Pacific Northwest
NATIONAL LABORATORY

# Synchrophasor Analytics using Cloud Based Machine Learning Platform

April 16, 2019

**Pavel Etingov, Jason Hou, Huiying Ren, Heng Wang**

NASPI Work Group Meeting

U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

# ML project for PMU data analysis

- Project is supported by the DOE through the GMLC program
- Develop a framework for PMU big data analysis
  - Event detection
  - Anomaly detection
  - Improved situational awareness
  - System identification (learning system dynamic behavior)
  - Advanced visualization
- Framework is based on the cloud technology and distributed computing:
  - PNNL institutional cloud system or Microsoft Azure
  - Apache SPARK for distributed big data analysis and Machine Learning (ML)

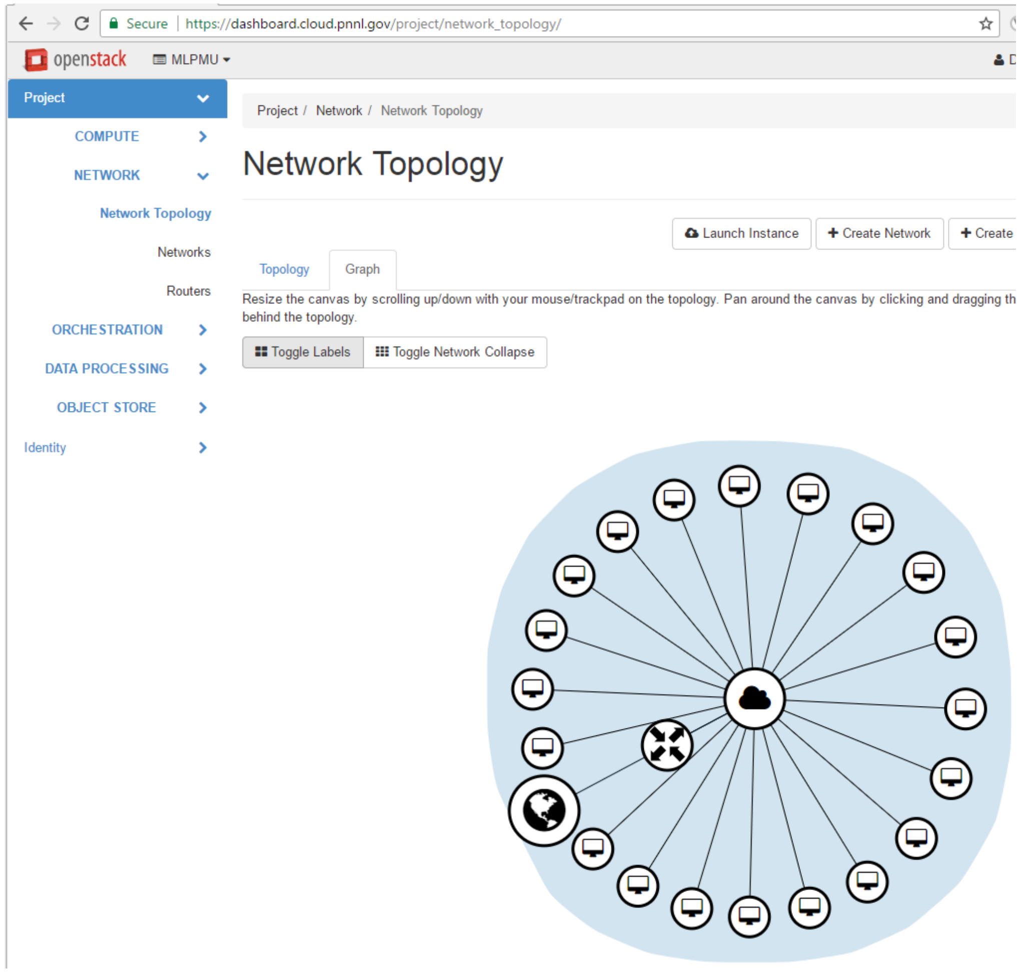
- PNNL
  - Jason Hou
  - Huiying Ren
  - Heng Wang
  - Troy Zuroske
  - Dimitri Zarzhitsky
  - Eric Andersen (PM)
  - Pavel Etingov

- Partners
  - LANL
  - LBNL
  - BPA

Pacific Northwest
NATIONAL LABORATORY

# PNNL Cloud Infrastructure

- PNNL cloud is based on OpenStack (a free and open-source software platform for cloud computing)
- Cloudera Apache Hadoop Distribution:
  - Apache Spark (an open-source cluster computing framework)
  - Apache Hive (a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis)
  - HBase (an open-source, non-relational, distributed database)

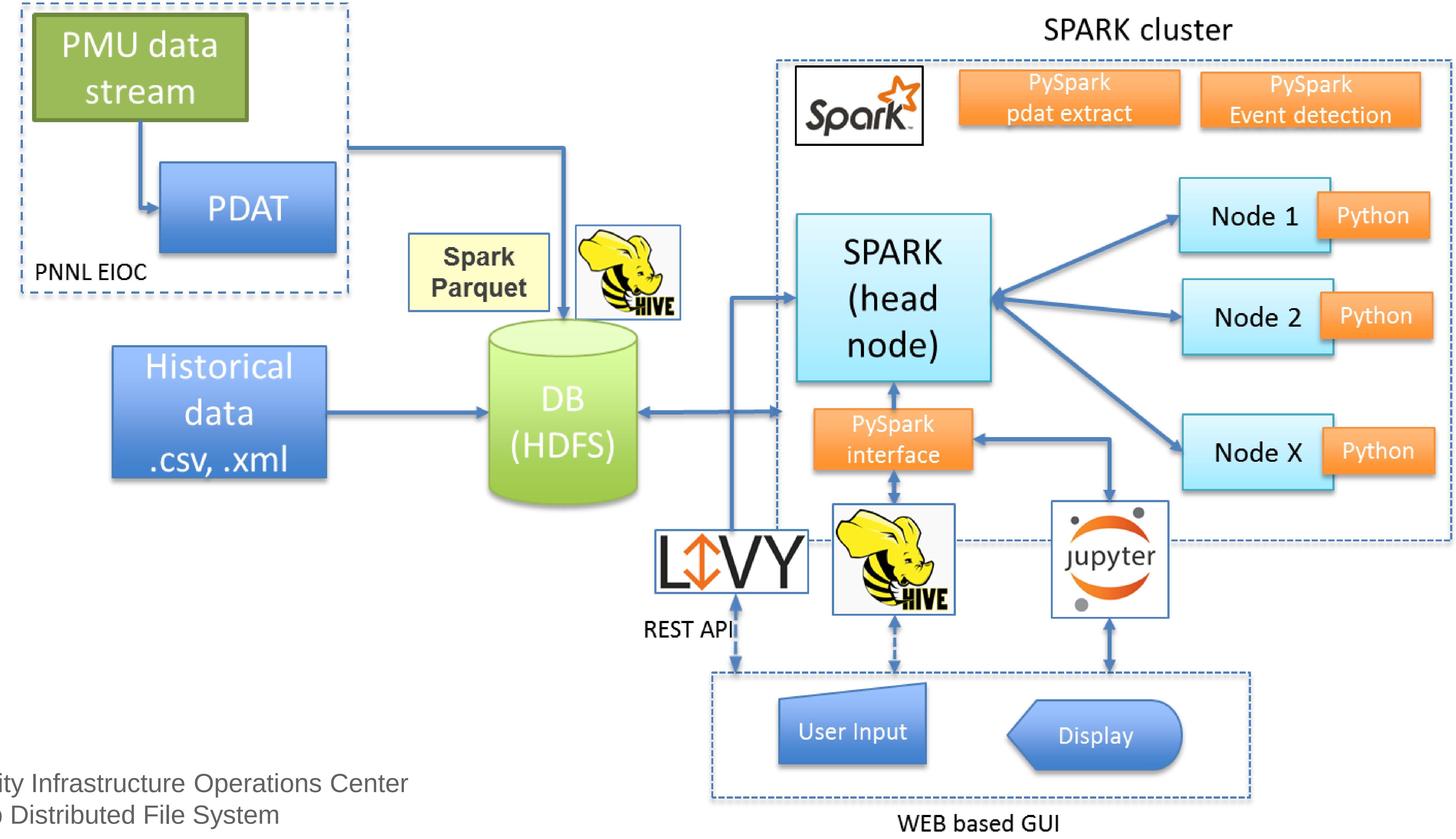



Spark research cluster
- 20 nodes
- RAM 512 Gb

# Apache Spark

- Large scale parallel data processing framework
- Extremely powerful (up to 100x faster than Hadoop)
- Large datasets distributed across multiple nodes within a computer cluster
- Support real time data stream
- Built-in Machine Learning library
- Support different languages (Scala, Java, Python, R)
- Support different data sources (SQL, Hive, HBase, Cassandra, Oracle, etc.)
- Open source and free
- Available through public cloud services (Amazon AWS, Microsoft Azure, IBM, etc.) and through new PNNL institutional cloud system.

# Cloud-based ML-PMU Framework




EIOC - Electricity Infrastructure Operations Center
HDFS- Hadoop Distributed File System

# PMU data stream

- PNNL receives PMU data stream from Bonneville Power Administration
  - 12 PMUs
  - Multiple channels (Voltage and Current Phasors, Frequency, ROCOF)
- PMU Data stored in PDAT format
  - PDAT format developed by BPA
  - Based on IEEE Std. C37.118.2-2011
  - Binary files
  - Each file contains 1 minute of data
  - One file ~ 5 MB

Data frame organization defined by IEEE C37.118.2

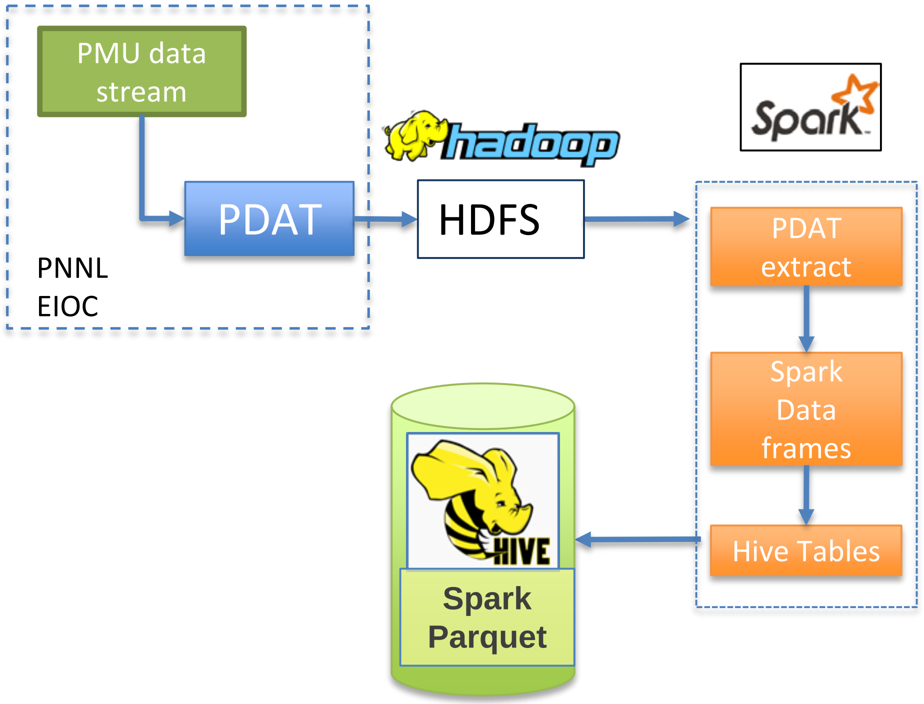
| No. | Field | Size (bytes) | Comment |
|---|---|---|---|
| 1 | SYNC | 2 | Sync byte followed by frame type and version number. |
| 2 | FRAMESIZE | 2 | Number of bytes in frame, defined in 6.2. |
| 3 | IDCODE | 2 | Stream source ID number, 16-bit integer, defined in 6.2. |
| 4 | SOC | 4 | SOC time stamp, defined in 6.2, for all measurements in frame. |
| 5 | FRACSEC | 4 | Fraction of Second and Time Quality, defined in 6.2, for all measurements in frame. |
| 6 | STAT | 2 | Bit-mapped flags. |
| 7 | PHASORS | 4 × PHNMR or 8 × PHNMR | Phasor estimates. May be single phase or 3-phase positive, negative, or zero sequence. Four or 8 bytes each depending on the fixed 16-bit or floating-point format used, as indicated by the FORMAT field in the configuration frame. The number of values is determined by the PHNMR field in configuration 1, 2, and 3 frames. |
| 8 | FREQ | 2 / 4 | Frequency (fixed or floating point). |
| 9 | DFREQ | 2 / 4 | ROCOF (fixed or floating point). |
| 10 | ANALOG | 2 × ANNMR or 4 × ANNMR | Analog data, 2 or 4 bytes per value depending on fixed or floating-point format used, as indicated by the FORMAT field in configuration 1, 2, and 3 frames. The number of values is determined by the ANNMR field in configuration 1, 2, and 3 frames. |
| 11 | DIGITAL | 2 × DGNMR | Digital data, usually representing 16 digital status points (channels). The number of values is determined by the DGNMR field in configuration 1, 2, and 3 frames. |
|  | *Repeat 6–11* |  | Fields 6–11 are repeated for as many PMUs as in NUM_PMU field in configuration frame. |
| 12+ | CHK | 2 | CRC-CCITT |

# Ongoing work

- Python (PySpark) modules:
  - PDAT data extraction
  - Data processing
    - Bad data
    - Missing points
    - Outliers
  - Event detection and classification
    - Frequency events
    - Voltage events
  - Feature extraction and analysis
    - Wavelet decomposition
    - State space models
    - Principal component analysis
    - Recurrent neural network

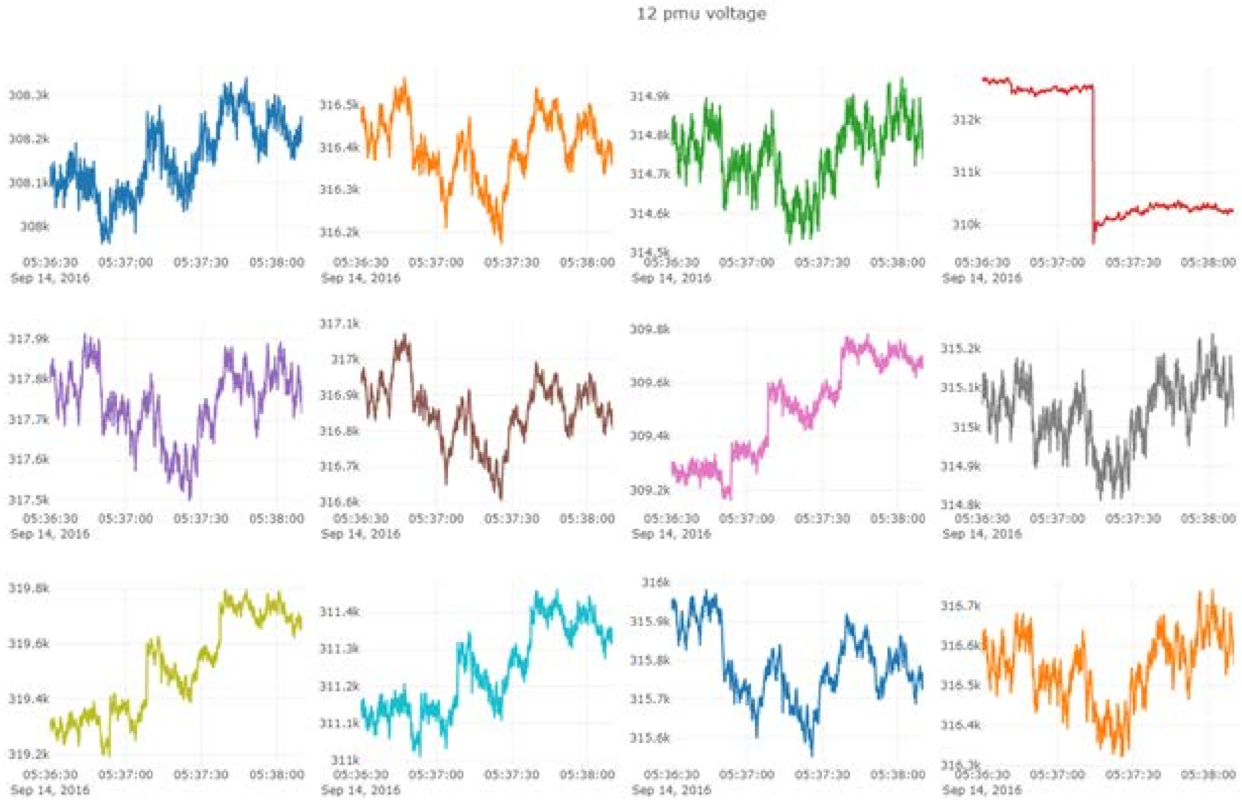
# PDAT data extraction

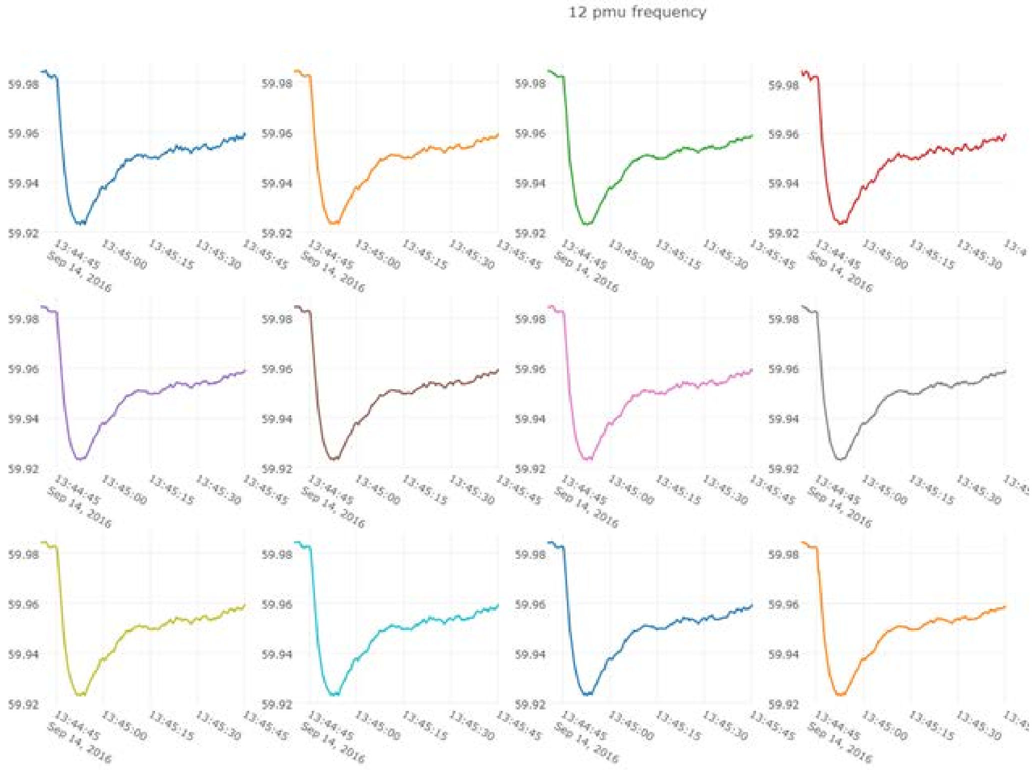
- Read information from PDAT and creates SPARK data frames
- Store information in Hive or Parquet tables
- Implemented in PySpark that allows parallel processing of multiple PDAT files
- Significantly increased performance
  - To read information for 1 hour takes about 20 seconds (20 nodes cluster)



# Event Detection

- Validation data consist of user-specified **delta frequency** and **event duration**
- Cross validation helps determine the optimal thresholds to reduce/avoid false alarms
- Spark usage significantly increases the computational throughput of the application
- Processing of 1 day data takes about 5-7 minutes (processing the same dataset using a PC takes about 1 hour)



# Machine Learning Algorithms for Offline and Online PMU Anomaly Detection

- Offline events detection: adopt wavelet decomposition to examine the PMU signals with multi-resolution analysis (MRA). The events can be detected at multiple temporal scales. ***Pros: yields high detection rate at multiple resolutions; Cons: requires long time period of data***
- Online events detection: learn the historical patterns of PMU signals and then predict for the future
  - Dynamic Linear Model (DLM): one of state-space models. ***Pros: fast forecasting with relatively short input time series; Cons: forecasts focus only on the near-term behaviors.***
  - Long short-term memory (LSTM): one of deep learning Recurrent Neural Networks (RNN). ***Pros: forecasts have high accuracy for relatively long time windows; Cons: needs long time period for training and may be computationally expensive***

# Wavelet-based MRA Event Detection Framework

PMU Signals → Wavelet-based multi-resolution analysis (MRA) → Wavelet coefficients extraction → Moving window based anomaly candidates → Anomaly scoring → Spatiotemporal correlation analysis → Final list of detected anomalies

**Wavelet-based multi-resolution analysis (MRA)**

Signals are transformed into time-frequency domain

MRA decomposes signals into approximation (A) and detail (D) components

**Wavelet coefficients extraction**

Extract the detail wavelet coefficients at scales matching the durations of events

**Moving window based anomaly candidates**

Identify anomaly candidates at multiresolution levels with moving-window-based anomaly detection

5-min bandwidth of moving window

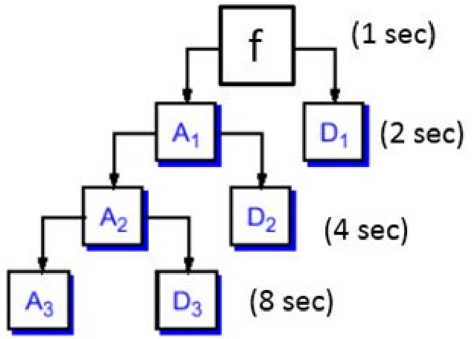
**Anomaly scoring**

Anomalous score is set to be 1 when an anomaly is detected at each resolution level.

The anomaly score matrices are the summation of scores at multi-resolution levels across all PMUs

**Spatiotemporal correlation analysis**

False alarms or localized events corresponds to relatively weak spatiotemporal correlations

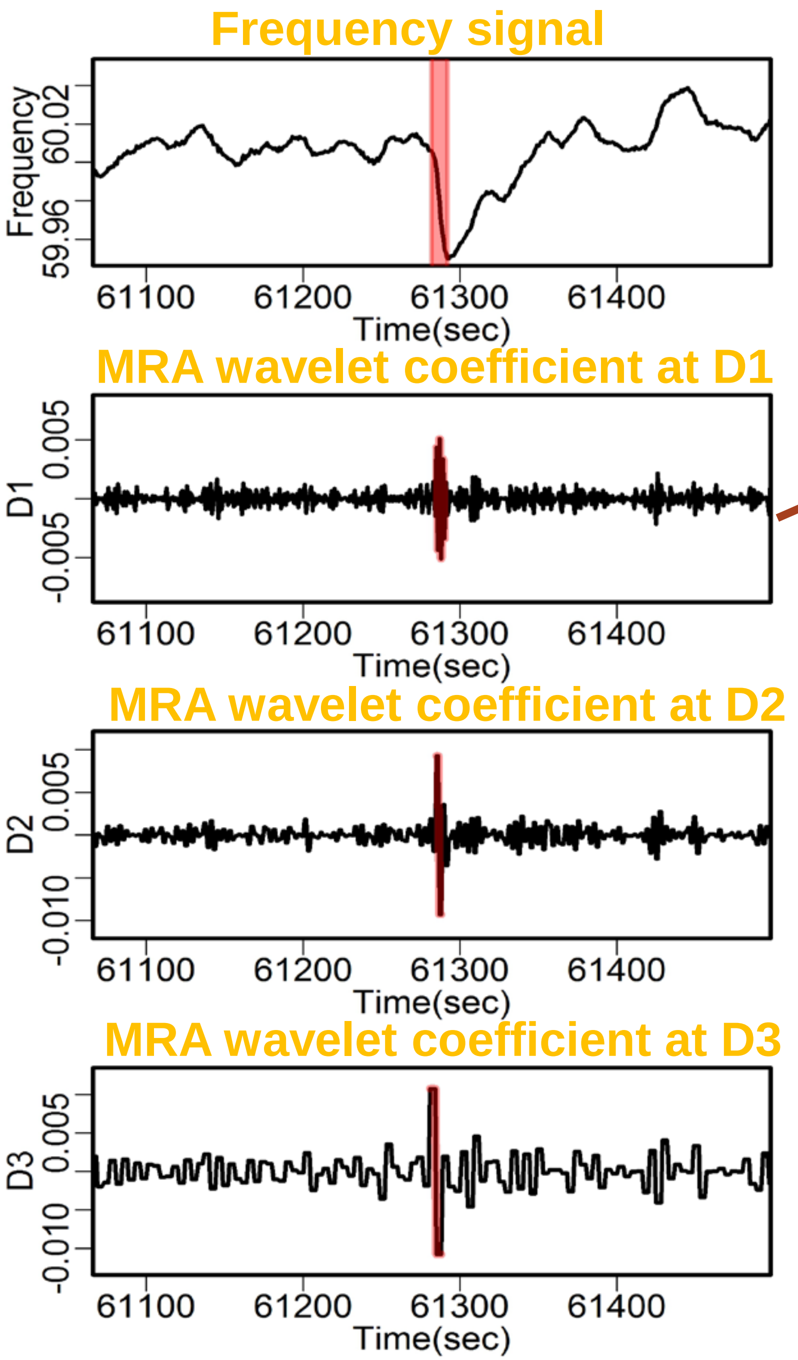


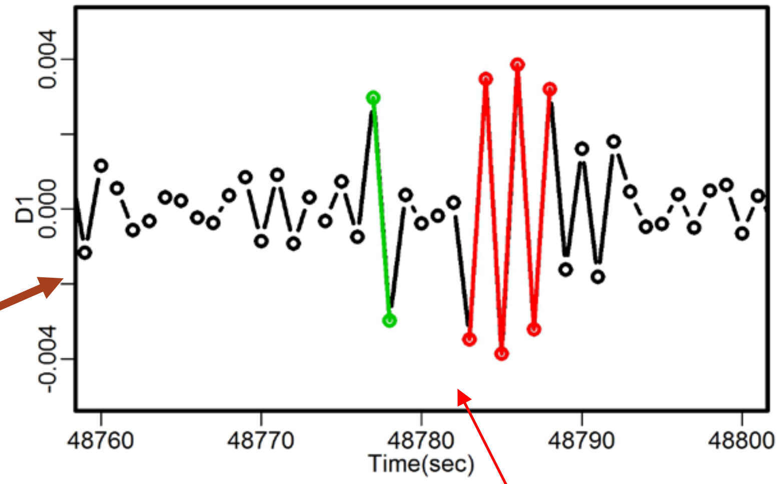

MRA decomposition Tree

# Assign Anomaly Scores on Decomposed Details

- The anomaly score matrices were calculated across 12 PMUs at multi-resolution levels for each PMU attribute.

- Red vertical lines correspond to a historical recorded event at multi-resolution levels

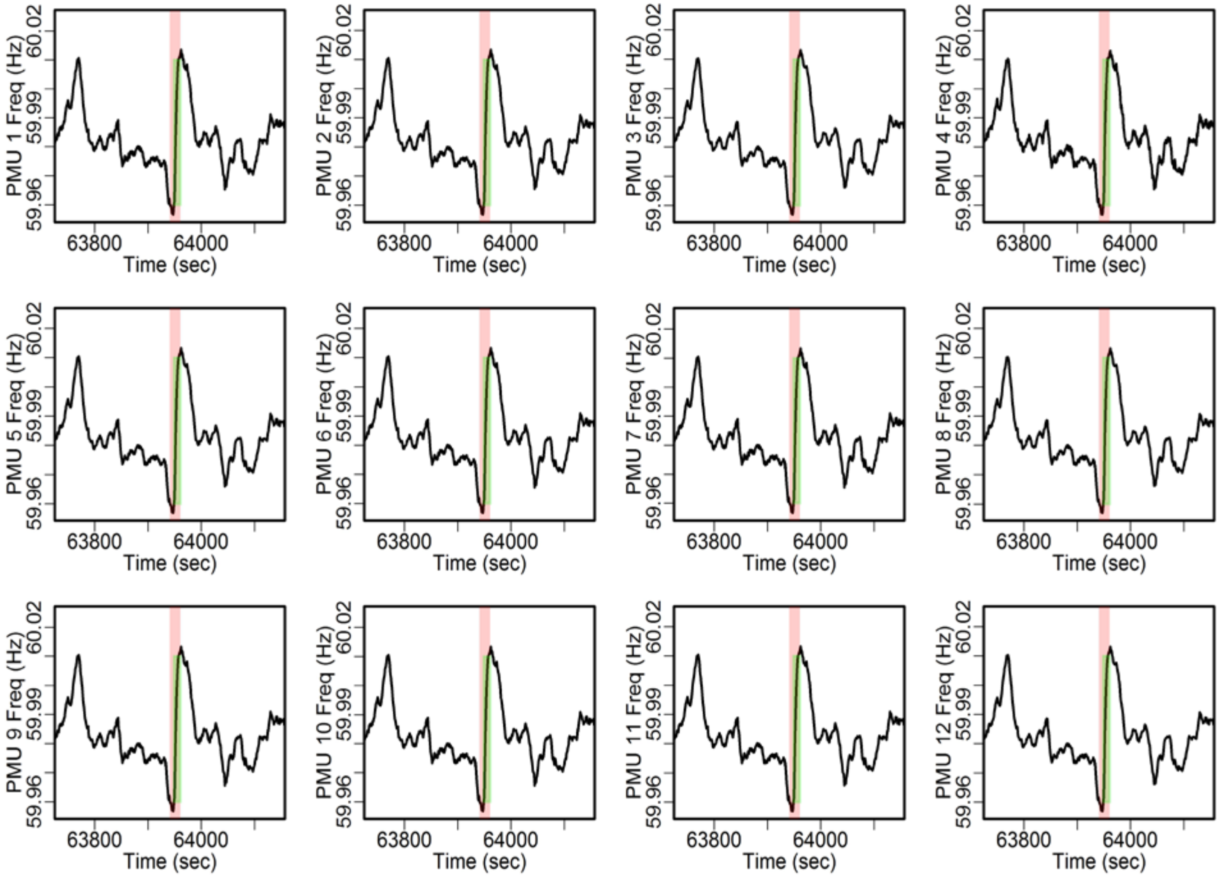


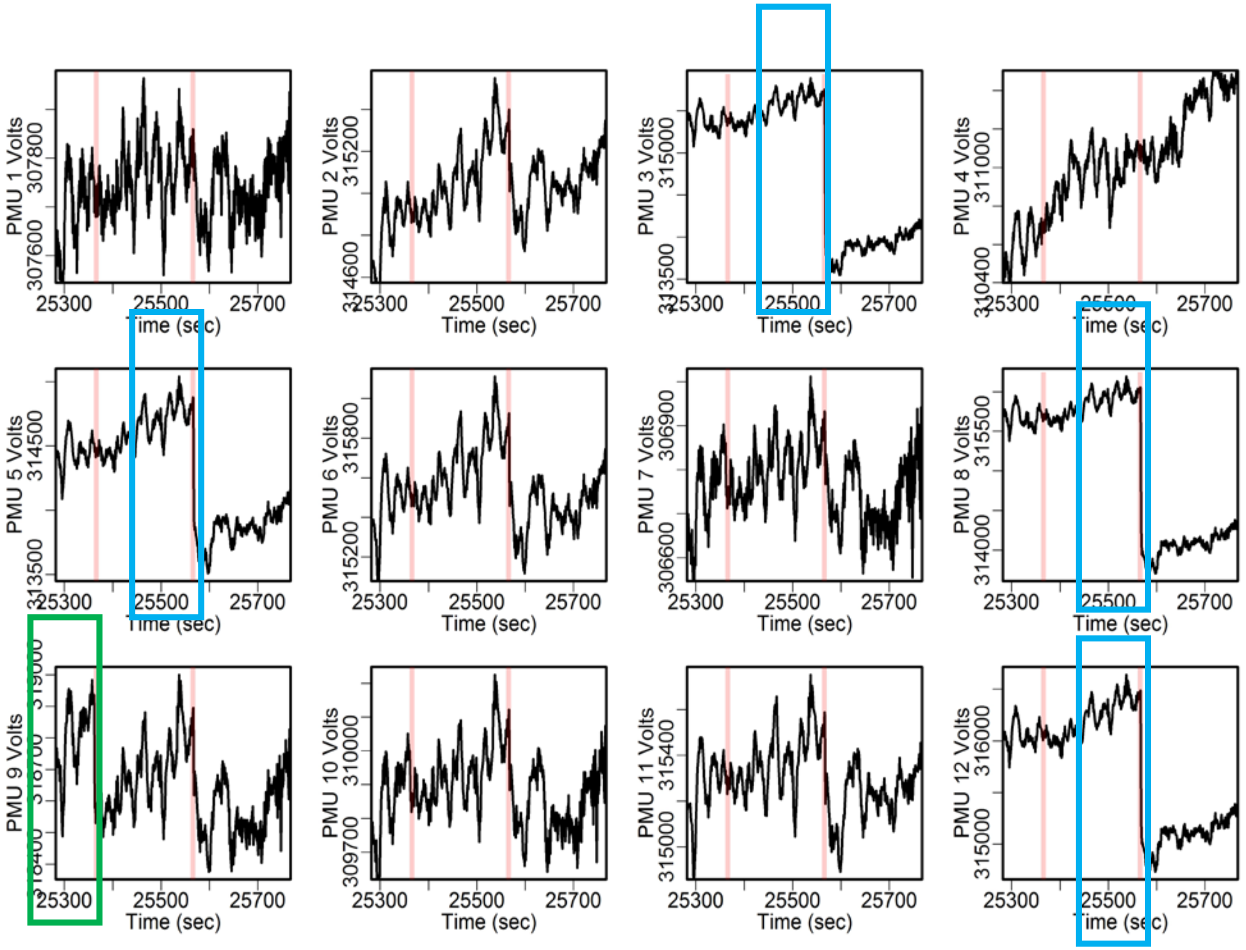

More than 3 sequential points exceeded the threshold and counted as an event. +1 added to the anomaly score matrices.

# Examples of MRA Detected Anomalies

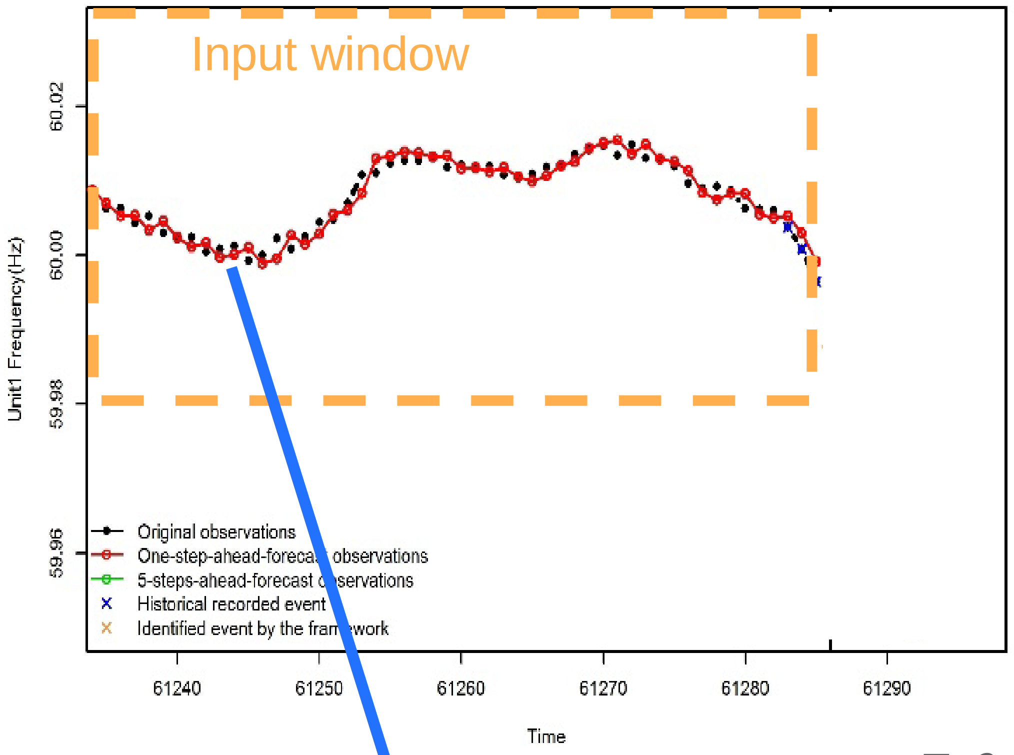
An example of detected system-wide anomaly (frequency signals) where the PMUs have consistent behaviors and strong cross-correlations.



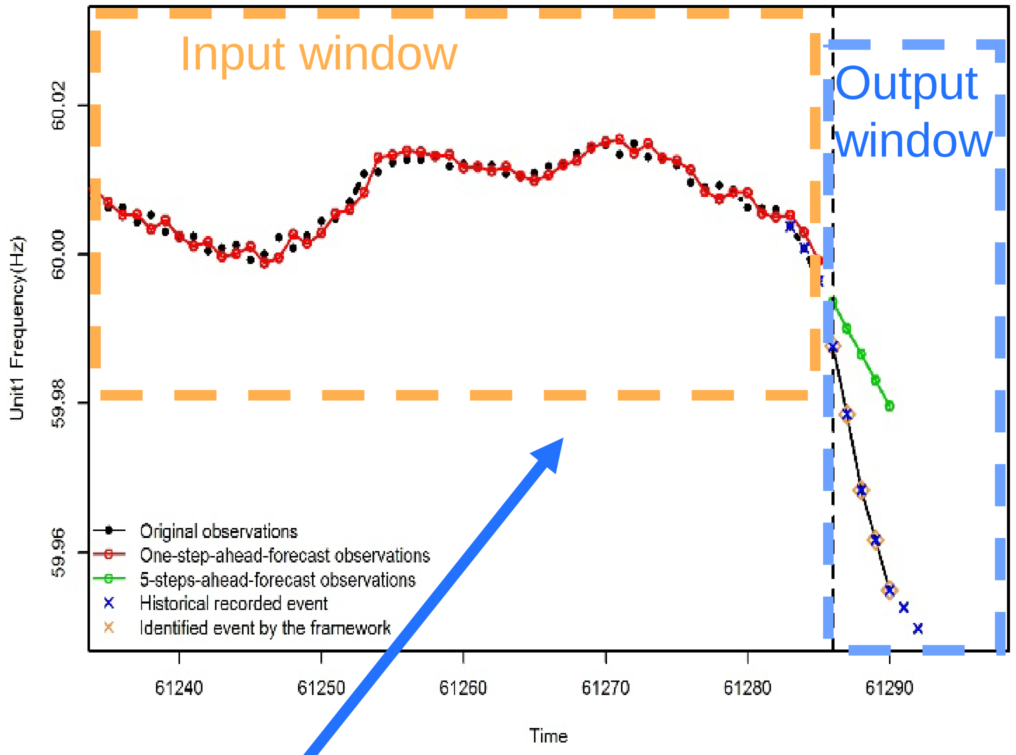

An example of detected local anomaly (voltage signals)



# DLM yields accurate forecasts in short time windows




DLM




Pure data-driven dynamic linear model training using Kalman Filter

$$y_t = F_t\theta_t + \vartheta_t, \qquad \vartheta_t \sim N(0, V_t)$$
$$\theta_t = G_t\theta_{t-1} + w_t, w_t \sim N(0, W_t)$$

$$\begin{array}{ccccccc} \theta_t & \longrightarrow & \theta_{t+1} & \longrightarrow & \cdots & \longrightarrow & \theta_{t+k} \\ | & & & & & & | \\ (Y_1, \ldots, Y_t) & & & & & & Y_{t+k} \end{array}$$

Predict 'normal' system behaviors by learning historical patterns.

The cumulative probability distribution (CDF) of prediction errors:

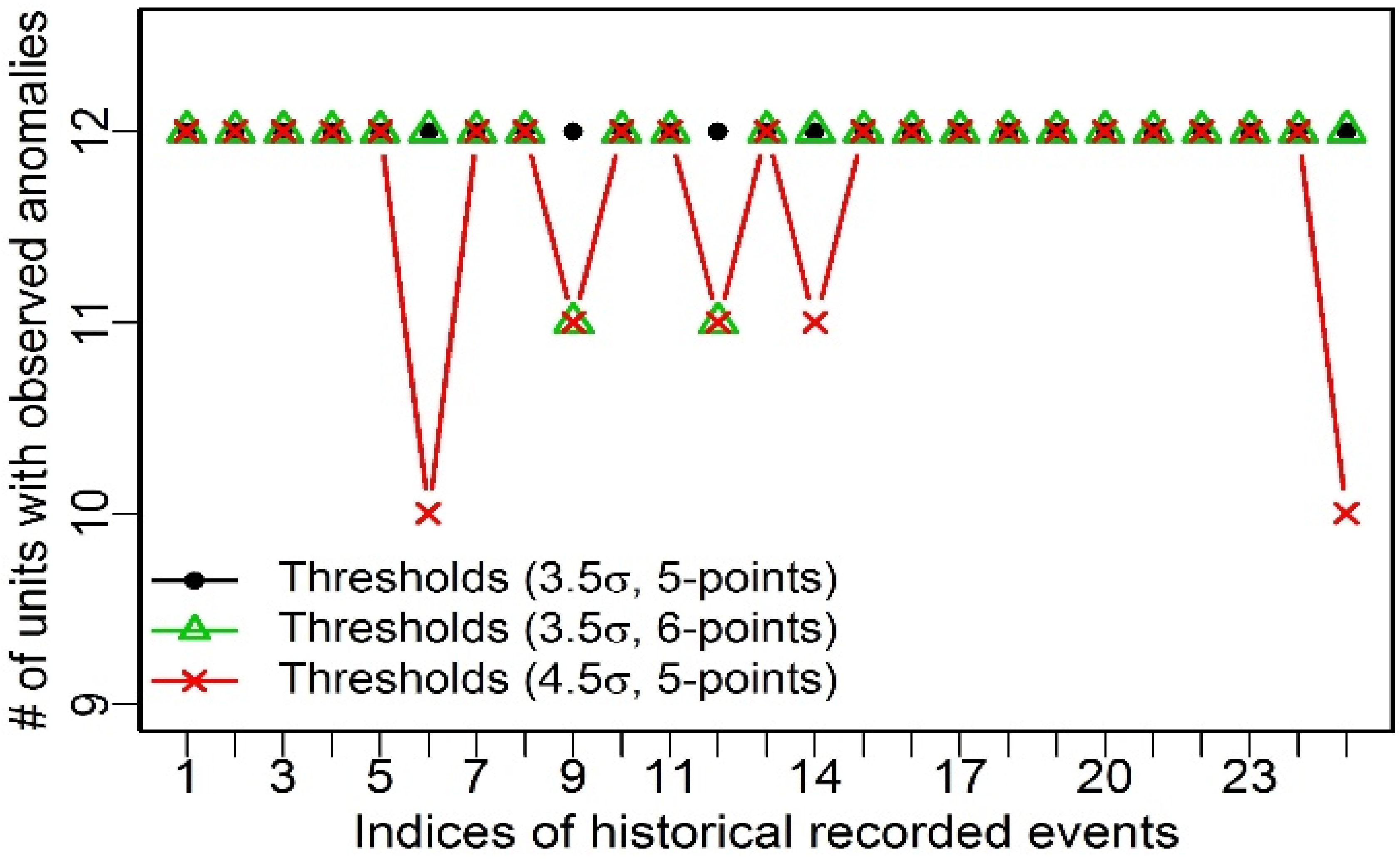
$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$$

The exceedance probability of a prediction error is then computed as:

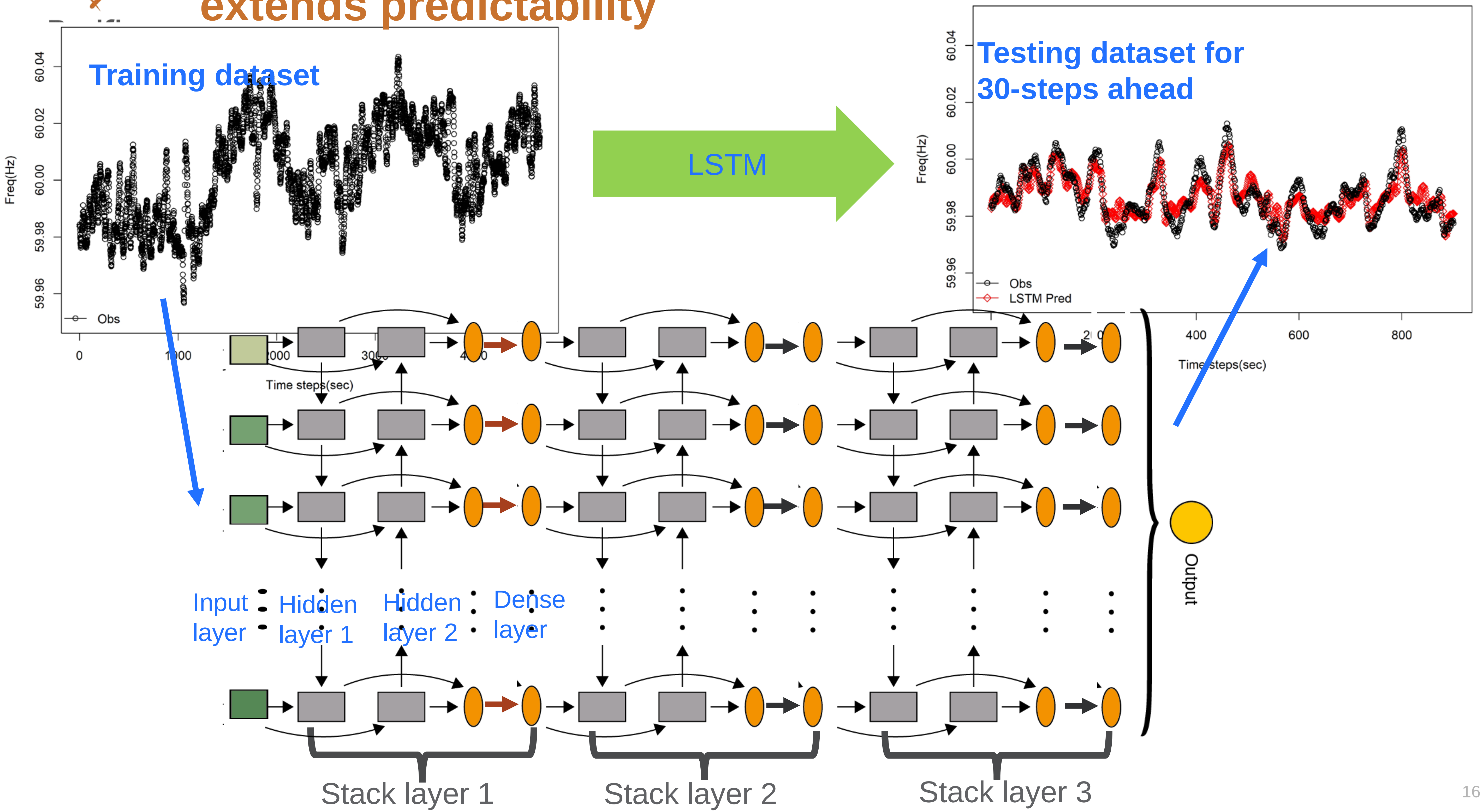
$$P_i(X \leq x) = \max(P_i(X \leq x), 1 - P_i(X \leq x))$$

# Criteria/Thresholds to Confirm an Event

- Threshold for exceedance probability: the prediction error is beyond X times of the corresponding standard deviation σ
- Threshold for duration: sequential points need to pass the screening in order to confirm an event.

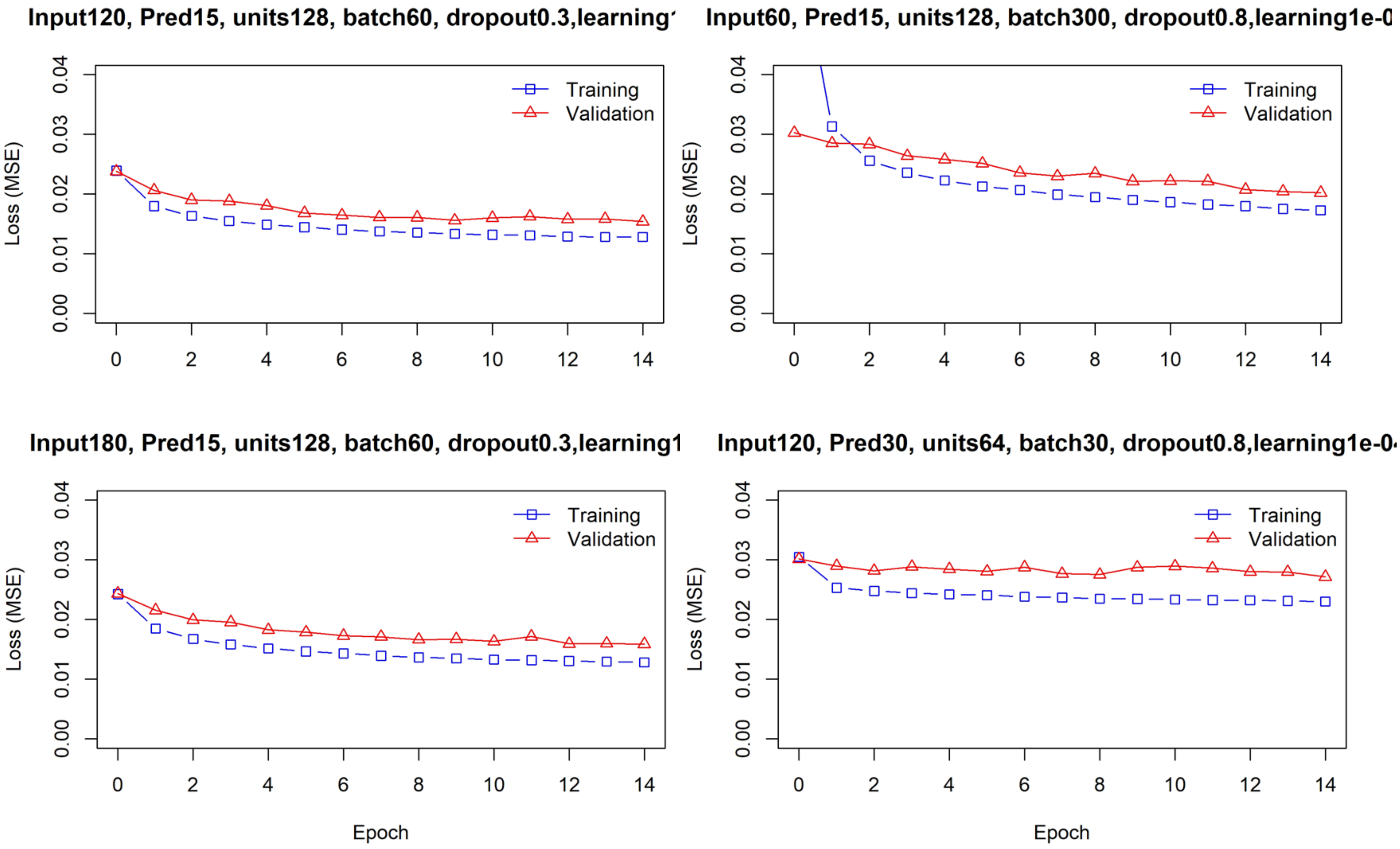


# LSTM extracts both short- and long-term patterns and extends predictability



# LSTM Model Evaluation: Training and Validation Loss

- Training data: 70%
- Validation data: 15%
- Testing data: 15%
- Loss function: Mean squared error (MSE)
- Model parameters: Input/output window, units, batch size, dropout rate

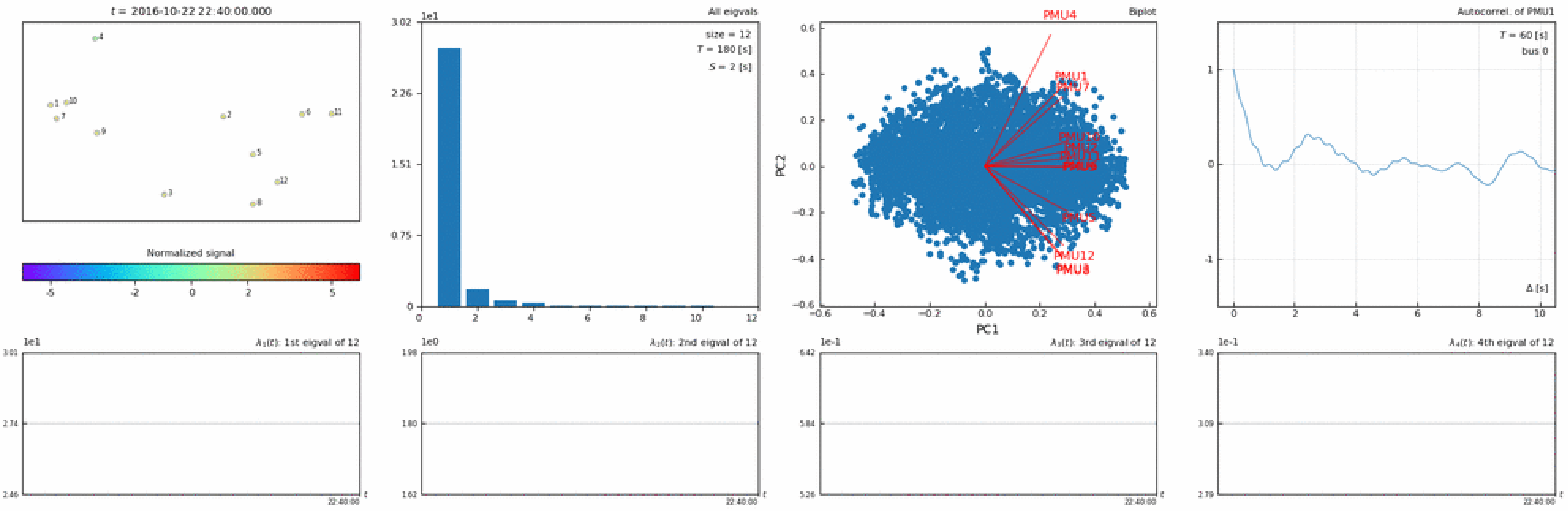


# Understanding the Spatial and Temporal Patterns in PMU Signals

- In additional to anomaly detection and classification practices, more data analytics can be used to understand the spatiotemporal behaviors of the PMU signals and the mechanisms
  - Block Principal Component Analysis of PMU attributes
  - Auto- and cross-correlation Analysis of PMU attributes
  - Taylor Diagram across hours, days, seasons
  - Spectra analysis and anomaly matching of ‘collocated’ PMU and weather attributes

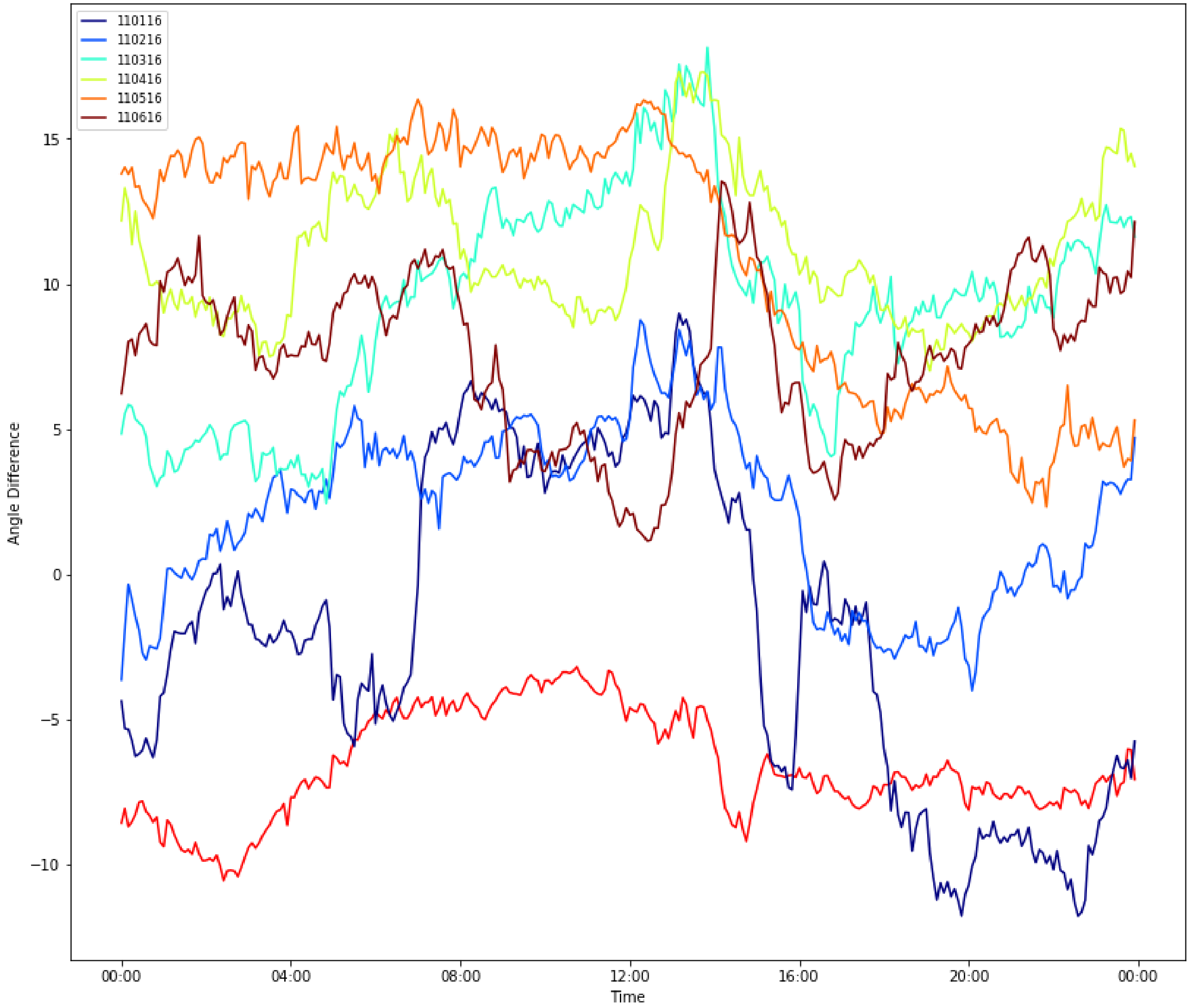
# Block Auto-correlation and Principal Component Analysis to Monitor Temporal Patterns in PMU attributes

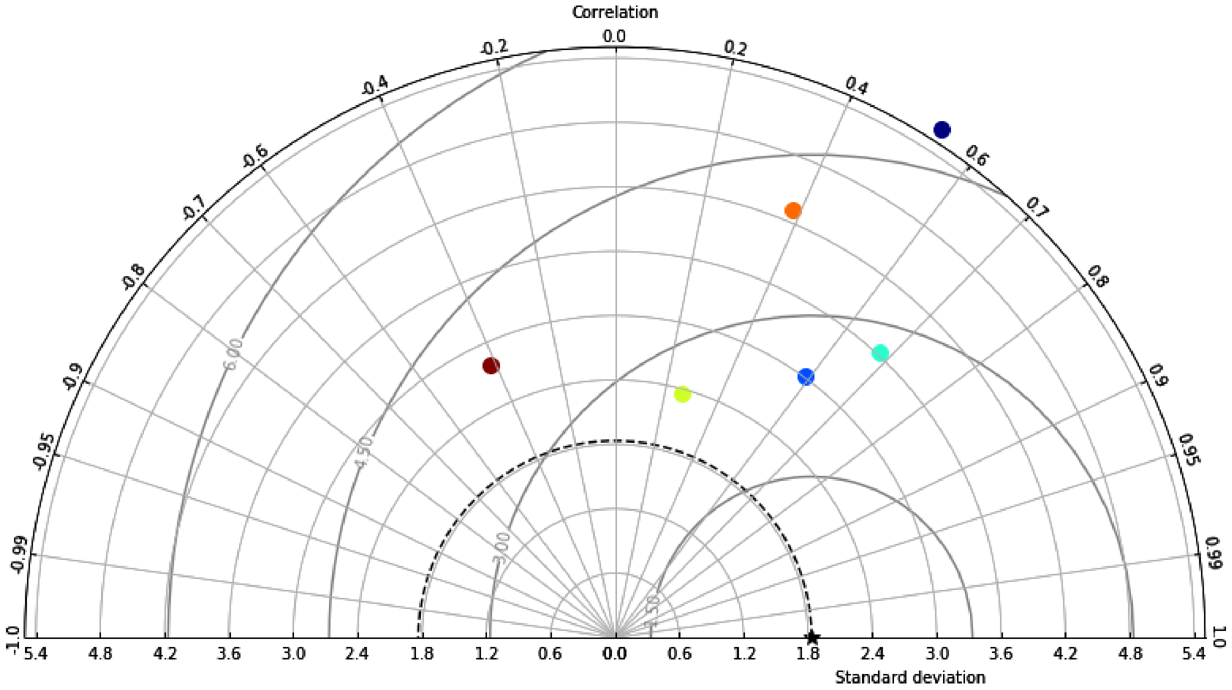
- Video demonstration



# Taylor Diagrams for Evaluating Temporal Similarities

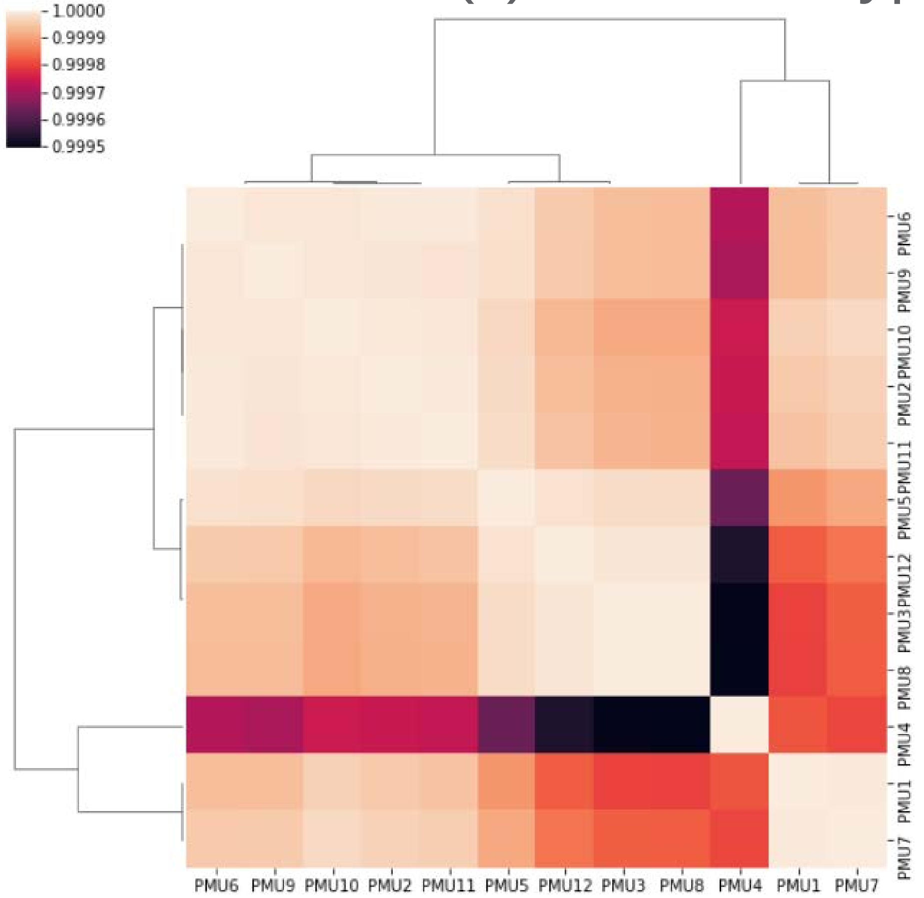
- Taylor Diagrams help identify similarities (in both absolute magnitudes and patterns) across hours, days, and seasons

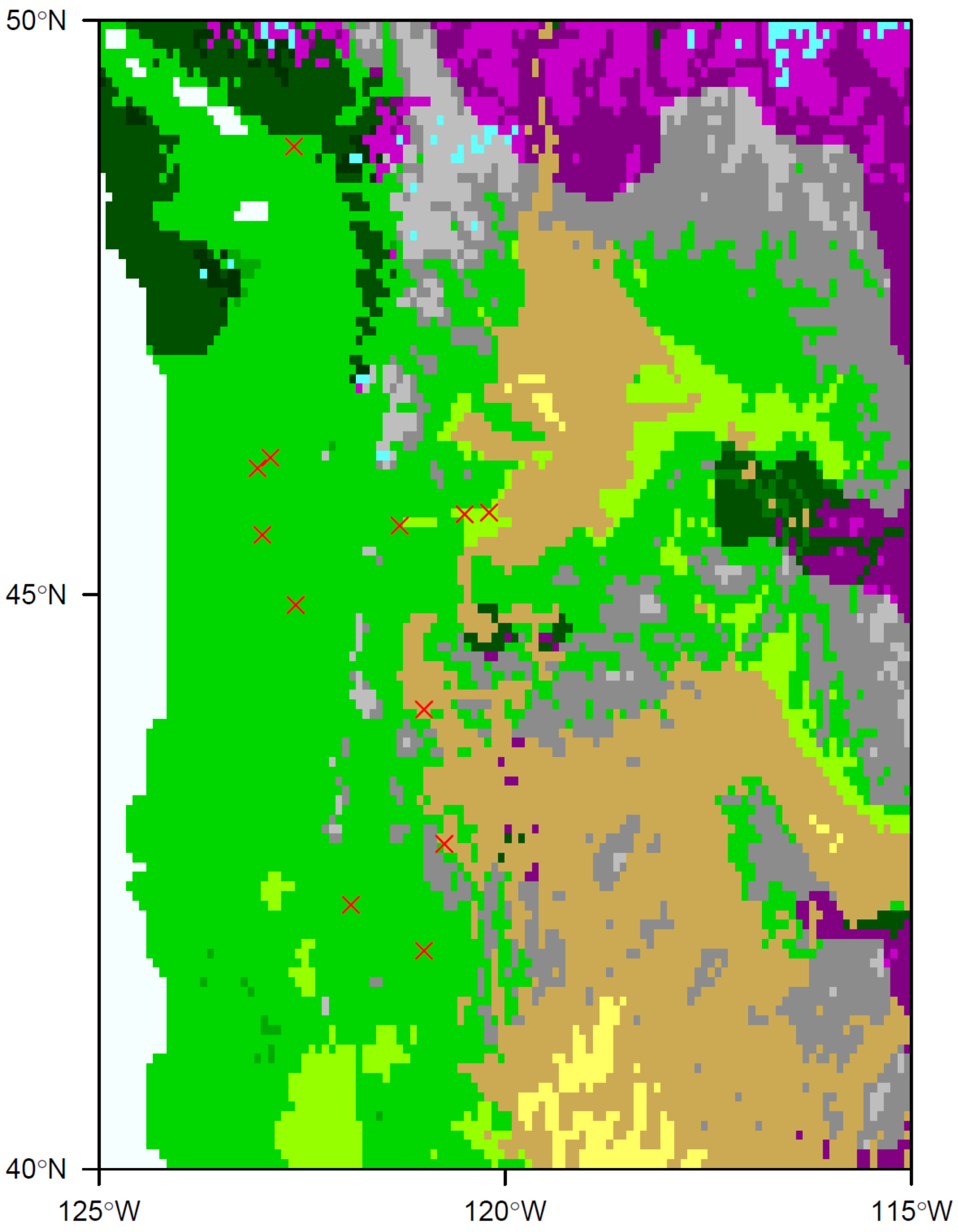


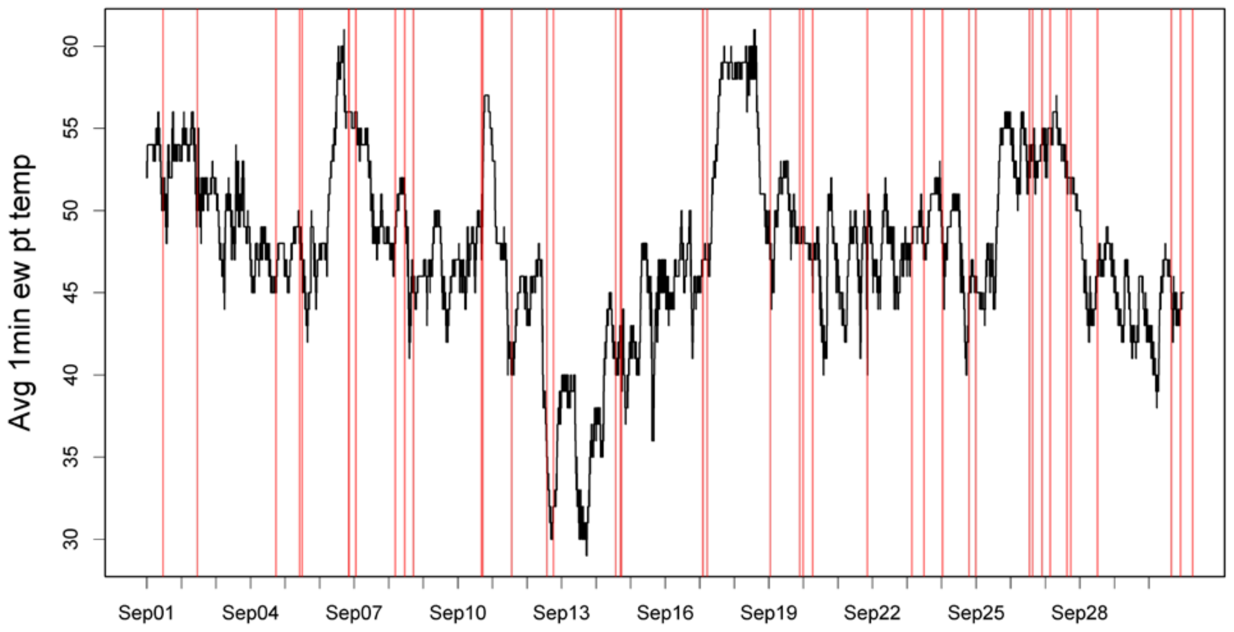
# Cross-Dependence Analysis of Spatial Patterns in PMU Attributes

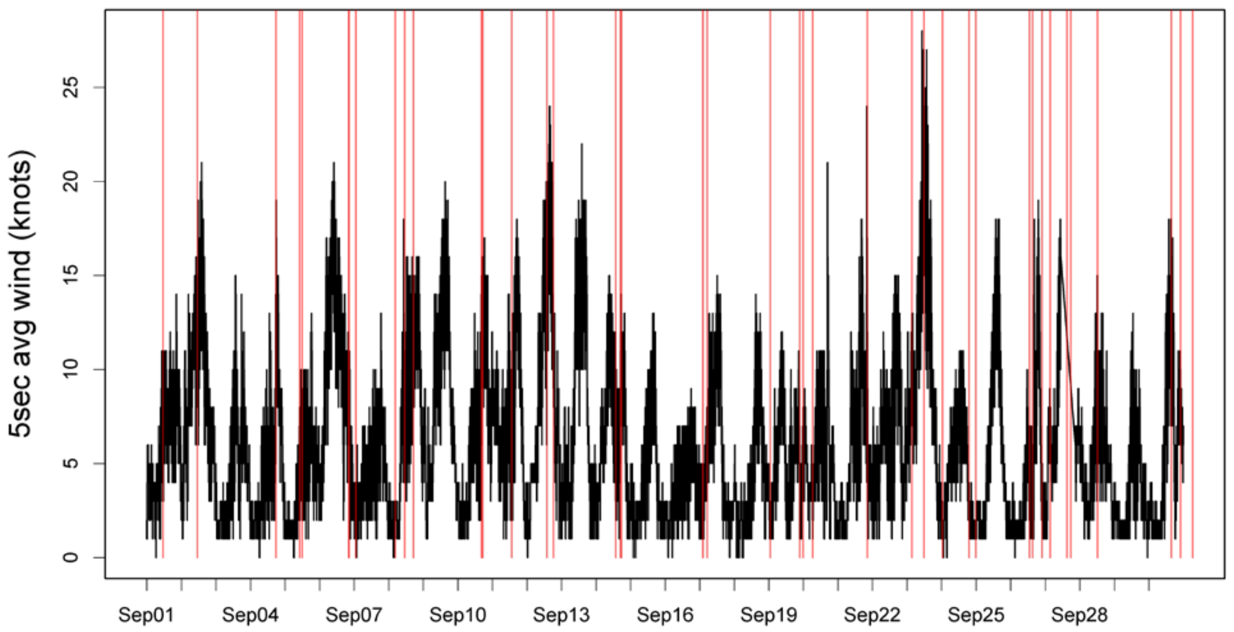
- ❑ Cross-correlation among 12 PMUs Oct 22nd, 2016 in frequency data
- ❑ The statistical clusters are found to be geographically clustered
- ❑ The anomalies within a cluster can be attributed to similar weather factor(s) or climate types



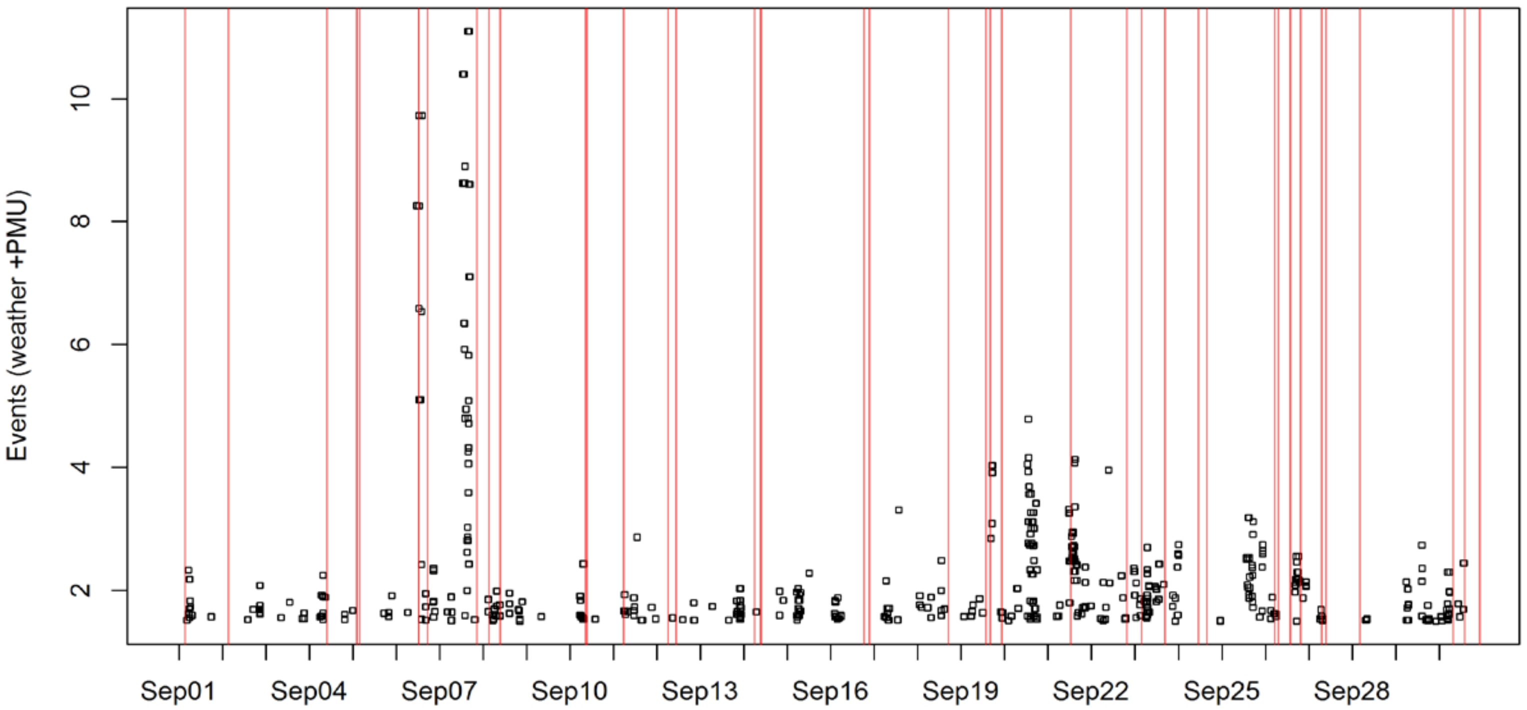
# Extreme Weather Drives PMU Events



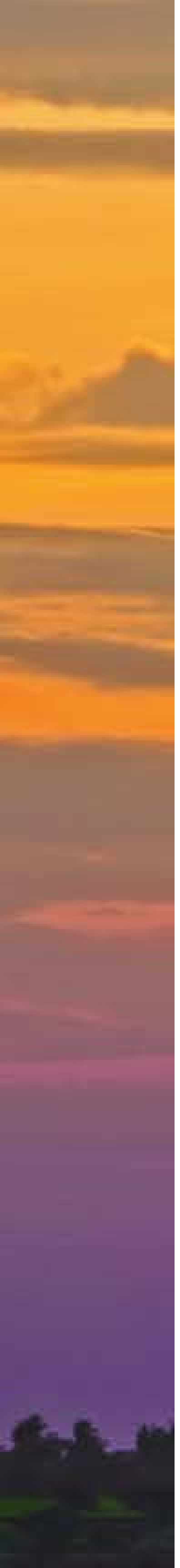

- The red vertical lines are the historical recorded events
- The black lines are the weather attributes (e.g., dew point, wind speed, precipitation)

Spectra analysis and anomaly matching of ‘collocated’ PMU and weather attributes



# Conclusions and Future Plans

- Spark cluster for ML and PMU (big data) analysis was deployed. It is based on the PNNL institution cloud system.
- PMU data have been collected and archived in PDAT format (PMU data stream from PBA to PNNL EIOC).
- Methodologies for both online and offline anomaly detection have been developed.
- Python (PySpark) modules are under development, with the following functions:
  - PDAT data extraction and preparation;
  - Event detection and classification with multiple resolution analysis, state space models, and deep recurrent neural networks;
  - Evaluation of spatial and temporal behaviors and identification of the potential driver.



Pacific Northwest
NATIONAL LABORATORY

# Thank you


Pavel.Etingov@pnnl.gov
Zhangshuan.Hou@pnnl.gov
Huiying.Ren@pnnl.gov
Heng.Wang@pnnl.gov