parameters are always reported to take the same unique values by all researchers. There are numerous systems with such properties in physical sciences. Here is a list of a few of them:

- Interaction of two electric charges: the magnitude of the force F resulting from the interaction in a free space of two charges q_1 and q_2 separated a distance r is expressed as

$$F = p\frac{q_1 q_2}{r^2}$$

where p is a parameter with a unique numerical value: $p = 1/4\pi\varepsilon_0$, ε_0 being equal to 8.854×10^{-12} SI.

- Interaction of two bodies: the magnitude of the force F of attraction between any two bodies is given by

$$F = p\frac{m_1 m_2}{r^2}$$

where p is a parameter with a unique numerical value: $p = 6.670 \times 10^{-11}$ SI, m_1 and m_2 are the masses of the two bodies and r is the distance between them.

- Equation of state of gases: for ideal gases, this equation is

$$PV = pnT$$

where p is a parameter with a unique numerical value: $p = 8.314\ 34 \times 10^3$ SI, P is the pressure, V is the volume, T is the absolute temperature and n is the number of moles. As the pressure gets higher and the temperature is close to the gas boiling point, the gas becomes non-ideal and is governed by the equation

$$\left(P + p_1\frac{n^2}{V^2}\right)(V - p_2 n) = pnT$$

where $p = 8.314\ 34 \times 10^3$ SI and p_1 and p_2 are two constant parameters that are gas-specific.

- The hydrogen bromine system: the reaction rate model for this system ($H_2 + Br_2 \rightarrow 2HBr$) is as follows:

$$r_{HBr} = \frac{1}{2}\left(\frac{p_1[H_2][Br_2]^{1/2}}{1 + p_2([HBr]/[Br_2])}\right)$$

where p_1 and p_2 are two uniquely determinable parameters: $p_1 \propto \exp(E_a/RT)$ with $E_a = 175$ SI, $R = 8.314\ 34 \times 10^3$ SI and $p_2 = 0.1$ (temperature-independent).

Identifiable systems, such as the ones presented above, all have one common property that can be expressed qualitatively in the following way.

Similar to what Cohen and Stewart did in their book *The Collapse of Chaos* (Cohen & Stewart, 1994), let us imagine that the information content of a system S can be measured by one single number $I(S)$. Considering the model M that is used to describe S as a mathematical system, its information content can also be measured by a number $I(M)$. Identifiable systems have the property that the two quantities of information $I(S)$ and $I(M)$ are almost equal. In more concrete terms, identifiable systems have models that explain practically all mechanisms governing the system behaviour. This is the case of all the foregoing examples of identifiable systems (*electrostatic interaction*, *gravitation*, *equation of state of gases*, *chemical system* H_2/Br_2, *etc*). However, in the case of a highly complex system, the quantity of information $I(M)$ is always strictly less than S, meaning that the model M does not account for all modes of the system behaviour. Jeppsson (1996) has expressed this fact very rightly for the case of the activated sludge WWT process: 'Though available models are quite complex they are still greatly simplifying the representation of many species of organisms. As the microbial population changes this needs to be reflected in changing kinetic parameters and even adding new state variables'. It is therefore the existence of a significant portion of the system behaviour not accounted for by the model M that renders the latter non-identifiable. WWT process models are then not identifiable and we have to live with this fact.

Model identifiability is not needed for systems control anyway

What is wrong with a system model that is not identifiable, but produces an acceptable performance in predicting the system behaviour? Nothing, if there is a mathematical guarantee of the model performance. A mathematical framework to help derive and establish such a guarantee is presented in Guergachi & Patry (2003). With this framework developed and the guarantee established, model identifiability becomes irrelevant.

Moreover, it should be noted that some emerging modelling technologies have also shown that model identifiability is not essential. Neural networks, for instance, are fundamentally non-identifiable, yet they

have been used extensively and successfully in several areas such as pattern recognition (Haykin, 1994).

VERIFIABILITY VERSUS OBSERVABILITY

Model verifiability requires that all state variables be directly measurable (Jeppsson, 1996). This is very demanding and not at all practical. Most systems, and especially complex ones, indeed have variables that are easily and directly measurable and others that are difficult or even impossible to measure. Because of that, it is suggested here to introduce a less stringent criterion called *model observability*. With this criterion, the system state variables do not have to be measured directly and separately. Rather, they are considered as 'hidden variables' and have somehow to be inferred from what can be measured (output).

Textbooks have defined the observability concept in several different ways which are all equivalent to each other (Mybeck, 1979; Ahmed, 1988; Borrie, 1992). Here we will consider the following definition.

Rewrite the model Equation (1) of the system S in a more general form:

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{f}(t,\mathbf{x},\mathbf{u},\mathbf{p}) \\ \mathbf{y} = \mathbf{h}(\mathbf{x}) \end{cases} \tag{4}$$

where $\mathbf{u}$ is a vector of input control variables (usually called either 'input' or 'control'), $\mathbf{y}$ is a vector of measured outputs (usually called just 'output') and $\mathbf{h}$ is a function of the state vector $\mathbf{x}$. The system S model is said to be *observable* if, given $\mathbf{y}(t)$ and $\mathbf{u}(t)$ for all $t \in [t_0, t_1]$, it is possible to uniquely determine the state vector $\mathbf{x}(t)$ for all $t \in [t_0, t_1]$. Thus the system model is observable if any state variable $x_{i(t)}$ can be determined for $t \in [t_0, t_1]$ from the knowledge of only the input and output over the interval $[t_0, t_1]$. The structure of an observable model must then be such that the output $\mathbf{y}(t)$ is affected in some manner by the change of any single state variable. In addition, the effect of any one state variable on the output must be distinguishable from the effect of any other state variable.

Here is a simple example to illustrate the concept of observability intuitively (adapted from Ahmed (1988)):

Example 1. Consider a system governed by the model equations:

$$\begin{cases} \dot{x}_1 = -p_1 x_3 + u_1 \\ \dot{x}_2 = u_2 \\ \dot{x}_3 = p_1 x_1 + u_3 \end{cases}$$

where x_1, x_2 and x_3 are the state variables, u_1, u_2 and u_3 are the inputs and p_1 is the model parameter ($\mathbf{x} = (x_1\ x_2\ x_3)^T$, $\mathbf{u} = (u_1\ u_2\ u_3)^T$, $\mathbf{p} = (p_1)$). Consider now the two following cases:

Case (i): the only variable that is measured is the sum of x_1 and x_2. In other words, the output $\mathbf{y}$ is a scalar and equal to $x_1 + x_2$. The model equations are then

$$\dot{x}_1 = -p_1 x_3 + u_1 \tag{I}$$

$$\dot{x}_2 = u_2 \tag{II}$$

$$\dot{x}_3 = p_1 x_1 + u_3 \tag{III}$$

$$y = x_1 + x_2 \tag{IV}$$

Can we reconstruct x_1, x_2 and x_3 from the knowledge of the values of only y and the controls u_1, u_2 and u_3? Yes. This is how we can proceed.

Add Equations (I) and (II). We get

$$\dot{y} = -p_1 x_3 + u_1 + u_2$$

Assuming of course that p_1 0, then

$$x_3 = \frac{u_1 + u_2 - \dot{y}}{p_1}$$

From Equation (III), we get x_1:

$$x_1 = \frac{1}{p_1}(\dot{x}_3 - u_3) = \frac{1}{p_1}\left(\frac{\dot{u}_1 + \dot{u}_2 - \ddot{y}}{p_1} - u_3\right)$$

From Equation (IV), we obtain x_2:

$$x_2 = y - x_1 = y - \frac{1}{p_1}\left(\frac{\dot{u}_1 + \dot{u}_2 - \ddot{y}}{p_1} - u_3\right)$$

Thus, this example shows that there is no need to measure directly and separately all the three state variables x_1, x_2 and x_3. If just the sum of x_1 and x_2 is measured (and the values of the control variables are assumed to be known to the operator, because she manipulates them), it is possible to uniquely reconstruct estimates $\hat{x}_1$, $\hat{x}_2$ and $\hat{x}_3$ for all the

three state variables, using the system input and output as the basis of this estimation. Because of this, the foregoing system (Equations (I)–(IV)) is observable. In practice, to check the observability criterion for linear systems, we just determine the rank of one matrix called the observability matrix (see for instance Ahmed (1988)). There is no need to go systematically through the above tedious algebraic calculations.

Case (ii): the only variable that is measured is x_1. In other words, the output **y** is a scalar and equal to x_1. The model equations are then

$$\dot{x}_1 = -p_1 x_3 + u_1 \quad \text{(I)}$$

$$\dot{x}_2 = u_2 \quad \text{(II)}$$

$$\dot{x}_3 = p_1 x_1 + u_3 \quad \text{(III)}$$

$$y = x_1 \quad \text{(IV)}$$

As said previously, it is easy to check the observability criterion of this linear system by computing the rank of the observability matrix (Ahmed, 1998) and establish that the system is not observable. In this example, however, the observability criterion is again examined intuitively:

From Equations (I) and (IV), we determine x_3:

$$x_3 = \frac{u_1 - \dot{y}}{p_1}$$

The variable x_1 is also uniquely determinable from Equation (IV):

$$x_1 = y$$

But the variable x_2 is, however, not uniquely determinable; any function

$$x_2 = x_{2_0} + \int u_2$$

with $x_{2_0} \in \Re$ is acceptable. Therefore, the model is not observable.

When a system model is observable, a *state observer* can usually be designed to generate an estimate of **x** using **u** and **y** as the basis for that estimation (Borrie, 1992).

The concepts of observability and observer were first introduced by Kalman in the early 1960s, but they have never been implemented in the area of WWT process mathematical modelling. The study of observability of linear systems is quite straightforward. It is, however, a challenging mathematical subject in the case of nonlinear systems such as WWT processes.

COMPLEX MODELS OR REDUCED-ORDER MODELS?

'An ''optimal'' model incorporates all of the important dynamic effects, is no more complicated in its structure than necessary, . . .' (Jeppsson, 1996). This is just another statement of the celebrated principle of simplicity commonly attributed to William of Ockham (1290?–1349?) and known as Occam's razor: 'If there are alternative explanations for a phenomenon, then, all other things being equal, we should select the simplest one' (Li & Vitányi, 1993).

In the case of the behaviour of biological WWT processes, however, we are faced with a complexity that is unparalleled in the chemical industry (Jeppsson, 1996). The reactions occurring in a bioreactor involve hundreds of different types of microorganisms biodegrading a multitude of different organic waste compounds. A simple bacterial cell in this bioreactor, *E. coli* for instance, has about 2,500 different *kinds* of macromolecules and contains about 24 million individual molecules (Madigan *et al.*, 1997). *E. coli*, as well as the other microorganisms present in the bioreactor, have to synthesize all of these molecules from the organic wastes so that they can grow and generate other organisms. In the course of this synthesis process, a large number of reactions take place involving the use of a multitude of different kinds of enzymes. It is the authors' opinion that a description with *scientific* accuracy of the bioreactor dynamics is unlikely to result in a model with a finite number of state variables and parameters.

However, even if such an accurate and highly complex model were possible to develop, it would be useless from an engineering viewpoint. The reason for this is not the

identifiability problem (as is suggested by Jeppsson (1996)), but is the *data scarcity*. If the size of the data set used for model identification is small while the number of model parameters is large (i.e. the model is highly complex), then the problem of *data overfitting* by the model would occur. On the other hand, if the model is too simple and, therefore, the number of parameters is sufficiently low compared to the size of the data set, then the explanatory power of the model would be so low that the value of the objective function (3) would become very high, meaning the model prediction of the true process behaviour is of a low quality. Consequently, the degree of complexity of a process model has to be adjusted to the amount of data that is available for the identification of this model. For any fixed amount of data, there is an *optimal* model complexity that has to be determined. Models that are more complex would cause overfitting, and models that are less complex would lead to a low prediction quality.

The mathematical framework that was developed in Guergachi & Patry (2003) defines all the necessary concepts and tools that help determine the optimal structure complexity of a WWT process model, corresponding to a fixed amount of data. In the following paragraphs, a qualitative explanation of the idea behind this framework is presented using some simple metaphors.

The identification procedure is viewed as an information transfer from a set of real data into the process model to be identified. Any data set Y carries a certain amount of information $I(Y)$ about the true process behaviour. This amount of information is characterized by the quantity and quality of the data. The quantity can be measured by the size N of the data set. The quality can be evaluated from a statistical point of view: the higher the statistical dependence among the elements of the data set, the less information Y carries. For a fixed size N of the data set Y, the latter contains maximum information when its elements are statistically independent. The process model M can be viewed as an information container. Its size is denoted $I(M)$. The more complex this model, the more information can be 'poured' into it from a real data set during the identification procedure.

Now consider a model M of the process under study and a fixed data set Y carrying an amount of information $I(Y)$. If M is too simple, the information container it represents will overflow during the identification procedure and, therefore, some of the information carried in the set Y will pour out and be lost. If, however, M is too complex, then the available amount of information $I(Y)$ will not be enough to fill up the model container completely. We will end up with an information container which is impressively large, but carrying very little information about the true process behaviour. Consequently, the best solution is to choose a degree of complexity for M such that $I(Y)$ matches $I(M)$.

THE ILLUSION OF MODEL VALIDATION

When a model is developed and identified, it needs to be validated. The procedures that are used for model validation have, however, been criticized not only in the area of WWT process mathematical modelling, but also in several other areas of science and engineering. Jeppsson (1996) pointed out that 'in strict sense, model validation is impossible'. Similarly, Zheng & Bennett (1995) noted that '*process* models, like any scientific hypothesis, cannot be validated in the absolute sense ... They can only be invalidated.' Konikow & Bredehoeft (1992) suggested that terms such as model verification and model validation convey a false sense of truth and accuracy and thus should be abandoned in favour of more realistic assessment descriptors such as history-matching and benchmarking. Several other researchers have expressed similar criticisms about the model validation issue (Oreskes *et al.*, 1994; Beck *et al.*, 1997: Bohlin, 1993).

An original approach to deal with this question of model validation is proposed in Guergachi & Patry (2003), where a mathematical framework for model quality evaluation is developed. The basic idea of this framework is explained below in more mathematical terms.

Let M be a model of the system S and suppose we are interested in the model predictions of the i_0th state variable x_{i_0} of S. We need to know the prediction accuracy of the model and also the risk of getting significant deviations between model predictions and the system response.

During the identification procedure, the model M 'sees' only a finite number of examples:

$x_{i_0}^{\text{data}}(t_1), x_{i_0}^{\text{data}}(t_2), \ldots, x_{i_0}^{\text{data}}(t_N)$

(the elements of the data set are called examples). However, the user expects the model to produce good predictions not only for the situations that it has seen before, but also for the other unseen situations that will occur in the real-world operation of the system. Consequently, the system modeller should strive to make sure that, by minimizing the objective function

$$J_{i_0}(\mathbf{p}) = \sum_{k=1}^{N} \left| x_{i_0}(\mathbf{p},t_k) - x_{i_0}^{\text{data}}(t_k) \right|^2 \tag{5}$$

or equivalently the arithmetic mean value:

$$R_{\text{emp}}(\mathbf{p}) = \frac{J_{i_0}(\mathbf{p})}{N} = \frac{1}{N} \sum_{k=1}^{N} \left| x_{i_0}(\mathbf{p},t_k) - x_{i_0}^{\text{data}}(t_k) \right|^2 \tag{6}$$

the expected time average:

$$R(\mathbf{p}) = \lim_{n \to +\infty} \frac{1}{n} \sum_{k=1}^{n} \left| x_{i_0}(\mathbf{p},t_k) - x_{i_0}^{\text{data}}(t_k) \right|^2 \tag{7}$$

will also become minimized. This is because the *true* measure of the model performance is not the empirical objective function $J_{i_0}(\mathbf{p})$ or the arithmetic mean $R_{\text{emp}}(\mathbf{p})$, but the expected average $R(\mathbf{p})$ of the infinite time sequence:

$$| x_{i_0}(\mathbf{p},t_1) - x_{i_0}^{\text{data}}(t_1) |^2, | x_{i_0}(\mathbf{p},t_2) - x_{i_0}^{\text{data}}(t_2) |^2,$$
$$| x_{i_0}^{\text{data}}(\mathbf{p},t_3) - x_{i_0}^{\text{data}}(t_3) |^2, \ldots$$

However, the value of $R(\mathbf{p})$ is not known. Therefore, we end up with the following situation:

- $R_{\text{emp}}(\mathbf{p})$ is merely an empirical measure of the model performance, but its numerical value is accessible to us;
- $R(\mathbf{p})$ is the exact measure of the model performance, but its value is inaccessible to us.

The whole question here is then how to infer information about the exact model performance measure $R(\mathbf{p})$ from the knowledge of the value of the empirical measure $R_{\text{emp}}(\mathbf{p})$. Addressing this question and developing a methodology for model quality evaluation can be found in Guergachi & Patry (2003).

CONCLUSION

A critical review of the existing methodologies for process mathematical modelling in the area of wastewater engineering was presented. The model identifiability issue was discussed: it was argued that identifiability is not required in mathematical modelling. The criterion of model verifiability was deemed very demanding, and it was suggested to replace it by a less stringent one: model observability. The issue of 'complex models versus reduced-order models' was also discussed; it was explained that this issue can be resolved by introducing a new concept: optimal model complexity. It was argued that the traditional procedures of model validation are not adequate and it was explained that a mathematical framework for model quality evaluation was needed.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support of CIDA and NSERC.

REFERENCES

Ahmed, N. U. 1988 *Elements of Finite-Dimensional Systems and Control Theory*. Longman Scientific & Technical, London.

Beck, M. B. 1986 Identification, estimation and control of biological wastewater treatment processes. *Proc. IEE* **133**, 254–264.

Beck, M. B., Ravetz, J. R., Mulkey, L. A. & Barnwell, T. O. 1997 On the problem of model validation for predictive exposure assessments. *Stochast, Hydrol. Hydraul.* **11**, 229–254.

Bohlin, T. 1993 Validation of identified models. In *Concise Encyclopedia of Environmental Systems* (ed. Young, P. C.), pp 645–650. Pergamon Press, Oxford.

Borrie, J. A. 1992 *Stochastic Systems for Engineers: Modeling, Estimation, and Control*. Prentice Hall, Englewood Cliffs, NJ.

Cohen, J. & Stewart, I. 1994 *The Collapse of Chaos*. Viking, New York.

Guergachi, A. & Patry, G. 2003 Using statistical learning theory to rationalize system model identification and validation. Part I: Mathematical foundations. *Complex Syst. J.* **14**(1), 63–90.

Haykin, S. 1994 *Neural Networks, A Comprehensive Foundation*. Macmillan, New York.

Henze, M., Grady, C. P. L., Gujer, W., Marais, G. v. R. & Matsuo, T. 1987 *Activated Sludge Model No. 1*. IAWPRC, London.

Jeppsson, U. 1996 *Modelling Aspects of Wastewater Treatment Processes*. Lund Institute of Technology, Stockholm.

Konikow, L. F. & Bredehoeft, J. D. 1992 Ground-water models cannot be validated. *Adv. Water Res.* **19**, 75–83.

Li, M. & Vitányi, P. 1993 *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, New York.

Madigan, M. T., Martinko, J. M. & Parker, J. 1997 *Biology of Microorganisms*. Prentice Hall, Engelwood Cliffs, NJ.

Mybeck, P. S. 1979 *Stochastic Models, Estimation and Control*. Academic Press, New York.

Oreskes, N., Shrader-Frechette, K. & Belitz, K. 1994 Verification, validation, and confirmation of numerical models in the earth sciences. *Science* **263**, 641–646.

Zheng, C. & Bennett, G. D. 1995 *Applied Contaminant Transport Modeling–Theory and Practice*. Van Nostrand Reinhold, New York.