Software for Identification of Ill-defined Systems: a Water Quality Example

K. Keesman
Environmental Systems Engineering Group, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

Abstract

This paper describes an information system (STEPS) designed to support the identification of ill-defined systems, and subsequent use for prediction of their behaviour. Ill-definedness is brought about by unavoidable inadequacies in model structure, usually in conjunction with sparse and unreliable empirical data. The uncertainty modelling used in STEPS is based on set-theoretic concepts, i.e. the uncertainties are expressed in terms of bounds, and not in terms of statistical parameters. The set-theoretic framework is outlined briefly. To assist the identification STEPS also contains recursive parameter estimation tools based on the stochastic concept rather than the set-theoretic concept. STEPS also provides support tools for data management, for model structure improvement and for the construction of predictions with the model. The information system is demonstrated by applying it to the identification of a simple dissolved oxygen model for a lake.
KEYWORDS: system identification, uncertainty analysis, predictions

1. INTRODUCTION

In modelling ill-defined systems it appears that the set-theoretic way of uncertainty modelling is a more appropriate choice than the stochastic way (Keesman and Van Straten [1]-[4]). In the latter case, several assumptions must be made, while in the former case the only assumptions are that the uncertainties are bounded (unknown-but-bounded model). Then the observation uncertainty, representing measurement and sampling uncertainty but also unidentifiable model error, belongs to a set. As a result each of the observations is specified as unknown-but-bounded. The observations with the associated uncertainty bounds span the so-called behaviour space. Consequently, estimation of the model parameters results not in a single 'optimal' parameter estimate, but in a set of equally acceptable parameter vectors. This set of parameter vectors spans the so-called behaviour-giving parameter space. The set of behaviour-giving parameter vectors (see Fig. 1) is consistent with the predefined parameter ranges (expressing the a priori parametric uncertainty), the model structure and the specified observation uncertainty bounds.

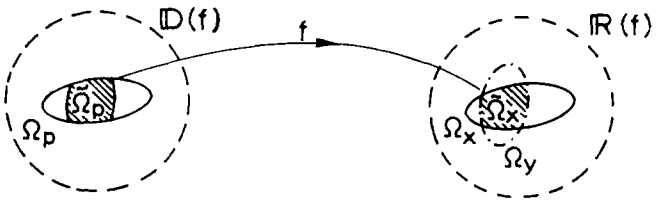

Fig. 1. Venn-diagram of the relation between parameter and observation space. D(f), R(f) : global parameter and observation space; f: vector function; Ω_p: parameter space; Ω_y: observation space; Ω_x: model response space.

During the modelling process of ill-defined systems it frequently appears that the presumed model structure of aggregated processes is not completely correct. This situation of so-called unstructured uncertainty can be handled according to two approaches.

First, on the basis of prior system knowledge a complex model, representing all kind of expected effects, can be set up with many unknown parameters. In this way unstructured uncertainty is, at least patially, converted into so-called structured uncertainty. The problem, then, is how to estimate the large number of parameters properly. Recently, Walter et al. [5] have proposed a method to estimate non-uniquely identifiable parameters from observations with bounded noise. Some years before Fedra et al. [6] proposed an identification method within the set-theoretic context based on Monte Carlo simulations (see also [1], [2]). This method yields a number of realizations of behaviour-giving parameter vectors out of a predefined parameter space on the basis of individual parameter intervals. In this way the presence of structural model error (unstructured uncertainty) is compensated for by the set of behaviour-giving parameter vectors (structured uncertainty). It appears, however, that this compensation is not always complete. Within the set-theoretic context Keesman and Van Straten [3] have found an estimate of the uncompensated model error.

The second approach to handle the presence of unstructured uncertainty is based on additional analysis of the system (posterior information). Depending on the results equations are added to the model to represent the omitted effects. Within this context of inductive modelling Beck and Young [7] have stressed the close relationship between model identification and parameter estimation by applying an extended Kalman filter (EKF) as a structural identification procedure (see also [8]-[10]). It must be noted then that this tool, which is essentially based on stochastic concepts, is applied for lack of a recursive identification algorithm for nonlinear-in-the-parameters models within a set-theoretic setting.

Fore-mentioned ideas form the basis of the proposed methodology of identification of ill-defined systems. Together with data management and

Paper received on 25 August 1988 and in final form on 11 January 1989 Referee: Dr. M.B. Beck

data analyses tools a computer-aided set-theoretic/stochastic identification framework for applications to ill-defined systems has been developed. In this paper we like to emphasize the organization of different procedures within this framework and not the algorithms itself. The paper further describes and illustrates the resulting information system (STEPS = Set-Theoretic Estimation in Poorly-defined Systems) by application to a simple water quality model.

2. ARCHITECTURE OF THE INFORMATION SYSTEM

For reasons of flexibility the information system (STEPS) consists of a preprocessor and a master program, containing different modules. The preprocessor has two main functions, i.e. formulation of a state-space model in a predefined code and selection of modules to be used in the master program. In Fig. 2 a window from STEPS is presented containing the available modules with a short description of the modules.

1 System Utilities & Data management
System specification and data file manipulation

2 Data Analyses
Information about data

3 Simulation
Model initialization and simulation

4 Parameter Identification
Search for sets of behaviour-giving parameter vectors

5 Model Structure Identification
Search for model modifications

6 Model Application
Prediction and verification of the model

Fig. 2. Available modules in master program.

Apart from a menu-driven application of the program, it is possible to load each of the modules separately. Each of the modules can then be included in a programmer 's code. The first three modules are supporting tools for the remaining modules, which form the body of the framework. An Input-Process-Output (IPO) scheme of these main modules is presented in Fig. 3.

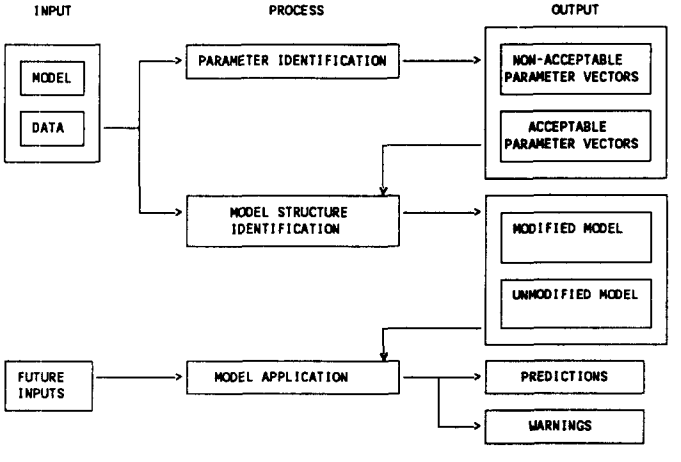

Fig. 3. IPO-scheme of main body.

2.1 Parameter identification

The aim of this module is to identify a set of behaviour-giving parameter vectors and to obtain information about the model validity. The inputs and outputs are presented in Fig. 4. The key tools are the behaviour definition procedure, space scanning, space identification, min-max estimation and model/data discrimination. It must be noted that the processes indicated within this framework can be handled less rigidly than the static representation of the scheme suggests. Different alternatives to the solution of the parameter identification problem will be emphasized, but first each of the presented processes are described in short.

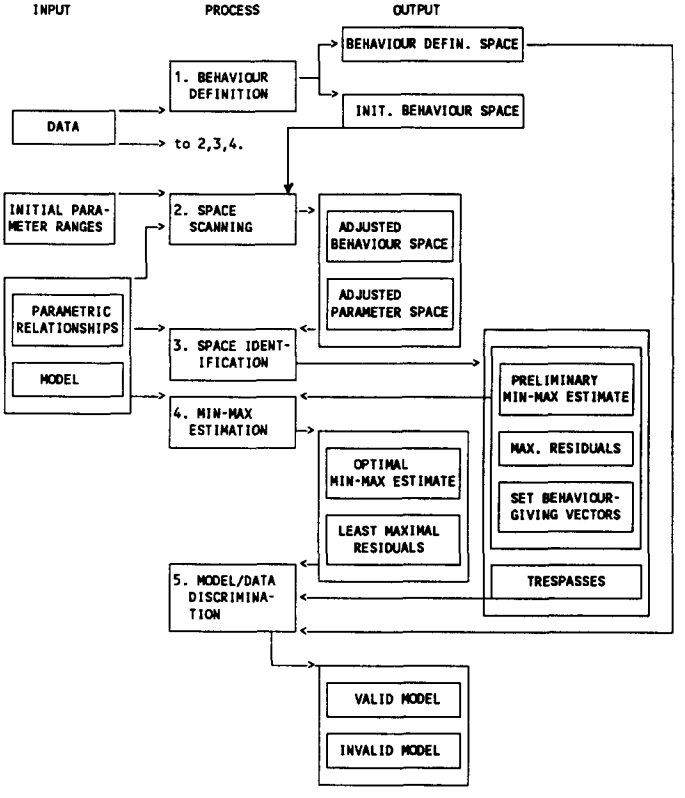

Fig. 4. IPO-scheme of parameter identification module.

1. behaviour definition

This procedure offers the possibility of specifying interactively a behaviour space by means of the observed output variables and associated fuzzy set-membership function ([1]). This membership function expresses the (decreasing) possibility of occurrence of the noise free observation with respect to the actual observation. The fuzzy set-membership function is characterized by the predefined observation error bounds and the shape of the function (for instance a rectangle, trapezium or triangle).

Unlike the initial behaviour-space, representing both measurement and structural model uncertainty, the behaviour definition space is only specified in view of discriminating between a valid and an invalid model. Therefore the behaviour definition space is specified under the hypothesis that the model is correct, which means that the bounds are chosen on the basis of measurement uncertainty alone.

2. space scanning

Space scanning (see [1]), on the basis of randomly within the predefined parameter space selected parameter vectors, is performed to indicate inconsistencies between the initial behaviour space and predefined parameter space. For the sake of efficiency of the subsequent space identification procedure it is important to indicate and remove outliers from the behaviour space in order to avoid an empty parameter set. An outlier is defined as an observation causing the behaviour-giving parameter space to be void under the assumption that the model is valid. The frequency of trespasses of the model responses with respect to the initial behaviour space at the various observation time instants provides information about the presence of outliers. Moreover, the efficiency is also improved by a preliminary indication where to find the behaviour-giving parameter space. This information is presented by so-called Box-and-Whisker plots ([11]), representing quartiles and limits associated with a certain behaviour-giving interval for each of the individual parameters.
Within this procedure the criterion for an acceptable parameter vector is specified in terms of a maximum number of allowable trespasses for each simulation.

3. space identification

On the basis of the adjusted behaviour and parameter space an algorithm can be run which will supply now a non-void set of behaviour-giving parameter vectors consistent with the behaviour and parameter space and the specified model. Algorithms suitable for solving the set-theoretic parameter estimation problem are the so-called polytope-bounding algorithms. However, these algorithms are only applicable for models which are linear-in-the-parameters (see Walter and Piet-Lahanier [12] for an overview). From the standpoint of robustness we chose an iterative random scanning procedure to solve the problem. This random scanning procedure is performed in conjunction with intermittent parameter space translations and rotations in order to improve the computational efficiency. The parameter space adjustment algorithm, which can be applied to nonlinear-in-the-parameters models, has been described in detail by Keesman and Van Straten [2]. This algorithm can be run automatically. The interpretations of the results to obtain information about parameter subspaces (see [4] for an interpretation of the results in terms of dominant directions), however, must be done interactively.

4. min-max estimation

Min-max estimation which results in a parameter estimate that minimizes the maximum deviation between model response and observation is performed for diagnostic purposes only ([1], [3]) using the 'pattern search' technique of Hooke and Jeeves [13]. The min-max estimation reveals 'hard' information about the minimal observation uncertainty bound to be specified to prevent a void behaviour-giving parameter space. From this point of view it can be interesting to perform min-max estimation prior to space identification (see [3]).

5. model/data discrimination

On the basis of results from preceding procedures analyses of: (i) critical data points, (ii) the min-max estimation and (iii) the model response space associated with behaviour-giving parameter vectors are performed to obtain information about the reliability of model and data.

It must be emphasized that supporting (most of the time graphical) information is provided and certainly it is not a final decision. That is, observations are indicated as outliers if a high percentage of trespasses occurs at the boundaries of the behaviour space at those observations. From the min-max estimation the most critical observation, associated with the maximal residual, indicates an outlier if the boundaries of the behaviour definition space, representing only the measurement and sampling error, are trespassed. If this observation appears to be unreliable, than it is considered as an outlier. Otherwise most likely the model structure is invalid. Analysis of the model response space with respect to the observations reveals the presence of uncompensated structural model error ([3]).

To emphasize the flexibility of the presented scheme (Fig. 4) some alternative approaches are discussed now.

Note that, due to wrong assumptions about the initial parameter space or behaviour space, it is quite possible that the resulting set of behaviour-giving parameter vectors is void. It is evident that a void set yields no information at all about where to find the behaviour-giving parameter vectors. To detect this situation the space-scanning experiment is performed. If this situation arises due to outliers in the observations Fourier filtering can be applied, as an alternative to the space scanning experiment, to remove these outliers. Of course it is also possible that one starts with large uncertainty bounds, which are adjusted iteratively using fuzzy set-theoretic information of parametric subspaces associated with smaller behaviour spaces. During these iterations information about outliers will come to light.

Another point to emphasize concerns the validity of the model. It was recognized ([1]) that, in spite of the invalidity of the model, the model could still be accepted from a practical point of view for an enlarged behaviour space. The associated behaviour-giving parameter space as well as the prediction uncertainty will reflect this concession.

2.2 Model structure identification

The aim of this module is to provide supporting information for model adjustments. The key tools within this module are the regionalized sensitivity analysis (RSA), the recursive parameter estimation, and the correlation analysis. The inputs and outputs to the processes are presented in Fig. 5.

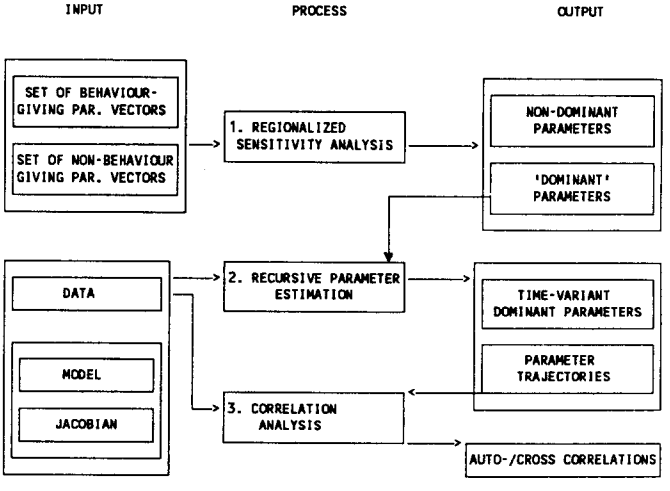

Fig. 5. IPO-scheme of model structure identification module.

1. regionalized sensitivity analysis

The sets of behaviour- and non-behaviour-giving parameter vectors are obtained from plain Monte Carlo simulation analysis on the basis of predefined parameter distributions. Of both sets, statistical analysis reveals dominant parameters and/or dominant parameter combinations at a certain confidence level ([14]).

2. recursive parameter estimation

The 'dominant' parameters resulting from RSA are input to the recursive parameter estimation tool (see [15]). As a recursive estimation tool the Extended Kalman Filter (EKF) is chosen, which can handle nonlinear model structures and state-parameter estimation. It must be noted that for application of the EKF additional linearized equations are required, which means that Jacobians of model and measurement vector functions must be supplied for an analytical treatment of the linearization. The EKF results in a number of time-variant dominant parameters with associated parameter trajectories. It is worth noting that, as an alternative to the EKF, also stochastic observers and stochastic approximation methods, requiring less detailed statistical information, could have been applied.

3. correlation analysis

The ultimate cross-correlations between time-variant parameter trajectories and system in- and outputs provide information about correlated effects (for instance [16]). The resulting auto-/cross-correlations, which are presented graphically, can sometimes be interpreted in terms of causal effects. It is worth noting that human intervention is indispensible in this stage of the identification. Supplementary data and additional theoretical information about the processes must then be employed to improve the model structure.

2.3 Model application

As yet the emphasis of the framework is on modelling for prediction or projection and not control. Projection refers to the situation where the internal description of the system dynamics must be changed according to future structural changes in the environmental system ([17]). Thomann [18] has stressed the importance of subsequent examination and verification of model predictive performance using the actual information of the systems' state. Therefore, STEPS contains, in addition to a model prediction tool, also a model verification tool (see Fig. 6).

1. model prediction

On the basis of the specified model structure, the set of behaviour-giving parameter vectors and the future inputs, bounded (as a result of the set-theoretic approach) predictions or projections are provided at desired time instants. There is also an option to obtain additional information at these time instants from frequency distributions.

2. model verification

The predictions are used to verify the model by looking at the trespasses when new observations are available. These observations are presented graphically in a plot of the prediction uncertainty bounds.

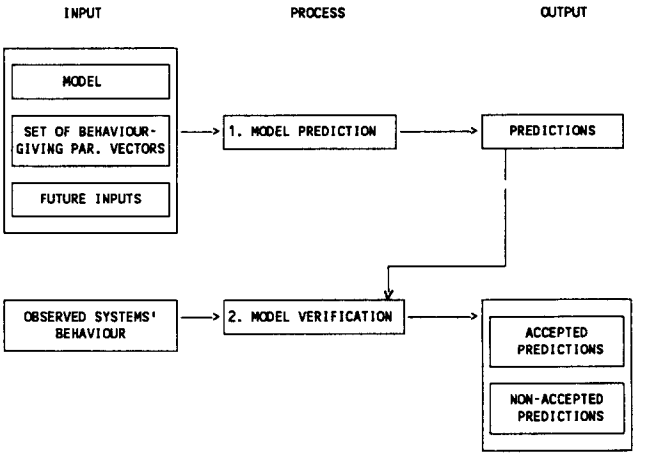

Fig. 6. IPO-scheme model application module.

3. A WATER QUALITY MODELLING EXAMPLE

3.1 System description

To illustrate STEPS a simple dissolved oxygen modelling example is chosen, which describes the DO-concentrations in a well-mixed lake. Hourly observed DO-concentrations are available from Lake De Poel and 't Zwet (The Netherlands) during a period of eight days (see Fig. 7). The changes in DO-concentrations are determined by reaeration exchange with the atmosphere, photosynthetic production from algae and water plants, and consumption due to respiration, biodegradation and sediment processes. The model equation is,

$$c(t) = Kr\,(Cs(t) - c(t)) + a\,I(t) - R \qquad (1)$$

where $c(.)$ = dissolved oxygen concentration (g/m^3)
$Cs(.)$ = saturation concentration (g/m^3)
$I(.)$ = radiation (W/m^2)
Kr = reaeration coefficient ($1/d$)
a = photosynthetic rate coefficient (g/mdW)
R = sink term (g/m^3d),

Note that the terms in the right hand side of (1) represent lumped processes. That is, by lack of detailed knowledge about the processes determining the rate of change of DO-concentrations, we are urged to aggregate processes, which intrinsically means that we incorporate some structural model uncertainty. In addition, the observed DO-concentrations will contain systematic error due to spatial concentration gradients in the lake, yielding non-representative samples, and fouling of the sensors. In such situations a set-theoretic approach to model fore-mentioned uncertainties is an appropriate choice. It must be noted that the system inputs, i.e. saturation concentration, as a function of the water temperature, and radiation (Fig. 8), are treated as deterministic variables, i.e. system input noise is lumped in an 'output' error.

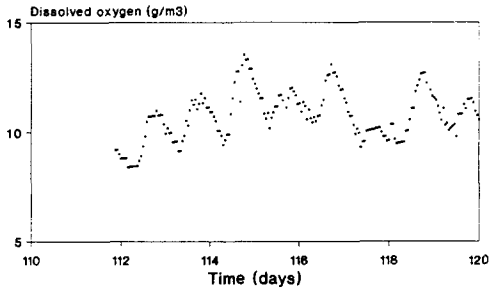

Fig. 7. Observed DO-concentrations in Lake De Poel and 't Zwet (The Netherlands) from 21-30 April 1983.

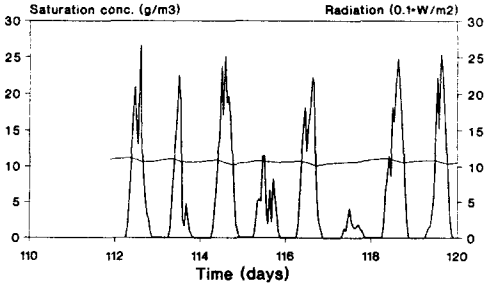

Fig. 8. System inputs to DO-model of Lake De Poel and 't Zwet.

3.2 Parameter space identification

On the basis of the complete data set a behaviour space is defined (see Fig. 4) in terms of the observed DO-concentrations plus/minus 1.5 g/m^3. Subsequently, a space scanning and a parameter space identification is performed. In previous papers ([1], [2]) different results with respect to this DO-model and data are presented. So, here we shall restrict ourselves to a graphical presentation (screen dump) of the set of behaviour-giving parameter vectors (Fig. 9). In this figure, each of the parameter vectors, containing the parameters Kr, a and R, is projected onto the faces of a three-dimensional box, that is spanned by the three predefined parameter intervals.

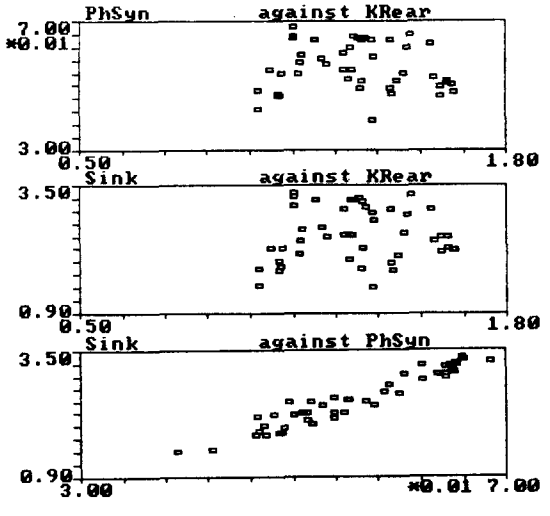

Fig. 9. Set of 45 behaviour-giving parameter vectors.

If there is knowledge about the bounds on the measurement uncertainty (behaviour definition space, see Fig. 4), then this knowledge can be explored in a model/data discrimination procedure (see [1]).

3.3 Recursive parameter estimation

In this example no sensitivity analysis (see Fig. 5) has been done in order to reduce the number of parameters to be estimated. So, all three parameters (Kr, a and R) and the state variable (c) are estimated simultaneously (see Eqn. 1). In a general feedback form for a single-output system the EKF-algorithm can be presented as,

$$x(k/k) = x(k/k-1) + K(k/k-1)\,\{y(k) - h[x(k/k-1)]\} \qquad (2)$$

where the first term on the right hand side is the predicted 'state' (usual state vector augmented with the unknown parameters) and the second term is a correction factor. The correction of the predicted state depends on the weighing matrix K(.) and the prediction error presented as the difference between the observation y(.) and the model output. To fit the DO-model (Eqn. 1, supplied by the user in a predefined code) to a continuous-discrete version of EKF the following state-space notation is used, i.e.

$$x(t) = f[x, u, t] + w(t) \qquad (3a)$$

$$y(k) = h[x, u, k] + v(k) \qquad (3b)$$

where x = state/parameter vector $[c, Kr, a, R]^T$
u = deterministic input vector $[Cs, I]^T$
w = system noise vector
y = observation (c_{meas})
v = observation noise
f,h = vector functions

The vector functions **f**[.] and **h**[.] have been chosen to take the following forms for the DO-model,

$$f[.] = \begin{bmatrix} Kr\,(Cs(t)-c(t)) + a\,I(t) - R \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad (4a)$$

$$h[.] = [\,c\,]_k \qquad (4b)$$

Substitution of (4a) in (3a) reveals that the parameter variations are modelled as essentially constant but with random fluctuations (random walk model). So the white noise vector w is composed of a component representing the uncertainty in model structure and inputs, and three components representing the uncertainty in the constant parameter model structure. A first parameter tracking, assuming a constant parameter model, i.e. $E\{w_i\}=0$ and $var\{w_i\}=0$; i=2,...,4, reveals the presence of possible outliers in the observed DO-concentration causing a jump in the parameter trajectories without demonstrable jumps in inputs at time instants 114.666, 115.666 and 118.875 (see Fig. 10).

Removal of these outliers, using one of the procedures in the data management tool, results in a smoother course of the parameter estimates. To explore the time-variability of the parameter trajectories more fully, the parameter models are modified using a nonzero additive white noise term in (3a), i.e. $E\{w_k w_j\} = Q\delta_{kj}$. The diagonals of the covariance matrix Q are chosen to be proportional to final variances of the associated parameter estimation errors of the preceding estimation run. The results of this time-variant parameter estimation can be seen in Fig. 11, where the initial values for parameters and parametric uncertainty also result from the preceding run.

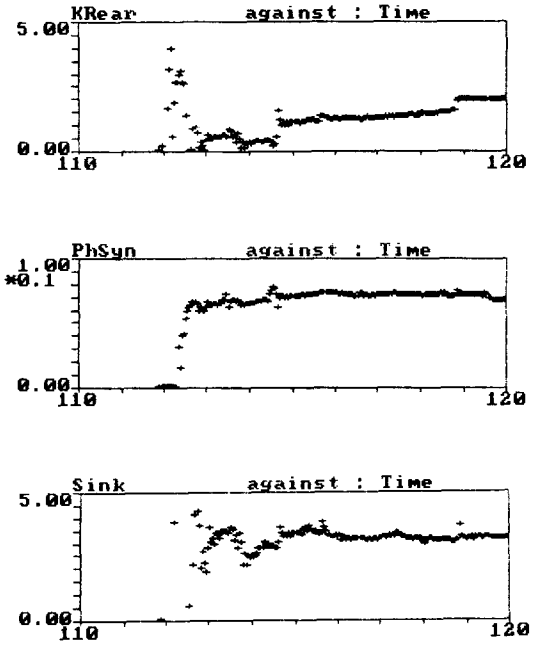

Fig. 10. Recursive estimates of the parameters Kr, a and R.

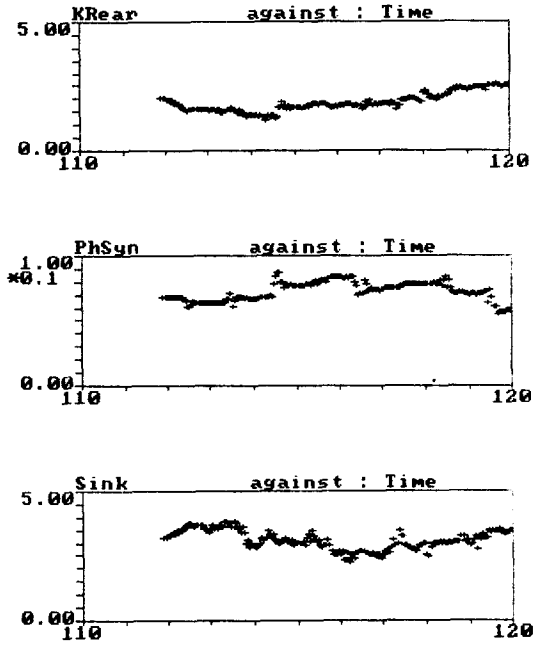

Fig. 11. Recursive estimates of the time-variant parameters Kr, a and R.

3.4 Correlation analysis

In previous papers ([1], [3]) we have noticed that the model is most likely invalid in view of set-theoretic criteria. This notion is confirmed by the nonstationary course of the parameter trajectories in Fig. 11 (see [7]). It is useful then to try to improve the model. Supporting information from the recursive estimates, representing the changes of the parameter values in time due to perturbations of the system, can be obtained by correlating these trajectories to observed system information.

From our example it appears that the recursive estimate of the sink term R is significantly correlated with the temperature ($r \approx -0.3$). Therefore possible improvement of the model performance can be obtained by incorporation of a temperature function in the dissolved oxygen consumption process. It is worth noting that the form of the final model adjustment is still a question to be answered. Knowledge about the processes, to be obtained from ecologists, biologists etc., is then indispensible. We have tried to improve the model by multiplying the sink term R with the temperature function θ^{T-20}, where $\theta<1$ as a result of the negative correlation. However, analyses of this extended model reveal that our adjustment does not significantly improve the model performance.

3.5 Model application

Even if we have to conclude that the model is invalid, we can still utilize it for predictions as all structural model uncertainty is represented by the identified parametric uncertainty. Alternatively, an explicit model error term can be added to the predictions (see [3]).

4. CONCLUDING REMARKS

The information system STEPS supports the identification of ill-defined systems for predictive purposes. Within the framework presented structural models, which are possibly nonlinear-in-the-parameters, can be handled easily using an algorithm based on Monte Carlo simulation, and set-theoretic (unknown-but-bounded) uncertainty models. Besides set-theoretic uncertainty models also stochastic uncertainty models (in for instance stochastic observers, stochastic approximation methods or the implemented Extended Kalman Filter) are used to obtain additional information from the data to improve the model. Predictions are based then on the (extended) model and the associated set of behaviour-giving parameter vectors resulting in prediction uncertainty bounds.

5. SYSTEM REQUIREMENTS

STEPS has the following hardware and software requirements:

- IBM PC, XT, AT or compatible computers
- minimum memory of 512 kB
- graphics display card: CGA, Hercules, EGA or VEGA
- 8087, 80287 or 80387 Math Coprocessor
- MS-DOS version 2.0 or more recent
- at least two 360 kB drives
- Epson printer
- Turbo Pascal 3.0 or 4.0 compiler

Optional,

- MATHPAK 87 version 2.0 or 3.0 for Turbo Pascal.

ACKNOWLEDGEMENTS

I am grateful to Gerrit van Straten for his comments and helpful suggestions. This research is supported by the Netherlands Technology Foundation (STW).

REFERENCES

1 Keesman K.J. and van Straten G. Modified set-theoretic identification of ill-defined water quality system from poor data. *Proc. of IAWPRC Symp. Systems Analysis in Water Quality Management*, Pergamon Press, Oxford, 297-308, 1987.

2 Keesman K.J., and van Straten, G. Embedding of random scanning and principal component analysis in set-theoretic approach to parameter estimation. *Proc. of 12th IMACS World Congress on Scientific Computation*, Paris, II.490-492, 1988.

3 Keesman K.J., and van Straten G. Identification and prediction propagation of uncertainty in models with bounded noise. To appear in *Int. J. Control.*

4 Keesman, K.J. On the dominance of parameters in structural models of ill-defined systems. To appear in *Appl. Math. Comp.*

5 Walter E., Piet-Lahanier H., and Happle J. Estimation of non-uniquely identifiable parameters via exhaustive modeling and membership set theory, *Math. Comp. Simul.* 1986, **28**, 479-490.

6 Fedra K., van Straten G., and Beck M.B. Uncertainty and arbitrariness in ecosystems modelling: A lake modelling example, *Ecol. Model.* 1981, **13**, 87-110.

7 Beck M.B., and Young, P.C. Systematic identification of DO-BOD model structure. *Proc. A.S.C.E., J. Environm. Eng. Div.* 1976, **102**(EE5), 909-927.

8 Beck M.B. Random signal analysis in an environmental sciences problem, *Appl. Math. Model.* 1978, **2**(1), 23-29.

9 Beck M.B. Lake eutrophication: identification of tributary nutrient loading and sediment resuspension dynamics, *Appl. Math. Comp.* 1985a, **17**, 433-458.

10 Whitehead P.G., Beck M.B., and O'Connell P.E. A systems model of streamflow and water quality in the Bedford Ouse river system - II. Water quality modelling, *Water Research* 1981, **15**, 1157-1171.

11 Tukey, J.W. *Exploratory Data Analysis.* Addison-Wesley, Reading, Massachusetts, 1977.

12 Walter E., and Piet-Lahanier H. Estimation of parameter bounds from bounded-error data: a survey. *Proc. of 12th IMACS World Congress on Scientific Computation*, Paris, II.467-472, 1988.

13 Hooke R., and Jeeves T.A. "Direct search" solution of numerical and statistical problems, *J. Assoc. Comp. Machinery* 1961, **8**, 212-229.

14 Hornberger G.M., and Spear R.C. An approach to the preliminary analysis of environmental systems, *J. Environm. Managem.* 1981, **12**, 7-18.

15 Whitehead P.G., and Hornberger G.M. Modeling algal behaviour in the river Thames, *Water Research* 1984, **18**, 945-953.

16 Beck M.B. Structure, failure, inference and prediction. In: *Identification and System Parameter Estimation*, (Eds. H.A. Barker and P.C. Young), Pergamon Press, Oxford, 1443-1448, 1985b.

17 van Straten G. *Identification, uncertainty assessment and prediction in lake eutrophication.* Ph.D. Thesis, Dept. Chem. Eng., University of Twente, The Netherlands, 1986.

18 Thomann R.V. Verification of water quality models, *Proc. A.S.C.E., J. Environm. Eng. Div.* 1982, **108**(EE5), 923-940.