

Detecting outliers in streaming time series data from ARM distributed sensors

Yuping Lu Jitendra Kumar Nathan Collier Bhargavi Krishna
University of Tennessee *Oak Ridge National Laboratory* *Oak Ridge National Laboratory* *Oak Ridge National Laboratory*
 Knoxville, TN, USA Oak Ridge, TN, USA Oak Ridge, TN, USA Oak Ridge, TN, USA
 yupinglu89@gmail.com jkumar@climatemodeling.org nathaniel.collier@gmail.com krishnab@ornl.gov

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

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 require high degree of accuracy to enable rigorous study of atmospheric processes. Outliers in collected data are common 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 sensors to identify any deviation from expected and known patterns and identify abnormal observation. The Pearson correlation coefficient, SSA and K-means methods were later combined together in a framework to detect outliers through 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, it is also essential to identify, address, and communicate any noise and outliers in the data to maintain high data quality. 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]. 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.

Different from the methods mentioned above, window-based detection breaks the time series data into overlapping subsequences with fixed window size [13]. 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 widely used algorithm for such problem [14].

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 of graph snapshots in temporal networks. Pearson correlation coefficient, which is explained in detail later, is a good candidate for such problem.

A number of approaches have been developed in literature for temporal outlier detection, especially for environmental sensor data. Birant et al. [15] discovered high wave heights values as 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. [16], [17] 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. [18] found anomalies from Australian district rainfall using rotated PCA. Wu et al. [19] detected precipitation outlier events while working on South American precipitation data set. Sun et al. [20] extracted locations which always have different temperature from their surroundings by exploring the South China area dataset from 1992 to 2002.

Within ARM program, the Data Quality Office (DQO) is charged with inspecting and assessing approximately 5,000 data fields on a daily to weekly basis. The objective of DQO is to quickly identify data anomalies and report them to site operators and instrument mentors so that corrective actions can be performed and thereby minimize the amount of unacceptable data collected. With focus on quick near real-time assessment of data, process relies heavily on univariate analysis and lacks rigorous detection of outliers. Objective of this study was to develop efficient and rigorous outlier detection technique for ARM time series data using univariate, multivariate and time series statistics techniques.

II. DATASETS

ARM data are stored and distributed in the Network Common Data Form (NetCDF) format which is self-describing and machine-independent [21], [22] and has good performance and data compression. It is commonly used to handle scientific data, especially in climate and Earth sciences, meteorology, oceanography, and remote sensing etc. All ARM data are publicly available and can be downloaded from ARM Data Center (<https://www.arm.gov/data>) 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 are available within ARM.

TABLE I
SGPMET DATASETS USED IN THIS STUDY

Facility	E1	E3	E4	E5	E6	E7
Begin Year	1996	1997	1996	1997	1997	1996
End Year	2008	2008	2010	2008	2010	2011
Facility	E8	E9	E11	E13	E15	E20
Begin Year	1994	1994	1996	1994	1994	1994
End Year	2008	2017	2017	2017	2017	2010
Facility	E21	E24	E25	E27	E31	E32
Begin Year	2000	1996	1997	2004	2012	2012
End Year	2017	2008	2001	2009	2017	2017
Facility	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 Oklahoma, United States. SGP is ARM's largest facility that comprises of a network of core and extended facilities. In our study we used MET data from 24 extended facilities where surface meteorological observations have been collected continuously and independently. While MET instruments collect a large array of direct and indirect measurements, we focused our analysis on five core meteorological 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 for our study and applied them to ARM time series data.

A. Data Pre-processing

Raw time series data from MET instruments are available at temporal resolution of one minute for all variables considered in this study. Data were pre-processed for in various analysis in our study. One minute temporal resolution time series was standardized with mean of zero and one standard deviation for Pearson Correlation analysis. A daily temporal resolution standardized time series was prepared for use with SSA based detection method. Multi-variate cluster analysis was conducted using standardized daily time series of all meteorological variables.

B. 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. Any abrupt change or break in correlation structure among meteorological behavior can be a sign of sensor malfunction and should be identified as an outlier. In addition, ARM SGP site comprise of multiple facilities making similar sets of measurement and any abrupt change in correlation structure not observed at other facilities will also indicate a potential outlier.

The Pearson correlation coefficient was first introduced by Karl Pearson [23] 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, if the value is close to 1, 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 facilities. 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 shows 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 air temperature and vapor pressure is positively correlated with correlation mean close to 0.75. And the Pearson correlation between atmospheric pressure and air temperature 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.

C. 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 anomalous observations can be indicative of extreme atmospheric events at the site and are important to identify. However, a range of natural inter-annual and intra-annual variability in meteorological 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 extreme events.

Singular Spectrum Analysis (SSA) is a popular method for time series data analysis [14], [24]. 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 [14]. For example, SSA can be applied to monitor volcanic activity [25]. It can also be used to extract trend [26]. SSA method 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.

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

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

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 .

$$X_i = (y_i, \dots, y_{L+i-1})^T \quad (1 \leq i \leq K)$$

$$\mathbf{X} = [X_1, \dots, X_K]$$

Thus the trajectory matrix is

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

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 \mathbf{S} are denoted by $\lambda_1, \dots, \lambda_L$ in the decreasing order of magnitude ($\lambda_1 \geq \dots \geq \lambda_L \geq 0$) and the corresponding eigenvectors by P_1, \dots, P_L . Let $d = \text{rank } \mathbf{X}$ and $V_i = \mathbf{X}^T P_i / \sqrt{\lambda_i}$ ($i = 1, \dots, d$). Thus, the trajectory matrix \mathbf{X} can then be written by its eigendecomposition,

$$\mathbf{X} = \mathbf{X}_1 + \dots + \mathbf{X}_d \quad (2)$$

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 Fast Fourier transform (FFT) to find the dominant frequency of each eigenvector [27]. 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 2. 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.

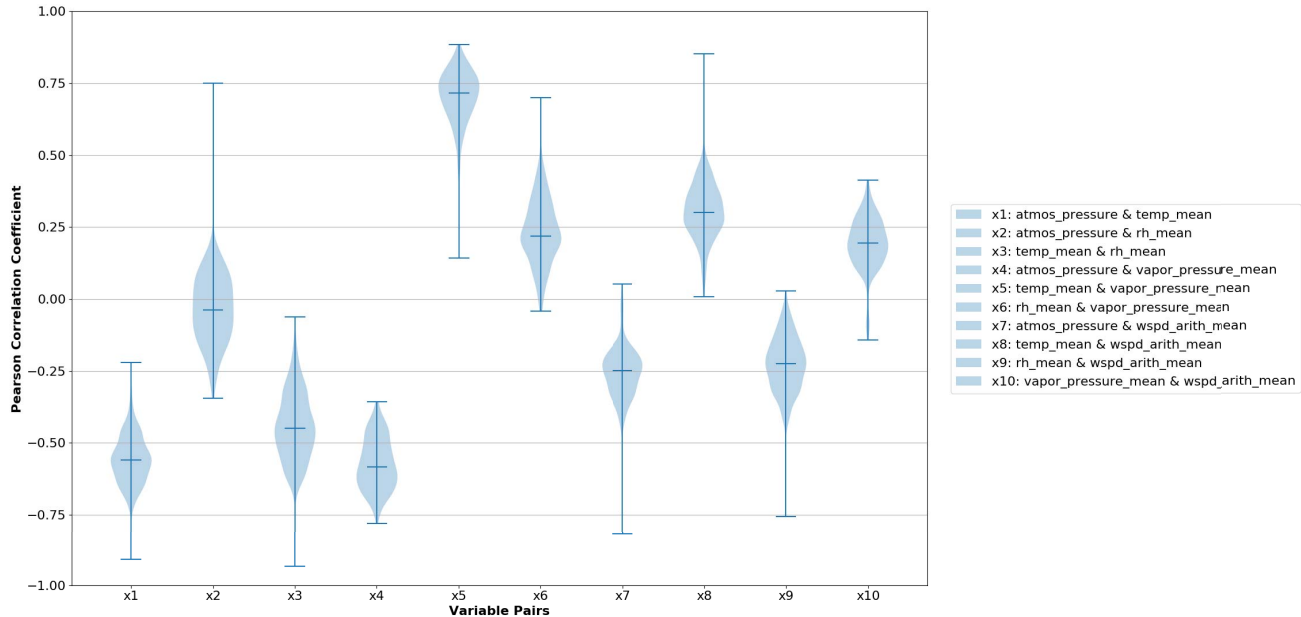


Fig. 1. Pearson Correlation patterns for ten meteorological variable pairs during spring season across all the years.

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, $R = Y - Yt$.

We applied SSA for analysis of all five meteorological variables across all facilities (Table I) to identify outliers in all meteorological observations.

Because SSA requires the time series data to be continuous, we corrected any missing values in the time series by replacing them with long term seasonality. We set $L = 400$ and isolated the signals corresponding to year and monthly frequency in the data. Thus $Yt = Yt[0] + Yt[1] + Yt[2]$. Figure 2 shows the result of SSA analysis for air temperature variable at facility E33. The first row of Figure 2 shows the raw daily time series (Yt) of air temperature, which shows no significant trend (orange line $Yt[0]$) at the site during period 2012 to 2017. The second and third rows show the annual ($Yt[1]$) and monthly ($Yt[2]$) frequencies of the temperature time series respectively. Temperature time series data shows strong annual and monthly frequencies at the sites which expected and reflective of long term weather patterns experienced at the SGP site. The last row shows the time series of residual after removing the trends, and annual and monthly frequencies from the data. While

some of the residuals may be reflective of natural variability, the anomalous positive or negative temperature residuals can be identified as outliers in the data. Multiple methods are available to set a threshold for extreme values in the residuals as outliers. We used the three sigma rule to extract outliers [28]. For example, the two peak points in Figure 2 are larger than three sigmas, thus are outliers.

D. *K-means*

Southern plains, where SGP site is located, are known to experience frequent extreme storms occurring most frequently during spring and early summer seasons. Identifying these extreme events is of interest for scientific users of the data to study and/or isolate these phenomena. However, meteorological variables during such events won't be captured by Pearson Correlation as they may still follow known correlation structure at seasonal scales or by SSA method since any individual variable may not show large deviation. Multivariate approach like k -means clustering have been widely used to identify weather and climate regimes [29], [30]. We used k -means clustering algorithm to delineate the weather regimes at SGP site. While extreme storms and weather events that often occur at sub-daily timescales may still fall within identified known weather regimes at the site, they often are out of norm extremes within the regime and of interest to us.

K -means is a partitioning clustering algorithm [9], [31]. It starts with user specified k centroids, and assigns the points to the nearest centroid. Then it computes new k centroids and assigns all data points to these centroids again. This process is repeated until convergence criteria is met.

We applied k -means clustering to ARM meteorological data set to defined weather regimes at SGP site. We then calculated

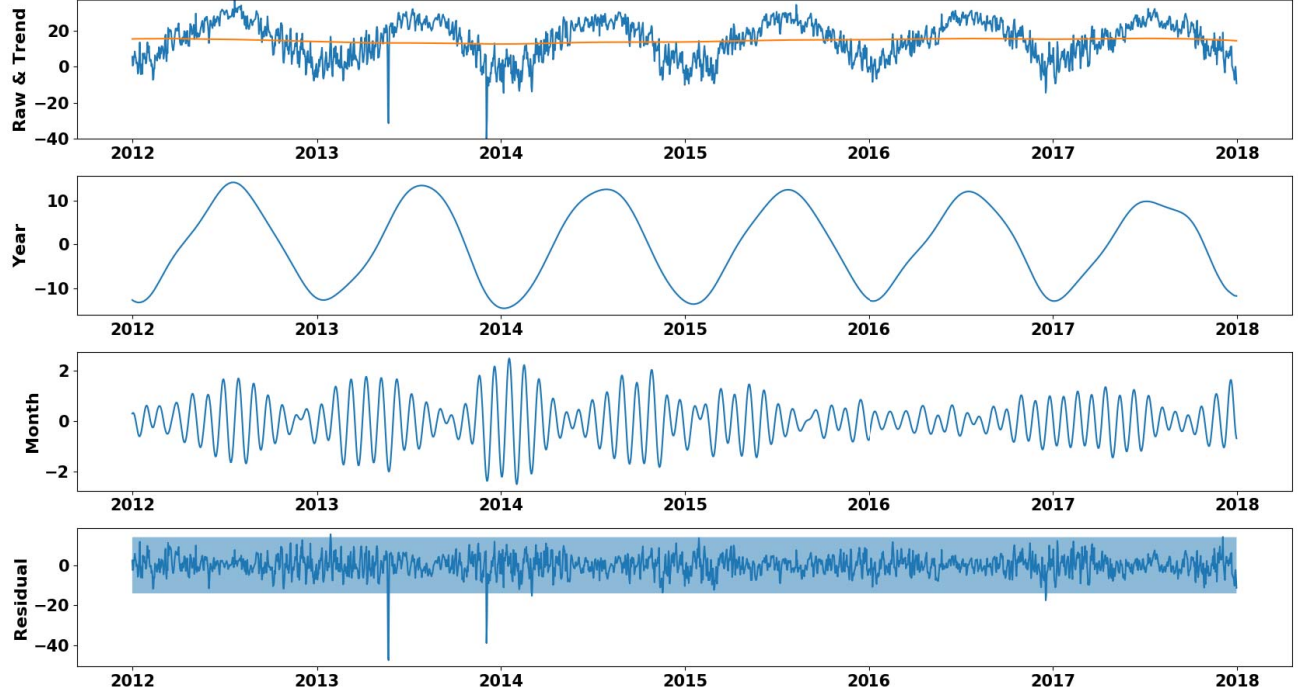


Fig. 2. Decomposition of air temperature data from MET instrument at facility E33 using SSA method to isolate various frequencies.

Algorithm 1: K-means Outlier Detection

Input : ARM time series data

Output: Outliers

```

1 outliers  $\leftarrow \emptyset$ 
2 df  $\leftarrow$  ARM time series data
3 data  $\leftarrow$  df['atmos_pressure', 'temp_mean',
  'rh_mean', 'vapor_pressure_mean', 'wspd_arith_mean']
4 number_of_clusters  $\leftarrow$  4
5 clusters  $\leftarrow$  K-means(data, number_of_clusters)
6 distances  $\leftarrow$  Distance between each point and its centroid
7 mean  $\leftarrow$  arithmetic mean of distances
8 sigma  $\leftarrow$  standard deviation of distances
9 threshold  $\leftarrow$  mean + 3 * sigma
10 for  $i$  in range(size of distances) do
11   if distances[ $i$ ] > threshold then
12     outliers  $\leftarrow$  outliers  $\cup$  distances[ $i$ ]
13   end
14 end
15 return outliers

```

the distance of each point within a cluster to its corresponding cluster centroid. Vector of distances within each cluster were used to identify points that are on fringes of the regime they belong to and considered outliers. All five meteorological variables were used in this analysis. Algorithm 1 describes the workflow.

Given known seasonal patterns at the site we set k to

four to determine weather regimes for four seasons. Figure 3 shows the four regimes at facility E33 that representing spring (cluster 1), winter (cluster 2), summer (cluster 3) and fall (cluster 4). Data points within each weather regime (or cluster) that are at significant distance from their clusters (identified by red squares in Figure 3) were identified as outlier (and may correspond to extreme weather events).

E. Evaluation of outlier detection

ARM data quality assurance program maintains a database of outliers that has been identified, inspected and documented for all historical data. However, recorded data quality issues are added manually for historical data when an issue is identified or reported and are known to be incomplete [32]. A description of the outlier event is included in these DQR which often are temporary change in operating conditions such as power failures, frozen and snow covered sensors, instrument degradation, or contamination. Most often extreme weather events are not captured and reported by the current system 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 [33] or the instrument mentor [34]. The Data Quality Reports (DQR) are stored and available as PostgreSQL database (<http://dq.arm.gov>). During study period of 1994-2017, across 24 facilities studied at SGP site, a total of 181 DQRs were reported for MET variables analyzed, each often spanning multiple day time period totaling 8540 days. The reported data quality issues covered all five variables:

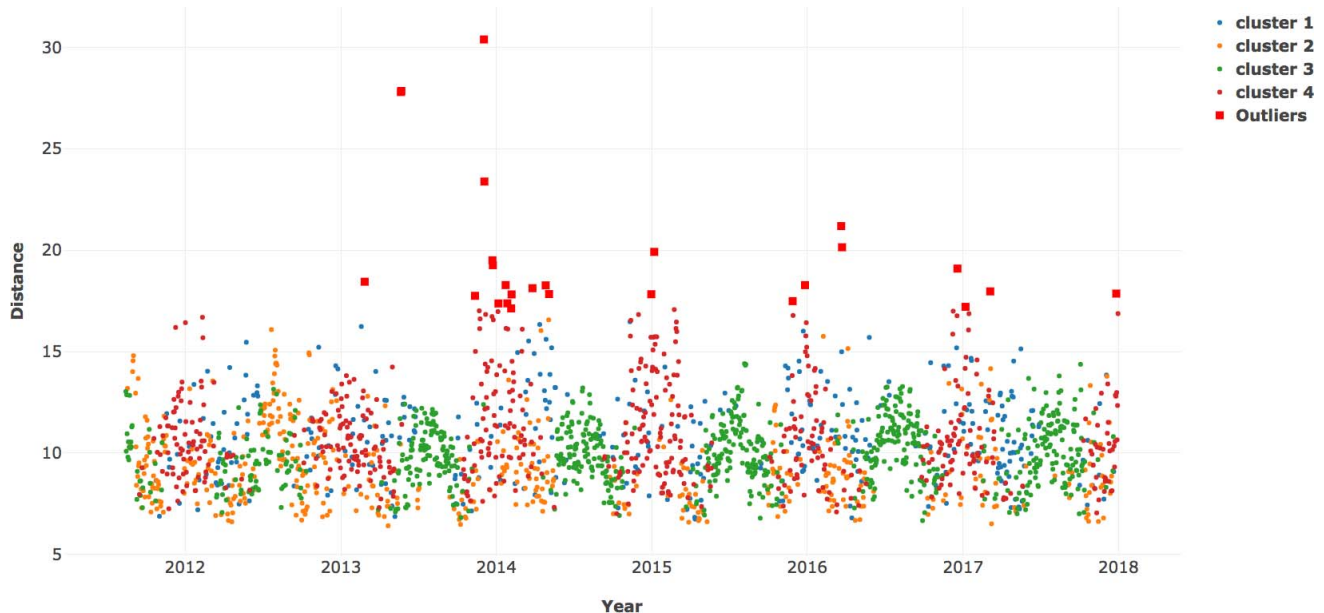


Fig. 3. Outliers detected using k -means method at facility E33. X-axis represent the daily meteorological time series, colored by cluster (weather regime) they belong to, while Y-axis show the distance of the data point from the centroid of its cluster (weather regime).

air temperature (41 events; 8217 days), vapor pressure (42 events; 8194 days), atmospheric pressure (12 events; 76 days), relative humidity (32 events; 8108 days), and wind speed (52 events; 265 days). We evaluated outliers identified by methods developed in this study against the DQRs in the database through database queries and calculated *Precision* and *Recall* metrics [35]. We treated outliers detected in DQR database as True Positives. The equation 3 and 4 show the calculation of *Precision* and *Recall*.

IV. RESULTS AND DISCUSSION

All three methods were applied to five meteorological variables across all facilities. The methods identified different sets of outlier events, with some events identified by more than one method (Figures 1,2,3).

Among three methods Pearson correlation was least effective with frequent false negatives. Pearson correlation is also an aggressive method that it may include many false positives. Those are all due to the fact that pairwise Pearson correlation method was applied at seasonal scale. Pearson correlation coefficient is a pairwise comparison method, however, if the two variables deviate in the same direction, their correlation may not change significantly and thus may go undetected. Due to seasonal nature of the analysis, it was not able to identify outliers that persisted at hours to days only. Univariate SSA method was very effective at identifying outliers with extreme high and low values in the time series but required the input data to be consistent with no missing values. k -means could be used to detect extreme storms and weather events but it was hard to tell which variable mainly caused the abnormality. However, these drawbacks could be easily

overcome by combining methods together to detect outliers from three different angles.

In our experiment, SSA method identified largest number of outlier events (922) (Table II) across the entire dataset, while k -means identified 508 events. While 378 events were identified as outliers by both the methods (intersection), 674 events were only identified by one of the methods (Table II). Figure 4 shows all the outliers detected by Pearson correlation, SSA and k -means methods at facility E33 for air temperature. When using Pearson Correlation, we used the interquartile range (IQR) method to extract outlier seasons that is those values beyond Tukey's fences as the three sigmas rule is too aggressive for Pearson correlation [36]. However, since the Pearson Correlation was applied at seasonal scale it identified only a few outlier seasons in the data. For example, at facility E33 Pearson Correlation analysis of temperature time series identified spring 2015 that experienced a severe frost event as outlier season (Figure 4). When combined together SSA and k -means methods had *Precision* of 11.10% which shows that many of the outliers detected are not within ARM DQR database, which is a known limitation of the current records that this current study is trying to address. Detected outliers also had low *Recall* which in addition to small number of true positives can be due to fact that DQR database often records a wide affected date range for an identified outlier instead of a precise date thus leading to large false negatives, all of which leads to low *Recall* values.

Overall, when combined together within a framework, set of methods applied allows to capture outlier events caused by a wide range of conditions.

$$\text{Precision} = \frac{\text{True Positives (Outliers detected in DQR database)}}{\text{True Positives} + \text{False Positives (Outliers detected not in DQR database)}} \quad (3)$$

$$\text{Recall} = \frac{\text{True Positives (Outliers detected in DQR database)}}{\text{True Positives} + \text{False Negatives (Undetected records in DQR database)}} \quad (4)$$

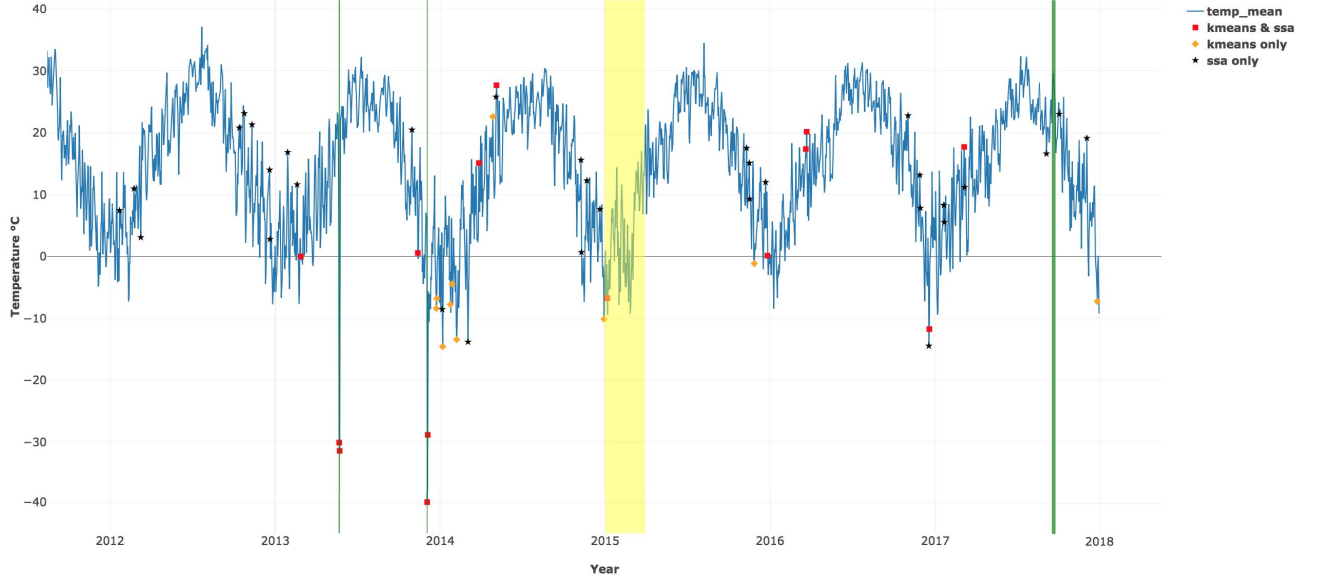


Fig. 4. Outliers detected at facility E33 for air temperature by Pearson correlation, SSA and k -means algorithms. The yellow shaded areas are outliers detected by Pearson correlation. Outliers detected by both SSA and k -means algorithms are shown by red squares, while those identified by SSA and k -means only are indicated by black stars and orange diamonds respectively. DQR records are denoted by the vertical green shaded areas.

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	Air Temperature	16.00%	1.20%
SSA	Vapor Pressure	20.70%	1.40%
SSA	Atmospheric Pressure	0.00%	0.00%
SSA	Relative Humidity	14.80%	0.50%
SSA	Wind Speed	0.60%	1.50%
Kmeans	All Variables	13.00%	1.90%
Combined	All Variables	11.10%	4.10%

V. CONCLUSIONS

In this paper we tested pairwise Pearson correlation, univariate SSA and multivariate k -means based method for detection of outliers in the data at ARM meteorological observations at SGP site. Combining the approaches within a framework for streaming data within ARM provides a platform to detect outliers from a wide range of sensor failure scenarios to extreme events. While each of the methods developed and

applied in this study has its strengths and limitations, our evaluation against existing database of data quality issue suggests that the framework is able to identify known outliers well. Although our current study focused on meteorological observations, it provides a framework for an efficient outlier detection of streaming datasets within ARM that can be extended to other classes of time series datasets not only tested MET data from SGP. In the future, we plan to analyze multiple classes of instruments like meteorological, radiometric, radar etc. simultaneously for improved detection of outliers. We also plan to develop multivariate SSA [37] and machine learning techniques to address this high dimensional problem in an operational data center environment.

The three algorithms and visualizations presented in this paper were implemented in Python. All codes and results are available on GitHub (<https://github.com/YupingLu/arm-pearson> and <https://github.com/YupingLu/arm-ssa>).

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. Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This manuscript has been authored by UT-

Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

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. Cheboli. (2010) Anomaly detection of time series. [Online]. Available: <http://hdl.handle.net/11299/92985>
- [14] N. Golyandina and A. Zhigljavsky, *Singular Spectrum Analysis for time series*. Springer Science & Business Media, 2013.
- [15] 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.
- [16] 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. Cite-seer, 2007, p. 503.
- [17] 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.
- [18] W. Drosowsky, "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.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] Unidata. (2014) Network common data form (netcdf) version 4.1.1. Boulder, CO: UCAR/Unidata. [Online]. Available: <https://doi.org/10.5065/D6H70CW6>
- [23] 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.
- [24] N. Golyandina and A. Korobeynikov, "Basic singular spectrum analysis and forecasting with r," *Computational Statistics & Data Analysis*, vol. 71, pp. 934–954, 2014.
- [25] 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.
- [26] T. Alexandrov, "A method of trend extraction using singular spectrum analysis," *arXiv preprint arXiv:0804.3367*, 2008.
- [27] 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.
- [28] F. Pukelsheim, "The three sigma rule," *The American Statistician*, vol. 48, no. 2, pp. 88–91, 1994.
- [29] F. M. Hoffman, W. W. Hargrove Jr, D. J. Erickson III, and R. J. Oglesby, "Using clustered climate regimes to analyze and compare predictions from fully coupled general circulation models," *Earth Interactions*, vol. 9, no. 10, pp. 1–27, 2005.
- [30] W. W. Hargrove and F. M. Hoffman, "Potential of multivariate quantitative methods for delineation and visualization of ecoregions," *Environmental management*, vol. 34, no. 1, pp. S39–S60, 2004.
- [31] 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.
- [32] R. McCord and J. Voyles, "The arm data system and archive," *Meteorological Monographs*, vol. 57, pp. 11–1, 2016.
- [33] 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.
- [34] T. S. Cress and D. L. Sisterson, "Deploying the arm sites and supporting infrastructure," *Meteorological Monographs*, vol. 57, pp. 5–1, 2016.
- [35] 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.
- [36] J. W. Tukey, *Exploratory data analysis*. Reading, Mass., 1977, vol. 2.
- [37] P. C. Rodrigues and R. Mahmoudvand, "The benefits of multivariate singular spectrum analysis over the univariate version," *Journal of the Franklin Institute*, vol. 355, no. 1, pp. 544–564, 2018.