Wide-Area Monitoring of Power Systems Using Principal Component Analysis and *k*-Nearest Neighbor Analysis

Lianfang Cai, Nina F. Thornhill, *Senior Member, IEEE*, Stefanie Kuenzel, *Member, IEEE*, and Bikash C. Pal, *Fellow, IEEE*

***Abstract*—Wide-area monitoring of power systems is important for system security and stability. It involves the detection and localization of power system disturbances. However, the oscillatory trends and noise in electrical measurements often mask disturbances, making wide-area monitoring a challenging task. This paper presents a wide-area monitoring method to detect and locate power system disturbances by combining multivariate analysis known as Principal Component Analysis (PCA) and time series analysis known as *k*-Nearest Neighbor (*k*NN) analysis. Advantages of this method are that it can not only analyze a large number of wide-area variables in real time but also can reduce the masking effect of the oscillatory trends and noise on disturbances. Case studies conducted on data from a four-variable numerical model and the New England power system model demonstrate the effectiveness of this method.**

***Index Terms*—Wide-area monitoring, electrical measurements, power system disturbances, security, stability, detection, localization, *k*-nearest neighbor, principal component analysis, real time.**

NOMENCLATURE

AI_{T^2}	Monitoring statistic built by applying *k*NN on T^2.
$AI_{T^2}^{\alpha}$	Detection threshold with confidence level α for AI_{T^2}.
$AI_{T^2,p}^{\circ}$	pth value of AI_{T^2} calculated online.
AI_Q	Monitoring statistic built by applying *k*NN on Q.
AI_Q^{α}	Detection threshold with confidence level α for AI_Q.
$AI_{Q,r}$	rth value of AI_Q calculated offline.
$AI_{Q,p}^{\circ}$	pth value of AI_Q calculated online.
a	Number of principal components.
$\boldsymbol{C}$	Covariance matrix of normalized variables.
$\mathbf{con}_{AI_Q,p}^{\circ}$	Vector of contributions of variables to AI_Q at the pth sampling time point online.
$\mathbf{con}_{AI_{T^2},p}^{\circ}$	Vector of contributions of variables to AI_{T^2} at the pth sampling time point online.
$\mathrm{CPV}(a)$	Ratio percentage of sum of $\lambda_1, \lambda_2, \cdots, \lambda_a$ over sum of $\lambda_1, \lambda_2, \cdots, \lambda_m$.
D^2	Square of Euclidean distance between two windows.
d	Derivarive operator.
$\boldsymbol{e}$	Vector of residual variables obtained by PCA.
e_i	ith residual varaible.
g	Number for a data window formulated offline.
$\boldsymbol{h}$	Vector of principal components obtained by PCA.
h_i	ith principal component.
$\mathbf{I}_m$	Identity matrix with dimension as $m \times m$.
k	Parameter for *k*NN.
L	Length of data window.
l	Temporary variable for counting from 1 to L.
m	Number of measured variables.
N	Size of modelling dataset.
n	Sampling time point for offline data.
p	Sampling time point for online data.
Q	Squared Prediction Error (SPE) statistic calcualted based on residual variables.
Q_n	nth value of Q calculated offline.
Q_p°	pth value of Q calculated online.
r	Number for a data window formulated offline, and $r \neq g$.
$\mathrm{r}_1, \mathrm{r}_2$	Specific values of r (constants).
s_i	ith sinusoidal signal.
$s_{i,t}$	Value of s_i at the time of t.
T^2	Hotelling’s statistic calculated based on principal components.
${T^2}_n$	nth value of T^2 calculated offline.
${T^2}_p^{\circ}$	pth value of T^2 calculated online.
t	Continuous time.
$\boldsymbol{U}$	Matrix with columns as $\boldsymbol{u}_1\ \boldsymbol{u}_2\ \cdots\ \boldsymbol{u}_m$.
$\boldsymbol{U}_{1:a}$	Matrix with columns as $\boldsymbol{u}_1\ \boldsymbol{u}_2\ \cdots\ \boldsymbol{u}_a$.
$\boldsymbol{u}_i$	ith eigenvector of $\boldsymbol{C}$.
$\boldsymbol{x}$	Vector of measured variables.
$\boldsymbol{x}_n$	nth vector value of $\boldsymbol{x}$ for offline modelling.
x_i	ith measured variable.
$x_{i,n}$	nth value of x_i for offline modelling.
$x_{i,t}$	Value of x_i at the time of t.
$\widetilde{\boldsymbol{x}}$	Vector of normalized variables.
$\widetilde{\boldsymbol{x}}_n$	nth vector value of $\widetilde{\boldsymbol{x}}$ calculated offline.
$\widetilde{\boldsymbol{x}}_p^{\circ}$	pth vecotr value of $\widetilde{\boldsymbol{x}}$ calculated online.
$\tilde{x}_i$	ith normalized variable.
$\tilde{x}_{i,n}$	nth value of $\tilde{x}_i$ calculated offline.
$\boldsymbol{Z}$	Embedding matrix of Q formulated offline.
$\boldsymbol{z}_r$	rth data window formulated offline (rth row of $\boldsymbol{Z}$).
$\boldsymbol{z}_g$	gth data window formulated offline (gth row of $\boldsymbol{Z}$).
$\boldsymbol{z}_p^{\circ}$	pth data window formulated online.
α	Confidence level for detection thresholds.

This work was supported by the U.K. Engineering and Physical Sciences Research Council under Grant “EP/L014343/1” “Stability and Control of Power Networks with Energy Storage (STABLE-NET).”

L. Cai and N. F. Thornhill are with the Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, SW7 2AZ, U.K. (e-mail: l.cai@imperial.ac.uk; n.thornhill@imperial.ac.uk).

S. Kuenzel is with the Department of Electronic Engineering, Royal Holloway, University of London, TW20 0EX, U.K. (e-mail: stefanie.kuenzel@rhul.ac.uk).

B. C. Pal is with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ, U.K. (e-mail: b.pal@imperial.ac.uk).

$\boldsymbol{\Lambda}$ Diagonal matrix with diagonal elements as $\lambda_1, \lambda_2, \cdots, \lambda_m$.

λ_i ith eigenvalue of $\boldsymbol{C}$.

$\boldsymbol{\Omega}$ Diagonal matrix with diagonal elements as $\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_a^{-1}$.

I. Introduction

WIDE-AREA monitoring of power systems plays a crucial role in understanding the system behavior and improving the system operating stability margin. It usually places much emphasis on the detection and localization of disturbances, because disturbances pose an increasingly severe threat to the system security and stability [1].

Generally, disturbances deteriorate the system health by making a power system deviate from the normal operating status. With more and more advanced measuring devices such as Phasor Measurement Units (PMUs) spreading across power systems, abundant measurements containing the information of the system operating status are available for analysis. How to extract such information from the measured data for disturbance detection and localization is an important issue for power system researchers [2]. Generally, the existing data-driven methods can be divided into three categories according to the applications: (1) for the protection of power system equipment, e.g., the wavelet coefficient energy based method [3] and the hidden Markov model based method [4]; (2) for the analysis of power quality especially the waveform of alternate voltage, e.g., the Hilbert-Huang transform based method [5] and the power quality state estimation based method [6]; (3) for the assessment of the system security and stability, typically by multivariate statistical analysis based methods [7]-[10].

Usually, the first two categories of methods take a univariate approach to analyze electrical variables separately. In contrast, the third category of methods use a multivariate approach to handle variables together, particularly suitable for wide-area monitoring of power systems where many variables need to be analyzed simultaneously. This work focuses on the latter.

Principal Component Analysis (PCA), one of the classical multivariate statistical analysis techniques, is well-known for its capability of compressing high-dimensional and correlated data without significant loss of information. It obtains Principal Components (PCs) that are uncorrelated and Residual Variables (RVs) by projecting physical variables onto a low-dimensional subspace that retains most of the variances of the projected variables [11]. To measure the variation of PCs within the PCA model and the variation of RVs not accounted for by the PCA model, two popular monitoring statistics were used respectively, that is, the Hotelling's T^2 statistic calculated as the sum of the squares of normalized PCs and the companion Squared Prediction Error (SPE or Q) statistic calculated as the sum of the squares of RVs [11]. The PCA model together with the T^2 and Q statistics, known as the PCA-based statistical monitoring method, have been widely applied for process monitoring in the chemical industry [11].

In 2013, Barocio *et al.* [7] introduced the PCA-based statistical monitoring method for the detection and visualization of power system disturbances and discussed its potential for wide-area monitoring of power systems. Subsequently, Liu *et al.* [8] focused on the geometric interpretation of T^2 and Q, and showed that by using frequency measurements T^2 detects generation mismatch events and Q detects islanding events. Recently, Rafferty *et al.* [9] considered the changing nature of frequency in a power system and developed a moving window PCA based statistical monitoring method updating the PCA model as well as T^2 and Q after obtaining a new window of frequency measurements. Although the existing works have led to some success in wide-area monitoring of power systems, one issue that affects the monitoring has not been considered.

Specifically, the above works require the amplitude of electrical measurements recorded before and after disturbances to be markedly different so that the amplitude of the T^2 values and that of the Q values calculated before and after disturbances can also be distinct and thus can be made use of to detect disturbances at the system-wide level. In practice, such a requirement cannot be met all the time, especially for the cases in power systems where electrical measurements often have oscillatory trends and noise [12], [13]. As exemplified in [14], the oscillatory trends and noise in measurements often mask disturbances, making the difference in the amplitude of measurements recorded before and after disturbances not distinguishable. As a result, there is also not much difference in the amplitude of the T^2 values and that of the Q values calculated before and after disturbances, and therefore it is difficult for T^2 and Q to detect disturbances using electrical measurements with oscillatory trends and noise.

k-Nearest Neighbor (*k*NN) analysis is a time series analysis method for the detection of anomalous data windows [15]-[18]. As stated in [14], *k*NN does not require the amplitude of measurements recorded before and after anomalies to be distinct while detecting anomalies. In a recent paper of the authors [19], *k*NN was introduced and adapted for real-time detection of power system disturbances. However, the method presented in [19] operates in a univariate manner to analyze variables separately and the online computational burden increases with the number of variables increasing.

Against this background, the motivation of this work is to integrate *k*NN with the PCA-based statistical monitoring method in order that a large number of variables can be analyzed in real time for wide-area monitoring of power systems, and at the same time, the masking effect of the oscillatory trends and noise in electrical measurements on disturbances can be reduced. More specifically, *k*NN is applied on T^2 and Q to obtain two new monitoring statistics for detecting disturbances. This paper will show that a *k*NN analysis in real time of T^2 and Q leads to more rapid detection of disturbances. The real-time implementation of *k*NN is achieved by building a recursive calculation strategy for the distance measure of *k*NN and a fast selection strategy for the *k*th smallest distance value. Finally, disturbance localization is performed by developing a contribution plot strategy which can quantify the contributions of variables to the new monitoring statistics. Case studies conducted on a four-variable numerical model and the New England power system model are used to demonstrate the effectiveness of the proposed method. It is worth noting that the proposed method is not relevant to protective relays since they fall into different categories, as stated previously.

The paper is organized as follows. Section II gives a brief description of wide-area monitoring based on PCA. Section III presents the wide-area monitoring method based on PCA and *k*NN. The application results and analysis of the two case studies are provided in Section IV. Discussions about the proposed method are given in Section V, while our conclusions are drawn in Section VI.

The following notational conventions are used throughout this contribution. Boldface capital and lower-case letters stand for matrices and column vectors respectively, while $\mathbb{R}$ denotes the field of real numbers. The transpose and inverse operators are denoted by $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{-1}$ respectively.

II. Wide-Area Monitoring Based on PCA

In this section, wide-area monitoring based on PCA [7]-[9], referred to as WAM-PCA here, is briefly introduced.

The symbol $\boldsymbol{x}^{\mathrm{T}} = [x_1\ x_2\ \cdots\ x_m\ ]$ denotes a vector of m electrical variables measured for monitoring, e.g., frequency, voltage amplitude, active power, reactive power. Historical measurements from the ambient condition are used to form the modelling data $\{\boldsymbol{x}_n\}_{n=1}^{N}$, where N denotes the dataset size and $\boldsymbol{x}_n$ denotes the nth vector value of $\boldsymbol{x}$. In what follows, PCA is used to analyze the measured variables together and to obtain PCs and RVs through multivariate analysis.

Firstly, the variables in the vector $\boldsymbol{x}$ are normalized with the sample means and sample variances calculated from $\{\boldsymbol{x}_n\}_{n=1}^{N}$ to make the obtained variables independent of their engineering units. The symbol $\tilde{\boldsymbol{x}}^{\mathrm{T}} = [\tilde{x}_1\ \tilde{x}_2\ \cdots\ \tilde{x}_m]$ denotes a vector of the m normalized variables. The covariance matrix of $\tilde{\boldsymbol{x}}^{\mathrm{T}}$ can be estimated based on the normalized data $\{\tilde{\boldsymbol{x}}_n\}_{n=1}^{N}$ and the eigenvalue decomposition of $\boldsymbol{C}$ can be implemented as:

$$\boldsymbol{C} = \frac{1}{N-1}\sum_{n=1}^{N} \tilde{\boldsymbol{x}}_n \tilde{\boldsymbol{x}}_n{}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}} = \sum_{i=1}^{m} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i{}^{\mathrm{T}} \qquad (1)$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with diagonal elements as the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_m$ of $\boldsymbol{C}$ in the descending order, while $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ is the eigenvector matrix with the column vectors as the eigenvectors $\boldsymbol{u}_1\ \boldsymbol{u}_2\ \cdots\ \boldsymbol{u}_m$ of $\boldsymbol{C}$.

Then, a vector of PCs can be obtained by:

$$\boldsymbol{h}^{\mathrm{T}} = [h_1\ h_2\ \cdots\ h_a] = \left(\boldsymbol{U}_{1:a}{}^{\mathrm{T}}\tilde{\boldsymbol{x}}\right)^{\mathrm{T}} \qquad (2)$$

where a is the number of PCs satisfying $a < m$, and $\boldsymbol{U}_{1:a}{}^{\mathrm{T}} = [\boldsymbol{u}_1\ \boldsymbol{u}_2\ \cdots\ \boldsymbol{u}_a]^{\mathrm{T}} \in \mathbb{R}^{a\times m}$ is called loading matrix. The sample covariance matrix of PCs is a diagonal matrix with the diagonal elements as $\lambda_1\ \lambda_2\ \cdots\ \lambda_a$.

Concurrently, a vector of RVs can be obtained by:

$$\boldsymbol{e}^{\mathrm{T}} = [e_1\ e_2\ \cdots\ e_m] = \left(\tilde{\boldsymbol{x}} - \boldsymbol{U}_{1:a}\boldsymbol{U}_{1:a}{}^{\mathrm{T}}\tilde{\boldsymbol{x}}\right)^{\mathrm{T}} \qquad (3)$$

The variation of PCs within the PCA model can be measured by the T^2 statistic:

$$T^2 = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{h} = \sum_{i=1}^{a}\left(h_i/\sqrt{\lambda_i}\right)^2 \qquad (4)$$

where $\boldsymbol{\Omega} \in \mathbb{R}^{a\times a}$ is a diagonal matrix with the diagonal elements as $\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_a^{-1}$.

Moreover, the variation of RVs not accounted for by the PCA model can be measured by the Q statistic:

$$Q = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = \sum_{i=1}^{m} e_i^2 \qquad (5)$$

III. Wide-Area Monitoring Based on PCA and KNN

Both T^2 and Q make use of the difference in their amplitude before and after disturbances for disturbance detection, requiring the amplitude of electrical measurements recorded before and after disturbances to be distinct. However, the oscillatory trends and noise in electrical measurements often have a masking effect on disturbances, making it difficult to satisfy this requirement. Thus, the detection performance of T^2 and Q will be adversely affected. In this section, *k*NN is introduced and applied on T^2 and Q to build two new monitoring statistics for improving the detection performance. The reason why *k*NN gives the improvement is because *k*NN does not require the amplitude of a time series before and after disturbances to be distinct [14]. Then, disturbance localization is performed by quantifying the contributions of variables to the new monitoring statistics. Both disturbance detection and localization constitute the subject of wide-area monitoring based on PCA and *k*NN, referred to as WAM-PCA*k*NN here. In the following, WAM-PCA*k*NN is presented in detail.

A. Disturbance Detection of WAM-PCAkNN

*k*NN adopts a certain type of distance measure to assess the similarity of two data windows in a time series, where a data window refers to a segment of data with the fixed length. Data windows with similar sequences of samples are called near neighbors. The similarity assessment is achieved by defining an Anomaly Index (AI) for each data window. Following the definition of AI in [14]-[16], this paper uses the distance of a data window to its *k*th nearest neighbor as AI of that data window. Anomalous data windows are those distinct from the underlying trend of the time series and the AI value for an anomalous data window will be much higher than that of any normal data window, which is the reason why *k*NN can be used for anomaly detection. A common distance measure to assess the similarity between data windows is Euclidean Distance (ED) [14]-[19], which can be written as:

$$D(\boldsymbol{\varphi}, \boldsymbol{\phi}) \triangleq \sqrt{\sum_{j=1}^{L}\left(\varphi_j - \phi_j\right)^2} \geq 0 \qquad (6)$$

where $\boldsymbol{\varphi}^{\mathrm{T}} = [\varphi_1\ \varphi_2\ \cdots\ \varphi_L]$ and $\boldsymbol{\phi}^{\mathrm{T}} = [\phi_1\ \phi_2\ \cdots\ \phi_L]$ denote two data windows with L measurements in each one, $D(\boldsymbol{\varphi}, \boldsymbol{\phi}) = 0$ indicates the maximum similarity.

This paper also uses ED to assess the similarity of two data windows. The reason why ED is used here instead of other types of distance measures such as Mahalanobis Distance (MD) is because the calculation of ED is much simpler which can facilitate the recursive calculation for the online detection.

If the T^2 or Q values obtained by (4) or (5) are viewed as a time series, the detection of power system disturbances can be achieved by detecting anomalous windows in this time series. Without loss of generality, Q is taken to illustrate the detection process, which also applies to T^2. The detection process includes: 1) the offline modelling; 2) the online detection.

1) The offline modelling

The offline modelling calculates a sequence of the AI values by using *k*NN to analyze the Q values. It then calculates a detection threshold based on the obtained AI values for determining whether disturbances occur or not.

Specifically, based on the modelling data $\{\boldsymbol{x}_n\}_{n=1}^{N}$, the Q values $\{Q_n\}_{n=1}^{N}$ are calculated by (5) and a matrix $\boldsymbol{Z}$ is built as:

$$\boldsymbol{Z}=\begin{bmatrix}\boldsymbol{z}_1{}^{\mathrm{T}}\\ \boldsymbol{z}_2{}^{\mathrm{T}}\\ \vdots\\ \boldsymbol{z}_{N-L+1}{}^{\mathrm{T}}\end{bmatrix}=\begin{bmatrix}Q_1 & Q_2 & \cdots & Q_L\\ Q_2 & Q_3 & \cdots & Q_{L+1}\\ \vdots & \vdots & \cdots & \vdots\\ Q_{N-L+1} & Q_{N-L+2} & \cdots & Q_N\end{bmatrix} \tag{7}$$

where $\boldsymbol{Z}$ is the embedding matrix of Q, its row $\boldsymbol{z}_r{}^{\mathrm{T}}$ denotes the rth data window of $\{Q_n\}_{n=1}^{N}$, and L denotes the window length. Two rows can be compared by the Square of ED (SED) as:

$$D^2(\boldsymbol{z}_g,\boldsymbol{z}_r)=\sum_{l=1}^{L}\left(Q_{g-l+L}-Q_{r-l+L}\right)^2 \tag{8}$$

The reason for using SED instead of directly using ED is due to the consideration of the calculation efficiency. This can be observed later in *2) The online detection*. For the rth row $\boldsymbol{z}_r{}^{\mathrm{T}}$, its AI value is calculated as the kth smallest SED value between it and all other rows except its near-in-time rows. The near-in-time rows of $\boldsymbol{z}_r{}^{\mathrm{T}}$ are those having at least one sample in common with $\boldsymbol{z}_r{}^{\mathrm{T}}$, e.g., $\boldsymbol{z}_L{}^{\mathrm{T}}$ is the last near-in-time row of $\boldsymbol{z}_1{}^{\mathrm{T}}$. The exclusion of the SED values between $\boldsymbol{z}_r{}^{\mathrm{T}}$ and its near-in-time rows during the calculation of AI is to avoid treating such near-in-time rows as near neighbors of $\boldsymbol{z}_r{}^{\mathrm{T}}$.

When all rows of $\boldsymbol{Z}$ obtain their corresponding AI values, a threshold is needed for the online detection. Based on the sequence of the obtained AI values $\{AI_{Q,r}\}_{r=1}^{N-L+1}$ where AI_Q denotes the new monitoring statistic built by applying kNN on the Q statistic and $AI_{Q,r}$ denotes the rth value of AI_Q, a threshold AI_Q^{α} with the confidence level α can be calculated as the δth highest value of this sequence, where δ is the integer nearest to $(1-\alpha)(N-L+1)$ [11].

Similar with AI_Q, another new monitoring statistic AI_{T^2} can be built by applying kNN on the T^2 statistic and the related detection threshold $AI_{T^2}^{\alpha}$ can also be determined.

2) The online detection

Next is the online detection, for which real-time calculation of AI is required. To meet this requirement, strategies for recursively calculating SED and for fast selecting the kth smallest SED value are built below.

The symbol $\boldsymbol{z}_p^{\circ}{}^{\mathrm{T}}=[Q_{p-L+1}^{\circ}\quad Q_{p-L+2}^{\circ}\quad\cdots\quad Q_p^{\circ}]$ denotes the data window of the L continuous Q values calculated based on the new measurements, where p denotes the sampling time point for the online data and the symbol "∘" is used to distinguish the online data from the offline data. Because all rows of $\boldsymbol{Z}$ in (7) are normal windows with the ambient characteristic, they can be taken as the reference data to test whether $\boldsymbol{z}_p^{\circ}{}^{\mathrm{T}}$ deviates from normal or not. If $\boldsymbol{z}_p^{\circ}{}^{\mathrm{T}}$ is anomalous, the SEDs between it and all rows of $\boldsymbol{Z}$ will be large. Accordingly, the AI value $AI_{Q,p}^{\circ}$ for $\boldsymbol{z}_p^{\circ}{}^{\mathrm{T}}$ which is the kth smallest SED value will also be large and will go beyond the threshold AI_Q^{α}. For the rth row $\boldsymbol{z}_r{}^{\mathrm{T}}$ of $\boldsymbol{Z}$, the SED between it and $\boldsymbol{z}_p^{\circ}{}^{\mathrm{T}}$ can be calculated as:

$$D^2(\boldsymbol{z}_p^{\circ},\boldsymbol{z}_r)=\sum_{l=1}^{L}\left(Q_{p-l+1}^{\circ}-Q_{r-l+L}\right)^2 \tag{9}$$

The calculation of (9) needs $(2L-1)$ additions and L multiplications. So, the online computation load relies largely on the window length L. To reduce the number of mathematical operations needed in (9), a recursive calculation strategy, called *Strategy* Γ here, is built using the result previously calculated.

(I) *Strategy* Γ for recursively calculating SED

For the window $\boldsymbol{z}_{p-1}^{\circ}{}^{\mathrm{T}}=[Q_{p-L}^{\circ}\quad Q_{p-L+1}^{\circ}\quad\cdots\quad Q_{p-1}^{\circ}]$ obtained a sampling time point earlier than $\boldsymbol{z}_p^{\circ}{}^{\mathrm{T}}$, the SED between it and the row $\boldsymbol{z}_{r-1}{}^{\mathrm{T}}$ of $\boldsymbol{Z}$ can be calculated as:

$$D^2(\boldsymbol{z}_{p-1}^{\circ},\boldsymbol{z}_{r-1})=\sum_{l=1}^{L}\left(Q_{p-l}^{\circ}-Q_{r-l+L-1}\right)^2 \tag{10}$$

Using (9) and (10), a recursive equation can be obtained as:

$$D^2(\boldsymbol{z}_p^{\circ},\boldsymbol{z}_r)=\begin{cases}D^2(\boldsymbol{z}_{p-1}^{\circ},\boldsymbol{z}_{r-1})+\left(Q_p^{\circ}-Q_{r-1+L}\right)^2\\ \quad-\left(Q_{p-L}^{\circ}-Q_{r-1}\right)^2,\ r>=2\\ \sum_{l=1}^{L}\left(Q_{p-l+1}^{\circ}-Q_{r-l+L}\right)^2,\ r=1\end{cases} \tag{11}$$

In comparison to (9), the calculation of $D^2(\boldsymbol{z}_p^{\circ},\boldsymbol{z}_r)$ in (11) only requires four addition and two multiplication operations for $r>=2$, which is beneficial to the real-time requirement. Here, the reason why SED instead of ED is used can be seen, which is due to the need of the recursive calculation.

Using (11), the sequence of the SED values $\{D^2(\boldsymbol{z}_p^{\circ},\boldsymbol{z}_r)\}_{r=1}^{N-L+1}$ can be calculated more efficiently. Then, the AI value $AI_{Q,p}^{\circ}$ for $\boldsymbol{z}_p^{\circ}{}^{\mathrm{T}}$ can be determined as the kth smallest SED value. A strategy for fast selection of the kth smallest element from a sequence is built and described below.

(II) *Strategy* ΓΓ for fast selection of the kth smallest SED

If k elements from a sequence are smaller than the rest, the maximum one of these k elements is the kth smallest element of the entire sequence. *Strategy* ΓΓ is built based on such consideration. Firstly, the first k elements of $\{D^2(\boldsymbol{z}_p^{\circ},\boldsymbol{z}_r)\}_{r=1}^{N-L+1}$ are sorted in the ascending order, denoted as $D^{2(1)},D^{2(2)},\cdots,D^{2(k)}$. Then, the $(k+1)$th element, denoted as $D^{2(*)}$, is compared with the k elements. If $D^{2(*)}$ is larger than $D^{2(k)}$, $D^{2(*)}$ is removed and the k elements remain unchanged; otherwise, $D^{2(k)}$ is removed and $D^{2(*)}$ is put into $D^{2(1)},D^{2(2)},\cdots,D^{2(k-1)}$ensuring the reserved k elements are still in the ascending order. After each element of $\{D^2(\boldsymbol{z}_p^{\circ},\boldsymbol{z}_r)\}_{r=k+1}^{N-L+1}$ is handled by such comparison, the maximum one of the ultimately reserved k elements is the kth smallest element of the entire SED sequence.

For the best case, $D^{2(*)}$ only needs to be compared with $D^{2(k)}$. For the worst case, $D^{2(*)}$ needs to be compared with all k elements, e.g., $D^{2(k-1)}<=D^{2(*)}<=D^{2(k)}$ and $D^{2(*)}$ is compared with $D^{2(1)},D^{2(2)},D^{2(3)},\cdots,D^{2(k-1)}$ in turn besides $D^{2(k)}$. To reduce the number of comparisons, the binary search is introduced to search the target position for $D^{2(*)}$. It begins by comparing $D^{2(*)}$ with the middle one of the k elements. If $D^{2(*)}$ is not larger than the middle one, the search continues on the former half of the k elements; otherwise, the search continues on the latter half. The search continues, eliminating half of the elements, and comparing $D^{2(*)}$ to the middle one of the remaining elements, until the target position is found. The number of comparisons is $\log_2(k)$ at most, smaller than k.

In addition, the binary search is also used to sort the first k elements of the SED sequence in the ascending order by putting them into target positions one by one. The only difference is that, when one of the first k elements is put into the target position, the maximum element does not need to be removed. The number of comparisons is $\log_2(k!)$ at most. Thus, the total number of comparisons for *Strategy* ΓΓ is $\log_2(k!)+(N-L+1-k)\cdot\log_2(k)$ at most. Through this strategy, the AI value $AI_{Q,p}^{\circ}$ for ${\boldsymbol{z}_p^{\circ}}^{\mathrm{T}}$ can be obtained as the maximum one of the ultimately reserved k elements and can be compared with the threshold AI_Q^{α} for the online detection.

Similarly, the AI value $AI_{T^2,p}^{\circ}$ can also be obtained by *Strategy* Γ and *Strategy* ΓΓ, and can be compared with the threshold $AI_{T^2}^{\alpha}$. Thus, disturbance detection of WAM-PCA*k*NN has been developed, which is summarized in Fig. 1.

B. Disturbance Localization of WAM-PCAkNN

Once a disturbance is detected, it needs to be located. Since the variables nearest to a local disturbance are usually affected most, identifying the variables affected most by the detected disturbance can provide a meaningful reference for disturbance localization. In the present study, contribution plots display the effectiveness in identifying such variables [20]. Usually, variables with largest contributions to monitoring statistics are the ones affected most by disturbances. In the following, a contribution plot strategy that can quantify the Contributions of Variables (CVs) to AI_Q and AI_{T^2} is built.

When the online detection is implemented, the AI value $AI_{Q,p}^{\circ}$ for ${\boldsymbol{z}_p^{\circ}}^{\mathrm{T}}$ is obtained as the kth smallest element of the SED sequence $\left\{D^2\left(\boldsymbol{z}_p^{\circ},\boldsymbol{z}_r\right)\right\}_{r=1}^{N-L+1}$, which can be calculated as:

$$AI_{Q,p}^{\circ}=D^2\left(\boldsymbol{z}_p^{\circ},\boldsymbol{z}_{\mathrm{r}_1}\right)=\sum_{l=1}^{L}\left(Q_{p-l+1}^{\circ}-Q_{\mathrm{r}_1-l+L}\right)^2 \tag{12}$$

where ${\boldsymbol{z}_{\mathrm{r}_1}}^{\mathrm{T}}=[Q_{\mathrm{r}_1}\quad Q_{\mathrm{r}_1+1}\quad\cdots\quad Q_{\mathrm{r}_1+L-1}]$ and r_1 is a constant denoting a specific value of r.

Similar with $AI_{Q,p}^{\circ}$, the AI value $AI_{T^2,p}^{\circ}$ can be calculated as:

$$AI_{T^2,p}^{\circ}=\sum_{l=1}^{L}\left({T^2}^{\circ}_{p-l+1}-{T^2}_{\mathrm{r}_2-l+L}\right)^2 \tag{13}$$

where r_2 is also a constant denoting a specific value of r.

Then, the CVs to AI_Q can be obtained by:

$$\begin{aligned}\mathbf{con}_{AI_Q,p}^{\circ}&=\sum_{l=1}^{L}\left|\frac{d\left(Q_{p-l+1}^{\circ}-Q_{\mathrm{r}_1-l+L}\right)^2}{d\tilde{\boldsymbol{x}}_{p-l+1}^{\circ}}\right|\\&=\sum_{l=1}^{L}\left|2\left(Q_{p-l+1}^{\circ}-Q_{\mathrm{r}_1-l+L}\right)\frac{dQ_{p-l+1}^{\circ}}{d\tilde{\boldsymbol{x}}_{p-l+1}^{\circ}}\right|\\&=\sum_{l=1}^{L}\left|4\left(Q_{p-l+1}^{\circ}-Q_{\mathrm{r}_1-l+L}\right)\left(\mathbf{I}_m-\boldsymbol{U}_{1:a}{\boldsymbol{U}_{1:a}}^{\mathrm{T}}\right)\tilde{\boldsymbol{x}}_{p-l+1}^{\circ}\right|\end{aligned} \tag{14}$$

where $\mathbf{con}_{AI_Q,p}^{\circ}\in\mathbb{R}^{m\times 1}$ is a column vector with the ith element as the contribution of the ith electrical variable to AI_Q at the pth sampling time point online, $\mathbf{I}_m$ denotes the $m\times m$ identity matrix, and d denotes the derivative operator.

Meanwhile, the CVs to AI_{T^2} can be obtained by:

$$\begin{aligned}\mathbf{con}_{AI_{T^2},p}^{\circ}&=\sum_{l=1}^{L}\left|\frac{d\left({T^2}^{\circ}_{p-l+1}-{T^2}_{\mathrm{r}_2-l+L}\right)^2}{d\tilde{\boldsymbol{x}}_{p-l+1}^{\circ}}\right|\\&=\sum_{l=1}^{L}\left|2\left({T^2}^{\circ}_{p-l+1}-{T^2}_{\mathrm{r}_2-l+L}\right)\frac{d{T^2}^{\circ}_{p-l+1}}{d\tilde{\boldsymbol{x}}_{p-l+1}^{\circ}}\right|\\&=\sum_{l=1}^{L}\left|4\left({T^2}^{\circ}_{p-l+1}-{T^2}_{\mathrm{r}_2-l+L}\right)\boldsymbol{U}_{1:a}\boldsymbol{\Omega}{\boldsymbol{U}_{1:a}}^{\mathrm{T}}\tilde{\boldsymbol{x}}_{p-l+1}^{\circ}\right|\end{aligned} \tag{15}$$

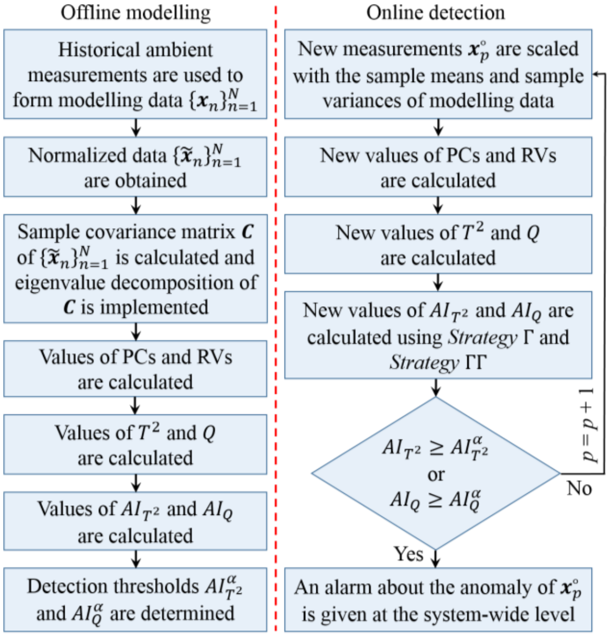
Offline modelling
Online detection
Historical ambient measurements are used to form modelling data $\{\boldsymbol{x}_n\}_{n=1}^{N}$
Normalized data $\{\tilde{\boldsymbol{x}}_n\}_{n=1}^{N}$ are obtained
Sample covariance matrix $\boldsymbol{C}$ of $\{\tilde{\boldsymbol{x}}_n\}_{n=1}^{N}$ is calculated and eigenvalue decomposition of $\boldsymbol{C}$ is implemented
Values of PCs and RVs are calculated
Values of T^2 and Q are calculated
Values of AI_{T^2} and AI_Q are calculated
Detection thresholds $AI_{T^2}^{\alpha}$ and AI_Q^{α} are determined
New measurements $\boldsymbol{x}_p^{\circ}$ are scaled with the sample means and sample variances of modelling data
New values of PCs and RVs are calculated
New values of T^2 and Q are calculated
New values of AI_{T^2} and AI_Q are calculated using *Strategy* Γ and *Strategy* ΓΓ
$AI_{T^2}\geq AI_{T^2}^{\alpha}$ or $AI_Q\geq AI_Q^{\alpha}$
No
$p=p+1$
Yes
An alarm about the anomaly of $\boldsymbol{x}_p^{\circ}$ is given at the system-wide level

Fig. 1. Disturbance detection of WAM-PCA*k*NN.

where $\mathbf{con}_{AI_{T^2},p}^{\circ}\in\mathbb{R}^{m\times 1}$ is a column vector with the ith element as the contribution of the ith electrical variable to AI_{T^2} at the pth sampling time point online.

Thus, a contribution plot strategy has been developed which quantifies the CVs to AI_Q and AI_{T^2} by (14) and (15) for identifying variables affected most by the detected disturbance. Through this strategy, a significant reference can be provided to locate the detected disturbance.

C. Parameter Settings for WAM-PCAkNN

1) Parameter k and window length L

For the parameter k, a relatively small value can better meet the real-time requirement from the online detection, because the total number of comparisons for selecting the kth smallest SED value in *strategy* ΓΓ will increase with k increasing. As stated in [14], a typical value of k is 3, which is also used in this paper. The reason why k is not set to be even smaller, e.g., $k=1$, is for avoiding excess false alarms in the normal condition. The window length L depends on specific applications. Here, L relates with the oscillation period. To be sufficient for characterizing all oscillations, L is supposed to be not smaller than the number of samples in the maximum oscillation period. This requires having an idea of typical oscillations in advance which can be achieved from the past system experience.

2) Number of PCs

To determine the number a of PCs, the Cumulative Percentage Variance (CPV) criterion, which is widely used in multivariate statistical monitoring [21], [22], is adopted. For the specific case here, it can be expressed as:

$$\mathrm{CPV}(a)=\frac{\sum_{i=1}^{a}\lambda_i}{\sum_{i=1}^{m}\lambda_i}\times 100\% \tag{16}$$

where a can be determined at the time when $\mathrm{CPV}(a)$ exceeds a certain constant. Usually, $\mathrm{CPV}(a)\geq 90\%$ is sufficient to signify that most variances of variables are captured by PCs while the remaining tiny variances are captured by RVs [23].

IV. Case Studies

In this section, WAM-PCA*k*NN is evaluated and compared with WAM-PCA in two case studies, involving data from a four-variable numerical model and the New England power system model.

A. Four-Variable Numerical Model

A four-variable numerical model which was also studied in [13] is given by:

$$x_{1,t} = 0.5s_{1,t} + 0.3s_{2,t} + 0.2s_{3,t} \quad (17)$$

$$x_{2,t} = 0.7s_{1,t} + 0.2s_{2,t} + 0.1s_{3,t} \quad (18)$$

$$x_{3,t} = 0.4s_{1,t} + 0.3s_{2,t} + 0.3s_{3,t} \quad (19)$$

$$x_{4,t} = 0.2s_{1,t} + 0.4s_{2,t} + 0.4s_{3,t} \quad (20)$$

where $s_{i,t} = \sin(2\pi f_i t)$ denotes the value of the sinusoidal signal s_i at the time of t for $i = 1,2,3,$ and $f_1 = 0.1$ Hz , $f_2 = 0.5$ Hz, $f_3 = 0.9$ Hz are the oscillation frequencies of the three sinusoidal signals s_1, s_2, s_3 respectively.

Suppose a disturbance occurs near x_1 at $t = 200$ seconds. The disturbance causes a local oscillation $s_{4,t} = \sin(2\pi f_4 t)$ with $f_4 = 1.5$ Hz, affecting x_1 much but x_2, x_3, x_4 little as:

$$x_{1,t} = 0.5s_{1,t} + 0.3s_{2,t} + 0.2s_{3,t} + 0.6s_{4,t} \quad (21)$$

$$x_{2,t} = 0.7s_{1,t} + 0.2s_{2,t} + 0.1s_{3,t} + 0.02s_{4,t} \quad (22)$$

$$x_{3,t} = 0.4s_{1,t} + 0.3s_{2,t} + 0.3s_{3,t} + 0.01s_{4,t} \quad (23)$$

$$x_{4,t} = 0.2s_{1,t} + 0.4s_{2,t} + 0.4s_{3,t} + 0.015s_{4,t} \quad (24)$$

The total simulation time is 300 seconds and data are sampled with the sampling frequency of 10 Hz. Thus, the first 2000 data points are from (17)-(20) representing the measurements under the ambient condition and the last 1000 data points are from (21)-(24) representing the measurements under the disturbance condition. The signal-to-noise ratios (SNRs) in the measurements of x_1, x_2, x_3, x_4 are 2dB, 10dB, 4dB and 2.5dB respectively. The ith SNR is calculated as

$$\mathrm{SNR}_i = 10\log_{10}\left(\sum_{n=1}^{3000} {x_{i,n}}^2 / \sum_{n=1}^{3000} {\omega_{i,n}}^2\right) \quad (25)$$

where $\omega_{i,n}$ denotes the value of the ith white noise variable ω_i at the nth sampling time point. The complete data are shown in Fig. 2. Among them, the first 1000 data points are used for modelling since they are ambient data, and the remaining 2000 data points are used for testing since they contain data subject to the disturbance effect highlighted in the rectangle of Fig. 2.

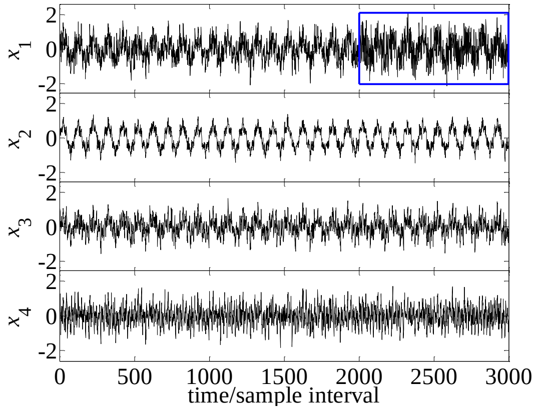

Fig. 2. The total simulation data in the first case study.

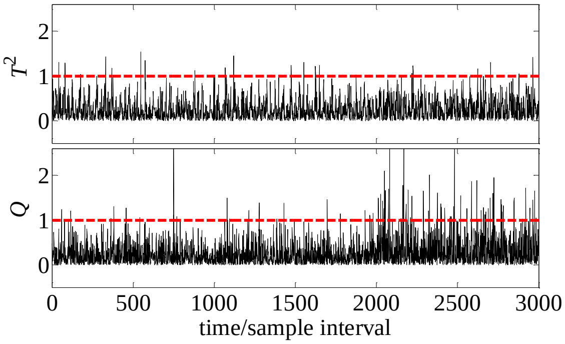

Fig. 3. The detection chart of WAM-PCA in the first case study, with values of T^2, Q as solid lines and detection thresholds as dashed lines.

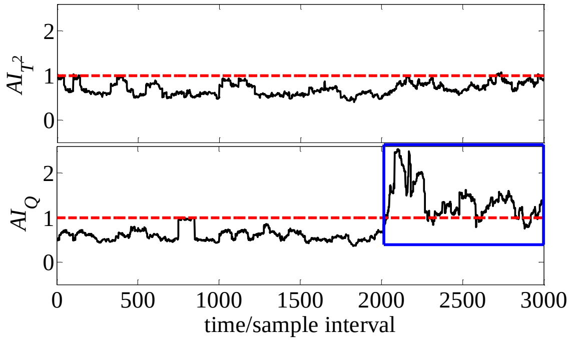

Fig. 4. The detection chart of WAM-PCA*k*NN in the first case study, with values of AI_{T^2}, AI_Q as solid lines and thresholds as dashed lines.

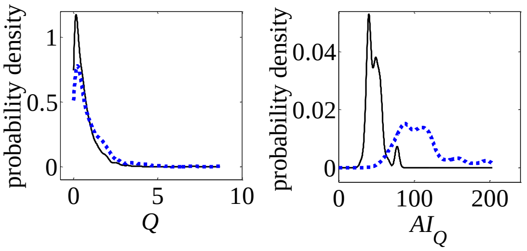

Fig. 5. The probability density values of Q and those of AI_Q before and after the disturbance in the first case study, with density values before the disturbance as solid lines and density values after the disturbance as dotted lines.

The window length L is set to 100 according to the sinusoidal component s_1 that has the maximum oscillation period with 100 samples. The number a of PCs is set to 2 by (16). The thresholds are determined with the confidence level of 99%.

The detection charts of WAM-PCA and WAM-PCA*k*NN are shown in Fig. 3 and Fig. 4 respectively. To facilitate the observation, the monitoring statistic values (solid lines) are normalized by their thresholds so that the thresholds (dashed lines) are equal to one. In Fig. 3, after the disturbance occurs at the 2000th sampling time point, most of the T^2 and Q values are still below the thresholds and thus WAM-PCA fails to detect the disturbance. By contrast, in the rectangle of Fig. 4, most of the AI_Q values significantly exceed the threshold after the disturbance and WAM-PCA*k*NN detects the disturbance at the 2026th sampling time point. The reason for such improvement of detection is that the overlap between the probability density values of AI_Q before and after the disturbance is much less than the overlap between the probability density values of Q before and after the disturbance, as illustrated in Fig. 5.

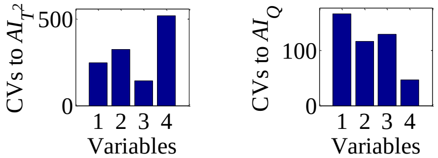

Fig. 6. The contributions of variables to AI_{T^2} and AI_Q at the detection time of AI_Q in the first case study.

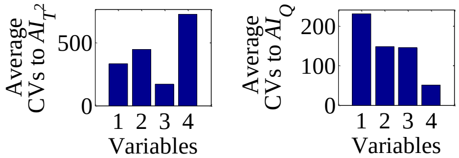

Fig. 7. The contributions of variables to AI_{T^2} and AI_Q averaged over the period from the detection time of AI_Q until the end in the first case study.

To evaluate the efficiency of the online detection, the time in calculating AI_{T^2} and AI_Q for each testing data sample is stored. The computations are carried out on an Intel(R) Core(TM) i7-4770 (3.40 GHz) with 16.0 GB RAM, Windows 7 Enterprise and MATLAB version R2014a. The maximum time in calculating AI_{T^2} and AI_Q are 0.003 seconds and 0.001 seconds respectively, small enough for the online detection in real time.

To identify the variables affected most by the detected disturbance, the contributions of variables to AI_{T^2} and AI_Q at the detection time of AI_Q and those averaged over the period from the detection time of AI_Q to the end of simulation are shown in Fig. 6 and Fig. 7 respectively. It can be seen from the right contribution plots of Figs. 6 and 7 that the contribution of x_1 to AI_Q is largest, suggesting that x_1 is the variable contributing most to the anomaly of AI_Q. This is in accordance with the fact that the disturbance occurs near x_1 and affects x_1 much more than the other variables. Thus, the contributions of variables to AI_Q provide a meaningful reference for localizing the disturbance. In comparison, the left contribution plots of Figs. 6 and 7 do not identify x_1 as the variable contributing most to AI_{T^2}. This is reasonable, since AI_{T^2} behaves normally from the detection time of AI_Q until the end of simulation and thus the largest contribution to AI_{T^2} at this time period is not supposed to come from x_1.

B. New England Power System Model

The New England power system model was described in [24] based on a single line diagram of the test system. The system is a 16-machine 68-bus system with 16 generators serving five geographical areas and eight tie lines connecting the areas to one another. The data used here were provided by authors of [24], which are 20Hz samples comprising measurements of active power (MW) and reactive power (MVAR) from the 16 generators (G1~G16) and the sending terminals of the 8 tie lines (L01, L16, L61, L62, L74, L76, L77, L86).

Two data sets were provided, one for the ambient condition (Data I) and another for the disturbance condition (Data II). According to the supplier of the data, Data I was generated by running the New England power system model normally with no disturbance, while Data II was generated by running the model with a local disturbance simulated. The disturbance is due to the step change in the voltage reference input of the automatic voltage regulator of the excitation system in G3. Four inter-area oscillations are present in both Data I and Data II, reflecting the property of the whole system. In addition, one local oscillation caused by the disturbance is present in Data II. The local oscillation is observable in the power measurements of G3 and G2 that is nearest to G3.

Table I lists the real and reactive power for the wide-area monitoring. The first 1000 samples of Data I are taken as modelling data, and the remaining 29000 samples of Data I are taken as testing data from the ambient condition. The 30000 samples of Data II are taken as testing data from the disturbance condition. To provide a compact demonstration, the normalized trends of the first 1200 samples of Data I (one-minute simulation episode) and those of Data II are shown in Fig. 8 and Fig. 9 respectively. Also for a compact demonstration, only reactive power measurements from representative generators G3, G2 (nearest to G3) and G14 (farthest from G3) rather than all power measurements are shown in Fig. 8 and Fig. 9. It can be observed from the comparison between Fig. 8 and Fig. 9 that the disturbance affects G3 and G2 much but it affects G14 little.

TABLE I
THE ACTIVE AND REACTIVE POWER FOR THE WIDE-AREA MONITORING

Variables	Description
x_1~x_{16} / x_{25}~x_{40}	Active/Reactive power of G1~G16
x_{17} / x_{41}	Active/Reactive power of sending terminal of L01
x_{18} / x_{42}	Active/Reactive power of sending terminal of L16
x_{19} / x_{43}	Active/Reactive power of sending terminal of L61
x_{20} / x_{44}	Active/Reactive power of sending terminal of L62
x_{21} / x_{45}	Active/Reactive power of sending terminal of L74
x_{22} / x_{46}	Active/Reactive power of sending terminal of L76
x_{23} / x_{47}	Active/Reactive power of sending terminal of L77
x_{24} / x_{48}	Active/Reactive power of sending terminal of L86

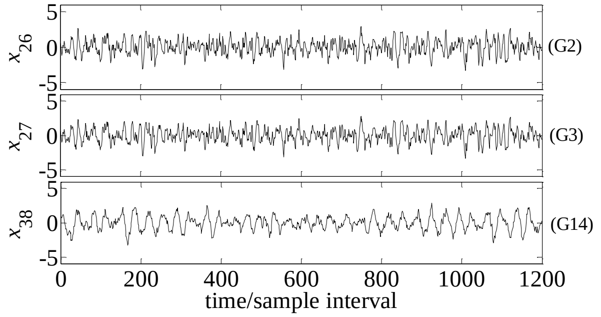

Fig. 8. The normalized trends of the first 1200 samples of Data I (reactive power measurements from G2, G3 and G14) in the second case study.

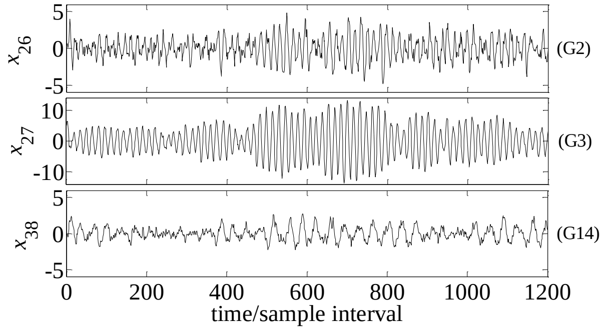

Fig. 9. The normalized trends of the first 1200 samples of Data II (reactive power measurements from G2, G3 and G14) in the second case study.

The simulated disturbance is difficult to detect and locate because the inter-area oscillatory trends in the data exhibit a masking effect on the local oscillation and thus on the disturbance. To detect and locate this disturbance can not only provide increased situational awareness of generators to the system operators but also give some reference about the time and the variables suitable for estimating the frequencies and damping ratios of different oscillations. Here, the expected detection result is that alarms should be rarely triggered for Data I whereas alarms should be constantly triggered for Data II. Besides, the expected localization result is that the active power and reactive power of G2 and G3 should be identified as the variables affected most by the detected disturbance.

Using the oscillation analysis method presented in [24] on the modelling data, it is found that the maximum oscillation period contains about 36 samples. Accordingly, the window length L is set to 36. The number a of PCs is set to 11 according to (16). The 99% confidence thresholds are also determined.

The detection charts of WAM-PCA and WAM-PCA*k*NN on Data I are shown in Fig. 10 and Fig. 11, respectively. Again, to facilitate the observation, the monitoring statistic values (solid lines) are normalized by their corresponding detection thresholds so that the thresholds (dashed lines) are equal to one. It can be observed from Fig. 10 and Fig. 11 that most of the values of T^2, Q, AI_{T^2}, AI_Q stay below their corresponding detection thresholds. These are the expected detection results, because Data I are from the ambient condition with no disturbance and few continuous alarms should be triggered for ambient data.

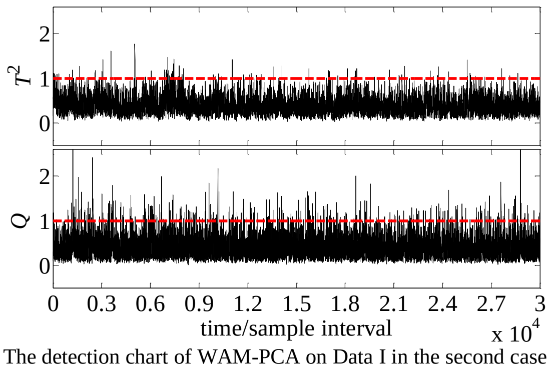

Fig. 10. The detection chart of WAM-PCA on Data I in the second case study, with values of T^2, Q as solid lines and thresholds as dashed lines.

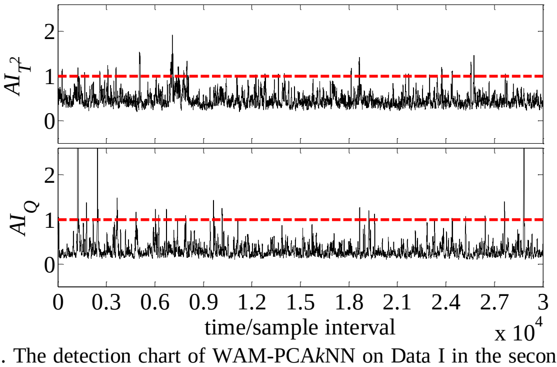

Fig. 11. The detection chart of WAM-PCA*k*NN on Data I in the second case study, with values of AI_{T^2}, AI_Q as solid lines and thresholds as dashed lines.

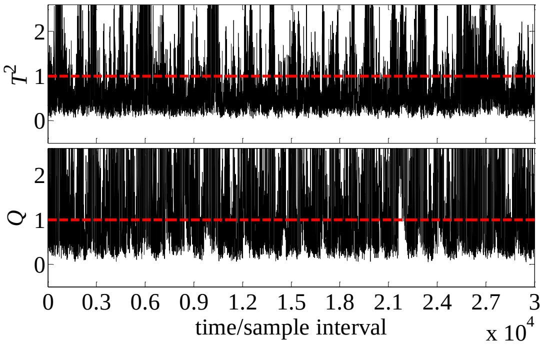

Fig. 12. The detection chart of WAM-PCA on Data II in the second case study, with values of T^2, Q as solid lines and thresholds as dashed lines.

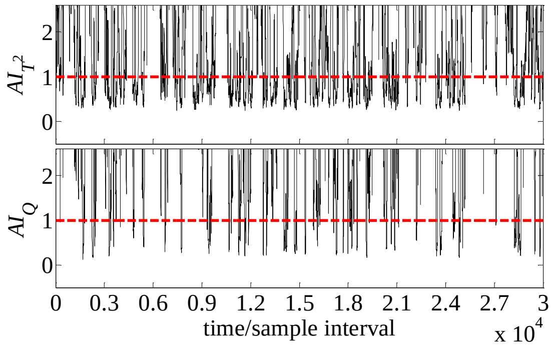

Fig. 13. The detection chart of WAM-PCA*k*NN on Data II in the second case study, with values of AI_{T^2}, AI_Q as solid lines and thresholds as dashed lines.

TABLE II
THE ALARM RATES OF WAM-PCA AND WAM-PCA*k*NN ON DATA I AND DATA II

Method / Data	WAM-PCA		WAM-PCA*k*NN	
	T^2	Q	AI_{T^2}	AI_Q
I	0.55%	2.20%	1.83%	1.64%
II	26.63%	62.60%	69.47%	91.45%

The detection charts of WAM-PCA and WAM-PCA*k*NN on Data II are shown in Fig. 12 and Fig. 13, respectively. Observing from Fig. 12 and Fig. 13, a large number of the values of T^2, Q, AI_{T^2}, AI_Q exceed their related detection thresholds. These are also the expected detection results, because Data II is from the disturbance condition and alarms are supposed to be constantly triggered for data subject to the disturbance effect. Moreover, by comparing the T^2 chart of Fig. 12 with the AI_{T^2} chart of Fig. 13, it can be found that the number of the AI_{T^2} values exceeding the detection threshold is much larger than the number of the T^2 values exceeding the threshold. Similarly, by comparing the Q chart of Fig. 12 with the AI_Q chart of Fig. 13, it can also be found that the number of the AI_Q values exceeding the detection threshold is much larger than the number of the Q values exceeding the threshold.

To quantify the results observed in Fig. 10 and Fig. 11 as well as the results observed in Fig. 12 and Fig. 13, Table II lists the alarm rates of WAM-PCA and WAM-PCA*k*NN on Data I and Data II. An alarm rate is calculated as the ratio percentage of the number of the triggered alarms over the dataset size. Taking the value ‘‘69.47%’’ in Table II as an example, it is obtained by dividing the number of the triggered alarms (that is, the number of the AI_{T^2} values exceeding the detection threshold in the AI_{T^2} chart of Fig. 13) by the dataset size which is

30000. It can be seen from Table II that almost all of the alarm rates calculated on Data I except for the alarm rate of Q are smaller than 2% and acceptable against the given confidence level 99%. Besides, by comparing the alarm rates calculated on Data II, AI_{T_2} achieves a much higher alarm rate than T^2, and AI_Q achieves a much higher alarm rate than Q.

The detection results in Fig. 10 and Fig. 11, Fig. 12 and Fig. 13, and Table II demonstrate that WAM-PCA*k*NN can significantly enhance the sensitivity of WAM-PCA in detecting disturbances by reducing the masking effect of the oscillatory trends on disturbances while behaving reliably in the ambient condition by triggering the appropriate quantity of false alarms acceptable against the given confidence level. Moreover, WAM-PCA*k*NN is suitable for the online detection in real time, since the maximum time on the calculation of AI_{T^2} and AI_Q is 0.003 seconds and 0.0021 seconds respectively, far smaller than the sampling interval of 0.05 seconds. Based on the above results and analysis, WAM-PCA*k*NN has a good potential for practical application, because its AI_{T^2} and AI_Q are system-wide monitoring statistics and the practical implementation of disturbance detection in a control room usually takes the form as a real-time traffic light with green or red indicators for the overall state of power systems.

After the disturbance is detected by AI_{T^2} and AI_Q, next is to identify the variables affected most by the detected disturbance so that a meaningful reference can be provided to locate the detected disturbance. The contributions of variables to AI_{T^2} and AI_Q at the 500th sampling time point of Data II and those averaged over a whole time period of Data II are shown in Fig. 14 and Fig. 15, respectively. It can be observed from both Fig. 14 and Fig. 15 that the 2nd, 3rd, 26th and 27th variables which are the active and reactive power of the generators G2 and G3 contribute most to the anomaly of AI_{T^2} and AI_Q and they are identified as the ones affected most by the detected disturbance. By observing the trends of reactive power measurements from the generators G2 and G3 in Fig. 8 and Fig. 9, it can be found that new oscillations arise in the disturbance condition, making the trends quite different from those in the ambient condition. Thus, the contribution plots in Fig. 14 and Fig. 15 correctly identify the variables affected most by the detected disturbance. This identification result can provide a meaningful reference for finally locating the detected disturbance.

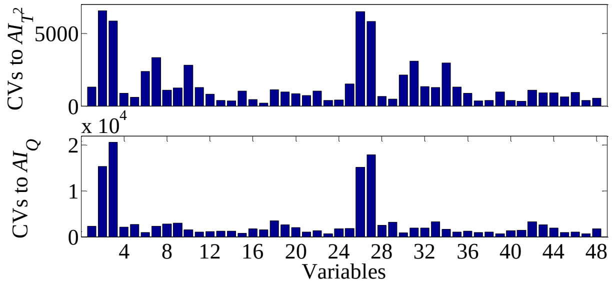

Fig. 14. The contributions of variables to AI_{T^2} and AI_Q at the 500th sampling time point of Data II in the second case study.

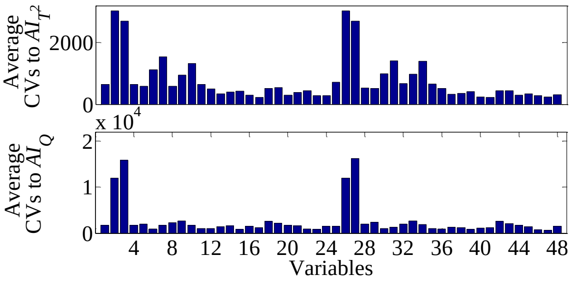

Fig. 15. The contributions of variables to AI_{T^2} and AI_Q averaged over the whole period of Data II in the second case study.

A point worth mentioning is that the ordinates in Figs. 8-15 and the abscissas in Figs. 8-13 have no unit, because Figs. 8-15 show the results obtained based on the normalized power measurements and the abscissas in Figs. 8-13 represent the sequence number of sampling with the interval of 0.05 seconds.

V. Discussions

An issue that can affect the performance of the proposed method in detecting and locating disturbances is limited sensor coverage, which has already been pointed out in [25] and [26]. Due to limited sensor coverage, the measured data may contain little disturbance information, making it difficult to detect and locate disturbances. Fortunately, this issue has been much relieved by the widespread PMUs. Besides, a feasible solution to this issue, as suggested in [25], is optimal sensor placement.

Another issue is the automatic update of a previously built detection model for new ambient conditions. To automatically identify the time when power systems enter new ambient conditions rather than relying on the experience of the system operators is a solution to this issue worth considering.

These two issues are outside the scope of the present work, but will make interesting topics for future study.

VI. Conclusion

A wide-area monitoring method (WAM-PCA*k*NN) has been proposed by combining Principal Component Analysis (PCA) with *k*-Nearest Neighbor (*k*NN) analysis to detect and locate power system disturbances in real time. The contribution is three-fold. Firstly, *k*NN has been combined with multivariate analysis PCA to build new system-wide monitoring statistics AI_{T^2} and AI_Q, which can not only monitor a large number of variables for the detection of disturbances but also reduce the masking effect of the oscillatory trends and noise in electrical measurements on disturbances. Secondly, the application of *k*NN has been extended from the offline analysis to the online analysis by developing *Strategy* Γ for recursively calculating the Square of Euclidean Distance (SED) and *Strategy* ΓΓ for fast selection of the *k*th smallest SED value. With the online *k*NN analysis, the real-time detection of disturbances can be achieved. Thirdly, a contribution plot strategy quantifying the contributions of variables to the anomaly of AI_{T^2} and AI_Q has been developed, which can identify the variables affected most by the detected disturbance and thus can provide a significant reference for finally locating the detected disturbance.

The analysis on the data from a four-variable numerical model and the New England power system model has illustrated that WAM-PCA*k*NN significantly improves the performance of the traditional wide-area monitoring method based on PCA (WAM-PCA) in detecting disturbances, e.g., AI_{T^2} of WAM-PCA*k*NN achieves the alarm rate of 69.47% under the disturbance condition whereas T^2 of WAM-PCA only achieves the alarm rate of 26.63%. Moreover, WAM-PCA*k*NN correctly identifies the variables affected most by the detected disturbance for guiding disturbance localization through the developed contribution plot strategy.

Acknowledgment

The authors gratefully acknowledge Dr. Karine Hay and Dr. Alex Golder for providing the simulated data of the New England power system model to support this paper.

References

[1] O. Samuelsson, M. Hemmingsson, A. H. Nielsen, K. O. H. Pederson, and J. Ramaussen, "Monitoring of power system events at transmission and distribution level," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 1007-1008, May 2006.

[2] CIGRE Technical Brochure 316, "Defense plan against extreme contingencies," Task Force C2.02.24, pp. 58-70, Apr. 2007.

[3] F. B. Costa, "Fault-induced transient detection based on real-time analysis of the wavelet coefficient energy," *IEEE Trans. Power Del.*, vol. 29, no. 1, pp. 140-153, Feb. 2014.

[4] Q. Huang, L. Shao, and N. Li, "Dynamic detection of transmission line outages using hidden Markov models," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 2026-2033, May 2016.

[5] N. Senroy, S. Suryanarayanan, and P. F. Ribeiro, "An improved Hilbert–Huang method for analysis of time-varying waveforms in power quality," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1843-1850, Nov. 2007.

[6] A. Farzanehrafat and N. R. Watson, "Power quality state estimator for smart distribution grids," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 2183-2191, Aug. 2013.

[7] E. Barocio, B. C. Pal, D. Fabozzi, and N. F. Thornhill, "Detection and visualization of power system disturbances using principal component analysis," in *Proc. 2013 IREP Symp. Bulk Power Syst. Dyn. Control IX*, Rethymnon, Greece, Aug. 2013, pp. 1-10.

[8] X. Liu *et al*., "Principal component analysis of wide-area phasor measurements for islanding detection—A geometric view," *IEEE Trans. Power Del.*, vol. 30, no. 2, pp. 976-985, Apr. 2015.

[9] M. Rafferty, X. Liu, D. M. Laverty, and S. McLoone, "Real-time multiple event detection and classification using moving window PCA," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2537-2548, Sep. 2016.

[10] L. Xie, Y. Chen, and P. R. Kumar, "Dimensionality reduction of synchrophasor data for early event detection: linearized analysis," *IEEE Trans. Power Syst.*, vol. 29, no. 6, pp. 2784-2794, Nov. 2014.

[11] E. L. Russell, L. H. Chiang, and R. D. Braatz, *Data-Driven Methods for Fault Detection and Diagnosis in Chemical Processes*. London, U.K.: Springer, 2012.

[12] J. J. Ayon, E. Barocio, and A. R. Messina, "Blind extraction and characterization of power system oscillatory modes," *Electr. Power Syst. Res*., vol. 119, pp. 54-65, Feb. 2015.

[13] J. Thambirajah, N. F. Thornhill, and B. C. Pal, "A multivariate approach towards interarea oscillation damping estimation under ambient conditions via independent component analysis and random decrement," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 315-322, Feb. 2011.

[14] I. M. Cecílio, J. R. Ottewill, H. Fretheim, and N. F. Thornhill, "Multivariate detection of transient disturbances for uni- and multirate systems," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 4, pp. 1477-1493, Jul. 2015.

[15] E. Keogh, J. Lin, and A. Fu, "HOT SAX: Efficiently finding the most unusual time series subsequence," in *Proc. 5th IEEE Int. Conf. Data Mining*, Houston, TX, USA, Nov. 2005, pp. 226-233.

[16] M. C. Chuah and F. Fu, "ECG anomaly detection via time series analysis," in *Proc. ISPA Int. Workshops Frontiers High Perform. Comput. Netw*., 2007, pp. 123–135.

[17] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Comput. Surv*., vol. 41, no. 3, pp. 1-58, Jul. 2009.

[18] G. O. Campos *et al.*, "On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study," *Data Mining Knowl. Discovery*, vol. 30, no.4, pp. 891-927, Jul. 2016.

[19] L. Cai, N. F. Thornhill, S. Kuenzel, and B. C. Pal, "Real-time detection of power system disturbances based on k-nearest neighbor analysis," *IEEE Access*, vol. 5, pp. 5631-5639, Mar. 2017.

[20] L. Cai, X. Tian, and S. Chen, "Monitoring nonlinear and non-Gaussian processes using Gaussian mixture model-based weighted kernel independent component analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 1, pp. 122-135, Jan. 2017.

[21] J. Valenzuela, J. Wang, and N. Bissinger, "Real-time intrusion detection in power system operations," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1052-1062, May 2013.

[22] W. Zhu *et al.*, "A novel KICA–PCA fault detection model for condition process of hydroelectric generating unit," *Measurement*, vol. 58, pp. 197-206, Dec. 2014.

[23] G. Li, S. Qin, and D. Zhou, "A new method of dynamic latent-variable modeling for process monitoring," *IEEE Trans. Ind. Electron.*, vol. 61, no.11, pp. 6438-6445, Nov. 2014.

[24] D. H. Wilson, K. Hay, and G. J. Rogers, "Dynamic model verification using a continuous modal parameter estimator," in *Proc. IEEE PowerTech Conference*, Bologna, Italy, June 2003, pp. 1-6.

[25] I. R. Cabrera, E. Barocio, R. J. Betancourt, and A. R. Messina, "A semi-distributed energy-based framework for the analysis and visualization of power system disturbances," *Electr. Power Syst. Res.*, vol. 143, pp. 339-346, Feb. 2017.

[26] P. Bhui and N. Senroy, "Application of recurrence quantification analysis to power system dynamic studies," *IEEE Trans. Power Syst.*, vol. 31, no. 1, pp. 581-591, Jan. 2016.

Lianfang Cai received the B.Eng. and Ph.D. degrees from the China University of Petroleum, Qingdao, China, in 2009 and 2014, respectively. Presently he is a Postdoctoral Research Associate at Imperial College London, U.K. His research interests include data-driven power system monitoring, simulation of power systems with energy storage, multivariate statistical modelling and data analysis.

Nina F. Thornhill (SM'93) received the B.A. degree in physics from Oxford University, Oxford, U.K., in 1976, the M.Sc. degree from Imperial College London, London, U.K., and the Ph.D. degree from University College London.

She is Professor of Process Automation in the Department of Chemical Engineering at Imperial College London where she holds the ABB Chair of Process Automation.

Stefanie Kuenzel (GS'11, M'14) received the M.Eng. and Ph.D. degrees from Imperial College London, London, U.K., in 2010 and 2014, respectively. Presently she is the Head of the Power Systems Group and Lecturer in the Department of Electronic Engineering at Royal Holloway, University of London and a visiting researcher at Imperial College London. Her current research interests include renewable generation and transmission, including HVDC.

Bikash C. Pal (M'00-SM'02-F'13) is Professor of Power Systems at Imperial College London. He is research active in power system stability, control and computation. Prof. Pal has graduated 20 PhDs and published 88 technical papers in IEEE Transactions and IET journals. He has co-authored two books and two awards winning IEEE Task Force/Working Group reports.

He is Editor-in-Chief of IEEE Transactions on Sustainable Energy and Fellow of IEEE for his contribution to power system stability and control.