Goodness-of-fit measures for numerical modelling in urban water management – a summary to support practical applications

Introduction

Modelling of biological, chemical and physical processes in the field of wastewater transport and treatment is a state-of-the-art tool for researchers and engineers and widely accepted and used for optimising and design of urban water systems.

The focus of research and application of modelling is on integrated modelling for simultaneous consideration of sewer, WWTP and receiving water for European Water Framework Directive Compliance, **rising model complexity** in combination with the need of **automated parameter estimation** for complex models. **In any case an assessment of the quality of the achieved model fit is required.**

Risks associated with frequently used goodness-of-fit measures

Among others Willmott (1984) and Legates and McCabe Jr. (1999) discuss the misuse of Coefficient of Determination R^2 as a frequently applied goodness-of-fit measure. The following example underlines the risk potential of R^2. Figure 1 shows the measured MLSS concentration in the biological reactor of a WWTP over several months including three different simulation scenarios. The result of a visual comparison is a significant difference (shift and drift to the measured data) of two of the three simulated series. The scenario V2 with a shift to the measured data leads to the assumption of incorrect choice of conversion factors whilst the scenario V3 let assume e.g. an incorrect estimation of sludge production related model parameters due to the drift in the estimated data. But all simulations lead to the **same R^2 of 0.78**, pretending the same good model fit (Table 1).

To overcome some problems with R^2 the results from the linear regression between the measured (M) and estimated (E) data should be included in the evaluation. The calculated parameters, slope **a** and intercept **b** from a minimisation of the function **E = aM + b** improve the representativeness of the Coefficient of Determination. Consequently R^2 should only be displayed together with a and b.

An example for practical application

Figure 2 shows measured concentration data and the results of three different simulation runs estimating some effluent concentration of a biological reactor of a WWTP.

A first visual check leads to the assumption that simulation V1 fits the measured data best. This is confirmed by the scatterplot and the plot of residuals. Both other show gross amplitude errors while all estimations reproduce the pattern in measured data quite well. Table 2 shows a quantitative assessment by means of the different goodness-of-fit measures. The last column shows the estimation that best fits. The best values for every estimation is marked with bold numbers, the whorst is marked yellow with italic numbers.

Conventional descriptive statistic measures are evaluated at first. Considering the mean, both simulation runs V3 and V2 are nearest to the measured mean. Interestingly this does not apply for the median value for which V1 is the best one. Both measures are easily calculated but do not adequately represent the quality of the model fit. The comparison of standard deviation shows clearly that model results V2 and V3 are less accurate than V1. The Coefficient of Determination is in a high range for all three simulation runs, although V2 drops out compared to the other ones. Slope and intercept from regression analysis confirm this conclusion as the deviations from 0 for the parameter b (intercept) and the 50% overestimation of parameter a (slope) with a value of 1.50. An interpretation of RMSE shows again the best fit with simulation V1. An evaluation is only possible with direct comparison of the three different RMSE. Both difference-based bounded error measures Coefficient of Efficiency E and Index of Agreement d in normal (E2, d2) and modified form (E1, d1) all identify V1 as the best fit simulation result whereas d is less sensitive than E in showing this trend. Furthermore negative E-values imply that simulation results represent a less accurate prediction than the mean of the measured data.

The modified form of the Index of Agreement (d1) including absolute rather than squared values shows a wider spreading in the values while the squared form d2 is rather equal close to the maximum value of 1. Considering this case the distinction between the different simulations is rather difficult since differences for error measurements are negligible, i.e. time series fit all quite well. Hence E may preferably used to achieve a certain conclusion regarding the accuracy of the modelling results.

The **Coefficient of Efficiency E** (originally published with j=2 by Nash and Sutcliffe (1970)) is shown in equation 1. This measure ranges from **-∞ to 1** as the best possible value. E reaches 0 when the square of the differences between measured and estimated values is as large as the variability in the measured data. In case of **negative E** values the **measured mean** is a **better predictor than the model**. E is widely used in hydrologic modelling but reaches in the originally published form with squared differences high values even with mediocre modelling results. Thus the use of the modified form with absolute differences (j=1) is here recommended.

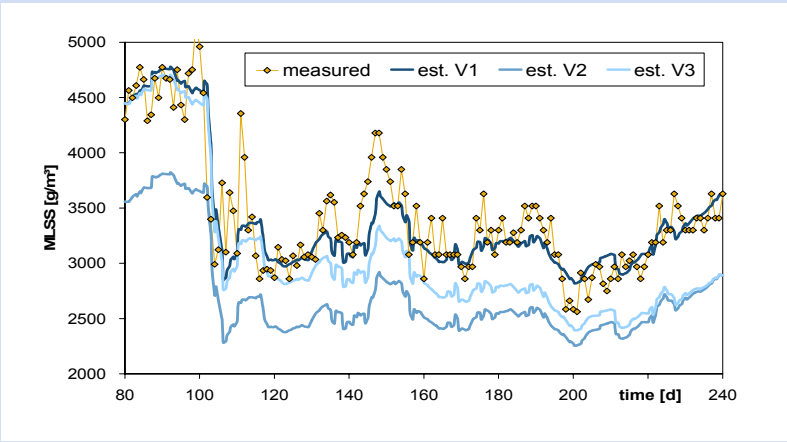
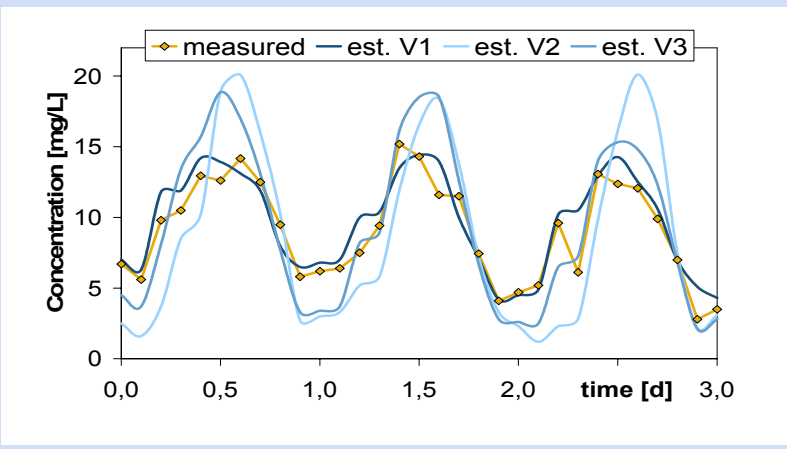


Figure 1: Measured and different estimated MLSS concentrations from a WWTP simulation

Table 1: Results of linear regression (all scenarios with R^2 = 0.78)

Scenario	Slope a	Intercept b
V1 (correct estimation)	1.0	0
V2 (estimation with shift)	0.8	0
V3 (estimation with drift)	1.15	-830



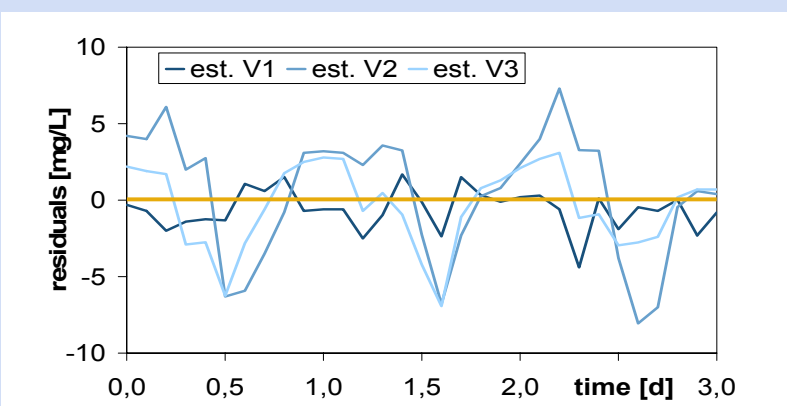


Table 2:
Goodness-of-fit measures for example simulations

Goodness-of-fit measure	Meas.	est. V1	est. V2	est. V3	best est.
Mean	9.04	9.64	8.63	9.42	V3,V2
Median	9.49	10.20	7.18	8.10	V1
Standard deviation	3.53	**3.43**	*6.45*	*5.66*	V1
Coefficient of Determination R^2		**0.86**	*0.67*	**0.88**	V3,V1
Linear regression, Par. a (E=aM+b)		**0.90**	*1.50*	*1.50*	V1
Linear regression, Par. b (E=aM+b)		**1.47**	*-4.90*	*-4.18*	V1
Root Mean Squared Error RMSE		**1.42**	*4.06*	*2.64*	V1
Coefficient of Efficiency E1		**0.65**	*-0.13*	*0.29*	V1
Coefficient of Efficiency E2		**0.83**	*-0.37*	*0.42*	V1
Index of Agreement d1		**0.83**	0.61	0.73	V1
Index of Agreement d2		**0.96**	0.82	**0.91**	V1, V3

Equation 1:
Nash-Sutcliffe-Coefficient (Coefficient of Efficiency)

$$E_j = 1 - \frac{\sum_{i=1}^{N} \left| M_i - E_i \right|^j}{\sum_{i=1}^{N} \left| M_i - \overline{M} \right|^j}$$

with j = 2 (original Coefficient of Efficiency with squared differences) or 1 (modified coefficient of efficiency with absolute differences)

Proposal for a formalised procedure

For a complete evaluation of the quality of a model fit the following routine is proposed:

1. **visual evaluation** (time series, scatterplot, plot of residuals)
2. test for fitting the time-dependent behaviour, when problems are evident (cross correlation, residence time distribution)
3. **use of descriptive statistics** for general balancing and distribution
4. **use of Goodness-of-fit measures** (E1, RMSE, R^2 (only together with calculation of linear regression coefficients))
5. further, more complex measures (e.g. systematic and unsystematic form of RMSE, see Willmott (1984).

Without a doubt this is a comprehensive list and the use will require additional effort for the analysis of the results.

As a **minimum** a **visual verification** including the subsequent quantitative assessment by the application of the **modified Coefficient of Efficiency (E1)** is recommended. If R^2 is used it is strongly recommended to include the calculation of slope and intercept of the linear regression.

Conclusions and Outlook

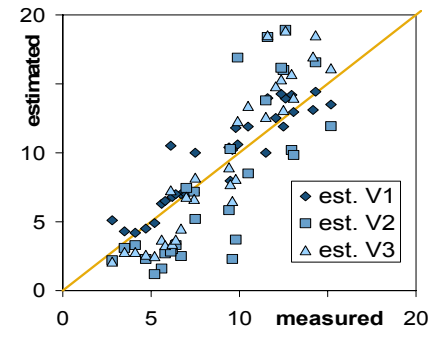
A comprehensive review of existing literature showed that **no common standard** for the assessment of the accuracy of modelling results can be identified. In most cases a **subjective method** (e.g. visual verification) is the preferred choice for an assessment of the model fit.

Based on the performed review and on elaborated examples two main approaches are recommended: for a distinct and systematic assessment a stepwise procedure is proposed. A minimum analysis should include the application of E1 (or R^2 with linear regression and evaluation of the regression parameters instead of simple R^2).

The authors believe this paper contributes to a 'Good modelling practice' among researchers and modelling engineers and hence increase the acceptance of numerical modelling in practice.

Figure 2: Example data for analysis of

- visual goodness of fit (top)
- plot of residuals from three different simulations (middle)
- scatter plot (right)



References

Langergraber, G., Rieger, L., Winkler, S., Alex, J., Wiese, J., Owerdieck, C., Ahnert, M., Simon, J., and Maurer, M. (2004). "A guideline for simulation studies of wastewater treatment plants." Water Science and Technology, 50(7), 131-138.

Legates, D. R., and McCabe Jr., G. J. (1999). " Evaluating the Use of "Goodness-of-Fit" Measures in Hydrologic and Hydroclimatic Model Validation." Water Res. Research, 35(1), 233-241.

Nash, J. E., and Sutcliffe, J. V. (1970). "River flow forecasting through conceptual models part I -- A discussion of principles." Journal of Hydrology, 10(3), 282.

Willmott, C. J. (1984). "On the evaluation of model performance in physical geography." Spatial Statistics and Models, G. L. Gaile and C. J. Willmott, eds., 443-460.

M. Ahnert, F. Blumensaat Institute for Urban Water Management, TU Dresden, 01069 Dresden, Germany.
G. Langergraber Institute of Sanitary Engineering and Water Pollution Control, University of Natural Resources and Applied Life Sciences, Vienna, Austria.
J. Alex ifak, Institut f. Automation und Kommunikation, Barleben, Germany.
D. Woerner iaks - Ingenieurbüro für Abfluss – Kläranlagen – Steuerung GmbH, Sonthofen, Germany.
T. Frehmann Emschergenossenschaft/Lippeverband, Essen, Germany.
N. Halft Department of Environmental Engineering, RWTH Aachen, Germany.
I. Hobus WiW - Wupperverbandsgesellschaft für integrale Wasserwirtschaft mbH, Wuppertal, Germany.
M. Plattes Centre de Ressources des Technologies pour l'Environnement (CRTE), CRP Henri Tudor, Esch-sur-Alzette, Luxembourg.
V. Spering Institute of Sanitary Engineering and Waste Management (ISAH), University of Hanover, Hanover, Germany.
S. Winkler Vienna University of Technology, Institute of Water Quality, Resources and Waste Management, Vienna, Austria.

Part of
Hochschulgruppe Simulation
a group of researchers from Germany, Austria, Luxembourg, Poland, the Netherlands and Switzerland

http://www.HSGSim.org