Comparative analysis of select techniques and metrics for data reconciliation in smart energy distribution network

Jeyanthi Ramasamy, Sriram Devanathan and Dhanalakshmi Jayaraman

ABSTRACT

Reliability of each state of process in many chemical process industries largely relies upon water and vitality supplies. In this way, there is great necessity to have an improved and controlled smart energy distribution network (SEDN) in industries. In SEDNs, sensor information related to flow control and optimization serves as a basis for modelling of energy management systems. Therefore, it is important to ensure that sensor data are accurate and precise. However, they are affected by random noise and measurement biases, which compromise the quality of measurements. Data Reconciliation (DR) is one such approach popularly used in industries to reduce the adverse impact of random errors present in pipe flow measurements. In this study, Python-based simulations of weighted least squares (WLS) and principal component analysis (PCA) based DR techniques are implemented on the selected flow streams of SEDN, and reconciled estimates are obtained. The results show that Root Mean Square Error (RMSE) is the best performance metric since it is more sensitive to small changes in the measurement values and the reconciled estimates. Further, it is observed that PCA-DR performs better than WLS-DR in reducing the random error (and thereby achieving greater precision of measured values).

Key words | data reconciliation, performance metrics, principal component analysis (PCA), smart energy management network

Jeyanthi Ramasamy (corresponding author)
Department of Electronics and Communication Engineering, Amrita School of Engineering,
Amrita Vishwa Vidyapeetham,
Bengaluru,
India
E-mail: *r_jeyanthi@blr.amrita.edu*

Sriram Devanathan
Center of Excellence in Advanced Materials & Green Technologies, Department of Chemical Engineering and Materials Science, Amrita School of Engineering,
Amrita Vishwa Vidyapeetham,
Coimbatore,
India

Dhanalakshmi Jayaraman
Department of Chemical Engineering,
SSN college of Engineering,
Chennai,
India

HIGHLIGHTS

- Application of data reconciliation (DR) techniques to treat random errors present in flow sensor data used by water distribution networks.
- Selection of best performing metric to evaluate data reconciliation (DR) techniques.
- Analyze the performance of selected DR techniques for small and large scale networks using Python-based simulation.

INTRODUCTION

In most chemical industries, utilities such as water and energy play an important role. A sensor-based smart energy distribution network (SEDN) is required to monitor the consumption of these supplies. SEDN can aid in providing better process quality, more efficient operation, more accurate forecasting of supply and demand. SEDN operation is usually supported by Supervisory Control and Data Acquisition (SCADA) systems (Park & Jung 2014; Quevedo *et al.* 2014; Krčová 2016). Therefore, measurement of process flow variables is an essential part of this process. The precision of the measurements is very important, without which modelling and analysis can be misleading.

doi: 10.2166/ws.2020.314

However, usually measured variables are contaminated by fixed and random errors (Schönenberger 2015; Tran *et al.* 2019). These random errors creep into measurements from various sources like high frequency pickups, low resolution, and signal converters (Câmara *et al.* 2017). Data reconciliation (DR) is an approach usually applied to treat random errors present in a measured variable under a constrained process environment. The most important difference between DR and other signal processing techniques is that DR uses process constraints; that is, mass and energy balances of the process network, while the latter do not.

At times it is not possible to measure all the process variables in a process due to practical difficulty. In such situations, unmeasured variables could be estimated through soft sensors or solving DR problem (Miao *et al.* 2009; Rieger *et al.* 2010; Quevedo *et al.* 2014; Narasimhan & Bhatt 2015; Xu *et al.* 2020). DR is a simple approach and it has even become an integral part of software packages like ASPEN, VALI, VisualMesa, SimSci DATACON, and Sigmafine (Narasimhan & Bhatt 2015; Camara *et al.* 2017). In the data reconciliation landscape, the techniques include Weighted Least Squared Data Reconciliation (WLS-DR), Quasi-Weighted Least Squared Data Reconciliation (QWLS-DR), Robust-DR, and a few more recent techniques, for dealing with random errors. (Rieger *et al.* 2010; Zhang *et al.* 2010; Fuente *et al.* 2015; Lin *et al.* 2019; Xie *et al.* 2019).

Principal Component Analysis (PCA) is a multivariate data processing technique extensively used for dimension reduction where data on a large number of variables are available. It is also widely used in behavioural modelling of large water management systems, monitoring of water distribution including leakage, abnormal use of water, illegal connections, process monitoring for multi input-multi output (MIMO) processes, soft sensor modelling, data reconciliation (DR), and gross error detection (GED). In Helness *et al.* (2019), Varshith *et al.* (2019), Bhattacharyya *et al.* (2017), Fuente *et al.* (2015), Narasimhan & Bhatt (2015), Narasimhan & Shah (2008), Park & Jung (2014), the use of PCA in data reconciliation and gross error detection techniques is illustrated.

In order to prove that the DR techniques are actually accomplishing the task of improving the precision of measured data, performance metrics are needed. In Spuler *et al.* (2015), various performance metrics were explained to evaluate regression methods applied for decoding neural signals. From this, Correlation Coefficient (CC), Global deviation (GD), Signal to Noise Ratio (SNR), and Root Mean Square Error (RMSE) are identified as the most suitable metrics to evaluate DR techniques. The other performance measures are Relative Error Reduction (RER), Measurement Relative Error (MRE), and Reconciled Relative Reduction Error (RRE), which are explained in Zhang *et al.* (2010). In order to find the best metric, factors that lead to deterioration of data should be looked at, and that is discussed in the following section of this study.

In this paper, integrated water supply networks are considered whose pipe flow streams are assumed to be contaminated by gaussian noise. Two DR techniques, WLS-DR and PCA-based DR, are applied to measurements, and reconciled estimates are obtained. The selected metrics are applied to evaluate the performance of DR techniques and the best metric is then found. Further, the same is implemented and evaluated for other datasets that have different variances and serially correlated errors.

DATA RECONCILIATION (DR) TECHNIQUES

Data reconciliation is generally applied to reduce the effect of random errors present in the process variables. The reconciled estimates are obtained from the information contained in the models. The objective function of a generalised DR problem (Zhang *et al.* 2010; Korpela *et al.* 2016; Syed *et al.* 2016; Srinivasan *et al.* 2017; Xie *et al.* 2019) is formulated as:

$$\min f(M,\ \hat{M}) = \sum_{i=1}^{n}\sum_{j=1}^{N}\rho\left(\frac{m_{ij}-\hat{m}_{ij}}{\sigma_i}\right), \text{ subject to } \mathrm{F}(\hat{M}) = 0 \tag{1}$$

where, M is the raw measurement matrix, $\hat{M}$ is the reconciled estimate of M, ρ is the cost function, m_j is the j^{th} measurement of i^{th} variable, $\hat{m}_{ij}$ is the j^{th} reconciled estimate of i^{th} variable, n is the number of process variables, N is the sample size, σ is the standard deviation of process variable and $\mathrm{F}(\hat{M})$ is the process constraint.

Optimised measurements of a system can be estimated which have lesser effect of random noise. Given below are

certain prerequisites which are essential in order to apply DR to a dataset.

i. The process constraints, which consist of material and energy balance equations, should be defined.
ii. The set of measured process variables must be specified and the inaccuracies in these measurements must be specified in terms of the associated variances and covariance.

Application of DR to linear steady state processes is discussed below. Consider *a(j)* to be the true measurement vector of a steady state system for each sample *j* where *j* = *1,2,3,…, N*. For a steady state process, equality constraints are derived as,

$$\tilde{C}a(j) = 0 \tag{2}$$

where $\tilde{C}$ is an incidence matrix of the process; $a(j)$ is the j^{th} true measurement vector and it can be termed as $[a_1, a_2, \ldots, a_n]^T$, N is the sample size, and n is the number of process variables. The process constraint matrix is derived from the incidence matrix $\tilde{C}$.

Let the measurement of process variables be *m(j)*, then the generalised measurement model (Narasimhan & Bhatt 2015) is represented in Equation (3):

$$m(j) = a(j) + r(j) + \delta(j) \tag{3}$$

where r(j) is a random error vector and $\delta(j)$ is the measurement bias.

The measurement model shown in Equation (3) represents a realistic model since in practice, over a long period, measurements may change only due to random error. Here, the true value vector $a(j)$, variance of measurements (σ^2) and process incidence matrix ($\tilde{C}$) are considered (realistic assumption) to be fixed.

There are a few other assumptions about random errors considered.

For independently identically distributed (i.i.d.) data,

i. random errors are normally distributed, i.e. $r(j) \sim N(0, \sigma^2)$,
ii. process variables and errors are not correlated i.e., E[a(j) r(j)] = 0 and
iii. errors are not serially correlated i.e., E[r(j)r(j − 1)] = 0.

Sometimes, data are serially correlated due to process dynamics, recycling, adding controller loop, etc. Serial correlation may affect the performance of DR techniques. Hence, performance evaluation of DR techniques for identically distributed but non-independent (non-*i.i.d.*) measurements is considered here (Lin *et al.* 2019; Jeyanthi & Devanathan 2020).

Assumptions considered for non-*i.i.d* are:

i. random errors are normally distributed, i.e. $r(j) \sim N(0, \sigma^2)$,
ii. process variables and errors are not correlated i.e., E[a(j) r(j)] = 0 and
iii. errors are serially correlated i.e., E[r(j)r(j − 1)] ≠0.

In non-*i.i.d.* data, the Auto Regressive Moving Average (ARMA) model is considered for the random error and its structure is expressed as in Equation (4):

$$\begin{aligned} r(j) = {} & \varphi_1 r(j-1) + \varphi_2 r(j-2) + \cdots\cdots\cdots\cdots \\ & + \varphi_p r(j-p) + \omega(j) - \theta_1\omega(j-1) - \quad \theta_1\omega(j-2) \\ & - \cdots\cdots\cdots\cdots\cdots - \theta_q\omega(j-q) \end{aligned} \tag{4}$$

where, φ is an auto regressive parameter, θ is a moving average parameter and ω is the white noise; p and q are the order of the model.

The weighted least squared-DR

The conventional approach to implement DR is the weighted least square data reconciliation (WLS-DR), which is used to get higher precision estimates of process variables from measurements which have noise added. The basic idea of this approach is to minimise the residuals (mass and energy imbalances resulting from measurement error) in the nodes (containers in the process where multiple flows converge) so that the estimates are much closer to true values. The generalised problem of steady state DR problem (Zhang *et al.* 2010; Fuente *et al.* 2015; Valle *et al.* 2018; Lin *et al.* 2019; Xie *et al.* 2019) is explained in the following Equation (5):

$$\min(M,\ A,\ R) = \sum_{i=1}^{n}\sum_{j=1}^{N} w_{ij}(m_{ij} - \hat{m}_{ij})^2 \tag{5}$$

subject to $f(A) = 0$ and $g(A) \leq 0$, where, M - raw measurement, A- reconciled estimate of M, w_{ij} - the j^{th} weight of the i^{th} measurement variable, and $f(a)$ and $g(a)$ represent the constraints of the process. The optimum value of w_{ij} will provide good reconciled estimates for m_{ij}. For a known covariance matrix (Λ), the reconciled estimates of the measurement of variables $\hat{a}(j)$ can be obtained by:

$$\hat{a}(j) = \Lambda C^T (C\Lambda C^T)^{-1} Cm(j) \tag{6}$$

where, $\hat{a}$ is reconciled estimates for $m(j)$, Λ is the error covariance matrix and C is the constraint matrix.

The obtained estimates are normally distributed and satisfy the process constraints.

Principal component analysis (PCA) based data reconciliation

PCA is a statistical technique that is generally used to reduce the dimensionality of data. PCA can also be used to identify linear relationships between variables. It identifies fewer uncorrelated variables, called principal components (PCs), from a large set of data. The goal of principal components is to explain the maximum amount of variance with the fewest PCs. Narasimhan & Bhatt (2015) have described an approach for applying PCA-based DR, a recent technique to obtain reconciled estimates. For a large data set, PCA-DR proves to be effective. This can be deployed when the error covariance matrix is known. The data matrix M is transformed as in Equation (7)

$$\mathrm{M_s} = \mathrm{\beth}^{-1} M = (\mathrm{\beth}^{-1} A + \mathrm{\beth}^{-1} R) \tag{7}$$

where $\beth$ is a decomposition of error covariance matrix (Λ) of the process network.

$$\Lambda = \beth \beth^{\mathrm{T}} \tag{8}$$

Therefore the covariance matrix of M_s can be calculated by Equation (9)

$$\Lambda_{\mathrm{M_s}} = \frac{1}{\mathrm{N}} \mathrm{M_s} \mathrm{M_s^T} \tag{9}$$

The positive square root of the eigenvalues of $\Lambda_{\mathrm{M_s}}$ are termed as singular values of M_s. The first highest γ singular values are positive and are equal to the rank of M_s while the remaining n- γ singular values are small and used to define the dependent variables in the process. SCREE plot (Narasimhan & Bhatt 2015) can be used to select the largest eigen values (γ) $\Lambda_{\mathrm{M_s}}$.

The singular value analysis (SVA) is a technique used generally in control loop selection. The highest singular values denote the principal components (PCs) of the process variables. The PCs explain the most influential variables of the process network. The PCs of the process are extracted by applying Singular Value Decomposition (SVD) of the transformed matrix

$$\mathrm{SVD(M_s)} = U\, S\, V^T \tag{10}$$

where, the columns of U are the input singular vectors and the columns of V are the output column vectors. S is a diagonal matrix of non-zero γ singular values in order from largest to smallest. U and V are also unitary matrices such that $UU^T = I$ and $VV^T = I$.

The decomposed matrix M_s can be expressed by vector containing highest singular values γ and the vector containing smallest singular values n-γ:

$$M_s = \sqrt{N} U_{1s} S_{1s} V_{1s}^T + \sqrt{N} U_{2s} S_{2s} V_{2s}^T \tag{11}$$

The reconciled estimates of M are derived from the first part of M_s, $\sqrt{N} U_{1s} S_{1s} V_{1s}^T$ and calculated as $\hat{A} = \beth \hat{M}_s$. The second part of $\mathrm{M_s}$ expresses the interaction among variables. So the identified constraint matrix of the process can be derived as $\hat{C} = U_{2s}^T \beth^{-1}$. It may be noted that the PCA-DR estimates do not satisfy the original constraints, but rather satisfy the identified constraint matrix $\hat{C}$.

Performance metrics

It is important to evaluate any technique so as to ensure that it is reliable even for a large set of data. The various performance metrics must capture the error properties like bias, scaling, and other types of errors present in the

measurement. They should be sensitive enough to capture these error properties and perform accurately. Among various performance metrics, Correlation Coefficient (CC), Root Mean Square Error (RMSE), Global Deviation (GD), Signal to Noise Ratio (SNR), and Relative Error Reduction (RER) are selected (Chai & Draxler 2014; Narasimhan & Bhatt 2015; Spuler *et al.* 2015; Xie *et al.* 2019). In this paper, two cases are taken for the study. The first is where only random errors (both *i.i.d* and non-*i.i.d*) are present in the measurements, and the second is where both random error and measurement bias are present in measurements.

The performance metrics are explained below.

Root mean square error (RMSE)

This is a measure of the square root of differences between estimated values and measured values. In short, it is a measure of the precision achieved by each DR technique in reducing the errors of different models for a particular dataset. It is always positive and a value of '0' would indicate a perfect fit to the data (i.e. 100% elimination of errors). RMSE is commonly used in climatology, soft sensor modelling, forecasting, and regression analysis to verify experimental results.

$$RMSE_i = \sqrt{\frac{\sum_{j=1}^{N} (\mathrm{m_{ij}} - \mathrm{a}_{ij})^2}{N}} \tag{12}$$

Global deviation (GD)

This is a measure of average squared difference between true and measured values.

$$\mathrm{GD_i} = \left(\frac{\sum_{\mathrm{j}=1}^{\mathrm{N}} (\mathrm{m_{ij}} - a_{ij})}{\mathrm{N}} \right)^2 \tag{13}$$

Correlation coefficient (CC)

The Pearson correlation coefficient is a statistical measure that calculates the strength and direction of the linear relationship between two variables. The value of CC is between −1 and 1. It cannot capture non-linear relationships between two variables.

$$\mathrm{CC_i} = \frac{\sum_{\mathrm{j}=1}^{\mathrm{N}} (\mathrm{m_{ij}} - \bar{m}_{ij})(\mathrm{a_{ij}} - \bar{a}_{ij})}{\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{N}} (\mathrm{m_{ij}} - \bar{m}_{ij})^2 \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{a_{ij}} - \bar{a}_{ij})^2}} \tag{14}$$

Signal to noise ratio (SNR)

With respect to statistics, SNR is defined as a measure of the ratio between variance of mean deviation from the true value of the measured variable to the variance of the measured variable:

$$\mathrm{SNR_i} = \frac{\mathrm{var}(\mathrm{m_{ij}} - a_{ij})}{\mathrm{var}(\mathrm{m_i})} \tag{15}$$

Relative error reduction (RER)

RER is another measure used to evaluate the performance of reconciliation techniques. This measure is the ratio of relative errors between raw measurement and reconciled estimates. Higher value of RER indicates a better reconciliation technique:

$$RER_i = \frac{\sum_i (MR_{ei} - RR_{ei})}{MR_{ei}} \text{ and,}$$
$$MR_{ei} = \frac{|a_{ij} - m_{ij}|}{a_{ij}},\ RR_{ei} = \frac{|a_{ij} - \hat{m}_{ij}|}{a_{ij}} \tag{16}$$

where, m is the measurements, a is the true value of the measurement, $\bar{m}$ is the mean of measurement variable I and $\hat{m}$ is the reconciled estimate. For raw measurements, RER is considered as zero.

SIMULATION

To analyse the performance of DR techniques, two benchmark process systems, a small scale recycle network and a large scale process network, have been chosen for the study. The selection of the benchmark systems (Valle *et al.* 2018; Varshith *et al.* 2019; Jeyanthi & Devanathan 2020) is based on the number of variables and interacting

nodes in the process. This would lend credence to the performance evaluation of the techniques included in the study.

Example 1

A small recycle network (Xie *et al.* 2019; Jeyanthi & Devanathan 2020) shown in Figure 1 consists of five interacting nodes with seven flow variables (F1, F2, F3, F4, F5, F6, and F7). The process constraints are linear and estimated from Equation (17).

Depending on interdependencies between nodes and direction of flow, base values are assumed and true values are generated. The true value vector ($a(j)$) for all variables is derived from the base values, $b = [10\ 30\ 30\ 10\ 20\ 10\ 10]^T$, process dynamics of each flow variable with variance of 0.01, and assuming the leaks and the losses across the nodes are null. The mass balance equation at each node is defined by Equation (17),

$$\text{Input flow variable of ith node} - \text{Output flow variable of 'i' th node} = 0 \tag{17}$$

The incidence matrix for this process is shown in Equation (18)

$$C = \begin{bmatrix} 1 & -1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{18}$$

Input and output flow variables are denoted as '1' and '-1' respectively. The process constraint matrix is derived from the incidence matrix by removing node '5', which has non-recycled variables in the network.

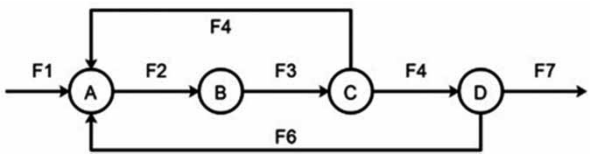

Figure 1 | Recycle network.

In order to evaluate the performance of each technique, a few flow variables with specific magnitudes are identified. The performance indices are calculated as explained in the previous section. The reconciled estimates are obtained using corresponding DR techniques as explained above, and the results are compared with raw data. The performance index calculation procedure is explained as follows:

Step 1: Obtaining raw measurement
Step 2: Applying DR technique
Step 3: Calculate reconciled estimates ($\hat{A}$)
Step 4: Calculate Performance Index

The performance of the recycled network is shown in Table 1. For all variables F1 to F7, SNR has not shown much variation amongst DR techniques. CC and GD have no significant index to show the performance improvement of DR techniques. RER has prominent changes for WLS and PCA-DR techniques, but it has negative values for non-performing reconciled estimates. RMSE has significant changes in different DR techniques. Also, it has a non-negative index. From the perception, it is obvious that RMSE is the best measurement to assess the performance of DR techniques. PCA-DR is sensitive to the magnitude and performs poorly for variables having higher base values F2, F3, and F5. WLS-DR performs well for feedback variables F4 and F6, and poorly for other variables.

Example 2

The large process network (Varshith *et al.* 2019) shown in Figure 2 consists of 11 nodes representing the balance equations and 28 flow variables. The base value of each variable is referred as in Table 2. The constraint matrix for this process is calculated as in Equation (17).

Figure 3 shows the performance indices of selected performance metrics of flow variables F7, F16, F22, and F27. It is observed that in Figure 3(a), RMSE is the best metric for capturing the minute changes in noise present in the data, followed by RER and CC. SNR and GD remain constant throughout, proving to be of no valid significance in evaluating the different data sets. As in the case of variables F16 and F22 shown in Figure 3(b) and 3(c) respectively, SNR and GD remain constant here as well.

Table 1 | Performance of DR techniques for recycle network

		Performance Metrics				
Flow variable	**DR technique**	**SNR**	**CC**	**GD**	**RER**	**RMSE**
1	Raw data	1.0107	0.0000	0.0001	0.0000	1.0065
	WLS-DR	1.0235	0.0000	0.0000	0.3288	0.6753
	PCA-DR	1.0482	0.0000	0.0003	0.5280	0.4745
2	Raw data	1.0107	0.0000	0.0001	0.0000	1.0065
	WLS-DR	1.0062	0.0000	0.0001	−0.3301	1.3390
	PCA-DR	1.0057	0.0000	0.0007	−0.3810	1.3904
3	Raw data	1.0107	0.0000	0.0001	0.0000	1.0065
	WLS-DR	1.0062	0.0000	0.0001	−0.3301	1.3390
	PCA-DR	1.0057	0.0000	0.0007	−0.3810	1.3904
4	Raw data	1.0107	0.0000	0.0001	0.0000	1.0065
	WLS-DR	1.0919	0.0000	0.0000	0.6529	0.3487
	PCA-DR	1.0482	0.0000	0.0003	0.5280	0.4745
5	Raw data	1.0107	0.0000	0.0001	0.0000	1.0065
	WLS-DR	1.0107	0.0000	0.0001	0.0000	1.0065
	PCA-DR	1.0125	0.0000	0.0000	0.0755	0.9305
6	Raw data	1.0107	0.0000	0.0001	0.0000	1.0065
	WLS-DR	1.0919	0.0000	0.0000	0.6529	0.3487
	PCA-DR	1.0482	0.0000	0.0003	0.5280	0.4745
7	Raw data	1.0107	0.0000	0.0001	0.0000	1.0065
	WLS-DR	1.0235	0.0000	0.0000	0.3288	0.6753
	PCA-DR	1.0482	0.0000	0.0003	0.5280	0.4745

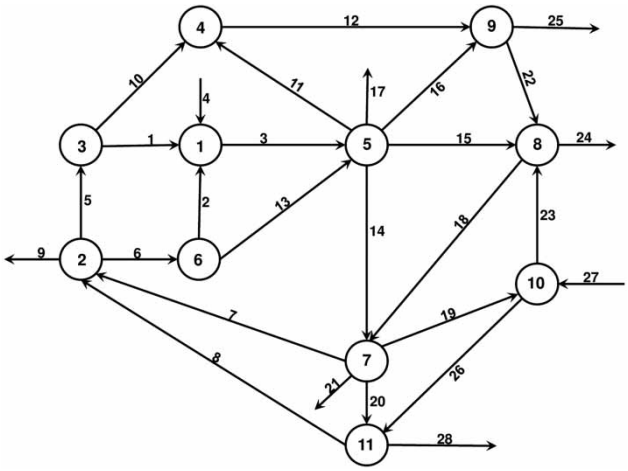

Figure 2 | Large process network.

In contrast, there is a variation in the evaluation outcome of RMSE, RER, and CC. RMSE and RER display improvement for PCA-DR data and CC does not show any improvement for WLS-DR data. The variable F27 shown in Figure 3(d) shows the variation for RMSE, RER, and CC. Since RMSE is majorly consistent and is sensitive to small changes in noise, it is being used as the main evaluation metric. RER almost complements RMSE, but is less sensitive at times.

To analyse the performance of PCA-DR and WLS-DR on the process data, flow variables F16-F1-F28-F5-F3-F18-F27 have been considered in the order of increasing base magnitude. Figure 4(a) shows the performance of the DR techniques when $\Sigma = \mathrm{I}$. The results obtained show that WLS-DR estimates are consistent and are base magnitude independent. The estimates obtained on performing DR are maximum likelihood estimates and they are more accurate when compared to the raw process measurement. PCA-DR performs linearly as the magnitude increases. For high magnitudes, its performance is less accurate when compared to the estimates of WLS-DR. Thus, we can effectively say that the performance of PCA-DR decreases as the base magnitude of the flow variables increases.

Table 2 | Base values for flow variables

Flow Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Base values	10	10	30	10	20	20	30	20	10	10	10	20	10	10
Flow variable	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Base values	10	5	5	40	5	5	10	20	20	10	5	30	45	15

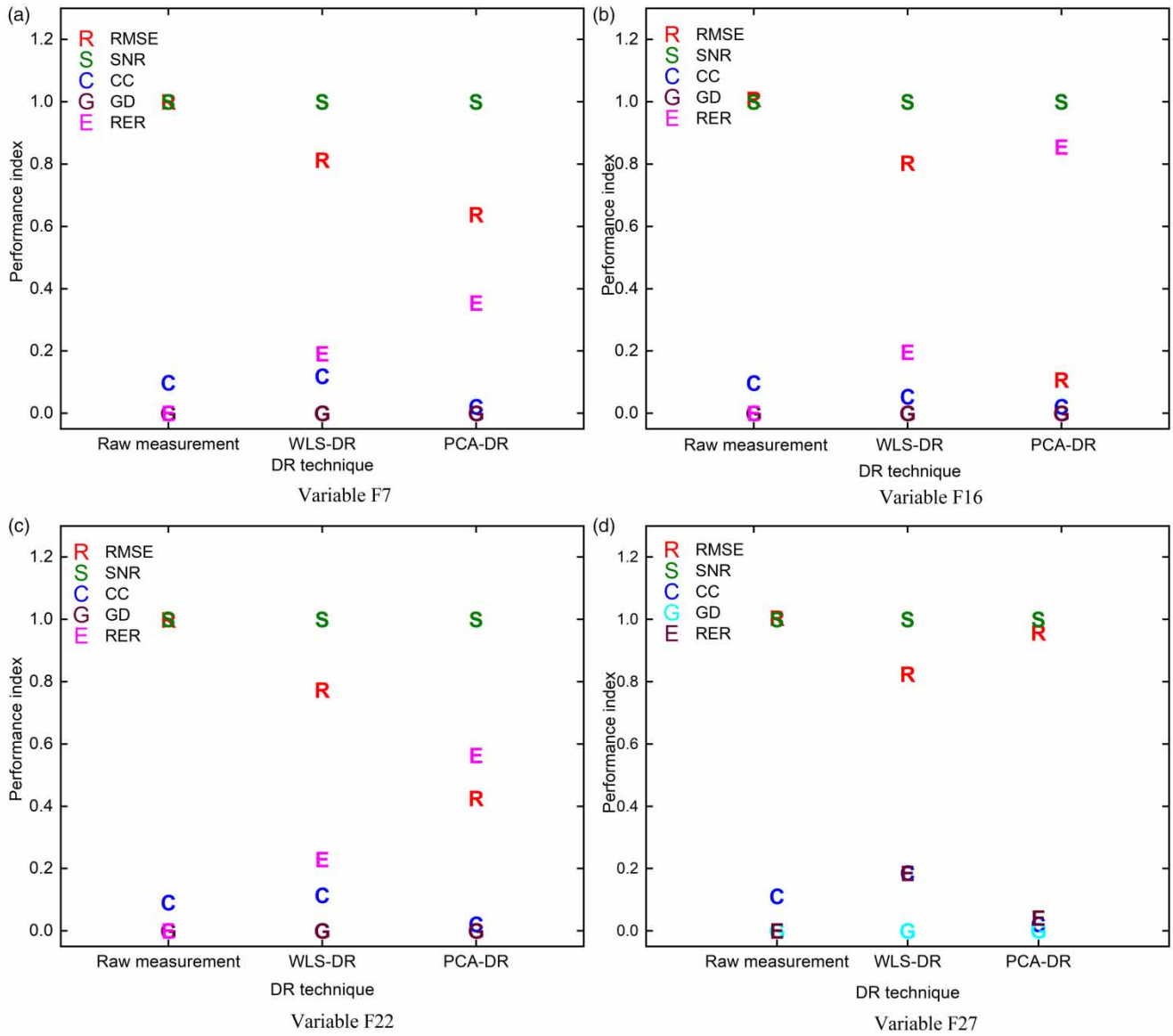

Figure 3 | Performance metrics of flow variable (*i.i.d*).

Figure 4(b) compares the performance of WLS-DR and PCA-DR for different variances in data for variable F16. It is seen that as the variance increases, the performance of WLS-DR decreases. PCA-DR performs well when compared to WLS-DR. This shows that increase in variance does not affect

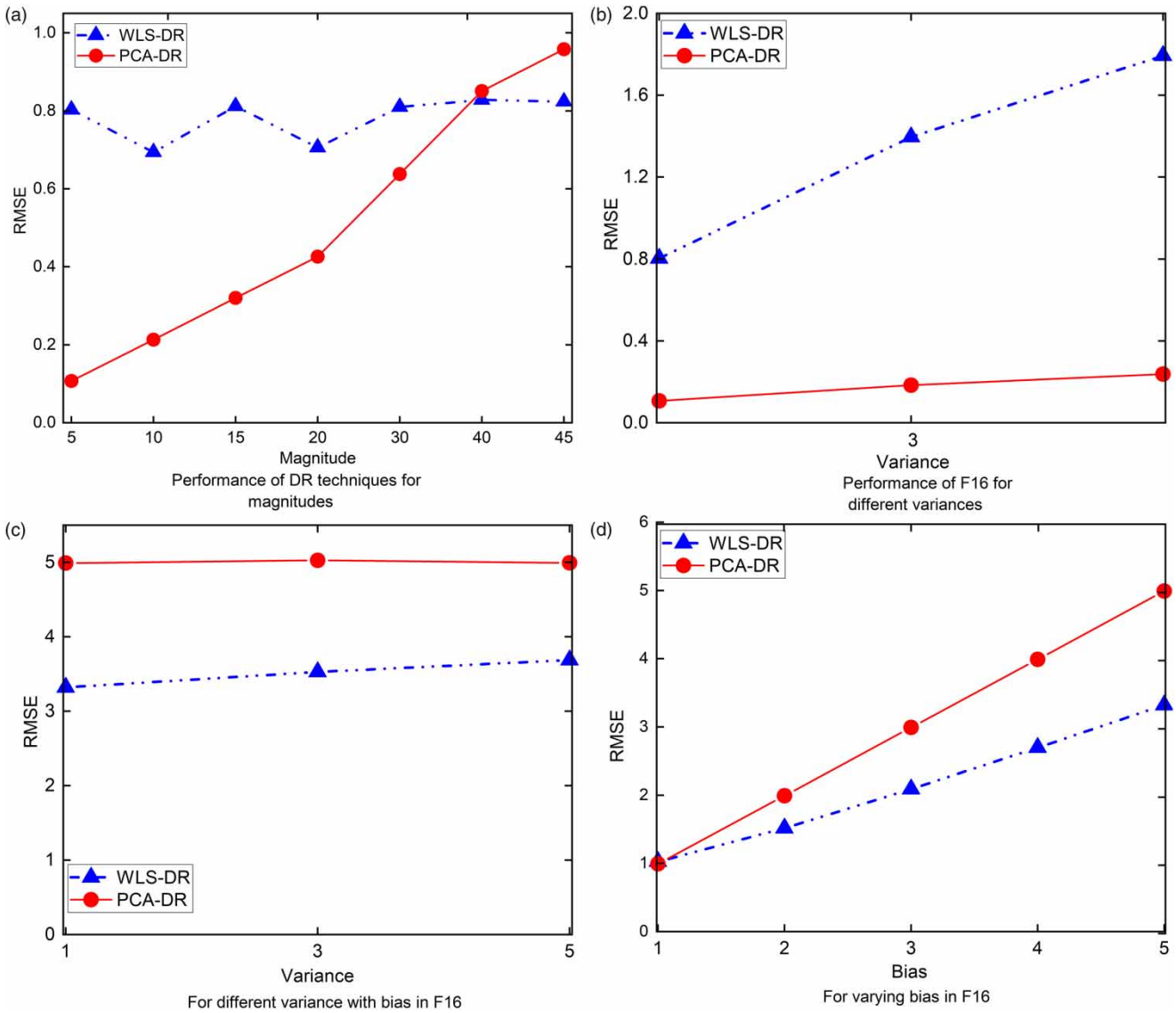

Figure 4 | Performance of DR techniques for different scenario.

the PCA-DR performance as it did for increase in magnitude.

Figure 4(c) shows the performance of DR techniques for different variance when bias (δ) of 5σ present in variable F16. As variance increases, RMSE of WLS-DR estimates also increases. PCA-DR estimates vary slightly and its RMSE remain around the gross error value. This indicates that PCA-DR can be used for detecting the presence of bias even when there are situations where variance changes.

Figure 4(d) shows that as the value of gross error increases, RMSE of PCA-DR is varying linearly with bias, and its magnitude is equal to that of the gross error. The RMSE of WLS-DR is also varying linearly with gross error, but not in a way that is equal to the magnitude of the gross error. The PCA-DR can be combined with gross error detection techniques to identify gross errors present in measurements.

Measurements are usually contaminated by random and gross errors. So, it is important to analyse the effect of gross

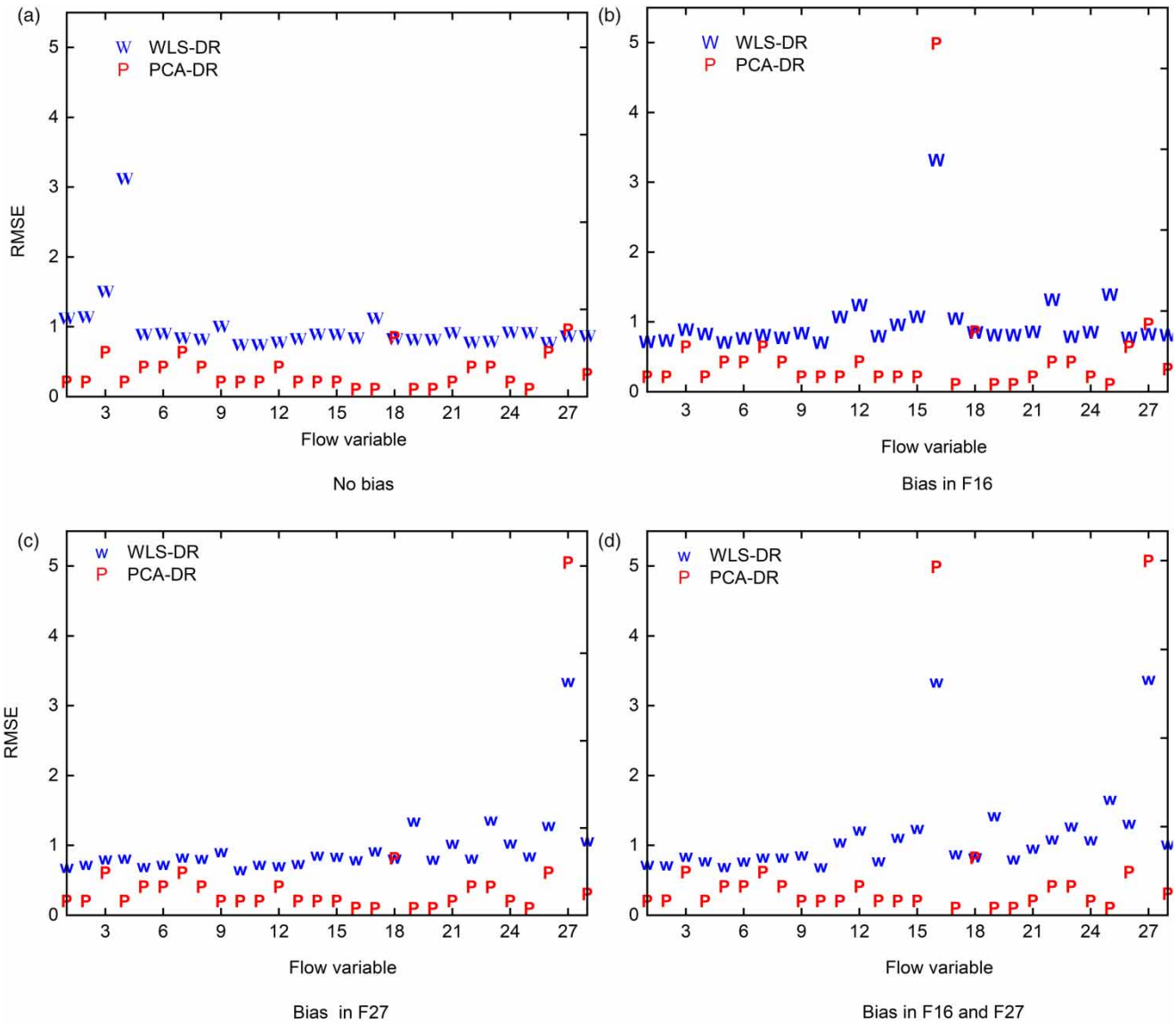

Figure 5 | Performance of DR techniques for the bias ($\delta = 5$ unit).

errors on the performance of DR techniques. Figure 5 shows two scenarios based on the presence of gross error (bias) of magnitude 5σ, in 5(a) variables F16 and F27 and 5(b) F16 and F27. The DR techniques didn't increase the accuracy of estimates. The WLS-DR distributed the gross error, and PCA-DR did not perform at all. The PCA-DR, as seen previously, depends on magnitude of variable in *i.i.d.* data; in the presence of bias, irrespective of base value, the result was the value of amount of gross error present.

To analyse the performance of the DR techniques on non-*i.i.d.* data as well, data with Auto Regressive Moving Average (ARMA) noise was simulated. The ARMA(1, 1) had φ_1 & θ_1 values as 0.4 and 0.2, while the ARMA(2, 2) had φ_1, φ_2, and θ_1, θ_2 values as 0.5, 0.4, and -0.4, 0.2 respectively. The DR techniques were then applied on this data and RMSE was calculated for evaluating the techniques. Figure 6 shows the comparison between DR techniques applied to *i.i.d.* and non-*i.i.d.* data. It is observed that PCA-DR

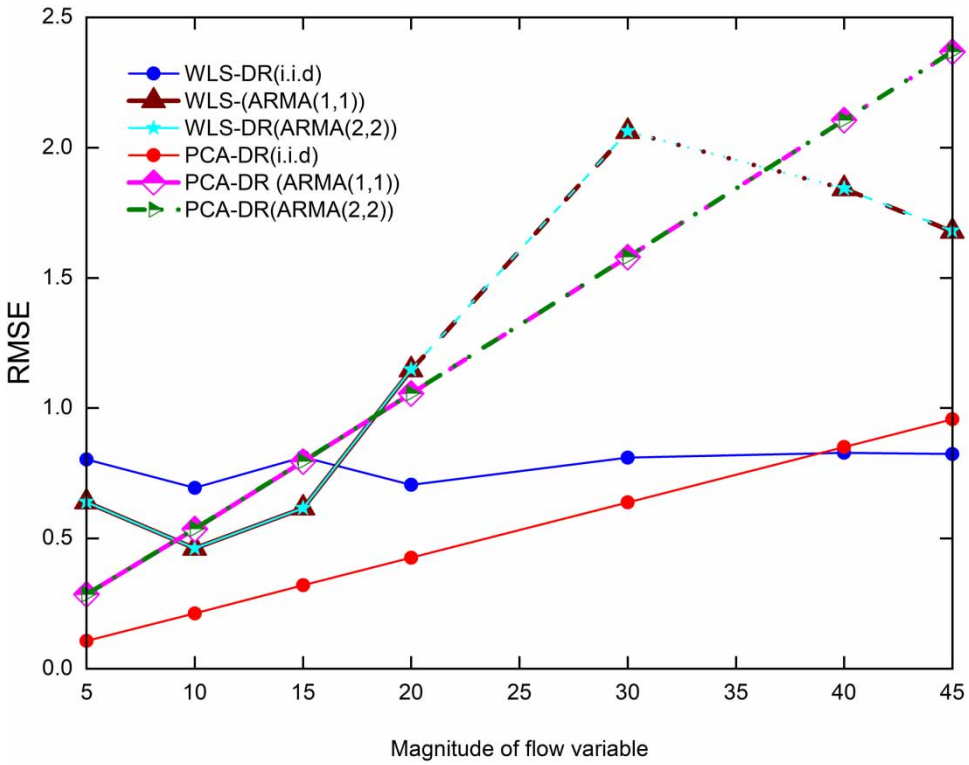

Figure 6 | Performance of DR techniques for *i.i.d.* and non-*i.i.d.* for varying magnitudes of flow variable.

estimates of non-*i.i.d.* data for both ARMA (1, 1) and ARMA (2, 2) noise are similar, and serial correlation makes the PCA-DR worsen due to erroneous variance estimation. WLS-DR when applied to the non-*i.i.d.* data shows high inconsistency in the estimates and gives inaccurate results, thus concluding that WLS-DR cannot be used for non-*i.i.d.* data. When comparing holistically with *i.i.d.* data, we can say that both PCA-DR and WLS-DR perform better when applied to *i.i.d.* data than to non-*i.i.d.* data.

CONCLUSION

The work presented here can be applied to monitoring of water inflows in urban smart water network management, waste water treatment and other areas where water supply is vital. We agree that the present work focuses on techniques that contribute significantly to augmenting the large body of knowledge in metrology. However, we would like to emphasize that these techniques find significant and direct applications in resource reconciliation in water supply networks. When it comes to water, while the specific challenges of commercial (process) use, domestic use, power use, mining use, utility use, etc. may differ, what they all have in common, when seen in the context of water supply networks, is that such networks contain the two basic constituent elements of streams (water flow) and nodes (convergence/divergence points of flows). The techniques outlined in this paper are then generally applicable for all such networks, as they employ a useful relationship constraint, namely the mass balances. Thus, it is evident that the current work has direct and important relevance to water supply, via the treatment it provides for resource reconciliation. In such networks, the reconciliation must, of necessity, address physical leaks as well as measurement biases – both of which form the core focus of data reconciliation and gross error techniques.

This work presents an approach that integrates the process of selecting a suitable data reconciliation technique with the process of selecting the best performance evaluation metric under different scenarios. The study reveals

that the RMSE is the best metric (among those considered), since it is very sensitive to small changes in the measurement and estimates. SNR, GD, and CC were seen to not capture significant changes in reconciliation. RER performs in a very different way from that of RMSE, and has negative values for poorly reconciled estimates.

From the two examples discussed in this paper, PCA-DR is a good technique when compared to WLS-DR, for variables with smaller magnitudes of *i.i.d.* data. When the difference in magnitudes between process variables increases PCA-DR performs less accurately compared to WLS-DR. As in the case of changing variance, PCA-DR is better than WLS-DR. For the biased data both techniques failed in reducing the random errors. For serially correlated (non-i.i.d.) data, PCA and WLS-DR both are performing poor, due to variance change.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Bhattacharyya, A., Yogi, V., Singla, S., Bhushan, M., Kelkar, M. G., Tiwari, A. P., Pramanik, M. & Belur, M. N. 2017 Adaptive, online models to detect and estimate gross error in SPNDs. In: *Proceedings of 2017 Indian Control Conference (ICC 2017)*, Guwahati, India. IEEE Control Systems Society. doi:10.1109/INDIANCC.2017.7846467.

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), 1–38. https://doi.org/10.3390/pr5040056.

Chai, T. & Draxler, R. 2014 Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* **7**, 1247–1250. doi:10.5194/gmd-7-1247-2014.

Fuente, M. J., Gutierrez, G., Gomez, E., Sarabia, D. & Prada, D. 2015 Gross error management in data reconciliation. In: *Proceedings of 9th International Symposium on Advanced Control of Chemical Processes*. IFAC, British Columbia, Canada. https://doi.org/10.1016/j.ifacol.2015.09.037.

Helness, H., Damman, S., Sivertsen, E. & Ugarelli, R. 2019 Principal component analysis for decision support in integrated water management. *Water Supply* **19** (8), 2256–2262. https://doi: 10.2166/ws.2019.106.

Jeyanthi, R. & Devanathan, S. 2020 Addressing higher order serial correlation in techniques for gross error detection. *Journal of Computational and Theoretical Nanoscience* **17**, 297–302. doi:10.1166/jctn. 2020.8665.

Korpela, T., Suominen, O., Majanne, Y., Laukkanen, V. & Lautala, P. 2016 Robust data reconciliation of combustion variables in multi-fuel fired industrial boilers. *Control Engineering Practice* **55**, 101–115. https://doi.org/10.1016/j.conengprac.2016.07.002.

Krōcová, S. 2016 Water supply systems and their influence on increasing operational safety in industry. *Perspectives in Science* **7**, 236–239. https://doi.org/10.1016/ j.pisc.2015.11.038.

Lin, Y., Kruger, U., Gu, F., Ball, A. & Chen, Q. 2019 Monitoring non stationary and dynamic trends for practical process fault diagnosis. *Control Engineering Practice* **84**, 139–158. https://doi.org/10.1016/j.conengprac.2018.11.020.

Miao, Y., Su, H., Gang, R. & Chu, J. 2009 Industrial process: data reconciliation and gross error detection. *Measurement+Control* **42** (7), 209–215. https://doi.org/10.1177 /002029400904200704.

Narasimhan, S. & Bhatt, N. 2015 Deconstructing principal components analysis using a data reconciliation perspective. *Computers and Chemical Engineering* **77**, 74–84. https://doi.org/10.1016/j.compchemeng.2015.03.016.

Narasimhan, S. & Shah, S. L. 2008 Model identification and error covariance matrix estimation from noisy data using PCA. *Control Engineering Practice* **16**, 146–155. https://doi.org/10.1016/j.conengprac.2007.04.006.

Park, S. & Jung, Y. S. 2014 Principal component analysis of water pipe flow data, 16th conference on water distribution system analysis, WDSA 2014. *Procedia Engineering* **89**, 395–400. doi:10.1016/j.proeng.2014.11.204.

Quevedo, J., Pascual, J., Puig, V., Saludes, J., Sarrate, R., Escobet, A., Espin, S. & Roquet, J. 2014 Flowmeter data validation and reconstruction methodology to provide the annual efficiency of a water transport network: the ATLL case study in Catalonia. *Water Supply* **14** (2), 337–346. https://doi.org/10.2166/ws.2013.203.

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 Environment Research* **82** (5), 426–433. https://www.jstor.org/stable/25679798.

Schönenberger, U. 2015 *Data Reconciliation and Gross Error Detection for Wastewater Treatment Processes*. Master Thesis, Process Engineering in Urban Water Management, ETH Zürich.

Spuler, M., Sanz, S. S., Birbaumer, N., Rosenstiel, W. & Murguialday, A. R. 2015 Comparing metrics to evaluate performance of regression methods for decoding of neural signals. In: *Proceedings of 37th Annual International*

Conference of EMBC'15, Milan, Italy. IEEE Engineering in Medicine and Biology Society, pp. 1083–1086. https://doi.org/10.1109/EMBC.2015.7318553.

Srinivasan, S., Billeter, J., Narasimhan, S. & Bonvin, D. 2017 Data reconciliation for chemical reaction systems using vessel extents and shape constraints. *Computers and Chemical Engineering* **101**, 44–58. https://doi.org/10.1016/j.compchemeng.2017.02.003.

Syed, M. S., Kerry, M., Dooley, K. M., Madron, F. & Knopf, C. 2016 Enhanced turbine monitoring using emissions measurements and data reconciliation. *Applied Energy* **173**, 355–365. https://doi.org/10.1016/j.apenergy.2016.04.059.

Tran, K. P., Nguyen, H. D., Tran, P. H. & Heuchenne, C. 2019 On the performance of CUSUM control charts for monitoring the coefficient of variation with measurement errors. *International Journal of Advanced Manufacturing Technology* **104** (5–8), 1903–1917. https://doi.org/10.1007/s00170-019-03987-6.

Valle, E. C., Kalid, R. A., Secchi, A. R. & Kiperstok, A. 2018 Collection of benchmark test problems for data reconciliation and gross error detection and identification. *Computers & Chemical Engineering* **111**, 134–148. https://doi.org/10.1016/j.compchemeng. 2018.01.002.

Varshith, C. R., Rishika, S. R., Ganesh, S. & Jeyanthi, R. 2017 Principal component analysis based data reconciliation for a steam metering circuit. In: *Advances in Intelligent Systems and Computing, Soft Computing and Signal Processing: Proceedings of ICSCSP 2018*. 898 (2), pp. 619–625. https://doi.org/10.1007/978-981-13-3393-4_63.

Xie, S., Yang, C., Yuan, X., Wang, X. & Xie, Y. 2019 A novel robust data reconciliation method for industrial processes. *Control Engineering Practice* **83**, 203–212. https://doi.org/10.1016/j.conengprac.2018.11.006.

Xu, Z., Ying, Z., Li, Y., He, B. & Chen, Y. 2020 Pressure prediction and abnormal working conditions detection of water supply network based on LSTM. *Water Supply* **20** (3), 963–974. https://doi.org/10.2166/ws.2020.013.

Zhang, Z., Shao, Z., Chen, X., Wang, K. & Qian, J. 2010 Quasi-weighted least squares estimator for data reconciliation. *Computers & Chemical Engineering* **34**, 154–162. https://doi.org/10.1016/j.compchemeng.2009.09.007.

First received 28 August 2020; accepted in revised form 30 October 2020. Available online 12 November 2020