# Power System Data Management and Analysis Using Synchrophasor Data


Rich Meier
Eduardo Cotilla-Sanchez
Oregon State University
Corvallis, OR 97331, USA

Ben McCamish
Washington State University
Vancouver, WA 98686, USA

David Chiu
University of Puget Sound
Tacoma, WA 98416, USA

Miles Histand
Jordan Landford
Robert B. Bass
Portland State University
Portland, OR 97201, USA



***Abstract*—This paper provides a synchrophasor data analysis methodology that leverages statistical correlation techniques in order to identify data inconsistencies, as well as power system contingencies. Included in this analysis are the techniques for data management that are used to process the high-granularity and high-cardinality data gathered from *synchrophasor* sensors – also know as *Phasor Measurement Units (PMUs)*. This research utilizes real, archived PMU data obtained from the Western Electric Coordinating Council (WECC) in order to show that this methodology is not only feasible, but extremely useful for power systems monitoring, decision support, and planning purposes. The results presented indicate preliminary identification of PMU data issues, as well as power system instabilities.**



***Keywords*—Synchrophasor, Correlation, Big Data, Protection, Smart Grid, PMU**


## I. Introduction

Recently, electric power grids have experienced increased penetration of non-dispatchable generation technologies, load congestion, demand for quality electric power, environmental concerns, and threats to cyber-security and physical infrastructure. Pressure from these issues compel engineers to create tools that leverage modern communications, signal processing and software applications to provide operators with insight into the operational state of electric power systems. As Horowitz, *et al.* discussed, there are multiple aspects to achieving the level of knowledge and control necessary to keep one of the world's greatest engineering feats stable and operational [1]. A few notable characteristics that apply to this research are: improved and accurate sensing and measurement capability, advanced control methods, and decision support. To this end, *synchrophasors* or *phasor measurement units (PMUs)*, have provided a solution to the sensing and measuring aspect. The massive amounts of data acquired by this technology should be leveraged to optimize the operation of electric power grids.

The purpose of our research is to develop computational processes that use the high-granularity data provided by PMUs to improve control and decision-making for operators of electric power grids. Similar efforts have emphasized using PMU data to monitor critical power paths [2], identify transmission line fault locations [3], isolate and mitigate low-frequency zonal oscillations [4], and predict critical slowing down of the network [5]. It has been stressed in [6] that synchrophasors are indeed a valid solution for wide area monitoring systems (WAMS) and state estimation – though there are still many challenges to address here. There is also the issue of data integrity which we strive to address in our correlation technique. Other research has addressed the possibility of obtaining erroneous data from PMUs and subsequently using this data for wide area montioring and decision making. This is a significant issue, and both [7] and [8] develop algorithms for analyzing and improving data integrity during fault or transient situations – a common culprit of bad data. As seen from recent work, using PMU data effectively is not only an electrical engineering and power systems challenge, but rather a multi-disciplinary issue that spans Big Data, computer science, mathematics, and systems engineering.

Our algorithm seeks to find correlation patterns among a wide variety of PMUs installed at different locations within a balancing area. We employ real-time playback of actual archived PMU data in order to mimic a real power system environment. Our correlation algorithm is implemented to analyze the incoming PMU data streams and detect data errors as well as identify power system contingencies. So far, our results show promising qualitative ability to distinguish between nominal operation, PMU sensing errors, and actual power system contingencies. We can conclude that this correlation technique, when paired with a fast database architecture as developed in [9], has real promise for being an advanced and automated tool to improve operator decision-making. The following work will first discuss PMU data gathering and manipulation in Section II. Next, Section III will introduce our correlation methodology. Some of our preliminary results will be outlined in Section IV, followed by possible future expansion on this technique in Section V. Finally, we conclude this work in Section VI.

## II. Data Gathering and Manipulation

Due to the relatively recent installation of PMU devices in the power system, one major component of this research is the management of big data. So far, pre-processing methodologies to handle high-cardinality data from PMUs are not widely available, and little progress has been made to streamline and consolidate these algorithms. Therefore, the two major capabilities necessary to maintain interoperability between raw power system data and our correlation methodology are described below – namely data playback and data storage.

### A. Historical Data Playback

The first step in creating a real-time, data-playback engine was analyzing the characteristics of the historic data sets obtained from Bonneville Power Administration (BPA). By

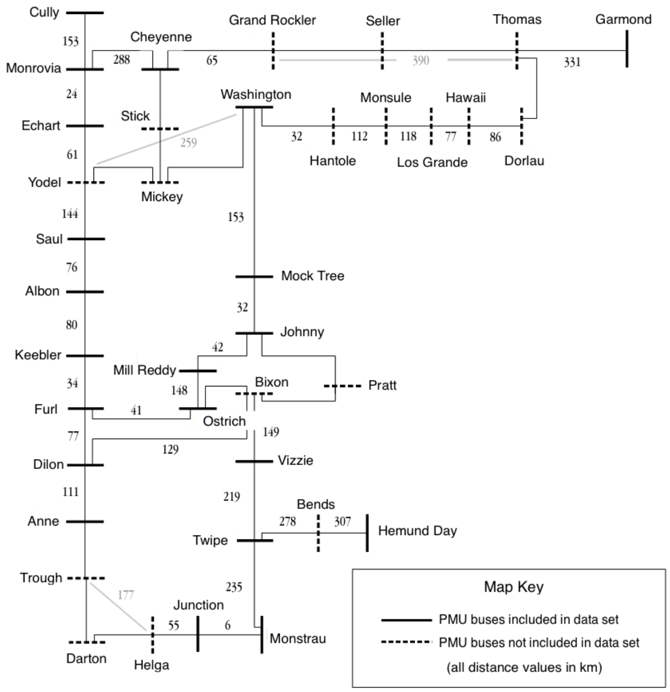


Fig. 1. The relative location of and distances between PMU sites in kilometers (not to scale).

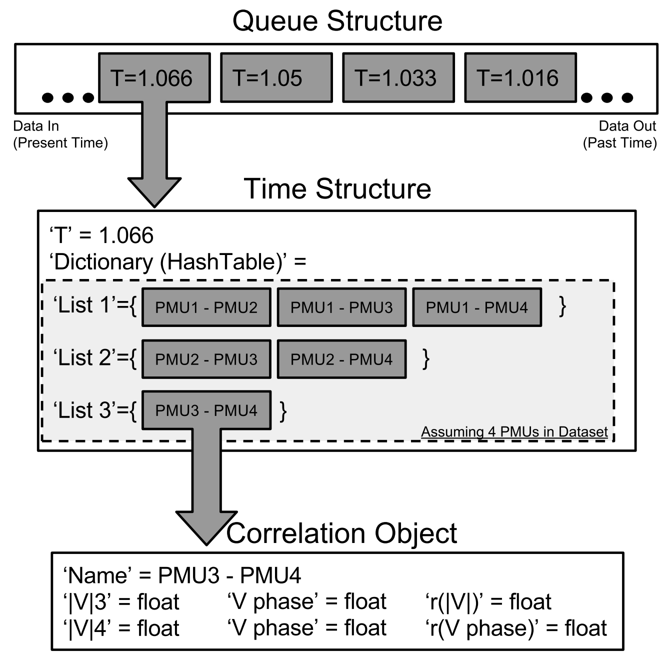


Fig. 2. The 3-tiered data structure used to store PMU data.

understanding the file traits, recorded power system attributes, data discretization rate, and topological layout of the PMUs (depicted in Figure 1), we were able to create a *Phasor Data Concentrator (PDC)* engine. Often, PMUs located at different points around the grid are grouped by zones, and consequently their data streams are multiplexed to a single data logging point. This central data entry point is what is commonly referred to as a PDC. Our PDC engine serves a similar purpose by recreating a single multiplexed data stream based on the archived operational PMU data, thus we mimic real-time power system operation.

The one-year data set we assessed includes information from August 2012 to August 2013. It totals 950 GB of positive-sequence voltage magnitude ($V$) and positive sequence voltage phase angle ($\phi$). Each measurement is represented by a *date/time* and its corresponding *phasor* value. These measurements are acquired every $0.0167$ seconds (60 Hz). The phase angle $\phi$ is a time-varying real number that oscillates within the range of $[-180, 180]$. The voltage, on the other hand, is a non-negative real number. Each file in the set typically holds one to five minutes of data from each of the 20 separate PMU sites. We opted to consolidate and standardize these files for ease of input into our PDC engine.

### B. Data Structure

As positive sequence voltage data is generated in the time-domain by our PDC, this data must be read into the working memory of our correlation algorithm. In an effort to minimize computational complexity, we developed a custom data structure in order to quickly append new data, reference data already stored, and account for multiple characteristics such as time, magnitude, phase, and correlation coefficients, for each of the 20 PMUs. This versatile data structure is depicted in Figure 2. As seen, the lowest layer of our data management system holds the actual values read in from the PDC feeder as well as the calculated correlation coefficients (referred to as "Correlation Objects"). These Correlation Objects are dynamically created based on the number of PMUs multiplexed. We have effectively reduced redundancy of data in the correlation technique since Correlation Objects represent combinations of two PMUs. The next layer of our management system is made up of Correlation Object combinations at a specific time stamp (referred to as "Time Structure"). This creates a triangular matrix of Correlation Objects that reference both a unique time as well as the PMU data and correlation value at that particular time. Finally, at the highest management level, each Time Structure is stored in a queue of dynamic length. The PMU and correlation data can effectively be monitored in the time domain for any desired window of time.

## III. Correlation Methodology

Besides pre-processing PMU data, the main contribution of this research is to provide a method that can be used for decision-making during grid operation, especially during contingency situations. In order to achieve this level of operator support, however, we must be able to distinguish between various PMU and power system events. We propose that a correlation technique can be used to flag specific data and power system events. Below we expand on the correlation methodology that was developed in order to identify events.

Correlation is a well known mathematical and statistical method for determining the compatibility of large data sets. Specifically, the Pearson Product-Moment correlation determines how well data is linearly correlated, and has been used successfully in other graph-based problems such as [10].

Given two independent input sets of data $X$ and $Y$ of length $N$ ($X$ and $Y$ being either the momentary magnitude or phase data values of two PMU site readings), the Pearson correlation yields a correlation coefficient $r$ between $-1$ and 1 based on the following equation:

$$r = \frac{\Sigma(XY) - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{(\Sigma(X^2) - \frac{(\Sigma X)^2}{N}) \times (\Sigma(Y^2) - \frac{(\Sigma Y)^2}{N})}} \tag{1}$$

Two modifications and application-specific improvements were made to this mathematical formula. First, the algorithm was made incremental. In this way, each data point could be read in from the PDC feeder and immediately incorporated into its correlation coefficient without the need to directly calculate each summation, average, and standard deviation repeatedly at each time step. This helps reduce run time. Second, the correlation algorithm was built to maintain correlation information over varying windows of time. The queue data structure actually maintains multiple separate pointers to end positions of each defined sliding window. This is seen clearly in Figure 3. It is worth noting here that separate lists for each sliding window are *not* created. Rather, pointers in a single list are maintained to minimize memory usage, as well as minimize copies of data already managed.

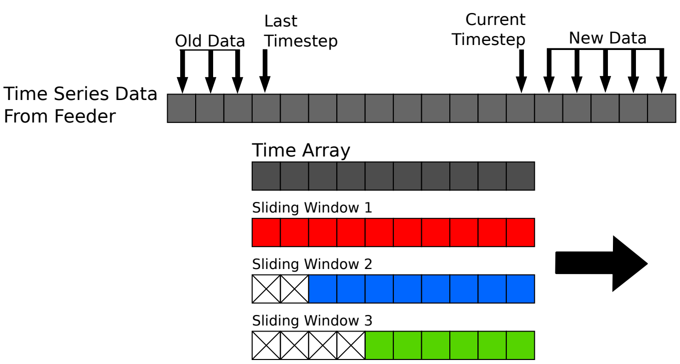


Fig. 3. The window sizing for data queue.

The addition of this latter feature allows for pairs of PMUs to be correlated over different time intervals concurrently. This design allows different events to be identified based on different sliding window sizes. This capability to correlate over multiple discrete periods of time is especially useful in determining if suspect correlations are due to data issues or are in fact real disturbances. In our methodology, large window sizes correspond to 1200, 600, and 60 data points (20 sec., 10 sec., and 1 sec. respectively). We also utilize smaller, multi-cycle window sizes (54, 48, 30, 18, 12, and 6 data points) in order to assist with identifying the difference between data events and power system contingencies. A visual depiction of the window sizes used in our case study (Section IV-B) can be seen in Figures 4 and 5.

We hypothesize that it is possible to characterize the difference between data inconsistencies and power system contingencies. Along with this ability, we also anticipate the ability to detect the nature of specific issues. We use this simple statistical measure, along with our proposed modifications, to achieve this end. An enumeration of the data and power

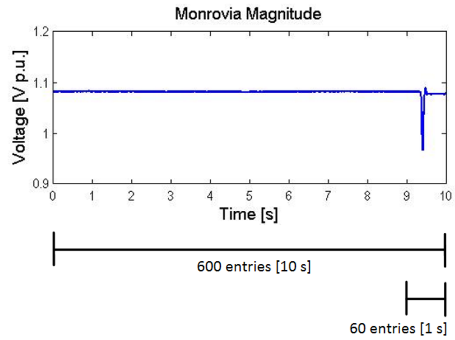


Fig. 4. Voltage magnitude plot during a lightning event (at the specific case-study bus) over large window sizes.


Fig. 5. Voltage magnitude plot during a lightning event (at the specific case-study bus) over small window sizes.

TABLE I. DATA AND POWER SYSTEM EVENT IDENTIFICATION

| Data Event | Expected Identifier/Description |
|---|---|
| Data Drop | V† and/or $\phi^*$ data = 0 |
| PMU Misread | Repeated values over multiple time steps |
| Loss of GPS Synchronization | $\phi$ drift, PMU units not synced |
| **Power System Event** | **Expected Identifier/Description** |
| Power Flow Contingency | Change in $\frac{d\phi}{dt}$ |
| Generator/Load Trip | Change in voltage and/or $\frac{d\phi}{dt}$ |
| Transmission Line Trip | Change in V and/or slight change in $\phi$ |
| Power Transformer Tap Change | Change in V |
| Miscalibration of Transformer | Pending further investigation |
| Capacitor/Reactor Switching | Change in V and $\phi$ |
| Inter & Intra Zone Oscillations | Slow-coherent change in V or $\phi$ |

† V = Positive sequence bus voltage magnitude
* $\phi$ = Positive sequence bus phase angle

system events we seek to identify can be found in Table I. The distinction here is important because, with any large-scale data set, there is a question of data validity. It is of strategic importance to identify false data originating from PMU inaccuracies, especially since these devices are used to inform higher-level applications such as state-space estimation and remedial action schema.

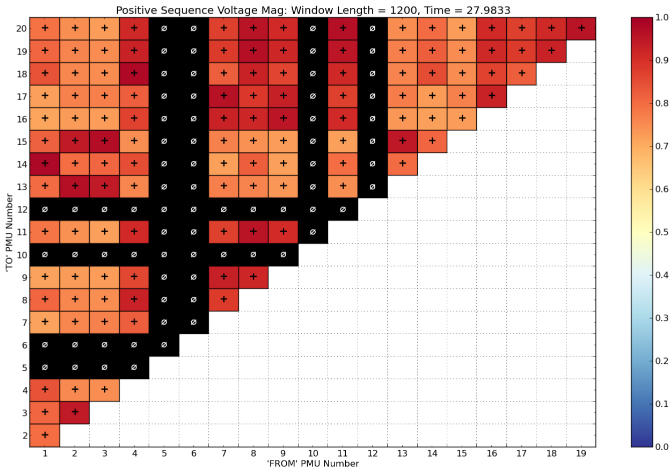


Fig. 6. Example PMU correlation visualization (topological distance) over 1200 data points (20 seconds).

## IV. Results

This section focuses on highlighting some of the preliminary qualitative information obtained by processing and analyzing the PMU data streams via our correlation methodology. We begin with a short introduction to our visualization technique and the analysis behind producing it. Next we follow up with a case study at one particular bus in the system (depicted in Fig. 1 and Fig. 11) – namely the Monrovia Bus. We demonstrate that the algorithm is capable of identifying the difference between data events and power system events.

### A. Visualization Structure

The purpose of this subsection is to introduce the layout of the visualization structure. A sample is seen in Figure 6. A few noteworthy points are as follows: each coordinate (square) represents the correlation coefficient of the two PMUs that make up its coordinates. The color of the square represents how close the correlation is to $1$ or $-1$, and the sign at the coordinate represents either positively correlated or inversely correlated PMU pairs. Typically a magnitude of correlation above $0.4-0.5$ is considered correlated. Thus any squares depicting blue shades would be considered de-correlated. It is important to keep in mind that this visualization is temporal, and represents different time window lengths as discussed in Section III (Fig 3).

The blacked out columns and rows with the null symbol represent PMUs that were offline during that particular time. This is typically given by the PMU node producing $0$ data. Our algorithm detects this, and in the later case study we remove these PMUs because they are consistently yielding unusable data.

One final addition made to the analysis and visualization of the PMU data was incorporating electrical distance into the spatial organization of each monitored bus. The notion of electrical distance has been proved useful in multiple power systems applications, but was developed most notably by Cotilla-Sanchez et al. in [11] for the purpose of multi-objective power network partitioning. In our data set, adjacent cells within the triangular visualization matrix are referenced to

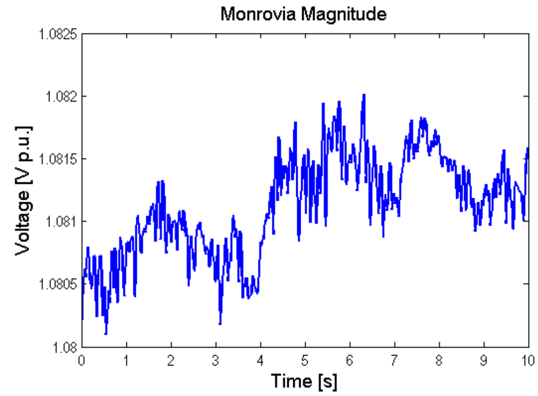


Fig. 7. Voltage magnitude plot of the Monrovia bus during nominal operating conditions.

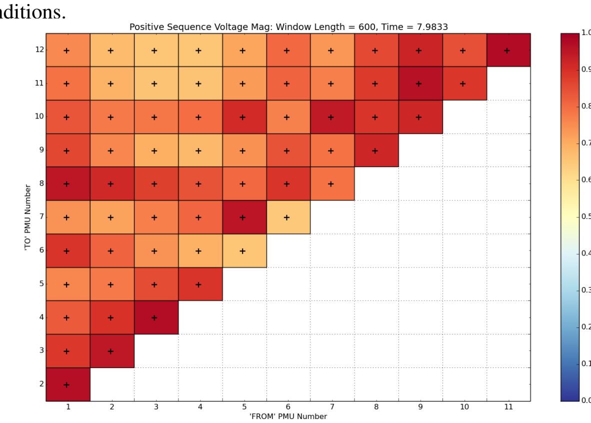


Fig. 8. Visual depiction of the correlation values during nominal operating conditions (electrical distance), taken over the same 10 second window as Figure 7.

PMU 1, either topologically, or electrically. We anticipate that this organization of PMUs will produce electrically coherent zones. As a result, the visualization will naturally cluster thus benefiting ease of analysis and application of advanced techniques such as pattern recognition.

### B. Monrovia Event Case Study

In order to demonstrate some preliminary identification of a subset of the events listed in Table I, we analyzed a set of known data and power system contingencies that occurred at the Monrovia Bus. We discuss below the analysis of known clean data sets, the PMU data drop and PMU data misread contingencies, and finally a known lightning event near the Monrovia bus.

*1) Clean Data:* Data sets void of any data issues or disturbances are categorized as 'clean' data sets. We know (from archived notes and comments) that these subsets of data exhibit the system operating under normal circumstances. Figure 7 depicts a single bus's positive sequence voltage magnitude plot over a ten second window. It is worth noting that the y-axis scale is zoomed in tightly around $1.081$ [PU].

Next, we characterized correlation when using a ten second window. Figure 8 depicts the correlation coefficients for every combination of PMUs in the network during nominal operation. The correlation matrix depicts how well coupled each

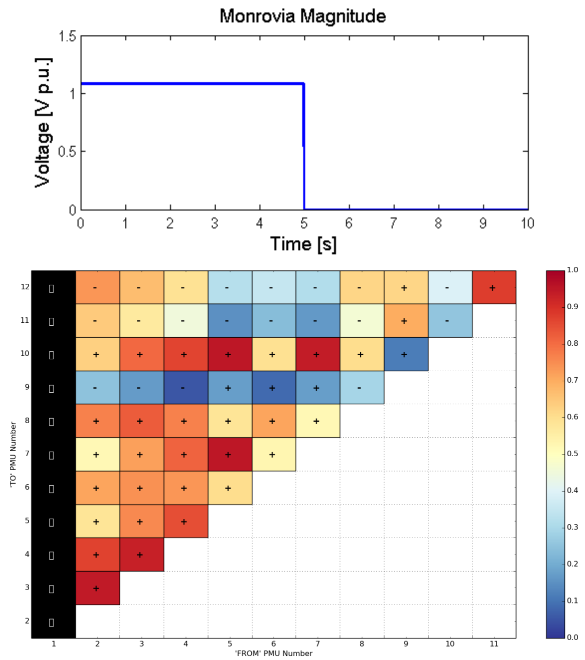


Fig. 9. A flagged "Data drop" event at the Monrovia bus with $\frac{1}{10}$ sec. sliding window (electrical distance).

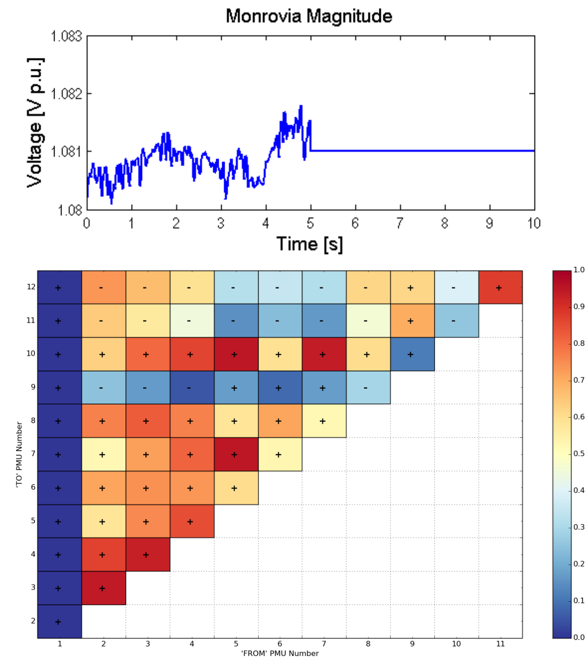


Fig. 10. A flagged "PMU Misread" event near the Monrovia bus with $\frac{1}{10}$ sec. sliding window (electrical distance).

PMU is to every other. As expected, we see that near-by buses are strongly correlated within distinct clusters, though there is some noise associated.

*2) PMU Data Events at Monrovia:* PMU data streams must be validated as accurate in order to ensure reliable and effective decisions are made during grid operation. Invalid data can be introduced in multiple ways, and so far we have created the ability to detect two specific data events. First, the PMU may go offline resulting in a constant stream of "zero" data (termed "data drop") as seen in Figure 9. Second, a PMU data stream may produce unreliable data, which is characterized by repeatedly producing the same measurement over a discrete window of time, as shown in Fig. 10.

Both of these data contingencies are flagged by our algorithm using the small window sizes, as typical data events occur in sub-second time frames. As seen in the images, the full-column pattern of null data and severe de-correlation indicates a data event at the Monrovia Bus.

*3) Power System Event at Monrovia:* The final type of event that our correlation technique is currently able to qualitatively characterize is when a power system contingency occurs. As will be discussed in Section V the ultimate goal is to flag specific events outlined in Table I, however for this case study we focus on a known lightning event near the Monrovia bus (indicated in Fig. 11).

Figures 4 and 5 depict the lightning event in the time domain. In this section we run the correlation algorithm, on this particular lighting strike, over two different window sizes (10 sec. and $\frac{1}{10}$ sec.). The results can be seen in Figures 12 and 13 respectively.

There are a couple of deductions that can be made from these power system contingency correlations. First, it is feasible to identify clusters of PMUs that change drastically from highly correlated (as in Fig. 8) to highly uncorrelated (as seen in Fig 12). This suggests that it would be possible to use higher order classification methods or pattern recognition to identify this type of contingency. Second, inspecting Figure 12 and referring to Figure 11, we see emergence of how electrical distance influences correlation. Since buses south of Monrovia are along a parallel path they are seen as, up to a certain distance, having a lower impedance when compared to the 'Cully' bus directly above. The gradient in correlation along the upper rows suggests this to be the case.

Furthermore, in order to rule out the possibility that these patterns of de-correlation were caused by invalid data we can directly compare correlations developed over smaller window sizes. If we compare Figure 13 to Figures 9 and 10, we see that there are no PMUs that are entirely de-correlated or blacked-out (due to null data) – both of which are distinct patterns of data events. We can thus safely classify this contingency as a power system contingency rather than a data issue.

## V. Future Work

So far, the synchrophasor data analysis techniques described in this research have shown potential for furthering automated grid control and decision support. Some next steps are described below.

There are several experiments that still need to be performed for one to better understand the correlation of PMU data. First, it is necessary to develop a noise band for nominal operating point data across many settings such as number of PMUs available, seasonal variables in the data, and other subsets of exogenous variables that occur even during normal circumstances. This will require further experimentation and characterization of the known generic and "clean" data in the data set. In order to establish a full understanding of this

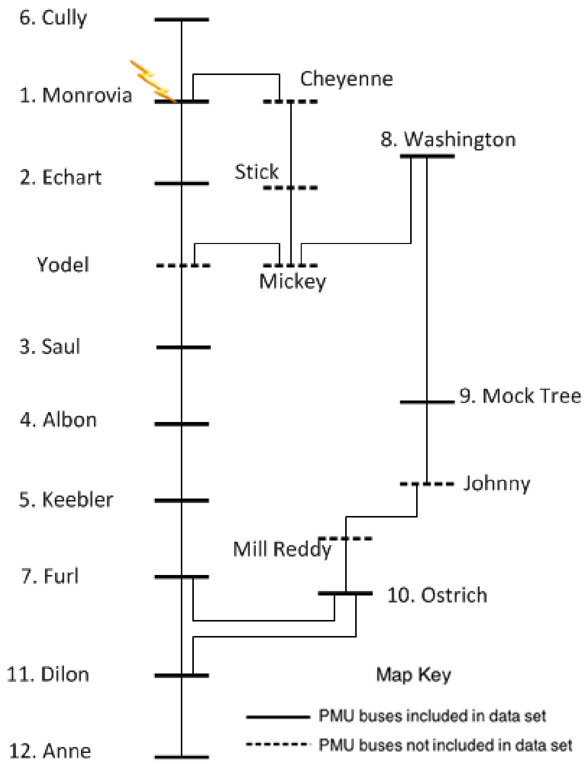


Fig. 11. Lightning event at Bus Monrovia. Bus number assignments align with correlation visualizations (electrical distance)

noise band, multiple different sliding window lengths ranging from sub-seconds to minutes will need to be fully analyzed. Once a noise band is developed, it will be possible to set up a quantitative classification technique (expanding on the current visual method) where, likely, some machine learning algorithms will become relevant.

There are also a few other questions and functionality that will need to be addressed. The speed and memory performance of the correlation algorithm and the corresponding data structure need to be analyzed. This will be important for validating real-time capability. Lastly, the use of different correlation techniques such as sinusoidal or quadratic correlation might result in better coefficients for different power system and data contingencies. These more sophisticated and potentially informative correlation methodologies should be explored.

## VI. Conclusion

This research has provided initial steps toward facilitating use of live PMU data streams for decision-making and improved control of the power system via de-correlation and identification of various event signatures. We have built and analyzed the machinery for processing multiple synchrophasors distributed around an electrical network, and we have leveraged one year of real, archived PMU data to test this capability. Using a correlation algorithm we have shown that it is useful and feasible to isolate PMU data inconsistencies from power systems contingencies, as well as begin to identify patterns in various de-correlations for the purpose of specific contingency flagging. Advancing this functionality and coupling it with an efficient database back-end shows particular promise in using

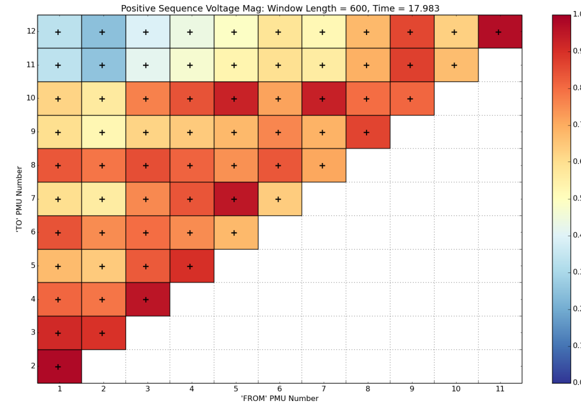


Fig. 12. Monrovia lightning event correlation over 10 sec. sliding window (electrical distance).

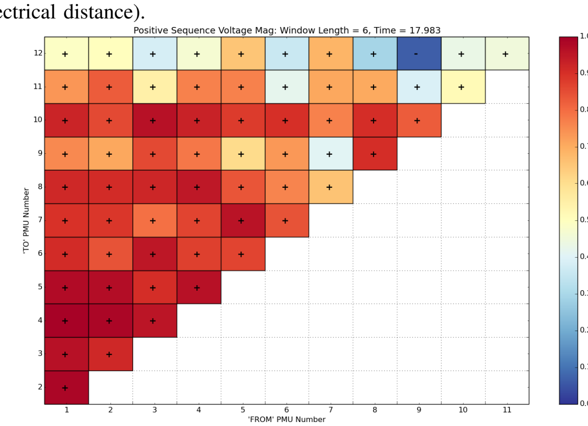


Fig. 13. Monrovia lightning contingency over a $\frac{1}{10}$ sec. sliding window (electrical distance).

a PMU network as an autonomous, wide area monitoring and control technique that will further grid operational awareness and decision support.

## Acknowledgment

The authors would like to thank the Bonneville Power Administration for providing archived PMU data and funding support. The authors also acknowledge Oregon BEST and their home institutions for funding support.

## References

[1] S. Horowitz, A. Phadke, and B. Renz, “The future of power transmission,” *IEEE Power and Energy Magazine*, vol. 8, no. 2, pp. 34–40, 2010.

[2] J. Chow, A. Chakrabortty, M. Arcak, B. Bhargava, and A. Salazar, “Synchronized phasor data based energy function analysis of dominant power transfer paths in large power systems,” *IEEE Transactions on Power Systems*, vol. 22, pp. 727–734, May 2007.

[3] T.-C. Lin, P.-Y. Lin, and C.-W. Liu, “An algorithm for locating faults in three-terminal multisection nonhomogeneous transmission lines using synchrophasor measurements,” *IEEE Transactions on Smart Grid*, vol. 5, pp. 38–50, Jan 2014.

[4] J. Ma, P. Zhang, H. jun Fu, B. Bo, and Z.-Y. Dong, “Application of phasor measurement unit on locating disturbance source for low-frequency oscillation,” *IEEE Transactions on Smart Grid*, vol. 1, pp. 340–346, Dec 2010.

[5] E. Cotilla-Sanchez, P. Hines, and C. Danforth, “Predicting critical transitions from time series synchrophasor data,” *Smart Grid, IEEE Transactions on*, vol. 3, pp. 1832–1840, Dec 2012.

[6] P. Ashton, G. Taylor, M. Irving, I. Pisica, A. Carter, and M. Bradley, “Novel application of detrended fluctuation analysis for state estimation using synchrophasor measurements,” *Power Systems, IEEE Transactions on*, vol. 28, pp. 1930–1938, May 2013.

[7] S. Das and T. Sidhu, “Detecting synchrophasors computed over fault/switching transients,” *IET Generation, Transmission & Distribution*, April 2014.

[8] G. Barchi, D. Macii, D. Belega, and D. Petri, “Performance of synchrophasor estimators in transient conditions: A comparative analysis,” *Instrumentation and Measurement, IEEE Transactions on*, vol. 62, pp. 2410–2418, Sept 2013.

[9] B. McCamish, D. Chiu, R. Meier, E. Cotilla-Sanchez, B. Bass, J. Landford, and M. Histand, “Managing pmu data sets with bitmap indexes,” in *IEEE Conference on Technologies for Sustainability*, (Portland, OR.), 2014.

[10] J. Runge, “On the graph-theoretical interpretation of pearson correlations in a multivariate process and a novel partial correlation measure,” pp. 1–20, Oct 2013.

[11] E. Cotilla-Sanchez, P. Hines, C. Barrows, S. Blumsack, and M. Patel, “Multi-attribute partitioning of power networks based on electrical distance,” *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4978–4987, 2013.