Data evaluation for wastewater treatment plants
Linear vs bilinear mass balances

Le, Q. H.; Carrera, P.; van Loosdrecht, M. C.M.; Volcke, E. I.P.

DOI
10.1016/j.compchemeng.2025.109012

Publication date
2025

Document Version
Final published version

Published in
Computers and Chemical Engineering

Citation (APA)
Le, Q. H., Carrera, P., van Loosdrecht, M. C. M., & Volcke, E. I. P. (2025). Data evaluation for wastewater treatment plants: Linear vs bilinear mass balances. *Computers and Chemical Engineering*, *195*, Article 109012. https://doi.org/10.1016/j.compchemeng.2025.109012

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Data evaluation for wastewater treatment plants: Linear vs bilinear mass balances

Q.H. Le[a], P. Carrera[a], M.C.M. van Loosdrecht[b], E.I.P. Volcke[a,*]

[a] Department of Green Chemistry and Technology, BioCo Research Group, Ghent University, Ghent 9000, Belgium
[b] Department of Biotechnology, Delft University of Technology, Delft 2600 AA, the Netherlands

ARTICLE INFO

Keywords:
Data reconciliation
Gross error detection
Wastewater treatment
Mass balances

ABSTRACT

While nowadays a lot of measurements are conducted at wastewater treatment plants, data reliability could further be improved, e.g., through data reconciliation. This study demonstrated the added value of data reconciliation to improve data quality in a full-scale wastewater treatment plant. Also, the effect of the mass balance setting (linear and bilinear mass balances) was quantitatively evaluated, considering data sets with missing measurements and with gross errors. The improvement in the precision of the key variables was higher with bilinear mass balances (40–80 %) compared to the linear setting (0–70 %). Besides, it delivered a higher number of improved key variables, especially when flow measurements were limited (minimum improved variables of 15 and 0, respectively). Bilinear mass balances were also more efficient in gross error detection and played a crucial role in cross-validation based on flow measurements, resulting in lower incorrectly-identified gross errors. Overall, it is recommended to use bilinear mass balances.

List of symbols and abbreviations

Symbol	Description	Units
x	Reconciled variable	kg.day^{-1}
y	Measured variable	kg.day^{-1}
u	Unmeasured variable	kg.day^{-1}
Q_i	Volumetric flow of the stream i	m^3.day^{-1}
mTP_i	Total phosphorus mass flow of stream i	kg.day^{-1}
$mCOD_i$	Chemical Oxygen Demand mass flow of stream i	kg.day^{-1}
mTN_i	Total nitrogen mass flow of stream i	kg.day^{-1}
$mTKN_i$	Total Kjeldahl nitrogen mass flow of stream i	kg.day^{-1}
DENI	Denitrified nitrogen in the activated sludge reactor	kg.day^{-1}
OC_{COD}	Required oxygen for the oxidation of COD	kg.day^{-1}
NITR	Required oxygen for nitrification	kg.day^{-1}
OC_{net}	Required total oxygen by the activated sludge unit	kg.day^{-1}
i_x	Improvement index of a variable x	%
σ_i^2	Standard error of the mean y	
Abbreviation	**Description**	
WWTP	Wastewater Treatment Plant	
p.e	Population equivalents	
K	Key variable	
M	Measured variable	
U	Unmeasured variable	

1. Introduction

Measurements provide the primary source of information for design, process optimisation, operator training, developing control strategies, benchmarking and simulation. The accuracy and reliability of data sets are, therefore, of great value. Data reconciliation is a proven technique to evaluate the consistency of collected data (Crowe, 1996; Özyurt and Pike, 2004). It involves a procedure of optimally adjusting estimates of the variables such that these estimates satisfy the conservation laws (mass balances) and other constraints (Crowe, 1996) and are therefore more accurate and reliable than the original values. Data reconciliation is often accompanied by statistical tests for gross error detection (measurement validation), which verify whether the deviation between each estimate and its measurement is acceptable compared to the measurement error. Data reconciliation has been applied in the field of (bio) chemical process engineering for decades (Madron et al., 1977; Madron and Veverka, 1992; van der Heijden et al., 1994).

Data quality is also crucial in wastewater treatment plants (WWTPs). Nowadays large amounts of data are generated, but data-rich is often equivalent to information-poor. Thus, data reconciliation is essential to improve plant data reliability. However, even though it is a mature technique, its application to wastewater treatment processes in general

[*] Corresponding author.
E-mail address: Eveline.Volcke@UGent.be (E.I.P. Volcke).

https://doi.org/10.1016/j.compchemeng.2025.109012
Received 16 July 2024; Received in revised form 15 November 2024; Accepted 19 January 2025
Available online 25 January 2025

and to full-scale WWTPs in particular remains limited. Most closely related to wastewater treatment were the applications of data reconciliation for microbiology by Strous et al. (1998) and later by Lotti et al. (2014). In these studies, data reconciliation was applied to long-term operating data sets from a lab-scale reactor to calculate the stoichiometry of carbon and nitrogen conversions by Anammox biomass. Inspired by the work of Barker and Dold (1995) and then Nowak et al. (1999), data reconciliation was performed on full-scale WWTPs by Behnami et al. (2016), Meijer et al. (2015) and Puig et al. (2008). These studies implemented measuring campaigns to collect additional data for various purposes. The collection of additional data was designed in such a way that data reconciliation could be applied to the resulting data sets. The reconciled data (where applicable) were then used for modelling, calculating operational conditions, benchmarking or performance evaluation. Kim et al. (2017) and Yoshida et al. (2015) applied data reconciliation for material flow analysis of certain substances in WWTPs such as total organic carbon, heavy metals and organic pollutants, using the software STAN (Cencic, 2016). The latter studies referred to data reconciliation as a proper approach for detecting error propagation and to obtain a balanced data set, in which all data satisfy the constraints or material balances.

For data reconciliation in industrial applications, the set of constraints which the balanced data need to fulfil could either be linear, bilinear and/or nonlinear mass balances, energy balances or any empirical equations (Câmara et al., 2017). As for the applications to wastewater treatment processes reported in literature, the constraints were usually in the form of linear or bilinear mass balances and the choice seemed to depend on the available tools to solve the data reconciliation problem. Behnami et al. (2016), Meijer et al. (2015), Lotti et al. (2014), Puig et al. (2008) and Strous et al. (1998) used conventional linear mass balances as constraints for data reconciliation. More specifically, Strous et al. (1998), Meijer et al. (2015) and Puig et al. (2008) relied exclusively on the linear data reconciliation method developed by van der Heijden et al. (1994), which was implemented in the software Macrobal (Hellinga and Romein, 1992). Kim et al. (2017) and Yoshida et al. (2015) used the software STAN for data reconciliation, which is a graph-based approach to generate relations between variables. The constraints were designed to be a combination of linear and bilinear mass balances depending on the input information (Cencic, 2016).

This work aimed to demonstrate the added value of applying data reconciliation in wastewater treatment, using a full-scale treatment plant as an illustration. In addition, the effect of the mass balance setup on data reconciliation was evaluated. The results of data reconciliation based on bilinear mass balances were compared to the ones of conventional data reconciliation based on linear mass balances, considering various collected data sets. Three quantitative performance indicators were defined in order to evaluate the reconciliation results and to facilitate the comparison between the bilinear setting and the linear setting, namely (i) precision improvement of key variables, (ii) number of reconciled key variables and (iii) gross error detection.

2. Materials and methods

2.1. Data reconciliation and gross error detection procedure

Various data reconciliation and gross error detection procedures described in literature were integrated into the general flow scheme visualized in Fig. 1. Details of the applied procedures of (1) data reconciliation and (2) gross error detection are provided in Supplementary Material S1.1. The whole evaluation procedure was implemented in MATLAB 2014a, The MathWorks, Inc., Natick, Massachusetts, United States. The consecutive steps are detailed below.

2.1.1. Data reconciliation

The objective function of the data reconciliation problem in this work is defined as the weighted least squares of the distance between the measurements vector and the vector of reconciled values weighted by the measurement error.

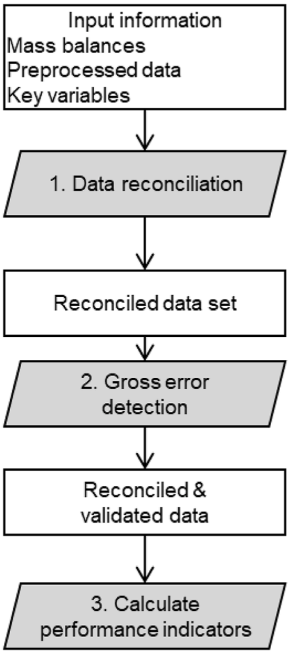

Fig. 1. Data reconciliation and gross error detection procedure proposed in this study.

$$\left.\begin{array}{l} \min \sum \dfrac{(y-x)^2}{\sigma_y^2} \\ \text{subject to } \mathrm{f}(\mathrm{x},\mathrm{u}) = 0 \end{array}\right\} \qquad (1)$$

where y is measured, x is reconciled and u is unmeasured variable. f(x, u) = 0 is a set of equality constraint equations in the form of mass balances. In this study, both overall mass balances and component mass balances are considered. σ_y^2 is a weighing factor, which is usually the standard error of the mean of y. If it is assumed that the measurement errors are normally distributed with zero mean, the solution for this constrained optimization problem gives maximum likelihood estimates of process variables, so they are minimum variance and unbiased estimators (Verheijen, 2010).

The implementation of the data reconciliation procedure using the raw data set and a set of constraints will result in a reconciled data set. This reconciled data set typically contains the following types of variables:

- reconciled measured variables, i.e., measured values of which the value is 'improved' through data reconciliation, The reconciled values are more reliable than the original values in the sense that they satisfy the constraints (mass balances). Besides a different (mean) value, the reconciled variables typically also have a higher accuracy (lower standard deviation or variance).
- reconciled unmeasured variables, i.e., values which were not measured but of which the value could be estimated from reconciled measured variables and from the constraints.
- unreconciled variables, be it measured or unmeasured, i.e., of which the values could not be calculated from (other) measured variables and the constraints

Measured variables for which an improved value can be found and unmeasured variables for which the value can be calculated through data reconciliation, are sometimes referred to as 'observable'. Variables for which no (new) value can be calculated through data reconciliation, are then referred to as ‘unobservable’.

Three types of input information are required for data reconciliation as applied in this work:

(i) a set of constraints in the form of mass balances, which are based on the plant configuration. The input form of mass balances determines the formulation of the objective function and the pre-processing of the input data.
   - In a linear setting, the input mass balances are in the form of linear combinations of variables of total mass flows (or flow rates if flow density is assumed unity for all the streams) and individual mass flows (such as mass flows of total phosphorus, COD or total nitrogen). The objective function is formed to determine the estimates of these variables such that the weighted sum of squares of the adjustments made to the mass flow of components is minimized (see example in Table 1).
   - In a bilinear setting, the input mass balances are in the form of both linear and bilinear combinations of variables. The mass balances of the total mass flows – or flow rates in case of constant density - are linear, while the individual mass flow balances are bilinear, by expressing each individual mass flow as a product of flow and corresponding concentration as separate variables. The objective function is formed to determine the estimates of all flows and concentrations such that the weighted sum of squares of adjustments made to flows and concentrations is minimized (Table 1).

(ii) a pre-processed raw measured data set, which contains mean values and standard errors of the means of the measured variables. The mean values of the total mass flows are calculated from time series data in the same way for both linear and bilinear data reconciliation. The distinction lies in the calculation of the input information of individual mass flows:
   - In a linear setting, the mean value of an individual mass flow is calculated as a product of the corresponding mean value of flow and the mean value of concentration. The standard error of the mean of each individual mass flow variable needs to be calculated following the error propagation rule.
   - In a bilinear setting, the mean value of flow rate and concentrations are treated as separate inputs.

(iii) a list of key process variables, which are the variables one aims to reconcile using data reconciliation. These variables can be either measured or unmeasured ones. In this study, key variables are defined in the form of total mass flow and individual component mass flows for both linear and bilinear data reconciliation, to provide the same basis for comparison.

2.1.2. Gross error detection

Three common tests reported in literature were used in this study (Table 2). The measurement test compares each of the new estimates x with the original measurements y (Eq. (1)). Secondly, one can substitute each of the measurements into the active constraints, f(y). In these two cases, the variances of the respective residuals can be derived easily. Thirdly, the sum-of-squares of the residuals of Eq. (1) has a known distribution and can therefore be used. Details of the gross error detection procedure are presented in Supplementary Material S1.3. The level of significance was set as 0.05 for the three tests.

Table 1
Difference between linear and bilinear mass-balance-based data reconciliation – mass balances and objective function set up for total mass flow and mass flow of total phosphorus as an example. Q = total mass flow with uniform density of all flow is assumed. TP = total phosphorus concentration, mTP is mass flow of total phosphorus. σ = standard error of the mean of the measurement. Unit: kg.d^{-1}.

	Linear	Bilinear
Key variables	Q_{in}, mTP_{in}	Q_{in}, mTP_{in}
Mass balances	$Q_{in} - Q_{out} = 0$ $mTP_{in} - mTP_{out} = 0$	$Q_{in} - Q_{out} = 0$ $Q_{in} \cdot TP_{in} - Q_{out} \cdot TP_{out} = 0$
Objective function	$minimize\left(\sum\frac{(\widehat{Q} - Q)^2}{\sigma_Q^2} + \sum\frac{(\widehat{mTP} - mTP)^2}{\sigma_{mTP}^2}\right)$	$minimize\left(\sum\frac{(\widehat{Q} - Q)^2}{\sigma_Q^2} + \sum\frac{(\widehat{TP} - TP)^2}{\sigma_{TP}^2}\right)$

Table 2
Basic tests in gross error detection (Verheijen, 2010).

Test	Description	Measure
Measurement test	Each individual measurement is considered	y - x
Nodal test	Each individual constraint misfit is considered	f(y)
Global test	Weighted sum of residuals squared gives an overall view	$\sum(y-x)^2$

2.1.3. Performance indicators

In this work, the results of linear and bilinear settings were compared by using three quantitative performance indicators, namely (i) the precision improvement of key variables, (ii) the number of reconciled key variables and (iii) the gross error detection efficiency.

Indicator 1. precision improvement of key variables

The precision improvement of a measured key variable x, noted i_x, is defined as the difference between the variance of the measurement, var(x), and the variance of the reconciled value, var(y), relative to the former value, and expressed as a percentage (Eq. (2)).

In case of an unmeasured key variable x, its precision improvement i_x is defined analogously (Eq. (2)), but this time comparing its reconciled value with its directly calculated value, i.e., calculated from other measured, non-reconciled variables using the available set of mass balances. var(x) was calculated by considering the variance of each measured variable involved in the calculation according to error propagation rules.

$$i_x = \frac{\mathrm{var}(x) - \mathrm{var}(y)}{\mathrm{var}(x)} \times \ 100 \tag{2}$$

The precision improvement i_x is also sometimes referred to as the effect of balancing. The value of i_x is always between 0 and 100 and is typically positive, which means that a better estimate (i.e., characterized by a smaller variance) is found for the key variable. The higher i_x, the larger the improvement of key variables through data reconciliation.

In a linear setting, the variables are total and individual mass flows and the variance of new estimates of (key) variables var(y) can directly be obtained from data reconciliation and will be used to calculate i_x. In a bilinear setting, for key variables in the form of individual mass flows, the variance of their reconciled values is calculated from the variance of the reconciled value of the corresponding flow rate and the reconciled value of the corresponding component concentration, following the general rule of error propagation (Supplementary Material S1.4). Note that, in the bilinear setting, a key variable in the form of a mass flow of a component is considered ‘improved’ if both corresponding flow rate and concentration are improved. Otherwise, it is termed ‘partially improved’.

Indicator 2. number of reconciled key variables

The number of reconciled key variables is used as an indicator in the comparison between the linear and bilinear mass balance approaches. Indeed, the number of reconciled (key) variables does not only depend on the redundancy of the data set but also depends on how variables are related by mass balances (van der Heijden et al., 1994).

Indicator 3. gross error detection efficiency

The gross error detection efficiency was quantified through the number of so-called type I errors in the measurement test. Type I errors refer to error-free variables that are incorrectly identified as containing gross errors.

2.2. WWTP under study

2.2.1. Plant configuration

The system under study is the municipal WWTP Houtrust, The Hague, The Netherlands. The plant has a capacity of 330,000 population equivalents (p.e), treating a daily flow rate of 60.000 $m^3.d^{-1}$. The plant consists of a A^2O process configuration with primary and secondary sludge digestion (Fig. 2). The three stage Phoredox process or A^2O process consists of an activated sludge reactor divided into three compartments with the sequence anaerobic-aerobic-anoxic. This allows for the removal of carbon, nitrogen and phosphorus from the wastewater. The main objectives of the plant were to ensure effective nutrient removal to meet effluent standards while minimizing operational costs, particularly in terms of energy consumption and sludge production.

2.2.2. List of key variables

Key variables were summarized in Table 3. The key process variables were defined to monitor COD, N, P transformations, according to the plant objectives. Additionally, the unmeasured variables were selected as a proxy for the activated sludge reactor performance and energy consumption. Fifteen variables were defined related to total mass flow (Q, $m^3 \cdot day^{-1}$) and individual mass flows (COD, total nitrogen and total phosphorus, $kg \cdot day^{-1}$) of the influent (stream 4), settled influent (stream 7), WWTP effluent (stream 17), waste activated sludge (stream 26), primary sludge (stream 28) and WWTP waste sludge (stream 36). All of the key variables of individual mass flow were considered measured if both flow and corresponding concentration were measured. Three key variables that cannot be measured were defined: required oxygen for the oxidation of COD (OC_{cod}, $kg \cdot day^{-1}$), denitrified nitrogen (DENI, $kg \cdot day^{-1}$) and the oxygen required for nitrification (NITR, $kg \cdot day^{-1}$) of the activated sludge units.

2.2.3. Mass balances

The set of constraints was expressed in the form of linear mass balances on the one hand and bilinear mass balances on the other hand (Table 4), considering the (combined) unit processes from Fig. 2. The subsystems over which mass balances were set up, were selected to ensure that they yield the maximum amount of independent mass balances, which contain all the key variables. The mass balances for the linear and bilinear case are essentially the same, but the bilinear terms of mass flow of components are regarded differently:

Table 3
Key variables defined in this study (15 in total). All of the key variables of individual mass flow were considered measured if both flow and corresponding concentration were measured. (U) indicates unmeasured key variables.

Number in process diagram & stream name		Total mass flow	Total phosphorus	COD	Total nitrogen
4	WWTP influent	Q_4	mTP_4	$mCOD_4$	
7	Settled influent	Q_7	mTP_7	$mCOD_7$	
26	Waste activated sludge (WAS)	Q_{26}	mTP_{26}	$mCOD_{26}$	
28	Primary sludge	Q_{28}*	mTP_{28}*	$mCOD_{28}$*	
36	WWTP waste sludge	Q_{36}	mTP_{36}	$mCOD_{36}$	
	denitrified nitrogen				DENI (U)
	required oxygen for the oxidation of COD			OC_{COD} (U)	
	the total required oxygen			NITR (U)	

- For the linear setting, the constraints are the mass balances in linear form, i.e. in terms of total mass flows and mass flow of components (individual mass flows).
- For the bilinear setting, the constraints are the mass balances in terms of total mass flows (as for the linear case) and the bilinear form of individual mass flows (i.e. for the components), which are the products of the flows and the corresponding concentrations.

The mass balances for total mass flows assumed the same density for all streams and thus reduced to flow rate balances. Individual mass flows were set up for total phosphorus (TP), COD and total nitrogen (TN). The mass balances were set up in a way that they contain all key variables in terms of total mass flow and mass flow of components (in the linear setting) or an equivalent product term of flow and concentration (in the

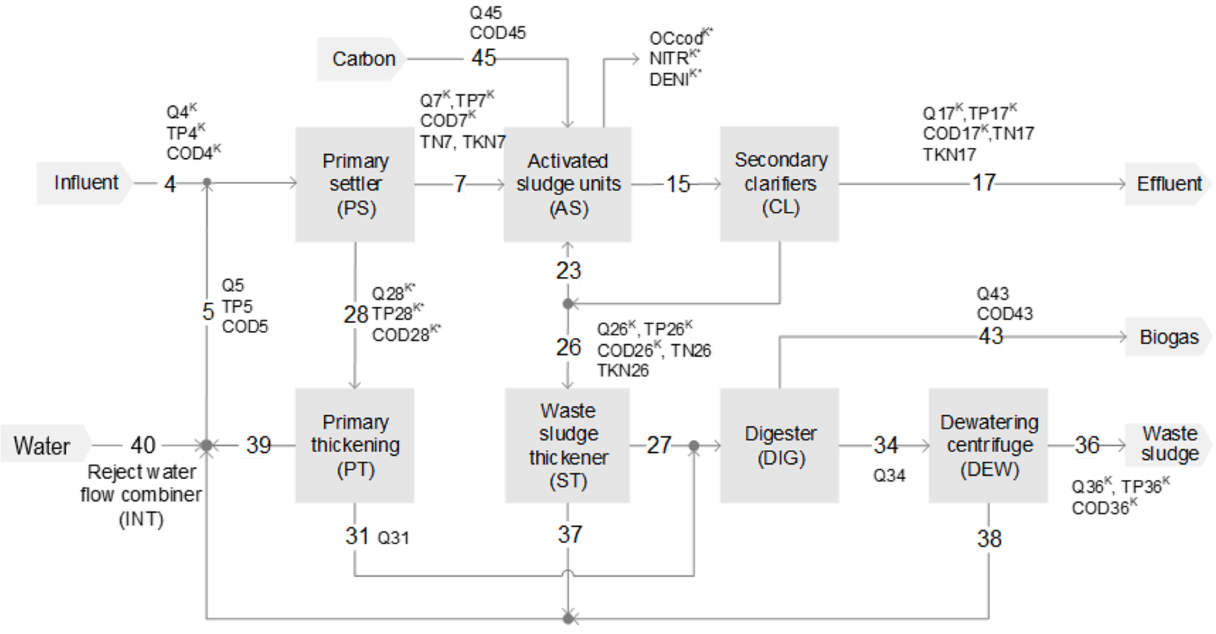

Fig. 2. Simplified process diagram of WWTP Houtrust. K: key variable.

Table 4
Mass balances around individual/combined unit processes, which serve as constraints for data reconciliation. Balances #1–7 are expressed in $m^3\cdot$ day^{-1}, #8–16 in kg$\cdot$ day^{-1}. Variables in bold are key variables.

#	Unit	Bilinear mass balance	Linear mass balance
	PS	$\mathbf{Q_4} + Q_5 - \mathbf{Q_7} - \mathbf{Q_{28}}$	
	AS	$Q_7 - Q_{17} - Q_{26}$	
	ST	$\mathbf{Q_{26}} - Q_{27} - Q_{37}$	
	PT	$\mathbf{Q_{28}} - Q_{31} - Q_{39}$	
	DIG	$Q_{27} + Q_{31} - Q_{34}$	
	DEW	$Q_{34} - \mathbf{Q_{36}} - Q_{38}$	
	INT	$Q_{37} + Q_{38} + Q_{39} + Q_{40} - Q_5$	
	PS	$\mathbf{Q_4}\cdot TP_4 + Q_5\cdot TP_5 - \mathbf{Q_7}\cdot TP_7 - \mathbf{Q_{28}}\cdot TP_{28}$	$\mathbf{mTP_4} + mTP_5 - \mathbf{mTP_7} - \mathbf{mTP_{28}}$
	AS	$\mathbf{Q_7}\cdot TP_7 - Q_{17}\cdot TP_{17} - \mathbf{Q_{26}}\cdot TP_{26}$	$\mathbf{mTP_7} - \mathbf{mTP_{17}} - mTP_{26}$
	ST	$\mathbf{Q_{26}}\cdot TP_{26} + \mathbf{Q_{28}}\cdot TP_{28} - Q_5\cdot TP_5 - \mathbf{Q_{36}}\cdot TP_{36}$	$\mathbf{mTP_{26}} + mTP_{28} - mTP_5 - mTP_{36}$
	PS	$\mathbf{Q_4}\cdot COD_4 + Q_5\cdot COD_5 - \mathbf{Q_7}\cdot COD_7 - \mathbf{Q_{28}}\cdot COD_{28}$	$\mathbf{mCOD_4} + mCOD_5 - \mathbf{mCOD_7} - \mathbf{mCOD_{28}}$
	AS	$\mathbf{Q_7}\cdot COD_7 + Q_{45}\cdot COD_{45} - Q_{17}\cdot COD_{17} - \mathbf{Q_{26}}\cdot COD_{26} - \mathbf{OC_{cod}} - 2.86\cdot\mathbf{DENI}$	$\mathbf{mCOD_7} + mCOD_{45} - \mathbf{mCOD_{17}} - \mathbf{mCOD_{26}} - \mathbf{OC_{cod}} - 2.86\cdot\mathbf{DENI}$
	ST	$\mathbf{Q_{26}}\cdot COD_{26} + \mathbf{Q_{28}}\cdot COD_{28} - Q_5\cdot COD_5 - \mathbf{Q_{36}}\cdot COD_{36} - Q_{43}\cdot COD_{43}$	$\mathbf{mCOD_{26}} + \mathbf{mCOD_{28}} - mCOD_5 - \mathbf{mCOD_{36}} - mCOD_{43}$
	AS	$\mathbf{Q_7}\cdot TN_7 - \mathbf{Q_{17}}\cdot TN_{17} - \mathbf{Q_{26}}\cdot TN_{26} - \mathbf{DENI}$	$mTN_7 - mTN_{17} - mTN_{26} - \mathbf{DENI}$
	AS	$\mathbf{Q_7}\cdot TKN_7 - \mathbf{Q_{17}}\cdot TKN_{17} - \mathbf{Q_{26}}\cdot TKN_{26} - \mathbf{NITR}$	$mTKN_7 - mTKN_{17} - mTKN_{26} - \mathbf{NITR}$
	WWTP	$OC_{net} - \mathbf{OC_{cod}} - 4.57\cdot\mathbf{NITR}$	

OCnet = required total oxygen by activated sludge unit (kg$\cdot$$day^{-1}$); OCcod = required oxygen for COD removal (kg$\cdot$ day^{-1}).
NITR = nitrified nitrogen (kg$\cdot$ day^{-1}); DENI = denitrified nitrogen (kg$\cdot$ day^{-1}).

bilinear setting). The external carbon source (stream 45) and the biogas (stream 43) were reasonably assumed to represent only COD; their total mass flow rates were neglected (in mass balances over #AS and #DIG, respectively).

Note that in this study, mass balances were set only with conservative variables, and referred to only one type of individual component. Nevertheless, the mass balances could also include kinetic relations, equality constraints or electron balances. For instance, kinetic relations between variables based on biological conversions could further improve the reconciliation results (Le et al., 2022). More detailed practical guidance on the selection of conservative quantities is provided by Meijer et al. (2015).

A key variable in terms of mass flow of component e.g. the mass flow of COD of primary sludge, $mCOD_{28}$, was considered improved in the linear setting if its new estimate was found by data reconciliation using linear mass balances as the constraints. This definition was applied to all the observable variables, i.e., both measured variables and unmeasured variables. In a bilinear setting, this key variable will only be considered improved if improved estimates of both the flow (Q_{28}) and the concentration (COD_{28}) are found by data reconciliation using the bilinear form of mass balances as the constraints.

2.2.4. Pre-processed data set

The pre-processed data set contains the mean values and standard errors of the mean, which are derived from measurements during a stable operation period of WWTP Houtrust. The time series data includes 50 daily measurements of flow rates and corresponding concentrations (24 h composite or grab samples) of total phosphorus (TP), COD, total nitrogen (TN) and total Kjeldahl nitrogen (TKN) under dry weather conditions (Meijer et al., 2015).

- For bilinear data reconciliation, data of flow rates (Q) and concentrations (TP, COD, TN, TKN) were used as the raw input data.
- For linear data reconciliation, data of flow rates (Q) or mass flow of components were used (mTP, mCOD, mTN, mTKN). The mass flow of components was derived as the product of the flow and the corresponding concentrations. Their error terms were calculated following the general rule of error propagation (Supplementary Material S1.4), assuming that there was no correlation between the error of flow measurements and the error of concentration measurements.

Several types of data sets were used in this study, namely the reference data set, reduced data sets and erroneous data sets (detailed in Supplementary Material S2).

- **Reference data set.** The original dataset consisted of 30 measurements, including total flows (10), TP (6), COD (8), TN (3) and TKN (3) (Supplementary Material, section S2.1). It served as a reference data set to compare the precision improvement of the key variables between the linear and bilinear settings (**indicator 1**). It was also used to derive reduced or erroneous data sets in this study.
- **Reduced data set.** Reduced data (Supplementary Material, section S2.2) were used to compare the number of reconciled key variables in dealing with missing data (**indicator 2**). 57 reduced data sets were formed by removing one or more measured variables from the reference data set. The reduced data sets were obtained from the reference data set by removing (A) all measured data (both flow rates and individual component concentrations); (B) only flow measurements; (C) only measurements of component concentrations, for one or more streams.
- **Erroneous data sets.** Erroneous data sets were used to compare the efficiency of linear and bilinear settings in detecting gross errors (**indicator 3**). Thirty erroneous data sets were formed by introducing an error of +20 %, +35 % and +50 % to the following measurements (one at a time): influent flow (Q_4), flow of WWTP waste sludge (Q_{36}), effluent flow (Q_{17}), flow of waste activated sludge (Q_{26}), total phosphorus of influent (TP_4), total phosphorus of settled influent (TP_7), total phosphorus of waste activated sludge (TP_{26}), COD of settled influent (COD_7), COD of waste activated sludge (COD_{26}) and COD of WWTP waste sludge (COD_{36}).

3. Results and discussion

3.1. Precision improvement of the key variables

The data reconciliation procedure was applied to the reference data set using linear and bilinear mass balances as constraints. For both the linear and bilinear setting, all 18 key variables were reconciled, meaning that more accurate estimates for their values were found by the data reconciliation procedure. Detailed results are provided in Supplementary Material S3. In both settings, the gross error detection procedure did not indicate any gross error in the reference data set.

The bilinear setting resulted in a more substantial improvement i (Eq. (2)) of key variables (40–80 %) overall compared to the linear setting (0–70 %) (Fig. 3A). For example, the flow of waste activated sludge (Q_{36}) is a measured key variable with a measured value of Q_{36} = 49 ± 5 $m^3\cdot day^{-1}$. The newly estimated value for this variable through bilinear data reconciliation was 51 ± 2 $m^3\cdot day^{-1}$ (i_{36} = 78 %) while linear data reconciliation resulted in the same value as the raw value ($i_{36}\approx$ 0 %). The flow of primary sludge (Q_{28}) was improved the most. Q_{28} is an unmeasured key variable of which the value was directly calculated value from raw data as 5194 ± 3326 $m^3\cdot day^{-1}$. The improved reconciled values of Q_{28} in linear and bilinear settings were 5051 ± 141 $m^3\cdot day^{-1}$ and 5067 ± 137 $m^3\cdot day^{-1}$, respectively. This corresponded with an improvement of i_{28} ≈100 % with both settings.

The larger improvement through bilinear data reconciliation compared to the linear setting was also expressed per stream (10–65 %, 40–70 %, respectively, Fig. 3B). For instance, the average improvement of key variables in the activated sludge units (AS) including required oxygen for the oxidation of COD (OC_{cod}, kg$\cdot day^{-1}$), the mass flow of denitrified nitrogen (DENI, kg$\cdot day^{-1}$) and the oxygen required for

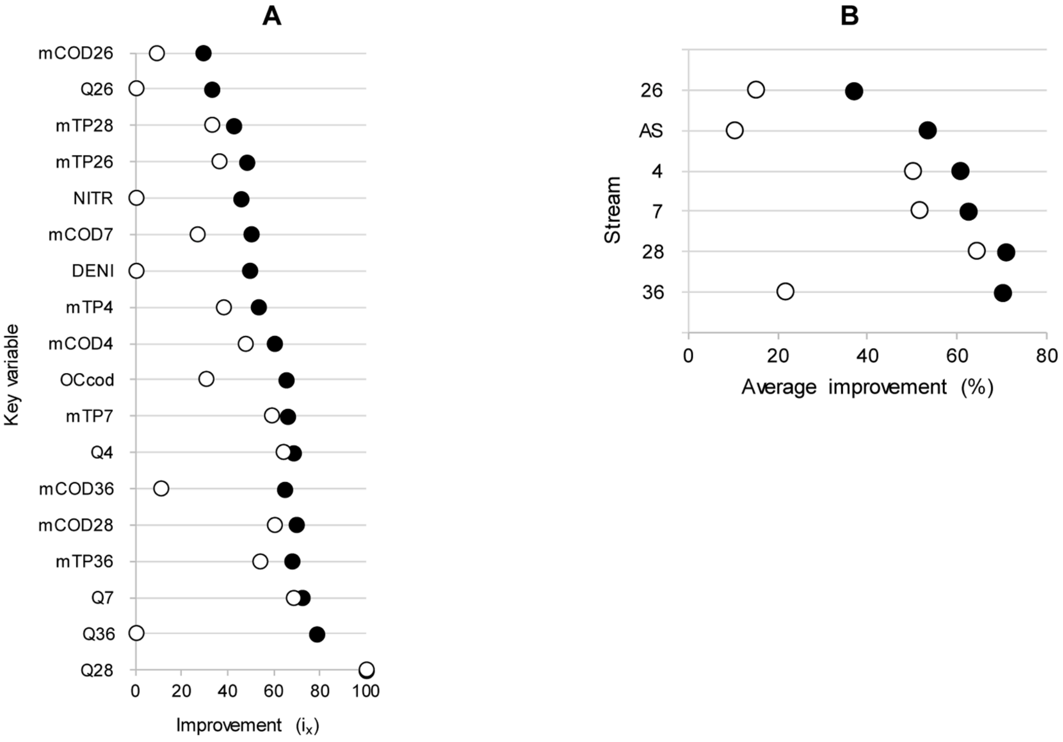

Fig. 3. Improvement of individual key variables (A) and average improvement of key variables in important streams (B) after the reconciliation procedure using linear (empty cycle) and bilinear setting (filled circle).

nitrification (NITR, kg·day^{-1}) were much better with the bilinear setting (i_{AS} = 53 %) compared to that with the linear setting (i_{AS} = 10 %). Note that, since these key variables in AS were unmeasured key variables, their improvements were calculated with regard to their directly calculated values as mentioned in Section 2.1.3.

The bilinear setting resulted in a larger improvement of the key variables because the new estimates of the key variables were derived from more measurements than in the linear setting. The more measured variables a variable can be calculated from (is constrained by), the more its value can be improved through data reconciliation (van der Heijden et al., 1994). Hence, in the bilinear setting, $mCOD_4$ was improved by i_{36} = 60 % in the linear setting this was only 48 %.

Only in the bilinear setting, variables of one type (such as $mCOD_4$ mentioned above) can be improved by multiple types of measurements (e.g. TP, COD, TN and Q). When considering linear mass balances, each type of measurement can only help to improve variables of that type. This principle holds only for conserved components (i.e., those not involved in any conversion process). For instance, previous studies with the linear setting (Behnami et al., 2016; Meijer et al., 2002; Puig et al., 2008;Meijer et al., 2015) demonstrated that flow rate measurements only contributed to improvements of flow rates, while contributions of the conserved total phosphorus measurements were limited to the improvement of the phosphorus mass flow. In bilinear setting, total phosphorus measurements could help to not only total phosphorus value but also others since they can be related to others by flow measurements. Since not all measured data were effectively utilised in the linear setting, a lower improvement of key variables was obtained.

3.2. *Effect of limited data on key variable identifiability*

For the reference case, all 18 key variables were reconciled for both the linear and bilinear settings. For the reduced data sets, the number of identified key variables is expected to be lower, since also the redundancy of the available data will likely be lower.

In case both flow and concentration were removed from the reference data set i.e. the reduced data sets A, both linear and bilinear settings had the same results (Fig. 4A). The number of improved key variables decreased with the reduction of the dataset, from 18 (ref. dataset) to 0 (dataset #1) improved key variables. In this case, there were the same losses in the redundancy for both the linear and bilinear settings. For example, when all flows and corresponding concentrations of stream 4 and 5 were removed from the reference data set to form the reduced data set #14A (Supplementary Material S2), the linear setting lost 6 measurements of the variables Q_4, $mCOD_4$, mTP_4, Q_5, $mCOD_5$ and mTP_5 and the bilinear lost the same number of measurements Q_4, COD_4, TP_4, Q_5, COD_5 and TP_5. Therefore, both settings were able to identify ten key variables in data set #14A.

In case one or more flows were removed from the reference data set, i.e. for the reduced data sets B, the bilinear setting resulted in a much higher number of improved key variables compared to the linear setting (minimum number of improved key variables of 15 and 0, respectively, Fig. 4B). For example, in the reduced data set #9B (Supplementary Material S2), the flows of settled influent (Q_7) and effluent (Q_{17}) were removed but their corresponding concentrations of TP, COD, TN and TKN remained. For the linear setting, ten variables Q_7, Q_{17}, mTP_7, mTP_{17}, $mCOD_7$, $mCOD_{17}$, mTN_7, mTN_{17}, $mTKN_7$ and $mTKN_{17}$ became unknown while in the bilinear setting only Q_7 and Q_{17} became unknown. Therefore, the bilinear setting offered much better results with all 18 reconciled key variables in contrast to seven reconciled key variables with the linear setting.

In case only the concentrations of one or more streams were removed from the reference data set, i.e. for the reduced data sets C, again, the linear and bilinear settings had the same results, showing a decrease from 18 to 5 improved key variables with the reduction of the dataset (Fig. 4C). The loss of redundancy in the data sets, in this case, was the same for both the linear and bilinear settings. For example, when the measurement of COD_7 was removed from the reference data set to form the reduced data set #13C (Supplementary Material S2), in the linear setting, $mCOD_7$ became an unknown variable and in the bilinear setting, COD_7 became an unknown variable. As a result, both settings had the

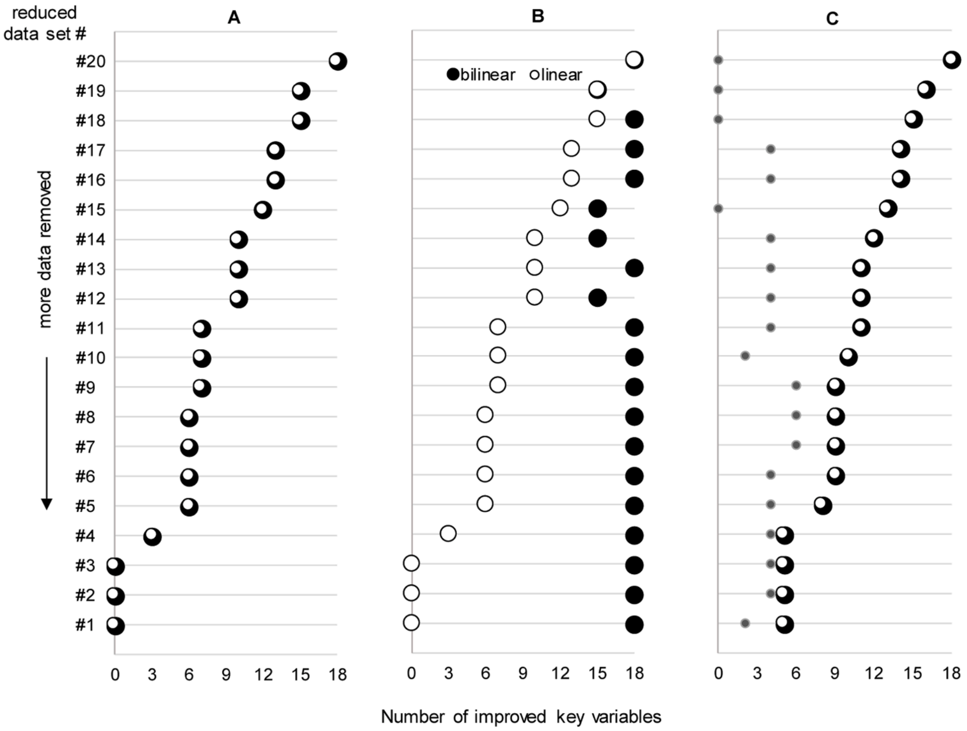

Fig. 4. Number of identifiable key variables when data reconciliation with linear (empty circles) and bilinear (filled circles) mass balances were applied to 60 reduced data sets of three types: (A) all measured data (both flow and corresponding component's concentrations) of one or more streams are removed; (B) only flow measurements of one or more streams were removed; (C) only measurements of component's concentrations were removed. The empty small dots are the number of "partially improved" key variables.

same performance in this case, with 11 reconciled key variables. In addition to the reconciled key variables, the bilinear setting also resulted in partially reconciled key variables as the mass flow of a component was calculated from a reconciled flow and an unreconciled concentration. For instance, four partially reconciled key variables were found in the reduced data set #13C with the bilinear setting including $mCOD_4$, $mCOD_{26}$, $mCOD_{36}$ and mTP_{26} because Q_4, Q_{26} and Q_{36} were reconciled but not the corresponding concentrations.

Overall, when measured data are limited, the bilinear setting will always lead to at least the same but in most cases higher number of key variables that can be reconciled, since it has at least an equal and mostly higher redundancy compared to the linear setting. With a higher redundancy, more variables can be expressed by other measured variables using the set of mass balances and this usually leads to a higher number of reconciled key variables.

The possibility to have a higher number of improved key variables with the bilinear setting is beneficial when there is a gross error in the data set. In such a case, the measurement of one or several suspected variables must be removed from the data set. The removal or missing measurement of a flow variable results in one unknown variable in the bilinear setting but at least two unknown variables in the linear setting as the flow becomes an unmeasured variable, and so does the mass flow of all components for the corresponding stream.

3.3. Gross error detection

The gross error detection results in case of one-at-a-time gross errors of +50 % are summarized in Fig. 5. More detailed results on gross error detection are summarised in Supplementary Material S3.3. A gross error is identified if at least one of the three tests (global, measurement or nodal test) indicates a gross error. The measurement test is used to report the suspected measurements.

Both linear and bilinear setting could detect gross errors in 8/11 and 10/11 cases, respectively. Both settings failed to detect a gross error in the effluent COD measurement (COD_{17}). The reason was that this variable was not constrained by other measured variables and its value could thus not be improved by data reconciliation. The linear setting, however, also failed to detect a gross error in other reconciled key variables of the waste activated sludge (COD_{26}) and the waste sludge (COD_{36}).

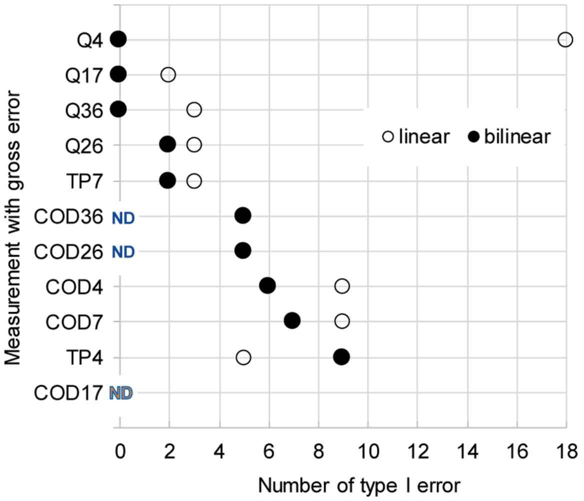

Fig. 5. Number of variables incorrectly identified to have a gross error (type I errors) in linear (empty circles) and bilinear (filled circles) settings when a genuine gross error of +50 % occurs in a single variable (flow or concentration measurement). The actual gross error was identified in all cases except the ones denoted by ND = no gross error detected.

It is important to note that the improvement of a measured variable is directly correlated to the possibility to detect gross errors in that variable (Narasimhan and Jordache, 2000). The higher the improvement of (key) variables, the more chance gross errors could be detected. For example, when a gross error occurred in COD_{26} or COD_{36}, the bilinear setting could pick up the gross error in these variables (Fig. 5). However, since the linear setting had a rather low improvement of $mCOD_{26}$ and $mCOD_{36}$ (about ≈10 % compared to 30–60 % of the bilinear setting), the linear setting could not detect any gross error in these cases.

Gross error detection using the bilinear setting mostly resulted in less type I errors, i.e. error-free variables that are incorrectly identified as containing gross errors, than using the linear setting (Fig. 5). For example, in case of an erroneous waste activated sludge flow rate (Q_{26}), the linear setting indicated two variables as suspected: the mass flow of total phosphorus in the effluent (mTP_7) and in the wasted activated sludge (mTP_{26}). This means that the gross error could originate from either one of the four variables Q_7, TP_7, Q_{26} or TP_{26}. In the same case, the bilinear setting indicated that three variables Q_{26}, TP_7 and TP_{26} were suspected. Since only Q26 contains a gross error, the number of type I errors of linear and bilinear settings were reported in Fig. 5 as three and two, respectively. In short, linear setting usually leads to a higher number of type I errors since it can distinguish between gross errors in flow or concentration. Similar results were also observed when gross errors of 35 and 20 % were introduced in the datasets (Supplementary Material S3.3).

When there is a gross error in the data set, the result of data reconciliation is affected by the smear effect (Narasimhan and Jordache, 2000), in which one erroneous variable would trigger the detection of gross error in many other measured variables, and therefore still produces a high number of type I errors. For this reason, process insights and expert knowledge are essential in isolating the found errors. Some common sources of measurement errors in WWTP provided by (Rieger et al., 2010) could be a good starting point for gross error elimination.

It is important to realize that not all measured variables will be checked by gross error detection but only the measured variables which are reconciled, i.e. the ones of which the measured value has been double-checked against mass balances. Gross errors in unreconciled measured variables cannot be detected, but will likely propagate into reconciled values (new estimates) of unmeasured variables. It is advised to double-check the measurements and carefully calibrate sensors to guarantee the error-free unreconciled measured variables (Heyen et al., 1996). Further, some reconciled variables may show a very low improvement after data reconciliation. It is likely that gross error detection also fails to detect an error in these variables. These variables with low improvement are usually referred to as practically unidentifiable variables in literature.

Using the bilinear setting for data reconciliation would be more useful if there are one or more variables subjected to gross error in a data set. The benefit would be two-fold. First, since the bilinear setting produces higher improvement, it has a higher chance to detect the gross error and to pinpoint the exact variable that contains an error. Second, more importantly, as the erroneous measurements were excluded from the data set, the bilinear approach, in many cases, was still able to maintain the same number of reconciled key variables as illustrated in Section 3.2. It means that when the redundancy of the data set was reduced by gross error elimination, it is highly likely that all the key variables will still be validated and reconciled using the bilinear mass balances.

4. Conclusions

Reliable data is crucial in many industries, including wastewater treatment. This study aimed to demonstrate the added value of applying data reconciliation to improve data quality in a full-scale wastewater treatment plant. In addition, the effect of mass balance setting on the result of data reconciliation has been evaluated. The bilinear mass balances hold the following advantages over linear mass balances:

- Data reconciliation improved data reliability with both linear and bilinear mass balances.
- Data reconciliation with bilinear mass balances resulted in a higher precision improvement of key variables of 40–80 % compared to the linear setting (0–70 %). Since a higher precision implies a higher chance of detecting errors in these variables, gross error detection and isolation were also more efficient with bilinear mass balances.
- Besides, bilinear mass balances delivered a higher number of improved key variables in comparison to linear mass balances, in particular when flow measurements were limited (minimum improved key variables of 0 and 15 for the linear and bilinear setting, respectively) and/or several measurements had to be removed due to gross errors.
- The bilinear setting played a crucial role in cross-validation, based on flow measurements. With a bilinear setting, a distinction could be made between gross errors in flow rates and in concentrations, resulting in less incorrectly detected gross errors. Flow measurements played a vital role because they were not only involved in the improvement of flow variables but also in that of all other types of variables.
- Data reconciliation was demonstrated for a wastewater treatment plant, but its application could be generalized for other processes with measurements of flows and concentrations.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Q.H. Le: Writing – review & editing, Writing – original draft, Visualization, Investigation, Formal analysis, Data curation, Conceptualization. **P. Carrera:** Writing – review & editing, Visualization, Formal analysis. **M.C.M. van Loosdrecht:** Writing – review & editing, Validation, Supervision. **E.I.P. Volcke:** Writing – review & editing, Validation, Supervision, Resources, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2025.109012.

Data availability

All the relevant data is included in Supplementary Information.

References

Barker, P.S., Dold, P.L., 1995. COD and nitrogen mass balances in activated sludge systems. Water Res. 29, 633–643. https://doi.org/10.1016/0043-1354(94)00155-Z.

Behnami, A., Shakerkhatibi, M., Dehghanzadeh, R., Benis, K.Z., Derafshi, S., Fatehifar, E., 2016. The implementation of data reconciliation for evaluating a full-scale petrochemical wastewater treatment plant. Environ. Sci. Pollut. Res. 23, 22586–22595. https://doi.org/10.1007/s11356-016-7484-5.

Câmara, M.M., Soares, R.M., Feital, T., Anzai, T.K., Diehl, F.C., Thompson, P.H., Pinto, J. C., 2017. Numerical aspects of data reconciliation in industrial applications. Processes 5, 56. https://doi.org/10.3390/pr5040056.

Cencic, O., 2016. Nonlinear data reconciliation in material flow analysis with software STAN. Sustain. Environ. Res. 26, 291–298. https://doi.org/10.1016/j.serj.2016.06.002.

Crowe, C.M., 1996. Data reconciliation–progress and challenges. J. Process Control 6, 89–98. https://doi.org/10.1016/0959-1524(96)00012-1.

Hellinga, C., Romein, B., 1992. Macrobal–a program for robust data reconciliation and gross error detection. Model. Control Biotech. Process. 1992, 459–460. https://doi.org/10.1016/S1474-6670(17)50415-2, 1992.

Heyen, G., Maréchal, E., Kalitventzeff, B., 1996. Sensitivity calculations and variance analysis in plant measurement reconciliation. Comput. Chem. Eng. 20, S539–S544. https://doi.org/10.1016/0098-1354(96)00099-3.

Kim, H., Jang, Y.C., Hong, Y.S., 2017. Substance flow analysis of mercury from industrial and municipal wastewater treatment facilities. Int. J. Appl. Eng. Res. 12, 5332–5338.

Le, Q.H., Verheijen, P.J.T., van Loosdrecht, M.C.M., Volcke, E.I.P., 2022. Application of data reconciliation to a dynamically operated wastewater treatment process with off-gas measurements. Environ. Sci. https://doi.org/10.1039/d2ew00006g.

Lotti, T., Kleerebezem, R., Lubello, C., van Loosdrecht, M.C.M., 2014. Physiological and kinetic characterization of a suspended cell anammox culture. Water Res. 60, 1–14. https://doi.org/10.1016/j.watres.2014.04.017.

Madron, F., Veverka, V., 1992. Optimal selection of measuring points in complex plants by linear models. AIChE J. 38, 227–236. https://doi.org/10.1002/aic.690380208.

Madron, F., Veverka, V., Vanecek, V., 1977. Statistical-analysis of material balance of a chemical reactor. AIChE J. 23, 482–486. https://doi.org/10.1002/aic.690230412.

Meijer, S.C.F., Van Der Spoel, H., Susanti, S., Heijnen, J.J., Van Loosdrecht, M.C.M., 2002. Error diagnostics and data reconciliation for activated sludge modelling using mass balances. Water Sci. Technol. 45, 145–156. https://doi.org/10.2166/wst.2002.0102.

Meijer, S.C.F., van Kempen, R.N.A., Appeldoorn, K.J., 2015. Plant upgrade using big-data and reconciliation techniques. Applications of Activated Sludge Models. IWA Publishing, pp. 357–410.

Narasimhan, S., Jordache, C., 2000. Data reconciliation and gross error detection : an intelligent use of process data. Annals of the New York Academy of Sciences. Gulf Publishing Company, Houston, Texas, US.

Nowak, O., Franz, A., Svardal, K., Muller, V., Kuhn, V., 1999. Parameter estimation for activated sludge models with the help of mass balances. Water Sci. Technol. 39, 113–120. https://doi.org/10.1016/S0273-1223(99)00065-7.

Özyurt, D.B., Pike, R.W., 2004. Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes. Comput. Chem. Eng. 28, 381–402. https://doi.org/10.1016/j.compchemeng.2003.07.001.

Puig, S., van Loosdrecht, M.C.M., Colprim, J., Meijer, S.C.F., 2008. Data evaluation of full-scale wastewater treatment plants by mass balance. Water Res. 42, 4645–4655. https://doi.org/10.1016/j.watres.2008.08.009.

Rieger, L., Takacs, I., Villez, K., Siegrist, H., Lessard, P., Vanrolleghem, P.A., Comeau, Y., 2010. Data reconciliation for wastewater treatment plant simulation studies-planning for high-quality data and typical sources of errors. Water Environ. Res. 82, 426–433. https://doi.org/10.2175/106143009X12529484815511.

Strous, M., Heijnen, J., Kuenen, J., Jetten, M., 1998. The sequencing batch reactor as a powerful tool for the study of slowly growing anaerobic ammonium-oxidizing microorganisms. Appl. Microbiol. Biotechnol. https://doi.org/10.1007/s002530051340.

van der Heijden, R.T.J.M., Heijnen, J.J., Hellinga, C., Romein, B., Luyben, K.C.A.M., 1994. Linear constraint relations in biochemical reaction systems: I. Classification of the calculability and the balanceability of conversion rates. Biotechnol. Bioeng. 43, 3–10. https://doi.org/10.1002/bit.260430103.

Verheijen, P.J.T., 2010. Data reconciliation and error detection. The Metabolic Pathway Engineering Handbook : Fundamentals. CRC Press/Taylor & Francis, Boca Raton.

Yoshida, H., Christensen, T.H., Guildal, T., Scheutz, C., 2015. A comprehensive substance flow analysis of a municipal wastewater and sludge treatment plant. Chemosphere 138, 874–882. https://doi.org/10.1016/j.chemosphere.2013.09.045.