Differences in the ASM model caused by data structure

Zbigniew Kowalewski[1,*]

[1]AGH University of Science and Technology

Abstract. The process of designing and exploiting municipal sewage treatment plants has become much simpler and more efficient thanks to mathematical modeling. The ASM model family is able to simulate the operation of existing or designed objects in a satisfactory manner. The basic problem in Poland is the insufficient amount of data for simulations coming from plant monitoring. It is provided to create unstable model results with difficulties in calibration and validation. The aim of this article is to confirm how the amount of data and its completeness will affect the quality of the simulation performed in the ASM model. The study object is a sewage treatment plant located in Chicago in the USA. It is a sewage treatment plant operating with activated sludge technology, with regular monitoring of the quality of raw and treated wastewater. For modeling, a variant of the ASM model built into the BioWin 5.2 software was used.

1 Introduction

Wastewater treatment plants (WWTP) are large and complicated systems. To design well-operated and economically reasonable treatment plant, a mathematical aided model is necessary [1]. Also, rational exploitation requires decisions made on the basis of mathematical models. Activated sludge models (ASM) and models being their development version belongs to the most popular used in the design, service, and analysis of a wastewater treatment plant [2]. ASM models contain equations for biological and chemical reactions, interactions between bacteria, and nitrogen and phosphorous transformation [3]. The problem is that such a complicated system needs a large amount of data. A precise measurement campaign lasts a long time and is expensive. There are some guidelines on how many measurements should be taken, from a week with hourly interval to months with daily intervals [4]. However, in Polish realities, such conditions are difficult to accomplish. The aim of this work is to verify how the size of the dataset will affect model results. Data for the study was collected from Calumet WWTP in Chicago. In the first stage of the study, the whole treatment plant was scaled to one single line with three aeration tanks, to simplify the model and to give it much more visual clarity. In the second stage, seven different exploitation variants were created (the ratio of returned sludge was changed). The aim of this study was to verify how the period of modelling will affect calculation results in different exploitation parameters. The third stage of the study was to verify how much shortened data

* Corresponding author: kowalew@agh.edu.pl

will affect the final output results. Nine new datasets were created from a raw initial dataset (containing measurements over 365 days). Three sets with 50, 100, 200 random measurement days were created. Then, with this datasets model the dynamic simulation was run.

2 Methods

2.1 Description of the Chicago Calumet WWTP

The wastewater treatment plant (WWTP) investigated in this study is located in the city of Chicago, Illinois, USA. The Calumet Water Reclamation Plant (CWRP) on the South Side is the oldest of the seven water treatment facilities in the Chicago area. In operation since 1922, it serves a population of more than 1 million people. The Calumet plant is designed for an average influent dry-weather flow rate of 1.34 million m^3 d^{-1}, the design maximum flow is 1.63 million m^3 d^{-1}. The plant is composed of three parallel lines called AB, C, and E1E2. Primary treatment units are 6 screens, 8 grit tanks and 12 primary clarifiers (46,5 m diameter, 4.65 m side wall depth). The main treatment line is divided into three routes. Line AB contains: 22 conventional one-pass aeration tanks (size 128 m L x 10.35 m W x 4.65 m D), 16 secondary clarifiers (27.3 m L x 27.3 m W x 3.6 m D), and 8 radial secondary clarifiers (26 m diameter, 3.6 m side wall depth). Line C contains 6 conventional aeration tanks (79.5 m L, 10.2 m W, 4.5 m D) and 8 secondary clarifiers (33 m diameter, 3.9 m side wall depth). Line E1E2 contains 20 conventional one-pass aeration tanks (size 128 m L x 10.35 m W x 4.65 m D) and 10 secondary clarifiers (45 m diameter, 4.5 m side wall depth). The anoxic and anaerobic zones are not assigned. Sludge return is external, from the secondary clarifier to the beginning of the aeration tanks. The last stage of the treatment process is disinfection using chlorine [5]. For the simulation study, only routine operational data from the Calumet WWTP were used, no additional measurements were performed.

Table 1. Descriptive statistics of raw wastewater from the Calumet WWTP.

	Flow [m^3 h^{-1}]	BOD5 [mg dm^{-3}]	VSS [mg dm^{-3}]	TSS [mg dm^{-3}]	TKN [mgN dm^{-3}]	TP [mgP dm^{-3}]	NO3 [mgN dm^{-3}]	pH [-]	Ca [mg dm^{-3}]	Mg [mg dm^{-3}]
Min.	4620	27	26	400	5	1	0.14	7.1	46	15
1st Qu.	7182	85	66	754	17	3.6	0.14	7.4	66	24
Median	8484	112	83.71	844	21	4.5	0.14	7.4	73	27
Mean	9611	121.2	83.62	881.3	20.98	5.01	0.296	7.45	73.1	27.01
3rd Qu.	11760	135	92	958	25	6	0.19	7.5	81	31
Max.	20706	460	246	1861	46	16.4	2.9	7.8	96	36

Raw data were collected from routine plant monitoring once a day, from the 1st January to the 31st of December in 2015. The input data (raw wastewater) for modelling contains the following parameters: flow [m^3 h^{-1}]; total carbonaceous biological oxygen demand, BOD5 [mg dm^{-3}]; volatile suspended solids, VSS [mg dm^{-3}], total suspended solids, TSS [mg dm^{-3}], total Kjeldahl nitrogen, TKN [mgN dm^{-3}]; total phosphorous, TP [mgP dm^{-3}]; nitrate, NO3 [mgN dm^{-3}]; pH [-]; calcium, Ca [mg dm^{-3}]; magnesium, Mg [mg dm^{-3}]. There were no data for alkalinity and dissolved oxygen, they were set up as constant at level 6 [mmol dm^{-3}] for alkalinity and 0 [mg dm^{-3}] for dissolved oxygen. The basic statistics of the raw input dataset are in Table 1. This dataset was used in a computer model. This set contains 365 measurements, subsequently from this dataset data used in the next part of the analysis were made. Three sets were sampled with 50 random measurements (50v1, 50v2, 50v3), three sets

with 100 random days (100v1, 100v2, 100v3) and three sets with 200 random days period (200v11, 200v2, 200v3).

2.2 Simulation environment and model configuration

In this work, simulations were carried out in BioWin® software version 5.2 developed by Envirosim® Canada. BioWin is a user-friendly platform, with a visual style (flow layout design). BioWin uses ASDM (Activated Sludge Digestion Model) for calculations. Presently, BioWin is one of three main types of software for modeling activated sludge in wastewater treatment plants, with two others being GPS-X® by Hydromantis® and Mike WEST® by DHI.

All simulation processes were carried out with standard equations and coefficients. No additional models combined with the ASDM model. Due to the plant`s large scale (48 aeration tanks, 52 secondary clarifiers), some reductions were taken. Whole treatment plant system and sewage flow were scaled to one technological line which contains: one grid chamber (volume – 1570 m3, percent of capture inert suspended solids – 65%), three primary settling tanks, three aeration chambers: first with AB line sizes, second with C line sizes third with E1E2 line sizes, the flow was split proportionally: 40% to the first and third reactor, and 20% to the second reactor. That kind of simplification makes the model much more useful and does not make changes to the calculation results. The aeration parameters were constant, the setpoint of dissolved oxygen was 2 [$mgO_2\ dm^{-3}$]. Return sludge in the secondary clarifiers was set up from a 0.5 to 2.5 ratio. All equation coefficients and fractions were set up as default. The calculation interval was set as one day, the time period was set to 365 days, and steady-state and dynamic (with seed values) simulations were performed. The Calumet WWTP configuration in the BioWin software is presented in Fig. 1. The blue line is wastewater, the dotted blue line is the returning sludge process, and the green lines are for sludge.

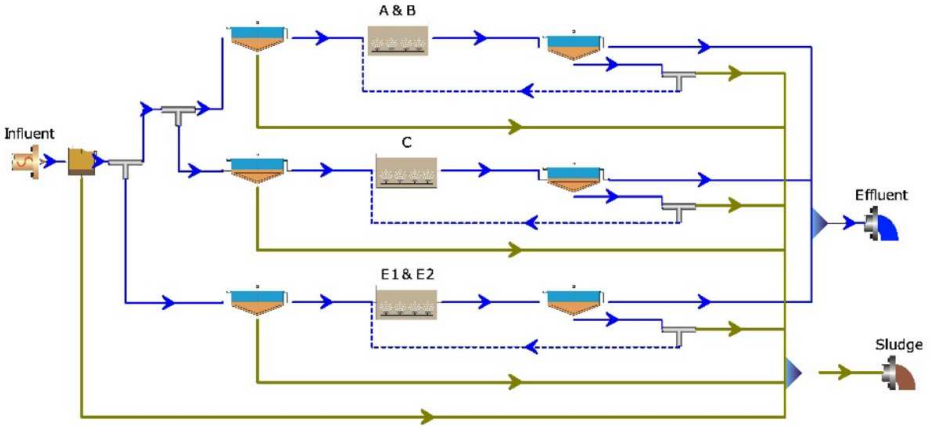

Fig. 1. BioWin configuration of the Calumet WWTP.

2.3 Statistical methods

A significance level $\alpha = 0.05$ was used in all analysis. CRAN R [6] software with RStudio GUI [7] was used for all analysis. For the prepared random data sets, a formula *sample* was used. Goodness-of-fit (GOF) measures between the observed and simulated values were performed with the *hydroGOF* R package [8]. Five statistics were used: MAE, RMSE, PBIAS, d, and KGE.

MAE, Mean Absolute Error is a model evaluation metric used with regression models. A smaller value indicates better model performance.

$$mae = \frac{1}{N}\sum_{i=1}^{N}|S_i - O_i| \tag{1}$$

RMSE, Root Mean Square Error gives the standard deviation of the model prediction error. A smaller value indicates better model performance.

$$rmse = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)^2} \tag{2}$$

PBIAS, Percent BIAS, measures the average tendency of the simulated values to be larger or smaller than observed values. 0.0 is the optimal value.

$$PBIAS = 100\frac{\sum_{i=1}^{N}(S_i - O_i)}{\sum_{i=1}^{N}O_i} \tag{3}$$

d, Index of Agreement developed as a standardized measure of the degree of model prediction error and varies between 0 and 1. A value of 1 indicates a perfect match, and 0 indicates no agreement at all. The authors present d as quite flexible and making it applicable to a wide range of model ‑ performance problems [9].

$$d = 1 - \frac{\sum_{i=1}^{N}(S_i - O_i)^2}{\sum_{i=1}^{N}(|S_i - \bar{O}| + |O_i - \bar{O}|)^2} \tag{4}$$

KGE, Kling-Gupta Efficiency is a variant (decomposition) of Nash-Sutcliffe Efficiency (NSE), which analyzes the importance of components like correlation, bias, and variability. A value brought closer to 1, indicates better model performance [10].

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{5}$$

For equations (1), (2), (3), (4), (5), N is a number of measurements, S_i is the simulated value, O_i is the observed value, r is the Pearson regression line correlation coefficient, α is (6) and β is (7).

$$\alpha = \frac{\sigma_S}{\sigma_O} \tag{6}$$

$$\beta = \frac{\mu_S - \mu_O}{\sigma_O} \tag{7}$$

Where: μ_S and σ_S are the mean and standard deviation of simulated values, while μ_O and σ_O are the mean and standard deviation of observed values.

3 Results

Due to the scarcity of output data, the goodness-of-fit models were tested for four parameters: total nitrogen – TN [mgN dm^{-3}], total phosphorous – TP [mgP dm^{-3}], total Kjeldahl nitrogen – TKN [mgN dm^{-3}] od pH [-]. Actual wastewater output statistics are in Table 2.

Table 2. Descriptive statistics of clear wastewater (output) from the Calumet WWTP.

	TN	TP	pH	TKN
Min.	4.7	0.7	7	1
1st Qr	8.9	2.5	7.1	1
Median	10.6	3.5	7.2	1
Mean	10.79	3.626	7.195	1.636
3rd Qr	12.5	4.7	7.2	2
Max.	19.1	10.3	7.4	8

To see how the exploitation parameters will change model results, the return sludge ratio was changed in every trial. The ratio had values: 0.5, 0.8, 1, 1.2, 1.5, 2.0, 2.5. Fit statistics between observed and the simulated data are presented in Table 3. For TN, TP and pH there were no great differences between the returning sludge ratio. Only TKN values changed across ratio differences. The statistics values of ratio 1.5 and 2 are nearly the same and got a better score than the rest of variants. Overall, there are no significant differences between the results, although the best results have a 1.5 ratio, so this configuration was used in model trials with shortened random data.

Table 3. Statistics of goodness-of-fit between ratio variants.

		ratio 1	ratio 0.5	ratio 0.8	ratio 1.2	ratio 1.5	ratio 2	ratio 2.5
TN	MAE	3.56	3.57	3.59	3.53	3.79	3.78	3.53
TN	RMSE	4.22	4.23	4.25	4.17	4.43	4.42	4.16
TN	PBIAS	30.1	30.2	30.3	29.8	32.9	32.8	29.8
TN	d	0.57	0.58	0.58	0.57	0.56	0.56	0.57
TN	KGE	0.41	0.4	0.4	0.42	0.4	0.4	0.42
TP	MAE	1.96	1.96	1.96	1.96	1.83	1.83	1.96
TP	RMSE	2.39	2.39	2.38	2.4	2.25	2.25	2.4
TP	PBIAS	-52.8	-52.8	-52.6	-53	-48.4	-48.5	-52.9
TP	d	0.48	0.48	0.49	0.48	0.5	0.5	0.48
TP	KGE	-0.07	-0.06	-0.05	-0.09	0	0	-0.09
pH	MAE	0.07	0.07	0.06	0.07	0.11	0.11	0.07
pH	RMSE	0.1	0.09	0.09	0.1	0.12	0.13	0.1
pH	PBIAS	0.1	0.1	-0.1	0.4	-1.4	-1.4	0.3
pH	d	0.66	0.66	0.66	0.64	0.53	0.53	0.64
pH	KGE	0.43	0.44	0.45	0.39	0.41	0.43	0.4
TKN	MAE	5.33	5.12	4.78	6.07	2.03	2.07	5.7
TKN	RMSE	5.72	5.49	5.11	6.52	2.21	2.25	6.1
TKN	PBIAS	324.8	312.4	291.6	370.4	120.9	123.5	347.7
TKN	d	0.19	0.19	0.21	0.17	0.4	0.39	0.18
TKN	KGE	-2.55	-2.41	-2.17	-3.06	-0.56	-0.58	-2.79

The results of MAE, RMSE, d, and KGE statistics for TN, TP, pH, TKN for the ratio1.5, and nine random data sets are presented in Fig. 2. In case the of pH parameters, GOF statistics had quite good results, PBIAS and mean errors are near 0, the d value in all cases is higher than 0.4 and KGE is not near 0. pH is quite a specific parameter, its vales normally are in the small range. In the case of TKN, the results are highly biased (over 100%), 50 elements had worse results, 100 and 200 datasets had the same value as a full set. The same situation is in the rest of the GOF's, the full set gives the best results but there are not big differences between all sets. The TN values overall are moderately biased but have a large MA and RMS error. The KGE value bars look similar to this from the pH, 100v have the same level, in 50v there are many differences, the best results are from full data. The d value is at a moderate-high level. For TP, PIAS is at the same, negative level of 50%, RMSE and MAE are at almost the same level for all variants. KGE is negative and has a nearly 0 value. The statistic of d is at the same acceptable level.

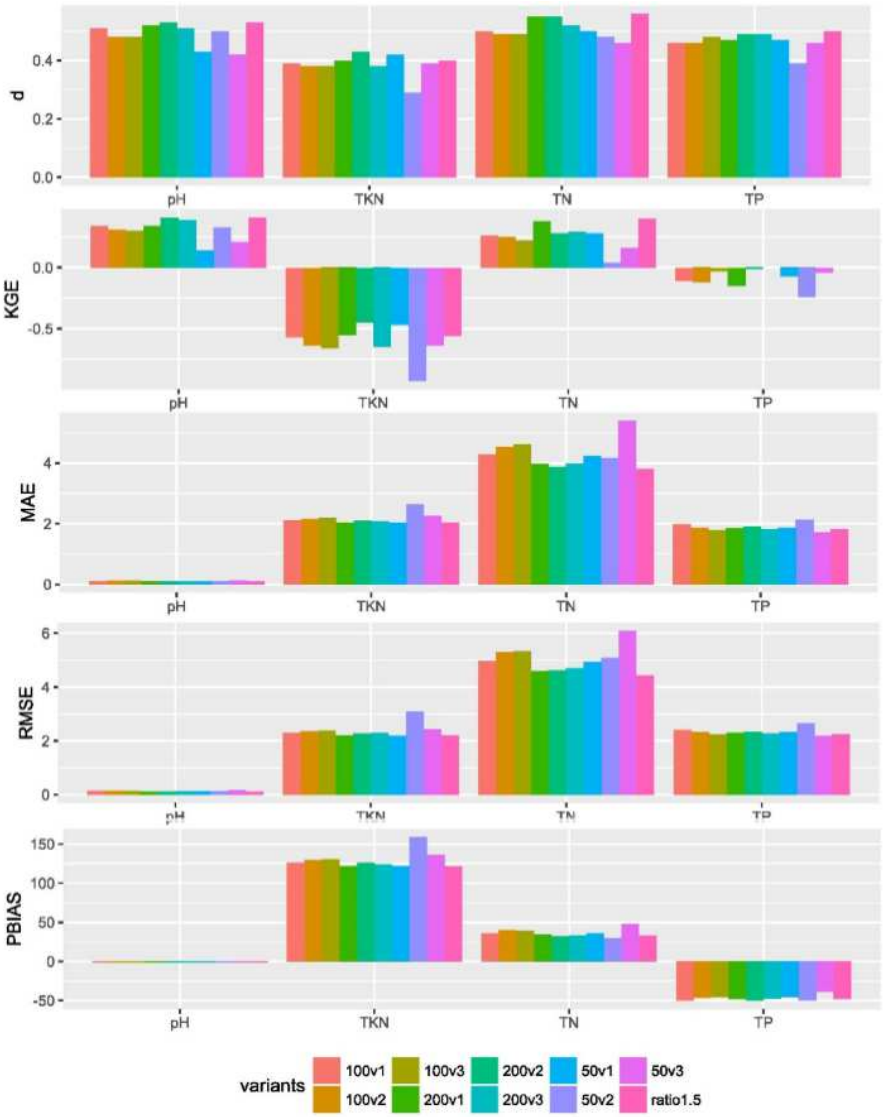

Fig. 2. Statistics of goodness-of-fit between full data (ratio 1.5) variant and partial data variants.

4 Conclusions

The created models were the "quick ones", without a special monitoring plan and changes in the monitoring program. No special calibrations or fractioning were done. The time cycle (365 days) and data interval (1 day) were quite large as an ASM model. The results from this model are even satisfying. The first conclusions are that in the case of specific data placement, parameters such as pH, TP or TN do not undergo significant change under the influence of a specific treatment process.

Only Kjeldahl nitrogen was sensitive to changes in the returning sludge ratio. The second conclusion concerns pH values, normally pH is a difficult modelling parameter [11], but in a situation were the range is small (in Calumet is from pH to 7.4), BioWin simulates pH values quite easily. The third conclusion, a big dataset gives better simulation results, in the case of one-day interval and one-year period differences between 365 days and 200 are not large, but the 50-day version gives the poorest results.

The Metropolitan Water Reclamation District of Greater Chicago (MWRD) for sharing the data of the Calumet WWTP.

References

1. A. Sochacki *et al.*, *Proc. a Polish-Swedish-Ukrainian Semin.*, (2009)
2. O. O. James, C. a O. Jiashun, F. Qian, and O. J. Oleyiblo, *Chinese J. Oceanol. Limnol.* **33** (2014)
3. E. Liwarska-Bizukojc, D. Olejnik, R. Biernacki, and S. Ledakowicz, *Bioprocess Biosyst. Eng.* **34** (2011)
4. P. Vanrolleghem, W. Schilling, W. Rauch, P. Krebs, and H. Aalderink, Water Sci. Technol. **39** (1999)
5. MWRDGC, (2015)
6. R Core Team (2013)
7. RStudio Team (2015)
8. M. Zambrano-Bigiarini (2017)
9. C. J. Willmott, S. M. Robeson, and K. Matsuura, Int. J. Climatol. **32** (2012)
10. H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, J. Hydrol. **377** (2009)
11. M. Fairlamb, R. Jones, I. Takacs, and C. Bye, **7** (2003)