Towards quantitative quality criteria to evaluate simulation results in wastewater treatment – A critical review

H. Hauduc[a,b], M. B. Neumann[b], D. Muschalla[b,c], V. Gamerith[d,b], S. Gillot[a] and P.A. Vanrolleghem[b]

[a] Cemagref, UR HBAN, Parc de Tourvoie, BP 44, F-92163 Antony Cedex, France. (E-mail: helene.hauduc@cemagref.fr; sylvie.gillot@cemagref.fr)
[b] model*EAU*, Département de génie civil et de génie des eaux, Université Laval, 1065 av. de la Médecine, Québec (QC), G1V 0A6, Canada. (E-mail: marc.neumann@gci.ulaval.ca; peter.vanrolleghem@gci.ulaval.ca)
[c] itwh – Institute for Scientific and Technical Hydrology, Engelbosteler Damm 22, 30167 Hannover, Germany. (E-mail: d.muschalla@itwh.de)
[d] Graz University of Technology, Institute of Urban Water Management and Landscape Water Engineering, Stremayrgasse 10/I, 8010 Graz, Austria. (E-mail: gamerith@sww.tugraz.at)

Abstract
A total of 31 quantitative quality criteria to compare measured with simulated time series in environmental modelling are critically reviewed. They are grouped using two classification schemes. A methodology to evaluate and compare the criteria is proposed and tested on a case study. This methodology includes a two stage cluster analysis using Kendall rank correlation and dendograms to identify independent quality criteria. Independent quality criteria allow for a comprehensive model assessment and are a prerequisite for multi-objective parameter estimation.

Keywords
Model quality evaluation; modelling objectives; model selection; parameter estimation

ABBREVIATIONS

AME: Absolute Maximum Error	MSRE: Mean Square Relative Error
ASM: Activated Sludge Model	MSSE: Mean Square Sorted Errors
CE: Coefficient of Efficiency	NSC: Number of Sign Change
$CE_{1,2}$: Nash-Sutcliffe	PBIAS: Percent Bias
COD: Chemical Oxygen Demand	PDIFF: Peak Difference
CrBal: Balance Criterion	PEP: Percent Error In Peak
HRT: Hydraulic Retention Time	PI: Coefficient of Persistance
IA: Index of Agreement	PE: Population Equivalent
MAE: Mean Absolute Error	RAE: Relative Absolute Error
MAER: Relative Mean Absolute Error	RMSE: Root Mean Square Error
MAPE: Mean Absolute Percent Error	RSR: RMSE-observation standard deviation ratio
MARE: Mean Absolute Relative Error	RVE: Relative Volume Error
MdAPE: Median Absolute Percent Error	SRT: Sludge Retention Time
ME: Mean Error	TKN: Total Kjeldhal Nitrogen
MPE: Mean Percent Error	TMC: Total Mass Controller
MRE: Mean Relative Error	TSS: Total Suspended Solids
MSDE: Mean Square Derivative Error	U^2: Theil's Inequality Coefficient
MSE: Mean Square Error	WWT: Wastewater Treatment
MSLE: Mean Square Logarithm Error	WWTP: Wastewater Treatment Plant

INTRODUCTION

In wastewater treatment (WWT) modelling, the evaluation of model quality is often based on qualitative comparisons between simulation results and observed data. Although such visual evaluation is useful, it does not provide an objective assessment of the quality of a calibration parameter set. Moreover, it cannot be used in an automatic calibration procedure.

Environmental sciences, hydrology in particular, widely use mathematical comparisons of predicted and observed values (Dawson *et al.*, 2007). In WWT several target constituents are usually considered simultaneously during model calibration (sludge production, total suspended solids (TSS), chemical oxygen demand (COD), nitrogen and phosphorus in the effluent …). Although a review of quality criteria is presented in Dochain and Vanrolleghem (2001), quantitative criteria are rarely determined in this field (Petersen *et al.*, 2002; Ahnert *et al.*, 2007; Sin *et al.*, 2008).

In order to facilitate the adoption of quantitative quality criteria for model evaluation and automated calibration, a literature review was undertaken covering a number of water-related disciplines (WWT, catchment hydrology, urban hydrology, climate sciences, environmental sciences…). Then, a methodology was set up to investigate the use of those criteria for WWT modelling, especially in view of determining suitable parameter values in an automated calibration procedure. The procedure is applied to a case study, and the quality criteria obtained are analysed. In particular the correlations between criteria are evaluated, to identify independent criteria. The use of independent criteria allows making better use of the available information in the data.

QUANTITATIVE QUALITY CRITERIA USED IN ENVIRONMENTAL SCIENCES

General methods to compare observed and predicted data

Depending on the modelling objectives, the goodness-of-fit of a model can be defined as the capability of the model to capture one or several characteristics of the observed data: mean, timing and magnitude of peaks or typical periodical variations (diurnal, weekly, seasonal…). For example, if a specific effluent limit of a plant is based on a monthly average it makes no sense to evaluate the accuracy of the fit of each single peak. However, if peak effluent limits have to be met, a criterion evaluating the fit of peaks should be used. Thus, to characterise the goodness-of-fit of the model, different quality criteria may be needed. These criteria vary in the way they are computed from the observed and predicted data:

- Criteria can be **averaged** over the number of data on which they were computed, which allows comparing results obtained on datasets of different sizes;
- **Absolute criteria** are expressed in the same units as the variables of interest;
- **Relative criteria** (divided by observed or predicted values or by the variance) are dimensionless; which allows to compare across different state variables;
- **Comparisons of residuals obtained with simple models** are used in several criteria to define the improvement of using the model over a simple model, such as a model defined as the mean of the observed values or the previous observed value (see e.g. Seibert, 2001). The model to be compared with can also be a model describing typical variations (e.g. daily mean time-series calculated from historical time-series), or a seasonal mean value (Legates and McCabe, 1999).

Other arithmetic operations can be applied to emphasise small or large errors or errors on specific parts of the time series:

- **Partitioning the dataset** according to different measurement magnitudes (e.g.: low, intermediate and high flows) and computing the quality criteria on each of these subsets (Perrin *et al.*, 2006; Moriasi *et al.*, 2007),
- **emphasising small errors or low magnitude values**: a power transformation of the data with an exponent lower than 1 (square root…) or a logarithmic transformation can be used,
- **emphasising large errors or high magnitudes values**: a power transformation of the data with an exponent larger than 1 or an exponential transformation can be used,
- **avoiding error compensation**: absolute values and even power values will avoid compensation of negative and positive errors when summing them.

These arithmetic operations are used to modify the general criteria to extract the required information, given a certain objective, e.g. give more importance towards errors at low magnitude, maximum errors or errors on peaks. It is important to note that all criteria discussed in this review are based on sums. Consequently, in case of datasets with variable time steps, the criteria will emphasise errors on more frequently sampled periods. A solution to overcome this problem would be to use weighted criteria inversely proportional to frequency (an isolated point will have a higher weight) (Willmott *et al.*, 1985).

Review and classification of quantitative criteria used in environmental sciences

The thirty-one quantitative quality criteria selected in the literature review are described in Appendix 1. They were grouped according to two classification systems (classes 1-6 and characteristics i-vi). The first classification scheme is inspired by Dawson *et al.* (2007) and groups the criteria into 6 main classes:

1. **Single event statistics:** In case modelling objectives require accurate simulation of events (e.g.: handle storm flows, toxic peaks), criteria are needed to characterise the goodness-of-fit of the model for this event. The single event statistics peak difference (Gupta et al., 1998) and percent error in peak (Dawson et al., 2007) aim at characterising the difference between the maximum observed and the maximum modelled value.
2. **Absolute criteria from residuals:** The absolute criteria are based on the sum of residuals (difference between observed O_i and predicted P_i values respectively at time step i), generally averaged by the number of data, n. A low value of this criterion means a good agreement between observation and simulation (with γ an exponent).
$$E_\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(O_i - P_i\right)^\gamma$$
3. **Residuals relative to observed values:** At each time step, the error is related to the corresponding observed or modelled value. A low value of this criterion means a good agreement between observation and simulation.
$$RE_\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{O_i - P_i}{O_i}\right)^\gamma$$
4. **Total residuals relative to total observed values:** For the following criteria, the sum of errors is related to the sum of observed values, without any correspondence in time step. A low value of this criterion means a good agreement between observation and simulation.
$$TRE_\gamma = \frac{\sum_{i=1}^{n}(O_i - P_i)^\gamma}{\sum_{i=1}^{n} O_i{}^\gamma}$$
5. **Agreement between distributional statistics of observed and modelled data:** These criteria are not based on error comparison, but on a comparison between cumulative modelled and observed data. These criteria originate from hydrology and aim at verifying whether the total water volume has been reproduced by summing the flows. In the wastewater field these criteria can be relevant for influent and effluent pollutant loads by summing the fluxes.
6. **Comparison of residuals with reference values and with other models:** These criteria compare the residuals with residuals obtained with a reference model $\tilde{P}$, such as a model describing the mean value ($\tilde{P}_i = \bar{O}$) or the previous value ($\tilde{P}_i = O_{i-1}$) (with α an exponent).
$$CE_{\alpha,\gamma} = 1 - \frac{\sum_{i=1}^{n}\left(O_i^\alpha - P_i^\alpha\right)^\gamma}{\sum_{i=1}^{n}\left(O_i^\alpha - \tilde{P}_i^\alpha\right)^\gamma}$$

In the second classification system the 31 quality criteria are classified along 6 main different characteristics of the adjustment of the predicted values to the observed dataset: i) criteria evaluating the mean error, ii) criteria evaluating the bias, iii) criteria that emphasise large errors, iv) criteria that emphasise small errors, v) criteria evaluating peak magnitudes and vi) criteria evaluating event dynamics (Table 1).

Table 1. List of criteria per characteristic (The quality criteria are described in detail in Appendix 1)

Characteristics	Quality criteria
Mean error	MAPE, TMC, RVE, $CE_{1,2}$, $CE_{1/2,2}$, RSR, U^2, MSE, RMSE, MARE, MSRE, MdAPE, MAE, MRE, MAER
Bias	PBIAS, ME, MPE, RAE, NSC, CrBal, MSSE, PI
Large errors	AME, R4MS4E, IA
Small errors	MSLE, $CE_{LN,2}$
Peak magnitude	PDIFF, PEP
Chronology of events	MSDE

METHOD

Automated criteria evaluation

The proposed evaluation procedure is summarised in Figure 1. The main steps and software used at each step are specified. Each step is described in the following paragraphs for the case study.

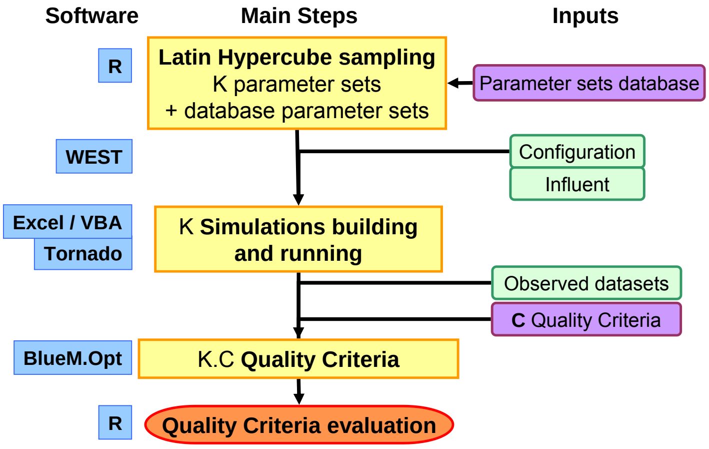

Figure 1. General framework of the automated criteria evaluation. In this case study, the number of simulations is K=5000 and the number of quality criteria is C=31. The parameter sets database is presented in Hauduc *et al.* (2011).

Description of the wastewater treatment plant (WWTP) used as case study

The procedure was tested using quality-controlled data (Rieger *et al.*, In Preparation) from a 250.000 population equivalent (PE) municipal WWTP located in France. It is configured in two parallel lanes that operate under similar conditions, each lane containing a plug-flow tank with a pre-denitrification zone. Aeration is controlled by a timer. Chemical phosphate removal is carried out with addition of alum. The sludge retention time (SRT) is 27 days and the hydraulic retention time (HRT) 8.8 hours.

The simulation period consists of 84 consecutive days, from February 15 to May 9 2009. This period was chosen because the added alum quantity was monitored during this time, allowing an estimation of the chemical sludge production by phosphorous precipitation. An advantage of this period is that it also includes high dynamics and varying operating conditions. The first half of this period exhibits typical operating conditions. However, on day 48, all aerators broke down for 3 days. Then from day 51 to 68 the aerators were running permanently. These varying operating conditions provide dynamic conditions that provide more information which improves the identification of model parameters. The target constituents (total suspended solids (TSS) in the biological reactor, TSS, COD, total Kjeldahl nitrogen (TKN), nitrate and ammonia in the effluent are measured daily as flow-proportional daily averages.

Activated Sludge Model n°1 (ASM1) (Henze *et al.*, 2000) was chosen to model this WWTP as there is no biological phosphorus removal and because it is the simplest and more commonly used model, for which parameter value ranges are known (Hauduc *et al.*, 2009; Hauduc *et al.*, 2011). However, a modified ASM1 that includes a different heterotrophic growth yield under anoxic conditions was preferred (Orhon *et al.*, 1996).

Implementation of the configuration

The configuration was implemented in WEST (Figure 2) (Vanhooren *et al.*, 2003). The two lanes were modelled as a single one with double volumes. Considering the U-shape of the aerobic tank, it was represented by an anoxic tank and three aerobic tanks, each with the same volume. As there is no sludge accumulation in the secondary settler, a point-settler model was chosen. The target constituents were sampled at the output through a modelled flow-proportional sampler so as to directly obtain flow-proportional daily averages as model outputs.

Biokinetic parameter ranges are available from a database of modelling projects (Hauduc *et al.*, 2011). However, no ranges are yet available for the fraction of non-settleable particulates. Furthermore, fractionation parameters were included in the Latin Hypercube sampling to consider the uncertainty on these parameters due to the lack of reliable fractionation experiments on the plant influent. To determine reasonable ranges for these parameters, a pre-calibration step was carried out using Cemagref's default parameters (Choubert *et al.*, 2009) (Table 2). The fractionation of the influent and the fraction of non-settleable particulates were roughly adjusted, starting respectively from usual values (Gillot and Choubert, 2010) and default WEST value (f_{ns}=0.005), to (visually) fit the sludge production and the mean effluent output data (Table 2). Parameter ranges were then defined as ±20% around these fractionation values and ±60% around the fraction of non-settleable particulates determined in the pre-calibration step. To keep a total COD and nitrogen fraction of 100%, XC_B and $XC_{B,N}$ were chosen to be the residual of the other fractions. Under these variations, XC_B varied from 42.4% to 61.8% while $XC_{B,N}$ varied from 73.8% to 82.4% (Table 2).

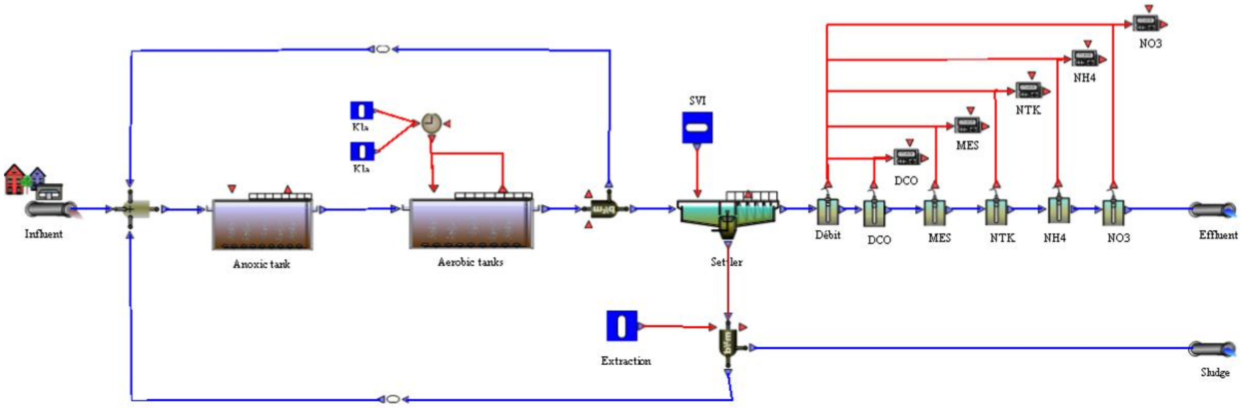

Figure 2. Layout of the WWTP in WEST. A combination of sensors and samplers for each target constituent provides daily mean flow-proportional values directly from the simulation.

Parameter sets sampling

The proposed procedure to evaluate quality criteria is based on the automated calibration procedure set up by Sin et al. (2008) and modified by Hauduc (2010). Monte Carlo simulations were carried out based on 5000 parameter sets sampled in a Latin hypercube, using R (http://www.r-project.org/). As presented in Table 2, all ASM1 parameters, except the temperature adjustment coefficients were considered. As no correlation between parameters could be identified in the modelling projects database (Hauduc *et al.*, 2011), the parameters are considered independent.

Simulations

The simulations were carried out in Tornado (Claeys *et al.*, 2006), the generic kernel of WEST.

To ensure correct initial steady-state conditions for the 84 days of dynamic simulation of each parameter set, 100 days (> 3 * SRT) were first simulated under pseudo steady-state conditions (alternating aeration periods, constant influent).

Quality criteria calculation and analysis

Quality criteria calculation. To automatically calculate the quality criteria, a modified version of BlueM.OPT (Bach *et al.*, 2009, http://www.bluemodel.org/) was used. BlueM.OPT is the optimization framework of BlueM, a software package for river basin management. For each target constituent in the dynamic simulation period, the 31 quality criteria were automatically calculated from the 5000 output files (containing the daily averaged simulation results) and the reference file (containing the observations).

Quality criteria analysis. The evaluation of quality criteria was performed with R. Among all calculated criteria, the first task was to highlight those that are non-correlated, i.e. that provide non-redundant information. To this aim, Pearson, Spearman and Kendall correlation coefficients between the criteria were calculated. Spearman and Kendall are rank correlation coefficients that non-parametrically measure the dependence of two variables, whereas Pearson correlation implies a linear relationship between variables.

As different quality criteria have different optimal values (-∞, 0, 1, +∞) (see also Appendix 1), all quality criteria were first mathematically transformed to obtain values between 0 and +∞ (or 0 and 1) with 0 being the optimal value (see Appendix 1). Thanks to these transformations, the criteria that have a negative correlation coefficient are anti-correlated, which means that when one criterion is improved, the other one is deteriorated, and criteria that have a positive correlation coefficient close to 1 are highly correlated, which means they provide essentially the same information. The similarity of the criteria will be described through a cluster analysis, leading to a dendrogram, presented in the results section.

Table 2. Cemagref ASM1 default parameter set at 20°C (Choubert *et al.*, 2009) and fractionation used in the modelling project, ASM1 parameter range values taken from the 25-75% percentiles of database results (Hauduc *et al.*, 2011) enlarged by 10% and fractionation ranges ±20% of the initial fractionation.

Kinetic parameters			
Parameter*	**Initial**	**Range**	**Unit**
$q_{XCB\_SB,hyd}$	3	1.98 - 3.3	$g\ XC_B.g\ X_{OHO}^{-1}.d^{-1}$
$K_{XCB,hyd}$	0.03	0.018 - 0.187	$g\ XC_B.g\ X_{OHO}^{-1}$
$\eta_{qhyd,Ax}$	0.4	0.36 - 0.55	-
$\mu_{OHO,Max}$	6	5.13 - 6.6	d^{-1}
$\eta_{\mu OHO,Ax}$	0.8	0.72 - 0.88	-
b_{OHO}	0.62	0.549 - 0.682	d^{-1}
$K_{O2,OHO}$	0.05	0.045 - 0.22	$g\ S_{O2}.m^{-3}$
$K_{SB,OHO}$	20	9 - 22	$g\ S_B.m^{-3}$
K_{NOx}	0.1	0.09 - 0.55	$g\ S_{NOx}.m^{-3}$
$\mu_{ANO,Max}$	0.8	0.594 - 0.99	d^{-1}
b_{ANO}	0.17	0.072 - 0.187	d^{-1}
q_{am}	0.08	0.063 - 0.088	$m^3.g\ X_{CB,N}^{-1}.d^{-1}$
K_{NHx}	0.1	0.675 - 1.1	$g\ S_{NHx}.m^{-3}$
$K_{O2,ANO}$	0.2	0.18 - 0.825	$g\ S_{O2}.m^{-3}$
$\theta_{KXCB,hyd}$	1		-
$\theta_{\mu OHO,Max}$	1.072		-
$\theta_{\mu ANO,Max}$	1.059		-
θ_{bOHO}	1.029		-
θ_{bANO}	1.027		-
$\theta_{qXCB\_SB,hyd}$	1.072		-
θ_{qam}	1.072		-

Stoichiometric parameters			
Parameter*	**Initial**	**Range**	**Unit**
$Y_{OHO,Ox}$	0.67	0.558 - 0.737	$g\ X_{OHO}.g\ XC_B^{-1}$
$Y_{OHO,Ax}$	0.54	0.496 - 0.594	$g\ X_{OHO}.g\ XC_B^{-1}$
Y_{ANO}	0.24	0.216 - 0.264	$g\ X_{ANO}.g\ S_{NOx}^{-1}$
$f_{XU\_Bio,lys}$	0.08	0.072 - 0.11	$g\ X_U.g\ X_{Bio}^{-1}$
Composition parameters			
i_{N_XBio}	0.086	0.0711 - 0.0946	$g\ N.g\ X_{Bio}^{-1}$
i_{N_XUE}	0.06	0.054 - 0.066	$g\ N.g\ X_{UE}^{-1}$
Settling parameters			
f_{ns}	0.003	0.001-0.005	-
Fractionation			From
S_U	7%	5.6-8.4%	DCO
S_B	25%	20-30%	DCO
$X_{U,Inf}$	8%	6.4-9.6%	DCO
X_{OHO}	8%	6.4-9.6%	DCO
XC_B	52%	*42.4-61.8%*	DCO
S_{NHx}	100%	-	NH4
$S_{B,N}$	22%	17.6-26.4%	Norg
$S_{U,N}$	0%	-	Norg
$XC_{B,N}$	78%	*73.8-82.4%*	Norg

* Standardised notation from Corominas *et al.* (2010)

RESULTS AND DISCUSSION

Results of the 5000 simulations on the case study example

The results of the 5000 simulations are presented for some of the target constituents in Figure 3: TSS in tanks, effluent COD, NH_4^+-N, and NO_3^--N. A histogram of the values obtained at day 30 is presented on the right side. Day 30 was chosen as representative of the normal plant operation, before the aerators broke down.

These graphs show the dependency of the model response to changes in the parameter set, compared to the observed values represented by the thick line. The histograms allow better visualisation of the spread of the modelled values compared to the observed values. It should be noted that the simulations from the 5000 parameter sets are often overestimated, especially for TSS in the tank (biomass prediction) and ammonia. This is probably due to non-suitable combinations of parameters (μ,b) for heterotrophs and autotrophs respectively, since no correlation between parameters was taken into account.

During the breakdown of the aerators, the model behaviour follows the observed values: nitrification cannot occur anymore so that the ammonia concentration increases and the nitrate concentration decreases to zero. Between days 51 to 68, the aerators are running permanently, resulting in low ammonia concentrations when nitrification is re-established, and to high nitrate concentrations. Note that for a number of parameter sets, no or only low nitrification activity is simulated, leading to high ammonia and low nitrate concentrations.

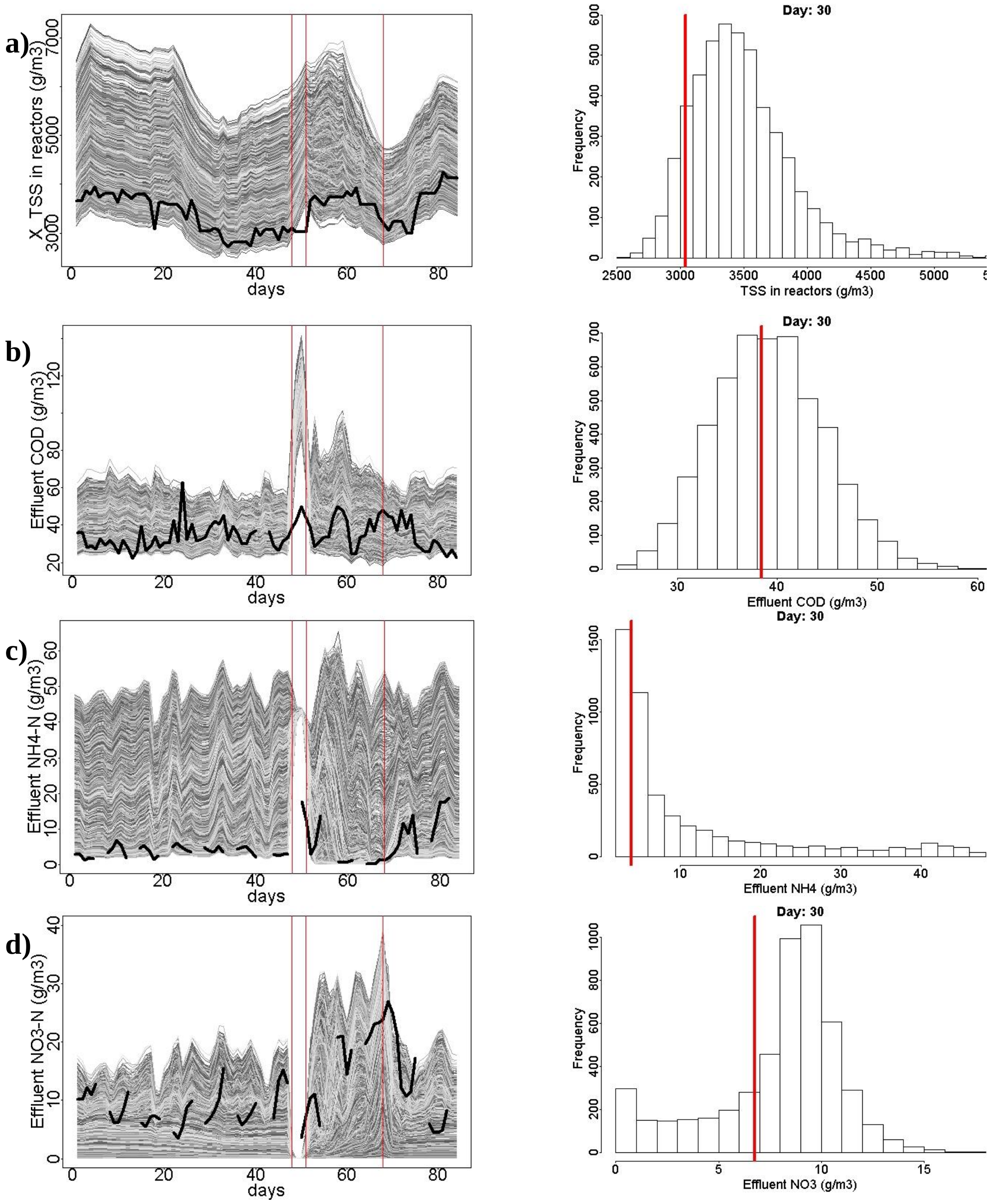

Figure 3. Results of the 5000 Simulations (left) and histograms of the values obtained at day 30 (right) for a) Effluent COD, b) Effluent TSS, c) Effluent NH_4^+ and d) Effluent NO_3^-. Bold lines correspond to the observed daily composite values. On day 48 all aerators broke down for 3 days, then from day 51 to 68 the aerators were running permanently.

Quality criteria ranges

The ranges for the relative quality criteria (classes 3 to 6) are calculated over the 5000 simulations and for all constituents. They are represented in Figure 4 using boxplots, with the thick bar indicating the median, boxes indicating 25-75% percentiles, and whiskers indicating the 5-95% percentiles. The target value is indicated on the left of each boxplot (0 or 1).

These boxplots indicate the ranges of quality criteria values that could possibly be found in a typical wastewater simulation study. These values can be used for example to set quality criteria levels in an automated calibration procedure.

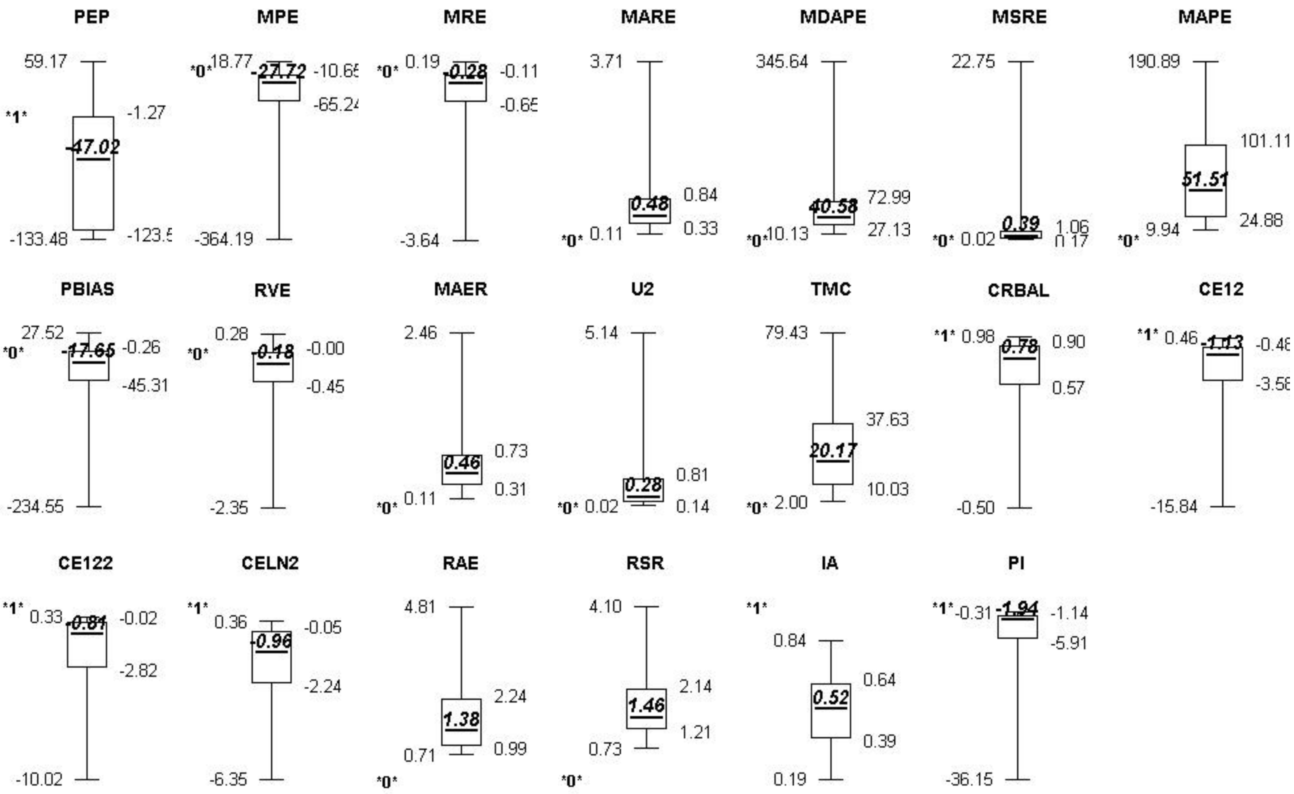

Figure 4. Boxplots of the calculated quality criteria over the 5000 simulation and for all target constituents. (Thick bar: median; boxes: 25-75% percentiles; whiskers: 5-95% percentiles; target value: *0* or *1* on the left of each boxplot).

Selection of non-correlated criteria for the case study

To describe the results of criteria correlation, a dendrogram is built from a cluster analysis (Figure 5 for TSS in the effluent), based on the dissimilarity measure calculated as 1 minus the transformed correlation coefficient. Consequently, a dissimilarity measure of 2 means that the criteria are completely anti-correlated, a dissimilarity measure of 1 means that the criteria are non-correlated and a dissimilarity measure of 0 means that the criteria are completely correlated.

For each target constituent, a dendrogram that represents the similarity between quality criteria was built, with each of the three methods: Pearson, Spearman and Kendall. The three methods lead to very similar dendrograms. However, the Kendall method better distinguishes between quality criteria and was therefore chosen. An example of the obtained dendrogram with the Kendall method is presented in Figure 5 for the effluent TSS constituent.

This dendrogram aims at helping a model user to choose non-redundant quality criteria for a planned modelling study. The selected quality criteria should be non-correlated, but the modeller should choose the number of criteria he would keep and which of them, depending on the modelling objectives and the properties of the criteria (to point out differences in peaks, bias, time lag…, see Appendix 1). For a better quality evaluation of simulations, it is advisable to choose at least anti-correlated criteria (at least 2 criteria on the example in Figure 5). As our case is an exploratory case study, there is no specific modelling objective. Consequently, an arbitrary cut-off of 0.5 was selected, and one representative criterion was chosen for each separate branch above this value. Using a cut-off value of 0.5 between 3 and 8 branches were identified depending on the constituent. Relative criteria are preferred, since they can be compared with other target constituents. For instance, RMSE is the most commonly used quality criterion in the WWTP field (Dochain and Vanrolleghem, 2001; Ahnert *et al.*, 2007; Sin *et al.*, 2008), but it is not a relative criterion. MSRE, that is close to RMSE, is then preferred.

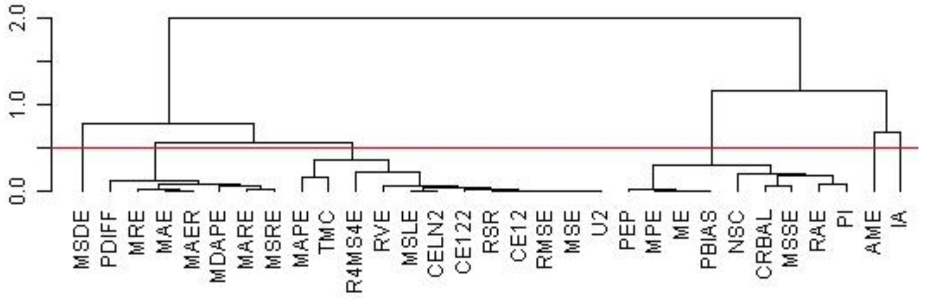

Figure 5. Dendrogram using the Kendall method of similarity between quality criteria for the effluent TSS constituent (0: correlated criteria; 1: non-correlated criteria; 2: anti-correlated criteria)

On the example of the effluent TSS constituent (Figure 5), the line fixed at 0.5 cuts 6 branches. Consequently, 6 quality criteria should be chosen, one per branch. Among the 6 branches, 3 branches lead to a single criterion (MSDE (1st from left to right), AME (5th) and IA (6th)). A choice among several criteria has to be made for only 3 branches. For the 2nd branch the MSRE (a relative criterion) is selected. In the 3rd branch the Nash-Sutcliff criterion ($CE_{1,2}$) commonly used in environmental sciences is chosen. Finally, for the 4th branch PBIAS is selected. The same procedure was carried out for the 5 other target constituents, leading to the resulting quality criteria presented in Table 3.

Table 3. List of non-correlated criteria (described in **Appendix 1**) for each constituent and finally chosen criteria

Constituant	Non-correlated criteria	Chosen criteria
TSS in Tank	PBIAS, MSDE, MSRE	MSDE, MSRE
Effluent COD	MSDE, MSRE, PBIAS, IA	MSDE, MSRE, IA
Effluent TSS	AME, $CE_{1,2}$, IA, MSDE, MSRE, PBIAS	AME, IA, MSDE, MSRE
Effluent TKN	MSDE, MSRE, PBIAS, MAPE, PEP, MSSE	-
Effluent NH4-N	MSDE, MSRE, PI, PBIAS, PEP, $CE_{1/2,2}$	MSDE, MSRE, PEP, $CE_{1/2,2}$
Effluent NO3-N	MSDE, MSRE, PBIAS, PEP, $CE_{1,2}$, IA, PI, $CE_{1/2,2}$	MSDE, MSRE, PEP, $CE_{1,2}$

This 1st clustering analysis leads to the selection of a total of 33 quality criteria for the 6 constituents, which is still a large number of criteria to deal with. Moreover, this methodology does not quantify the similarity between criteria of different constitutents. Consequently, a 2nd clustering analysis was performed on these 33 quality criteria leading to the dendrogram presented in Figure 6. This dendrogram underlines that the study of NH4-N and TKN profiles essentially provides the same information. Consequently, only the NH4-N constituent should be studied. Again, fixing a limit to 0.5 leads to a subset of 17 quality criteria that are summarised in Table 3.

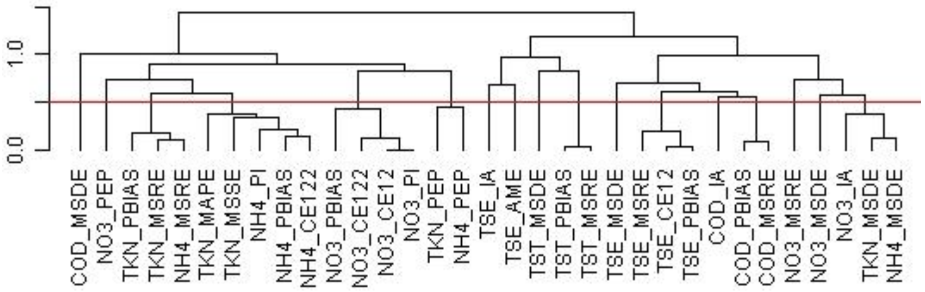

Figure 6. Dendrogram using the Kendall method of the 33 selected quality criteria (TST for TSS in tank, TSE for effluent TSS)

CONCLUSIONS

- Thirty-one criteria have been compiled and structured using two classification schemes. First they were grouped into six main *classes* following the way they are calculated: 1) single event statistics, 2) absolute criteria from residuals, 3) residuals relative to observed values, 4) total residuals relative to total observed values, 5) agreement between distributional statistics of observed and modelled data, and 6) comparison of residuals with reference values and with other models. In the second classification scheme 6 main *characteristics* of the adjustment of predicted values to observations were distinguished: i) mean error, ii) bias, iii) large errors, iv) small errors, v) peaks and vi) events dynamics.

- The criteria were evaluated on a WWTP modelling case study, based on simulation of 5000 Latin Hypercube Sampled parameter sets using parameter ranges found in literature.
- A two-step cluster analysis using Kendall correlation and dendrograms is proposed to select criteria that provide non-redundant information.

OUTLOOK

It is suggested to use the identified independent criteria for multi-objective parameter estimation, either by the use of a linear combination of the criteria (van Griensven *et al.*, 2002) or the use of Pareto optimisation methods (Yapo *et al.*, 1998; Muschalla *et al.*, 2008).

ACKNOWLEDGEMENTS

At the time of this study Hélène Hauduc was a Ph.D. student at Cemagref (Antony, France) and modelEAU (Université Laval, Québec, Canada). Peter Vanrolleghem holds the Canada Research Chair in Water Quality Modelling.

LITERATURE

Ahnert M., Blumensaat F., Langergraber G., Alex J., Woerner D., Frehmann T., Halft N., Hobus I., Plattes M., Spering V. and Winkler S. (2007). Goodness-of-fit measures for numerical modelling in urban water management – A summary to support practical applications. In: *Proceedings of 10th LWWTP Conference*, Vienna, Austria, 9-13 September 2007.

Bach M., Froehlich F., Heusch S., Hübner C., Muschalla D., Reußner F. and Ostrowski M. (2009). BlueM – a free software package for integrated river basin management. In: *Proceedings of Annual Meeting of the Society for Hydrologists and Water Managers*, Kiel, Germany, March 2009.

Choubert J.-M., Stricker A.-E., Marquot A., Racault Y., Gillot S. and Héduit A. (2009) Updated Activated Sludge Model n1 parameter values for improved prediction of nitrogen removal in activated sludge processes: Validation at 13 full-scale plants. *Water Environment Research*, **81**, 858-865.

Claeys F., de Pauw D.J.W., Benedetti L., Nopens I. and Vanrolleghem P.A. (2006). Tornado: A versatil and efficient modelling & virtual experimentation kernel for water quality systems. In: *Proceedings of Summit on Environmental Modelling and Software (iEMSs2006)*, Burlington, Vermont, USA, July 9-12 2006.

Corominas L., Rieger L., Takács I., Ekama G., Hauduc H., Vanrolleghem P.A., Oehmen A., Gernaey K.V. and Comeau Y. (2010) New framework for standardized notation in wastewater treatment modelling. *Water Science and Technology*, **61**(4), 841-857.

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 and Software*, **22**(7), 1034-1052.

Dawson C.W., Abrahart R.J. and See L.M. (2010) HydroTest: Further development of a web resource for the standardised assessment of hydrological models. *Environmental Modelling and Software*, **25**(11), 1481-1482.

Dochain D. and Vanrolleghem P. (2001) *Dynamical Modelling and Estimation in Wastewater Treatment Processes*, IWA Publishing, London, UK.

Elliott J.A., Irish A.E., Reynolds C.S. and Tett P. (2000) Modelling freshwater phytoplankton communities: an exercise in validation. *Ecological Modelling*, **128**(1), 19-26.

Gillot S. and Choubert J.-M. (2010) Biodegradable organic matter in domestic wastewaters: comparison of selected fractionation techniques. *Water Science and Technology*, **62**(3), 630-639.

Gupta H.V., Sorooshian S. and Yapo P.O. (1998) Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Water Resources Research*, **34**(4), 751-763.

Hauduc H. (2010) *Modèles biocinétiques de boues activées de type ASM: Analyse théorique et fonctionnelle, vers un jeu de paramètres par défaut [ASM type biokinetic activated sludge models: theoretical and functional analysis, toward a default parameter set]*. PhD thesis, Génie Civil et Génie des Eaux, Université Laval/AgroParisTech, Québec (CA) / Paris (FR).

Hauduc H., Gillot S., Rieger L., Ohtsuki T., Shaw A., Takács I. and Winkler S. (2009) Activated sludge modelling in practice - An international survey. *Water Science and Technology*, **60**(8), 1943-1951.

Hauduc H., Rieger L., Ohtsuki T., Shaw A., Takács I., Winkler S., Heduit A., Vanrolleghem P.A. and Gillot S. (2011) Activated sludge modelling: Development and potential use of a practical applications database. *Water Science and Technology*, **63**(8), In press.

Henze M., Grady C.P.L., Gujer W., Marais G.v.R. and Matsuo T. (2000). Activated Sludge Model No.1. In *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*. edited by M. Henze, et al., IWA Publishing, London, UK.

Legates D.R. and McCabe G.J. (1999) Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, **35**(1), 233-241.

Moriasi D.N., Arnold J.G., Van Liew M.W., Bingner R.L., Harmel R.D. and Veith T.L. (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, **50**(3), 885-900.

Muschalla D., Schneider S., Gamerith V., Gruber G. and Schro?ter K. (2008) Sewer modelling based on highly distributed calibration data sets and multi-objective auto-calibration schemes. *Water Science and Technology*, **57**(10), 1547-1554.

Orhon D., Sozen S. and Artan N. (1996) The effect of heterotrophic yield on the assessment of the correction factor for anoxic growth. *Water Science and Technology*, **34**(5-6), 67-74.

Perrin C., Andréassian V. and Michel C. (2006) Simple benchmark models as a basis for model efficiency criteria. *Large Rivers*, **17**(Arch. Hydrobiol. Suppl. 161/1-2), 221-244.

Perrin C., Michel C. and Andréassian V. (2001) Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hydrology*, **242**(3-4), 275-301.

Petersen B., Gernaey K., Henze M. and Vanrolleghem P.A. (2002) Evaluation of an ASM1 model calibration procedure on a municipal-industrial wastewater treatment plant. *Journal of Hydroinformatics*, **4**(1), 15-38.

Power M. (1993) The predictive validation of ecological and environmental models. *Ecological Modelling*, **68**(1-2), 33-50.

Rieger L., Gillot S., Langergraber G., Ohtsuki T., Shaw A., Takacs I. and Winkler S. (In Preparation) *Guidelines for Using Activated Sludge Models*, IWA Publishing, London, UK.

Seibert J. (2001) On the need for benchmarks in hydrological modelling. *Hydrological Processes*, **15**(6), 1063-1064.

Sin G., De Pauw D.J.W., Weijers S. and Vanrolleghem P.A. (2008) An efficient approach to automate the manual trial and error calibration of activated sludge models. *Biotechnology and Bioengineering*, **100**(3), 516-528.

van Griensven A. and Bauwens W. (2003) Multiobjective autocalibration for semidistributed water quality models. *Water Resources Research*, **39**(12), SWC91-SWC99.

van Griensven A., Francos A. and Bauwens W. (2002) Sensitivity analysis and auto-calibration of an integral dynamic model for river water quality. *Water Science and Technology*, **45**(9), 325-332.

Vanhooren H., Meirlaen J., Amerlinck Y., Claeys F., Vangheluwe H. and Vanrolleghem P.A. (2003) WEST: modelling biological wastewater treatment. *Journal of Hydroinformatics*, **5**(1), 27-50.

Willmott C.J., Ackleson S.G., Davis R.E., Feddema J.J., Klink K.M., Legates D.R., O'Donnell J. and Rowe C.M. (1985) Statistics for the evaluation and comparison of models. *Journal of Geophysical Research*, **90**(C5), 8995-9005.

Yapo P.O., Gupta H.V. and Sorooshian S. (1998) Multi-objective global optimization for hydrologic models. *Journal of Hydrology*, **204**(1-4), 83-97.