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

M. Ahnert [1] *, F. Blumensaat [1], G. Langergraber [2], J. Alex [3], D. Woerner [4], T. Frehmann [5], N. Halft [6], I. Hobus [7], M. Plattes [8], V. Spering [9], S. Winkler [10]

[1] Institute for Urban Water Management, TU Dresden, 01069 Dresden, Germany.(E-mail: Markus.Ahnert@tu-dresden.de) / [2] Institute of Sanitary Engineering and Water Pollution Control, University of Natural Resources and Applied Life Sciences, Vienna, Austria. / [3] ifak, Institut f. Automation und Kommunikation, Barleben, Germany. / [4] iaks - Ingenieurbüro für Abfluss – Kläranlagen – Steuerung GmbH, Sonthofen, Germany. / [5] Emschergenossenschaft/Lippeverband, Essen, Germany. / [6] Department of Environmental Engineering, RWTH Aachen, Germany. / [7] WiW - Wupperverbandsgesellschaft für integrale Wasserwirtschaft mbH, Wuppertal, Germany. / [8] Centre de Ressources des Technologies pour l'Environnement (CRTE), CRP Henri Tudor, Esch-sur-Alzette. / [9] Institute of Sanitary Engineering and Waste Management (ISAH), University of Hanover, Hanover, Germany / [10] Vienna University of Technology, Institute of Water Quality, Resources and Waste Management, Vienna, Austria

*corresponding author

Abstract The use of goodness-of-fit measures to quantify the accuracy is controversially discussed in recent publications of modelling in the field of hydrologic and ecological research and practise. In the field of activated sludge modelling however there is neither a standard recommendation for the use of objective evaluations nor is there a discussion ongoing on this issue. This paper summarises the findings of a review that can not presented due to the limiting page count for poster presentations. The work has been prepared within the framework of HSG-Sim (Hochschulgruppe Simulation, http://www.hsgsim.org), a group of researchers from Germany, Austria, Luxembourg, Poland, the Netherlands and Switzerland.

Keywords Activated sludge modelling, Coefficient of Efficiency, goodness of fit measures, urban water modelling,

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 present review is based on comprehensive discussions about model fits and their assessment in the framework of HSG-Sim. One objective was the development of a guideline for simulation studies (Langergraber et al., 2004). This work is now continued by the IWA Task Group on Good Modelling Practice (http://www.modeleau.org//GMP_TG). Within this context the issue of finding a common basis for a more standardised evaluation of goodness-of-fit measures was discussed.

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. A detailed version of this paper is published on the HSG-Sim web site with a complete overview about the investigated goodness-of-fit measures and further examples.

Current practice using goodness-of-fit measures

The use of goodness-of-fit measures in the field of water management and activated sludge modelling was analysed during a web-based search in the database of Google Scholar™. The total number of publications (about 2'000'000 for "model AND water" and about 19'000 for "activated AND sludge AND model", respectively) shows the potential sources for an investigation about goodness-of-fit measures. The results lead to the conclusion that only the regression based error measures (Pearson's Correlation Coefficient (r) and the Coefficient of Determination (R^2)) has an increased relevance (used in 47 % of "model AND water" and 17 % of "activated AND sludge AND model" papers, respectively). All other criteria are of subsidiary importance. It is known that in most publications only verbal quotes about the quality of the model fit are made, e.g. '*the model fits well; reached an acceptable level of accuracy; modelling results correspond to the measured values…*'. Often this yields in a different interpretation of model accuracy and underlines the need for a general and formalised procedure for the assessment of modelling results. It is strongly believed that such an approach will not only formalise and facilitate the process of model evaluation but also increase the level of acceptance among practitioners.

Several attempts have been made to define standards for model performance evaluation. Beside this there exist some compilations of error measures in form of free available software packages (e.g. IRENE) described by (Fila et al., 2003) as a standalone software or HydroTest described by (Dawson et al., 2007) as a web-based tool for online evaluation). This may be helpful but the variety of available measures and the differences in interpreting results depending on the characteristic of the time series makes a goodness-of-fit evaluation very complex.

Based on this analysis and the comments in the corresponding publications a combination of the findings from (Willmott, 1984) and (Legates and McCabe Jr., 1999) is derived as most appropriate procedure. This includes a combination of visual evaluation, descriptive statistics and difference based measures. An adaptation to the field of urban water modelling seems possible since characteristics of analysed time series are very similar.

An example for practical application

To illustrate the different possibilities of interpreting simulation results 3 selected goodness-of-fit measures are compared for different simulations. Figure 1 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 (Figure 1). Both other show gross amplitude errors, simulation V2 shows a small time offset, while all estimations reproduce the pattern in measured data quite well. Generally time-dependent problems within the estimation have to be eliminated before any further analysis of model fit starts. Several tools such as the calculation of cross-correlation, the analysis of residence time distribution with tracer experiments are available to do so. Another important aspect is to consider the response time of used online-sensors for data acquision (e.g. Rieger et al., 2003).

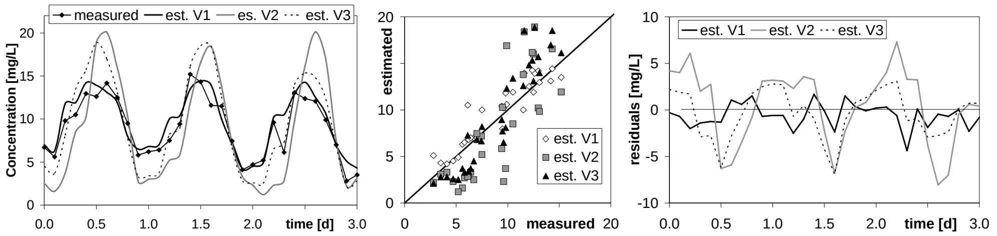

Figure 1 Example data for analysis of goodness of fit (left), scatter plot (middle) and plot of residuals (right) from three different simulations

Table 1 shows a quantitative assessment by means of the different goodness-of-fit measures. The last column shows the estimation that best fits. 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.

Table 1 Goodness-of-fit measures for example simulations (grey marked cells are hints for ill model fits while bold values show a good evaluation)

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 R²		**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

Both difference-based bounded error measures Coefficient of Efficiency E and Index of Agreement d in normal (E_2, d_2) and modified form (E_1, d_1) 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 (d_1) including absolute rather than squared values shows a wider spreading in the values while the squared form d_2 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 the following equation. This measure ranges from -∞ to 1 as the best possible value.

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

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

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.

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 (E_1, 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 (E_1) is recommended. If R^2 is used it is strongly recommended to include the calculation of slope and intercept of the linear regression.

References

Dawson, C. W., Abrahart, R. J., and See, L. M. (2007). "HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts." *Environmental Modelling & Software*, 22(7), 1034.

Fila, G., Bellocchi, G., Acutis, M., and Donatelli, M. (2003). "IRENE: a software to evaluate model performance." *European Journal of Agronomy*, 18(3-4), 369.

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.

Rieger, L., Alex, J., Winkler, S., Boehler, M., Thomann, M., and Siegrist, H. (2003). "Progress in sensor technology - progress in process control? Part I: Sensor property investigation and classification." *Water Science And Technology*, 47(2), 103-112.

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.