This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2018.2875954, IEEE Sensors Journal

Monitoring Influent Measurements at Water Resource Recovery Facility Using Data-Driven Soft Sensor Approach

Tuoyuan Cheng, Fouzi Harrou, *Member, IEEE*, Ying Sun and TorOve Leiknes

***Abstract*—Monitoring inflow measurements of water resource recovery facilities (WRRFs) is essential to promptly detect abnormalities and helpful in the decision making of the operators to better optimize, take corrective actions, and maintain downstream processes. In this paper, we introduced a flexible and reliable monitoring soft sensor approach to detect and identify abnormal influent measurements of WRRFs to enhance their efficiency and safety. The proposed data-driven soft sensor approach merges the desirable characteristics of principal component analysis (PCA) with k-nearest neighbor (KNN) scheme. PCA performed effective dimension reduction and revealed interrelationships between inflow measurements, while KNN distances demonstrated superior detection capacity, robustness to underlying data distribution, and efficiency in handling high-dimensional dataset. Furthermore, nonparametric thresholds derived from kernel density estimation further enhanced detection results of PCA-KNN approach when compared with parametric counterparts. Moreover, the radial visualization plot is innovatively employed for fault analysis and diagnosis in combination with PCA and delineated interpretable visualization of anomalies and detector performances. The effectiveness of these soft sensor schemes is evaluated by using real data from a coastal municipal WRRF located in Saudi Arabia. Also, we compared the proposed soft sensor scheme with the conventional PCA-based approaches, including standard prediction error, Hotelling's T^2, and joint univariate methods. Results demonstrate that this soft sensor-based monitoring approach outperforms conventional PCA-based methods.**

***Index Terms*—Water resource recovery facility, Influent measurements, Process monitoring, K-nearest neighbor, Radial visualization**

I. Introduction

WATER resource recovery facilities (WRRFs) are sophisticated systems that have to sustain long-term qualified performance, regardless of temporally volatile volumes or qualities of the incoming wastewater [1], [2]. Municipal WRRFs have to manage rainfall and snowmelt, while industrial counterparts often treat wastewater with frequent sudden shifts in unique composition and temperature due to production processes [3]. Nevertheless, WRRFs have limited storage for inflow, which cannot be rejected or abandoned. Dynamic and nonlinear influent measurements (IMs) together with tightening discharge regulations and rising operational costs demand maximized efficiency from practitioners.

For decades, challenges are proposed to operators. Updated instrumentation, control, and automation are offering overwhelmingly voluminous data, which are often unexploited, forming the "data-rich, information-poor" dilemma [4]. Timely analysis and modeling would extract valuable information to support process understanding, online prediction, process monitoring and predictive control.

Water resource recovery has been given vital importance and seen as a promising solution to the water scarcity in water-stressed countries, such as the Kingdom of Saudi Arabia (KSA). As initial conditions provided to WRRFs, IMs influence treatment systems, ongoing processes and product characteristics, and accordingly are fundamental in operation thus emphasized, recorded, and monitored in KSA. Anomalies or faults in IMs need to be detected and diagnosed promptly for decision making to avoid unexpected system crash, maintain steady product quality, support efficient downstream processes, improve reliability and reduce labor costs [5]. Accordingly, the detection and identification of abnormal events in IMs of WRRFs are of primary importance to keep WRRFs operating with the desired performance.

Modern WRRFs need to improve process quality while guaranteeing efficient and fault-free operations, though they are continuously subjected to unexpected influent changes. Conventionally, only parameters from the process are monitored, offline and univariately, which often ignored interaction between correlated variables and was not providing satisfying results in practice. Since the process generally takes more than one day, multivariate data-driven soft sensors targeting the influent (the upstream compared to the process) would be promising. Operators can take timely responses to detected anomalies from the IMs. Moreover, information from multiple variables is utilized simultaneously, where cross-correlation among variables are considered in the decision rules, and the total number of monitoring control charts are minimized. This may serve as a better "best practice" to the current situation in WRRFs. In the process industry, data-driven soft sensors are becoming more and more popular [6]. Applications include prediction, reconstruction, and sensor monitoring.

All over the years, methods are developed for prediction as well as fault detection and diagnosis, including mechanistic model-based or analytical methods, and model-free or data-

T. Cheng and T. Leiknes are with Water Desalination and Reuse Center, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia, e-mail: tuoyuan.cheng@kaust.edu.sa

F. Harrou, and Y. Sun are with Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia, e-mail: fouzi.harrou@kaust.edu.sa

1558-1748 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2018.2875954, IEEE Sensors Journal

derived methods [7]. Analytical models, utilizing first principals, could theoretically explain both linear and nonlinear behavior, reveal process mechanism, but request prior parameters for calibration and would be challenged by costly high-dimensional computation and ill-conditioned problems [8]. Data-derived methods, nowadays are more common in the environmental field, even though not as widespread as, for example, in petrochemical industry where soft-sensors are extensively utilized for billions of dollars were once wasted annually due to abnormal events [7], [9]. Environmental data have been utilized by data-derived methods for prediction of downstream pollutants concentration in river networks [10], and sludge bulking monitoring in WRRFs [11].

Time series analysis approaches have widely been applied to model WRRFs. Berthouex et al. [12] proposed an exponentially weighted moving average model for forecasting WRRF performance.Huo et al. [13] applied time series techniques to evaluate and predict IMs. Escalas-Cañellas et al. [14] proposed a time series model estimating and evaluating influent water temperatures at WRRFs. Stationarity is usually assumed in time series models, while the wastewater treatment processes are often nonstationary. Additionally, these approaches depend on the accuracy of the model used.

Machine learning algorithms turn out to play a considerable role in the literature. Artificial neural network (ANN) simulation for the monitoring and control of an anaerobic WRRF is presented by Wilcox et al. [15]. The ANN procedure has been trained over 80 h with various monitoring data and exhibited suitable ability in monitoring alkalinity level. Dias et al. [16] investigated ANN and fuzzy neural models for monitoring and prediction. The approach showed good detection performance when applied to data from the IWA/COST benchmark simulation model. ANN models have also been used to monitor key parameters (turbidity, temperature, pH, oxidation-reduction potential, and UV light intensity) in WRRFs [17]. Zhu et al. [18] introduced a hybrid approach based multiple linear regression, artificial neural networks to predict influent biochemical oxygen demand, which is expensive and difficult to measure with sensors. Other methods examined for process monitoring include fuzzy models [19], and support vector regression [20]. Such machine learning methods depend on the availability of input data, and their implementation is no easy task, especially for real-time applications.

Latent variables are employed for monitoring and prediction via projection latent structures (PLS) and principal component analysis (PCA). Amaral et al. [21] applied PLS regression for activated sludge process monitoring. PLS methods have been utilized to predict deterioration of sludge sedimentation properties [22]. An approach to predict the influent chemical oxygen demand (COD) using PLS models has been applied to a newsprint mill WRRF [23].Wang et al. [24] proposed a statistical approach based on combined PCA and multiple regression to model a WRRF. Ebrahimi et al. [25], suggested a multivariate approach based on PCA to predict quality parameters such as BOD_5 and total phosphorus to analyze WRRFs performance.

PCA as a popular multivariate statistical dimension reduction technique applied to the visualizing of dataset variation and composition, can delineate normal operational conditions (NOCs) and detect faults with simplicity and interpretability. However, it is well known that the conventional PCA-based squared prediction error (SPE), Hotelling's T^2, and joint univariate methods, assuming Gaussian distribution among process observations, does not guarantee a satisfactory performance in anomaly detection, in particular when detecting incipient changes [26].

Thereby, the overarching goal of this study is introducing new soft sensor-based monitoring strategy with improved detectability compared to conventional PCA-based methods. Here, a data-driven soft sensor approach merging the feature-extraction capability of PCA and the classification capacity of k-nearest neighbors (KNN) is proposed to detect and identify abnormal influent measurements of WRRFs. Our scheme alleviates the drawbacks of the conventional PCA-based indices because it employs KNN-based detection scheme, which demands no prior assumptions on the underlying data structure, and can cope with nonlinearity and multimodality in data [27]. These properties are favored in practical situations where collected data are non-Gaussian distributed or linearly non-separable, which implies KNN as a competitive alternative to traditional PCA-based fault detection approach. In the proposed scheme, fault-free residuals obtained from PCA model is fed to KNN for training. Then KNN classifier would discriminate between NOCs and faults in testing data by computing their distances to neighbors. Moreover, threshold selection methods by conventional parametric and kernel density estimation (KDE) based nonparametric approaches are also evaluated. Conventional PCA-based SPE and T^2 approaches are used as benchmarks for comparison. The aim of this research is not only sensing abnormal events in influent measurements of WRRFs but also identifying the type (source) of abnormalities, such that operators can respond accordingly by making any necessary and take corrective actions to protect the system. To assist fault diagnosis, anomalies are analyzed via the radial visualization (RadViz) plot, which is intuitive in high dimensional data interpretation. To evaluate the proposed soft sensor-based monitoring schemes, data from a coastal municipal WWRF located in KSA is experimented with.

The PCA and KNN based soft sensors used for multivariate process monitoring are briefed in the following section. Then the proposed monitoring scheme is introduced in Section III. The performance of the developed data-driven soft sensor approach is assessed via real data application in Section IV, and conclusions are presented in Section V.

II. Methods

This section introduces an overview of conventional PCA-based modeling and monitoring, together with the KNN-based algorithm.

A. Principal component analysis (PCA)

PCA has become a popular modeling technique to extract information from multivariate process data by relating process

1558-1748 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

variables [28], [29], [30], [31]. Let $\mathbf{X} = \left[\mathbf{x}_1^T, \dots, \mathbf{x}_n^T\right]^T \in R^{n\times m}$ be a centered and scaled measurement matrix with n measurements and m variables. The data matrix $\mathbf{X}$ is factorized into two orthogonal parts,

$$\mathbf{X} = \mathbf{T}\mathbf{W}^T = \sum_{i=1}^{k} \mathbf{t}_i \mathbf{w}_i^T + \sum_{i=k+1}^{m} \mathbf{t}_i \mathbf{w}_i^T = \widehat{\mathbf{X}} + \mathbf{E} \qquad (1)$$

where $\widehat{\mathbf{X}}$ is the approximation or prediction and $\mathbf{E}$ is the residual. $\mathbf{T} = [\mathbf{t}_1 \ \mathbf{t}_2 \cdots \mathbf{t}_m] \in R^{n\times m}$ and $\mathbf{W} \in R^{m\times m}$ represent a matrix of the transformed uncorrelated variables (i.e. principal components, PCs) and a corresponding loading matrix, respectively.

Generally speaking, PCA is a method for decoupling signals and noises where collinearity often presents. Given certain correlation (redundancy) in data $\mathbf{X}$, the first k PCs (where $k < m$) can capture most of the variability in $\mathbf{X}$. This part of the variability commonly arises from the true underlying signals. The remaining $m - k$ PCs capture the variability that arises from noise. In other words, the information unexplained by the first k PCs formed the residual data matrix.

When implementing the PCA algorithm, the singular value decomposition is applied to the observed data to compute the loading vectors. The loading matrix is calculated from the covariance matrix $\mathbf{S}$ of the input data $\mathbf{X}$ as:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} = W\Lambda W^T \ \text{ with } \ WW^T = W^TW = I_n. \qquad (2)$$

which can be reformulated as:

$$\mathbf{S} = \left[\begin{array}{cc} \widehat{W} & \widetilde{W} \end{array}\right] \left[\begin{array}{cc} \widehat{\Lambda} & 0 \\ 0 & \widetilde{\Lambda} \end{array}\right] \left[\begin{array}{c} \widehat{\mathbf{W}}^T \\ \widetilde{\mathbf{W}}^T \end{array}\right] \qquad (3)$$

Here, $\Lambda = diag(\sigma_1^2, \dots, \sigma_m^2)$ is a diagonal matrix comprised of decreasingly ordered eigenvalues of $\mathbf{S}$ [32]. The eigenvalues λ_i are equal to the variance of the PC t_i, namely σ_i^2. Besides, $\widehat{\mathbf{W}} \in R^{m\times k}$ is comprised of eigenvectors corresponding to the first k largest eigenvalues in $\Lambda \in R^{k\times k}$, while $\widetilde{\mathbf{W}} \in R^{m\times(m-k)}$ represents the remaining $m-k$ eigenvectors associated to the rest of eigenvalues. From equation (1),

$$\mathbf{X} = \widehat{\mathbf{T}}\widehat{\mathbf{W}}^T + \widetilde{\mathbf{T}}\widetilde{\mathbf{W}}^T = \widehat{\mathbf{X}} + \mathbf{E} \qquad (4)$$

where $\widehat{\mathbf{T}} = [\mathbf{t}_1, \dots, \mathbf{t}_k]$ is the principal component score matrix ($n \times k$), which describes the values of variables in the transformed $n \times k$ basis space spanned by $\widehat{\mathbf{W}}$. Here $\widetilde{\mathbf{T}} = [\mathbf{t}_{k+1}, \dots, \mathbf{t}_m]$ is obtained by choosing the last $m - k$ PCs in $\mathbf{T}$ such that $\widetilde{\mathbf{T}}$ represents only the variability of random errors.

In this study, cumulative percent variance (CPV) procedure is applied here to properly select the number of retained PCs:

$$CPV = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \times 100. \qquad (5)$$

In this CPV procedure, k is determined by counting PCs until the cumulative variance explains the desired percentage (i.e., 80%) of the total variance. In this way, the true signal variation would be decoupled from the noise variation, and the two types of variation shall be monitored separately.

B. Parametric and nonparametric PCA-based anomaly detection methods

After a reference PCA model is designed using anomaly-free data, it can be used for monitoring new datasets via SPE and Hotelling's T^2. The SPE statistic monitors residual subspace, while the T^2 statistic monitors changes in the PCs subspace. The residuals of a fitted model are defined by:

$$\mathbf{e} = \mathbf{x} - \widehat{\mathbf{x}}, \qquad (6)$$

where $\widehat{\mathbf{x}}$ is the predicted value of $\mathbf{x}$ by PCA. If the PCA model describes the observed fault-free data adequately and the process is in NOC, then residuals should be approximately around zero due to noise and uncertainty, in a normal distribution.

The SPE scheme is a widely used criterion for measuring the goodness-of-fit of a data sample to its PCA model. The SPE statistic is computed as [26]:

$$SPE = \mathbf{e}^T\mathbf{e}. \qquad (7)$$

The SPE statistics gives a signal of an anomaly when:

$$SPE > SPE_\alpha, \qquad (8)$$

where SPE_α is a threshold or upper control limit (UCL), defined by:

$$SPE_\alpha = \varphi_1 \left[\frac{h_0 c_\alpha \sqrt{2\varphi_2}}{\varphi_1} + 1 + \frac{\varphi_2 h_0 (h_0 - 1)}{\varphi_1^2}\right], \qquad (9)$$

where $\varphi_i = \sum_{j=k+1}^{m} \lambda_j^i, \quad \text{for} \ \ i = 1, 2, 3, \ h_0 = 1 - \frac{2\varphi_1\varphi_3}{3\varphi_2^2}$, and c_α is the confidence limits for the $1 - \alpha$ percentile in a normal distribution. The T^2 statistic is defined by [33]:

$$T^2 = \sum_{i=1}^{k} \frac{t_i^2}{\sigma_i^2}, \qquad (10)$$

where σ_i^2 is the estimated variance of the corresponding PC $\mathbf{t}_i$. The T^2 quantifies changes in PCs subspace. The T^2 statistics would give a signal of an anomaly when:

$$T^2 > T_\alpha^2, \qquad (11)$$

where T_α^2 is a threshold or UCL, defined by:

$$T_\alpha^2 = \chi_{1-\alpha,k}^2 \qquad (12)$$

The parametric UCL of SPE and T^2 requires the assumption that observed values are temporally non-correlated and normally distributed, which are usually denied in many cases. One alternative approach is to adopt the kernel density estimation for nonparametric SPE and T^2 thresholds [34]. After the density of a statistic generated in NOCs (anomaly-free) is calculated by KDE, the nonparametric UCL or decision threshold is then set as the corresponding $(1 - \alpha)$-th quantile of this estimated distribution.

C. K-nearest neighbor (KNN)

The k-nearest neighbor algorithm is a widely used nonparametric classification technique that quantifies the similarity be-

1558-1748 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

tween test observations and their corresponding similar training sets [35]. The technique does not make prior assumptions about the underlying data structure, which is preferred when the collected data are non-Gaussian distributed or not linearly separable [36], [37]. KNN-based approaches have widely used in the literature [36], [38], [39], [37]. For instance, in [39], KNN-based forecasting approach is introduced to predict the IMs of WRRF. In [38], KNN is used for detecting anomalies in meteorological data. In KNN, each training tuple delineates a point in a multidimensional feature space. For every test sample, the detector searches the feature space for k training points that are closest to the new sample, which consists the k-nearest neighbors of the tuple. Closeness is represented by distance measures, such as Euclidean or Manhattan distance. In this paper, residuals from PCA are employed to compute distances to their nearest neighbor, namely, $k = 1$, to minimize computational cost. Significant distances may imply anomalies and therefore utilized for fault detection. The flow of KNN distance based anomaly detection are given below:

Step 1 For every residual of observation $\mathbf{x}_i$ in the training dataset, find its Manhattan and Euclidean distances to its nearest neighbor in the training set D_i, based on which we have sample distributions of distances;

Step 2 From the distribution of D_i, parametric UCL of KNN distance is defined as the mean value plus three times the sample standard deviation ($D_{\alpha,p} = \mu_D + 3\sigma_D$, with $\alpha = 2pnorm(-3) \approx 0.0013$), where μ_D and σ_D are the mean and standard deviation of KNN distances using anomaly-free training data, and normality is assumed implicitly in this three-sigma methodology.

Setp 3 For every residual of new observation in the testing dataset, compute its Manhattan and Euclidean distance D_* to its nearest neighbor in the training set. Anomaly is detected if D_* exceeds the UCL, $D_{\alpha,p}$.

The conventional parametric three-sigma control chart is appropriate only when the normality assumption is valid. Otherwise, results would be unreliable or even misleading. To overcome this drawback, kernel density estimation can be used to estimate the distribution of KNN distances [34]. The KNN algorithm with a nonparametric threshold is summarized next.

Step 1 For every residual of observation $\mathbf{x}_i$ in the training dataset, find its Manhattan and Euclidean distances to its nearest neighbor in the training set D_i, based on which we have sample distributions of distances;

Step 2 From the distribution of D_i, non-parametric UCL of KNN distance, $D_{\alpha,np}$ is defined as the $(1-\alpha)$-th quantile of the estimated distribution of KNN distances obtained by KDE, where $\alpha = 2pnorm(-3) \approx 0.0013$ is the same false alarm rate as in the parametric approaches.

Step 3 For every residual of new observation in the testing dataset, compute its Manhattan and Euclidean distance D_* to its nearest neighbor in the training set. Anomaly is detected if D_* exceeds the UCL, $D_{\alpha,np}$.

III. Proposed PCA-KNN anomaly detection scheme

Firstly, PCA model is built under NOCs using fault-free training data and applied on testing data (with faults) to generate residuals, which serve as the input to KNN model for anomaly detection. Then the KNN-based scheme is used to quantify distances between actual residuals from testing observations and residuals from fault-free training samples. Data from NOCs should have KNN distances closer to zero, whereas anomalies would display larger KNN distance values. To set UCL for decisions, both parametric and nonparametric procedures are adopted. The parametric ones are derived from the three-sigma rule, while the nonparametric ones involve quantiles from KDE.

In summary, the proposed soft sensors-based monitoring approach consists of two main phases: model construction based on the anomaly-free training dataset, and online anomaly detection of the testing dataset using the KNN-based monitoring methods. In model identification, the objective is to find a suitable PCA model for estimating the PCs space or its complementary part, the residual space (for anomaly detection, the latter is critical). The main steps of the proposed PCA-KNN algorithm are summarized next.

I **Phase 1: Model building**

Step 1 Collect the training dataset (fault-free data), representative of a normal situation.

Step 2 Scale the data to zero mean and unit variance.

Step 3 Construct PCA model based on the training data:

- ○ Select the number of PCs to be kept in the PCA model by using CPV procedure.
- ○ Decompose the scaled matrix $\mathbf{X}$ as two complementary parts, the predictions and residuals, as given in Equation (1).

Step 4 Compute the control limits (parametric and nonparametric)of the PCA-KNN statistic, $D_{\alpha,p}$ and $D_{\alpha,np}$.

II **Phase 2: Anomaly detection**

Step 1 Pretreat the testing dataset by scaling with the mean and standard deviation of the training data.

Step 2 Compute residuals and KNN distances.

Step 3 Declare a fault when the KNN statistic exceeds the control limits previously computed using the training data, $D_{\alpha,p}$ and $D_{\alpha,np}$.

IV. Case study: monitoring a WRRF in Saudi Arabia

A. Data description

To validate the proposed data-driven soft sensor method, experimental investigations are conducted, using historical records from the WRRF based in King Abdullah University of Science and Technology (KAUST) located in Thuwal, Saudi Arabia (Figure 1). This plant is an advanced facility with a

1558-1748 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2018.2875954, IEEE Sensors Journal

Fig. 1. a) Headwork of the WRRF at KAUST, Thuwal, KSA, b) vortex flow grit chamber, c) sampling sites.

daily treatment capacity of 9500 m^3. Since water is precious throughout the region, all wastewater (stormwater, gray water, and black water), as well as condensate load from KAUST campus and community, are delivered online to the WRRF for water resource recovery. Treated wastewater is used for irrigation, thus greatly reducing potable water demand and desalination energy consumption of the university. Operators maintained daily measurements of twenty-one variables as listed in Table I. The obtained dataset contains seven years (from Sep. 1st, 2010 to Sep. 1st, 2017) of daily observation over twenty-one variables with few not available data (within 1%, 132 out of 63950) imputed by R package Amelia [40].

In this study, influent measurements, such as flow, temperature, pH, conductivity, TSS, COD, NH_3N, NO_3N, BOD_5, and Cl are measured daily by on-line sensors (IQ SensorNet, https://www.ysi.com/IQSN2020). The IQ SensorNet is installed inside the headwork and next to the sampling sites. The whole sensor instrument monitors the water quality continuously and it is connected to the display panel and controller for storing and executing all system settings. To use it, the probe (Figure 2(a)) is dipped into wastewater after which physicochemical reactions took place, and signals are passed from the probe through the cable to the display panel (Figure 2(b)). The sensors have sensitivities of 0.01 L/min for flow, of 0.01 Celsius for temperature, and of 0.01 mg/L for water quality parameters. The installed sensors are reported to be used twice per day, and the probes within are renewed per two years. Others IMs including TDS, CaHardness, MgHardness, TotalAlkalinity, FOG, TKN, PO_4P, and Boron are measured daily by off-line analysis.

IMs can be sketched by descriptive statistics as given in Table II. The distribution of each parameter is captured by its mean value or the first order moment, while the spread in the dataset is delineated by standard deviations, extremes, and quartiles. Skewness and kurtosis are computed to exhibit symmetry and shape of the investigated time series distributions.

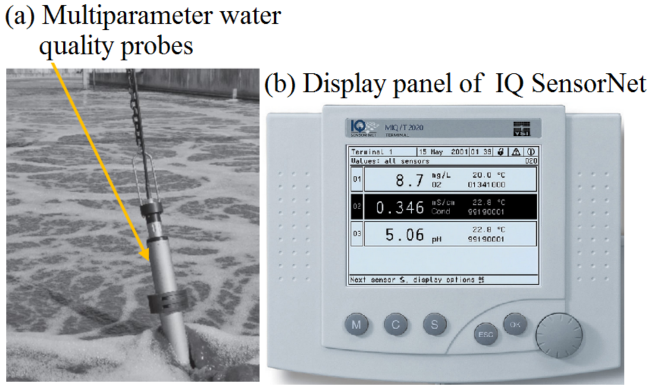

Fig. 2. Multiparameter sensor capable of measuring and recording multiple parameters simultaneously. (a) Multiparameter water quality sensors and (b) Display panel of IQ SensorNet.

The hierarchically-clustered heatmap based on Pearson correlation coefficients reveals linear relationship and hierarchically closeness among variables [41]. Figure 3 shows that IM parameters were ranked and segmented into five blocks by cutting branches and positive relationships are evident compared to negative ones. Since pH is a logarithmic value, it is relatively independent of others and only showed a weak connection with the second "Flow" block. The inflow from the whole campus area contributed the majority of the inflow from the whole university, and both of them showed a positive relationship with temperature, indicating higher water consumption during hot seasons. Boron concentration is generally provided by the flow from a desalination plant, which is another component of the total inflow.

Central in the matrix lies the third "CNP" block, where carbon and nutrient contents are included. Interestingly, total suspended solids is also found within, and displayed a positive correlation with BOD_5 or COD, implying organic composition. In the fourth "TDS" block encompassing conductivity and TDS, it can be inferred that the inorganic ionic content is principally controlled by chloride and magnesium. The fifth block contains "miscellaneous" characteristics, in which the recycled inflow is statistically independent, whereas alkalinity, calcium hardness, and FOG formed a closer cluster. Besides, alkalinity and FOG also featured positive correlation with the third "CNP" block, while calcium hardness disclosed positive correlation with the fourth "TDS" block.

1558-1748 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2018.2875954, IEEE Sensors Journal

No.	Variable name	Measurement scopes	limit
1	InFlow-LS1	Wastewater inflow, from the whole campus area, in m^3/day	-
2	InFlow-LS8	Wastewater inflow, from a desalination plant, in m^3/day	-
3	InFlow-DP	Wastewater inflow, recycled from WRRF itself, in m^3/day	-
4	InFlow-Total	Wastewater inflow, from the whole university, in m^3/day	2500-6000
5	Temp	Temperature, in Celsius	-
6	pH	potential of hydrogen, unitless	6-9
7	Conductivity	Conductivity, in μS/cm	< 2850
8	TDS	Total dissolved solid, in mg/L	< 2000
9	TSS	Total suspended solid, in mg/L	< 312
10	CaHardness	Calcium hardness, in mg/L	-
11	MgHardness	Magnesium hardness, in mg/L	-
12	TotalAlkalinity	Total alkalinity, in mg/L	< 200
13	BOD_5	5-day Biochemical Oxygen Demand, in mg/L	< 264
14	COD	chemical oxygen demand, in mg/L	< 527
15	FOG	Fat, oils and grease, in mg/L	-
16	TKN	Total Kjeldahl Nitrogen, in mg/L	< 40
17	NH_3N	Ammonia, in mg/L	< 25
18	NO_3N	Nitrate, in mg/L	< 10
19	PO_4P	Phosphate, in mg/L	-
20	Cl	Chloride, in mg/L	-
21	Boron	Boron, in mg/L	< 2.5

TABLE I
MONITORED INFLUENT MEASUREMENTS IN KAUST WRRF.

	mean	std	min	0.25	0.5	0.75	max	skewness	kurtosis
InFlow-LS1	3021.61	535.45	2228.00	2660.00	2851.00	3244.00	5249.00	1.33	1.89
InFlow-LS8	279.31	156.08	64.00	156.00	228.00	366.00	867.00	1.13	0.92
InFlow-DP	47.43	95.79	0.00	9.00	10.00	36.00	749.00	3.80	18.67
InFlow-Total	3512.92	611.00	2558.00	3036.00	3389.00	3853.00	5642.00	0.86	0.34
Temp	29.31	1.59	25.99	28.10	29.20	30.60	32.50	0.27	-0.85
pH	7.40	0.25	6.59	7.25	7.36	7.52	8.56	0.91	3.20
Conductivity	669.37	284.69	264.00	537.00	625.00	719.00	2466.00	3.19	15.37
TDS	459.48	209.08	174.00	363.00	429.00	492.00	1809.00	3.37	16.69
TSS	68.66	27.29	12.00	49.00	64.00	82.00	187.00	1.14	2.10
CaHardness	72.75	30.78	20.00	52.00	72.00	94.00	176.00	0.49	0.17
MgHardness	41.91	27.46	6.00	24.00	36.00	48.00	156.00	1.81	3.59
TotalAlkalinity	120.88	24.98	68.00	100.00	120.00	136.00	196.00	0.22	-0.33
BOD_5	99.03	36.95	27.00	71.00	92.00	123.00	224.00	0.58	0.10
COD	152.99	61.83	42.00	102.00	157.00	189.00	329.00	0.50	-0.25
FOG	54.36	53.05	2.90	14.30	37.10	77.10	351.40	1.86	5.25
TKN	17.91	6.18	2.10	13.80	17.30	21.90	37.90	0.30	0.45
NH_3N	11.84	4.10	0.94	9.30	12.00	14.50	23.60	-0.06	0.47
NO_3N	4.17	1.68	0.10	2.90	4.20	5.10	9.80	0.49	0.47
PO_4P	8.25	2.86	1.30	6.40	8.10	10.00	23.50	1.41	6.59
Cl	126.09	75.62	45.00	91.00	107.00	137.00	654.00	4.29	23.37
Boron	1.15	0.33	0.50	0.90	1.10	1.30	2.50	1.40	2.63

TABLE II
DESCRIPTIVE STATISTICS OF THE TRAINING DATASET.

B. PCA Modeling

The training data is autoscaled before building the PCA model. Here, we used CPV to determine the number of PCs that explains at least 80% of the total variability in the data. Seven PCs (capturing 80.01% of variance) are selected to build the PCA model (See Figure 4).

The transformation matrix between variables and PCs is shown in Figure 4 as a heatmap, with variables ordered identically to the clustered correlation heatmap (Figure 3). The highest loading is given by the first PC which accounted for 32.54% of total dataset variance. As the dominant data pattern provider, it is related to the "Flow" and "CNP" blocks. The second PC explaining 17.55% of total variance is found linked to the "TDS" block and the calcium hardness. The third component incorporated the "Flow" and "CNP" blocks. The fourth component involved the "pH" and "miscellaneous" blocks while the fifth one is majorly comprised of the "miscellaneous" block. The sixth PC is notably influenced by the recycled flow and pH. The seventh PC, as a fine-tuning part in PCA modeling, is influenced by all blocks except the "TDS" block.

Time series of typical variables from each block together with their predictions produced by PCA modeling are given in Figure 5. Generally, the PCA reconstruction by seven components reproduced the trends of 21 variables well and therefore justified dimension reduction in this case and further application in anomaly detection. It can be seen that the influent is weakly alkaline as a whole, whereas its pH may reach over 9, forming a skewness equal to 0.91 from Table II. The inflow from the whole university is fluctuating annually, whose local maximum values are recorded in summer but extreme values seen in winter or the raining season. Boron

1558-1748 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2018.2875954, IEEE Sensors Journal

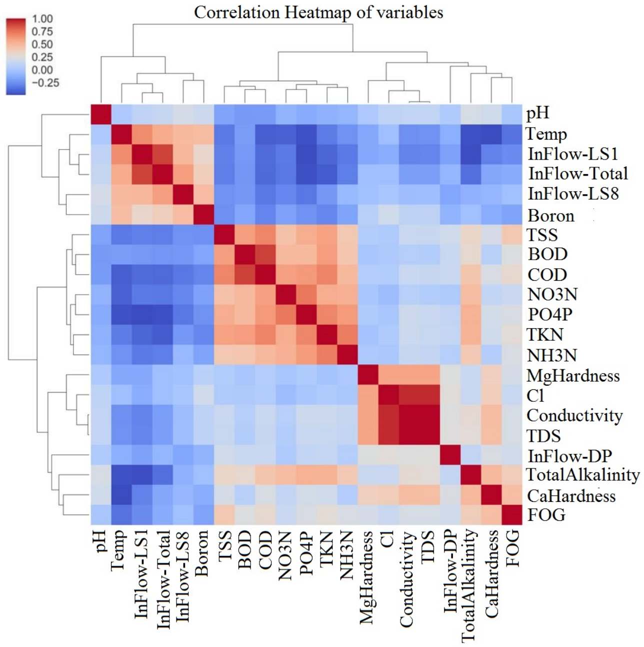

Fig. 3. Hierarchically-clustered heatmap based on Pearson correlation coefficients of the IM variables.

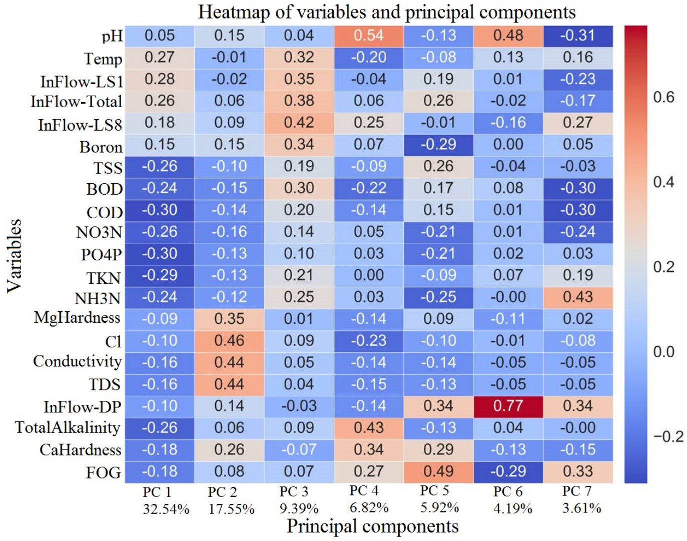

Fig. 4. HeatMap of IMs and retained PCs.

concentration, as well as conductivity, are both leptokurtic and positively skewed, and their extreme values may signify anomalies caused by pollution of seawater origin. Although the BOD$_5$ content showed a rising trend over the years, it is usually below 150 mg/L, and the upper limit of 264 mg/L is rarely reached. The nitrate content, however, has flagged anomalies by crossing 10 mg/L for certain times. Total alkalinity is also oscillating annually, with a local maximum located in winter. Moreover, the limit of total alkalinity at 200 mg/L has been exceeded many times since the average sits on 120 mg/L already.

C. Detection results

The testing dataset covers the period from May 15, 2011, to September 1st, 2017, and contains several real abnormal events such as seawater intrusion, discharge from construction area over the limit, and hypochlorite dosage (see Table III). Faults are identified and reported by operators from the WRRF, based on the combination of both downstream process performances and their judgments in practice. Due to several intensive rainfalls and saline water intrusion into the lift station, several variables were reported to be strongly skewed, producing anomalies identified by the operators.

Fault types	Counts
Discharge from construction area	3
InFlow-Total (Rainfall)	5
Internal circulation	3
Lift station maintenance/flushing	2
Lime dumping from RO plant	9
NaOCl dosage	26
Others	6
pH over limit	9
Seawater Intrusion	45
TDS over limit	11
Total alkalinity over limit	36
Water supply shutdown	2
Total	157

TABLE III
SUMMARY OF ABNORMAL EVENTS IN KAUST WRRF.

In this section, the proposed PCA-based KNN soft sensor is compared with the conventional PCA-based SPE, T^2, and residuals-based univariate soft sensors. For PCA-based KNN approach, both Euclidean and Manhattan distances, are used to measure the difference between a new sample and the normal training data. For each soft sensor-based monitoring scheme, detection results are provided via both parametric (p) and nonparametric (np) control limit. Moreover, the performance of detection procedures is quantitatively compared using the following metrics: the probability of detection or the true positive rate (TPR), the probability of false alarm or the false positive rate (FPR), precision or the positive predictive value, and the area under curve (AUC).

Algorithm	TPR	FPR	Precision	AUC
PCA-KNN$_{np}^{Manh}$	**0.882**	**0.067**	**0.931**	**0.908**
PCA-KNN$_{np}^{Eucl}$	0.873	0.072	0.925	0.901
PCA-SPE_{np}	0.853	0.057	0.939	0.898
PCA-KNN$_{p}^{Manh}$	0.971	0.235	0.774	0.868
PCA-KNN$_{p}^{Eucl}$	0.951	0.237	0.771	0.857
PCA-T_p^2	0.657	0.059	0.928	0.799
PCA-T_{np}^2	0.402	0.005	0.969	0.699
PCA-Residuals$_{np}$	0.765	0.533	0.480	0.616
PCA-SPE_p	0.196	0.001	0.964	0.598

TABLE IV
PERFORMANCES OF ANOMALY DETECTION SCHEMES.

1558-1748 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

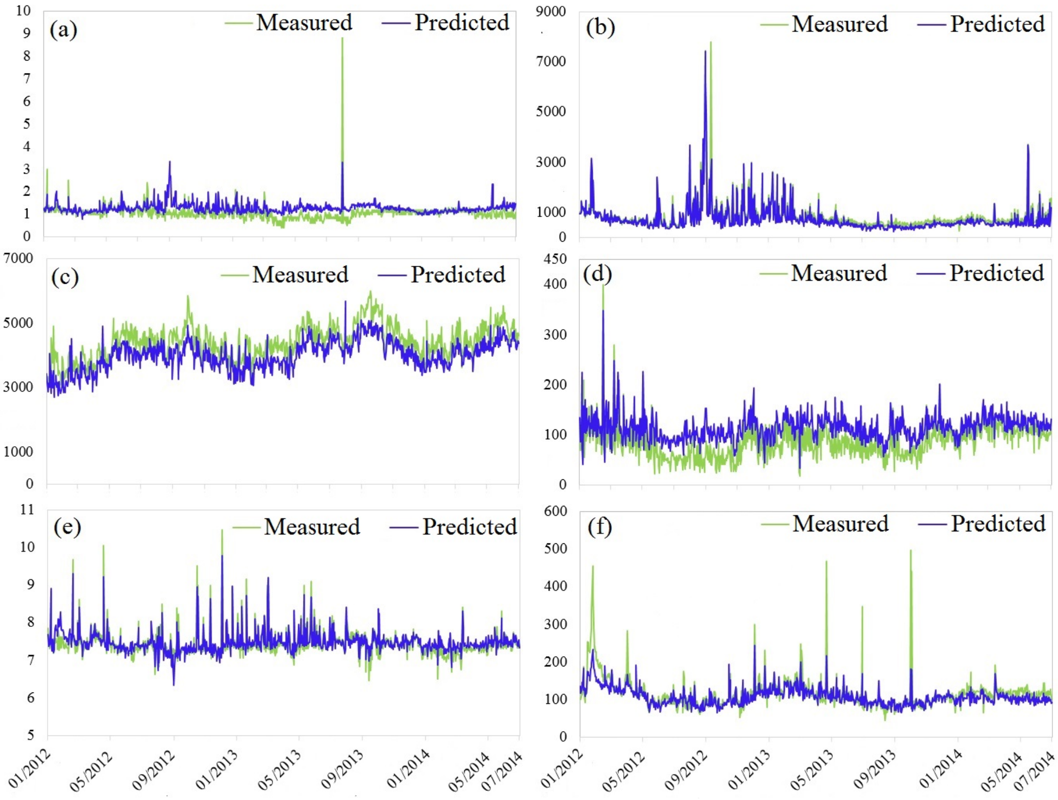

Fig. 5. Historical observations and PCA predictions of typical IMs, (a) boron, (b) conductivity, (c) total inflow, (d) BOD$_5$, (e) pH, and (f) total alkalinity.

Firstly, it is shown in this case study that, PCA is competent in providing acceptable approximations of data in reduced subspace and in decoupling the signal variability from the noise variability. Moreover, it is demonstrated that significant enhancement in the detection of abnormal events in IMs can be reached by using PCA-based nonparametric UCLs instead of the parametric counterparts. The control limits of parametric PCA-based SPE and T^2 charts are determined with the assumption that residuals are normally distributed, which is unreliable when applied to IMs (see Table V). From the Table IV, it can be seen that the conventional PCA-based SPE, T^2 with high FPRs are unsuitable for monitoring IMs.

It can be noticed that the PCA-Residuals$_{np}$ scheme fails in accurately detecting abnormal events in the monitored IM data (see Table IV). PCA-Residuals$_{np}$ scheme is based on a joint detection method comprised of several nonparametric univariate schemes, where each scheme is inspecting a single process variable, and the joint anomaly detection scheme provides a signal of potential anomalies when at least one individual monitoring scheme detects an anomaly. When looking at multivariate data, this approach ignores the interaction between correlated variables and therefore results in a misleading analysis.

From Table IV, nonparametric T^2 exhibits poor anomaly detection performance with high missed detection rate (i.e., TPR=0.402 and FPR=0.005). This result is expected because T^2 scheme detects changes in the principal components space which already possess the majority of total variance in contrast to residuals. In other words, the control limit defined by T^2 is large and moderate anomalies cannot be detected.

As shown in Tables IV, the detection efficiency is greatly enhanced by using the proposed PCA-KNN approach. This fact is due to the flexibility of PCA and the sensitivity of KNN algorithm to small changes in the features. KNN is intuitive and straightforward in implementation and computation complexity and provides a non-parametric approach without assumptions about data distribution or convexity. Another advantage is that when KNN is applied for unsupervised learning, the number of clusters is not required and it is robust to the sequence of input. Furthermore, KNN can handle large dataset with high dimensionalities while keeping robustness to noise. All of the above merits endorse the outperforming of PCA-KNN methods.

In summary, the conventional parametric PCA-based charts provide unsuitable detection performance for the assumption that process variables follow a multivariate normal distribution is not satisfied. This can be confirmed by both skewness and kurtosis of data or statistics as shown in Table II and Table V. Compared to other statistics, KNN distances reserved stronger capacity to represent the dataset and support efficient detectors when normality assumption is violated.

D. Anomaly analysis/diagnosis and detector comparison by radial visualization (RadViz)

To assist fault diagnosis, anomalies are analyzed via RadViz plot, an intuitive tool in high dimensional data exploration. Until recently, RadViz plot has not been utilized in fault

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2018.2875954, IEEE Sensors Journal

	Mean	Std	Min	25%	50%	75%	Max	Skewness	Kurtosis
SPE	12.90	24.29	0.48	6.66	9.65	13.55	536.03	13.52	223.66
T2	14.43	108.68	0.59	4.28	7.15	11.47	4444.02	34.58	1293.41
Manhattan	8.06	3.89	2.25	6.23	7.66	9.10	65.04	6.11	63.46
Euclidean	2.31	1.28	0.68	1.72	2.16	2.62	20.86	6.96	75.11

TABLE V
DESCRIPTIVE STATISTICS OF SPE, T^2, MANHATTAN AND EUCLIDEAN KNN DISTANCES.

analysis or diagnosis. Radviz was developed to directly explore and interpret complex high dimensional datasets such as stock exchange trends or microarray data via 2D projection. Radviz combines the advantages of projection methods, reducing the dimensionality, with that of scatterplots, where the value of each point can be inferred from the distance to the axis. In Radviz, original attributes on IMs are set as ordered dimensional anchors around the circumference of a circle, and observed points are plotted on the interior such that attributes with higher values obtain higher weights, increased attraction, and shorter distances from corresponding anchors [42]. In this study, to enable visualization and interpretation of detected multidimensional anomalies, R package RadViz with an optimized ordering of anchors is employed [42].

In Figure 6, datasets are shown by RadViz plot with their scores on principal components set as anchors and categories marked by shapes and colors. Proximal observations hinted similar pattern, and exterior records signified potential problems of their nearby attributes [42]. The simplification derived from dimension reduction by PCA and the visualization by radial scatterplot are combined here, which together with prior knowledge of PC composition from Figure 4, enabled faults interpretation. Gray points of normal operating conditions are generally scattered in the central area, while various anomalies surrounded them from the second to the sixth PCs in a 'V' shape manner. Anomalies caused by "discharge from construction area" and "TDS over the limit" were closer to the second and third anchors, exactly matching their roles as "TDS" and "Inflow" block-indicators. Rainfall, internal circulation, and lift station flushing let their records closer to the "Inflow" and the fifth pillar, one that's related to FOG and recycled flow. Events including "lime dumping from RO plant", "pH over limit", "total alkalinity over limit" and "water supply shutdown" were accumulated towards the sixth "pH" PC, and affected by the second "TDS" anchor as well as the fourth "pH, alkalinity and calcium hardness" anchor. Interestingly, "seawater intrusion" points were diverged into two locations, either joining ones with "TDS over the limit" or ones with "total alkalinity over the limit." The dosage of sodium hypochlorite seems to raise complex consequences along the "V," or they might be concurrently happening instead of being the causation.

The RadViz plot can also be adopted for the visualization of soft-sensor performances and detailed illustration of Table IV, as implemented in Figure 7. The nonparametric KNN soft-sensor achieved the best balance between precision and sensitivity, perceiving the most anomalies and the least false alarms, with some miss on the ambiguous "sodium hypochlorite dosage" and "seawater intrusion." The nonparametric SPE soft-sensor output worse results on events like "sodium hypochlorite dosage," "alkalinity over the limit," "rainfall" and "alkalinity over the limit." The parametric T^2 soft-sensor, though performed better than its nonparametric counterpart as shown in Table IV, played worst in comparison, showing unsatisfying outcome on "sodium hypochlorite dosage" and "alkalinity over the limit" and triggering most false positives.

V. CONCLUSIONS

The monitoring of IMs is essential for WRRFs operations and is especially valued in KSA. However, IMs following non-Gaussian distributions with high skewness and kurtosis are not suitable for conventional parametric PCA based SPE approach. Hotelling's T^2 displayed insensitivity to small and moderate anomalies since retained PCs captured the majority of variance in NOCs. Joint univariate methods monitor single process variables and may overlook the interactions among correlated variables, resulting in insufficient results. In this study, a multivariate data-driven soft-sensor based on PCA-KNN is proposed and validated by historical IMs data from a WRRF in KSA. PCA performed effective dimension reduction and revealed interrelationships between IMs, while KNN distances demonstrated superior sensing capacity, robustness to underlying data distribution, and efficiency in handling high-dimensional dataset. Nonparametric thresholds derived from KDE further enhanced detection results when compared with parametric ones. Moreover, RadViz is applied for fault analysis and diagnosis in combination with PCA, and delineated innovative interpretable visualization of anomalies and detector performance.

For future work, extensions of kernel PCA and a nonlinear principal component regression would be investigated to capture nonlinear and dynamic relationships among variables. Prior knowledge of faults could be introduced to assist sensing techniques and achieve multi-class fault diagnosis or classification. A higher sampling frequency of IMs may reveal diurnal trends and enable fine-tuned soft-sensing in WRRFs.

ACKNOWLEDGEMENT

This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No: OSR-2015-CRG4-2582.

REFERENCES

[1] M. W. Sweeney and J. C. Kabouris, "Modeling, instrumentation, automation, and optimization of water resource recovery facilities," *Water Environment Research*, vol. 86, no. 10, pp. 1314–1331, 2014.

[2] M. Grossi, R. Lazzarini, M. Lanzoni, A. Pompei, D. Matteuzzi, and B. Riccò, "A portable sensor with disposable electrodes for water bacterial quality assessment," *IEEE Sensors Journal*, vol. 13, no. 5, pp. 1775–1782, 2013.

[3] F. Woodard, *Industrial waste treatment handbook*. Elsevier, 2001.

[4] G. Olsson, "Ica and me–a subjective review," *Water Research*, vol. 46, no. 6, pp. 1585–1624, 2012.

[5] T. Li, M. Winnel, H. Lin, J. Panther, C. Liu, R. O'Halloran, K. Wang, T. An, P. K. Wong, S. Zhang *et al.*, "A reliable sewage quality abnormal event monitoring system," *Water research*, vol. 121, pp. 248–257, 2017.

1558-1748 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

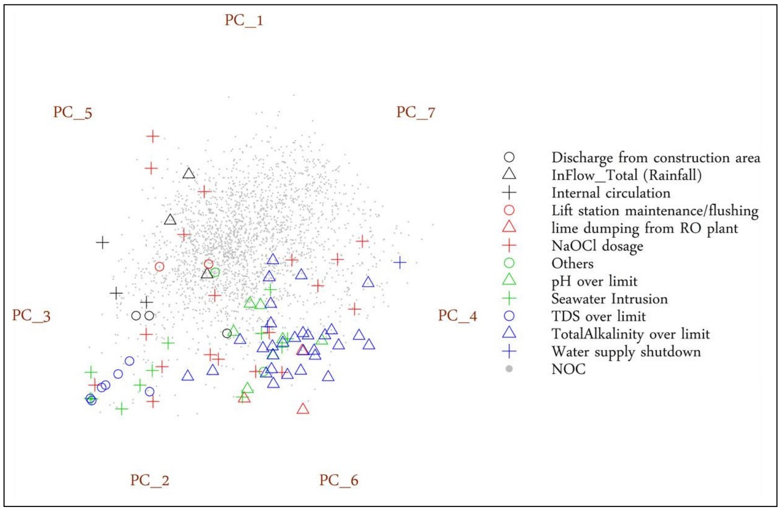

Fig. 6. RadViz of true anomalies.

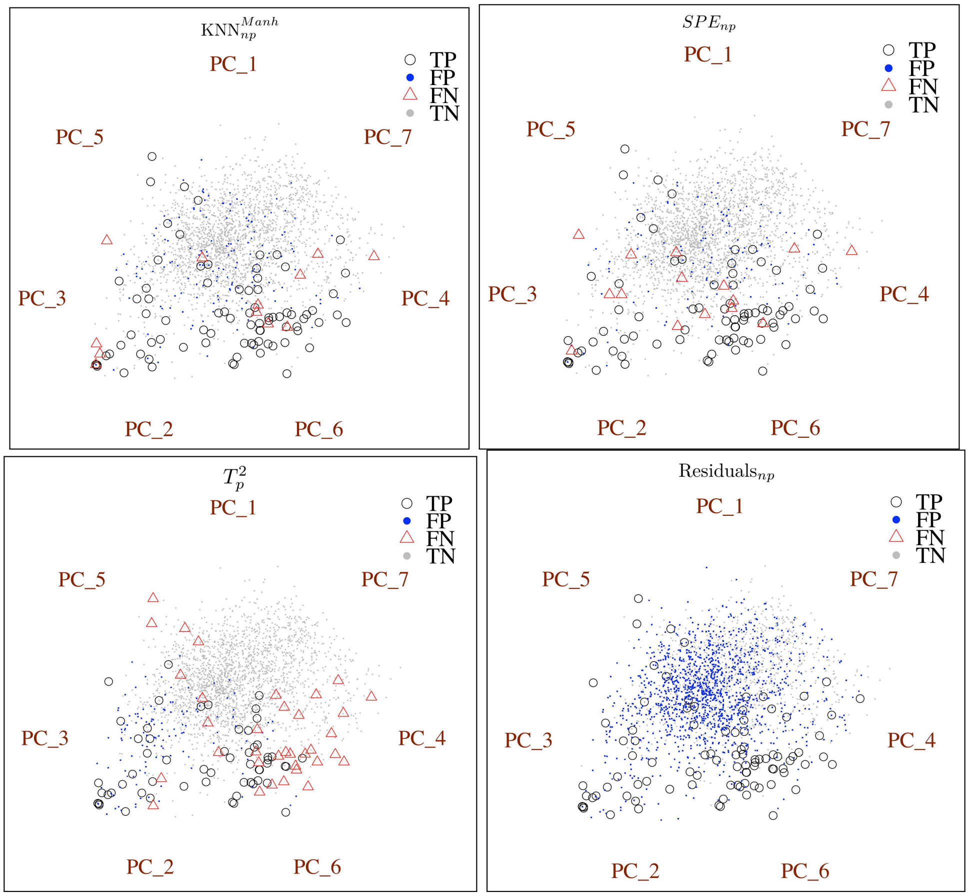

Fig. 7. RadViz of detector performances, including KNN_{np}^{Manh}, SPE_{np}, T_p^2, and Residuals_{np}. TP, FP, FN, and TN are respectively true positive, false positive, false negative and true negative metrics.

[6] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven soft sensors in the process industry," *Computers & chemical engineering*, vol. 33, no. 4, pp. 795–814, 2009.

[7] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, "A review of process fault detection and diagnosis: Part iii: Process history based methods," *Computers & chemical engineering*, vol. 27, no. 3, pp. 327–346, 2003.

[8] H. Haimi, M. Mulas, F. Corona, and R. Vahala, "Data-derived soft-sensors for biological wastewater treatment plants: An overview," *Environmental modelling & software*, vol. 47, pp. 88–107, 2013.

[9] J. Xu and Z. Li, "A review on ecological engineering based engineering management," *Omega*, vol. 40, no. 3, pp. 368–378, 2012.

[10] M. Álvarez-Cabria, J. Barquín, and F. J. Peñas, "Modelling the spatial and seasonal variability of water quality for entire river networks: Relationships with natural and anthropogenic factors," *Science of The Total Environment*, vol. 545, pp. 152–162, 2016.

[11] A. G. Capodaglio, H. V. Jones, V. Novotny, and X. Feng, "Sludge bulking analysis and forecasting: application of system identification and artificial neural computing technologies," *Water Research*, vol. 25, no. 10, pp. 1217–1224, 1991.

1558-1748 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2018.2875954, IEEE Sensors Journal

[12] P. Berthouex and G. E. Box, “Time series models for forecasting wastewater treatment plant performance,” *Water Research*, vol. 30, no. 8, pp. 1865–1875, 1996.

[13] J. Huo, W. L. Seaver, R. B. Robinson, and C. D. Cox, “Application of time series models to analyze and forecast the influent components of wastewater treatment plants (wwtps),” in *Impacts of Global Climate Change*, 2005, pp. 1–11.

[14] A. Escalas-Cañellas, C. J. Ábrego-Góngora, M. G. Barajas-López, D. Houweling, and Y. Comeau, “A time series model for influent temperature estimation: application to dynamic temperature modelling of an aerated lagoon,” *Water research*, vol. 42, no. 10-11, pp. 2551–2562, 2008.

[15] S. Wilcox, D. Hawkes, F. Hawkes, and A. Guwy, “A neural network, based on bicarbonate monitoring, to control anaerobic digestion,” *Water Research*, vol. 29, no. 6, pp. 1465–1470, 1995.

[16] A. Dias, M. Alves, and E. Ferreira, “Application of computational intelligence techniques for monitoring and prediction of biological wastewater treatment systems,” in *Proceedings of the Int. IWA Conf. on Automation in Water Quality Monitoring, 3: Gent, Belgium*. Springer, 2007, pp. 1–8.

[17] C.-H. Lin, R.-F. Yu, W.-P. Cheng, and C.-R. Liu, “Monitoring and control of uv and uv-tio2 disinfections for municipal wastewater reclamation using artificial neural networks,” *Journal of hazardous materials*, vol. 209, pp. 348–354, 2012.

[18] J.-J. Zhu, L. Kang, and P. R. Anderson, “Predicting influent biochemical oxygen demand: Balancing energy demand and risk management,” *Water research*, vol. 128, pp. 304–313, 2018.

[19] S. Marsili-Libelli, “Control of sbr switching by fuzzy pattern recognition,” *Water Research*, vol. 40, no. 5, pp. 1095–1107, 2006.

[20] F. Fang, B. Ni, W. Li, G. Sheng, and H. Yu, “A simulation-based integrated approach to optimize the biological nutrient removal process in a full-scale wastewater treatment plant,” *Chemical Engineering Journal*, vol. 174, no. 2-3, pp. 635–643, 2011.

[21] A. Amaral and E. Ferreira, “Activated sludge monitoring of a wastewater treatment plant using image analysis and partial least squares regression,” *Analytica Chimica Acta*, vol. 544, no. 1-2, pp. 246–253, 2005.

[22] S.-P. Mujunen, P. Minkkinen, P. Teppola, and R.-S. Wirkkala, “Modeling of activated sludge plants treatment efficiency with plsr: a process analytical case study,” *Chemometrics and intelligent laboratory systems*, vol. 41, no. 1, pp. 83–94, 1998.

[23] L. Eriksson, P. Hagberg, E. Johansson, S. Rännar, O. Whelehan, A. Åström, and T. Lindgren, “Multivariate process monitoring of a newsprint mill. application to modelling and predicting COD load resulting from de-inking of recycled paper,” *Journal of chemometrics*, vol. 15, no. 4, pp. 337–352, 2001.

[24] X. Wang, H. Ratnaweera, J. A. Holm, and V. Olsbu, “Statistical monitoring and dynamic simulation of a wastewater treatment plant: a combined approach to achieve model predictive control,” *Journal of environmental management*, vol. 193, pp. 1–7, 2017.

[25] M. Ebrahimi, E. L. Gerber, and T. D. Rockaway, “Temporal performance assessment of wastewater treatment plants by using multivariate statistical analysis,” *Journal of environmental management*, vol. 193, pp. 234–246, 2017.

[26] F. Harrou, Y. Sun, M. Madakyaru, and B. Bouyedou, “An improved multivariate chart using partial least squares with continuous ranked probability score,” *IEEE Sensors Journal*, vol. 18, no. 16, pp. 6715–6726, 2018.

[27] Q. P. He and J. Wang, “Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 20, no. 4, pp. 345–354, 2007.

[28] P. B. Garcia-Allende, O. M. Conde, J. Mirapeix, A. M. Cubillas, and J. M. López-Higuera, “Data processing method applying principal component analysis and spectral angle mapper for imaging spectroscopic sensors,” *IEEE Sensors Journal*, vol. 8, no. 7, pp. 1310–1316, 2008.

[29] F. Harrou, M. N. Nounou, H. N. Nounou, and M. Madakyaru, “Statistical fault detection using PCA-based GLR hypothesis testing,” *Journal of loss prevention in the process industries*, vol. 26, no. 1, pp. 129–139, 2013.

[30] A. Perera, N. Papamichail, N. Bârsan, U. Weimar, and S. Marco, “On-line novelty detection by recursive dynamic principal component analysis and gas sensor arrays under drift conditions,” *IEEE Sensors Journal*, vol. 6, no. 3, pp. 770–783, 2006.

[31] F. Harrou, M. Nounou, and H. Nounou, “Statistical detection of abnormal ozone levels using principal component analysis,” *ternational Journal of Engineering & Technology*, vol. 12, no. 6, pp. 54–59, 2012.

[32] J. E. Jackson and G. S. Mudholkar, “Control procedures for residuals associated with principal component analysis,” *Technometrics*, vol. 21, no. 3, pp. 341–349, 1979.

[33] S. Joe Qin, “Statistical process monitoring: basics and beyond,” *Journal of chemometrics*, vol. 17, no. 8-9, pp. 480–502, 2003.

[34] S. J. Sheather and M. C. Jones, “A reliable data-based bandwidth selection method for kernel density estimation,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 683–690, 1991.

[35] J. Xiong, Q. Zhang, G. Sun, X. Zhu, M. Liu, and Z. Li, “An information fusion fault diagnosis method based on dimensionless indicators with static discounting factor and KNN,” *IEEE Sensors Journal*, vol. 16, no. 7, pp. 2060–2069, 2016.

[36] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, “Top 10 algorithms in data mining,” *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008.

[37] A. Dairi, F. Harrou, Y. Sun, and M. Senouci, “Obstacle detection for intelligent transportation systems using deep stacked autoencoder and k-nearest neighbor scheme,” *IEEE Sensors Journal*, vol. 18, no. 12, pp. 5122–5132, 2018.

[38] D. J. Hill and B. S. Minsker, “Anomaly detection in streaming environmental sensor data: A data-driven modeling approach,” *Environmental Modelling & Software*, vol. 25, no. 9, pp. 1014–1022, 2010.

[39] M. Kim, Y. Kim, H. Kim, W. Piao, and C. Kim, “Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant,” *Frontiers of Environmental Science & Engineering*, vol. 10, no. 2, pp. 299–310, 2016.

[40] J. Honaker, G. King, M. Blackwell *et al.*, “Amelia II: A program for missing data,” *Journal of statistical software*, vol. 45, no. 7, pp. 1–47, 2011.

[41] D. Fang, G. Zhao, X. Xu, Q. Zhang, Q. Shen, Z. Fang, L. Huang, and F. Ji, “Microbial community structures and functions of wastewater treatment systems in plateau and cold regions,” *Bioresource technology*, vol. 249, pp. 684–693, 2018.

[42] Y. Abraham, B. Gerrits, M.-G. Ludwig, M. Rebhan, and C. Gubser Keller, “Exploring glucocorticoid receptor agonists mechanism of action through mass cytometry and radial visualizations,” *Cytometry Part B: Clinical Cytometry*, vol. 92, no. 1, pp. 42–56, 2017.

1558-1748 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.