Detecting outliers in streaming time series data from ARM distributed sensors

Yuping Lu
University of Tennessee
Knoxville, TN, USA
yupinglu89@gmail.com

Nathan Collier
Oak Ridge National Laboratory
Oak Ridge, TN, USA
nathaniel.collier@gmail.com

Michael A. Langston
University of Tennessee
Knoxville, TN, USA
langston@tennessee.edu

Jitendra Kumar
Oak Ridge National Laboratory
Oak Ridge, TN, USA
jkumar@climatemodeling.org

***Abstract*—The Atmospheric Radiation Measurement (ARM) Data Center at ORNL collects data from a number of permanent and mobile facilities around the globe. The data is then ingested to create high level scientific products. High frequency streaming measurements from sensors and radar instruments at ARM sites requires high degree of accuracy to enable rigorous study of atmospheric processes. Outliers in collected data are common, however, due to instrument failure or extreme weather events. Thus, it is critical to identify and flag them. We employed multiple univariate, multivariate and time series techniques for outlier detection methods and studied their effectiveness. First, we examined Pearson correlation coefficient which is used to measure the pairwise correlations between variables. Singular Spectrum Analysis (SSA) was applied to detect outliers by removing the anticipated annual and seasonal cycles from the signal to accentuate anomalies. K-means was applied for multivariate examination of data from collection of sensor to identify any deviation from expected and know patterns and identify abnormal observation. The Pearson correlation coefficient, SSA and K-means methods were later combined together in a framework to detect outliers though a range of checks. We applied the developed method to data from meteorological sensors at ARM Southern Great Plains site and validated against existing database of known data quality issues.**

***Index Terms*—outlier detection, time series, clustering, atmospheric science**

I. Introduction

The Atmospheric Radiation Measurement (ARM) user facility was founded by the U.S. Department of Energy (DOE) in 1989 [1]. Since then, its aim is to be the platform for the observation and study of Earth's climate. ARM facility collects large volume of datasets from instruments deployed in different ground stations across the globe [2]. The ARM Data Center (ADC) is responsible for ingesting the collected data and creating high level scientific data products for distribution and dissemination to scientific research community, especially to inform and improve the representation of atmospheric, cloud and aerosols processes in global climate models (GCMs) [3]. They also develop a large number of high level data products, also called "Value Added Products" (VAPs), quality of which are highly dependent on the correctness of the raw data. Data are transferred from individual site to ADC in a streaming near-real-time fashion and the raw data is ingested, processed to produce VAPs and made available to users via a web-based data discovery interface with a lag time of less than an hour. Along with expediency, its also essential to maintain high quality of data and identify, address, and communicate any noise and outliers in the data. Thus an effective and efficient outlier and noise detection is crucial for ARM to provide scientific users with high quality data for research.

Outlier detection, also called anomaly detection or intrusion detection, is a common task in many application domains that include time series data, streaming data, distributed data, spatio-temporal data, and network data [4]. Temporal data is a broad concept that includes commercial transactions, sensor data, astronomy data, computer network traffic, medical records, judicial records, social network data and many others. Common techniques for outlier detection include signal processing, classification, clustering, nearest neighbor, density, statistical, information theory, spectral decomposition, and visualization. Among all these techniques, time series data outlier detection and temporal network outlier detection are especially useful for ARM data.

Outlier detection in time series data was first studied by Fox in 1972 [5]. Common types of outliers are additive outliers, level shifts, temporary changes, and innovative outliers. One common approach is the discriminative method which is based on a similarity function. For example, the normalized longest common subsequence (NLCS) is a similarity measurement widely used in the field of data mining [6]–[8]. Commonly used clustering methods such as K-means [9], dynamic clustering [8], single-linkage clustering [10], principal component analysis (PCA) [11], and self-organizing map (SOM) [12] are also popular. The choice of the clustering algorithm depends on the problem itself as each has different size and complexity. Three unsupervised parametric models, finite state automata (FSA), Markov models, and Hidden Markov Models (HMMs), are often seen in outlier detection as well. An outlier is detected if the FSA in the current state could not reach the final state [7]. The history size in the Markov model could be either fixed or flexible. HMMs are easy to interpret but not function well with big datasets [7]. Researchers also tried supervised methods such as neural networks [13], support vector machines (SVMs) [14], and decision trees [15] to detect outliers.

Different from the methods mentioned above, window-based detection is breaking the time series data into overlapping subsequences with fixed window size [16]. Each window is assigned an anomaly score, and then a final score for the

times series data is calculated by aggregating the window scores. Subspace based analysis for univariate time series data is similar to window-based detection. The subspace based transformation is to convert a univariate time series into a multivariate time series with fixed window size. It then transforms the multivariate time series back to univariate time series. Singular Spectrum Analysis is a good example of this idea [17].

ARM data also belongs to the class of temporal data as we can sequentially create a time series of network changes or graph snapshots at different periods. Each period forms a graph snapshot using various graph distance metrics from a set of nodes. Many challenges exist for outlier detection for temporal data. First, the algorithm or model needs to be chosen carefully as the properties of each data and network are different. Second, the temporal data has space and time dimensions which make it complex to analysis. Third, its scale is massive, and efficient algorithm is crucial for fast outlier detection. One common problem for temporal data is to detect outlier graph snapshots from a series graph snapshots in temporal networks. Spearman's correlation coefficient is a good candidate for such problem. It is the rank correlation between two sorted lists of graph vertices which are ordered by PageRank or other properties [18]. Similar to Spearman's correlation coefficient, Pearson correlation coefficient and mutual information are also commonly used. Jaccard similarity is the size of intersection vertex set divided by the size of union vertex set [19]. Graph edit distance describes the necessary changes to make graph G_1 isomorphic to graph G_2. It can defined as $d(G_1, G_2) = |V_{G_1}| + |V_{G_2}| - 2|V_{G_1} \cap V_{G_2}| + |E_{G_1}| + |E_{G_2}| - 2|E_{G_1} \cap E_{G_2}|$ [18]. The spectral distance is the difference between the adjacency spectrum of graph G_1 and G_2, written as $\sigma(G_1, G_2) = \sum_{i=1}^{n} |\lambda_i(G_1) - \lambda_i(G_2)|$ [20]. Entropy distance is defined by the entropy-like measurement between two graphs [21]. All these metrics are also common seen in temporal network outlier detection.

A number of approaches have been developed in literature for temporal outlier detection, especially for environmental sensor data. Birant et al. [22] discovered that locations with high wave heights are outliers while studying the wave height values from the east of the Mediterranean Sea, the Marmara Sea, the Black sea, and the Aegean Sea. Hill et al. [23], [24] filtered out measurement errors in the wind speed data stream from Water and Environmental Research Systems (WATERS) Network Corpus Christi Bay testbed with dynamic Bayesian networks. Drosdowsky et al. [25] found anomalies from Australian district rainfall using rotated PCA. Wu et al. [26] detected precipitation outlier events while working on South American precipitation data set. Sun et al. [27] extracted locations which always have different temperature from their surroundings by exploring the South China area dataset from 1992 to 2002.

II. Datasets

ARM data are stored and distributed in the Network Common Data Form (NetCDF) format which is self-describing and machine-independent [28], [29] and has good performance and data compression. It is commonly used to handle scientific data, especially those from the climatology, meteorology, oceanography and GIS projects. All ARM data are publicly available and can be downloaded from ARM Data Archive (http://www.archive.arm.gov) where a large range of datasets ranging from meteorology, to atmospheric profiles, to weather radars to satellite observations are available. Datasets are collected at a number of different locations using large number of diverse instruments.

TABLE I
SGPMET DATASETS USED IN THIS STUDY

Instrument	E1	E3	E4	E5	E6	E7
Begin Year	1996	1997	1996	1997	1997	1996
End Year	2008	2008	2010	2008	2010	2011
Instrument	E8	E9	E11	E13	E15	E20
Begin Year	1994	1994	1996	1994	1994	1994
End Year	2008	2017	2017	2017	2017	2010
Instrument	E21	E24	E25	E27	E31	E32
Begin Year	2000	1996	1997	2004	2012	2012
End Year	2017	2008	2001	2009	2017	2017
Instrument	E33	E34	E35	E36	E37	E38
Begin Year	2012	2012	2012	2012	2012	2012
End Year	2017	2017	2017	2017	2017	2017

In this study, we used the data from Surface Meteorology Systems (MET) collected at the ARM Southern Great Plains (SGP) site in Oklohoma, United States. SGP is ARM's largest and central facility and comprises of a network of core and extended facilities. In our study we used MET data from 24 extended facilities where surface meterological observations have been collected continuously and independently. While MET instruments collect a large array of direct and indirect measurements, we focused on our analysis on five core meterological variables: air temperature (*temp_mean*), vapor pressure (*vapor_pressure_mean*), atmospheric pressure (*atmos_pressure*), relative humidity (*rh_mean*) and wind speed *wspd_arith_mean*. These five core meteorological variables are inputs for a large number of derived datasets produced by the ARM and are often essential set of data for most atmospheric analysis, hence focus of our study. Table I provides details of sites and available time series for the datasets used.

III. Methodology

From the many outlier detection methods introduced in the first section, we carefully selected Pearson correlation coefficient, Singular Spectrum Analysis and K-means and tested them on ARM data.

The various algorithms required differing levels of preprocessing. The first level raw data was stored in minute level. It was normalized for pairwise comparison algorithm. Some algorithms might not need so much detail information to extract outliers. Thus we created a second level data by averaging the 1440 minute data points into one day point

from the raw data. The second level data could save a lot of running time and is easier for Plotly [30] and Matplotlib [31] to visualize. The third level data was especially created for multivariate method by standardization all the 5 variables into the same scale based on the second level data. Below we will talk about each algorithm in detail.

A. Pearson correlation coefficient

Co-located meteorological variables measure different aspect of the atmospheric conditions at any location, and driven by atmospheric physics are inherently correlated with each others. Any atmospheric phenomena at the location would affect all variables in an expected and correlated fashion. Analysis of historical time series data would provide us the baseline correlation structure and patterns for the location. In addition, any abrupt change or break in correlation structure among meteorological behavior can be a sign of sensor malfunction and should be identified a outlier.

The Pearson correlation coefficient was first introduced by Karl Pearson [32] and can be used to measure the linear correlation between two variables. The Pearson correlation coefficient is calculated from the covariance of two variables divided by the multiplication of the standard deviation of those two variables. This normalization results in a value between [-1, 1]. If the value is close to -1, it means those two variables are highly negatively related. On the other hand, then the two variables are strongly positively related. If the value is near 0, it means those two variables do not have linear relation.

We performed a pairwise comparison of the five variables using Pearson correlation using data from all 24 extended sites. Atmospheric dynamics are strongly driven by seasons and the correlation patterns among meteorological variables can have season specific patterns. We performed our analysis seasonally by separating the data among Winter, Spring, Summer and Fall seasons. Figure 1 show the distribution of pairwise correlation for Spring season. All variables show strong correlations which are normally distributed. The long tails of the distribution are potentially due to outlier data points. For example, the Pearson correlation between *temp_mean* and *vapor_pressure_mean* is positively correlated with correlation mean close to 0.75. And the Pearson correlation between *atmos_pressure* and *temp_mean* is negatively correlated with correlation mean close to -0.60. These highly correlated Pearson correlation coefficients are stored as the expected values between two variables. We then compare each Pearson correlation of two variables from a specific season in a specific year from a specific instrument individually. If this pairwise Pearson correlation of two variables deviates far away from our expected historical correlation, we treat it as an outlier. This method would allow to check incoming datastream on near-real-time basis to identify outliers.

B. Singular Spectrum Analysis

Univariate time series analysis of meteorological variables can be applied to identify any unexpected variability and extreme values observed by the instruments. These anamolous observations can be indicative of extreme atmospheric events at the site and are important to identify. However, a range of natural inter- and intra-annual variability in meterological times series is also expected and it's important to not erroneously flag them as outliers. We applied Singular Spectrum Analysis for time series of analysis of meteorological observations to identify exteme events.

Singular Spectrum Analysis (SSA) is a popular method for time series data analysis [17], [33]. The general idea is to use a subset of the decomposition of trajectory matrix to approximate the original data. Many applications can be found in [17]. For example, SSA can be applied to monitor volcanic activity [34]. It can also be used to extract trend [35]. Different from the classic SSA method, we defined our own version of SSA which is designed to remove any number of modes of specified periodicity from the time series. This is meant to remove known seasonalities from the data in order to isolate true anomalous values more accurately. We provided a schematic of the algorithm used in Figure 2 and a sample application in Figure 3.

Assume we have an ARM time series data Y of length T

$$Y = (y_1,\ \ldots,\ y_T)$$

where $T > 2$ and y_i is not empty. Let L $(1 < L \leq T/2)$ be the window size and $K = T - L + 1$. In general, the algorithm contains two main parts: decomposition and reconstruction. The first step is to form the trajectory matrix $\mathbf{X}$ from vector Y by embedding subsets of Y. These subsets of Y X_i are lagged vectors of length L.

$$X_i = (y_i,\ \ldots,\ y_{L+i-1})^T \quad (1 \leq i \leq K)$$
$$\mathbf{X} = [X_1,\ \ldots,\ X_K]$$

Thus the trajectory matrix is

$$\mathbf{X} = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ y_3 & y_4 & y_5 & \cdots & y_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_T \end{pmatrix} \quad (1)$$

where $x_{ij} = y_{i+j-1}$. We can see from equation 1 that matrix $\mathbf{X}$ has equal elements on anti-diagonals and therefore it is a Hankel matrix. Then we perform the singular value decomposition (SVD) on $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ where the eigenvalues of S are denoted by $\lambda_1, \ldots, \lambda_L$ in the decreasing order of magnitude $(\lambda_1 \geq \ldots \geq \lambda_L \geq 0)$ and the corresponding eigenvectors by $P_1, \ldots, P_L$. Let $d = rank\ \mathbf{X}$ and $V_i = \mathbf{X}^T P_i / \sqrt{\lambda_i} (i = 1, \ldots, d)$. Thus, the trajectory matrix X can then be written by its eigendecomposition,

$$\mathbf{X} = \mathbf{X_1} + \ldots + \mathbf{X_d} \quad (2)$$

where $\mathbf{X_i} = \sqrt{\lambda_i} P_i V_i^T$.

Next we choose a subset of eigenpairs to form an approximation of the trajectory matrix. It is at this point that our version of the algorithm differs. Given that the time series we are studying has seasonality at known frequencies, we use

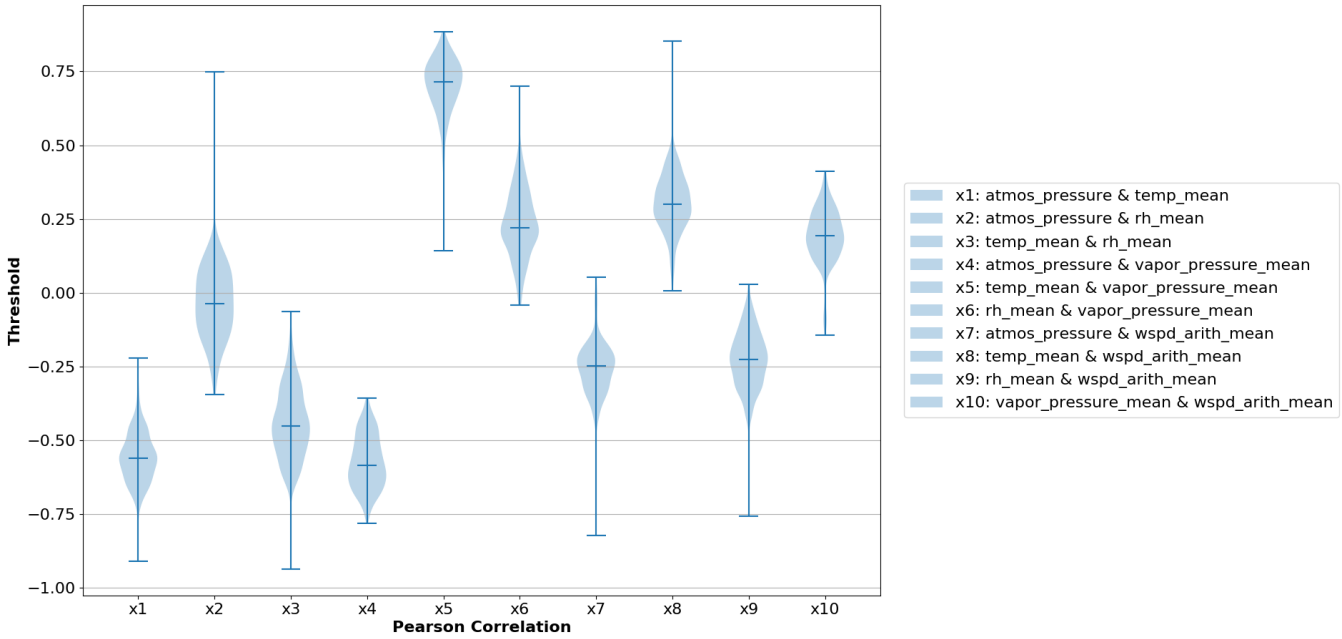

Fig. 1. Correlation patterns for five meteorological variables during spring season.

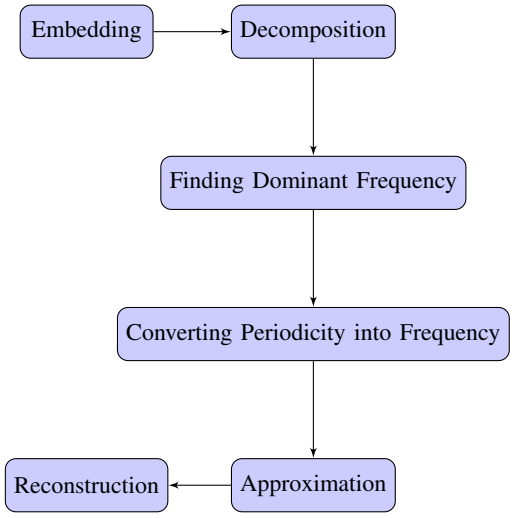

Fig. 2. Flowchart of SSA

Fast Fourier transform (FFT) to find the dominant frequency of each eigenvector [36]. We then approximate the trajectory matrix by including modes which match the frequencies of the seasonality we wish to remove. For example, we anticipate that the temperature data will have a annual and possibly monthly cycle, as shown in Figure 3. SSA allows us to tease out these contributions in additive fashion. In this example, the signals from the year, month, and residual sum together to form the original raw data. This residual is then the noise in the raw data with the seasonality removed as doing so exposes large anomalies which are possible outliers.

Once the eigenpairs are chosen, we proceed with the classical definition of the method. If I represents a set of indices corresponding to the eigenmodes to remove, we approximate the trajectory matrix

$$\mathbf{Xt} = \sum_{i \in I} \mathbf{X_i}$$

An approximation Yt to the original signal Y can be obtained from $\mathbf{Xt}$ by inverting the process used to form the trajectory matrix, Equation (1). Each column of $\mathbf{Xt}$ represents a shifted approximation to Yt, thus we average each shifted column. Finally the deseasonalized residual is the difference between the original signal and the reconstruction, $R = Y - Yt$.

In this paper, we chose the *temp_mean* data from instrument E33 as Y to illustrate SSA. Because SSA requires the time series data to be continuous, we replaced the empty points with the average temp_mean value for that day in a year. We set $L = 400$ and picked a single year and month as the periodicity groups. Thus $Yt = Yt[0] + Yt[1] + Yt[2]$. Figure 3 is a visualization of the result. The first row is the raw data Y. The orange line $Yt[0]$ is the trend. As we can see from the figure, the trend is pretty flat from 2012 to 2017. The second row and third row are $Yt[1]$, $Yt[2]$ respectively. The Year data matches the pattern of the raw data. The last row is the residual. Those peak values outside the blue shaded area are outliers.

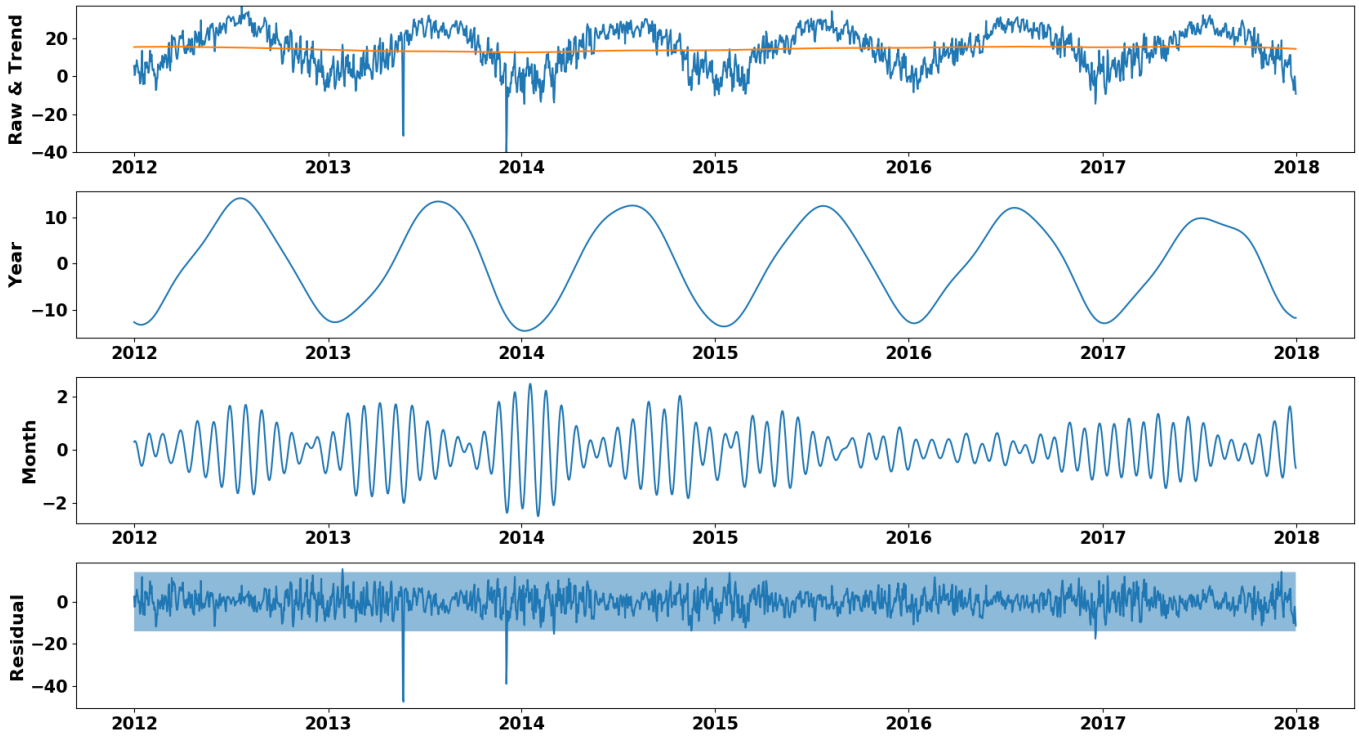

Fig. 3. Example of SSA application on ARM data. The full decomposition of *temp_mean* data from instrument E33.

C. K-means

K-means is a partitioning clustering algorithm [9], [37]. It starts with the k centroids user specified, and assigns the points to the nearest centroid. Then it computes new k centroids and assign the rest points to these centroids again. The process repeats until it converges.

Algorithm 1: K-means Outlier Detection

Input : ARM time series data
Output: Outliers

```
outliers ← ∅
df ← ARM time series data
data ← df['atmos_pressure','temp_mean',
  'rh_mean','vapor_pressure_mean','wspd_arith_mean']
number_of_clusters ← 4
clusters ← K-means(data, number_of_clusters)
distances ← Distance between each point and its centroid
mean ← arithmetic mean of distances
sigma ← standard deviation of distances
threshold ← mean + 3 * sigma
for i in range(size of distances) do
    if distances[i] > threshold then
        outliers ← outliers ∪ distances[i]
    end
end
return outliers
```

In this paper, we did not stop after clustering ARM data with K-means. We transformed the generated clusters into a vector of distance between each point and its corresponding centroid. If the distance of a point is larger than the threshold, this point will be filtered out as an outlier. Algorithm 1 describes the whole process. Unlike SSA, we used all the 5 variables mentioned in Datasets section together to extract outliers.

Again, we used data from instrument E33 as an example for K-means. Here we set k to 4 as each year has 4 seasons. Figure 4 visualized the outliers detected from E33. Y axes in this figure is the distance metric. Those red squares which are far away from the centroids are outliers.

IV. Results and Discussion

The three algorithms and visualizations are implemented in Python in this paper. All codes and results are available on GitHub (https://github.com/YupingLu/arm-pearson and https://github.com/YupingLu/arm-ssa). Multiple methods are available to set a threshold for extreme values as outliers. We used the three sigma rule to extract outliers [38]. For example, if the distance of one point is larger than three sigmas, we treat this point as an outlier in Algorithm 1.

Pearson correlation coefficient is a pairwise comparison method which is used to detect abnormality of correlation between two variables. However if the two variables suddenly change in the same direction, their correlation may still be normal similar to their "supposed" value. It is the same case if only a few outlier points inside a big quantity of data points. As we performed the Pearson correlation coefficient on the seasonal level, it is not possible to track down to the exact

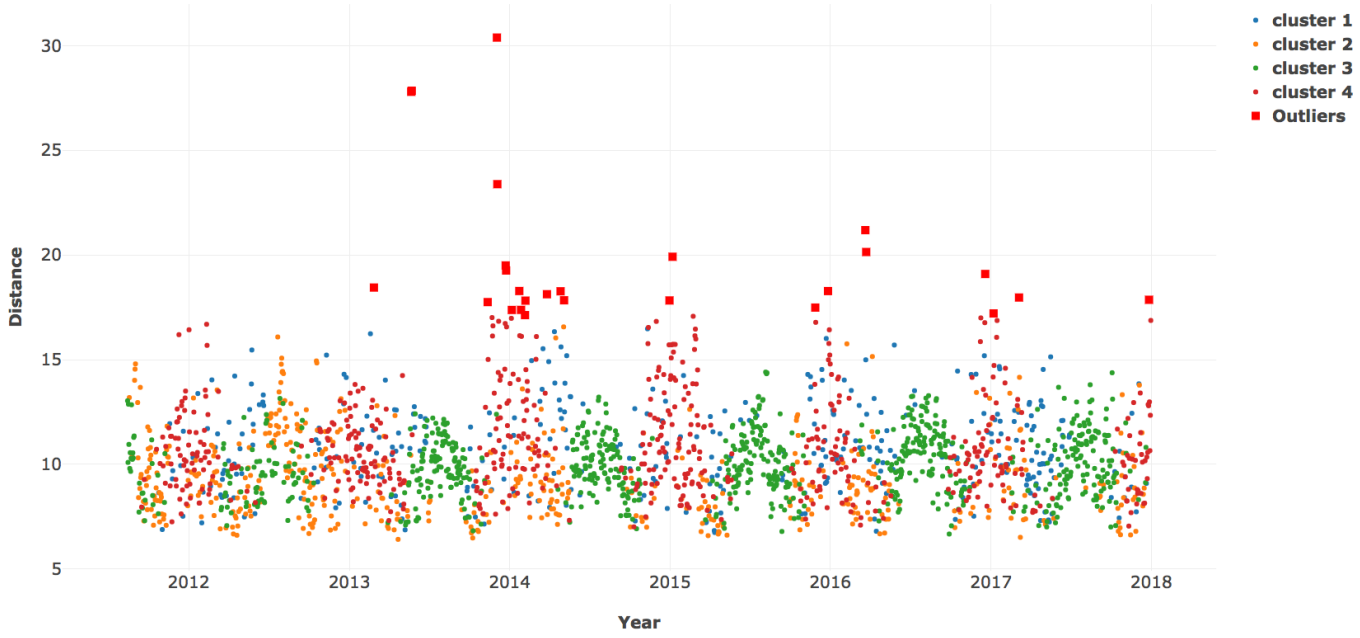

Fig. 4. Outliers detected using K-means for E33

day. SSA is a univariate method to detect outliers for each variable in the ARM data. It can quickly catch those high peak and drop points. But it requires the time series data to be continuous with no missing values. K-means is a commonly used multivariate method for clustering. Here we used it for outlier detection. The problem is that the detected outliers could be just one type of variable or multiple types of variable. It is hard to tell which is the case and get the detail for future correction. As we also averaged the raw minute level data into day level data, some outliers may be averaged out by this process.

One outlier may only be detected by SSA or Pearson correlation coefficient or K-means. Thus we combined all the three methods together as a whole framework. SSA and K-means are used directly to detect outliers. Pearson correlation coefficient can mainly be used to detect the main variables which caused the anomaly from the K-means results by comparing the pairwise correlations. Figure 5 was the result of detected outliers for *temp_mean* from E33. The red squares stand for the common outliers detected by both K-means and SSA. The orange diamonds are the ones detected by K-means excluding the common outliers. And the black stars represents the outliers detected by SSA excluding the common outliers. We can see from the figure that more outliers have been detected compared to Figure 3 and Figure 4. Thus we applied this framework on all the test data. Table II shows the number of detected outliers. The size of common detected outliers is 378 by this framework.

The current data quality or outlier detection is maintained as data quality reports (DQRs) stored in the DQR database

TABLE II
COMPARISON OF SSA AND K-MEANS OUTLIER SET SIZE

	Outlier Set Size
SSA	922
K-means	508
Intersection	378
Symmetric Difference	674

TABLE III
PRECISION AND RECALL OF SSA AND K-MEANS

Method	Variable	Precision	Recall
SSA	temp_mean	16.00%	1.20%
SSA	vapor_pressure_mean	20.70%	1.40%
SSA	atmos_pressure	0.00%	0.00%
SSA	rh_mean	14.80%	0.50%
SSA	wspd_arith_mean	0.60%	1.50%
Kmeans	5 together	12.90%	1.90%
Combined	5 together	11.10%	4.10%

with each entry manually entered [39]. A description of an event which changed the normal data is included in these DQRs. The event could be temporary operating conditions such as power failures and frozen and snow covered sensors, instrument degradation, and contamination. It could also be an extreme weather event that has never been observed before. Each DQR entry also contains a specific time range affected, list of data projects, and specific measurements. And these entries are usually submitted by either the Data Quality Office [40] or the instrument mentor [41]. It is easy to notice that this method is not efficient as it requires a lot of labor. It is nearly also impossible to detect all the outliers due to the complexity

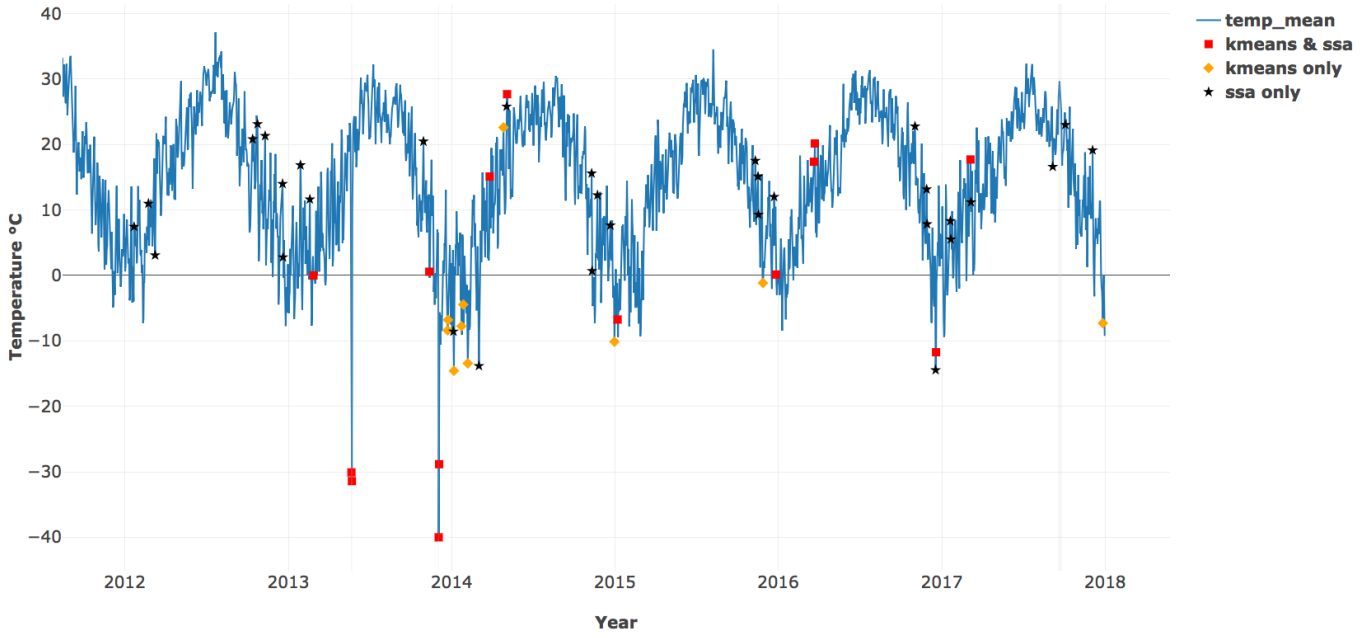

Fig. 5. Outliers detected for E33 *temp_mean* using combined algorithms

and high volume of the ARM data.

Currently, not many outliers entries stored in DQR database. Here we used manually detected 181 outliers in the DQR database as the ground truth to compare with the results from our framework. Precision and recall which were first defined in [42] were used as the comparison metric. They are commonly used to measure the quality of classification tasks [43]. Precision is calculated from True Positives divided by the sum of True Positives and False Positives. On the other hand, recall is measured from True Positives divided by the sum of True Positives and False Negatives. We treated outliers in DQR database as True Positives. Thus detected outliers not in the DQR database are False Positives. Undetected values which in the DQR database are False Negatives, and which not in the DQR database are True Negatives. Table III contains the statistics of the comparison.

Precision attempts to answer the proportion of positive identifications was actually correct. The Combined precision is 11.10% which shows that many outliers detected by the framework are not in the DQR database. Recall tries to solve the proportion of actual positives was identified correctly. The number is 4.10% which is even smaller than precision. One reason is the same as precision that the size of True Positives is much small. The other reason is that DQR database records the whole possible affected time range which makes the size of False Negatives large. It could be possible that only a few days of the data recorded during that time range are actually outliers.

V. Conclusions

In this paper we tested pairwise Pearson correlation coefficient, univariate SSA and multivariate K-means and combined them as a framework to detect outliers in the ARM data. Each method has its own drawbacks. But our experiments showed that this framework worked well compared to the manually Data Quality Report method. Currently, we only tested MET data from SGP. And we analyzed data from each instrument independently. We will apply this framework on other types of data from other facilities in the future. Meanwhile, other methods will be examined to test data from multiple instruments together such as graph theory methods [44] and machine learning methods.

Acknowledgment

This research was supported by the Atmospheric Radiation Measurement (ARM) user facility, a U.S. Department of Energy (DOE) Office of Science user facility managed by the Office of Biological and Environmental Research.

References

[1] "Arm research facility," https://www.arm.gov/, accessed: 2018-06-22.

[2] G. M. Stokes and S. E. Schwartz, "The atmospheric radiation measurement (arm) program: Programmatic background and design of the cloud and radiation test bed," *Bulletin of the American Meteorological Society*, vol. 75, no. 7, pp. 1201–1222, 1994.

[3] K. Gaustad, T. Shippert, B. Ermold, S. Beus, J. Daily, A. Borsholm, and K. Fox, "A scientific data processing framework for time series netcdf data," *Environmental modelling & software*, vol. 60, pp. 241–249, 2014.

[4] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250–2267, 2014.

[5] A. J. Fox, "Outliers in time series," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 350–363, 1972.

[6] S. Budalakoti, A. N. Srivastava, M. E. Otey *et al.*, "Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications"*, vol. 39, no. 1, p. 101, 2009.

[7] V. Chandola, V. Mithal, and V. Kumar, "Comparative evaluation of anomaly detection techniques for sequence data," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 743–748.

[8] K. Sequeira and M. Zaki, "Admit: anomaly-based data mining for intrusions," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 386–395.

[9] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[10] L. Portnoy, E. Eskin, and S. Stolfo, "Intrusion detection with unlabeled data using clustering," in *In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001*. Citeseer, 2001.

[11] M. Gupta, A. B. Sharma, H. Chen, and G. Jiang, "Context-aware time series anomaly detection for complex systems," in *Workshop Notes*, vol. 14, 2013.

[12] F. A. González and D. Dasgupta, "Anomaly detection using real-valued negative selection," *Genetic Programming and Evolvable Machines*, vol. 4, no. 4, pp. 383–403, 2003.

[13] D. Dasgupta and F. Nino, "A comparison of negative and positive selection algorithms in novel pattern detection," in *Systems, man, and cybernetics, 2000 IEEE international conference on*, vol. 1. IEEE, 2000, pp. 125–130.

[14] X. Li, J. Han, and S. Kim, "Motion-alert: automatic anomaly detection in massive moving objects," in *International Conference on Intelligence and Security Informatics*. Springer, 2006, pp. 166–177.

[15] D.-K. Kang, D. Fuller, and V. Honavar, "Learning classifiers for misuse detection using a bag of system calls representation," in *International Conference on Intelligence and Security Informatics*. Springer, 2005, pp. 511–516.

[16] D. Cheboli. (2010) Anomaly detection of time series. [Online]. Available: http://hdl.handle.net/11299/92985

[17] N. Golyandina and A. Zhigljavsky, *Singular Spectrum Analysis for time series*. Springer Science & Business Media, 2013.

[18] P. Papadimitriou, A. Dasdan, and H. Garcia-Molina, "Web graph similarity for anomaly detection," *Journal of Internet Services and Applications*, vol. 1, no. 1, pp. 19–30, 2010.

[19] J. J. Jay, J. D. Eblen, Y. Zhang, M. Benson, A. D. Perkins, A. M. Saxton, B. H. Voy, E. J. Chesler, and M. A. Langston, "A systematic comparison of genome-scale clustering algorithms," in *BMC bioinformatics*, vol. 13, no. 10. BioMed Central, 2012, p. S7.

[20] I. Jovanović and Z. Stanić, "Spectral distances of graphs," *Linear Algebra and its Applications*, vol. 436, no. 5, pp. 1425–1435, 2012.

[21] B. Pincombe, "Anomaly detection in time series of graphs using arma processes," *Asor Bulletin*, vol. 24, no. 4, p. 2, 2005.

[22] A. Kut and D. Birant, "Spatio-temporal outlier detection in large databases," *Journal of computing and information technology*, vol. 14, no. 4, pp. 291–297, 2006.

[23] D. J. Hill, B. S. Minsker, and E. Amir, "Real-time bayesian anomaly detection for environmental sensor data," in *Proceedings of the Congress-International Association for Hydraulic Research*, vol. 32, no. 2. Citeseer, 2007, p. 503.

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

[25] W. Drosdowsky, "An analysis of australian seasonal rainfall anomalies: 1950–1987. ii: Temporal variability and teleconnection patterns," *International Journal of Climatology*, vol. 13, no. 2, pp. 111–149, 1993.

[26] E. Wu, W. Liu, and S. Chawla, "Spatio-temporal outlier detection in precipitation data," in *Knowledge discovery from sensor data*. Springer, 2010, pp. 115–133.

[27] S. Yuxiang, X. Kunqing, M. Xiujun, J. Xingxing, P. Wen, and G. Xiaoping, "Detecting spatio-temporal outliers in climate dataset: A method study," in *Geoscience and Remote Sensing Symposium, 2005. IGARSS'05. Proceedings. 2005 IEEE International*, vol. 2. IEEE, 2005, pp. 4–pp.

[28] R. Rew and G. Davis, "Netcdf: an interface for scientific data access," *IEEE computer graphics and applications*, vol. 10, no. 4, pp. 76–82, 1990.

[29] Unidata. (2014) Network common data form (netcdf) version 4.1.1. Boulder, CO: UCAR/Unidata. [Online]. Available: https://doi.org/10.5065/D6H70CW6

[30] P. T. Inc. (2015) Collaborative data science. Montreal, QC. [Online]. Available: https://plot.ly

[31] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[32] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[33] N. Golyandina and A. Korobeynikov, "Basic singular spectrum analysis and forecasting with r," *Computational Statistics & Data Analysis*, vol. 71, pp. 934–954, 2014.

[34] E. Bozzo, R. Carniel, and D. Fasino, "Relationship between singular spectrum analysis and fourier analysis: Theory and application to the monitoring of volcanic activity," *Computers & Mathematics with Applications*, vol. 60, no. 3, pp. 812–820, 2010.

[35] T. Alexandrov, "A method of trend extraction using singular spectrum analysis," *arXiv preprint arXiv:0804.3367*, 2008.

[36] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of computation*, vol. 19, no. 90, pp. 297–301, 1965.

[37] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[38] F. Pukelsheim, "The three sigma rule," *The American Statistician*, vol. 48, no. 2, pp. 88–91, 1994.

[39] R. McCord and J. Voyles, "The arm data system and archive," *Meteorological Monographs*, vol. 57, pp. 11–1, 2016.

[40] R. A. Peppler, K. E. Kehoe, J. W. Monroe, A. K. Theisen, and S. T. Moore, "The arm data quality program," *Meteorological Monographs*, vol. 57, pp. 12–1, 2016.

[41] T. S. Cress and D. L. Sisterson, "Deploying the arm sites and supporting infrastructure," *Meteorological Monographs*, vol. 57, pp. 5–1, 2016.

[42] J. W. Perry, A. Kent, and M. M. Berry, "Machine literature searching x. machine language; factors underlying its design and development," *Journal of the Association for Information Science and Technology*, vol. 6, no. 4, pp. 242–254, 1955.

[43] D. L. Olson and D. Delen, *Advanced data mining techniques*. Springer Science & Business Media, 2008.

[44] J. D. Phillips, W. Schwanghart, and T. Heckmann, "Graph theory in the geosciences," *Earth-Science Reviews*, vol. 143, pp. 147–160, 2015.