Submitted to the Annals of Applied Statistics

ONLINE MONITORING OF AIR QUALITY USING PCA-BASED SEQUENTIAL LEARNING

BY XIULIN XIE[1,a], NICOLE QIAN[2,b] AND PEIHUA QIU[3,c]

[1]*Department of Statistics, Florida State University,* [a]*xiulin.xie@ufl.edu*

[2]*Buchholz High School, Gainesville, Florida ,* [b]*starrynight386@gmail.com*

[3]*Department of Biostatistics, University of Florida,* [c]*pqiu@ufl.edu*

Air pollution surveillance is critically important for public health. One air pollutant, ozone, is extremely challenging to analyze properly, as it is a secondary pollutant caused by complex chemical reactions in the air, and does not emit directly into the atmosphere. Numerous environmental studies confirm that ozone concentration levels are associated with meteorological conditions, and long-term exposure to high ozone concentration levels is associated with the incidence of many diseases, including asthma, respiratory, and cardiovascular diseases. Thus, it is important to develop an air pollution surveillance system to collect both air pollution and meteorological data and monitor the data continuously over time. To this end, statistical process control (SPC) charts provide a major statistical tool. But, most existing SPC charts are designed for cases when the in-control (IC) process observations at different times are assumed to be independent and identically distributed. The air pollution and meteorological data would not satisfy these conditions due to serial data correlation, high dimensionality, seasonality, and other complex data structure. Motivated by an application to monitor the ground ozone concentration levels in the Houston-Galveston-Brazoria (HGB) area, we developed a new process monitoring method using principal component analysis and sequential learning. The new method can accommodate high dimensionality, time-varying IC process distribution, serial data correlation, and nonparametric data distribution. It is shown to be a reliable analytic tool for online monitoring of air quality.

1. Introduction. Ozone has become one of the most harmful pollutants, and long-term exposure to high ozone concentration levels can cause many health problems, including asthma, respiratory and cardiovascular diseases (Carey et al. (2013), Jenkin and Clemitshaw (2000), Health Effects Institute (2019)). It is growing into a major threat to public health in many areas of the world. For example, Houston has been classified as a severe ozone nonattainment area under the Clean Air Act. To demonstrate the ozone pollution in Houston, Figure 1 presents four pictures of the Houston downtown area on a clear day versus ozone days. From the pictures, it can be seen that heavy smog was lingering over the Houston downtown area on ozone days.

Unlike other air pollutants, ozone is not emitted directly into the atmosphere from manufacturing factories and other industrial operations. Instead, it is produced by complex chemical reactions of nitrogen oxides and volatile organic compounds under some weather conditions (Jenkin and Clemitshaw (2000), World Health Organization (1976)). It has been confirmed through environmental research that meteorological conditions can substantially influence air quality, especially ozone concentration (Gorai et al. (2015), Jacob and Winner (2009), Liu et al. (2020)). Thus, to monitor the ozone concentration levels effectively, it is

Keywords and phrases: Air pollution surveillance, Dynamic processes, High-dimensional data, Principal component analysis, Seasonality, Self-starting charts.

FIG 1. *Pictures of the Houston Downtown area on a clear day (upper-left panel) and ozone days (remaining panels).*

important to monitor the observed data of certain meteorological variables properly. However, meteorological data often have complicated structures, including high dimensionality, dynamic longitudinal pattern (e.g., seasonality), and serial data correlation (Ordóñez et al. (2005), Zhao et al. (2009)).

Because of the importance of effective monitoring of ozone concentration levels, the Texas Commission on Environmental Quality (TCEQ) has established a surveillance system to collect meteorological data in Houston. They also developed a parametric model to predict ozone concentration levels based on several meteorological variables, such as wind speed, air temperature, and solar radiation (Environmental Protection Agency (1999)). However, this model cannot properly accommodate the complicated data structure of the observed meteorological data, including serial data correlation and complex dynamic patterns like seasonality and day-of-the-week variation. In addition, the observed meteorological data often have high dimensionality, and the related meteorological variables are usually associated with each other (Abdul-Wahab et al. (2005), Zhang and Fan (2008)). To address the issue of high dimensionality, principal component analysis (PCA) has been used in environmental studies to investigate the association between ozone concentration levels and meteorological variables (e.g., Abdul-Wahab et al. (2005), Statheropoulos et al. (1998)). The related PCA-based methods, however, are all retrospective in the sense that the time interval of process observations needs to be pre-specified. Thus, these methods cannot effectively monitor the air quality of a region sequentially over time.

To monitor a sequential process online, a major statistical tool is the statistical process control (SPC) charts (cf., Montgomery (2012), Qiu (2014)), including various Shewhart charts, cumulative sum (CUSUM) charts, exponentially weighted moving average (EWMA) chart, and charts based on change-point detection (CPD) (cf., Hawkins et al. (2003), Page (1954), Roberts (1959), Shewhart (1931)). Early SPC charts are designed mainly for monitoring processes with a single quality variable. For monitoring multiple quality variables, a number of multivariate SPC charts have also been developed, including the Hotelling's T^2 chart, multivariate CUSUM chart, multivariate EWMA chart, and more (cf., Crosier (1988), Hotelling

(1947), Lowry et al. (1992), Chapter 7 in Qiu (2014)). A related SPC problem is to monitor profiles that describe the functional relationship between response variables and predictors. See, for instance, Chicken et al. (2009), Qiu et al. (2010) and Noorossana et al. (2011) for related discussions. To monitor high-dimensional processes, some control charts based on variable selection have been developed (cf., Capizzi and Masarotto (2011), Wang and Jiang (2009), Zou and Qiu (2009)). These methods require the sparsity assumption that shifts in a high-dimensional process can only occur in a small number of quality variables. Some other control charts designed for monitoring high-dimensional processes are based on the maximum, summation, or other summaries of the CUSUM charting statistics constructed for monitoring individual quality variables (cf., Mei (2010), Tartakovsky et al. (2006), Zou et al. (2015)). In addition, some PCA-based control chart have been developed for monitoring high-dimensional processes, where the PCA technique is used for reducing the dimensionality of quality variables (cf., Ferrer (2007), Jackson (1991), Kourti and MacGregor (1996)).

However, most of these existing charts require various assumptions on the observed data, including that the IC process distribution does not change over time, the process observations are independent at different observation times, and the IC process observations follow a parametric (e.g., normal) distribution. These assumptions are rarely valid in applications like air quality monitoring. In the SPC literature, it has been well demonstrated that such control charts would become unreliable to use when one or more of their assumptions are violated (Apley and Tsung (2002), Qiu (2018), Qiu and Xiang (2014)). To address this issue, some new SPC charts have been developed recently. For instance, to monitor univariate correlated data, many control charts have been developed based on parametric time series modelling and sequential monitoring of the resulting residuals (cf., Apley and Tsung (2002), Capizzi and Masarotto (2008), Knoth and Schmid (2004), Psarakis and Papaleonida (2007)). One limitation of these residual-based charts is that their performance is sensitive to the assumed parametric time series models that could be invalid in practice (cf., Qiu et al. (2020)). As discussed in some previous studies (e.g., Ku et al. (1995), Vanhatalo and Kulahci (2016)), the traditional PCA-based control charts would also become unreliable to use when the process under monitoring has a time-varying IC distribution and the observed data are serially correlated. So, some alternative PCA-based control charts have been developed to monitor different processes with data autocorrelation. For instance, the dynamic PCA method accommodates autocorrelation by modelling time-lagged data together with the data at the current observation time (e.g., Ku et al. (1995), Tsung (2000)). The PCA method based on latent variables models the data correlation using latent factors (e.g., Dong and Qin (2018), Li et al. (2014)). These alternative PCA methods can accommodate different types of dynamic data correlation, including both serial correlation and cross-component correlation in multivariate cases. But, they are designed for cases when the IC process distribution does not change over time. To accommodate time-varying IC process distribution, some moving window PCA methods have been developed by using a moving window of a pre-specified size to perform PCA (e.g., Lennox et al. (2001), Wang et al. (2005)). For an overview on PCA-based control charts, see De Ketelaere et al. (2015). However, the existing PCA-based control charts have several fundamental limitations. First, almost all of them are Shewhart charts that make decisions about the process performance at a given time point based solely on the observed data at that time point, and all historical data are ignored by them. Consequently, they are less effective in detecting relatively small and persistent shifts, compared to other types of charts such as the CUSUM, EWMA, and CPD charts (cf., Qiu (2014)). Second, a large IC dataset is often needed for these existing charts for estimating certain IC parameters in order to have a reliable IC performance, which is usually unavailable in the current air quality monitoring problem. Third, they cannot handle cases with serial data correlation and time-varying IC process distributions that are hard to describe by a parametric distribution family. However,

such cases are realistic in the air quality monitoring applications. As far as we know, there are no existing methods that can properly accommodate the complex data structure in air quality monitoring applications. This paper aims to fill the gap.

In this paper, we suggest a flexible method for sequential monitoring of high-dimensional dynamic processes with serially correlated data. The serial data correlation is assumed to be short-ranged in the sense that the correlation between two process observations is weaker when their observation times are farther away, which should be reasonable in air quality monitoring applications. Because of this assumption, the correlation between two process observations can be ignored if the two observation times are at least b_{max} apart, where b_{max} denotes the autocorrelation range. The autocorrelation is assumed such that the correlation between two process observations depends on the distance between the two observation times only, which is routinely assumed in the SPC literature on monitoring processes with serially correlated data (e.g., Apley and Tsung (2002), Capizzi and Masarotto (2008), Xie and Qiu (2023)). The proposed method also requires an initial IC dataset to be available before online process monitoring, from which an initial estimate of the IC temporal pattern of the process under monitoring can be obtained. During online process monitoring, the observed data at the current observation time are first standardized using the estimated IC temporal pattern, and the PCA procedure is then applied to the standardized data for dimension reduction. After proper data decorrelation for the selected principal components of the observed data, a multivariate CUSUM chart is constructed for process monitoring. In addition, our proposed chart is self-starting (cf., Hawkins (1987)) in the sense that if the related process is declared to be IC at the current observation time, then the observed data at that time point is combined with the available IC data and the estimate of the IC temporal pattern of the process gets updated using the combined IC data for process monitoring at the next observation time. For updating the estimates of the related IC quantities, recursive formulas are derived to substantially save the computing time, memory, and storage requirement for storing all previous data, since the estimate updates need to be implemented at each observation time. Unlike the existing PCA-based charts, the new method allows the IC process distribution to vary over time and be unconstrained to a specific parametric distribution family. Numerical studies show that it is reliable and effective for air quality monitoring applications.

The remainder of the paper is organized as follows. Our proposed new method is described in detail in Section 2. Some simulation studies to evaluate its numerical performance are presented in Section 3, in comparison with several representative existing methods. Its applications to monitor two air quality data collected in the Houston area are discussed in Sections 4. Some remarks conclude the paper in Section 5. Proof of a theoretical result and some numerical examples are given in a supplementary file.

2. High-Dimensional Process Monitoring by PCA-Based Sequential learning. Assume that $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ is a vector of p numerical quality variables for monitoring a sequential dynamic process. Its observation at time n is denoted as $\mathbf{X}_n = (X_{n1}, X_{n2}, \ldots, X_{np})'$. To monitor the sequential process $\{\mathbf{X}_n, n \geq 1\}$ online, an initial IC dataset $\mathcal{X}_{IC}^{(0)} = \{\mathbf{X}_{-m_0+1}, \mathbf{X}_{-m_0+2}, \ldots, \mathbf{X}_0\}$ of size m_0 is assumed to be available in advance. The time period of the initial IC dataset is then set as a baseline time interval. The main goal of online process monitoring is to detect any substantial deviation in the future temporal pattern of the process from its regular temporal pattern in the baseline time interval as promptly as possible. If seasonality is present in the temporal pattern of the process, then the baseline time interval should contain at least one whole season. Then, our proposed method proceeds in the following several steps, as demonstrated in Figure 2. First, it computes initial estimates of the regular temporal pattern of the dynamic process under monitoring and other IC quantities from $\mathcal{X}_{IC}^{(0)}$. Second, at the current time point n during online

process monitoring, the observed data $\mathbf{X}_n$ are first standardized using the estimated regular temporal pattern, and then the PCA procedure is applied to the standardized data to reduce their dimensionality and obtain the principal component (PC) data $\mathbf{Y}_n$. Third, the PC data $\mathbf{Y}_n$ at time n is decorrelated with the PC data at previous observation times. Fourth, a control chart is applied to the decorrelated PC data. If the control chart does not give a signal at time n, then the observed data $\mathbf{X}_n$ at time n are combined with the IC data at the previous time point, denoted as $\mathcal{X}_{IC}^{(n-1)} = \{\mathbf{X}_j, -m_0 + 1 \le j \le n-1\}$. The estimates of the regular temporal pattern and other IC quantities now get updated using the combined IC data $\mathcal{X}_{IC}^{(n)} = \{\mathbf{X}_j, -m_0 + 1 \le j \le n\}$. The online process monitoring then proceeds to the next time point, $n+1$. A more detailed description of the proposed method is given in several subsections below.

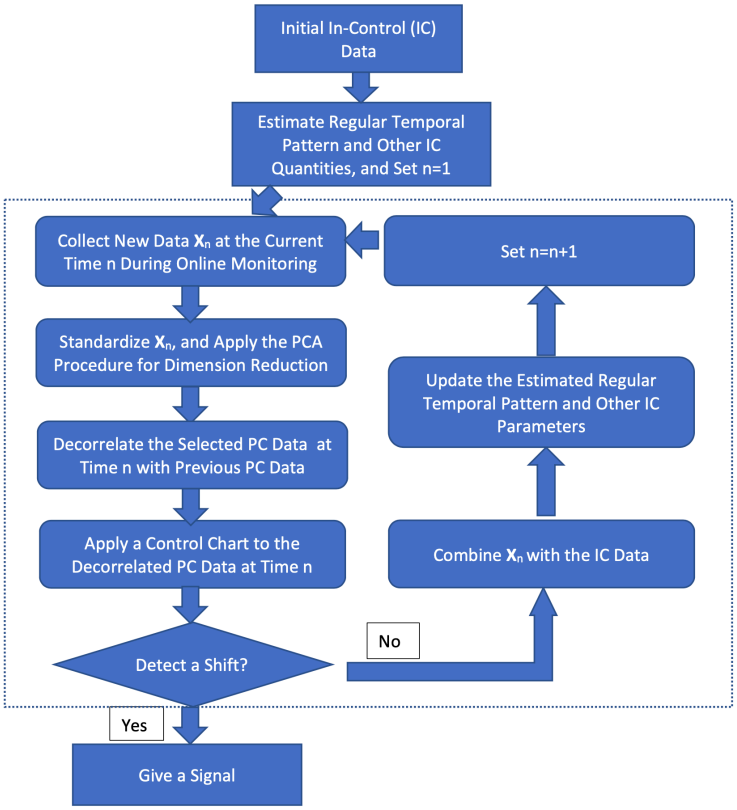

FIG 2. *Diagram of the proposed method for monitoring high-dimensional dynamic processes based on PCA and sequential learning. The dashed rectangle highlights the proposed sequential learning procedure.*

2.1. *Initial estimation of the regular temporal pattern and other IC quantities.* Assume that process observations in the initial IC dataset $\mathcal{X}_{IC}^{(0)}$ follow the nonparametric longitudinal model

$$\mathbf{X}_j = \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_j, \qquad \text{for } j = -m_0 + 1, -m_0 + 2, \ldots, 0, \tag{1}$$

where $\boldsymbol{\mu}_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jp})$ is the mean of $\mathbf{X}_j$, and $\boldsymbol{\epsilon}_j$ is the p-dimensional zero-mean error term. In Model (1), it is assumed that the serial correlation is short-ranged and stationary in the sense that $\mathrm{Cov}(\boldsymbol{\epsilon}_j, \boldsymbol{\epsilon}_{j'}) \approx \mathbf{0}$ when $|j - j'| > b_{max}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_j, \boldsymbol{\epsilon}_{j'})$ depends on j and j' through $|j - j'|$, where b_{max} denotes the autocorrelation range. No parametric forms are imposed on the error distribution and the temporal pattern of the mean vectors $\{\boldsymbol{\mu}_j\}$.

In Model (1), the mean vector $\boldsymbol{\mu}_j$ can be estimated by the local linear kernel (LLK) smoothing procedure (cf., Xiang et al. (2013)). In matrix notation, let $\mathbf{W}^{(0)} = (X_{-m_0+1,1}, \ldots, X_{0,1}, \ldots, X_{-m_0+1,p}, \ldots, X_{0,p})'$, $\mathbf{Z}_j = [(1, -m_0+1-j)', \ldots, (1,-j)']'$, and $\mathbf{K}_j = \text{diag}\{K(\frac{i-j}{h_l}), i = -m_0+1, -m_0+2, \ldots, 0, l = 1, 2, \ldots, p\}$, where $K(\cdot)$ is a kernel function and $\{h_l, l = 1, 2, \ldots, p\}$ are bandwidths. Then, the initial estimate of $\boldsymbol{\mu}_j$, for $j = -m_0+1, -m_0+2, \ldots, 0$, is given by:

$$\widehat{\boldsymbol{\mu}}_j^{(0)} = \left[[\mathbf{S}_j^{(0)}]^{-1}\mathbf{R}_j^{(0)}\right]' (I_{p\times p} \otimes \boldsymbol{\xi}_1), \tag{2}$$

where $\boldsymbol{\xi}_1 = (1,0)'$, $\mathbf{S}_j^{(0)} = (I_{p\times p} \otimes \mathbf{Z}_j)'\mathbf{K}_j(I_{p\times p} \otimes \mathbf{Z}_j)$, and $\mathbf{R}_j^{(0)} = (I_{p\times p} \otimes \mathbf{Z}_j)'\mathbf{K}_j\mathbf{W}^{(0)}$. In the above LLK procedure, the kernel function $K(\cdot)$ is usually chosen to be the Epanechnikov kernel function, i.e., $K(u) = \frac{3}{4}(1-u^2)I(|u| \le 1)$, since it was shown to be optimal in estimating $\boldsymbol{\mu}_j$ under some regularity conditions (Epanechnikov (1969), Yang and Qiu (2018)). Regarding the initial estimate $\widehat{\boldsymbol{\mu}}_j^{(0)}$ in (2), its different components can be computed separately since $\mathbf{K}_j$ is a diagonal matrix. Thus, the computation is relatively simple.

For choosing the bandwidths $\{h_l, l = 1, 2, \ldots, p\}$, it has been well discussed in the literature that the conventional cross-validation (CV) procedure would not perform well when process observations at different time points are serially correlated, since the CV procedure cannot properly distinguish the data correlation structure from the data mean function (e.g., Altman (1990), Opsomer et al. (2001)). Thus, we propose choosing them by using a modified cross-validation (MCV) procedure as suggested by De Brabanter et al. (2011) in the univariate regression setup with correlated data. For each bandwidth h_l, for $l = 1, 2, \ldots, p$, the MCV score is defined to be:

$$\text{MCV}(h_l) = \frac{1}{m_0}\sum_{j=-m_0+1}^{0} (X_{jl} - \widehat{\mu}_{-j,l})' (X_{jl} - \widehat{\mu}_{-j,l}),$$

where $\widehat{\mu}_{-j,l}$ is the leave-one-out estimate of μ_{jl} by (2) when the observation X_{jl} is excluded in the computation and the kernel function $K(\cdot)$ is modified into

$$K_\varepsilon(u) = \frac{4}{4-3\varepsilon-\varepsilon^3}\begin{cases} \frac{3}{4}(1-u^2)I(|u| \le 1), & \text{when } |u| \ge \varepsilon, \\ \frac{3(1-\varepsilon^2)}{4\varepsilon}|u|, & \text{when } |u| < \varepsilon, \end{cases}$$

where $\varepsilon \in (0,1)$ is a small constant. The modified kernel function $K_\varepsilon(u)$ equals 0 at $u = 0$ and is small around $u = 0$, to diminish the impact of data autocorrelation on bandwidth selection. Then, h_l can be chosen by minimizing the above MCV score.

2.2. *Dimension reduction by PCA for the initial IC data.* PCA is a popular statistical tool for reducing the dimensionality of a dataset by projecting the original dataset into a lower-dimensional space without losing much information. The first principal component (PC) of the p-dimensional random vector $\mathbf{X}$ is defined to be the linear combination $\mathbf{u}_1'\mathbf{X}$ with the maximum variance, where $\mathbf{u}_1$ is a coefficient vector with a unit length, and the second PC is defined to be the linear combination $\mathbf{u}_2'\mathbf{X}$ with the maximum variance, where $\mathbf{u}_2$ is a coefficient vector with a unit length that is orthogonal to $\mathbf{u}_1$, and so forth. The PCs can be obtained by the eigenvalue-eigenvector decomposition of the covariance matrix of $\mathbf{X}$. In the current research problem, dimension reduction is needed because p is assumed large, and the related computation would be exceptionally extensive otherwise. To this end, the PCA procedure is considered, which is described below for analyzing the initial IC data.

For the lth quality variable in $\mathbf{X}$, its standardized observation at time j is defined to be

$$\widetilde{X}_{jl} = \left(X_{jl} - \widehat{\mu}_{jl}^{(0)}\right)/\widehat{\sigma}_{jl}^{(0)}, \text{ for } l = 1, 2, \ldots, p, \ j = -m_0+1, -m_0+2, \ldots, 0,$$

where $\widehat{\mu}_{jl}^{(0)}$ is defined in (2), and $\widehat{\sigma}_{jl}^{(0)}$ is an initial estimate of the standard deviation of X_{jl} defined to be $\widehat{\sigma}_{jl}^{(0)} = \sqrt{\alpha_{jl}^{(0)}/\beta_{jl}^{(0)}}$ in which

$$\alpha_{jl}^{(0)} = \sum_{i=-m_0+1}^{0} \left(X_{jl} - \widehat{\mu}_{jl}^{(0)}\right)^2 K\left(\frac{i-j}{g_l}\right),$$

$$\beta_{jl}^{(0)} = \sum_{i=-m_0+1}^{0} K\left(\frac{i-j}{g_l}\right),$$

and g_l is a bandwidth that can be chosen by the MCV procedure discussed earlier. Then, the initial estimate of the covariance matrix of the standardized observations $\widetilde{\mathbf{X}}_j = (\widetilde{X}_{j1}, \widetilde{X}_{j2}, \ldots, \widetilde{X}_{jp})'$ can be defined to be

$$\widehat{\mathbf{\Sigma}}^{(0)} = \frac{1}{m_0} \sum_{j=-m_0+1}^{0} \widetilde{\mathbf{X}}_j \widetilde{\mathbf{X}}_j'.$$

For $\widehat{\mathbf{\Sigma}}^{(0)}$, its eigenvalue-eigenvector decomposition is assumed to be $\widehat{\mathbf{\Sigma}}^{(0)} = [\widehat{\mathbf{U}}^{(0)}]\widehat{\mathbf{\Lambda}}^{(0)}[\widehat{\mathbf{U}}^{(0)}]'$, where $\widehat{\mathbf{\Lambda}}^{(0)} = \text{diag}(\widehat{\lambda}_1^{(0)}, \widehat{\lambda}_2^{(0)}, \ldots, \widehat{\lambda}_p^{(0)})$ is a diagonal matrix with its diagonal elements $\widehat{\lambda}_1^{(0)} \geq \widehat{\lambda}_2^{(0)} \geq \ldots \geq \widehat{\lambda}_p^{(0)}$ being the eigenvalues of $\widehat{\mathbf{\Sigma}}^{(0)}$, and $\widehat{\mathbf{U}}^{(0)} = (\widehat{\mathbf{u}}_1^{(0)}, \widehat{\mathbf{u}}_2^{(0)}, \ldots, \widehat{\mathbf{u}}_p^{(0)})$ is an orthonormal matrix with its columns being the corresponding eigenvectors of $\widehat{\mathbf{\Sigma}}^{(0)}$.

The PCs of $\widetilde{\mathbf{X}}_j$ are the linear combinations $(\widehat{\mathbf{u}}_l^{(0)})'\widetilde{\mathbf{X}}_j$, for each j and $l = 1, 2, \ldots, p$. To determine the number of PCs to use in the subsequent analysis, we suggest using the percentage of data variation explained by the selected PCs as a criterion, as recommended in the literature (e.g., Johnson and Wichern (2008)). To this end, let v be a pre-specified thresholding percentage value. Then, the first $d^{(0)}$ PCs will be selected, where $d^{(0)}$ is defined by

$$d^{(0)} = \min\left\{q : \left(\sum_{l=1}^{q} \widehat{\lambda}_l^{(0)} \Big/ \sum_{l=1}^{p} \widehat{\lambda}_l^{(0)}\right) \times 100\% > v, 1 \leq q \leq p\right\}. \tag{3}$$

In the literature, it has been well explained that $\widehat{\lambda}_l^{(0)}$ measures the percentage of information in the standardized observations $\{\widetilde{\mathbf{X}}_j, -m_0 + 1 \leq j \leq 0\}$ that can be explained by the lth PC, for each l. Thus, Expression (3) implies that we select the first $d^{(0)}$ PCs such that the percentage of information contained in these PCs about the standardized initial IC data is at least the pre-specified value of v. In practice, v is often selected to be 80%, 90% or 95%, depending on the specific research problem, and the resulting $d^{(0)}$ is usually much smaller than p to achieve the dimension reduction purpose. Then, instead of using the original standardized observations $\{\widetilde{\mathbf{X}}_j, -m_0 + 1 \leq j \leq 0\}$, we can use the related PCs $\{\mathbf{Y}_{-m_0+1}, \mathbf{Y}_{-m_0+2}, \ldots, \mathbf{Y}_0\}$ for process monitoring, where $\mathbf{Y}_j = (\widehat{\mathbf{U}}_{d^{(0)}}^{(0)})'\widetilde{\mathbf{X}}_j$ and $\widehat{\mathbf{U}}_{d^{(0)}}^{(0)} = (\widehat{\mathbf{u}}_1^{(0)}, \widehat{\mathbf{u}}_2^{(0)}, \ldots, \widehat{\mathbf{u}}_{d^{(0)}}^{(0)})$ is the $(p \times d^{(0)})$-dimensional PCA projection matrix.

As mentioned at the beginning of Subsection 2.1, two original process observations are allowed to be correlated if their observation times are within b_{max} apart and their serial correlation does not change over time. Because the selected PCs are linear combinations of different components of the original process observations, they would share these serial correlation properties. Let $\boldsymbol{\gamma}^{(0)}(s) = \text{Cov}(\mathbf{Y}_j, \mathbf{Y}_{j+s})$, for each j and s. Then, $\boldsymbol{\gamma}^{(0)}(s)$ can be estimated by the following moment estimate:

$$\widehat{\boldsymbol{\gamma}}^{(0)}(s) = \frac{1}{m_0 - s} \sum_{j=-m_0+1}^{-s} \mathbf{Y}_{j+s}\mathbf{Y}_j', \qquad \text{for } 0 \leq s \leq b_{max}. \tag{4}$$

2.3. *PCA-based sequential learning and online process monitoring.* Next, we discuss online monitoring of the p-dimensional dynamic process with the observations $\{\mathbf{X}_n, n \geq 1\}$. When the process is IC, it is assumed that it has the regular temporal pattern described by Model (1). Namely, its observations follow the model

$$\mathbf{X}_n = \boldsymbol{\mu}_n + \boldsymbol{\epsilon}_n, \qquad \text{for } n \geq 1, \tag{5}$$

where $\boldsymbol{\mu}_n = \boldsymbol{\mu}_{n^*}$, n^* is an integer in $[-m_0+1, 0]$, $n = n^* + lm_0$, $l \geq 1$ is an integer, and the error term $\boldsymbol{\epsilon}_n$ has the same covariance structure as that in Model (1).

Data standardization and dimension reduction: At the current time point n, we first standardize the observation $\mathbf{X}_n = (X_{n1}, X_{n2}, \ldots, X_{np})'$ by

$$\widetilde{X}_{nl} = \left(X_{nl} - \widehat{\mu}_{nl}^{(n-1)}\right) \Big/ \widehat{\sigma}_{nl}^{(n-1)}, \qquad \text{for } l = 1, 2, \ldots, p,$$

where $\boldsymbol{\mu}_n^{(n-1)} = (\widehat{\mu}_{n1}^{(n-1)}, \widehat{\mu}_{n2}^{(n-1)}, \ldots, \widehat{\mu}_{np}^{(n-1)})'$ and $\{\widehat{\sigma}_{nl}^{(n-1)}\}$ describes the updated regular temporal pattern of the process under monitoring obtained at the previous time point $n-1$, which is defined in Equation (8) below. Then, the PCA procedure can be applied to the standardized observation $\widetilde{\mathbf{X}}_n = (\widetilde{X}_{n1}, \widetilde{X}_{n2}, \ldots, \widetilde{X}_{np})'$ for dimension reduction. Let

$$\mathbf{Y}_n = \left(\widehat{\mathbf{U}}_{d^{(n-1)}}^{(n-1)}\right)' \widetilde{\mathbf{X}}_n,$$

where $\widehat{\mathbf{U}}_{d^{(n-1)}}^{(n-1)}$ is the $(p \times d^{(n-1)})$-dimensional PCA projection matrix updated at the previous time point $n-1$, which is defined in the part "Update of the IC parameter estimates" below. Then, $\mathbf{Y}_n$ is a $d^{(n-1)}$-dimensional vector, with $d^{(n-1)}$ being an integer that is often much smaller than p.

Data decorrelation: Before monitoring the PCA-transformed observations $\{\mathbf{Y}_n, n \geq 1\}$, they should be decorrelated properly across different time points, since the conventional control charts in the SPC literature are designed mainly for monitoring processes with uncorrelated observations (cf., Qiu (2014), Chapter 7). To this end, the observation $\mathbf{Y}_n$ at the current time point n needs to be decorrelated with its previous b_{max} observations, since the serial correlation is assumed to be short-ranged, as discussed in Subsection 2.1. Because the sequential data decorrelation needs to be implemented at each observation time, reduction of computing time is substantially important. For that purpose, the concept of spring length, originally discussed in Chatterjee and Qiu (2009), will be considered in the proposed method. This concept is based on the restarting mechanism of a CUSUM chart. At the current time n, the spring length T_n is defined to be the number of observation times from the last time that the CUSUM charting statistic was reset to zero to the current time n. By combining this concept and the assumed short-range serial correlation, the observation $\mathbf{Y}_n$ at the current time point n only needs to be decorrelated with its previous $b_n = \min\{T_{n-1}, b_{max}\}$ observations. Here, T_{n-1} (instead of T_n) is used because the CUSUM chart has not yet made a decision about the process status at n during the data decorrelation at time n. Since T_{n-1} is often a single-digit integer (cf., Chatterjee and Qiu (2009)), much computing time can be saved by using it. The entire data decorrelation procedure is then briefly described below.

- When $n = 1$ or $b_n = 0$, the decorrelated observation of $\mathbf{Y}_n$ is defined to be $\mathbf{e}_n = \left[\widehat{\gamma}^{(n-1)}(0)\right]^{-1/2} \mathbf{Y}_n$, where $\widehat{\gamma}^{(n-1)}(0)$ is defined in (9) below.
- When $n > 1$ and $b_n > 0$, the estimated covariance matrix of $(\mathbf{Y}'_{n-b_n}, \mathbf{Y}'_{n-b_n+1}, \ldots, \mathbf{Y}'_n)'$ is defined to be

$$\widehat{\boldsymbol{\Sigma}}_{n,n} = \begin{pmatrix} \widehat{\gamma}^{(n-1)}(0) & \cdots & \widehat{\gamma}^{(n-1)}(b_n) \\ \vdots & \ddots & \vdots \\ \widehat{\gamma}^{(n-1)}(b_n) & \cdots & \widehat{\gamma}^{(n-1)}(0) \end{pmatrix} =: \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_{n-1,n-1} & \widehat{\boldsymbol{\Sigma}}_{n-1,n} \\ \widehat{\boldsymbol{\Sigma}}'_{n-1,n} & \widehat{\gamma}^{(n-1)}(0) \end{pmatrix}.$$

Then, the decorrelated observation of $\mathbf{Y}_n$ is defined to be

$$\mathbf{e}_n = \widehat{\mathbf{D}}_n^{-1/2}\left[-\widehat{\mathbf{\Sigma}}'_{n-1,n}\widehat{\mathbf{\Sigma}}^{-1}_{n-1,n-1}\mathbf{B}_{n-1} + \mathbf{Y}_n\right],$$

where $\widehat{\mathbf{D}}_n = \widehat{\gamma}_{n-1}(0) - \widehat{\mathbf{\Sigma}}'_{n-1,n}\widehat{\mathbf{\Sigma}}^{-1}_{n-1,n-1}\widehat{\mathbf{\Sigma}}_{n-1,n}$, $\mathbf{B}_{n-1} = (\mathbf{Y}'_{n-b_n}, \mathbf{Y}'_{n-b_n+1}, \ldots, \mathbf{Y}'_{n-1})'$, and $\widehat{\mathbf{\Sigma}}^{-1}_{n,n}(n)$ can be computed recursively by using the following formula: for $n \geq 2$,

$$\widehat{\mathbf{\Sigma}}^{-1}_{n,n} = \begin{pmatrix} \widehat{\mathbf{\Sigma}}^{-1}_{n-1,n-1} + \widehat{\mathbf{\Sigma}}^{-1}_{n-1,n-1}\widehat{\mathbf{\Sigma}}_{n-1,n}\widehat{\mathbf{D}}^{-1}_{n-1}\widehat{\mathbf{\Sigma}}'_{n-1,n}\widehat{\mathbf{\Sigma}}^{-1}_{n-1,n-1}, & -\widehat{\mathbf{\Sigma}}^{-1}_{n-1,n-1}\widehat{\mathbf{\Sigma}}_{n-1,n}\widehat{\mathbf{D}}^{-1}_{n-1} \\ -\widehat{\mathbf{D}}^{-1}_{n-1}\widehat{\mathbf{\Sigma}}'_{n-1,n}\widehat{\mathbf{\Sigma}}^{-1}_{n-1,n-1}, & \widehat{\mathbf{D}}^{-1}_{n-1} \end{pmatrix}.$$

By using the above sequential data decorrelation procedure, the resulting decorrelated observations $\{\mathbf{e}_n, n \geq 1\}$ would be asymptotically uncorrelated with each other, and each would have an asymptotic mean of $\mathbf{0}$ and an asymptotic identity covariance matrix.

It should be pointed out that the data decorrelation algorithm described above is implemented on the PCA-transformed data $\mathbf{Y}_n$, for $n \geq 1$, whose dimension d_n is usually much smaller than the dimension p of the original data. In addition, recursive computation has been employed in the algorithm. It can be checked that the computational complexity to compute $\mathbf{e}_n$ at time n is $4((b_n+1)d_n)^2 + O((b_n+1)d_n))$, which is relatively low since both b_n and d_n are usually small.

PCA-based sequential process monitoring: After implementing the PCA and data decorrelation procedures, the original process observations $\{\mathbf{X}_n, n \geq 1\}$ have been transformed to the asymptotically uncorrelated ones $\{\mathbf{e}_n, n \geq 1\}$ with lower dimensions. Because each element of $\mathbf{e}_n$ is a linear combination of the original process observations, its distribution would be close to $N(\mathbf{0}, I_{d^{(n-1)}})$, under some regularity conditions. Consequently, the distribution of $\mathbf{e}'_n\mathbf{e}_n$ would be close to $\chi^2_{d^{(n-1)}}$. Thus, we suggest the following CUSUM chart for sequential process monitoring:

$$C_n = \max\left[0, C_{n-1} + \frac{\mathbf{e}'_n\mathbf{e}_n - d^{(n-1)}}{\sqrt{2d^{(n-1)}}} - k\right], \tag{6}$$

where $C_0 = 0$, and $k > 0$ is an allowance constant. The chart gives a signal when

$$C_n > \rho, \tag{7}$$

where $\rho > 0$ is a control limit.

The performance of the CUSUM chart (6)-(7) can be measured by the IC average run length (ARL), denoted as ARL_0, and the out-of-control (OC) ARL, denoted as ARL_1 (Qiu (2014), Chapter 3). By definition, ARL_0 is the average number of observation times from the beginning of online process monitoring to the signal time of the chart when the process is IC, and ARL_1 is the average number of observation times from the occurrence of a shift to the signal time of the chart after the process becomes OC. Usually, ARL_0 is pre-specified, and the chart performs better for detecting a given shift when its ARL_1 is smaller. In the chart (6)-(7), there are two parameters k and ρ to choose in advance. The allowance constant k is usually pre-specified, and the control limit ρ is chosen to achieve the pre-specified value of ARL_0. If the distribution of $\mathbf{e}_n$ is exactly a multivariate normal distribution, then we can determine the control limit ρ easily, using Monte Carlo simulations. But, when the distribution of the original process observations is quite skewed and the dimension p is quite small, the distribution of $\mathbf{e}_n$ could be substantially different from a normal distribution. In such cases, the proposed chart, with its control limit determined based on the normal distribution assumption, would be unreliable to use in the sense that its actual ARL_0 value could be quite different from the pre-specified ARL_0 value (cf., Qiu (2018)). To make the proposed chart more robust, we suggest determining its control limit ρ by using a bootstrap procedure

(cf., Chatterjee and Qiu (2009)), as described below. First, compute the decorrelated PCA-transformed observations $\{\mathbf{e}_j, j = -m_0+1, -m_0+2, \ldots, 0\}$ from the initial IC data $\mathcal{X}_{IC}^{(0)}$, as discussed above in the part "Data decorrelation". Second, a bootstrap sample is drawn randomly with replacement from these decorrelated PCA-transformed observations. Third, the CUSUM chart (6)-(7) with a given control limit ρ is applied to the bootstrap sample to obtain a run length (RL) value. Fourth, the bootstrap resampling procedure is repeated for $B = 1,000$ times, and the average of the B RL values is used to approximate the actual ARL_0 value of the chart. Fifth, ρ can then be searched by a numerical algorithm, such as the bisection search algorithm discussed in Qiu (2014), so that the pre-specified ARL_0 value is reached.

Update of the IC parameter estimates: If the control chart (6)-(7) does not give a signal at time n, then the observation $\mathbf{X}_n$ can be combined with the IC data $\mathcal{X}_{IC}^{(n-1)} = \{\mathbf{X}_j, -m_0 + 1 \le j \le n-1\}$ at the previous time $n-1$, and the combined IC data is denoted as $\mathcal{X}_{IC}^{(n)}$. The estimates of the IC parameters obtained at time $n-1$ can then be updated properly using the combined IC data. To this end, the estimate of the IC mean $\boldsymbol{\mu}_j$ can be updated by the following formula: for $n \ge 1$,

$$\widehat{\boldsymbol{\mu}}_j^{(n)} = \left(\left[\mathbf{S}_j^{(n)} \right]^{-1} \mathbf{R}_j^{(n)} \right)' \left(I_{p\times p} \otimes \boldsymbol{\xi}_1 \right), \qquad \text{for } j = -m_0+1, -m_0+2, \ldots, 0, \tag{8}$$

where $\mathbf{S}_j^{(n)}$ and $\mathbf{R}_j^{(n)}$ can be updated recursively by

$$\mathbf{S}_j^{(n)} = \mathbf{S}_j^{(n-1)} + (\mathbf{k}_n \otimes \mathbf{Z}_n)\left(I_{p\times p} \otimes \mathbf{Z}_n\right)',$$
$$\mathbf{R}_j^{(n)} = \mathbf{R}_j^{(n-1)} + (\mathbf{k}_n \mathbf{X}_n) \otimes \mathbf{Z}_n,$$

$\mathbf{Z}_n = (1, n^* - j)'$, $\mathbf{k}_n = \text{diag}\left[K(\frac{n^*-j}{h_1}), K(\frac{n^*-j}{h_2}), \ldots, K(\frac{n^*-j}{h_p})\right]$, and n^* is an integer in $[-m_0+1, 0]$ that is related to n through $n = n^* + lm_0$, for $l \ge 1$, as discussed at the beginning of Subsection 2.3.

The estimate of the standard deviation of X_{jl} can be updated by: for $n \ge 1$,

$$\widehat{\sigma}_{jl}^{(n)} = \sqrt{\alpha_{jl}^{(n)} \Big/ \beta_{jl}^{(n)}}, \text{ for } l = 1, 2, \ldots, p, \; j = -m_0+1, -m_0+2, \ldots, 0,$$

where $\alpha_{jl}^{(n)}$ and $\beta_{jl}^{(n)}$ can be updated recursively by

$$\alpha_{jl}^{(n)} = \alpha_{jl}^{(n-1)} + \left(X_{jl} - \widehat{\mu}_{jl}^{(n)}\right)^2 K\left(\frac{n^*-j}{g_l}\right),$$
$$\beta_{jl}^{(n)} = \beta_{jl}^{(n-1)} + K\left(\frac{n^*-j}{g_l}\right).$$

The correlation estimates can be updated as follows: for $n \ge 1$,

$$\widehat{\gamma}^{(n)}(s) = \frac{1}{m_0+n-s} \mathbf{Y}_n \mathbf{Y}'_{n-s} + \frac{m_0+n-s-1}{m_0+n-s} \widehat{\gamma}^{(n-1)}(s), \text{ for } 0 \le s \le b_{max}. \tag{9}$$

The eigenvalues and eigenvectors of the sample covariance matrix can also be updated by using the incremental PCA algorithm (Weng et al. (2003)) with the following formulas:

$$\widehat{\lambda}_l^{(n)} = \|\mathbf{v}_l^{(n)}\|, \;\; \widehat{\mathbf{u}}_l^{(n)} = \frac{\mathbf{v}_l^{(n)}}{\|\mathbf{v}_l^{(n)}\|}, \text{ for } l = 1, 2, \ldots, p,$$

where

$$\mathbf{v}_l^{(n)} = \frac{m_0+n-1}{m_0+n}\mathbf{v}_l^{(n-1)} + \frac{1}{m_0+n}\mathbf{P}_l^{(n)}[\mathbf{P}_l^{(n)}]'\widehat{\mathbf{u}}_l^{(n-1)},$$

$\mathbf{P}_l^{(n)} = \mathbf{P}_{l-1}^{(n)} - \widehat{\mathbf{u}}_{l-1}^{(n)}[\mathbf{P}_{l-1}^{(n)}]'\widehat{\mathbf{u}}_{l-1}^{(n)}$, and $\mathbf{P}_1^{(n)} = \widetilde{\mathbf{X}}_n$.

Finally, the proposed CUSUM chart (6)-(7) depends on the number of selected PCs, $d^{(n-1)}$, which should also be updated if the chart does not give a signal at time n. This quantity affects many components of the proposed method, from data decorrelation, computation of the control limit ρ of the chart, to decision making about the process status. It will add much computing burden if its value changes frequently over time. For this reason, we suggest updating the value of $d^{(n-1)}$ only at times when the CUSUM charting statistic gets restarted (i.e., $C_n = 0$ in (6)). The reason for this consideration is that the evidence in the observed data for a process shift is considered by the chart to be weak and thus updates of the IC parameter estimates are the most needed at such times (cf., Qiu (2014), Chapter 4). Based on our numerical experience, the value of $d^{(n-1)}$ changes slightly (e.g., from 4 to 5) when n is small, and stabilizes when n gets large, which is justified theoretically by Theorem 1 below.

THEOREM 1. *In the IC Model (1), let $\boldsymbol{\mu}_j = \boldsymbol{\mu}((j+m_0)/m_0)$, for each j. Then, $\boldsymbol{\mu}(t)$ is the IC mean function, for $t \in [0,1]$. The variance function $\sigma^2(t)$ can be defined similarly in $[0,1]$. Assume that both functions are twice continuously differentiable in $[0,1]$, and there are constants $\delta > 5$ and $0 < C_\epsilon < \infty$ such that $E(|\epsilon_{jl}|^\delta) < C_\epsilon$, for all j and $l = 1,2,\ldots,p$. The time series of each component of $\{\boldsymbol{\epsilon}_j\}$ is assumed to be weakly stationary with absolutely summable auto-covariances (i.e., $\sum_{s=-\infty}^{\infty} |Cov(\epsilon_{jl}, \epsilon_{j-s,l})| < \infty$). The kernel function $K(\cdot)$ is assumed to be a Lipschitz-1 continuous density function with the support $[-1,1]$, and the bandwidths $\{h_l\}$ and $\{g_l\}$ are assumed to satisfy the conditions that $h_l = o(1)$, $log^2(m_0)/(m_0h_l^2) = o(1)$, $g_l = o(1)$, and $log^2(m_0)/(m_0g_l^2) = o(1)$, for $l = 1,2,\ldots,p$. Then, we have*

$$\left|\sum_{l=1}^{k}\widehat{\lambda}_l^{(0)} - \sum_{l=1}^{k}\lambda_l\right| = O_{\mathbb{P}}\left(h_{\max}^4 + \log^2(m_0)/(m_0h_{\min}^2) + g_{\max}^4 + \log^2(m_0)/(m_0g_{\min}^2) + m_0^{-1/2}\right),$$

where $k = 1,2,\ldots,p$, $\{\lambda_l, l = 1,2,\ldots,p\}$ are the eigenvalues of the covariance matrix of the standardized IC process observations, $h_{\max} = \max\{h_1,h_2,\ldots,h_p\}$, $h_{\min} = \min\{h_1,h_2,\ldots,h_p\}$, $g_{\max} = \max\{g_1,g_2,\ldots,g_p\}$, and $g_{\min} = \min\{g_1,g_2,\ldots,g_p\}$.

By Theorem 1, we have $\lim_{n\to\infty} d^{(n-1)} = d_v$, where

$$d_v = \min\left\{q : \left(\sum_{l=1}^{q}\lambda_l \Big/ \sum_{l=1}^{p}\lambda_l\right) \times 100\% > v, 1 \le q \le p\right\}.$$

So, the value of $d^{(n-1)}$ will indeed stabilize when n gets large. The proof of Theorem 1 is given in the supplementary file.

As in other self-starting control charts (Hawkins (1987)), the updating mechanism described above could contaminate the IC parameter estimates if a shift cannot be detected in a timely manner, since the OC process observations collected after the undetected shift could be used in computing the IC parameter estimates. See related discussions in Section 4.5 of Qiu (2014). Based on our numerical experience, if the initial IC sample size is relatively large, then the impact of such potential contamination on the chart performance is negligible.

However, if the initial IC sample size is small (e.g., < 100), then a process shift (or drift) can be missed permanently by the chart if the shift cannot be detected soon after its occurrence. To partially overcome this limitation, one solution is to update the IC parameter estimates only in cases when the chart gets re-started at the current observation time. As explained earlier, the chance for the process to be OC is small in such cases, and thus contamination of the IC parameter estimates by OC process observations would also be unlikely.

Practical guidelines: To use the proposed chart (6)-(7), the following practical guidelines are provided. i) The PCA procedure is considered in the proposed method for reducing the computing burden of process monitoring. When p is relatively small (e.g., $p \leq 10$), the benefit to use the PCA procedure would be limited. In such cases, we suggest using the proposed chart without considering the PCA data transformation. ii) the performance of the proposed chart would depend on the initial IC data size m_0. Based on extensive numerical studies (cf., Tables 1 and 2 in Section 3), we suggest choosing $m_0 \geq 10p$ to ensure a reliable IC performance of the proposed chart. iii) The proposed chart assumes that serial correlation can be ignored when two observation times are at least b_{max} apart. In practice, however, b_{max} is usually unknown. Based on extensive numerical studies, the performance of the proposed chart can hardly be improved when b_{max} is chosen larger than 20, and its performance could be negatively affected in certain cases if b_{max} is chosen smaller than 10. Thus, we suggest choosing $b_{max} \in [10, 20]$.

3. Simulation Studies. In this section, we evaluate the numerical performance of the proposed method by Monte Carlo simulations. The proposed chart (6)-(7) is denoted as PCA-D-C, where the letter "D" means that it considers dynamic processes, and the following letter "C" implies that serial correlation in the observed data is allowed. In the simulation studies, the following four different cases are considered:

Case I: IC process observations $\{\mathbf{X}_n, n \geq 1\}$ are independent and identically distributed (i.i.d.) with the IC distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (0, 0, \ldots, 0)'$ and $\boldsymbol{\Sigma} = (\sigma_{l_1 l_2})_{p\times p}$ with $\sigma_{l_1 l_2} = 0.5^{|l_1 - l_2|}$, for $l_1, l_2 = 1, 2, \ldots, p$.

Case II: IC process observations $\{\mathbf{X}_n, n \geq 1\}$ are generated from Model (5). Their means and correlation structure are specified in Model (1), where the first five components of $\boldsymbol{\mu}_j$ are

$$[tanh(j/m_0), \exp(j/m_0), j/m_0, \cos(2\pi j/m_0), 0]\,, \text{for } j = -m_0 + 1, -m_0 + 2, \ldots, 0,$$

the remaining components are replicated from the first five components (e.g., the sixth component is $tanh(j/m_0)$, the seventh component is $\exp(j/m_0)$, and so forth), the error terms $\{\boldsymbol{\epsilon}_j\}$ are i.i.d. with the distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma} = (\sigma_{l_1 l_2})_{p\times p}$ with $\sigma_{l_1 l_2} = 0.5^{|l_1 - l_2|}$, for $l_1, l_2 = 1, 2, \ldots, p$.

Case III: Same as Case II, except that the error terms $\{\boldsymbol{\epsilon}_j\}$ are assumed to follow the vector autoregressive (VAR) model $\boldsymbol{\epsilon}_j = 0.2\boldsymbol{\epsilon}_{j-1} + \boldsymbol{\eta}_j$, where $\boldsymbol{\epsilon}_0 = \mathbf{0}$, $\{\boldsymbol{\eta}_j\}$ are i.i.d. with the distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma} = (\sigma_{l_1 l_2})_{p\times p}$ with $\sigma_{l_1 l_2} = 0.5^{|l_1 - l_2|}$, for $l_1, l_2 = 1, 2, \ldots, p$.

Case IV: Same as Case III, except that each component of $\boldsymbol{\eta}_j$ has the standardized χ^2_3 distribution and the covariance matrix of $\boldsymbol{\eta}_j$ remains the same as that in Case III.

For the four cases described above, Case I is the conventional case with i.i.d. IC process observations generated from a zero-mean normal distribution. Cases II-IV consider three different dynamic processes. The one in Case II has independent and normally distributed IC observations at different observation times. Case III is the same as Case II, except that process observations are serially correlated in Case III. Case IV considers a scenario when process observations are serially correlated and the IC process distribution is skewed. In all cases, the p quality variables are mutually correlated.

For comparison purposes, besides the proposed chart PCA-D-C, the following four alternative methods are also considered:

- A simplified version of PCA-D-C, denoted as PCA-ND-C: This chart is the same as PCA-D-C, except that the IC process distribution is assumed to be time-independent. Namely, the process under monitoring is assumed to be non-dynamic.
- A simplified version of PCA-D-C, denoted as PCA-D-NC: This chart is the same as PCA-D-C, except that process observations at different time points are assumed to be uncorrelated and thus the data decorrelation procedure described in Subsection 2.3 is not implemented.
- The moving-window PCA method discussed in Wang et al. (2005), denoted as MWPCA: The MWPCA method performs PCA at each observation time, using process observations within a moving window of size w of the current observation time. Its charting statistic is

$$M_n^2 = (\mathbf{X}_n - \widehat{\boldsymbol{\mu}}_w)' \widehat{\mathbf{U}}_{d,w} \widehat{\boldsymbol{\Lambda}}_{d,w}^{-1} \widehat{\mathbf{U}}_{d,w}' (\mathbf{X}_n - \widehat{\boldsymbol{\mu}}_w),$$

  where $\widehat{\boldsymbol{\mu}}_w$ is the sample mean of $\mathcal{X}_{n,w} = \{\mathbf{X_{n-w+1}}, \mathbf{X_{n-w+2}}, \ldots, \mathbf{X_n}\}$, $\widehat{\boldsymbol{\Lambda}}_{d,w}$ is a diagonal matrix whose diagonal elements are the first d eigenvalues of the sample covariance matrix of $\mathcal{X}_{n,w}$, $\widehat{\mathbf{U}}_{d,w}$ is a $p \times d$ matrix whose columns are the corresponding eigenvectors, and d is the number of PCs determined similarly to that in (3). The MWPCA method assumes that IC process observations are independent and normally distributed. Its control limit is chosen to be $F_{d,w-d}(\alpha)[d(w^2-1)]/[w(w-d)]$, where $F_{d,w-d}(\alpha)$ is the $(1-\alpha)$th quantile of the F distribution with d and $w-d$ degrees of freedom, and $\alpha = 1/ARL_0$.
- The multivariate control chart discussed in Zou et al. (2015) for monitoring high-dimensional data, denoted as ZCUSUM: This chart was developed based on a goodness-of-fit test of the CUSUM statistics of the p individual quality variables. Its charting statistic is defined to be

$$L_n = \sum_{l=1}^{p} \left\{ \log \left[\frac{U_{n(l)}^{-1} - 1}{(p-1/2)(l-3/4) - 1} \right] \right\}^2 I_{\{U_{nl} > (l-3/4)/p\}}$$

  where $U_{n(1)} \le U_{n(1)} \le \cdots \le U_{n(p)}$ are the order statistics of $\{U_{n1}, U_{n2}, \ldots, U_{np}\}$, U_{nl} is the cdf of S_{nl}, for $l = 1, 2, \ldots, p$, and S_{nl} is the CUSUM statistic of the lth quality variable with an allowance constant k_l. The control limit of the ZCUSUM chart can be determined through simulations, since it assumes that the IC process distribution is normal.

In the simulation studies, the nominal ARL_0 values of all charts are fixed at 200. In MWPCA, the moving window size w is chosen to be $10p$, as suggested by Wang et al. (2005). By the suggestion in Zou et al. (2015), the allowance constants $\{k_l, l = 1, 2, \ldots, p\}$ of ZCUSUM are all chosen to be 0.5. In the three charts PCA-D-C, PCA-ND-C and PCA-D-NC, b_{max} is chosen to be 10, and the allowance constants are all chosen to be 0.5. Their control limits are chosen by the bootstrap procedure described in the paragraph below Expression (7). For charts MWPCA, PCA-D-C, PCA-ND-C and PCA-D-NC, v is chosen to be 90% for determining the number of PCs (cf., (3)).

3.1. *IC performance.* We first study the IC performance of the related control charts. To compute the actual ARL_0 value of a chart, an IC dataset of size m_0 is first generated from an IC model. Then, a control chart is applied to a sequence of 2,000 IC process observations for online process monitoring, and its RL value is recorded. The online process monitoring is then repeated 1,000 times, and the average of the 1,000 RL values is used as the estimate of the actual conditional ARL_0 value, conditional on the IC data. Finally, all steps described above, starting from the generation of the IC data, to computation of the estimate of the actual conditional ARL_0 value, are repeated 100 times. The actual ARL_0 value of the chart is then estimated by the average of the 100 estimates of the conditional ARL_0 value. We first

assume that $p = 100$ (i.e., there are 100 quality variables to monitor), and the IC sample size m_0 is fixed at 1,200. The results of the estimated actual ARL_0 values of the five charts in various cases are presented in Table 1, along with their standard errors. From the table, it can be seen that (i) MWPCA is reliable to use in Case I when process observations are i.i.d. and normally distributed, but unreliable in all other cases because some of its model assumptions are invalid. (ii) ZCUSUM is relatively reliable in Case I, but unreliable in all other cases, because it requires a large IC dataset to estimate its IC parameters in Case I, and some of its model assumptions are invalid in all other cases. (iii) PCA-ND-C and PCA-D-NC are reliable only when their model assumptions are valid. For instance, PCA-D-NC is reliable in Cases I and II when its "no serial correlation" assumption is valid, and unreliable in Cases III and IV when this assumption is invalid. (iv) In comparison, the proposed chart PCA-D-C has a reasonably good performance in all cases considered, since its estimated actual ARL_0 values are always within 10% of the nominal ARL_0 level of 200.

TABLE 1
Actual ARL_0 values and their standard errors (in parentheses) of the five control charts when $p = 100$, $m_0 = 1,200$, and the nominal ARL_0 values of all charts are fixed at 200.

Cases	PCA-D-C	PCA-ND-C	PCA-D-NC	MWPCA	ZCUSUM
I	190 (3.45)	193 (3.29)	190 (3.40)	193 (2.94)	163 (1.99)
II	189 (3.64)	67 (1.01)	194 (3.11)	143 (2.13)	56 (0.92)
III	187 (3.46)	75 (1.35)	136 (2.27)	125 (1.97)	48 (0.95)
IV	181 (3.29)	70 (1.44)	147 (2.45)	96 (1.87)	46 (0.93)

From the description of its construction in Section 2, the IC performance of the proposed chart PCA-D-C may depend on the IC sample size m_0 and the dimensionality p. To see how m_0 and p affect the IC performance of PCA-D-C, next we let p change between 50 and 100, m_0 change among 400, 800, 1,200, 1,600, and 2,000, and the remaining setups keep unchanged from those in the example of Table 1. The calculated actual ARL_0 values of PCA-D-C are presented in Table 2. From this table, we can have the following conclusions. First, when p gets larger, the necessary IC data should also be larger. In Case I, for instance, the estimated actual ARL_0 value is already within 10% of the nominal ARL_0 level when $m_0 = 800$ and $p = 50$, while the necessary IC data size m_0 needs to be 1,200 to have a similar estimated actual ARL_0 value when $p = 100$. Second, it seems that PCA-D-C has a better IC performance when m_0 is larger. Based on our numerical experience, its IC performance is quite reliable when $m_0 \geq 10p$. It should be pointed out that compared to the proposed method, the competing method ZCUSUM requires a larger IC dataset to achieve a reliable IC performance in cases when its model assumptions are valid. For instance, Table 1 shows that its actual ARL in Case I is about 20% smaller than the nominal ARL_0 value when $m_0 = 1,200$ and $p = 100$. As a matter of fact, Zou et al. (2015) showed that when $p = 100$ and IC observations were independent and normally distributed, ZCUSUM required more than 4,000 IC observations to achieve a reliable IC performance, since a large IC dataset was required to obtain reliable estimates of its IC parameters.

3.2. *OC performance.* Next, we study the OC performance of the related control charts in cases where $p = 100$, $m_0 = 1,200$, and the nominal ARL_0 values are all set at 200. To this end, it is assumed that a mean shift occurs at the beginning of process monitoring and the shifted mean becomes $\boldsymbol{\mu}_n + \delta\mathbf{a}$, where $\mathbf{a} = (1, 1, \ldots, 1, 0, \ldots, 0)'$ has ηp elements being 1 and the remaining elements being 0, and δ denotes the shift size. Two scenarios are considered in this example. In Scenario I, we fix the number of variables that have shifts to be 20 (i.e., $\eta = 0.2$), and let δ change among 0.2, 0.4, 0.6, 0.8, and 1. In Scenario II, the shift size

TABLE 2
Actual ARL_0 values and their standard errors (in parentheses) of the control chart PCA-D-C when p changes between 50 and 100, m_0 changes among 400, 800, 1,200, 1,600 and 2,000, and the nominal ARL_0 value is fixed at 200.

p	Cases	m_0= 400	800	1200	1600	2000
50	I	169 (2.99)	192 (3.14)	194 (3.35)	196 (3.41)	201 (3.43)
	II	165 (3.11)	190 (3.56)	196 (3.48)	205 (3.62)	204 (3.31)
	III	161 (3.09)	184 (3.51)	193 (3.38)	196 (3.64)	202 (3.44)
	IV	158 (2.94)	185 (3.20)	187 (3.26)	192 (3.01)	197 (3.23)
100	I	148 (3.09)	178 (3.86)	190 (3.45)	193 (3.21)	197 (3.53)
	II	149 (3.17)	176 (3.97)	189 (3.64)	195 (3.57)	201 (3.27)
	III	136 (3.86)	166 (3.40)	187 (3.46)	205 (3.97)	201 (3.30)
	IV	130 (3.07)	156 (2.97)	181 (3.29)	186 (2.66)	194 (3.12)

δ is fixed at 0.6, and η changes among 0.1, 0.2, 0.3, 0.4, and 0.5. To make the comparison among different charts as fair as possible, their control limits have been adjusted properly so that their actual ARL_0 values all equal the nominal ARL_0 value of 200. Also, for detecting a given shift, their procedure parameters have been searched so that their ARL_1 values reach the minimum. Namely, their optimal OC performance is compared here, which is common in the SPC literature for a fair comparison of the OC performance of different control charts (e.g., Qiu and Xie (2022)). The computed optimal ARL_1 values of the five charts are presented in Figure 3. From the figure, we can have the following conclusions: i) all charts perform reasonably well in Case I, ii) the charts PCA-D-C and PCA-D-NC perform better than the other three charts in Case II when the process is dynamic and process observations at different time points are independent, and iii) the proposed chart PCA-D-C performs better than the other four charts in Cases III and IV when the process is dynamic and there is serial correlation in process observations.

In practice, the true shift is usually unknown and thus consideration of the optimal OC performance of a chart becomes impractical (Knoth et al. (2021)). Next, we consider an example in which the procedure parameters of all charts are chosen to be the same as those used in the example of Table 1 and all other setups are the same as those in the previous example. The ARL_1 values of the five charts in Cases I-IV of Scenario I considered in Figure 3 are shown in Figure S.1 in the supplementary file. From the figure, it can be seen that similar conclusions to those in the previous example can be made here about the OC performance of the five charts.

In the previous two examples, mean shifts are assumed to occur at the beginning of process monitoring and the related results are called zero-state results in the literature (cf., Section 4.2, Qiu (2014)). In practice, however, a real shift usually occurs after the beginning of process monitoring, making the zero-state OC performance impractical (Knoth et al. (2023)). Next, we examine the steady-state OC performance of the related control charts when a shift occurs at the time $\tau = 100$ and other setups are the same as those in the example of Figure S.1. The computed ARL_1 values of the charts are presented in Figure S.2. From the figure, it can be seen that similar conclusions to those in the example of Figure S.1 can be made here regarding their steady-state OC performance.

As discussed in Section 2, the updating mechanism of the proposed chart has the limitation that a process shift (or drift) could be missed permanently if it cannot be detected early after it occurs. To over this limitation, we have proposed to update the IC parameter estimates only when the CUSUM chart restarts. This modified chart is denoted as Mod-PCA-D-C. To study its performance, we consider a new example in which a mean drift occurs at the beginning of process monitoring in Case III, the OC mean function is $\boldsymbol{\mu}_n + n\psi_1(1, 1, \ldots, 1)'$ (i.e., linear mean drift) or $\boldsymbol{\mu}_n + n^2\psi_2(1, 1, \ldots, 1)'$ (i.e., quadratic mean drift), where ψ_1 and

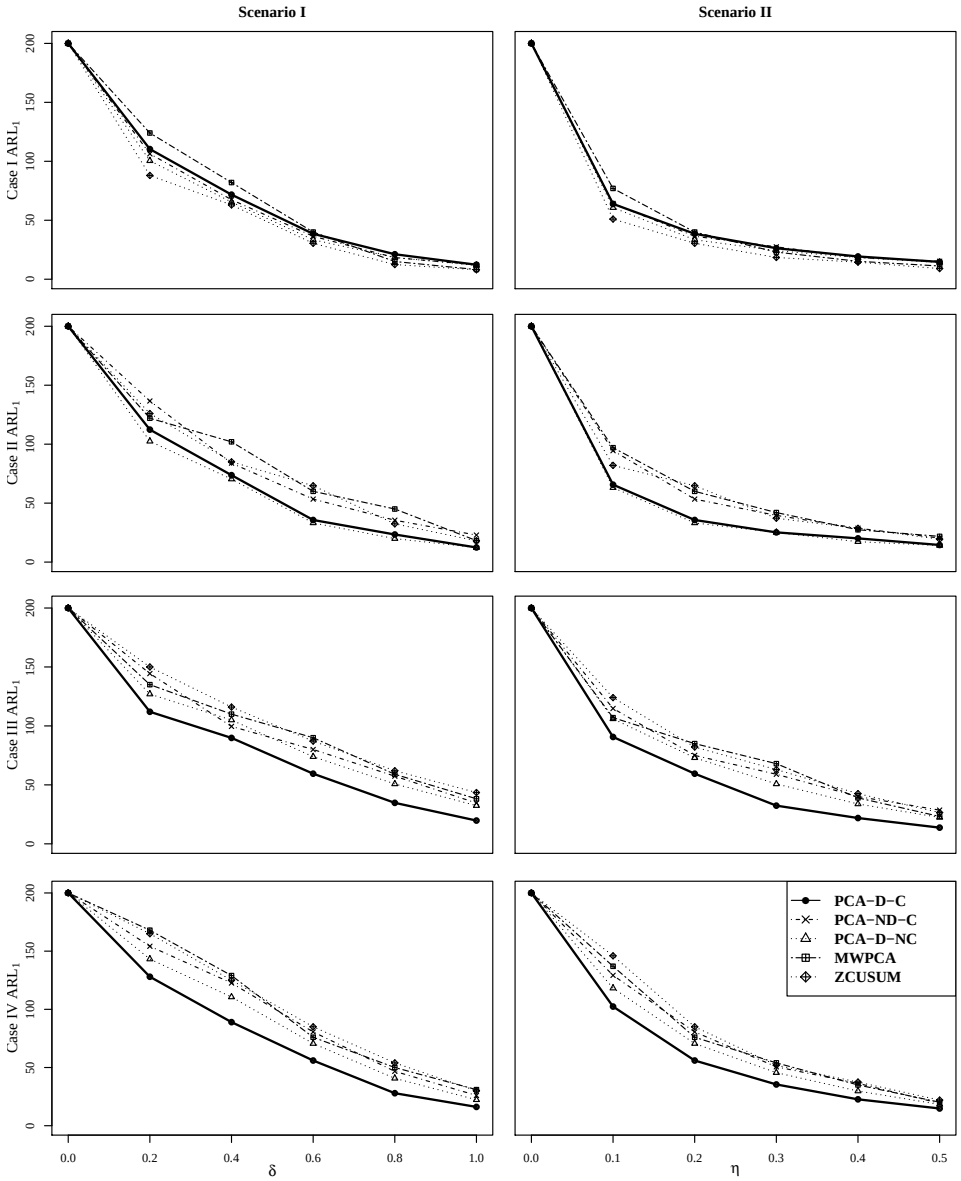

FIG 3. *Optimal ARL_1 values of the five charts when their nominal ARL_0 values are all fixed at 200, $p = 100$, and $m_0 = 1,200$. In Scenario I, $\eta = 0.2$, and the shift size parameter δ changes among 0.2, 0.4, 0.6, 0.8 and 1.0. In scenario II, $\delta = 0.6$ and η changes among 0.1, 0.2, 0.3, 0.4, and 0.5.*

ψ_2 are the drift sizes. In cases when the nominal ARL_0 values of the charts Mod-PCA-D-C and PCA-D-C are fixed at 200, ψ_1 changes among $\{0, 0.001, 0.005, 0.01, 0.05, 0.1\}$, and ψ_2 changes among $\{0, 0.0001, 0.0005, 0.001, 0.005, 0.01\}$, their ARL_1 values are presented in Figure S.3. From the figure, it can be seen that (i) the modification can indeed improve the performance of PCA-D-C in all cases considered, and (ii) the improvement is generally small, especially when the drift size is relatively large.

4. Online Monitoring of Ozone Pollution In the Houston Area. Houston is one of the largest cities in the U.S. that has been suffering severe ozone pollution (cf., Figure 1). Under the Clean Air Act, it has been classified as a severe ozone non-attainment area. This is mainly due to its large number of chemical manufacturing facilities and metal recycling facil-

ities that generate abundant precursor pollutants (Sun et al. (2015)). In addition, the Houston-Galveston-Brazoria (HGB) area has relatively high average temperatures and abundant sunshine that provides ideal meteorological conditions to accelerate chemical reactions for ozone formation (Gorai et al. (2015)). In recent years, ozone pollution has attracted a lot of attention from the Texas governments, and they have been making various efforts to improve air quality and prevent ozone pollution in the HGB area. For instance, the Clean-Air Plan has been implemented in the HGB area to improve air quality by taking several preventive measures, including requiring manufacturing facilities to install pollution-control equipment, limiting industrial emissions, and improving the accessibility of public transportation (Sexton and Linder (2015)). The government agency TCEQ also developed an air pollution surveillance system to collect environmental data in the HGB area. In this section, we apply the proposed method to two specific ozone datasets collected in the Houston area.

4.1. *Application to a standard ozone dataset in the Houston area.* The first dataset is saved in the UC Irvine Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets/ozone+level+detection`) that contains observations of 72 meteorological variables in the HGB area. These variables have been verified empirically to be associated with ozone concentration, and used by environmental researchers for predicting ozone concentration levels (e.g., Draxler (2000), Sun et al. (2015)). In addition, the dataset contains a binary variable indicating whether a given day was declared to be an ozone day, based on certain meteorological criteria. Thus, it has been used as a standard dataset to justify analytic models for predicting ozone condition from meteorological variables. See Zhang and Fan (2008) for a detailed description about this dataset.

The dataset introduced above contains observations of the 72 meteorological variables collected from January 1, 1999 to December 31, 2000. Among the 72 variables, one of them is not reliable to use because most of its values are either zero or missing. So, in our data analysis, it is excluded and the remaining 71 variables, including air temperature, wind speed resultant, relative humidity, sea level pressure, and many more, are used. Some of these 71 variables have a few missing values, and they are imputed as follows: For each variable, the LLK procedure discussed in Subsection 2.1 is considered, and its missing values are replaced by their LLK estimates defined in Equation (2). To have an intuitive impression about the dataset, observations of all 71 variables are displayed in Figure 4 by an image constructed as follows. Because different variables could have different scales, each variable has been re-scaled by subtracting its minimum from each of its observations, and then dividing the difference by its range. Thus, the rescaled variable has a range of $[0, 1]$. Then, the 71 rescaled variables, denoted as V1, V2, ..., V71, are shown in the 71 rows of the image, with a darker color denoting a larger value. Besides the image in Figure 4, Figure 5 shows the observations of the following six representative variables: daily average wind speed, daily average temperature, relative humidity, sea level pressure, K index, and total totals index. Among these variables, the variables K index and total totals index measure the thunderstorm potential and the severe weather occurrence potential, respectively.

From Figures 4 and 5, it can be seen that there is a quite obvious yearly seasonality in the observed data. Also, from the binary variable about the daily ozone status, the HGB area seemed to have a good environmental status in the first six months of each year, since there were only 7 and 8 ozone days in that period of time in 1999 and 2000, respectively. However, starting from early July, the environmental status in year 2000 was much worse than that in year 1999. During the two months of July and August, for instance, there were 10 ozone days in 1999, and 23 ozone days in 2000. For this reason, the data of non-ozone days in the year 1999 are used as the IC data in this section for estimating the regular longitudinal pattern of the related variables, and the data in the year 2000 are used for online process monitoring. It

is our hope that the environmental deterioration in the second half of the year 2000 can be detected promptly by our proposed method, so that proper interventions can be applied in a timely manner to protect the environment in the HGB area.

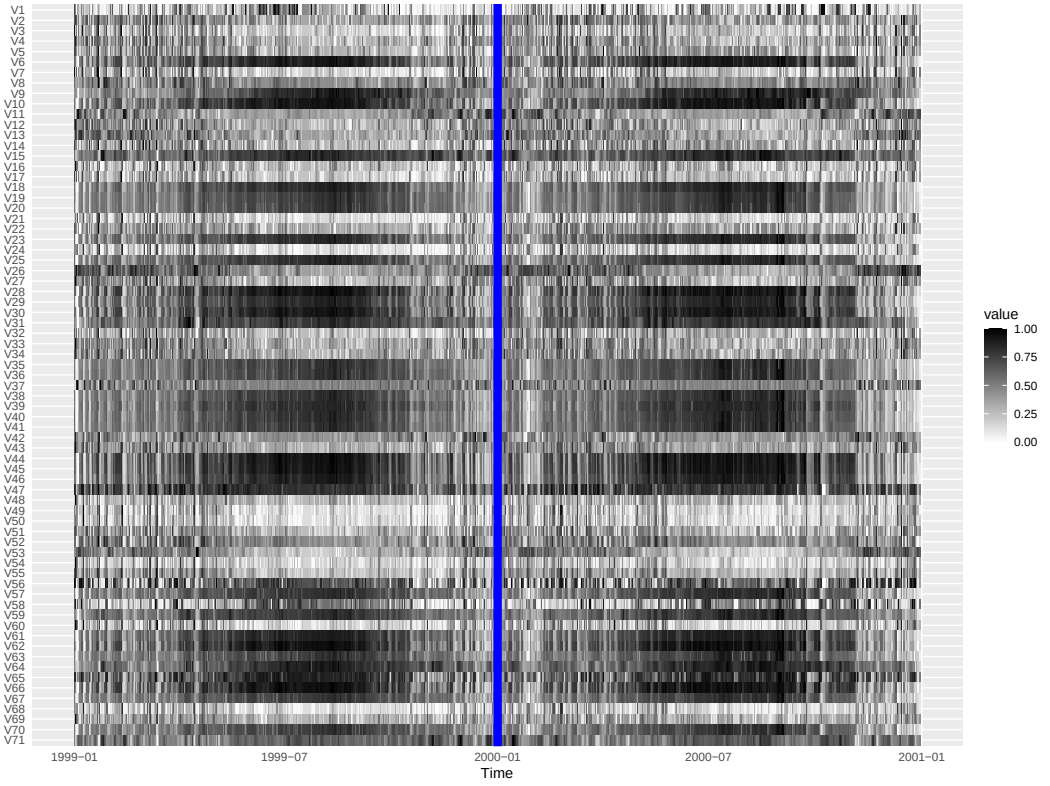

FIG 4. *Observations of the 71 rescaled meteorological variables in the environmental data of the HGB area in 1999 and 2000. A darker color denotes a larger value. The vertical solid line separates the initial IC data from the data for online process monitoring.*

To apply the proposed method, we first compute the initial LLK estimates $\widehat{\mu}_{jl}^{(0)}$ and $\widehat{\sigma}_{jl}^{(0)}$ from the initial IC data, for each j and l, as discussed in Subsections 2.1 and 2.2. Then, the standardized IC observations $\widetilde{X}_{jl} = (X_{jl} - \widehat{\mu}_{jl}^{(0)})/\widehat{\sigma}_{jl}^{(0)}$ are computed for all j and l. For the second-year data, the standardized observations $\widetilde{X}_{nl} = (X_{nl} - \widehat{\mu}_{jl}^{(n-1)})/\widehat{\sigma}_{nl}^{(n-1)}$ are computed as discussed in Subsections 2.3, for all n and l. The standardized observations of the six representative variables considered in Figure 5 are shown in Figure 6, where the vertical dotted line in each plot separates the initial IC data from the data for online process monitoring, and the small triangles denote the ozone days in year 2000. It can be seen that the standardized data of these representative variables are quite stable throughout 1999 and the first several months of the year 2000, before becoming unstable in the remaining months of 2000.

To check the normality of the IC data, the Shapiro-Wilk test is applied to the standardized IC data, and the test gives a p-value of 2.2×10^{-26}, implying that the data distribution is significantly non-normal. To check the serial correlation, the Durbin-Watson test is applied to the standardized data of individual variables, and the p-values of the test for all variables are < 0.003, implying a significant autocorrelation in the IC data. We also verified the validity of the assumption of short-ranged serial correlation for the standardized IC data using the R function **acf()** from the package **stats**. The analysis shows that for all variables, the

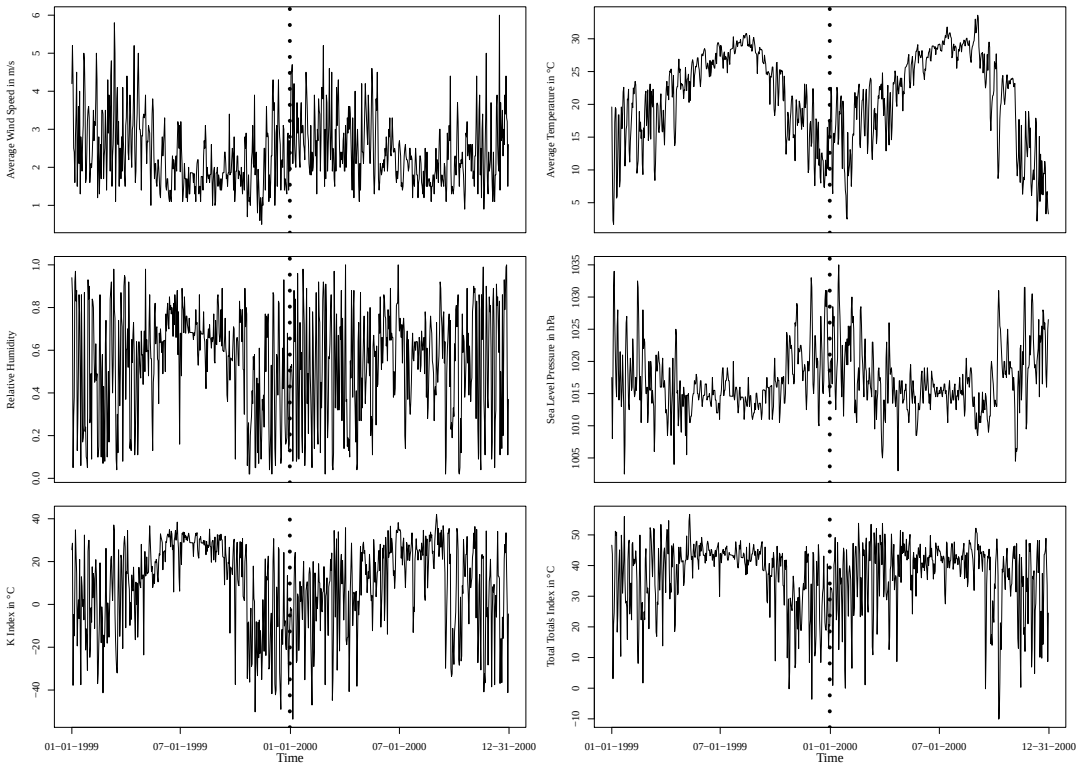

FIG 5. *Original observations of the six representative meteorological variables: daily average wind speed, daily average temperature, relative humidity, sea level pressure, K index, and total totals index in the environmental data of the HGB area in the years 1999 and 2000. The vertical dotted line in each plot separates the initial IC data from the data for online process monitoring.*

autocorrelation coefficients are not statistically significant beyond a lag of 5 observations, indicating that the short-ranged serial correlation assumption holds in this application. The PCA procedure discussed in Subsection 2.2 is then applied to the standardized IC data with $v = 90\%$ (cf., (3)), and the first 15 PCs are selected. To check the stationarity of the serial correlation in the observations of the selected PCs, the augmented DickeyFuller (ADF) test is used for individual PCs and the p-values of the test for all PCs are < 0.02, implying that the stationarity assumption cannot be rejected in this example.

Next, we apply the five charts considered in Section 3 to this dataset. In all control charts, their nominal ARL_0 values are fixed at 200, and their procedure parameters (e.g., the allowance constant k) are chosen to be the same as those in the example of Table 1. Their control limits are computed in the same way as that discussed in Section 3. The five charts are shown in Figure 7. From the plots in the figure, it can be seen that MWPCA gives signals at many observation times, starting in early January of 2000. Since its model assumptions are violated in this example, this result may not be reliable. The remaining four charts PCA-D-C, PCA-ND-C, PCA-D-NC, and ZCUSUM give their first signals on 7/22/2000, 9/6/2000, 7/26/2000, and 8/4/2000, respectively. Therefore, among these four charts, the proposed chart PCA-D-C gives the earliest first signal in this example. If we check the standardized observations of the six representative meteorological variables shown in Figure 6, then it can be seen that these variables seem to have some systematic changes around the first signal time of PCA-D-C, and the frequency of ozone days starts to intensify around that specific time as well. Therefore, the proposed chart PCA-D-C seems quite effective in detecting shifts in the related meteorological variables in this example. As a sidenote, Figure 6 shows that some

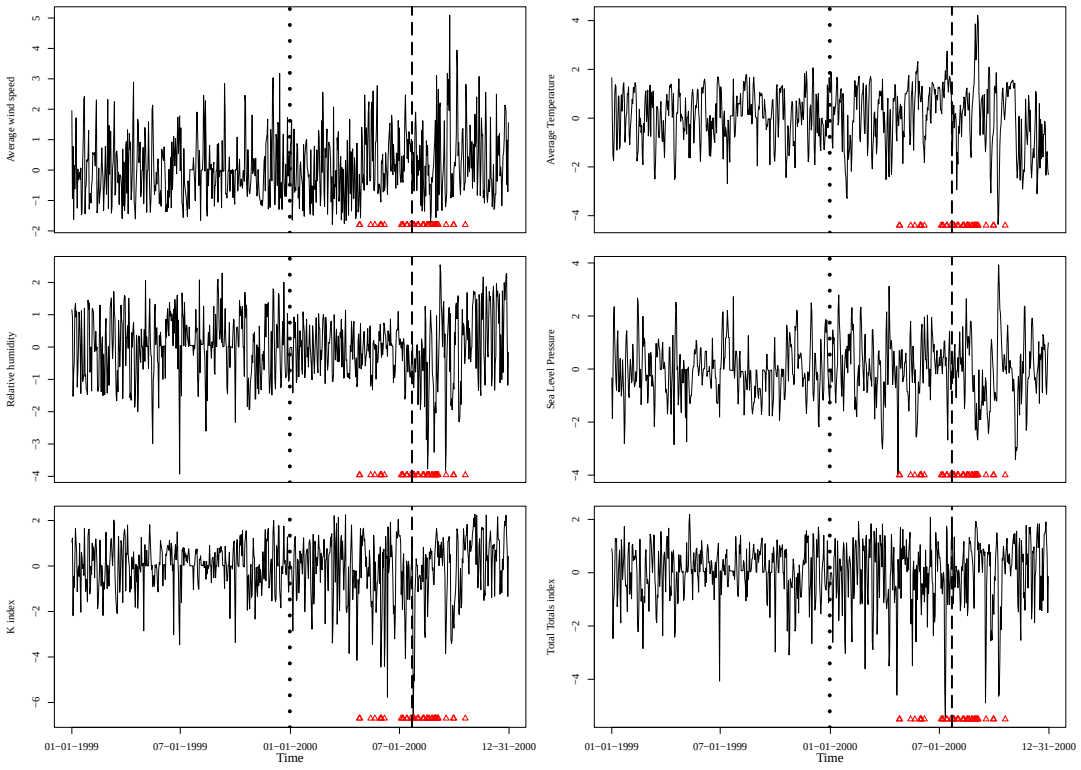

FIG 6. *Standardized observations of the six representative meteorological variables considered in Figure 5. The vertical dotted line in each plot separates the initial IC data from the data for online process monitoring, the small triangles denote the ozone days in the year 2000, and the vertical dashed line indicates the signal time of the proposed process monitoring method PCA-D-C.*

ozone days exist before the first signal time of PCA-D-C, and thus they are not detected by PCA-D-C. This is intuitively reasonable because the signal of a CUSUM chart like PCA-D-C is usually given after the cumulative information in the observed data provides a strong evidence of a shift. Therefore, its first signal is usually given after the shift occurs, and our goal is to make the CUSUM chart react to the shift as quickly as possible.

4.2. *Pollution surveillance after the COVID-19 pandemic.* The COVID-19 pandemic emerged recently had a profound impact on people's daily lives worldwide. To prevent its spread, our government implemented various measures, such as the closure of factories and the promotion of remote work. Despite the significant challenges posed by the pandemic, it had an unexpected positive outcome in terms of the environment. The reduction in traffic and the closure of many factories led to a significant decrease in pollution levels (Venter et al. (2020)). As the pandemic situation gradually improved and many factories reopened, industrial production resumed to compensate for the losses incurred during the pandemic, resulting in an increase in pollution levels post COVID-19 pandemic (Alava and Singh (2022), Barua and Nath (2021)). It is therefore important to monitor pollution levels continuously such that the government can take proactive measures to reduce overproduction and pollution caused by industrial activities to mitigate the adverse effects of pollution on human health and the environment.

In this example, we have selected a total of 21 variables observed in the Houston area to monitor, including some major air quality variables, such as the ozone concentration

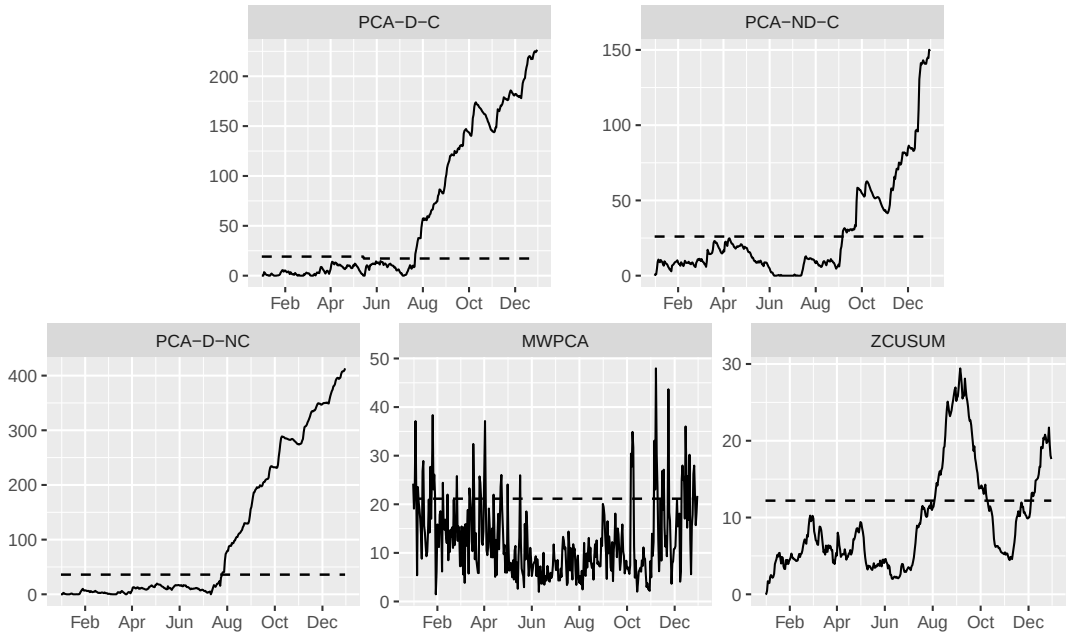

FIG 7. *Five control charts for online monitoring of the meteorological data in the HGB area during January 1 and December 30, 2000. The horizontal dashed line in each plot denotes the control limit of the related control chart.*

levels, Nitrogen Dioxide (NO2), and Carbon Monoxide (CO), and some important meteorological variables, such as the air temperature, dew point temperature (DEW), and wind speed. The observed data of the air quality variables are downloaded from the U.S. Environmental Protection Agency website (https://www.epa.gov), and the observed data of the meteorological variables are downloaded from the Weather Underground website (https://www.wunderground.com). The downloaded data span from January 1, 2020 to December 31, 2021. The observed data of four representative variables, including the ozone concentration levels, CO, daily average wind speed, and DEW, are shown in Figure 8. Then, the first-year data are used as the initial IC data to represent the environmental conditions during the COVID-19 pandemic, and the data in the second year are used for online process monitoring.

To use the proposed method, we first compute the initial IC estimates $\widehat{\mu}_{jl}^{(0)}$ and $\widehat{\sigma}_{jl}^{(0)}$ from the initial IC data, for each j and l. The estimated IC mean functions for the four representative variables are shown in Figure 8 by the solid lines. Then, the initial IC data are standardized using $\{\widehat{\mu}_{jl}^{(0)}\}$ and $\{\widehat{\sigma}_{jl}^{(0)}\}$. The normality assumption for the standardized initial IC data is checked using the Shapiro-Wilk test, which gives a p-value of 2.2×10^{-26}, indicating a significant violation of the normality assumption. The autocorrelation in the standardized initial IC data is also checked using the Durbin-Watson test, which gives p-values $< 1 \times 10^{-4}$ for 20 out of the 21 variables, indicating a significant autocorrelation in the data. In addition, the autocorrelation coefficients for all variables are not statistically significant beyond a lag of 10 observations, implying that the short-ranged serial correlation assumption is reasonable in this application.

Next, the five control charts PCA-D-C, PCA-ND-C, PCA-D-NC, MWPCA, and ZCUSUM are used to monitor the observed data in the second year in the same way as that in Figure 7. The five charts are shown in Figure 9. From the figure, it can be seen that the first signals of the charts PCA-D-C, PCA-ND-C, PCA-D-NC, MWPCA, and ZCUSUM are on

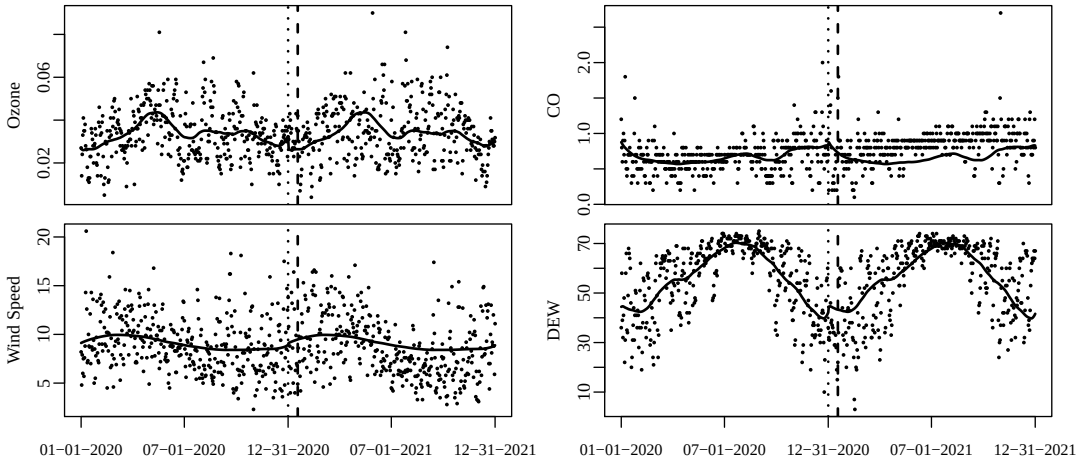

FIG 8. *Original observations of four representative variables: ozone concentration levels, CO, daily average wind speed, and DEW in the years 2020 and 2021. In each plot, the solid curve denotes the estimated IC mean function of the related variable, the vertical dotted line separates the initial IC data from the data for online process monitoring, and the vertical dashed line denotes the first signal time of the proposed chart PCA-D-C.*

1/15/2021, 2/16/2021, 1/17/2021, 1/17/2021, and 1/21/2021, respectively. Thus, the proposed chart PCA-D-C gives the earliest signal in this example. By checking the original process observations of the four representative variables shown in Figure 8, it can be seen that the temporal pattern of the observations is indeed quite different from the estimated IC mean functions around the first signal time of PCA-D-C that is indicated in the plots by the vertical dashed lines.

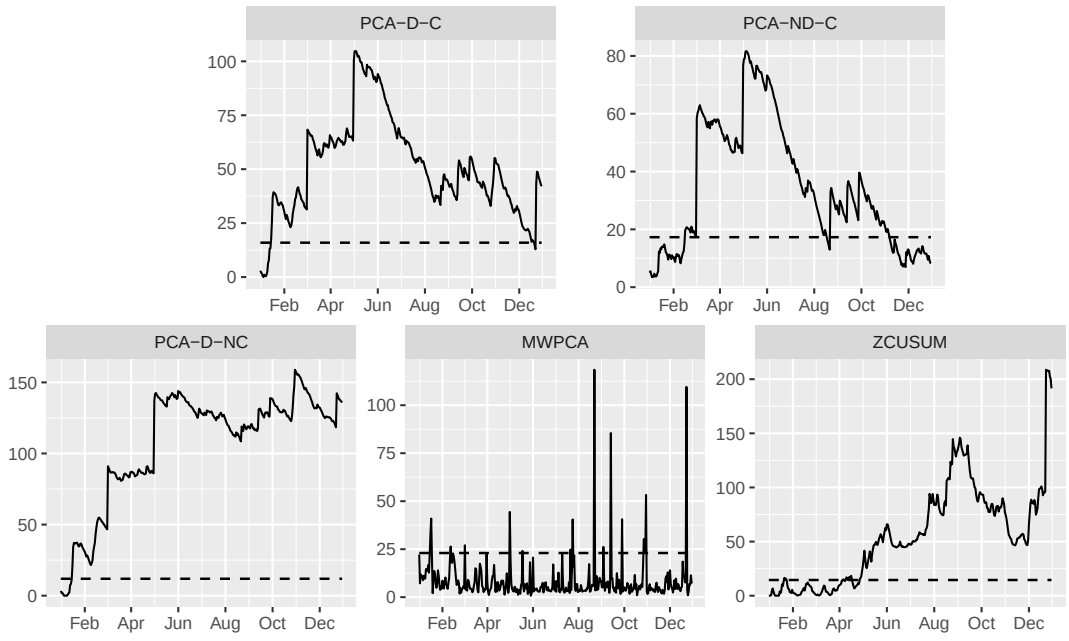

FIG 9. *Five control charts for online monitoring of the air quality data in the Houston area during January 1 and December 31, 2021. The horizontal dashed line in each plot denotes the control limit of the related control chart.*

5. Concluding Remarks. In recent years, ozone pollution has become a major global public health risk factor. It has attracted much attention from governments, and a huge amount of resource has been spent on reducing the negative impact of ozone pollution on public health. In all the effort to handle ozone pollution, early detection of the deterioration of ozone concentration levels is especially important, since it can help governments and individual people to take proper interventions in a timely manner to minimize the impact of ozone pollution on public health. However, this is a challenging task because of the complexity of the ozone data. Motivated by the ozone data observed in the Houston area, we have developed a new method in this paper for air quality surveillance. The proposed online process monitoring chart (i.e., the PCA-D-C chart) is flexible in the sense that the IC process distribution is allowed to be time-varying and the IC process observations at different time points are allowed to have serial correlation. Numerical studies have shown that it has a reliable and effective performance in different cases considered. Because of its generality, besides air quality surveillance, the proposed new method should also be useful for many other applications, including infectious disease surveillance, seismic surveillance, and earthquake monitoring.

We would like to point out that the current version of the proposed method still has some issues to address in future research. For instance, this method mainly concerns the first shift in the process distribution. In the current air quality surveillance problem, the underlying air quality process cannot be stopped after a signal by the proposed method. Thus, continuing online monitoring of the process after the signal is important, which has not been considered by the current method. Furthermore, after the control chart gives a signal, it is important to figure out which variables have the detected shifts and when the shifts start. In the SPC literature, there have been some discussions on post-signal fault diagnosis for high-dimensional data (e.g., Li et al. (2020), Xiang et al. (2022)). It should be checked whether these existing methods are appropriate to use in the current air quality surveillance problem. In addition, the current method assumes that observation times are equally spaced. However, observation times could be unequally spaced in some applications. For instance, medical facilities (e.g., clinics) are usually closed during weekends and holidays. Thus, the quality variables measuring the performance of these medical facilities would not have observed data on such dates. In such cases, ARL_0 and ARL_1 are obviously inappropriate for measuring the performance of the proposed chart (6)-(7), and construction of the related chart should take into account the inequality of the observation times. Also, description and estimation of serial correlation for time series data with unequally spaced observation times could be challenging. Last but not least, the current method assumes that serial correlation in the observed data is short-ranged. However, there could be some processes in practice with long range data dependence (cf., Beran (1992)). All these issues will be studied carefully in our future research.

Acknowledgments: The authors thank the editor, the associate editor, and three referees for their constructive comments and suggestions, which improved the quality of the paper greatly. This research is supported in part by an NSF grant.

SUPPLEMENTARY MATERIAL

Supplementary Materials of the Paper Titled "Online Monitoring of Air Quality Using PCA-Based Sequential Learning"

To save some space in the paper with the above title, the proofs of Theorem 1 and some extra numerical results are presented in this supplementary file.

REFERENCES

Abdul-Wahab, S. A., Bakheit, C. S., and Al-Alawi, S. M. (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software* **20,** 1263–1271.

Alava, J. J. and Singh, G. G. (2022). Changing air pollution and CO2 emissions during the COVID-19 pandemic: Lesson learned and future equity concerns of post-COVID recovery. *Environmental Science & Policy* **130,** 1–8.

Altman, N. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association* **85,** 749–759.

Apley, D. W. and Tsung, F. (2002). The autoregressive t2 chart for monitoring univariate autocorrelated processes. *Journal of Quality Technology* **34,** 80–96.

Barua, S. and Nath, S. D. (2021). The impact of COVID-19 on air pollution: Evidence from global data. *Journal of Cleaner Production* **298,** 126755.

Beran, J. (1992). Statistical methods for data with long-range dependence. *Statistical Science* **7,** 404–416.

Capizzi, G. and Masarotto, G. (2008). Practical design of generalized likelihood ratio control charts for autocorrelated data. *Technometrics* **50,** 357–370.

Capizzi, G. and Masarotto, G. (2011). A least angle regression control chart for multidimensional data. *Technometrics* **53,** 285–296.

Carey, I. M., Atkinson, R. W., Kent, A. J., van Staa, T., Cook, D. G., and Anderson, H. R. (2013). Mortality associations with long-term exposure to outdoor air pollution in a national english cohort. *American Journal of Respiratory and Critical Care Medicine* **187,** 1226–1233.

Chatterjee, S. and Qiu, P. (2009). Distribution-free cumulative sum control charts using bootstrap-based control limits. *The Annals of Applied Statistics* **3,** 349–369.

Chicken, E., Pignatiello, J. J., and Simpson, J. R. (2009). Statistical process monitoring of nonlinear profiles using wavelets. *Journal of Quality Technology* **41,** 198–212.

Crosier, R. B. (1988). Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics* **30,** 291–303.

De Brabanter, K., De Brabanter, J., Suykens, J. A. K., and De Moor, B. (2011). Kernel regression in the presence of correlated errors. *Journal of Machine Learning Research* **12,** 1955–1976.

De Ketelaere, B., Hubert, M., and Schmitt, E. (2015). Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data. *Journal of Quality Technology* **47,** 318–335.

Dong, Y. and Qin, S. J. (2018). A novel dynamic PCA algorithm for dynamic data modeling and process monitoring. *Journal of Process Control* **67,** 1–11.

Draxler, R. R. (2000). Meteorological factors of ozone predictability at houston, texas. *Journal of the Air & Waste Management Association* **50,** 259–271.

Environmental Protection Agency (1999). Guideline for developing an ozone forecasting pro- gram. *Washington: Environmental Protection Agency* .

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications* **14,** 153–158.

Ferrer, A. (2007). Multivariate statistical process control based on principal component analysis (MSPC-PCA): Some reflections and a case study in an autobody assembly process. *Quality Engineering* **19,** 311–325.

Gorai, A. K., Tuluri, F., Tchounwou, P. B., and Ambinakudige, S. (2015). Influence of local meteorology and NO2 conditions on ground-level ozone concentrations in the eastern part of texas, USA. *Air quality, atmosphere, & health* **8,** 81–96.

Hawkins, D. M. (1987). Self-starting cusum charts for location and scale. *Journal of the Royal Statistical Society. Series D (The Statistician)* **36,** 299–316.

Hawkins, D. M., Qiu, P., and Kang, C. W. (2003). The changepoint model for statistical process control. *Journal of Quality Technology* **35,** 355–366.

Health Effects Institute (2019). State of global air 2019. *Boston: Health Effects Institute* .

Hotelling, H. (1947). Multivariate quality control. In *Techniques of Statistical Analysis (C. Eisenhart, M. Hastay, and W.A. Wallis, eds.)*, pages 111–184. McGraw Hill.

Jackson, J. E. (1991). *A User's Guide to Principal Components*. New York: John Wiley & Sons.

Jacob, D. J. and Winner, D. A. (2009). Effect of climate change on air quality. *Atmospheric Environment* **43,** 51–63.

Jenkin, M. E. and Clemitshaw, K. C. (2000). Ozone and other secondary photochemical pollutants: chemical processes governing their formation in the planetary boundary layer. *Atmospheric Environment* **34,** 2499–2527.

Johnson, R. A. and Wichern, D. W. (2008). *Applied Multivariate Statistical Analysis (6th Edition)*. Upper Saddle River, NJ: Pearson.

Knoth, S., Saleh, N. A., Mahmoud, M. A., Woodall, W. H., and Tercero-Gómez, V. G. (2023). A critique of a variety of "memory-based" process monitoring methods. *Journal of Quality Technology* **55,** 18–42.

Knoth, S. and Schmid, W. (2004). Control charts for time series: A review. In Lenz, H.-J. and Wilrich, P.-T., editors, *Frontiers in Statistical Quality Control 7*, pages 210–236. Physica-Verlag HD.

Knoth, S., Tercero-Gómez, V. G., Khakifirooz, M., and Woodall, W. H. (2021). The impracticality of homogeneously weighted moving average and progressive mean control chart approaches. *Quality and Reliability Engineering International* **37,** 3779–3794.

Kourti, T. and MacGregor, J. F. (1996). Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology* **28,** 409–428.

Ku, W., Storer, R. H., and Georgakis, C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **30,** 179–196.

Lennox, B., Montague, G. A., Hiden, H. G., Kornfeld, G., and Goulding, P. R. (2001). Process monitoring of an industrial fed-batch fermentation. *Biotechnology and Bioengineering* **74,** 125–135.

Li, G., Qin, S. J., and Zhou, D. (2014). A new method of dynamic latent-variable modeling for process monitoring. *IEEE Transactions on Industrial Electronics* **61,** 6438–6445.

Li, W., Xiang, D., Tsung, F., and Pu, X. (2020). A diagnostic procedure for high-dimensional data streams via missed discovery rate control. *Technometrics* **62,** 84–100.

Liu, Y., Zhou, Y., and Lu, J. (2020). Exploring the relationship between air pollution and meteorological conditions in china under environmental governance. *Scientific Reports* **10,** 14518.

Lowry, C. A., Woodall, W. H., Champ, C. W., and Rigdon, S. E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics* **34,** 46–53.

Mei, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams. *Biometrika* **97,** 419–433.

Montgomery, D. C. (2012). *Introduction to Statistical Quality Control*. New York: John Wiley & Sons.

Noorossana, R., Saghaei, A., and Amiri, A. (2011). *Statistical Analysis of Profile Monitoring*. Boca Raton, FL: Chapman Hall/CRC.

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16,** 134–153.

Ordóñez, C., Mathis, H., Furger, M., Henne, S., Hüglin, C., Staehelin, J., and Prévôt, A. S. H. (2005). Changes of daily surface ozone maxima in switzerland in all seasons from 1992 to 2002 and discussion of summer 2003. *Atmospheric Chemistry & Physics* **5,** 1187–1203.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika* **41,** 100–115.

Psarakis, S. and Papaleonida, G. E. A. (2007). SPC procedures for monitoring autocorrelated processes. *Quality Technology & Quantitative Management* **4,** 501–540.

Qiu, P. (2014). *Introduction to Statistical Process Control*. Boca Raton, FL: Chapman Hall/CRC.

Qiu, P. (2018). Some perspectives on nonparametric statistical process control. *Journal of Quality Technology* **50,** 49–65.

Qiu, P., Li, W., and Li, J. (2020). A new process control chart for monitoring short-range serially correlated data. *Technometrics* **62,** 71–83.

Qiu, P. and Xiang, D. (2014). Univariate dynamic screening system: An approach for identifying individuals with irregular longitudinal behavior. *Technometrics* **56,** 248–260.

Qiu, P. and Xie, X. (2022). Transparent sequential learning for statistical process control of serially correlated data. *Technometrics* **64,** 487–501.

Qiu, P., Zou, C., and Wang, Z. (2010). Nonparametric profile monitoring by mixed effects modeling. *Technometrics* **52,** 265–277.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics* **1,** 239–250.

Sexton, K. and Linder, S. H. (2015). Houston's novel strategy to control hazardous air pollutants: A case study in policy innovation and political stalemate. *Environmental Health Insights* **9,** 1–12.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. New York: D. Van Nostrand Company.

Statheropoulos, M., Vassiliadis, N., and Pappa, A. (1998). Principal component and canonical correlation analysis for examining air pollution and meteorological data. *Atmospheric Environment* **32,** 1087–1095.

Sun, W., Palazoglu, A., Singh, A., Zhang, H., Wang, Q., Zhao, Z., and Cao, D. (2015). Prediction of surface ozone episodes using clusters based generalized linear mixed effects models in houston–galveston–brazoria area, texas. *Atmospheric Pollution Research* **6,** 245–253.

Tartakovsky, A. G., Rozovskii, B. L., Blazek, R. B., and Kim, H. (2006). Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology* **3,** 252–340.

Tsung, F. (2000). Statistical monitoring and diagnosis of automatic controlled processes using dynamic PCA. *International Journal of Production Research* **38,** 625–637.

Vanhatalo, E. and Kulahci, M. (2016). Impact of autocorrelation on principal components and their use in statistical process control. *Quality and Reliability Engineering International* **32,** 1483–1500.

Venter, Z., Aunan, K., Chowdhury, S., and Lelieveld, J. (2020). COVID-19 lockdowns cause global air pollution declines. *Proceedings of the National Academy of Sciences* **117,** 18984–18990.

Wang, K. and Jiang, W. (2009). High-dimensional process monitoring and fault isolation via variable selection. *Journal of Quality Technology* **41,** 247–258.

Wang, X., Kruger, U., and Irwin, G. W. (2005). Process monitoring approach using fast moving window PCA. *Industrial and Engineering Chemistry Research* **44,** 5691–5702.

Weng, J., Zhang, Y., and Hwang, W.-S. (2003). Candid covariance-free incremental principal component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25,** 1034–1040.

World Health Organization (1976). Photochemical oxidants: Environmental health criteria 7. *Geneva: World Health Organization* .

Xiang, D., Qiu, P., and Pu, X. (2013). Nonparametric regression analysis of multivariate longitudinal data. *Statistica Sinica* **23,** 769–789.

Xiang, D., Qiu, P., Wang, D., and Li, W. (2022). Reliable post-signal fault diagnosis for correlated high-dimensional data streams. *Technometrics* **64,** 323–334.

Xie, X. and Qiu, P. (2023). Control charts for dynamic process monitoring with an application to air pollution surveillance. *The Annals of Applied Statistics* **17,** 47–66.

Yang, K. and Qiu, P. (2018). Spatiotemporal incidence rate data analysis by nonparametric regression. *Statistics in Medicine* **37,** 2094–2107.

Zhang, K. and Fan, W. (2008). Forecasting skewed biased stochastic ozone days: analyses, solutions and beyond. *Knowledge and Information Systems* **14,** 299–326.

Zhao, X., Zhang, X., Xu, X., Xu, J., Meng, W., and Pu, W. (2009). Seasonal and diurnal variations of ambient PM2.5 concentration in urban and rural environments in beijing. *Atmospheric Environment* **43,** 2893–2900.

Zou, C. and Qiu, P. (2009). Multivariate statistical process control using LASSO. *Journal of the American Statistical Association* **104,** 1586–1596.

Zou, C., Wang, Z., Zi, X., and Jiang, W. (2015). An efficient online monitoring method for high-dimensional data streams. *Technometrics* **57,** 374–387.