Environmental Science
Water Research & Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d4ew00315b

Effective measuring campaigns for reliable and informative full-scale WWTP data†

Q. H. Le,[ab] P. Carrera,[ab] M. C. M. van Loosdrecht[c] and E. I. P. Volcke *[ab]

Sensor availability and costs are nowadays no longer limiting data gathering at wastewater treatment plants (WWTPs). However, one should be aware that a higher amount of measured data gathered does not necessarily imply that also more information is obtained. In this light, this contribution assesses the general applicability and the added value of a structured experimental design approach for planning measurement campaigns at WWTPs, in view of mass-balance-based data reconciliation. To this end, the results from full-scale WWTP case studies available in the literature were compared to those obtained with the developed structured experimental design procedure. Planning measurement campaigns comprises the selection of (additional) measurements to meet a pre-set main goal. The need for a structured experimental design procedure replacing past expert judgment approaches became clear from the fact that three out of five case studies available in the literature failed to meet the main goal and/or performed unnecessary additional measurements. Translating the main goal into specific key variables was found essential in this respect. The general applicability of the procedure was proven with three outcomes. First, the procedure, involving well-defined steps, could be applied to different WWTP layouts. Second, it ensured the fulfilment of various main goals. Third, it provided useful outcomes, *i.e.*, optimal measurement campaigns, which reduced the need for additional measurements (40–70% less) compared to expert knowledge approaches, hence more information could be obtained with less analytical data. Overall, the experimental design procedure proved a fast and useful tool ensuring the success of subsequent mass-balance-based data reconciliation.

Received 18th April 2024,
Accepted 23rd December 2024

DOI: 10.1039/d4ew00315b

rsc.li/es-water

Water impact

Nowadays, large amounts of data are generated in wastewater treatment plants (WWTPs) but data-rich does not always mean information-rich. In this light, this contribution assesses the general applicability and the added value of a structured experimental design approach for planning measurement campaigns at WWTPs in view of mass-balance-based data reconciliation for reliable plant data gathering.

1 Introduction

Measurements provide the main source of information in view of wastewater treatment plant (WWTP) design and optimization, process evaluation, operator training, modelling and simulation, and benchmarking simulation. To verify and improve the quality of collected data, data reconciliation has become a proven technique.[1,2] Data reconciliation is directed towards finding better estimates of key process variables, which may be either measured variables or unmeasured variables that are calculated from measured variables. Key variables are process variables, the reconciliation (= identification) of which ensures the fulfilment of the main goal of the study, which is typically stated by the plant operator. The reconciled values are more reliable than the original values in the sense that they satisfy conservation laws (*e.g.* mass balances) and other constraints. Besides a different (mean) value, the reconciled variables typically also have a higher accuracy, *i.e.*, a lower standard deviation. While data reconciliation has been widely applied in the field of (bio) chemical process engineering,[3–5] this concept has received relatively little attention so far in wastewater treatment process engineering (Table A1, Appendix A, ESI†). Nevertheless, published studies so far clearly show the added value of data reconciliation for different purposes such as data validation for modelling and plant assessment,[6–8] variable classification and redundancy analysis for sensor placement,[9,10] or mass flow analysis of different process variables.[11,12]

In order to guarantee that key variables can be identified through data reconciliation, it is vital that the available measurements satisfy redundancy and steady-state

[a] Department of Green Chemistry and Technology, Ghent University, 9000 Ghent, Belgium. E-mail: Eveline.Volcke@UGent.be; Tel: +32 (0) 9 264 61 29
[b] Centre for Advanced Process Technology for Urban REsource recovery (CAPTURE), Frieda Saeysstraat 1, 9052 Gent, Belgium
[c] Department of Biotechnology, Delft University of Technology, 2600 AA Delft, The Netherlands

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4ew00315b

conditions. The data redundancy requirement means that one or more variables in the data set can be calculated from other (measured) variables using the available set of constraints (including mass balances), and are therefore identifiable,[5] in the sense that they can be reconciled (identified). In case there are no or not sufficient initially measured data available, additional measurements need to be carried out to ensure the required degree of redundancy and thus the possible identification of key variables. Experimental design involves the determination of sets of (additional) measurements to fulfil this goal. This concept has been proven useful for optimal sensor placement for different applications, including chemical processes[13,14] or water networks.[15,16] Nevertheless, experimental design formulated as a structured, optimization problem has still limited attention in wastewater treatment processes.[17,18] Instead, in available case studies from practice (Table A1, ESI†), providing sufficient redundancy has been interpreted quite intuitively, by ensuring that the number of constraints (independent mass balances) was higher than the number of unknown variables, *i.e.* aiming at an overdetermined system. In this way, redundancy was considered a 'global property' of the system while in reality, it is a property of individual variables.[5] It is therefore unclear whether the experimental design approaches followed in the case studies previously reported in the literature guarantee the identifiability of all specified key process variables.[17]

In order to overcome the shortcomings of previous studies, Le *et al.*[17] presented a more formal, structured experimental design procedure, including a comprehensive redundancy analysis to unambiguously check the identifiability of all key variables. The search for optimal sets of additional measurements is solved as a multi-objective optimisation problem minimising the cost of additional measurements and maximising the accuracy of the improved estimates of key variables. The results are visualized in a Pareto-optimal front, which represents the optimal solutions (= sets of measurements) taking into account the trade-off between their cost and accuracy. This is a valuable outcome for measurement planning, as it allows for compliance with the main goal with an optimal use of resources. However, so far the results obtained with the experimental design procedure of Le *et al.*[17] have not yet been compared with those obtained in previously published studies.

In this work, the added value and general applicability of the structured experimental design procedure of Le *et al.*[17] in view of mass-balance-based data reconciliation were scrutinized by comparing them with previous expert judgment approaches for WWTP measurement campaign planning. In particular, the procedure was assessed in terms of applicability to different layouts and main goals, redundancy and identifiability of key variables, relevance of the mass balances and the number and type of additional measurements. To this end, the experimental design procedure of Le *et al.*[17] was applied to five full-scale WWTP case studies available in the literature dealing with experimental design in view of mass-balance-based data reconciliation.[19–22] Going beyond the mere detection of mistakes from the past, this work demonstrates why a rigorous experimental approach is needed for future measurement campaigns and how this can be performed.

2 Materials and methods

The experimental design procedure of Le *et al.*[17] (Appendix B, ESI†) was applied to five available case studies from the literature dealing with experimental design in view of mass-balance-based data reconciliation (Table 1) to evaluate the effectiveness of the proposed measurement campaigns in fulfilling their main goal.

First of all, experimental design was conducted independently of what was proposed in the previous studies. This involves the translation of the main goal of the measurement campaign into key variables and the determination of optimal sets of additionally measured variables (besides initially available ones) that guarantee the identifiability of these key variables. Typical examples of key variables concern influent and effluent mass flow rates (*e.g.*, total phosphorus, nitrogen) of the activated sludge process or the waste sludge mass flow rate. The oxygen requirements for carbon and nitrogen removal are usually important as well.[17] These would be appropriate key variables if one wants to get reliable data for monitoring plant performance or perform model simulations. The sets of additional measurements obtained by solving the multi-objective optimization problem are also referred to as (optimal) solutions and belong to the Pareto optimal front.

The application of the experimental design procedure required three main types of input information: (a) main goal and associated key variables, (b) mass balances and (c) initially measured data set and potential additionally measured variables, with estimated cost and variance of the measured variables. These inputs were obtained from the five case studies.

a. **Key variables**. The main goal of each case study was translated into key variables, the identification of which ensured that this goal was fulfilled. Key variables can be initially measured or not; their identification means that their value can be calculated from other measured variables through which they are related by mass balances. As a result, the value of this variable can be reconciled by using mass balances. This implies that key variables need to be conservative quantities, *i.e.*, fulfil material conservation laws (mass balances). For the case studies from the literature in which the key variables were not specified explicitly, they were deduced in this study from the given main goal.

b. **Mass balances**. The mass balances used in this work correspond with the incidence matrices from previous studies. They were represented for each case study in equation form. It can be noted that all the studies used steady-state mass balances and calculated average operating conditions for data reconciliation. For dynamic processes,

Table 1 Literature overview of case studies on the design of measurement campaigns for full-scale WWTPs in view of subsequent data reconciliation based on mass balances, serving as a benchmark

#	WWTP	Type/capacity	Configuration	Main goal of the study	Were key variables specified?	How was the measurement campaign carried out?
1	WWTP Katwoude,[20] average data of one year	Municipal WWTP 86 300 p.e.[a]	A2/O process with limited biological phosphorus removal and mainly chemical phosphorus removal	Reliable data for model calibration	Partially Only total oxygen consumption, and the amount of nitrified nitrogen and denitrified nitrogen were explicitly defined as key variables Variables involved in SRT calculation were not defined as key variables but implied to be so	Measurement campaign was not implemented Average data of one year from SCADA* and routine lab analysis of the plant was used
2	WWTP Katwoude,[20] 8 day measurement campaign	Same as previous	Same as previous	Reliable data for model calibration	Yes Seven process flow rates	8 day measurement campaign was carried out with 24 h-composite samples (where available) and grab samples (at peak flow) combined with data from SCADA[b] and routine lab analyses
3	WWTP Deventer[22]	Municipal WWTP 182 000 p.e.	Modified UCT-process according to the BCFS-concept	Reliable data for calculating sludge retention time and operational conditions for benchmarking	Partially Only total oxygen consumption, and the amount of nitrified nitrogen and denitrified nitrogen were explicitly defined as key variables Variables involved in SRT calculation were not defined as key variables but implied to be so	Intensive measurement campaign was carried out on three separate days with 24 h-composite samples (where available) and grab samples combined with data from SCADA and routine lab analyses
4	WWTP Houtrust[21]	Municipal WWTP 330 000 p.e.	A2/O process with primary and secondary sludge fermentation	Reliable data for model validation and calibration	Yes 15 flow variables and six mass flows of COD and total phosphorus	The plant was monitored for six weeks. Collected comprehensive data set consisting of 24 h-composite samples (where available), grab samples, data from SCADA and routine lab analyses
5	WWTP Tabriz[19]	Petrochemical WWTP 4800 m^3 per day	Following steps: oil separation coagulation & flocculation - activated sludge - sand filter	Reliable data for evaluating the performance of individual unit processes	No Only flow measurements were balanced by data reconciliation Mass flows of COD were reported to be balanced, but they were calculated from balanced flows and measured COD concentrations	Four sampling runs were carried out and combined with data from SCADA and routine lab analyses

[a] p.e. = population equivalent. [b] SCADA = supervisory control and data acquisition system.

other approaches such as moving-time window data reconciliation could be used,[23] but they were out of the scope of this study.

c. **Initially measured data set and potential additionally measured variables**, with estimated cost and variance of measured variables. The additionally measured variables were proposed relative to a set of initially available data to ensure the fulfilment of the main goal. From the previous studies, however, it was not always clear which of the presented measured data were initially available and which ones were proposed additionally. For the previous studies in which the initially measured data were not clearly specified, the initially measured data set was assumed. Since the costs of additionally measured variables (flows and concentrations) were not specified in any of the previous studies, the costs of all measured variables were assumed equal. This assumption implies that the cost is proportional to the number of additionally measured variables. The use of measurement-

specific costs would not limit the applicability of the procedure but may deliver different optimal solutions in terms of cost. The uncertainty of the measured variables was expressed in terms of their standard deviation, the magnitude of which could be derived from data provided in previous studies.

The set(s) of additionally measured variables proposed by applying the experimental design procedure from Le *et al.*[17] was subsequently compared to those actually carried out in the previous studies. The previous studies defined the experimental design based on expert knowledge, and not explicitly as a multi-objective optimization problem like Le *et al.*[17] Thus, the outcome was a single set of additional measurements. This was compared with the optimal solutions obtained in this study in terms of the number of additionally measured variables and in terms of the accuracy of key process variables. Only additionally measured variables that were used for data reconciliation were considered in the comparison. A detailed analysis of previous approaches was made by answering the following questions:

– Main goal and key variables: Were the key variables defined in previous studies? If yes, to what extent did they reflect the main goal of the measurement campaign?

– Mass balance setup: Were the mass balances relevant?

– Experimental design results: Was the set of (additional) measured variables implemented in the previous study relevant – did it allow the identification of key variables? Are there any alternative sets of additionally measured variables which may be better in terms of the number of required additionally measured variables and/or resulting accuracy of key variables?

3 Results

In what follows, the case studies from literature, in which a WWTP measurement campaign was planned in view of data reconciliation, are analysed one by one. Table 2 summarizes the results. The comparison with respect to the main goal and the key variables was evaluated by analysing the defined and identified key variables (*A* and *B*). The relevance of the mass balance setup was assessed by evaluating the number of defined and relevant mass balances (*C* and *D*). The experimental design results were assessed by considering the initial dataset, potential additionally measured variables, the additional measurements actually performed and the missing crucial variables (*E*–*I*). The sets of optimal solutions, as well as the minimum number of additional variables needed to fulfil the main goal were evaluated as well (*J*–*M*).

3.1 Case study 1: Meijer *et al.*,[20] average data of one year

3.1.1 Plant configuration and measured data. WWTP Katwoude (The Netherlands) has a capacity of 86 300

Table 2 Overview of the results from the case studies presented in the literature, in comparison with the experimental design results from this study

	Number of	Case study 1 (ref. 20) average data of one year	Case study 2 (ref. 20) 8 day data	Case study 3 (ref. 22)	Case study 4 (ref. 21)	Case study 5 (ref. 19)
Main goal and key variables						
A	Key variables (number of which defined in a previous study)	6 (3)	7 (7)	11 (3)	21 (21)	9 (0)[a]
B	Key variables identified in a previous study	3	0	11	17	9
Mass balance setup						
C	Mass balances set up by previous studies	8	12	14	20	8
D	Relevant mass balances among *C*	8	6	14	19	8
Experimental design results						
E	Initially available measured variables	20	8	9	11	2
F	Potential additionally measured variables, *i.e.* initially unmeasured variables in mass balances related to key variables[b]	2	12	25	29	17
G	Additionally measured variables obtained (measurement campaign) in a previous study	0[c]	21	25	27	17
H	Relevant additionally measured variables obtained in a previous study, *i.e.* contributing to the identification of key variables	NA[c]	4	25	24	17
I	Missing essential additionally measured variable in a previous study, *i.e.*, required for the identification of all key variables	2	2	0	2	0
J	Number of Pareto-optimal solutions found using an experimental design procedure	1	6	8	12	10
K	Minimum number of additionally measured variables needed to identify all key variables (*i.e.* for the Pareto-optimal solution with minimum number of additionally measured variables)	2	7	11	18	7
L	Additionally measured variables for the most accurate Pareto-optimal solution (*i.e.* for the optimal solution with a maximum number of additionally measured variables)	2	12	25	29	17
M	Maximum potential reduction of additionally measured variables, compared with a previous study with a Pareto-optimal solution with a minimum number of additionally measured variables[d]	NA[c]	70%	56%	38%	59%

[a] Only considering flows. [b] As checked in step 3 of the experimental design procedure.[17] [c] No additional measurements were performed in this case study. [d] $M = (G + I - K)/(G + I)$. Note: in case the additionally measured variables proposed in a previous study (*G*) were not sufficient for the identification of key variables, the number of missing essential additionally measured variables (*I*) was added.

population equivalents (p.e.) and was built according to anaerobic/anoxic/oxic design (A2/O design) with limited biological phosphorus removal and mainly chemical phosphorus removal (Fig. 1). In the first case study by Meijer *et al.*,[20] average data of one year from SCADA and routine lab analysis of the plant (Table C1, Appendix C1, ESI†) were used for data reconciliation. Twenty measured variables were initially available. No additional data were gathered (no measurement campaign).

3.1.2 Main goal and key variables. The main goal of Meijer *et al.*[20] was to obtain reliable data for model calibration. For this purpose, they proposed to collect additional data to increase data redundancy by aiming at more mass balance equations than the number of unknown variables. The following three key variables were defined explicitly: net oxygen consumption (OC_{net}, kg per day), amounts of nitrified nitrogen (NITR, kg per day) and denitrified nitrogen (DENI, kg per day). The calculation of SRT was also put forward as one of the main goals. The additional key variables related to this main goal were not predefined by Meijer *et al.*[20] but defined in this study as the effluent flow rate (Q_{ef}), the excess sludge flow rate (Q_{ex}) and the mass flow of total phosphorus in the influent ($m_{TP_{in}}$). Three more variables were involved in the SRT calculation, namely the phosphate concentration in the total suspended solids (TSS) fractions of Q_{ef}, Q_{ex} and the reactor. However, the latter variables are non-conservative and are therefore not considered as key variables. Overall, there were 9 variables related to the main goal, 6 of which were key variables (Table 2). The key variables defined by the previous study were not considered sufficient to fulfil the main goal, since 3 variables (related to the SRT) were not defined.

3.1.3 Mass balance setup. Meijer *et al.*[20] set up 8 mass balances for data reconciliation (Table 3). Total mass flow (#1 and #2) and total phosphorus (#3 and #4) balances were set up over the whole WWTP and around the centrifuge. COD (#5, #8), TKN (#6) and NOx (#7) mass balances were set up over the whole WWTP. The mass balance setup was considered relevant since all key variables were included and the mass balances were all independent from each other.

3.1.4 Experimental design results. The experimental design procedure of Le *et al.*[17] was applied in this study to determine sets of additionally measured variables that guarantee the identifiability of key variables. There were two potential additionally measured variables, namely the effluent flow rate (Q_{ef}) and the centrifuge outflow rate (Q_{cent}). The application of our experimental design procedure (step 4: feasibility evaluation) learned that only Q_{ef} could be identified based on the 20 measured variables initially available. All 6 key variables could only be identified when both Q_{ef} and Q_{cent} were added to the initially available data. However, no additional measurements were carried out by Meijer *et al.*[20] Contradictory enough, they stated that all key variables were identified using data reconciliation – apparently, identifiability was not checked when applying the experimental design procedure as it was not part of their procedure.

By discarding the effluent flow (Q_{ef}) and centrifuge outflow (Q_{cent}) from the set of potential additionally measured variables and by performing again a feasibility evaluation (*i.e.*, checking variable identifiability without Q_{ef} and Q_{cent} as additional measurements), it was concluded in this study that they were essential. Q_{ef} was required to identify the flow of excess sludge (Q_{ex}) and the mass flow of total phosphorus in the influent ($m_{TP_{in}}$). Q_{cent} was required to identify the amount of denitrified nitrogen (DENI), the amount of nitrified nitrogen (NITR) and the total oxygen consumption (OC_{net}).

3.2 Case study 2: Meijer *et al.*,[20] data from an 8-day measurement campaign

3.2.1 Plant configuration and measured data. The plant configuration was the same as in case study 1, but the main

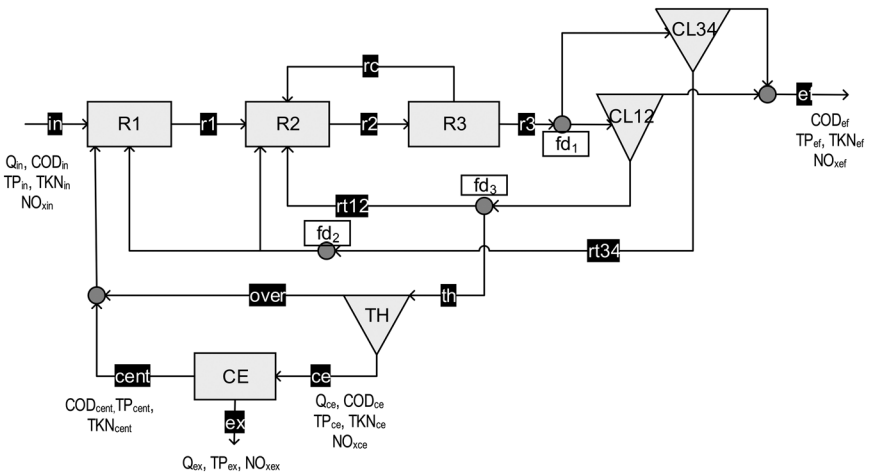

Fig. 1 WWTP Katwoude, adapted from Meijer *et al.*[20] Unit processes are indicated in grey: R1 = mixed non-aerated selector, R2 = completely mixed anoxic reactor, R3 = aerated carrousel reactor, CL12, CL34 = four clarifiers were operated in pairs. fd_1, fd_2, fd_3 = flow dividers, TH = sludge thickener and CE = centrifuge. The black boxes refer to the name of the streams. The measured variables are indicated by name at their respective positions.

Table 3 Mass balances over the whole WWTP and centrifuge unit (CE) set up by Meijer *et al.*[20] Key variables are indicated in bold

#	Unit process	Mass balances	Unit
1	WWTP	$\boldsymbol{Q_{in}} - \boldsymbol{Q_{ef}} - \boldsymbol{Q_{ex}}$	Flow (m^3 per day)
2	CE	$Q_{ce} - Q_{cent} - \boldsymbol{Q_{ex}}$	
3	WWTP	$\boldsymbol{m_{TP_{in}}} - m_{TP_{ef}} - m_{TP_{ex}}$	Total phosphorus (kg per day)
4	CE	$m_{TP_{ce}} - m_{TP_{cent}} - m_{TP_{ex}}$	
5	WWTP	$m_{COD_{in}} - m_{COD_{ef}} - m_{COD_{ce}} + m_{COD_{cent}} - \boldsymbol{OC_{cod}} - 2.87 \cdot \boldsymbol{DENI}$	COD and nitrogen (kg per day)
6	WWTP	$m_{TKN_{in}} - m_{TKN_{ef}} - m_{TKN_{ce}} + m_{TKN_{cent}} - \boldsymbol{NITR}$	
7	WWTP	$\boldsymbol{DENI} - \boldsymbol{NITR} + m_{NO_{x_{ef}}} + m_{NO_{x_{in}}} - m_{NO_{x_{ex}}}$	
8	WWTP	$\boldsymbol{OC_{net}} - \boldsymbol{OC_{cod}} - 4.57 \cdot \boldsymbol{NITR}$	

Q = flow, m_{TP} = total phosphorus mass flow, m_{TKN} = Kjeldahl nitrogen mass flow, m_{COD} = COD mass flow, m_{NO_x} = NO_3 mass flow rate. OC_{net} = net oxygen consumption (kg per day), OC_{cod} = oxygen for COD removal (kg per day), NITR = nitrified nitrogen (kg per day), DENI = denitrified nitrogen (kg per day).

goal and key variables were different. The measured data are summarised in Table C2a, Appendix C2, ESI.† The initially measured variables were not specified by Meijer *et al.*;[20] they were assumed in this study to be the same as in case study 1.

3.2.2 Main goal and key variables. The main goal of the second case study was to have reliable data for model calibration and validation. In order to achieve this goal, Meijer *et al.*[20] defined 7 key variables, all of which were flow rates: influent flow rate (Q_{in}), return sludge flow rate from clarifiers 1 and 2 (Q_{rt12}), return sludge flow rate from clarifiers 3 and 4 (Q_{rt34}), recycle flow rate (Q_{rc}), thickener inflow rate (Q_{th}), centrifuge inflow rate (Q_{cent}) and excess sludge (Q_{ex}) flow rate.

3.2.3 Mass balance setup. Meijer *et al.*[20] set up 12 mass balances (Table 4). Flow balances were set up over the selector (R1), denitrification reactor (R2), aerated carousel (R3), clarifiers (CL), thickeners (TH) and centrifuge (CE). Total phosphorus balances were set up over R1, R2, R3 and CL. Ammonium was balanced over R1 and R2.

3.2.4 Experimental design results. Twelve potential additionally measured variables were available (Q_{r1}, Q_{r2}, Q_{r3}, Q_{rc}, Q_{rt12}, Q_{rt34}, Q_{ef}, Q_{th}, Q_{ce}, Q_{over}, Q_{cent}, and Q_{ex}), making up 2^{12} = 4096 possible combinations of additionally measured variables. These were all evaluated by applying the experimental design procedure of Le *et al.*,[17] resulting in 232 solutions, 6 of which were on the Pareto-front (Fig. 2, detailed in Table C2b, Appendix C2, ESI†). The Pareto-optimal solutions identified in this study involved 7 to 12 additional measured variables, all of which were flow rates.

For comparison, Meijer *et al.*[20] used 24 measured variables for data reconciliation, 21 of which were measured additionally during an 8-day measurement campaign. The set of additionally measured variables of Meijer *et al.*[20] was not presented in Fig. 2 since it did not satisfy the defined main goal and therefore was not a solution. In fact, our findings suggest that none of the key variables could be identified with the proposed set of additionally measured variables. In order to identify the seven key variables, the proposed set of measured data (Table C2a, Appendix C2, ESI†) should be complemented with the flow from R3 to R2 (Q_{r2}) and the centrifuge output flow (Q_{cent}), making up two additionally measured variables.

From the experimental design procedure, it is clear that only flow measurements can help in identifying flows.[17] However, Meijer *et al.*,[20] with the aim of identifying only total flows, included 6 mass balances for total phosphorus (TP) and total Kjeldahl nitrogen (TKN) in the system of mass balances (Table 4) and 17 corresponding concentration measurements of TP and TKN, 15 of which were measured additionally. The measurements of TP and TKN, however, will not contribute to the identification of flow variables as there is no direct relation to total flows in the mass balances.

Table 4 Mass balances around selector (R1), denitrification reactor (R2), aerated carousel (R3), clarifiers (CL), thickeners (TH) and centrifuge (CE), adapted from the incidence matrix of Meijer *et al.*[20] Variables in bold are key variables

		Unit process	Mass balance	Unit
1	R1	Selector	$\boldsymbol{Q_{in}} + \boldsymbol{Q_{rt34}} + Q_{over} + \boldsymbol{Q_{cent}} - Q_{r1}$	Flow (m^3 per day)
2	R2	Denitrification reactor	$Q_{r1} + \boldsymbol{Q_{rc}} + \boldsymbol{Q_{rt12}} - Q_{r2}$	
3	R3	Aerated carousel	$Q_{r2} - \boldsymbol{Q_{rc}} - Q_{r3}$	
4	CL	Clarifiers	$Q_{r3} - Q_{ef} - \boldsymbol{Q_{rt12}} - \boldsymbol{Q_{rt34}} - \boldsymbol{Q_{th}}$	
5	TH	Thickeners	$\boldsymbol{Q_{th}} - Q_{over} - Q_{ce}$	
6	CE	Centrifuge	$Q_{ce} - \boldsymbol{Q_{cent}} - \boldsymbol{Q_{ex}}$	
7	R1	Selector	$m_{TP_{in}} + m_{TP_{rt34}} + m_{TP_{over}} + m_{TP_{cent}} - m_{TP_{r1}}$	Total phosphorus (kg per day)
8	R2	Denitrification reactor	$m_{TP_{r1}} + m_{TP_{rc}} + m_{TP_{rt12}} - m_{TP_{r2}}$	
9	R3	Aerated carousel	$m_{TP_{r2}} - m_{TP_{rc}} - m_{TP_{r3}}$	
10	CL	Clarifiers	$m_{TP_{r3}} - m_{TP_{ef}} - m_{TP_{rt12}} - m_{TP_{rt34}} - m_{TP_{th}}$	
11	R1	Selector	$m_{NH_{in}} + m_{NH_{rt34}} + m_{NH_{over}} + m_{NH_{cent}} - m_{NH_{r1}}$	Ammonium (kg per day)
12	R2	Denitrification reactor	$m_{NH_{r1}} + m_{NH_{rc}} + m_{NH_{rt12}} - m_{NH_{r2}}$	

Q = flow, m_{TP} = total phosphorus mass flow, m_{NH} = ammonium mass flow.

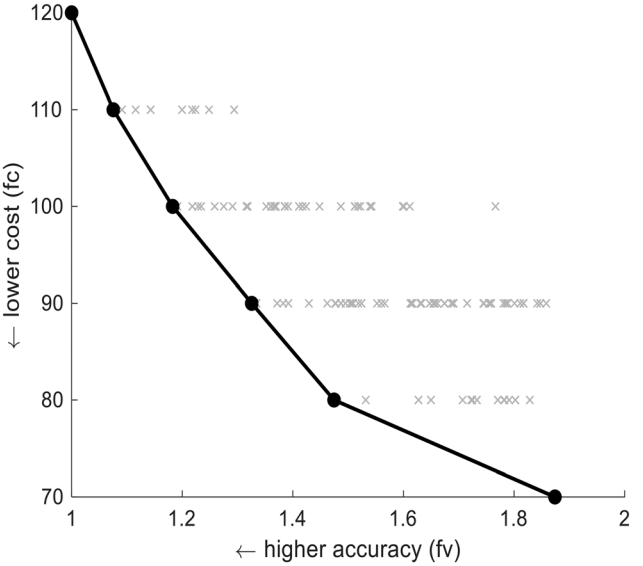

Fig. 2 Pareto optimal solutions for the setup of Meijer *et al.*[20] determined by the experimental procedure of Le *et al.*[17] and expressed in terms of accuracy and costs. The line with the filled circles (black) denotes the Pareto-optimal front. 'x' = a solution.

Therefore, setting up total phosphorus and ammonium mass balances and performing 17 concentration measurements of TP and TKN, as proposed by Meijer *et al.*,[20] were irrelevant for the reconciliation of flow rates. The result of the flow balancing would be the same with or without the TP and TKN mass balances and measurements.

In brief, only 6 of the 12 mass balances set in a previous study were relevant. In addition, Meijer *et al.*[20] proposed 21 additionally measured variables, only 4 of which were relevant and still missing 2 essential ones. Using the experimental design procedure from Le *et al.*,[17] the minimum number of additionally measured variables was 7. As a result, the potential reduction in the number of additionally measured variables could be up to 70% (= [21 + 2 − 7]/[21 + 2]) compared to the proposed set of Meijer *et al.*[20]

3.3 Case study 3: Puig *et al.*[22]

3.3.1 Plant configuration and measured data. The Deventer WWTP (The Netherlands) has a capacity of 182 000 p.e. and was built as a modified UCT-process according to the BCFS-concept[24] (Fig. 3). The measured data are summarised in Table C3a, Appendix C3, ESI.†

3.3.2 Main goal and key variables. The main goals of Puig *et al.*[22] were to calculate the solids retention time (SRT) by different methods and to calculate variables related to operating conditions such as the amount of nitrogen nitrified (NITR, kg per day), amount of nitrogen denitrified (DENI, kg per day) and total oxygen consumption (OC_{net}, kg per day).

Four methods for SRT calculation were considered. The first one was the classical SRT calculation obtained as the ratio of the sludge mass TSS in the reactor to the sludge mass TSS flow rate leaving the reactor through the waste sludge stream ('was'), the secondary settler effluent ('ef') and the stripping reactor effluent streams ('se') (eqn (1)).

$$\mathrm{SRT_{classical}} = \frac{V_{\mathrm{reactor}} \times \mathrm{TSS_{reactor}}}{Q_{\mathrm{was}} \times \mathrm{TSS_{was}} + Q_{\mathrm{ef}} \times \mathrm{TSS_{ef}} + Q_{\mathrm{se}} \times \mathrm{TSS_{se}}} \quad (1)$$

Assuming that the COD and total phosphorus fractions of the sludge (TSS) in the reactor, in the waste activated sludge and in the effluent streams are constant and the same, the SRT can also be calculated based on their particulate COD content (eqn (2)) or based on their particulate total phosphorus fraction (eqn (3)):

$$\mathrm{SRT_{CODtss\_output}} = \frac{V_{\mathrm{reactor}} \times \mathrm{COD_{TSS,reactor}}}{Q_{\mathrm{was}} \times \mathrm{COD_{TSS\_was}} + Q_{\mathrm{ef}} \times \mathrm{COD_{TSS\_ef}} + Q_{\mathrm{se}} \times \mathrm{COD_{TSS\_se}}} \quad (2)$$

$$\mathrm{SRT_{TPtss\_output}} = \frac{V_{\mathrm{reactor}} \times \mathrm{TP_{TSS,reactor}}}{Q_{\mathrm{was}} \times \mathrm{TP_{TSS\_was}} + Q_{\mathrm{ef}} \times \mathrm{TP_{TSS\_ef}} + Q_{\mathrm{se}} \times \mathrm{TP_{TSS\_se}}} \quad (3)$$

The total phosphorus fraction of the sludge (TP_{TSS}) represents particulate phosphorus and thus equals the difference between total phosphorus (TP) and soluble phosphate (PO_4). By taking this into account and by substituting the total phosphorus mass balance over the plant (mass balance #7, Table 5), eqn (3) is rewritten as eqn (4):

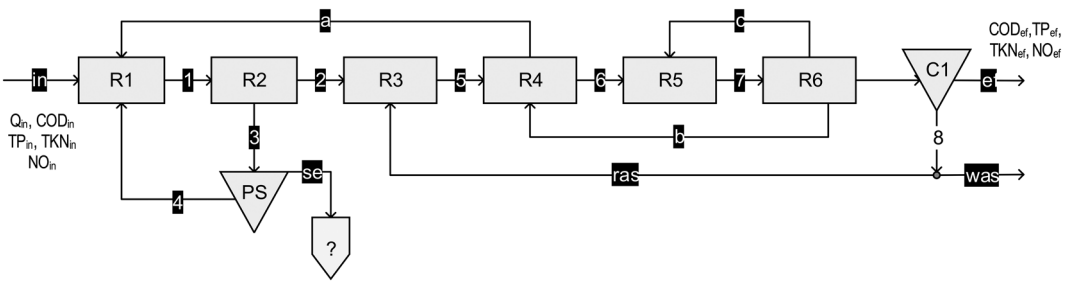

Fig. 3 Flow diagram of the Deventer WWTP, The Netherlands (adapted from Puig *et al.*[22]). R1 and R2 = two anaerobic reactors, R3 = a contact tank, R4 = an anoxic reactor, R5 = an alternatively aerated reactor, R6 = aerated rector, C1 = six secondary settlers (in parallel) and PS = stripping reactor. Measured variables are indicated.

Table 5 Mass balances around the anaerobic tank 1 selector (R1), anaerobic tank 2, the denitrification reactor (R2), the anoxic contact tank (R3), the anoxic tank (R4), the alternate aerated tank (R5), the aerobic tank (R6), and the clarifiers (C1), translated from the incidence matrix of Puig *et al.*[22] Variables in bold are key variables

	Unit process	Mass balance	Unit
1	WWTP	$\boldsymbol{Q_{in} - Q_{se} - Q_{ef} - Q_{was}}$	Flow (m^3 per day)
2	R1 + R2	$Q_{in} - \boldsymbol{Q_{se}} - Q_2 + Q_a$	
3	R3 + R4	$Q_2 + Q_{ras} - Q_a - Q_6 + Q_b$	
4	R5	$Q_6 - Q_7 + Q_c$	
5	R6	$Q_7 - Q_c - Q_b - Q_8$	
6	C1	$-Q_{ras} + Q_8 - \boldsymbol{Q_{ef}} - \boldsymbol{Q_{was}}$	
7	WWTP	$\boldsymbol{m_{TP_{in}} - m_{TP_{se}} - m_{TP_{ef}} - m_{TP_{was}}}$	Total phosphorus (kg per day)
8	R1 + R2	$\boldsymbol{m_{TP_{in}} - m_{TP_{se}}} - m_{TP_2} + m_{TP_a}$	
9	R3 + R4	$m_{TP_2} + m_{TP_{ras}} - m_{TP_a} - m_{TP_6} + m_{TP_b}$	
10	C1	$m_{TP_8} - m_{TP_{ras}} - \boldsymbol{m_{TP_{ef}} - m_{TP_{was}}}$	
11	WWTP	$m_{TKN_{in}} - m_{TKN_{se}} - m_{TKN_{ef}} - m_{TKN_{was}} - \mathbf{NITR}$	COD and nitrogen (kg per day)
12	WWTP	$\mathbf{NITR} - \mathbf{DENI} - NO_{ef} - NO_{se} - NO_{was}$	
13	WWTP	$m_{COD_{in}} - m_{COD_{se}} - m_{COD_{ef}} - m_{COD_{was}} - 2.78 \times \mathbf{DENI} - \mathbf{OC_{cod}}$	
14	WWTP	$\boldsymbol{-OC_{cod}} - 4.57 \times \mathbf{NITR} + \mathbf{OC_{net}}$	

Q = flow, m_{TP} = total phosphorus mass flow, m_{TKN} = Kjeldahl nitrogen mass flow, m_{COD} = COD mass flow, m_{NO_x} = NO_3 mass flow. OC_{net} = total oxygen consumption (kg per day), OC_{cod} = oxygen for COD removal (kg per day), NITR = nitrified nitrogen (kg per day), DENI = denitrified nitrogen (kg per day).

$$SRT_{TPtss\_input} = \frac{V_{reactor} \times TP_{TSS,reactor}}{Q_{in} \times TP_{in} - Q_{was} \times PO_{4_{was}} - Q_{ef} \times PO_{4_{ef}} - Q_{se} \times PO_{4_{se}}}, \quad (4)$$

in which the SRT is calculated based on the total phosphorus mass entering the plant and the phosphate concentrations in the waste activated sludge and effluent streams, besides the particulate total phosphorus mass in the reactor.

The key variables corresponding with the defined main goals were not specified by Puig *et al.*[22] but were deduced in this study. Eleven variables related to SRT (eqn (1)–(4)) and operating conditions (NITR, DENI and OC_{net} in mass balances #11–#14, Table 5) were conservative and therefore defined as key variables, namely: the flow rates and mass flows of total phosphorus in the influent (Q_{in}, $m_{TP_{in}}$), effluent (Q_{ef}, $m_{TP_{ef}}$), excess sludge (Q_{was}, $m_{TP_{was}}$), stripped effluent (Q_{se}, $m_{TP_{se}}$), amount of nitrified nitrogen (NITR, kg per day), denitrified nitrogen (DENI, kg per day) and total oxygen consumption of WWTP (OC_{net}, kg per day).

From the SRT calculations (eqn (2)–(4)), it is clear that there are more variables related to the main goal, namely the mass flows of total suspended solids (TSS), orthophosphate (PO), total particulate phosphorus TSS (TP_{TSS}) and particulate COD (COD_{TSS}). They could not, however, be defined as key variables since they are not conservative quantities, which means that no mass balances can be set up for these compounds. As a result, the SRT calculations were based on both the measured and reconciled variables.

3.3.3 Mass balance setup. Puig *et al.*[22] set up 14 mass balances (Table 5). Flows were balanced over the anaerobic tank 1 (R1), the anaerobic tank 2 (R2), the anoxic contact tank (R3), the anoxic tank (R4), the alternate aerated tank (R5), the aerobic tank (R6), and the clarifiers (C1). Total phosphorus was balanced over R1, R2, R3, R4, C1 and the whole WWTP. Total Kjeldahl nitrogen was balanced over the whole WWTP. All the mass balances were considered relevant in this study.

3.3.4 Experimental design results. There were 25 potential additionally measured variables (Q_{se}, Q_2, Q_{ras}, Q_a, Q_6, Q_7, Q_c, Q_b, Q_8, Q_{ef}, Q_{was}, TP_{se}, TP_2, TP_{ras}, TP_a, TP_6, TP_b, TP_8, TP_{was}, COD_{se}, COD_{was}, NO_{se}, NO_{was}, TKN_{se} and TKN_{was}), which implies 2^{25} = 33 554 432 possible combinations of additionally measured variables. They were all evaluated in this study through the experimental design procedure of Le *et al.*,[17] resulting in 4944 solutions, 8 of which were Pareto-optimal (Fig. 4, detailed in Table C3b, Appendix C3, ESI†).

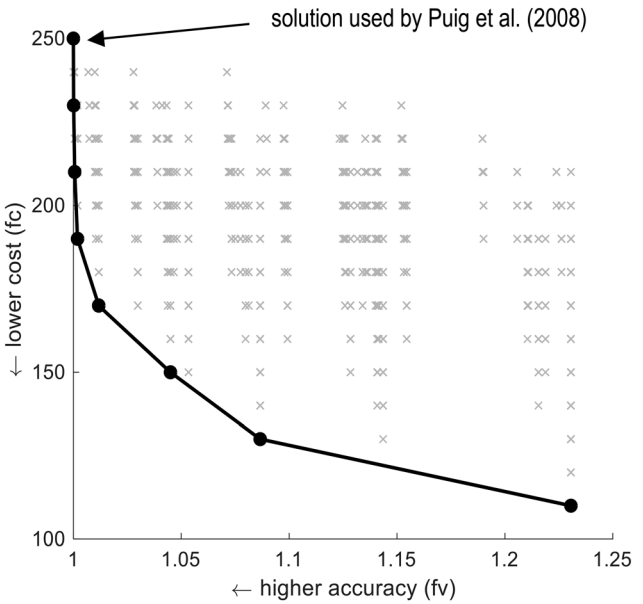

Fig. 4 Pareto optimal solutions for the setup of Puig *et al.*[22] determined using the experimental procedure of Le *et al.*[17] and expressed in terms of accuracy and costs. The line with the filled black circles denotes the Pareto-optimal front. 'x' = a solution.

The set of measured data used by Puig *et al.*[22] for data reconciliation contained 25 additionally measured variables (indicated in Fig. 4 and detailed in Table C3a, Appendix C3, ESI†). This solution satisfied the main goal and, moreover, belonged to the Pareto-optimal front. More specifically, it was the most accurate but also the most expensive Pareto-optimal solution. In this case study, the minimum number of additionally measured variables was 11. As a result, the potential reduction in the number of additionally measured variables was 56% compared to the proposed set of Puig *et al.*[22] Thus, the set of additional measurements was relevant and allowed the identification of all the key variables. Nevertheless, alternative solutions were found in this study involving fewer measurements and lower cost.

3.4 Case study 4: Meijer *et al.*[21]

3.4.1 Plant configuration and measured data. WWTP Houtrust (The Netherlands) has a yearly average loading of 330 000 p.e. It consists of an A2/O process configuration with primary and secondary sludge digesters (Fig. 5). The measured data are summarised in Table C4a, Appendix C4, ESI.†

3.4.2 Main goal and key variables. The main goal of Meijer *et al.*[21] was to have reliable data for model calibration and validation. The following 21 key variables were defined: all 15 flow rates (Q_4, Q_5, Q_7, Q_{15}, Q_{17}, Q_{23}, Q_{26}, Q_{27}, Q_{28}, Q_{31}, Q_{34}, Q_{35}, Q_{37}, Q_{38}, and Q_{40}, corresponding with the stream numbers in Fig. 5) and 6 mass flows of total phosphorus (TP) and COD, namely those of the waste activated sludge ($m_{TP_{26}}$ and $m_{COD_{26}}$), the raw influent (m_{TP_4} and m_{COD_4}) and the settled influent (m_{TP_7} and m_{COD_7}). All the defined key variables were considered appropriate with regard to the main goal.

3.4.3 Mass balance set up. Meijer *et al.*[21] set up 20 mass balances, consisting of four groups according to Table 6.

3.4.4 Experimental design results. A number of 2^{29} = 536 870 912 possible combinations of additionally measured variables were evaluated with the experimental design procedure, resulting in 4824 solutions. The Pareto-optimal front contains 12 solutions (detailed in Table C4b, Appendix C4, ESI†).

Meijer *et al.*[21] used 34 measured variables for data reconciliation, 27 of which were measured additionally. However, only 24 additionally measured variables actually contributed to the identification of key variables. The TSS mass flow balance around the waste sludge thickener (#20 in Table 6) did not contribute to the identification of any key variables. Therefore, this mass balance and the three associated TSS measurements (TSS_{26}, TSS_{37} and TSS_{27}) were not necessary in this case study.

Moreover, 4 key variables could not be identified with the measured data from Meijer *et al.*:[21] the total influent flow rate (Q_4), the return activated sludge flow rate (Q_{23}), the influent COD mass flow (m_{COD_4}) and the influent mass flow of total phosphorus (m_{TP_4}). So, the main goal was not entirely achieved by their measurement campaign. Still, Meijer *et al.*[21] reported that Q_4 and Q_{23} were balanced by data reconciliation – no results were reported for balancing m_{COD_4} and m_{TP_4}.

The set of measured data applied for data reconciliation by Meijer *et al.*[21] missed two crucial additionally measured variables: the settled influent flow rate (Q_7) to balance m_{COD_4} and m_{TP_4} and the inflow rate to the secondary clarifiers (Q_{15}) to balance Q_{23}. These two variables were found essential to identify all the key variables during the redundancy analysis performed in this study. The addition of these two variables (Q_7 and Q_{15}) to the measured data set used by Meijer *et al.*[21] would have resulted in a solution (indicated by 'x' in Fig. 6 and detailed in Table C4b, Appendix C4, ESI†), *i.e.*, would have allowed the identification of all key variables. However, the latter solution is not a Pareto-optimal solution since it has the same number of additionally measured variables but about 38% lower accuracy than the most expensive Pareto-optimal solution (accuracy fv = 1.38). In this case study, the minimum additionally measured variables were 18. The potential reduction in the number of additionally measured

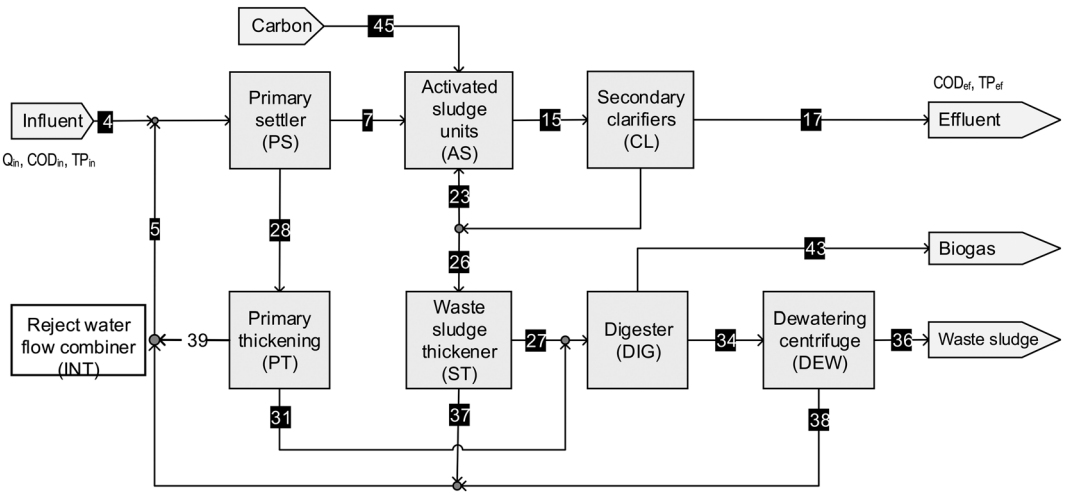

Fig. 5 Flow diagram of the Houtrust WWTP, The Netherlands, adapted from Meijer *et al.*[21] Measured variables are indicated.

Table 6 Mass balances around groups and individual process units by Meijer *et al.*[21] Variables in bold are key variables

	Process unit	Mass balance	
1	WWTP	$\boldsymbol{Q_4} + Q_{40} - \boldsymbol{Q_{17}} - Q_{35}$	Total flow (m^3 per day)
2	Water line	$\boldsymbol{Q_7} - \boldsymbol{Q_{17}} - \boldsymbol{Q_{26}}$	
3	Sludge line	$\boldsymbol{Q_{26}} + \boldsymbol{Q_{31}} - \boldsymbol{Q_{35}} - \boldsymbol{Q_{37}} - \boldsymbol{Q_{38}}$	
4	Rejected water line	$\boldsymbol{Q_5} - \boldsymbol{Q_{37}} - \boldsymbol{Q_{38}} - Q_{39} - \boldsymbol{Q_{40}}$	
5	Activated sludge units	$\boldsymbol{Q_7} + \boldsymbol{Q_{23}} - \boldsymbol{Q_{15}}$	
6	Primary settler	$\boldsymbol{Q_4} + \boldsymbol{Q_5} - \boldsymbol{Q_7} - \boldsymbol{Q_{28}}$	
7	Primary thickening	$\boldsymbol{Q_{28}} - \boldsymbol{Q_{31}} - Q_{39}$	
8	Secondary clarifier	$\boldsymbol{Q_{15}} - \boldsymbol{Q_{23}} - \boldsymbol{Q_{26}} - \boldsymbol{Q_{17}}$	
9	Waste sludge thickener	$\boldsymbol{Q_{26}} - \boldsymbol{Q_{27}} - \boldsymbol{Q_{37}}$	
10	Digestor	$\boldsymbol{Q_{27}} + \boldsymbol{Q_{31}} - \boldsymbol{Q_{34}}$	
11	Dewatering	$\boldsymbol{Q_{34}} - \boldsymbol{Q_{35}} - \boldsymbol{Q_{38}}$	
12	WWTP	$\boldsymbol{m_{TP_4}} - m_{TP_{17}} - m_{TP_{35}}$	Total phosphorus (kg per day)
13	Water line	$\boldsymbol{m_{TP_7}} - m_{TP_{17}} - \boldsymbol{m_{TP_{26}}}$	
14	Secondary clarifier	$m_{TP_{15}} - m_{TP_{23}} - m_{TP_{26}} - m_{TP_{17}}$	
15	Primary settler	$\boldsymbol{m_{TP_4}} + m_{TP_5} - \boldsymbol{m_{TP_7}} - m_{TP_{28}}$	
16	Sludge line	$\boldsymbol{m_{TP_{26}}} + m_{TP_{28}} - m_{TP_5} - m_{TP_{35}}$	
17	Primary settler	$\boldsymbol{m_{COD_4}} + m_{COD_5} - \boldsymbol{m_{COD_7}} - m_{COD_{28}}$	COD (kg per day)
18	Sludge line	$\boldsymbol{m_{COD_{26}}} + m_{COD_{28}} - m_{COD_5} - m_{COD_{35}} - m_{COD_{43}}$	
19	Digester	$m_{COD_{27}} + m_{COD_{31}} - m_{COD_{34}} - m_{COD_{43}}$	
20	Waste sludge thickener	$m_{TSS_{26}} - m_{TSS_{37}} + m_{TSS_{27}}$	TSS (kg per day)

Q = flow, m_{TP} = total phosphorus mass flow, m_{COD} = COD mass flow, m_{TSS} = mass flow of total suspended solids.

variables could be up to 38% compared to the proposed set of Meijer *et al.*[21]

Overall, 19 of the 20 mass balances set by Meijer *et al.*[21] were considered relevant. However, the set of additional measurements did not allow the identifiability of the key variables. On the one hand, unnecessary measurements were performed. On the other hand, crucial variables were missing.

3.5 Case study 5: Behnami *et al.*[19]

3.5.1 Plant configuration and measured data. The Tabriz petrochemical WWTP (Fig. 7) has an average design flow rate of 4800 m^3 per day. The measured data are summarized in Table C5a, Appendix C5, ESI.†

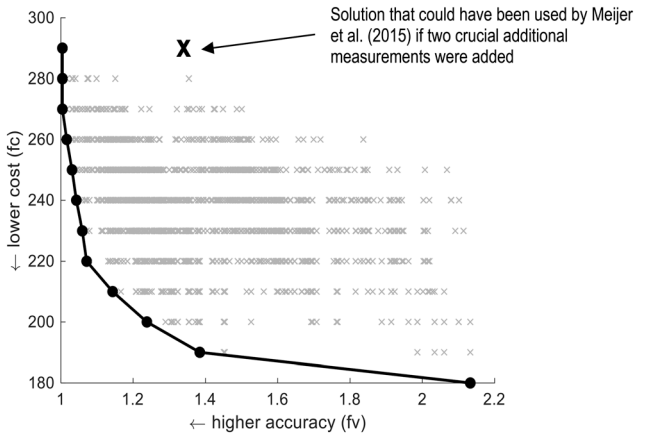

Fig. 6 Pareto optimal solutions for the setup of Meijer *et al.*[21] determined using the experimental procedure of Le *et al.*[17] and expressed in terms of accuracy and costs. The line with the filled grey circles denotes the Pareto-optimal front. 'x' = a solution.

3.5.2 Main goal and key variables. The main goal of Behnami *et al.*[19] was stated in a general way as to have reliable data for evaluating the performance of the individual WWTP units. They did not specify which variables had to be identified. The key variables were defined in this study from the presented data reconciliation results. The key variables comprised the flow rates of influent process wastewater (Q_{in1}), screened influent (Q_1), oil separator (API) outflow (Q_2), equalization (Q_3), dissolved air flotation (DAF) (Q_4), aeration (Q_5), clarifier 1 (Q_6), clarifier 2 (Q_7), and treated effluent (Q_8). The COD measurements of individual unit processes also appeared in the data reconciliation results of Behnami *et al.*[19] However, further analysis indicated that the reconciled mass flows of COD were calculated from the reconciled flow rates and the measured COD concentrations. Therefore, the variables of COD mass were not defined as key variables in this study.

3.5.3 Mass balance setup. Eight flow balances were set up in Table 7, based on the incidence matrix provided by Behnami *et al.*[19] All the mass balances were considered relevant.

3.5.4 Experimental design results. From the 17 possible additionally measured variables (Q_1, Q_2, Q_3, Q_4, Q_5, Q_6, Q_7, Q_8, Q_9, Q_{10}, Q_{11}, Q_{12}, Q_{13}, Q_{14}, Q_{15}, Q_{16}, and Q_{17}), 2^{17} = 131 072 possible combinations of additionally measured variables were analysed and 2612 solutions were found using the experimental design procedure. The result is a Pareto-front with 10 solutions (Fig. 8, detailed in Table C5b, Appendix C5, ESI†).

Behnami *et al.*[19] used 19 flow measured variables in data reconciliation, 17 of which were considered measured additionally compared to the initially measured flows in the measurement campaign (detailed in Table C5a, Appendix C5, ESI†). This data set satisfied the main goal to identify all key variables. This set also belongs to the Pareto-optimal front

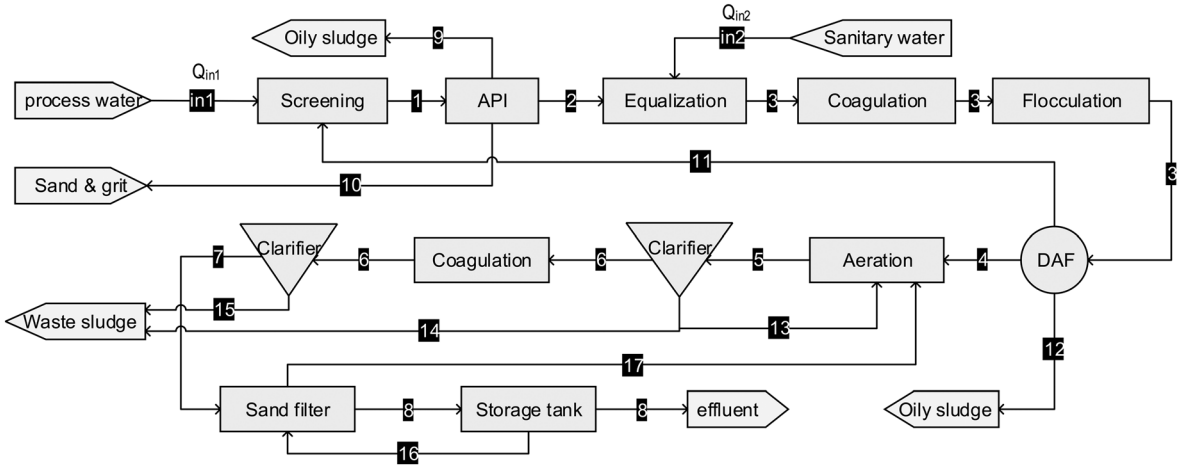

Fig. 7 Flow diagram of the Tabriz petrochemical WWTP, Iran.[19] Measured variables are presented.

obtained using the experimental design procedure. The set of additionally measured variables implemented by Behnami *et al.*[19] provided the highest accuracy (accuracy fv = 1) but with the highest cost (cost fc = 170) (Fig. 8 and detailed in Table C5b, Appendix C5, ESI†). In this case study, the minimum additionally measured variables given by the experimental design procedure would be 7. Therefore, the maximum potential reduction in the number of additionally measured variables could be up to 59%. Overall, a similar conclusion to the case study from Puig *et al.*[22] could be drawn. The set of additional measurements was relevant and allowed the identification of all the key variables, but alternative solutions involving fewer measurements and lower cost were found in this study.

4 Discussion

4.1 Specification of key process variables

Translating the main goal of a study into key variables constitutes the first step in the well-defined, structured experimental design procedure we advocate for Le *et al.*[17] The identification of key variables means that their value can be calculated – by applying data reconciliation – from measured variables through which they are related by mass balances. As a result, two important considerations need to be kept in mind during the key variable selection: the selection of conservative variables and the need for their identifiability.

First, only conservative variables can possibly be identified using data reconciliation and therefore qualify as key variables. Some variables related to the main goal cannot appear in the mass balances because they are not expressed in conservative quantities, so they cannot be put forward as key variables. For example, the mass flow of orthophosphate and total suspended solids in case study 3 (ref. 22) and case study 1 (ref. 20) were non-conservative.

A second important requirement to keep in mind during key variable selection is that all key variables must be

Table 7 Mass balances around groups and individual process units. Variables in bold are key variables

	Process unit	Mass balance	
1	Screening	$\boldsymbol{Q_{in1}} - \boldsymbol{Q_1} + Q_{11}$	Flow (m^3 per day)
2	API	$\boldsymbol{Q_1} - \boldsymbol{Q_2} - Q_9 - Q_{10}$	
3	Equalization	$\boldsymbol{Q_2} + Q_{in2} - \boldsymbol{Q_3}$	
4	DAF	$\boldsymbol{Q_3} - \boldsymbol{Q_4} - Q_{11} - Q_{12}$	
5	Aeration	$\boldsymbol{Q_4} - \boldsymbol{Q_5} + Q_{13} + Q_{17}$	
6	Clarifiers	$\boldsymbol{Q_5} - \boldsymbol{Q_6} - Q_{13} - Q_{14}$	
7	Clarifiers	$\boldsymbol{Q_6} - \boldsymbol{Q_7} - Q_{15}$	
8	Sand filter	$\boldsymbol{Q_7} - \boldsymbol{Q_8} + Q_{16} - Q_{17}$	

Q = total mass flow (density is assumed to be the same for all streams).

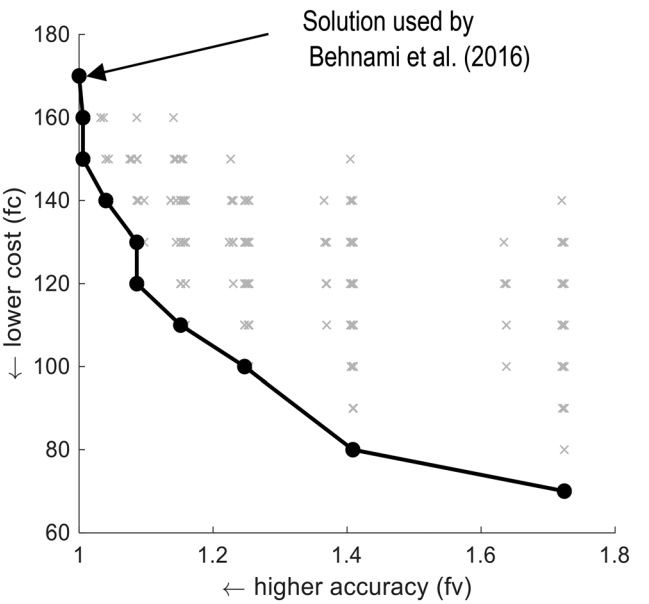

Fig. 8 Pareto optimal solutions for the setup of Behnami *et al.*[19] determined using the experimental procedure of Le *et al.*[17] and expressed in terms of accuracy and costs. The line with the filled circles denotes the Pareto-optimal front. 'x' = a solution.

identifiable for the set of mass balances considered. This is checked through a redundancy analysis of the system of mass balances, which is an integral part (step 4, Fig. B1, Appendix B, ESI†) of the experimental design procedure of Le *et al.*[17] In case one or more key variables are not identifiable, the set of mass balances needs to be reviewed first. In some cases, the problem can be solved by adding mass balances. However, it could also be that some variables related to the main goal, even if they are conservative, cannot be identified because practical constraints make it impossible to close the corresponding mass balances (*e.g.* measurement of the gas phase components of an open tank with a large surface area).

Redundancy analysis is an essential part of experimental design, as it removes the dependent mass balances and checks the presence of all the key variables in the independent mass balances (detailed mathematical procedure can be found in the study by Le *et al.*[17]). This ensures the identifiability of all the key variables. The absence of this analysis might lead to unnecessary mass balances with dependent constraints irrelevant for the identification of the key variables, and associated irrelevant additional measurements. None of the case studies reported in the literature so far performed such redundancy analysis. As a result, irrelevant mass balances were set up in case studies 2 and 4 of Meijer *et al.*[20,21] Subsequently, unnecessary additional measurements were performed.

Besides avoiding the use of mass balances which are irrelevant (not related to key variables) and or redundant (linearly dependent on other mass balances), another point of attention is to take into account the maximum amount available of independent mass balances containing key variables. For instance, in case study 5 of Behnami *et al.*,[19] more mass balances could have been defined and additional associated variables could have been identified. More specifically, in the latter study, only 15 flows were actually reconciled and used for further calculation and process evaluation, while more than 100 variables were additionally measured. These additional measurements were not exploited to their full potential. In case also the mass balances of COD, phosphorus and nitrogen would have been set up, more key variables could have been defined and identified (reconciled) for this case study.

4.2 From more measurements to more information

Increasing the number of measured variables does not always lead to more information in view of data reconciliation. In all case studies, additional measurements and associated mass balances were proposed such that the number of constraints (independent mass balances) was higher than the number of unknown variables, *i.e.* aiming at an overdetermined system. However, the identifiability of a variable depends on how it appears in the set of mass balances. As a result, aiming at an overdetermined system of mass balances does not guarantee the identifiability of all key process variables.[5] It involves the risk of adding measurements without added value (oversized measurement campaigns) while missing out on some critical measurements. For a small number of key variables and mass balances, the identifiability of key variables could be deduced as such. However, for a more extensive set of key variables and mass balances, a clear and straightforward experimental design procedure is essential.

In the studies of Puig *et al.*[22] and Behnami *et al.*,[19] additional measurements were collected for all unknown variables that appear in the set of mass balances. The redundant data sets collected in these case studies corresponded with the most expensive (but most accurate) Pareto-optimal solutions identified with the experimental design procedure in this study. The study of Meijer *et al.*[21] involved a relatively complex set of mass balances, which made it challenging to find the right additionally measured variables without a structured experimental design approach. As a consequence, the additional measurements performed in the latter case study missed two crucial additionally measured variables and not all key variables could be identified.

As a side note, it could be mentioned that the measurement accuracy will also influence their usefulness and added value. In the experimental design procedure in this study, the measurement accuracy is taken into account through their variance, which is incorporated in the objective function. Adding redundant sensors for variables which are already observable may lead to improved precision of the reconciled values or to improved sensor fault isolation, as demonstrated by Villez *et al.*[18] through the concept of 'surplus redundancy'. However, such a procedure was considered beyond the scope of the present study.

Overall, it is clear that more measurements do not necessarily lead to more information. Data gathering should only be done if one knows where to use the data for, *i.e.* once the main goal and key variables have been defined. Rather than measuring more, one needs to measure the right things. Balancing the number of measurements (costs) and the obtained accuracy of identified variables by staying on the Pareto-optimal front will avoid excess costs for additional measurements that do not add information. Besides the measurement costs as such, also overhead and costs associated with data management cannot be neglected – the costs for sensors are just the tip of the iceberg. Digitalisation of the water industry, a topic which has attracted a lot of interest,[25] should therefore never be the goal as such.

4.3 Application of the experimental design procedure to other WWTPs

The potential and general applicability of the experimental design procedure of Le *et al.*[17] was demonstrated in this study through its application to five different case studies. On the one hand, the procedure could be successfully applied to different plant layouts with different main goals. On the other hand, the provided solutions ensured the fulfilment of the main goal (and the identifiability of the key variables).

Additionally, the optimal solutions implied fewer additional measurements (about 40–70% less) compared to the previous approaches from the evaluated case studies.

The application of the design procedure is straightforward. It consists of seven steps, the first three of which require inputs from the user to organize all the collected information in one preformat input file: to translate the main goal(s) into key variables (step 1), to set up mass balances relating key variables to other, measured variables (step 2) and to inventory initially available data (step 3). Step 4 to step 7 are fully automated and are directed in finding the (optimal) solutions for any problem that can be formulated in the first three steps. For all case studies considered in this contribution, these last four steps took at most 10 seconds. This fast implementation makes one effectively and iteratively rework the set of mass balances and recheck input data and/or the key variable definition in case one or more key variables cannot be balanced/estimated.

The experimental design procedure is very flexible in providing alternative additional measurement sets. For a user-defined set of potential additionally measured variables, the experimental design procedure proposes several alternatives, all of which are Pareto-optimal. The user can then select a solution from the Pareto-front based on the available budget and expected accuracy. Note that additional approaches can be used to define the optimization problem in accordance with the main goal of the study. For instance, Villez *et al.*[18] defined the sensor placement procedure as a trade-off between observability and cost in WWTPs, while other studies in the chemical engineering field also considered objectives such as reliability (= low probability of faults), precision or estimability.[26,27] Nevertheless, the Pareto optimal front is considered an excellent option to visualize the solutions.

If problems are expected with the measurement of specific streams, *e.g.* because of safety issues or difficult access, these can be avoided upfront by discarding them from the set of potential additionally measured variables. Application of the experimental design procedure will then indicate whether the discarded variables are essential (in step 4: feasibility evaluation, see Fig. B1, Appendix B, ESI†) and if not, will propose alternative solutions. For example, in case studies 2 and 3, measuring internal recycling flows, which may be problematic, could be avoided by excluding them from the list of potential additionally measured variables.

While this study deals with measurement campaigns for WWTPs, similar experimental design methodologies for application to water distribution networks and sewer systems could be developed in the future.

5 Conclusions

In the framework of digitalization of the water industry, rather than measuring more, one needs to measure the right things in order to obtain more information. This contribution validates a structured procedure for defining 'the right things'. More specifically, the general applicability and added value of the experimental design procedure of Le *et al.*[17] for planning measurement campaigns on WWTPs was demonstrated in this study, leading to the following insights:

– The application of the experimental design procedure was straightforward and could easily be adapted for different WWTP configurations and different main goals.

– Translating the main goal of a study into key variables is essential to find appropriate additionally measured variable sets. The key variables should be conserved quantities and need to appear in the set of mass balances considered for the system under study, in order to be identifiable during future data reconciliation. In three out of the five case studies from the literature applying expert judgement approaches, the main goal was not translated well into specific key variables and thus they were not well identified with additional measurements.

– A redundancy analysis, to check the identifiability of key variables for the considered set of mass balances, is an essential part of the proposed experimental design procedure. The optimal sets of additionally measured variables proposed using the procedure thus guarantee the identifiability of all the key variables through subsequent data reconciliation. This was not always the case in the literature case studies. This showed that more measurements do not necessarily lead to more information.

– Even though adequate additional measurements were proposed using the expert judgement approach, there were often too many measurements. With the structured experimental design procedure, about 40% to 70% fewer measurements were needed.

Data availability

All the data used for this research have been included in the manuscript and the ESI.†

Conflicts of interest

There are no conflicts to declare.

References

1 C. M. Crowe, Data reconciliation - Progress and challenges, *J. Process Control*, 1996, **6**, 89–98, DOI: **10.1016/0959-1524(96)00012-1**.
2 D. B. Özyurt and R. W. Pike, Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes, *Comput. Chem. Eng.*, 2004, **28**, 381–402, DOI: **10.1016/j.compchemeng.2003.07.001**.
3 F. Madron, V. Veverka and V. Vanecek, Statistical-Analysis of Material Balance of a Chemical Reactor, *AIChE J.*, 1977, **23**, 482–486, DOI: **10.1002/aic.690230412**.
4 F. Madron and V. Veverka, Optimal selection of measuring points in complex plants by linear models, *AIChE J.*, 1992, **38**, 227–236, DOI: **10.1002/aic.690380208**.
5 R. T. J. M. van der Heijden, J. J. Heijnen, C. Hellinga, B. Romein and K. C. A. M. Luyben, Linear Constraint Relations in Biochemical Reaction Systems .1. Classification of the

Calculability and the Balanceability of Conversion Rates, *Biotechnol. Bioeng.*, 1994, **43**, 3–10, DOI: **10.1002/bit.260430103**.

6 S. Lee, S. Rao, M. J. Kim, I. J. Esfahani and C. K. Yoo, Assessment of environmental data quality and its effect on modelling error of full-scale plants with a closed-loop mass balancing, *Environ. Technol.*, 2015, **36**, 3253–3261, DOI: **10.1080/09593330.2015.1058859**.

7 V. Monje, H. Junicke, D. J. Batstone, K. Kjellberg, K. V. Gernaey and X. Flores-Alsina, Prediction of mass and volumetric flows in a full-scale industrial waste treatment plant, *Chem. Eng. J.*, 2022, **445**, 136774, DOI: **10.1016/j.cej.2022.136774**.

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

9 A. Spindler, Structural redundancy of data from wastewater treatment systems. Determination of individual balance equations, *Water Res.*, 2014, **57**, 193–201, DOI: **10.1016/j.watres.2014.03.042**.

10 K. Villez, P. A. Vanrolleghem and L. Corominas, Optimal flow sensor placement on wastewater treatment plants, *Water Res.*, 2016, **101**, 75–83, DOI: **10.1016/j.watres.2016.05.068**.

11 L. Benedetti, G. Dirckx, D. Bixio, C. Thoeye and P. A. Vanrolleghem, Substance flow analysis of the wastewater collection and treatment system, *Urban Water J.*, 2006, **3**, 33–42, DOI: **10.1080/15730620600578694**.

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

13 M. Carnero, J. L. Hernández and M. Sánchez, Optimal Sensor Location in Chemical Plants Using the Estimation of Distribution Algorithms, *Ind. Eng. Chem. Res.*, 2018, **57**, 12149–12164, DOI: **10.1021/acs.iecr.8b01680**.

14 J. Wang, Z. Wang, X. Ma, A. Smith, F. Gu and C. Zhang, *et al.*, Locating Sensors in Large-Scale Engineering Systems for Fault Isolation Based on Fault Feature Reduction, *J. Franklin Inst.*, 2020, **357**, 8181–8202, DOI: **10.1016/j.jfranklin.2020.05.037**.

15 A. Soldevila, J. Blesa, S. Tornil-Sin, R. M. Fernandez-Canti and V. Puig, Sensor placement for classifier-based leak localization in water distribution networks using hybrid feature selection, *Comput. Chem. Eng.*, 2018, **108**, 152–162, DOI: **10.1016/j.compchemeng.2017.09.002**.

16 C. Winter, V. R. Palleti, D. Worm and R. Kooij, Optimal placement of imperfect water quality sensors in water distribution networks, *Comput. Chem. Eng.*, 2019, **121**, 200–211, DOI: **10.1016/j.compchemeng.2018.10.021**.

17 Q. H. Le, P. J. T. Verheijen, M. C. M. van Loosdrecht and E. I. P. Volcke, Experimental design for evaluating WWTP data by linear mass balances, *Water Res.*, 2018, **142**, 415–425, DOI: **10.1016/j.watres.2018.05.026**.

18 K. Villez, P. A. Vanrolleghem and L. Corominas, A general-purpose method for Pareto optimal placement of flow rate and concentration sensors in networked systems – With application to wastewater treatment plants, *Comput. Chem. Eng.*, 2020, 139, DOI: **10.1016/j.compchemeng.2020.106880**.

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

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

21 S. C. F. Meijer, R. N. A. van Kempen and K. J. Appeldoorn, Plant upgrade using big-data and reconciliation techniques, in *Applications of Activated Sludge Models*, ed. D. Brdjanovic, S. C. F. Meijer, C. M. Lopez-Vazquez, C. M. Hooijmans and M. C. M. van Loosdrecht, IWA Publishing, 2015, pp. 357–410.

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

23 Q. H. Le, P. J. T. Verheijen, M. C. M. van Loosdrecht and E. I. P. Volcke, Application of data reconciliation to a dynamically operated wastewater treatment process with off-gas measurements, *Environ. Sci.*, 2022, **8**, 2114–2125, DOI: **10.1039/d2ew00006g**.

24 M. C. M. Van Loosdrecht, F. A. Brandse and A. C. De Vries, Upgrading of waste water treatment processes for integrated nutrient removal the BCFS® process, *Water Sci. Technol.*, 1998, **37**, 209–217, DOI: **10.1016/S0273-1223(98)00290-X**.

25 S. Vairavamoorthy, *The rise of digital water - The Source*, 2018.

26 M. Bhushan and R. Rengaswamy, Comprehensive design of a sensor network for chemical plants based on various diagnosability and reliability criteria. 2, Applications, *Ind. Eng. Chem. Res.*, 2002, **41**, 1840–1860, DOI: **10.1021/ie010437v**.

27 M. Carnero, J. Hernández and M. Sánchez, A new metaheuristic based approach for the design of sensor networks, *Comput. Chem. Eng.*, 2013, **55**, 83–96, DOI: **10.1016/j.compchemeng.2013.04.007**.