Environmental Modelling & Software 68 (2015) 196–204

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Review

Efficiency criteria for environmental model quality assessment: A review and its application to wastewater treatment

H. Hauduc [a, b, g, h, i], M.B. Neumann [a, c, d], D. Muschalla [a, e], V. Gamerith [e, a], S. Gillot [f], P.A. Vanrolleghem [a, *]

[a] modelEAU, Département de génie civil et de génie des eaux, Université Laval, 1065 av. de la Médecine, Québec, QC G1V 0A6, Canada
[b] Irstea UR HBAN, 1 rue Pierre-Gilles de Gennes, F-92761 Antony Cedex, France
[c] Basque Centre for Climate Change, BC3, Alameda Urquijo, 4 – 4°, 48008 Bilbao, Spain
[d] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[e] Graz University of Technology, Institute of Urban Water Management and Landscape Water Engineering, Stremayrgasse 10/I, 8010 Graz, Austria
[f] Irstea UR MALY, centre de Lyon-Villeurbanne, F-69926 Villeurbanne Cedex, France
[g] Université de Toulouse, INSA, UPS, INP, LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France
[h] INRA, UMR 792 Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France
[i] CNRS, UMR 5504, F-31400 Toulouse, France

ARTICLE INFO

Article history:
Received 5 October 2014
Received in revised form
10 February 2015
Accepted 12 February 2015
Available online

Keywords:
Model quality assessment
Modeling objective
Parameter estimation
Dissimilarity between criteria

ABSTRACT

In various cases in environmental modeling, modelers need to account for multiple variables and multiple objectives in systems with many processes occurring at different time scales. To assist the modeler to choose a relevant pool of efficiency criteria, a method is proposed to identify dissimilar criteria. A total of 30 efficiency criteria used in environmental modeling are critically reviewed and classified into six groups according to different modeling objectives. After accounting for equivalence of functional form 18 criteria remain for further analysis. To quantify the dissimilarity for the remaining criteria a methodology based on the ratio of shared parameter sets in regions of good performance is proposed. Then, for a wastewater treatment plant case-study the dissimilarity of efficiency criteria is analyzed as a function of target variables and operating conditions.

1. Introduction

The evaluation of the quality of a model is often based on visual comparisons between simulation results and observed data. Visualization can be done directly by comparing the time series or by the use of scatterplots (Ritter and Munoz-Carpena, 2013). Although visual comparison allows the modeler to evaluate easily many aspects of the model quality, it lacks objectivity and cannot be used in an automatic calibration procedure. It is therefore recommended to use both visualization and quantitative metrics (Bennett et al., 2013; Chiew and McMahon, 1993; Houghton-Carr, 1999). A recent position paper by Bennett et al. (2013) provides a comprehensive review of methods for measuring quantitative performance of environmental models. It further proposes a procedure for model performance evaluation including i) definition of model aim; ii) characterization of dataset; iii) visual overview of the overall performance; iv) selection of basic performance criteria and v) refinements and improvement of the model.

Environmental sciences, hydrology in particular, widely use mathematical comparisons of predicted and observed values (Bennett et al., 2013; Dawson et al., 2007). In contrast to hydrology, many applications of environmental modeling use multiple variables pertaining to processes occurring at various timescales. For example, in wastewater treatment (WWT) several target constituents are usually considered simultaneously during model calibration (total suspended solids (TSS), chemical oxygen demand (COD), O_2 consumption, sludge production, NH_4–N, N_{Tot} or PO_4–P in the effluent …) and thus different criteria for assessing multi-variable model quality have to be used. Furthermore, the fitting objective may be different for different target constituents: for example a modeler may want to capture the mean value of the biological tank TSS, but the dynamics of NO_3–N and PO_4–P effluent concentrations. Each efficiency criterion allows emphasizing a

* Corresponding author.
E-mail addresses: helene.hauduc@gmail.com, Helene.Hauduc@Insa-Toulouse.Fr (H. Hauduc), marc.neumann@bc3research.org (M.B. Neumann), d.muschalla@tugraz.at (D. Muschalla), gamerith@hydroconsult.net (V. Gamerith), sylvie.gillot@irstea.fr (S. Gillot), peter.vanrolleghem@gci.ulaval.ca (P.A. Vanrolleghem).

http://dx.doi.org/10.1016/j.envsoft.2015.02.004
1364-8152/© 2015 Elsevier Ltd. All rights reserved.

Abbreviations

Efficiency criteria

AME	Absolute Maximum Error
CE	Coefficient of Efficiency
$CE_{1,2}$	Nash–Sutcliffe
CrBal	Balance Criterion
IA	Index of Agreement
MAE	Mean Absolute Error
MAER	Mean Absolute Error Relative
MARE	Mean Absolute Relative Error
MdAPE	Median Absolute Percent Error
ME	Mean Error
MPE	Mean Percent Error
MRE	Mean Relative Error
MSDE	Mean Square Derivative Error
MSE	Mean Square Error
MSLE	Mean Square Logarithm Error
MSRE	Mean Square Relative Error
MSSE	Mean Square Sorted Errors
NSC	Number of Sign Changes
PBIAS	Percent Bias
PDIFF	Peak Difference
PEP	Percent Error In Peak
PI	Coefficient of Persistence
RAE	Relative Absolute Error
RMSE	Root Mean Square Error
RSR	RMSE–observation standard deviation ratio
RVE	Relative Volume Error
TMC	Total Mass Controller
U^2	Theil's Inequality Coefficient

Others

ASM	Activated Sludge Model
COD	Chemical Oxygen Demand
HRT	Hydraulic Retention Time
PE	Population Equivalent
SRT	Sludge Retention Time
TKN	Total Kjeldahl Nitrogen
TSS	Total Suspended Solids
WWT	Wastewater Treatment
WWTP	Wastewater Treatment Plant

different aspect of model behavior, but is imperfect to catch other characteristics of the model. Therefore depending on the modeling objectives, using a single criterion may lead to favor models that do not reproduce important behavior of the data (Bennett et al., 2013). Consequently, the calibration and performance assessment of a typical wastewater treatment plant model requires a “multi-variable, multi-objective” approach.

Aggregated efficiency criteria (multi-criteria, multi-objective, multi-variable) have been used in different studies and are based on the sum of normalized efficiency criteria. To sum several efficiency criteria, van Griensven and Bauwens (2003) proposed to sum them with appropriate weights to put emphasis on certain criteria/measurements. Brun et al. (2002) and Sin et al. (2008) normalized the sum of squared errors by the mean of the measurement and the standard error of the measurement respectively (χ^2 criterion). Dochain and Vanrolleghem (2001) show how this weighting method is generalized by using the inverse of the covariance matrix of the measurement errors of the different variables. This means that the multi-criteria problem is turned into a single criterion one. However, aggregating criteria that emphasize different aspects of model behavior results in the loss of the individual information they provide (Efstratiadis and Koutsoyiannis, 2010).

The aim of this study is therefore to assist modeler in the choice of a relevant pool of efficiency criteria to be used in a multi-objective problem. A critical review and classification of efficiency criteria was first undertaken covering a number of water-related disciplines (wastewater treatment, catchment hydrology, urban hydrology, climate sciences, environmental sciences …). The specificity of each criterion and class of criteria to measure the performance of the model for describing particular characteristics of the observed data is discussed. Then, the selected efficiency criteria are computed for a full scale WWTP case study, which has been modeled with 5000 different parameter sets. A procedure is proposed to identify dissimilar and thus complementary criteria based on the ratio of shared parameter sets in regions of good model performance. The dissimilarity between criteria is tested against three factors: i) the functional form; ii) the system behavior (dynamics, stiffness, degree of non-linearity: i.e. the operating region of the model) and iii) the choice of target variables including the experimental design (location of measured data in space and time).

2. Review of quantitative efficiency criteria used in environmental sciences and engineering

2.1. General methods to compare observed and predicted data

Depending on the modeling objectives, the model performance can be defined as the capability of the model to capture one or several characteristics of the observed data: mean, timing or magnitude of peaks or typical periodical variations (diurnal, weekly, seasonal …). For example, if a specific effluent limit of a wastewater treatment plant is based on a monthly average there is little sense in evaluating 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 characterize the performance of the model, different efficiency criteria may be needed. Characteristics of 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 for comparison across different state variables;
- **Comparison with a reference model** is used to define the improvement of using the model over a simpler model, such as a model defined as the mean of the observed values or the previous observed value (see e.g. Seibert, 2001), or a model describing typical variations such as an average diurnal or seasonal variation (Legates and McCabe, 1999).

Other arithmetic operations can be applied to emphasize 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 efficiency criteria on each of these subsets (Perrin et al., 2006; Moriasi et al., 2007),
- **Emphasizing small errors or low magnitude values**: a power transformation of the data with an exponent lower than 1 or a logarithmic transformation can be used,
- **Emphasizing large errors or high magnitude 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 or power values avoid compensation of negative and positive errors.

These arithmetic operations are used to modify the general metrics (e.g. error) to extract the required information, given a certain objective (e.g. to give more importance to errors at low magnitude of the variables, to emphasize maximum errors or errors on peaks). It is important to note that most of the criteria discussed in this review are based on sums (except single events statistics). Consequently, in case of datasets with variable time steps, the criteria will emphasize errors on more frequently sampled periods. A solution to overcome this problem is the use of weighted criteria inversely proportional to the sampling frequency, resulting in higher weights for isolated points (Willmott et al., 1985).

2.2. Review and classification of quantitative criteria used in environmental sciences

A literature review of efficiency criteria from water-related disciplines (catchment hydrology, urban hydrology, climate sciences, environmental sciences …) leads to a pool of thirty different quantitative efficiency criteria (Table 1). After theoretical analysis they were grouped into the following six classes:

2.2.1. Single event statistics

When the modeling objectives require accurate simulation of single events (e.g.: storm flow peaks, toxicity peaks), criteria are needed to characterize the goodness-of-fit of the model for this event. The single event statistics peak difference (PDIFF) (Gupta et al., 1998) and percent error in peak (PEP) (Dawson et al., 2007) aim at characterizing the difference between the observed and the modeled peak. However, they do not evaluate whether the peaks occurred at the same time. Consequently, in case of multiple events occurring in a given time-series, the corresponding peaks must first be tagged.

2.2.2. Absolute criteria from residuals

Absolute criteria are based on the sum of residuals (difference between observed O_i and predicted P_i values at time step i), generally averaged with the number of data, n. A low value of this criterion means a good agreement between observation and simulation. The general form of these efficiency criteria is presented in Equation (1) (where γ is an exponent):

$$\mathrm{E}_\gamma = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^\gamma \tag{1}$$

The simplest efficiency criterion of this class is the mean error (ME) with $\gamma = 1$, which allows identifying the existence of systematic bias, i.e. the characteristic of a model leading to systematic over- or under-prediction (Power, 1993). However, with this criterion errors can compensate each other, and no information on the magnitude of the errors is obtained. This can be solved by using $|O_i - P_i|$ to obtain the mean absolute error (MAE) which indicates the average magnitude of the model error (accuracy) (Willmott et al., 1985), but does not indicate the direction of the deviation.

The mean square error (MSE) with $\gamma = 2$ also avoids error compensations and furthermore emphasizes high errors, but is more widely applied in the form of the root mean square error ($\mathrm{RMSE} = \mathrm{MSE}^{0.5}$) in the same units as the variables (Willmott et al., 1985). It indicates the overall agreement between predicted and observed data and it can be used in conjunction with the MAE to provide information on the prominence of outliers in the dataset (Bennett et al., 2013). To put even more emphasis on the larger errors, the fourth root mean quadruples error can be used (R4MS4E) (Dawson et al., 2007).

The mean square logarithm error (MSLE) is the MSE calculated with the natural logarithm of the predicted and observed value, which emphasizes small errors (Dawson et al., 2010). For this metric, a small number ε (negligible) is introduced to avoid a zero value in the logarithm, in the same way as in the denominator of some metrics.

The absolute maximum error (AME) indicates the maximum error of the model and is very sensitive to outliers (Gupta et al., 1998).

The mean square of sorted errors (MSSE) is calculated based on sorted observed and predicted data (Van Griensven and Bauwens, 2003). Observations and predictions are sorted independently one from the other to allow comparison of empirical density distributions. This criterion is then insensitive to the timing of the events.

The number of sign changes (Gupta et al., 1998), or equivalently the number of runs (a run is a series of residuals with the same sign, Dochain and Vanrolleghem (2001)), counts the number of times the residual ($\mathrm{O_i} - \mathrm{P_i}$) sign changes. The minimum value is zero and the maximum n, the length of the dataset. A value close to zero indicates a systematic error (over-estimating or under-estimating model) but a more consistent model. A value close to n indicates a random error. This criterion should be analyzed in association with other criteria to evaluate the adequacy of a model and in particular to evaluate whether the residuals behave as random, independent measurement errors (Dochain and Vanrolleghem, 2001).

2.2.3. Criteria evaluating event dynamics

The mean square derivative error (MSDE) is the square of the differences of predicted and observed variations between two time steps (Dawson et al., 2010). This criterion penalizes noisy time series and series with a timing error; it thus allows evaluating the peak's timing.

2.2.4. Residuals normalized with observed values

At each time step, the error is related to the corresponding observed value, which provides a dimensionless criterion. Furthermore these criteria give more weights to low magnitude measurements. A low value of this criterion means a good agreement between observation and simulation. The general form of these efficiency criteria is presented in Equation (2) (where γ is an exponent, ε is a small (negligible) value added to handle zero data).

$$\mathrm{RE}_\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{O_i - P_i}{O_i + \varepsilon}\right)^\gamma \tag{2}$$

The mean percentage error (MPE) (Power, 1993) and mean relative error (MRE) (Dawson et al., 2007) provide the average relative model error with $\gamma = 1$. However, negative and positive errors can compensate for each other. This is overcome by the mean absolute relative error (MARE) (Petersen et al., 2002) and by the mean square relative error (MSRE) with $\gamma = 2$, which furthermore emphasizes larger relative errors (Dawson et al., 2007).

Table 1
List of efficiency criteria (O for observed data; P for predicted data; n for number of data; ε is a small value added when necessary to handle zero data values). In *Unit* column, C stands for concentration or any other target variable unit.

Criteria name	Equation	Characteristics				Emphasizes			
		Minimum value	Maximum value	Optimal value	Unit	Mean	Large errors	High magnitudes	Low magnitudes
1 – Single event statistics									
Peak Difference	$\mathrm{PDIFF} = \max(\{O_i\}) - \max(\{P_i\})$	$-\infty$	$+\infty$	0	C			+	
Percent Error In Peak	$\mathrm{PEP} = 100 \times \frac{\max(O_i)-\max(P_i)}{\max(O_i)}$	$-\infty$	$+\infty$	0	%			+	
2 – Absolute criteria									
Mean error	$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)$	$-\infty$	$+\infty$	0	C	+			
Mean Absolute error	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert O_i - P_i\rvert$	0	$+\infty$	0	C	+			
Root Mean Square Error	$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(O_i-P_i)^2}{n}}$	0	$+\infty$	0	C		+		
Mean Square Error	$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2$	0	$+\infty$	0	C^2	+			
Fourth root mean quadruples error	$\mathrm{R4MS4E} = \sqrt[4]{\frac{\sum_{i=1}^{n}(O_i-P_i)^4}{n}}$	0	$+\infty$	0	C		++		
Mean Square Logarithm Error	$\mathrm{MSLE} = \frac{1}{n}\sum_{i=1}^{n}(\ln(O_i+\varepsilon) - \ln(P_i+\varepsilon))^2$	0	$+\infty$	0	–				+
Absolute Maximum Error	$\mathrm{AME} = \max(\lvert O_i - P_i\rvert)$	0	$+\infty$	0	C		++		
Mean Square Sorted Errors	$\mathrm{MSSE} = \frac{1}{n}\sum_{j=1}^{n}(O_j - P_j)^2$	0	$+\infty$	0	C^2		+		
Number of Sign Changes	NSC	0	$+\infty$	0	–				
3 – Derivative error									
Mean Square Derivative Error	$\mathrm{MSDE} = \frac{1}{n-1}\sum_{i=1}^{n}((O_i - O_{i-1}) - (P_i - P_{i-1}))^2$	0	$+\infty$	0	C				
4 – Relative error criteria									
Mean Percent Error	$\mathrm{MPE} = \frac{100}{n}\sum_{i=1}^{n}\frac{O_i-P_i}{O_i+\varepsilon}$	$-\infty$	$+\infty$	0	%	+			+
Mean Relative Error	$\mathrm{MRE} = \frac{\mathrm{MPE}}{100}$	$-\infty$	$+\infty$	0	–	+			
Mean Absolute Relative Error	$\mathrm{MARE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert O_i-P_i\rvert}{O_i+\varepsilon}$	0	$+\infty$	0	–	+			+
Median Absolute Percent Error	$\mathrm{MdAPE} = \mathrm{Median}\left(100 \times \frac{\lvert O_i-P_i\rvert}{O_i+\varepsilon}\right)$	0	$+\infty$	0	%	+			+
Mean Square Relative Error	$\mathrm{MSRE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{O_i-P_i}{O_i+\varepsilon}\right)^2$	0	$+\infty$	0	–		+		+
5 – Sum of residuals relative to sum of observed values									
Percent Bias	$\mathrm{PBIAS} = 100 \times \frac{\sum_{i=1}^{n}(O_i-P_i)}{\sum_{i=1}^{n}O_i}$	$-\infty$	$+\infty$	0	%	+			
Relative Volume Error	$\mathrm{RVE} = \frac{\mathrm{PBIAS}}{100}$	$-\infty$	$+\infty$	0	–	+			
Total Mass Controller	$\mathrm{TMC} = 100 \times \left\lvert\frac{\sum_{i=1}^{n}O_i}{\sum_{i=1}^{n}P_i} - 1\right\rvert$	0	$+\infty$	0	%	+			
Balance Criterion	$\mathrm{CrBal} = 1 - \left\lvert\sqrt{\frac{\sum_{i=1}^{n}P_i}{\sum_{i=1}^{n}O_i}} - \sqrt{\frac{\sum_{i=1}^{n}O_i}{\sum_{i=1}^{n}P_i}}\right\rvert$	$-\infty$	1	1	–		+		
Mean Absolute Error Relative	$\mathrm{MAER} = \frac{\sum_{i=1}^{n}\lvert O_i-P_i\rvert}{\sum_{i=1}^{n}O_i}$	0	$+\infty$	0	–	+			
Theil's Inequality Coefficient	$\mathrm{U}^2 = \frac{\sum_{i=1}^{n}(O_i-P_i)^2}{\sum_{i=1}^{n}O_i^2}$	0	$+\infty$	0	–		+		
6 – Comparison of residuals with reference values and with other models									
Coefficient of Efficiency (Nash–Sutcliffe)	$CE_{1,2} = 1 - \frac{\sum_{i=1}^{n}(O_i-P_i)^2}{\sum_{i=1}^{n}(O_i-\overline{\mathrm{O}})^2}$	$-\infty$	1	1	–		+		
RMSE–observation standard deviation ratio (RSR)	$\mathrm{RSR} = \frac{\sqrt{\sum_{i=1}^{n}(O_i-P_i)^2}}{\sqrt{\sum_{i=1}^{n}(O_i-\overline{\mathrm{O}})^2}}$	$-\infty$	1	0	–	+			
Coefficient of Efficiency variations	$CE_{1/2,2} = 1 - \frac{\sum_{i=1}^{n}(\sqrt{O_i}-\sqrt{P_i})^2}{\sum_{i=1}^{n}(\sqrt{O_i}-\overline{\sqrt{\mathrm{O}}})^2}$	$-\infty$	1	1	–				+
	$CE_{ln,2} = 1 - \frac{\sum_{i=1}^{n}(\ln(O_i+\varepsilon)-\ln(P_i+\varepsilon))^2}{\sum_{i=1}^{n}(\ln(O_i+\varepsilon)-\overline{\ln(O+\varepsilon)})^2}$	$-\infty$	1	1	–				++
Relative Absolute Error	$\mathrm{RAE} = 1 - \frac{\sum_{i=1}^{n}\lvert O_i-P_i\rvert}{\sum_{i=1}^{n}\lvert O_i-\overline{\mathrm{O}}\rvert}$	0	$+\infty$	0	–	+			
Index of Agreement	$\mathrm{IA} = 1 - \frac{\sum_{i=1}^{n}(O_i-P_i)^2}{\sum_{i=1}^{n}(\lvert P_i-\overline{\mathrm{O}}\rvert-\lvert O_i-\overline{\mathrm{O}}\rvert)^2}$	0	1	1	–	+	+		
Coefficient of Persistence	$\mathrm{PI} = 1 - \frac{\sum_{i=1}^{n}(O_i-P_i)^2}{\sum_{i=1}^{n}(O_i-O_{i-1})^2}$	$-\infty$	1	1	–		+		

An alternative criterion is the median of the absolute relative error (MdAPE) expressed in percentage (Dawson et al., 2007). This criterion is less affected by outliers and the form of the errors distribution.

2.2.5. *Sum of residuals normalized with sum of observed values*

For the criteria presented in this section, the sum of errors is related to the sum of observed values, without any correspondence to the time step. The general form of these efficiency criteria is presented in Equation (3) (with γ an exponent). A low value of this criterion means a good agreement between observation and simulation. These criteria correspond to the visual comparison of predicted and observed cumulative plots. In the wastewater field these criteria can be relevant for analyzing influent and effluent pollutant loads by summing the fluxes.

$$\mathrm{TRE}_\gamma = \frac{\sum_{i=1}^{n}(O_i - P_i)^\gamma}{\sum_{i=1}^{n}O_i^\gamma} \quad (3)$$

The percent bias (PBIAS) (Dawson et al., 2007) and relative volume error (RVE) are the sum of errors related to the sum of observed values, expressed in percentage or as relative value. These criteria measure an overall adequacy between distributional statistics of predicted and observed data. The total mass controller (TMC) criterion used by van Griensven and Bauwens (2003) is a transformation of the RVE, however with the loss of information on the direction of the deviation.

Perrin et al. (2001) use the balance criterion (CrBal) which is a combination of TMC and RVE. The difference between the inversed fractions penalizes larger differences between observed and predicted cumulative values.

For these criteria the errors can be compensated. This is overcome by the relative mean absolute error (MAER) (Elliott et al., 2000) and by Theil's inequality coefficient used by Power (1993) and Elliott et al. (2000), which is the mean square error divided by the sum of observed data. It emphasizes larger errors.

2.2.6. Comparison of residuals with reference values or with other models

These criteria compare the residuals with residuals obtained with a reference model $\tilde{\mathbf{P}}$, such as a model describing the mean value ($\tilde{\mathbf{P}}_\mathbf{i} = \overline{\mathbf{O}}$) or the previous value ($\tilde{\mathbf{P}}_\mathbf{i} = \mathbf{O}_{\mathbf{i}-1}$). The general form of these efficiency criteria is presented in Equation (4) (with α and γ an exponent).

$$\mathrm{CE}_{\alpha,\gamma} = 1 - \frac{\sum_{i=1}^{n}\left(O_i^\alpha - P_i^\alpha\right)^\gamma}{\sum_{i=1}^{n}\left(O_i^\alpha - \tilde{P}_i^\alpha\right)^\gamma} \tag{4}$$

The first criterion is the Nash–Sutcliffe criterion ($\mathrm{CE}_{1,2}$), a widely used criterion in hydrology. The values range between −∞ and 1. A value of zero means that the model is not better than the “no knowledge” model, which is characterized by the mean value of observations. This criterion is sensitive to extreme values. From a functional analysis it follows that it is equivalent to the RMSE-observation standard deviation ratio (RSR) which is the RMSE of the predicted data divided by the RMSE of the no knowledge model (mean of observed values) (Moriasi et al., 2007). Most importantly it leads to the same optimal parameter set (same location of the minimum). Its values are in the magnitude of the target constituent unit and can be compared to the RAE (see below) to indicate the influence of larger errors.

The second criterion $\mathrm{CE}_{1/2,2}$ is close to the Nash–Sutcliffe criterion, but it is calculated from the root values, which emphasizes low magnitudes and the third criterion $\mathrm{CE}_{\ln,2}$ is calculated from the logarithms of the values, which emphasizes very low magnitudes (Perrin et al., 2001).

The relative absolute error (RAE) compares the sum of absolute residuals to the residuals of the no knowledge model (mean of observed values) (Legates and McCabe, 1999). This criterion does not allow error compensation.

The index of agreement (IA) is the ratio of the sum of squared errors (SSE) and the largest potential error with respect to the mean of observed values (Willmott et al., 1985). This parameter is sensitive to the model mean and to the peak values, and is insensitive to low magnitude values.

The coefficient of persistence (PI) is close to the Nash–Sutcliffe criterion, but the simplistic model used is the last observed value instead of the mean of observed values (Moriasi et al., 2007).

Krause et al. (2005) use relative deviations for these criteria in order to make these criteria less sensitive to the effect of magnitude variations in the dataset.

3. Material and methods

3.1. Dissimilarity analysis

Model simulations are performed for a large number of parameter sets (n = 5000), sampled from ranges defined appropriately by the modeler (see section 3.2.2). For each simulation the efficiency criteria are calculated. For each efficiency criterion the parameter sets that lead to the (α = 1%) best performance are selected. In a pairwise manner it is tested how many sets n_{cp} (number of common parameter sets) are shared within the subsets of two criteria. Thus, the distance d is obtained with following equation (5):

$$d = 1 - \frac{n_{cp}}{n \cdot \alpha} \tag{5}$$

Two efficiency criteria with a distance of d = 0 means that the parameter sets leading to the best model performance for both criteria are the same. This implies that the two criteria contain very similar information. Only the α% best parameter sets are considered because the criteria could be very dissimilar for many of the parameter sets tested in the Monte Carlo analysis that are poor-performing and thus would normally not be relevant to the selection of the best parameter set. It should however be noted that in this study, the choice of the α = 1% best parameter sets to compute cluster analysis and determination of similar sets of criteria was chosen arbitrarily by the authors. This threshold could be reduced with increasing number of Monte Carlo simulations (n). Testing the sensitivity of dissimilarity towards changes in the threshold value was out of the scope of the paper.

A cluster analysis is then applied to the distance measure leading to a dendrogram characterizing dissimilarity between criteria. The hierarchical cluster analysis is performed with the “hclust” function of R (R Core Team, 2012, http://www.r-project.org/), in which the distance between two groups is re-computed following the Lance–Williams formula as the distance between the most remote pair of elements (Lance and Williams, 1967).

3.2. Case study

3.2.1. Description of the wastewater treatment plant (WWTP)

The studied municipal WWTP is located in France and has a capacity of about 250.000 population equivalent (PE) and is configured in two parallel lanes operating under similar conditions. Each lane consists of a plug-flow tank with a predenitrification zone. The simulation period consists of 84 consecutive days, the first half of this period exhibiting typical operating conditions. On day 48, all aerators broke down for 3 days. Then from day 51–68 the aerators were running permanently. On day 68 normal operation is reinstated. Differences in similarity between criteria for two periods (normal condition (day 1–48) and disturbed condition (day 48–68)) are explored. The target constituents include total suspended solids (TSS) in the biological reactor as well as TSS, COD, total Kjeldahl nitrogen (TKN), nitrate and ammonia in the effluent. All constituents are measured daily as flow-proportional daily averages (supplementary material).

3.2.2. Model and parameter ranges

The 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 most commonly used model, for which parameter value ranges are known (Hauduc et al., 2011, 2009). Biokinetic parameter ranges were obtained from a database of modeling projects (Hauduc et al., 2011). As no correlation between parameters could be identified from this extensive modeling projects database (Hauduc et al., 2011), the parameters were considered to be independent. Furthermore, ranges for wastewater fractionation parameters were included as no reliable fractionation information on the plant influent was available. Overall, 14 kinetic, 4 stoichiometric, 2 compositional, 1 settling and 5 fractionation parameters were characterized by value ranges (Table 1 in supplementary material).

3.3. Simulations and efficiency criteria calculation

The proposed procedure is based on Monte Carlo simulation of a large number of (n = 5000) parameter sets. The parameter sets are sampled in a Latin hypercube from the ranges provided in Table 1 of supplementary material. The sampling is performed with the R software (R Core Team, 2012, http://www.r-project.org/).

The n = 5000 simulations of the WWTP model were carried out in Tornado (Claeys et al., 2006), the generic kernel of WEST software (mikebydhi.com). To ensure correct initial steady-state conditions for the 84 days of dynamic simulation of each parameter set, 100 days (>3 times the Sludge Retention Time (SRT)) were first simulated under pseudo steady-state conditions (alternating aeration periods, constant influent).

4. Results and discussion

4.1. Results of the Monte Carlo simulation

The results of the n = 5000 simulations are presented for selected target constituents in Fig. 1: TSS in tanks, effluent COD, NH_4–N, and NO_3–N.

These graphs show the dependency of the model response (gray shaded lines) to changes in the parameter set, compared to the

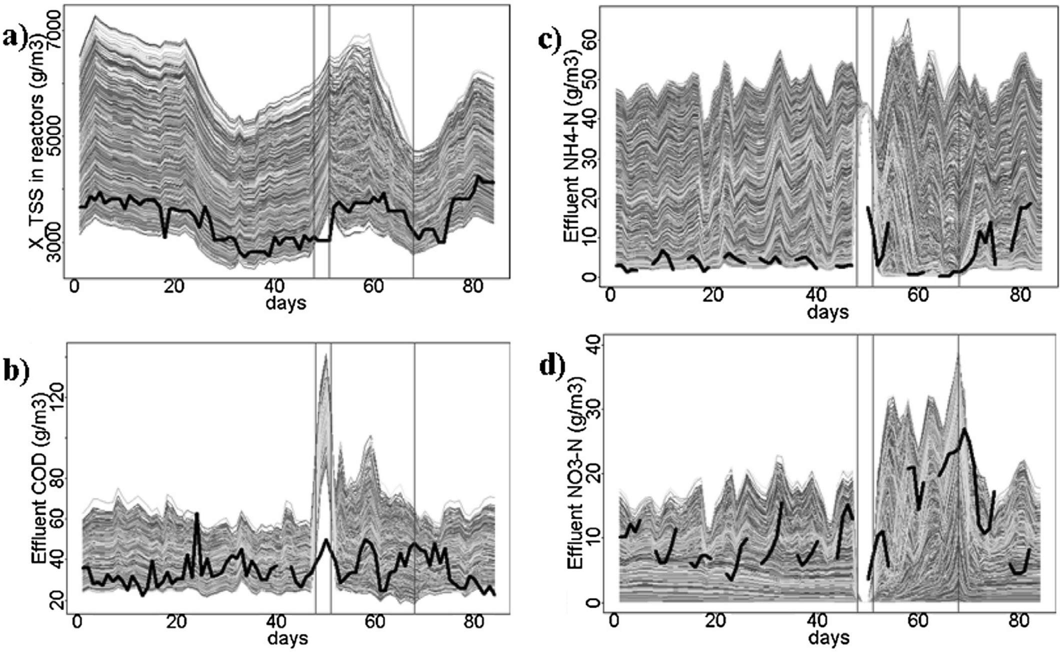

Fig. 1. Results of the n = 5000 simulations for a) Effluent TSS, b) Effluent COD, c) Effluent NH_4–N and d) Effluent NO_3–N. Bold lines correspond to the observed daily composite values. On day 48 all aerators broke down for 3 days, then from day 51 to 68 the aerators were running permanently (days 48, 51 and 68 indicated by vertical lines).

observed values represented by the thick line. During the breakdown of the aerators, the model behavior seems to follow the observed tendency of target constituents: As nitrification cannot occur anymore the ammonium (NH_4–N) concentration increases and the nitrate concentration (NO_3–N) decreases to zero. Between days 51–68, the aerators are running permanently, resulting in low ammonium concentrations when nitrification is re-established, and to high nitrate concentrations.

4.2. *Analysis of dissimilar criteria*

4.2.1. *Equivalence due to functional form*

The analysis of functional forms is performed based on the 30 reviewed quantitative efficiency criteria. It should be noted that criteria presented in the classes 5 and 6 of Table 1, namely the total relative error criteria and the comparison of residuals with reference values and with other models, are generally computed from an absolute criterion and a metric based only on observed data. Those criteria then have a different significance than the absolute criteria, but are highly correlated to the absolute criteria from which they are computed. This analysis leads to the 18 non-equivalent groups of criteria listed in Table 2 that were used for further analysis in the case study.

4.2.2. *Dissimilarities due to the choice of variables*

The location of sensors within the plant and the frequency of measurements (continuous sensors, daily grab samples, 2 samples a week …) may affect the relevance of some efficiency criteria (e.g. single event statistics, event dynamics …). Some variables may also be correlated (e.g. nitrate and ammonium, total suspended solids and volatile suspended solids), leading to higher similarity between criteria for some of the variables.

For each target constituent in the dynamic simulation periods of the case study, a selection of 16 non-equivalent efficiency criteria based on Table 2 (section 4.2.1.) were automatically calculated (the authors have chosen to not include TMC and CrBal in this calculation). To analyze the results of the case study efficiency criteria, for each target constituent and dataset partition (day 1–47 and day 48–84), a dendrogram was built from a cluster analysis based on the distance measures calculated with equation (5). The dendrograms computed for the efficiency criteria for effluent COD, NH_4–N, NO_3–N and reactor TSS are presented in Fig. 2. These dendrograms provide a global view of the relationship between efficiency criteria for the case study by identifying the efficiency criteria that are most similar to each other: a node height close to zero means they provide essentially the same information, whereas a node height close to one means they provide completely different information.

Scatterplots for the $\alpha = 1\%$ best parameter sets of selected pairs of criteria are shown in Fig. 3 to illustrate the relationship among

Table 2
Identification of non-equivalent criteria. Group-representative criteria selected for the case study are in bold.

Efficiency criteria	Functional form characteristic
PEP, PDIFF	Errors in peaks
MPE, MRE	Relative error
MAE, MAER, RAE	Absolute error
MARE	Absolute relative error
RMSE, MSE, U^2, $CE_{1,2}$, RSR	Squared error (large errors)
MSRE	Squared relative error (large errors)
R4MS4E	Quadrupled error (very large errors)
AME	Maximum error (very large errors)
MdAPE	Median relative error
$\mathbf{CE_{1/2,2}}$	Error of variables roots (low magnitudes)
MSLE, $CE_{ln,2}$	Error of variables logarithm (very low magnitudes)
MSSE	Error in predicted distribution
MSDE	Error in timing
PBIAS, ME, RVE	Global error
TMC	Adequacy of observed and predicted cumulative values
CrBal	Adequacy of observed and predicted cumulative values (large errors)
IA	Comparison of model prediction to mean of observed values
PI	Comparison of model prediction to last observed value

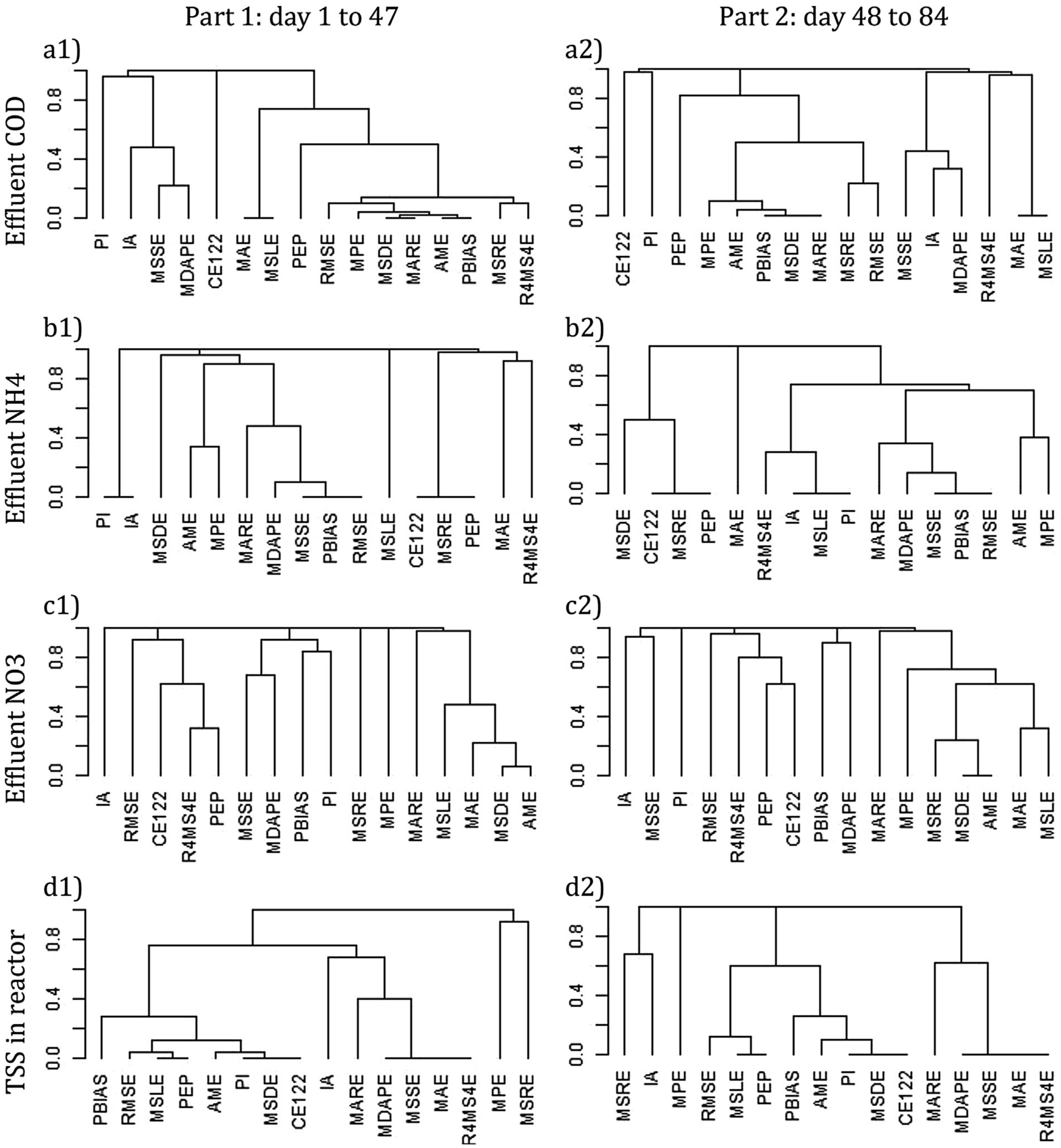

Fig. 2. Dendrograms (0: similar criteria; 1: dissimilar criteria) obtained for effluent COD, NH_4–N, NO_3–N and TSS in reactor; and for the two partitions of the dataset (left (a1–d1): normal operating condition; right (a2–d2) disturbed operating condition).

the different efficiency criteria and the computation of the distance measure *d*. Fig. 3 a and b represent the scatterplots of MSLE and MAE for effluent COD and NH_4–N respectively. For effluent COD these two efficiency criteria have an overlapping rate of $OL = 68\%$, meaning that 68% of their $\alpha = 1\%$ best parameter sets are shared. The distance for these two efficiency criteria computed with equation (5) is 0.32 ($=1 - 0.68$). In the dendrogram a1 (Fig. 2) the two efficiency criteria are then close to each other. Note that the distance in the dendrogram (close to zero for this pair of criteria) is not equal to the actual calculated distance d that feeds the hierarchical cluster analysis (0.32) as it has been re-computed following the Lance–Williams formula (section 3.1). The same pair of efficiency criteria for effluent NH_4–N shows an overlapping rate of only 2% (Fig. 3 b). This low overlapping rate indicates that the efficiency criteria provide different information, and are consequently complementary. This leads to a large distance in the dendrogram for MAE and MSLE (closest node distant from 1) as can be seen in dendrograms b1 and b2 of Fig. 2.

This behavior of the couple MAE/MSLE may be explained by analyzing the functional form and the data. The functional difference between MAE and MSLE is the logarithm used for observed and predicted result in MSLE, the difference then being squared. These operations allow emphasizing errors on low magnitude results, whereas MAE operations change any prominence of low or high magnitudes or small or large errors. The analysis of Fig. 1 shows that predicted values are much more variable for effluent NH_4–N than for COD. The dissimilarity of MAE and MSLE for effluent NH_4–N could then reveal the prominence of errors on low

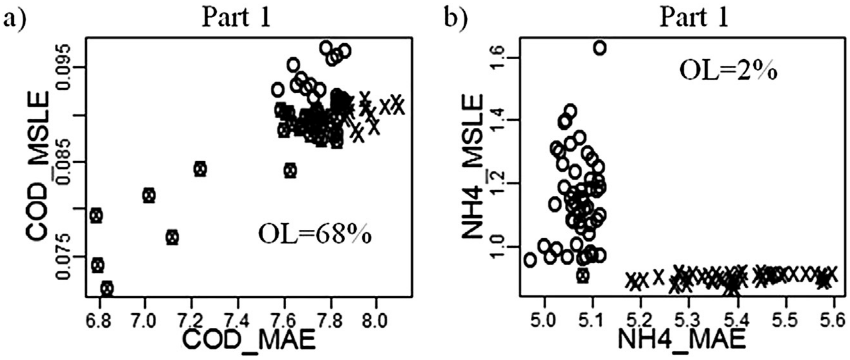

Fig. 3. Scatterplots for selected pairs of efficiency criteria for COD and NH_4: circles represent the $\alpha = 1\%$ best parameter sets of the efficiency criterion on the x-axis and crosses represent the $\alpha = 1\%$ best parameter sets of the efficiency criterion on the y-axis. Overlap (OL) occurs where crosses and circles coincide. The distance d is defined as 1 − OL.

magnitudes in the computation of the MSLE metric for effluent NH_4–N. If one of the objectives of the modeling project is to accurately predict the effluent NH_4–N, it is then relevant to choose the MSLE criteria over or in addition to the MAE, whereas for effluent COD the simpler MAE criterion would be sufficient to assess the model quality. This example illustrates that a same pair of criteria that are not similar in their functional form, may or may not be similar for a particular variable.

4.2.3. Dissimilarities due to operating conditions

The similarity between two efficiency criteria can also be affected by the system behavior, here, for instance, the operating mode. For example, in case of a narrow range of data values (constant aeration, nitrification/denitrification control by sensors …), using transformations such as power or logarithm will not allow catching different model behavior and will result in similar criteria.

The clustering for the two operational periods (Fig. 2) is very similar for all target constituents. The main difference is found for the behavior of R4MS4E, which is very similar to many other efficiency criteria (RMSE and AME among them) for effluent COD of part 1 (dendrogram a1), whereas it is dissimilar from any other efficiency criterion for effluent COD of part 2 (dendrogram a2). The similarity of R4MS4E, RMSE and AME in part 1 suggests that the errors of the models are quite homogenous, whereas some larger errors appear in the simulation of part 2, which is also revealed by the larger dissimilarity between AME and RMSE. The behavior is totally different for effluent NH_4–N, where R4MS4E is very dissimilar to any other criteria in part 1 and is quite similar to some of the criteria in part 2 and among them, of MSLE. This reveals that in part 1 large errors exist but are not necessarily correlated to high or low magnitude values, whereas the similarity of R4MS4E and MSLE in part 2 shows that the large errors are related to low magnitude values.

4.3. Discussion on choosing dissimilar criteria

The modeler has to carefully choose the efficiency criteria in view of the objectives of the modeling project and specificities of the plant and dataset. This choice may lead to several relevant criteria depending on objectives and on target variables. As an example, RMSE is often selected by modelers as it quantifies the global error of the model in the same unit as the target constituent (Bennett et al., 2013; Boyle et al., 2000; Legates and McCabe, 1999; Ritter and Munoz-Carpena, 2013). However RMSE tends to overemphasize fitting of peaks and higher values, which often leads to biased simulations in case of datasets with a wide range of values. Other absolute criteria may then be preferred or should be combined with a total relative error criterion, such as MPE, depending on the modeling objectives (Boyle et al., 2000; Ritter and Munoz-Carpena, 2013). In this case-study, the modeler may evaluate the ability of the model to reproduce i) the average TSS in the biological tanks (by applying RMSE and MPE criteria to the biweekly TSS measurements), and ii) the diurnal dynamics of nitrate to ensure an hourly effluent limit (by applying PDIFF and MSDE criteria to the continuous sensor measurements of effluent nitrate). This leads to a multi-criteria, multi-objective and multi-variable study.

Contrary to aggregated efficiency criteria as presented in the introduction, the use of Pareto optimization methods and multiobjective evolutionary algorithms (Efstratiadis and Koutsoyiannis, 2010; Muschalla et al., 2008; Yapo et al., 1998) allow taking advantage of information provided by each individual criterion selected. The dissimilarity analysis presented here could be used to avoid using similar criteria thus reducing the dimensionality of the computationally expensive Pareto optimization. The choice of one criterion in a pool of similar criteria is then essentially led by the preference of the modeler and/or for results visualization or discussion.

The aim of this case study was to illustrate the different kind of similarities in criteria that depend on the dataset under study. It should therefore be noted that the results obtained by the dissimilarity analysis are always case study dependent. They depend on the operating region of the model (determined by inputs, parameter values, temporal and spatial resolution we are interested in) and the target variables including the experimental design (location of measured data in space and time). We suggest an approach with which the modeler can determine dissimilarity in his/her case study starting from the 18 identified non-equivalent criteria. The only case-study independent conclusions that can be made are those presented in paragraph 4.2.1 “equivalence due to functional form”.

5. Conclusions

Thirty efficiency criteria to evaluate the environmental models were compiled and grouped into six *classes*: 1) single event statistics, 2) absolute criteria from residuals, 3) derivative errors, 4) relative error criteria, 5) total relative error criteria, and 6) comparison of residuals with reference values and with other models.

In a first step criteria with equivalent functional form were identified, leading to 18 groups. From each group a representative criterion was sub-selected for quantitative evaluation in an illustrative wastewater treatment plant modeling case study considering four target variables and two operating conditions.

A methodology was proposed to quantify dissimilarity between remaining criteria. It is based on assessing the ratio of shared parameter sets in the regions of best model performance for pairwise criteria. The application of this methodology to the WWTP modeling case study illustrated how dissimilarity between efficiency criteria depends not only on their functional form but also on the system behavior and on the experimental design. Varying any of these factors can change the dissimilarity between criteria.

The proposed methodology can assist the modeler to choose a relevant pool of dissimilar efficiency criteria in the presence of multiple objectives and variables.

Acknowledgments

At the time this study was initiated, Hélène Hauduc was a Ph.D. student at Irstea (former Cemagref, Antony, France) and modelEAU (Université Laval, Québec, Canada). This work was also made possible through the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Peter Vanrolleghem holds the Canada Research Chair in Water Quality Modeling.

Charles Perrin (Irstea, Antony, France) is kindly acknowledged for his advice on this article.

The authors would like to thank the three anonymous reviewers and the Editor for their comments, which have greatly improved the quality of this paper.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2015.02.004.

References

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20. http://dx.doi.org/10.1016/j.envsoft.2012.09.011.

Boyle, D.P., Gupta, H.V., Sorooshian, S., 2000. Toward improved calibration of hydrologic models: combining the strengths of manual and automatic methods. Water Resour. Res. 36, 3663–3674.

Brun, R., Kuehni, M., Siegrist, H., Gujer, W., Reichert, P., 2002. Practical identifiability of ASM2d parameters – systematic selection and tuning of parameter subsets. Water Res. 36, 4113–4127.

Chiew, F., McMahon, T., 1993. Assessing the adequacy of catchment streamflow yield estimates. Aust. J. Soil Res. 31, 665–680. http://dx.doi.org/10.1071/SR9930665.

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

Dawson, C.W., Abrahart, R.J., See, L.M., 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ. Model. Softw. 22, 1034–1052.

Dawson, C.W., Abrahart, R.J., See, L.M., 2010. HydroTest: further development of a web resource for the standardised assessment of hydrological models. Environ. Model. Softw. 25, 1481–1482.

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

Efstratiadis, A., Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrol. Sci. J. 55, 58–78. http://dx.doi.org/10.1080/02626660903526292.

Elliott, J.A., Irish, A.E., Reynolds, C.S., Tett, P., 2000. Modelling freshwater phytoplankton communities: an exercise in validation. Ecol. Model. 128, 19–26.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. Water Resour. Res. 34, 751–763.

Hauduc, H., Gillot, S., Rieger, L., Ohtsuki, T., Shaw, A., Takács, I., Winkler, S., 2009. Activated sludge modelling in practice - an international survey. Water Sci. Technol. 60, 1943–1951.

Hauduc, H., Rieger, L., Ohtsuki, T., Shaw, A., Takács, I., Winkler, S., Heduit, A., Vanrolleghem, P.A., Gillot, S., 2011. Activated sludge modelling: development and potential use of a practical applications database. Water Sci. Technol. 63, 2164–2182.

Henze, M., Grady, C.P., Gujer, W., Marais, G.v. R., Matsuo, T., 2000. Activated sludge model No.1. In: Activated Sludge Models ASM1, ASM2, ASM2d and ASM3, Scientific and Technical Report No.9. IWA Publishing, London, UK.

Houghton-Carr, H.A., 1999. Assessment criteria for simple conceptual daily rainfall-runoff models. Hydrol. Sci. J. 44, 237–261. http://dx.doi.org/10.1080/02626669909492220.

Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97.

Lance, G.N., Williams, W.T., 1967. A general theory of classificatory sorting strategies 1. hierarchical systems. Comput. J. 9, 373–380. http://dx.doi.org/10.1093/comjnl/9.4.373.

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50, 885–900.

Muschalla, D., Schneider, S., Gamerith, V., Gruber, G., Schröter, K., 2008. Sewer modelling based on highly distributed calibration data sets and multi-objective auto-calibration schemes. Water Sci. Technol. 57, 1547–1554.

Perrin, C., Andréassian, V., Michel, C., 2006. Simple benchmark models as a basis for model efficiency criteria. Large Rivers 17, 221–244.

Perrin, C., Michel, C., Andréassian, V., 2001. Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. J. Hydrol. 242, 275–301.

Petersen, B., Gernaey, K., Henze, M., Vanrolleghem, P.A., 2002. Evaluation of an ASM1 model calibration procedure on a municipal-industrial wastewater treatment plant. J. Hydroinf. 4, 15–38.

Power, M., 1993. The predictive validation of ecological and environmental models. Ecol. Model. 68, 33–50.

R Core Team, 2012. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ritter, A., Munoz-Carpena, R., 2013. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J. Hydrol. 480, 33–45. http://dx.doi.org/10.1016/j.jhydrol.2012.12.004.

Seibert, J., 2001. On the need for benchmarks in hydrological modelling. Hydrol. Process. 15, 1063–1064.

Sin, G., De Pauw, D.J.W., Weijers, S., Vanrolleghem, P.A., 2008. An efficient approach to automate the manual trial and error calibration of activated sludge models. Biotechnol. Bioeng 100, 516–528. http://dx.doi.org/10.1002/bit.21769.

Van Griensven, A., Bauwens, W., 2003. Multiobjective autocalibration for semi-distributed water quality models. Water Resour. Res. 39, SWC91–SWC99.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., Rowe, C.M., 1985. Statistics for the evaluation and comparison of models. J. Geophys. Res. 90, 8995–9005.

Yapo, P.O., Gupta, H.V., Sorooshian, S., 1998. Multi-objective global optimization for hydrologic models. J. Hydrol. 204, 83–97.