






# Synchrophasor Data Compression Under Disturbance Conditions via Cross-Entropy-Based Singular Value Decomposition

Weikang Wang , *Student Member, IEEE*, Chang Chen , Wenxuan Yao , *Member, IEEE*, Kaiqi Sun , *Member, IEEE*, Wei Qiu , and Yilu Liu , *Fellow, IEEE*

**Abstract**—The increasing deployment of phasor measurement units and the advances of their reporting rates are challenging the present data centers in terms of storing and analyzing large-volume data. Under power system disturbance conditions, it is difficult to retain critical information while compressing the synchrophasor data effectively. This article combines the cross entropy and the singular value decomposition, proposing a novel model to compress the synchrophasor data to an extremely small size yet keep superior accuracy. The proposed model is extensively tested and compared with the state-of-the-art algorithms using the simulated and the FNET/GridEye field-collected data. The result indicates that the proposed algorithm has superior performance in compressing the data while retaining critical information under disturbance conditions.

**Index Terms**—Cross-entropy, data compression, phasor measurement unit (PMU), singular value decomposition (SVD), synchrophasor.

## I. INTRODUCTION

PHASOR measurement units (PMUs) have been increasingly deployed in the past decades since their invention, which overperform traditional supervisory control and data

acquisition system (SCADA), thanks to their high reporting rates, rich measurement types, and high accuracy. These features enable many advanced applications that help ensure the reliability of power grids [1]–[4]. Typically, a PMU collects phasors including voltage magnitude, voltage angle, current magnitude, current angle, frequency, etc. A typical PMU collects GPS-synchronized phasors and streams them to a phasor data concentrator (PDC) at 10–120 Hz reporting rate [4], [5]. On the other hand, grid structures nowadays are complex [6]–[10], which require more and more PMUs to cover the transmission system. For example, according to [11], to cover the transmission system in the USA, more than 1100 PMUs are required. Clearly, the high reporting rate and the large number of PMUs will result in a huge amount of data. For example, assuming there are 1100 PMUs reporting data at 30 Hz via the IEEE C37.118 protocol [12], over 700 gigabytes (GB) data will be generated per day. Furthermore, using advanced PMUs, which report at 120 Hz, the total data volume can exceed 2.8 terabytes (TB) per day. Realizing the challenge from the large-volume PMU data, data compression techniques need to be exploited to efficiently compress [13] and store the data.

In general, compression techniques can be categorized into lossy and lossless approaches [14]. Lossless compression focuses on exploring the statistics of the data and using efficient bit-wise encoding techniques to compress it. Lossless compression allows compressed data to be compressed with no information loss. Comprehensive comparisons are conducted among well-known lossless compression models including Deflate, Bzip2, Lempel-Ziv 77 (LZ77), Lempel-Ziv-Markov-algorithm, and the Szip [15], [16]. These works imply using the Szip model may achieve the best compression performance for synchrophasors. However, lossless compression methods can hardly reach a high compression ratio (CR) since the dimension of the synchrophasor data is ignored.

On the contrary, lossy compression emphasizes trading controllable errors for a better CR. For synchrophasor data compression, the lossy compression models mainly rely on two philosophies. The first and most straightforward way is to compress the data by analyzing each PMU independently. Models such as discrete wavelet transformation (DWT) [17], [18], improved

Manuscript received May 24, 2020; accepted June 23, 2020. Date of publication June 29, 2020; date of current version January 4, 2021. This work was supported in part by the Engineering Research Center Shared Facilities supported by the Engineering Research Center Program of the National Science Foundation and DOE under NSF Award EEC-1041877, in part by CURENT Industry Partnership Program, and in part by the U.S. Department of Energy, Advanced Grid Modeling (AGM) program. Paper no. TII-20-2599. (*Corresponding author: Wenxuan Yao.*)

Weikang Wang, Chang Chen, Kaiqi Sun, and Wei Qiu are with the Department of Electrical Engineering and Computer Science, The University of Tennessee, Knoxville, TN 37996 USA (e-mail: wwang72@utk.edu; cchen75@utk.edu; ksun8@utk.edu; qwei4@utk.edu).

Wenxuan Yao is with Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA (e-mail: yaow1@ornl.gov).

Yilu Liu is with the Department of Electrical Engineering and Computer Science, The University of Tennessee, Knoxville, TN 37996 USA, and also with Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA (e-mail: liu@utk.edu).

Color versions of one or more of the figures in this article are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TII.2020.3005414

DWT [19]–[21], exception compression-swing door trending (SDT) [22], etc., are proposed. Among these models, the SDT method can achieve good CR with small normalized mean square error. Another way to compress the data exploits the linearity in the synchrophasor data. Within an interconnected power grid, synchrophasors may contain high linearity. For example, a principal component analysis (PCA) based model is proposed to use dimensionality reduction to perform early event detection [23]. This work lays the basis of using dimensionality reduction approaches to analyze PMUs' data in the modern smart grid and it implies its potential to compress the PMU data as well. Toward this end, a statistical change detection (SCD)-PCA-DWT/DCT model is proposed to compress the synchrophasors [24]. In this work, the PCA compression achieves good performance on various measurements. Similarly, another PCA-based algorithm is proposed to compress the distribution-level synchrophasors [25]. In this work, the data can be compressed at good CRs with controllable errors. Recently, a multiscale PCA model is proposed to decide different parameters for the PCA model. It first performs a spatial-wise cluster analysis, then uses different PCA models to compress corresponding clusters. This work can achieve good CR under ambient conditions and acceptable CR under generator trip conditions. However, this model was not extensively tested with disturbance events such as oscillations, and interarea oscillations, where the clusters can have a similar density.

Though it seems promising, there are several issues in using PCA to compress the synchrophasor data. The first one is the choice of the principal components or the compression space  $R$ . Some have proposed to separate the data into ambient condition and fault conditions and decide the compression space  $R$ , respectively. However, the choice of principal components is still not well defined. In [24], 80% static normalized cumulative variance (NCV) is used under ambient conditions, while 95% NCV is used under fault conditions. A similar score selection philosophy is implemented in another work [25]. This rather static rule has a chance of losing important information during data compression, even if the RMSE seems acceptable. As shown in Fig. 1, when the data is compressed and reconstructed under 99.0% score, there is still a significant information loss. For the reconstructed data, the maximum frequency is reduced from above 60.25 to 60.08 Hz, while the minimum frequency is elevated from 59.80 to 59.88 Hz. This error can result in inaccurate frequency response assessments, which are required in standard BAL-003 [26] by North American Electric Reliability Corporation. Another issue is the effects of disturbances. A related work proposes to use SCD, where the deviations from the measuring values to the nominal values are quantified given a time window [24]. In this work, the “nominal values” are calculated by averaging the measured values of the current device in the time window. However, from the wide-area standpoint, using the average measurements of a single device may lead to several drawbacks. First, under islanding conditions, the measured quantities of some locations can go way off from other locations due to desynchronization. Using the SCD method, the islanded devices may still report itself as running under nominal conditions if their data does not contain large excursions. This

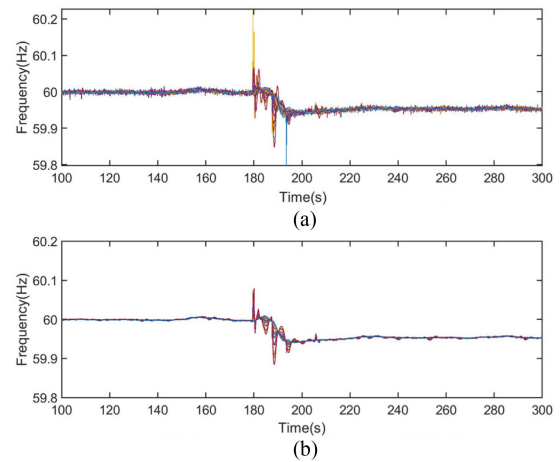


Fig. 1. Frequency data comparison (110 units).

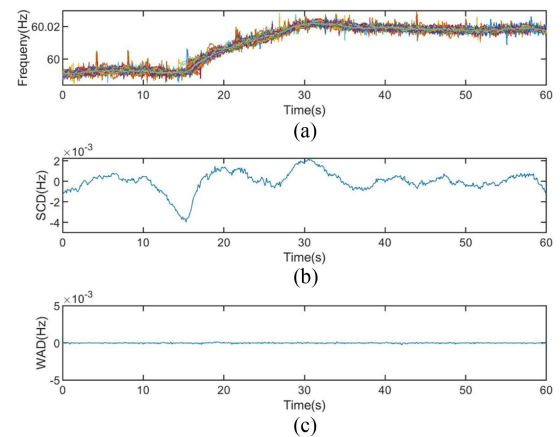


Fig. 2. Frequency ramping under low load condition.

will affect the calculation of the SCD because if all devices report themselves as running under nominal conditions, no statistical change will be detected. Second, using the average value of the past several seconds may be insufficient to measure the chaos in the data since it can be sensitive to normal frequency changes under low load conditions. In such a scenario, the synchrophasor measurements can contain high linearity even if the SCD algorithm reports a statistical change. Fig. 2 shows the frequency, the SCD, and the wide-area deviation (WAD) of a typical frequency ramping. Here, the WAD is represented by the difference between the measurements and the system medians. As seen in Fig. 2, a scheduled load change causes the frequency to ramp up to 60.024 Hz, while the frequencies are still synchronized across the grid. High linearity is observed in the data as the WAD is very smooth. However, the SCD reports a statistical change, which cannot accurately reflect the real system dynamics.

This article tries to address the abovementioned issues by evaluating and compressing the synchrophasor data via cross entropy and the state-of-the-art PCA variant, singular value decomposition (SVD) [25]. First, this article exploits a machine learning concept, cross entropy, to evaluate the patterns within

the synchrophasor data. Then, it generates compression periods according to the evaluation result. Finally, the proposed model compresses the synchrophasor data using the SVD algorithm, under relative error thresholds.

## II. SYNCHROPHASOR DATA EVALUATION USING CROSS ENTROPY

### A. Cross Entropy for Synchrophasor Data

Cross entropy [27] is a widely used concept in machine learning. It is commonly used in machine learning loss functions to measure the difference between the model outputs and the ground truth. For synchrophasors, the cross entropy can be written as follows:

$$H(M^t, \bar{M}^t) = - \sum_i P(\bar{M}^t) \log P(M_i^t) \quad (1)$$

where  $i$  and  $t$  represent the device ID and the time index, respectively.  $M$  is the distribution of the measurements,  $\bar{M}$  is the distribution of the nominal value,  $H(M^t, \bar{M}^t)$  represents the cross entropy of  $M^t$  with respect to  $\bar{M}^t$ ,  $P(x)$  is the probability of sample  $x$ .

Cross entropy can be used as an ancillary criterion for data compression due to its indication of off-nominal patterns. Under disturbance scenarios, the synchrophasor signals can have obvious off-nominal patterns. The off-nominal patterns represent the extent to which the system runs “chaotically.” The introduction of cross entropy helps describe how different the synchrophasor data is from the nominal patterns. Since the off-nominal patterns are observed during fault conditions and the nominal patterns are observed during ambient conditions, the evaluation of the synchrophasor data can be generalized as a bipartite classification problem. A simplified cross-entropy function for a bipartite classification can be written as follows:

$$H_B(M_i^t) = - \{ M_i^t \log [P(M_i^t)] + (1 - M_i^t) \log (1 - M_i^t) \} \quad (2)$$

where  $H_B(M_i^t)$  represents the bipartite cross entropy of  $M_i^t$ .

Now, the target is to identify the chunk of data that has off-nominal patterns, evaluate its cross entropy, and separate it from others that have nominal patterns. Therefore, (2) can be further simplified as (3), since presumably only the logarithmic distance between the distribution  $M_i^t$  and the target 0 is concerned

$$H_B(M_i^t) = -M_i^t \log [P(M_i^t)]. \quad (3)$$

In the compression algorithm, the nominal value of the frequency data is defined as the median frequency of an interconnected grid and the distribution  $M$  is calculated by subtracting the system median frequency value using PMUs’ reported actual frequencies. The nominal value of the voltage magnitude is the normal voltage magnitude per unit (p.u.), and its distribution  $M$  is calculated by subtracting each PMU’s normal voltage magnitude (p.u.) using this PMU’s actual voltage magnitude. Finally, the nominal value of the phase angle is the median of unwrapped angles. It is worth noting that phase angles may vary greatly compared to the frequency and voltage magnitudes. Therefore, for a certain PMU, it is required to subtract its phase angle value

---

**Algorithm 1:** Calculate the Cross-Entropy of the Synchrophasor Data, and Generate Partitions According to the Cross-Entropy Levels.

---

**Input:**  $Sdata^l$ : the synchrophasor data, where  $l$  is the length of the data;  $Ndata^l$ : the nominal-value data;  $Threshold\_entropy$ : the entropy threshold separating the ambient and the disturbance conditions;  $ew\_size$ : the window size to pre-partition the data;  $mw\_size$ : the window size to merge the pre-partition results.

**Output:**  $Partitions$ : the partitions generated by the algorithm

**Initialization:**  $c \leftarrow 0, j \leftarrow 0, Partitions \leftarrow [], k \leftarrow 1, Merged\_partitions \leftarrow []$ .

```

while  $j < l$  do
   $entropy \leftarrow Sdata_j - Ndata_j$ 
  if  $entropy > Threshold\_entropy$  then
    if  $c = 0$  then
       $s \leftarrow j$ 
    end if
     $c \leftarrow ew\_size$ 
  else
    if  $c > 0$  then
       $c \leftarrow c - 1$ 
    if  $c = 0$  then
       $e \leftarrow j$ 
      Append  $[s, e]$  to  $Partitions$ 
    end if
  end if
  while  $k < l$  do
    if  $Partitions[k][0] - Partitions[k-1][1] < mw\_size$  then
      Append  $[Partitions[k-1][0], Partitions[k][1]]$  to  $Merged\_partitions$ 
    end if
     $k \leftarrow k - 1$ 
  end while

```

---

at the first timestamp from all rest phase angle values [28], then normalize it through a  $[0, 1]$  range.

### B. Cross-Entropy-Based Evaluation

The purpose of the cross-entropy analysis of the synchrophasor data is to identify the periods that are chaotic, i.e., they contain low linearity. Identifying these periods are crucial to the data compression because the dimensionality reduction-based compression models exploit the high linearity of the synchrophasor data to achieve optimal compression performance. Therefore, if a chunk of data is of high cross entropy, a lower CR is required to maintain the information in the data, otherwise, a higher CR may be used to achieve a superior CR without losing too much information.

This article proposes a cross-entropy-based synchrophasor measurement evaluation approach combining the information from a wide area. Algorithm 1 shows a general partitioning



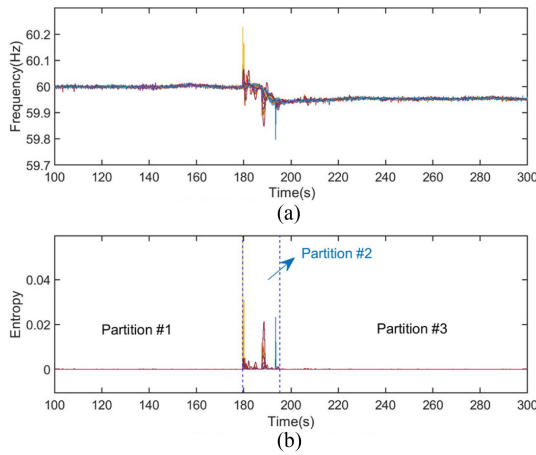


Fig. 3. Synchronphasor data partitioning via cross entropy (110 units).

method that calculates the cross entropy and generates the partitions for a chunk of synchronphasor data. However, on some occasions, it will generate partitions that are temporally close to each other. This is because high-linearity and low-linearity periods are interweaved under fault conditions. Although the number of partitions may not affect the compression performance directly, more partitions can result in excessive overheads that may take considerably large space when the data chunk is small. To avoid excessive overheads, the temporally close partitions are merged to reduce the number of partitions.

Fig. 3 shows the entropy distribution of a chunk of frequency measurements. As is seen from Fig. 3, the entropy of the ambient periods is around  $10^{-7}$ , which is relatively trivial compared to that of a fault period. Meanwhile, the entropy of the fault period goes up to over  $10^{-3}$ , which is 104 times larger than that of an ambient period. With entropy being calculated, generating partitions becomes a rather straightforward task. Fig. 3 also shows the merged result from step 2. For the frequency data, the nominal value at timestamp always equals to the median frequency of the grid. Then this article uses the proposed algorithms to partition the frequency data into chunks. As is seen from the figure, the frequency data is partitioned into three chunks. The first and the last partitions reflect the system-wide frequency distribution under ambient conditions, while the second partition reflects that under a generator trip condition. The first and the last partitions also imply high linearity, as their data are more “concentrated.” In the meantime, the second partition shows low linearity, as its data contains more excursions.

### III. SYNCHROPHASOR DATA COMPRESSION VIA SINGULAR VALUE DECOMPOSITION CONSIDERING DISTURBANCES

Using dimensionality reduction models to compress the synchronphasor data is not a new area of research. However, as aforementioned, the optimal decision of the compression space  $R$  is more of a human experience-based tradeoff between compression performance and accuracy.

A static choice of  $R$  is relatively biased in terms of the type, the volume, the resolution, and the entropy of the synchronphasor

data. The choice of  $R$  for a small synchronphasor network may not work for a large synchronphasor network assuming higher linearity exists when the number of devices is larger. On the other hand, under disturbance-involved power system dynamics, the linearity of the synchronphasor data can change drastically [28], which makes the choice of  $R$  rather difficult. Information vanishing is likely to happen if improper  $R$  is chosen. To address this issue, this article exploits the local characteristics of the synchronphasor data, proposing a dynamic singular value decomposition model to decide the best  $R$  for each data chunk. On the other hand, the proposed model also uses a relative evaluation methodology, which is capable of tracing very small fluctuations in the data.

#### A. Singular Value Decomposition

SVD is a widely accepted dimensionality reduction algorithm, which decomposes a large matrix  $M_{m \times n}$  into three smaller matrix  $U_{m \times n}$ ,  $\Sigma_{n \times n}$ , and  $V_{n \times n}$ . The SVD algorithm can be represented as follows:

$$M_{m \times n} = U_{m \times n} \Sigma_{n \times n} V_{n \times n}^T \quad (4)$$

where  $m$  is the number of samples,  $n$  is the number of PMUs,  $U_{m \times n}$  is the left singular vectors,  $\Sigma_{n \times n}$  is the diagonal matrix that represents the singular values, and  $V_{n \times n}$  is the right singular vectors.

The compression algorithm takes the top  $K$  singular vectors out of the  $N$  singular vectors. Therefore, the SVD reduces the problem to

$$M'_{m \times n} = U_{m \times k} \Sigma_{k \times k} V_{k \times n}^T \quad (5)$$

The CR is calculated by measuring the total number of values of the original matrix and the reduced matrix. Therefore, the CR is calculated by

$$CR = \frac{m \times n}{m \times k + k \times k + n \times k} = \frac{m \times n}{k(m + k + n)} \quad (6)$$

#### B. Model Tuning via Local Characteristics

In this article, a local characteristic evaluation methodology is proposed to address the vanishing of information caused by using a static NCV score as the threshold. The local characteristic LC is represented by

$$LC^{s,e} = \max_{i=1, \dots, n} |M_i^{s,e} - \text{median}(M^{s,e})| \quad (7)$$

where  $LC^{s,e}$  is the local characteristic of the  $i$ th measurement within a period,  $s$  is the start time,  $e$  is the end time,  $M_i^{s,e}$  is the  $i$ th measurement data within the  $[s, e]$  time period.

The idea of introducing the system median is it represents the most common distribution of the data. By calculating the maximum absolute deviation between the measurement and the system median, the proposed method recognizes the largest excursions that are caused by the disturbances.

This article calculates a proportion of the LC as the criteria to decide the tolerated reconstruction error threshold (TRET) by

$$\text{TRET}^{s,e} = \lambda \cdot LC^{s,e} \quad (8)$$

TABLE I  
PERFORMANCE COMPARISON OF SIMULATED DATA

Sig.		BF			LF			TF			LT		
		CR	MAE	ARMSE	CR	MAE	ARMSE	CR	MAE	ARMSE	CR	MAE	ARMSE
SVD-	VM	5.8	$1.2 \times 10^{-16}$	$7.4 \times 10^{-20}$	5.7	$3.2 \times 10^{-3}$	$1.7 \times 10^{-5}$	4.6	$2.2 \times 10^{-4}$	$3.5 \times 10^{-6}$	5.7	$7.4 \times 10^{-4}$	$1.4 \times 10^{-5}$
SCD	VH	3.8	$4.6 \times 10^{-5}$	$5.4 \times 10^{-6}$	11.5	$5.1 \times 10^{-5}$	$3.6 \times 10^{-7}$	11.5	$1.6 \times 10^{-5}$	$1.1 \times 10^{-6}$	11.5	$5.7 \times 10^{-6}$	$5.5 \times 10^{-7}$
	F	1.05	$1.7 \times 10^{-4}$	$2.5 \times 10^{-5}$	7.6	$2.0 \times 10^{-4}$	$3.4 \times 10^{-6}$	11.5	$9.6 \times 10^{-5}$	$3.7 \times 10^{-7}$	11.5	$1.5 \times 10^{-4}$	$3.0 \times 10^{-6}$
SVD-	VM	<b>9.5</b>	<b><math>3.5 \times 10^{-17}</math></b>	<b><math>2.5 \times 10^{-20}</math></b>	4.7	<b><math>7.7 \times 10^{-4}</math></b>	$4.8 \times 10^{-5}$	4.7	<b><math>1.2 \times 10^{-4}</math></b>	$2.9 \times 10^{-6}$	4.7	<b><math>2.0 \times 10^{-4}</math></b>	$3.4 \times 10^{-6}$
CE	VH	2.7	$4.3 \times 10^{-5}$	$1.37 \times 10^{-7}$	11.1	<b><math>5.8 \times 10^{-6}</math></b>	$2.0 \times 10^{-7}$	9.7	<b><math>8.3 \times 10^{-6}</math></b>	$2.2 \times 10^{-7}$	11.2	$5.2 \times 10^{-6}$	$3.1 \times 10^{-7}$
	F	<b>1.0</b>	<b><math>2.7 \times 10^{-18}</math></b>	<b><math>4.0 \times 10^{-18}</math></b>	7.7	<b><math>1.0 \times 10^{-4}</math></b>	$3.3 \times 10^{-6}$	11.4	<b><math>2.3 \times 10^{-5}</math></b>	$4.9 \times 10^{-7}$	7.8	<b><math>4.1 \times 10^{-5}</math></b>	$2.4 \times 10^{-6}$

where  $\lambda$  is a static coefficient that represents the TRET in percentage. This article uses 0.05 as the value of  $\lambda$  throughout the performance evaluation. It is noted that the choice of  $\lambda$  is subject to the requirements of users. Users may choose a smaller  $\lambda$  to preserve more information or a larger  $\lambda$  to get bigger **CR** per the requirements of applications.

#### IV. PERFORMANCE EVALUATION

##### A. Simulated Data

This article uses the “savnw” 23-bus system provided by PSSE 33 [29], assuming each bus is equipped with a PMU, which measures the bus voltage (VM), voltage phase angle (VH), and frequency ( $F$ ) at a reporting rate of 120 Hz. In this article, all simulations last 60 s. Since synchrophasor data may subject to local distribution or transmission characteristics, it is common to observe noises in such data [30]. To make the simulated data more authentic, white Gaussian noises of 75 dB, which equals to the observed average noise level in the field-collected data [31] as well as random phase and frequency jumps [32] are added to the simulated synchrophasor dataset.

In the simulation, this article considers disturbances including bus fault (BF), line fault (LF), transformer switch off (TF), and line trip (LT).

In this article, two criteria are considered to evaluate the reconstruction performance, which are maximum absolute error (MAE) and average root mean square error (ARMSE). The MAE is calculated by

$$\text{MAE} = \max_{m,n} |M_{m \times n} - M'_{m \times n}| \quad (9)$$

while the ARMSE is calculated by

$$\text{ARMSE} = \frac{M_{m \times n} - M'_{m \times n}}{\sqrt{mn}}. \quad (10)$$

Table I shows the comparison of the proposed cross-entropy-based SVD (SVD-CE) approach and the state-of-the-art statistical change detection-based SVD (SVD-SCD) algorithm. As is seen from Table I, the proposed SVD-CE generally outperforms the SVD-SCD algorithm. For the comparison of the voltage magnitude data under BF, the performance of the SVD-CE algorithm has better CR while keeping lower recovery errors. This is because, under BF, steps changes happen after the short-circuits on the buses, which cannot be easily detected by the SVD-SCD algorithm. It is also seen from Table I that the SVD-SCD can achieve better CRs on some occasions but they

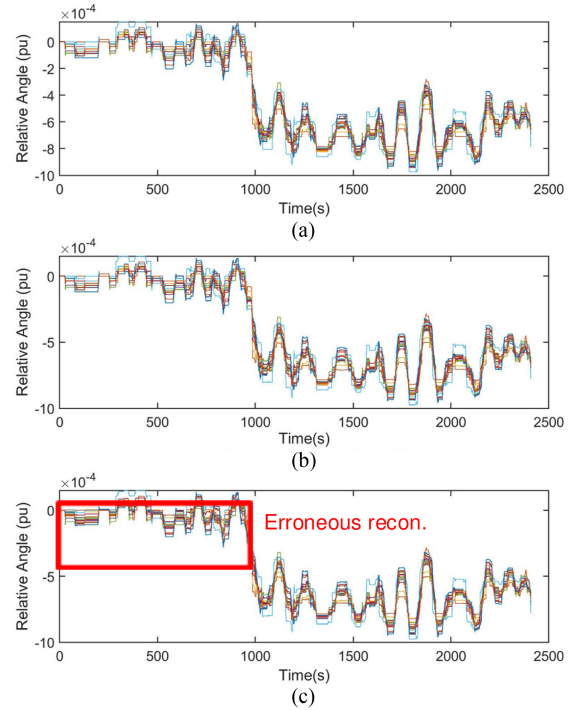
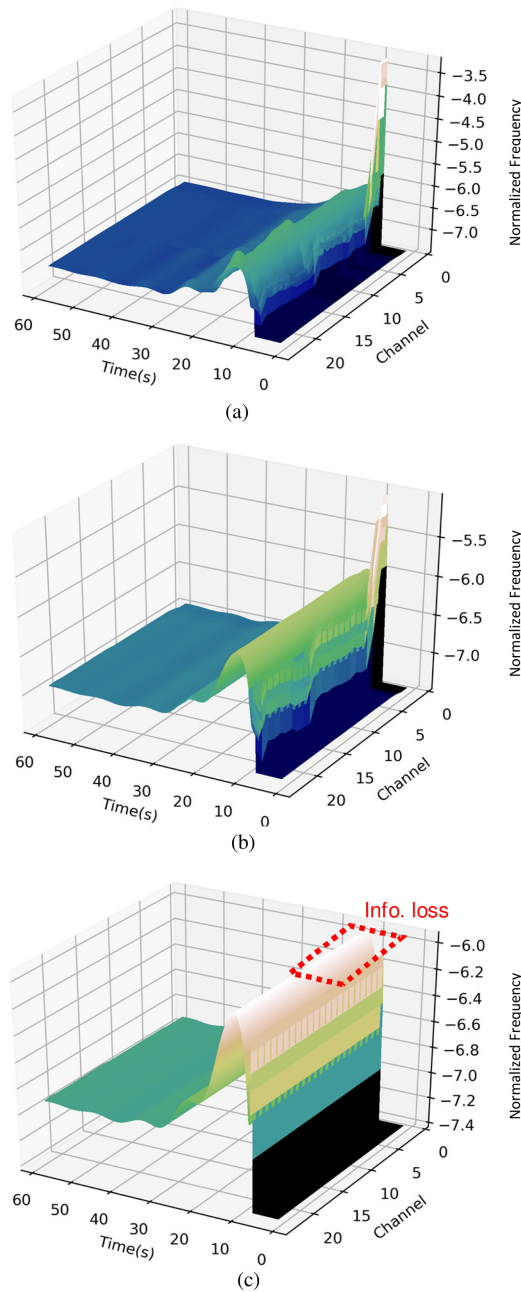


Fig. 4. Reconstruction performance angle BF (23 units).

usually imply unaffordable information loss. This is because the SVD-SCD algorithm may generate a single data chunk, where the high-linearity and low-linearity periods are interweaved. During events where high-linearity periods outnumber low-linearity ones, the overall linearity of the data chunk may rise. Under these scenarios, the SVD-SCD algorithm can compress the data more aggressively. Moreover, this feature can also result in lower ARMSEs among the high-linearity periods, since the algorithm tends to fit the high-linearity periods but undermine the low-linearity periods. These observations mostly happen under the LF and the LT conditions because under both conditions the high-linearity periods well outnumber the low-linearity periods. As opposed to it, the proposed SVD-CE algorithm differentiates the high-linearity periods and the low-linearity periods using their cross entropies. It assigns the best CR to each period, while keeps superior reconstruction accuracy. As is shown in Table I, the proposed algorithm can restrain the MAEs within lower ranges, while maintaining comparable CRs.

As a case study, Fig. 4 shows the recovered angle data under a BF. As is seen from Fig. 4, the proposed SVD-CE algorithm



**Fig. 5.** Frequency reconstruction performance LF & LT (23 units). (a) Original frequency measurements. (b) Reconstruction via SVD-CE. (c) Reconstruction via SVD-SCD.

reconstructs the data more accurately, while the SVD-SCD algorithm has human-eye perceivable errors at many time instants. This is because the SVD-SCD algorithm tends to find fewer principal components and causes the loss of critical information. In this case, the average CR of the SVD-CE algorithm is 2.7, while that of the SVD-SCD algorithm is 3.8. Therefore, the SVD-CE algorithm reaches a comparable CR rate while keeps relatively lower errors.

**Fig. 5** shows the reconstruction of frequency data under a complex disturbance, where an LT follows an LF. As seen in

**Fig. 5(a)**, there are a few frequency spikes in the original data around the fault location. After the disturbances, the frequency first rises, then drops to a steady level as frequency responses take place. As seen in **Fig. 5(b)**, the proposed SVD-CE approach can reconstruct the data to reflect the dynamics under the disturbance. However, as seen in **Fig. 5(c)**, the SVD-SCD algorithm overgeneralizes the data. It only includes the trend of the frequency while loses critical information around the fault location.

In conclusion, for the simulated data, the proposed model can maintain critical disturbances information while achieving a comparable CR rate. Nonetheless, as seen from **Table I**, the CR improvement by the proposed SVD-CE algorithm is usually not obvious. This is because the simulated data contains less noise. The cleaner simulated data are mainly caused by the simplicity of the simulation system. In this simulation system, there are few renewable sources or power electronic interfaces, making the noises caused by harmonic pollution, etc., less obvious than those in a real, complex system. Therefore, there has not seen a significant improvement in the CRs on all signals.

### B. Field Data

In this section, the field-collected frequency and voltage phase angle from the U.S. eastern interconnection provided by the distribution-level wide-area monitoring system FNET/GridEye are used for performance evaluation [33]. For the field data, this article evaluates the dynamic conditions including generator trip (GT), frequency ramping (FR), oscillation, and forced oscillation (FO).

**Table II** shows the performance comparison of the filed collected data. As is shown from **Table II**, under real-world scenarios, the proposed SVD-CE algorithm generally outperforms the SVD-SCD algorithm. Under GT and FO scenarios, although the CRs of the proposed algorithm is similar to the SVD-SCD algorithm, their errors are much less than the SVD-SCD algorithm because the SVD-CE algorithm has a stronger ability to pinpoint the high-entropy periods in the real-world scenarios, which have lower linearity for synchrophasor data in different locations. By specifying these periods, the algorithm can fit these periods with a tailor-made error threshold instead of an NCV.

As is seen from **Table II**, under LT scenarios, the proposed algorithm can achieve a much better result than the SVD-SCD algorithm. The reason is that static NCV does not work well when LT events are present. The NCV tends to force the algorithm to find a much higher compression dimension to meet the NCV. However, the proposed algorithm finds the lowest possible compression dimension while keeping relatively low reconstruction errors. Moreover, under LT scenarios, there are tiny phase steps in the point of wave (POW) measurements that cause frequency deviations [34]. Under LT scenarios, the tripping of the line usually causes tiny phase steps in the POW, which in result causes tiny local angle variations. As is shown in **Fig. 6**, the LT event causes angle variations at around 14.8 s. The magnitudes of these variations are very small, making them hard to recognize via human eyes. However, these tiny variations

TABLE II  
PERFORMANCE COMPARISON OF FIELD DATA

Sig.	GT			LT			FR			FO		
	CR	MAE	ARMSE	CR	MAE	ARMSE	CR	MAE	ARMSE	CR	MAE	ARMSE
SVD-VH	49.6	$1.2 \times 10^{-4}$	$1.7 \times 10^{-6}$	17.1	$3.7 \times 10^{-5}$	$1.8 \times 10^{-6}$	<b>38.8</b>	$1.7 \times 10^{-4}$	$3.4 \times 10^{-6}$	40.2	$6.8 \times 10^{-6}$	$2.9 \times 10^{-6}$
SCD-F	9.3	$1.3 \times 10^{-3}$	$3.6 \times 10^{-5}$	11.9	$2.2 \times 10^{-4}$	$2.1 \times 10^{-5}$	92.0	$8.3 \times 10^{-5}$	$2.4 \times 10^{-8}$	1.7	$2.9 \times 10^{-4}$	$4.4 \times 10^{-5}$
SVD-VH	49.4	$3.2 \times 10^{-5}$	$1.5 \times 10^{-6}$	19.0	$3.6 \times 10^{-5}$	$1.7 \times 10^{-6}$	<b>41.9</b>	$1.7 \times 10^{-5}$	$3.7 \times 10^{-6}$	52.1	$6.8 \times 10^{-4}$	$4.4 \times 10^{-6}$
CE-F	9.5	$4.1 \times 10^{-4}$	$2.6 \times 10^{-5}$	15.5	$2.2 \times 10^{-4}$	$1.7 \times 10^{-5}$	92.0	$8.3 \times 10^{-5}$	$2.4 \times 10^{-8}$	1.6	$2.1 \times 10^{-4}$	$8.0 \times 10^{-6}$

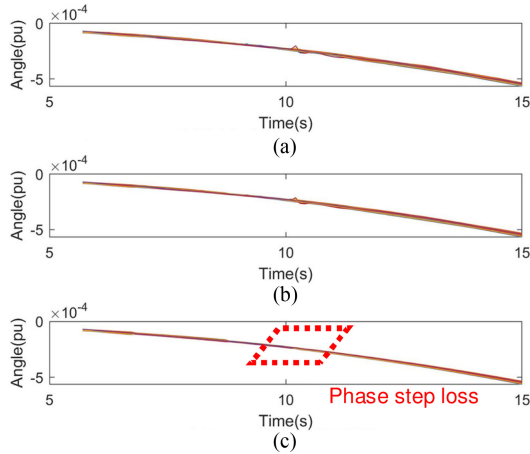


Fig. 6. Reconstruction performance angle LT (110 units).

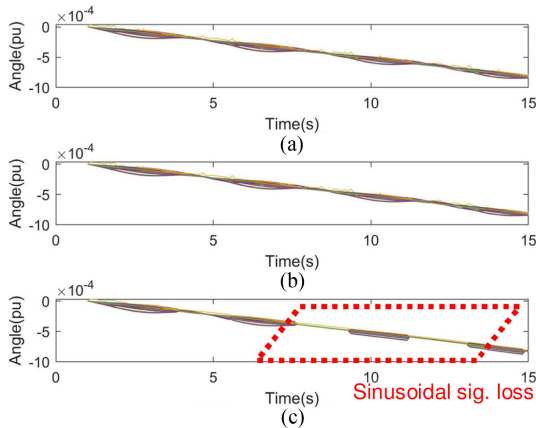


Fig. 7. Reconstruction performance angle FO (110 units).

represent the intrinsic characteristics of the POW data, which is of great significance in terms of deciding the credibility of the LT event. As is seen from Fig. 6, the SVD-CE algorithm successfully preserves these tiny variations that the SVD-SCD algorithm does not reflect. On the other hand, it also suggests, under LT scenarios, the SVD-CE algorithm achieves better CR while keeping good reconstruction accuracy.

Moreover, as Fig. 7 shows, under FO scenarios, the proposed SVD-CE algorithm successfully retains the oscillation information, while the SVD-SCD algorithm only retrains partial oscillation information. Moreover, under FO scenarios, the information

loss of the SVD-SCD algorithm can greatly affect the event analysis, since it loses critical sinusoidal signals at multiple points. As a result, the reconstructed data of the SVD-SCD algorithm would fail the classic forced oscillation analysis. Therefore, even though the SVD-SCD algorithm achieves better CR, it fails to retain critical information, which is crucial to event analysis.

### C. Data Compression Under Complex Disturbance Conditions

In field operations, the characteristics of disturbances are complex. Complex disturbances can greatly affect the dynamics of an interconnected power grid; thus, they can seriously deteriorate the performance of the compression algorithms.

A notable feature of the proposed method is its superior information retaining ability under complex disturbances. Fig. 8 shows the performance comparison during an GT event. In Fig. 8(a) and (b), the x-axis represents the PMU channels, while the y-axis represents the time index. A GT disturbance is observed at time index 192 and it causes system-wide frequency drops in all channels. As seen in Fig. 8, the proposed SVD-CE method achieves good reconstruction performance. This is because it can successfully identify the disturbances period and perform efficient compression strategies on disturbance period and ambient periods, respectively. For the SVD-SCD method, it loses critical information during the disturbance and post-disturbance periods. The high reconstruction error during the disturbance period is caused by its inability to bound the TRET. While the high reconstruction error during post-disturbance period is due to the fact that the SVD-SCD approach treats the postdisturbance characteristics by finding a much smaller compression space regardless of the important local characteristics. As opposed to it, the proposed SVD-CE algorithm still retains more information. A similar example is shown in Fig. 9, where an LD disturbance is involved. As seen from Fig. 9(c), the SVD-SCD approach finds less accurate representations of the data. It still oversimplifies the data and causes inaccuracies in the postdisturbance period. On the other hand, on some channels, this approach results in erroneous reconstructions including spikes that are not presented in the original data.

A more complex disturbance scenario is shown in Fig. 10, where an LT happens at 4.0 s and a GT happens at 16.0 s. As seen from Fig. 10(c), the SVD-SCD algorithm is able to accurately preserve the LT information at 4.0 s. However, it fails



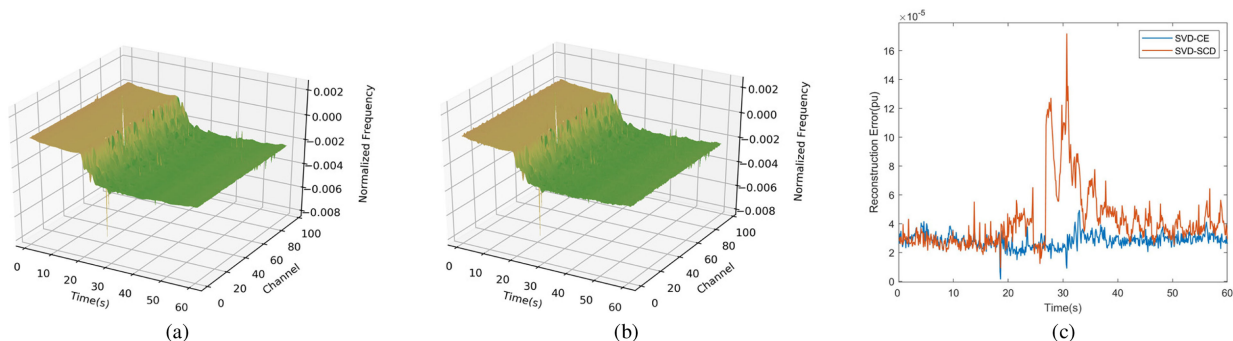


Fig. 8. Reconstruction performance on frequency under simple GT (110 units). (a) Original. (b) SVD-CE. (c) Reconstruction error comparison.

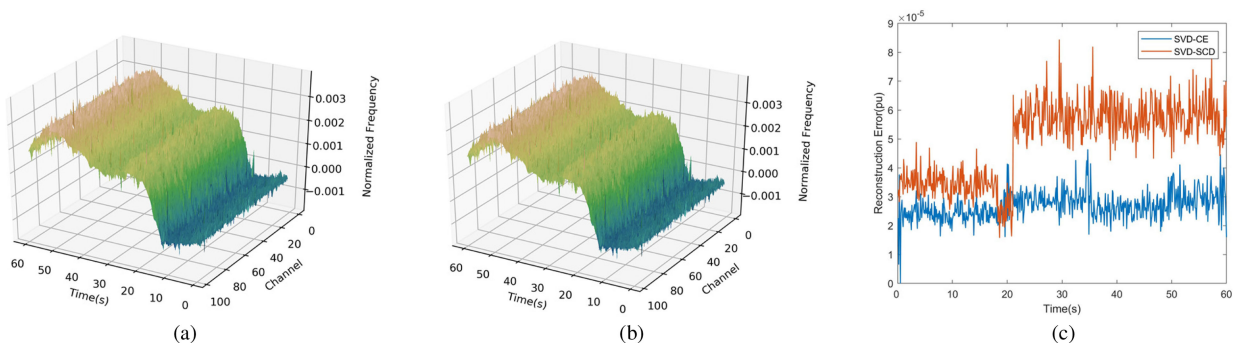


Fig. 9. Reconstruction performance on frequency under simple LD (110 units). (a) Original. (b) SVD-CE. (c) Reconstruction error comparison.

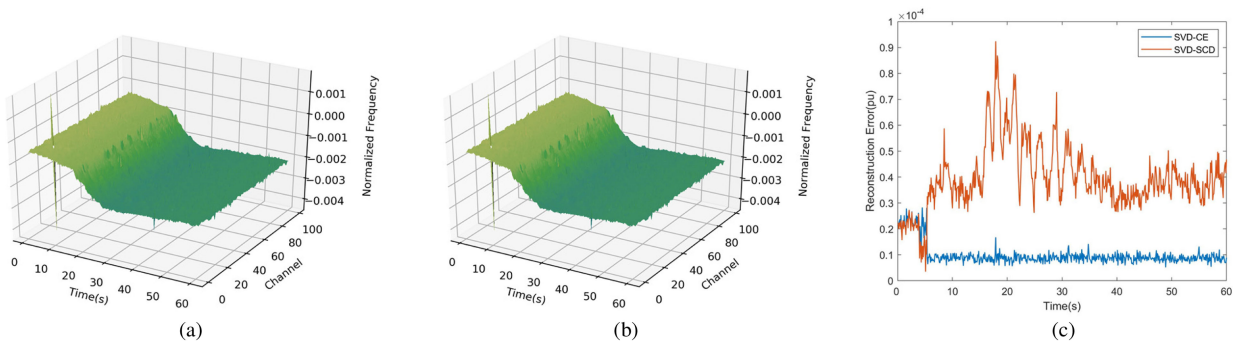


Fig. 10. Reconstruction performance on frequency under complex LT & GT (110 units). (a) Original. (b) SVD-CE. (c) Reconstruction error comparison.

to generalize the complex disturbance due to the high reconstruction errors during the GT and post-GT periods. However, the proposed SVD-CE algorithm succeeds in identify the LT and the GT and compresses the PMU data in a more accurate manner. As seen from Fig. 10(b), the reconstructed data by the proposed approach is almost identical to the original data.

Finally, Fig. 11 shows the reconstruction results under a continued FO. As seen, the reconstruction performance of the proposed model is superior. The reason is that during the FO event, the off-nominal characteristics from the wide-area are obvious. Therefore, the SVD-CE based method successfully recognizes the disturbance periods and selects proper CR to

compress the data. For the SVD-SCD approach, the statistics of the synchrophasor data keep changing, thus the SCD algorithm is constantly triggered and the whole period is recognized as a disturbance period. However, like Fig. 8(c) without bounding the TRET, the reconstruction error of the SVD-SCE approach is still much higher than that of the proposed SVD-CE method.

## V. DISCUSSION

### A. Online Implementation and Compression Time

Since the matrix factorization algorithms like SVD are usually costly in time consumption [35], it is necessary to evaluate its



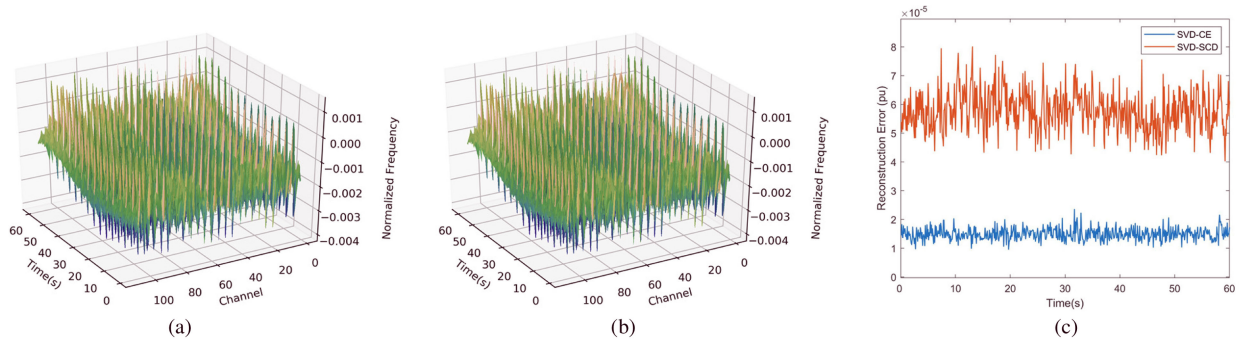


Fig. 11. Reconstruction performance on frequency under simple FO (110 units). (a) Original. (b) SVD-CE. (c) Reconstruction error comparison.

TABLE III  
AVERAGE TIME CONSUMPTION

	Field Data (10Hz)		Simulated Data (120Hz)	
	Data Length	Compression Time	Data Length	Compression Time
Ambient	18.3s	46.0ms	1.6s	7.2ms
	28.0s	68.3ms	3.4s	18.8ms
	46.3s	114.3ms	4.8s	25.6ms
Event	2.6s	49ms	0.3s	2.1ms
	7.6s	98.9ms	0.7s	5.7ms
	12.2s	150.1ms	1.1s	7.6ms

execution time so that it can be implemented online. Due to the high reporting rate of PMU, immediately processing the data once it arrives at the PDC may introduce unaffordable time overhead. Therefore, to overcome this issue, this article follows a widely adopted batch-processing strategy [36] to perform an online data compression.

The experimental computer is equipped with an Intel Core i7-8700 3.20 GHz CPU, 16.0 GB memory, and Python 3.7. In this article, when the PDC receives the measurement data, it pushes the newly received data into a cache, which is limited in capacity. Then, once the limit of the capacity is exceeded, the PDC invokes the compression algorithm to compress the cached data chunk asynchronously. Here, the capacity of the data chunk is defined as 600, which equals to 60 s data at 10 Hz reporting rate. The choice of the 60 s windows is primarily due to the requirement of disturbance analytical applications [35], [37] and it equals to around 3 MB data in the memory. However, it is worth noting that the actual data chunk size depends on the calculational power of the computer. Although a larger data chunk size may be possible, generally, it should not exceed the throughput limit of the hard disk. Table III shows the average time consumption of the proposed algorithm to compress certain-length raw measurement data. As Table III shows, utilizing the batch compression strategy, both 10 and 120 Hz data can be compressed in a short time. It is worth to note the compression time of the field-collected 10 Hz data is greater than that of the simulated 120 Hz data. This is because the field-collected has greater variations, thus it introduces greater entropy. Given this fact, the SVD-CE algorithm tends to search exhaustively for the best CR, so it takes a relatively long time to execute. However, under both scenarios, the compression procedure catches up well with the data collection procedure.

### B. Choice of Coefficient $\lambda$

In this article, the  $\lambda$  is set to 0.05, tolerating a maximum of 5% reconstruction error. Setting  $\lambda$  to 0.05 is mainly due to the requirement from compliance investigation. However, the determination of the  $\lambda$  depends on how the reconstructed will be used. For example, if there is no concern about the accuracy of each PMU's data, the  $\lambda$  may be set to a much large value, e.g., 0.2. The advantages of setting the  $\lambda$  to a larger value are it speeds up the execution of the compression procedure, and it usually acquires a high CR. However, under this condition, some important information will not be retained during disturbance periods. On the other hand, if there is a stringent requirement of data accuracy, e.g., compliance purposes, the  $\lambda$  shall be set to a lower value. Otherwise, the inaccurate reconstructed data could eventually result in a financial loss of power companies because it brings inaccuracy to the compliance investigation. Moreover, in this article, the performance evaluation indicates that the 0.05 value of the  $\lambda$  is independent of data types, data volumes, disturbance types, etc., and using this value meets the requirement of compliance standards. However, it is urged to keep the  $\lambda$  within 0.1 (10%), where a series of data loss is observed under data reconstruction.

## VI. CONCLUSION

This article proposed a synchrophasor data compression method using cross-entropy-based SVD. First, this article illustrated the characteristics of the synchrophasor data under power system dynamics and explained the motivation of using the cross entropy to measure the chaotic extent of the synchrophasor data. Second, the cross-entropy-based data compression approach was proposed and implemented to compress the simulation

and field-collected synchrophasor data. The comparison result suggested that the proposed algorithm can achieve better compression and reconstruction performance compared to the state-of-the-art SVD-SCD method.

Since the proposed model is based on a dimensionality reduction technique, SVD, it requires complex matrix manipulations. They can consume a relatively long time if the input data is large. Moreover, finding the best CR iteratively will take a rather long time in real-world scenarios. Since the linearity of the synchrophasor data is decided by the off-nominal values, it may be interesting to exploit recent machine learning advances to estimate the optimal CRs. Another way to facilitate the compression model is utilizing the parallel computational capability of GPUs. GPUs can accelerate matrix operations for more than 100x. Since classic matrix operations all rely on CPUs, it will be significant to exploit the power of GPUs in the compression world.

## REFERENCES

- [1] J. Zhao, G. Zhang, M. L. Scala, and Z. Wang, "Enhanced robustness of state estimator to bad data processing through multi-innovation analysis," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1610–1619, Aug. 2017.
- [2] Z. Lin, F. Wen, Y. Ding, and Y. Xue, "Data-driven coherency identification for generators based on spectral clustering," *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 1275–1285, Mar. 2018.
- [3] W. Yu, W. Yao, X. Deng, Y. Zhao, and Y. Liu, "Timestamp shift detection for synchrophasor data based on similarity analysis between relative phase angle and frequency," *IEEE Trans. Power Del.*, vol. 35, no. 3, pp. 1588–1591, Jun. 2020.
- [4] W. Wang, J. Zhao, W. Yu, and Y. Liu, "FNETVision: A WAMS big data knowledge discovery system," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2018, pp. 1–5.
- [5] W. Yao *et al.*, "A Fast load control system based on mobile distribution-level phasor measurement unit," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 895–904, Jan. 2020.
- [6] C. Chen, H. Yang, W. Wang, M. Mandich, W. Yao, and Y. Liu, "Harmonic transmission characteristics for ultra-long distance AC transmission lines based on frequency-length factor," *Elect. Power Syst. Res.*, vol. 182, 2020, Art. no. 106189.
- [7] C. Chen, X. Ma, H. Yang, W. Wang, and Y. Liu, "Study on the power-frequency waves distribution characteristics for half-wavelength transmission lines based on the frequency-length factor," *Math. Problems Eng.*, vol. 2020, pp. 1–14, 2020.
- [8] N. Tong, L. Chen, W. Wang, X. Lin, and Z. Li, "Local-measurement-based high-speed protection for half-wavelength UHV lines," *IEEE Trans. Power Del.*, to be published, doi: 10.1109/TPWRD.2020.2969960.
- [9] K. Sun *et al.*, "VSC-MTDC system integrating offshore wind farms based optimal distribution method for financial improvement on wind producers," *IEEE Trans. Ind. Appl.*, vol. 55, no. 3, pp. 2232–2240, May/Jun. 2019.
- [10] T. Li, Y. Li, M. Liao, W. Wang, and C. Zeng, "A new wind power forecasting approach based on conjugated gradient neural network," *Math. Problems Eng.*, vol. 2016, pp. 1–9, 2016.
- [11] Summary of the North American Synchrophasor Initiative (NASPI) Activity Area, U.S. Dept. Energy, Washington, DC, USA, 2012.
- [12] IEEE Standard for Synchrophasor Measurements for Power Systems, IEEE Standard C37.118.1-2011, 2011.
- [13] S. Das and T. Singh Sidhu, "Application of compressive sampling in synchrophasor data communication in WAMS," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 450–460, Feb. 2014.
- [14] K. Sayood, *Introduction to Data Compression*, 4th ed. New York, NY, USA: Elsevier, 2012.
- [15] R. Klump, P. Agarwal, J. Tate, and H. Khurana, "Lossless compression of synchronized phasor measurements," in *Proc. IEEE PES Gen. Meeting*, 2010, pp. 1–7.
- [16] P. Top and J. Breneman, "Compressing phasor measurement data," in *Proc. North Amer. Power Symp.*, 2013, pp. 1–4.
- [17] S. Santoso, E. Powers, and W. Grady, "Power quality disturbance data compression using wavelet transform methods," *IEEE Trans. Power Del.*, vol. 12, no. 3, pp. 1250–1257, Jul. 1997.
- [18] J. Ning, J. Wang, W. Gao, and C. Liu, "A wavelet-based data compression technique for smart grid," *IEEE Trans. Smart Grid*, vol. 2, no. 1, pp. 212–218, Mar. 2011.
- [19] E. Hamid and Z. Kawasaki, "Wavelet-based data compression of power system disturbances using the minimum description length criterion," *IEEE Trans. Power Del.*, vol. 17, no. 2, pp. 460–466, Apr. 2002.
- [20] J. Khan, S. Bhuiyan, G. Murphy, and J. Williams, "Data denoising and compression for smart grid communication," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 2, pp. 200–214, Jun. 2016.
- [21] J. Khan, S. M. A. Bhuiyan, G. Murphy, and M. Arline, "Embedded-Zerotree-wavelet-based data denoising and compression for smart grid," *IEEE Trans. Ind. Appl.*, vol. 51, no. 5, pp. 4190–4200, Sep./Oct. 2015.
- [22] F. Zhang, L. Cheng, X. Li, Y. Sun, W. Zhao, and W. Zhao, "Application of a real-time data compression and adapted protocol technique for WAMS," *IEEE Trans. Power Syst.*, vol. 30, no. 2, pp. 653–662, Mar. 2015.
- [23] L. Xie, Y. Chen, and P. Kumar, "Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis," *IEEE Trans. Power Syst.*, vol. 29, no. 6, pp. 2784–2794, Nov. 2014.
- [24] P. Gadde, M. Biswal, S. Brahma, and H. Cao, "Efficient compression of PMU data in WAMS," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2406–2413, Sep. 2016.
- [25] J. Souza, T. Assis, and B. Pal, "Data compression in smart distribution systems via singular value decomposition," *IEEE Trans. Smart Grid*, vol. 8, no. 1, pp. 275–284, Jan. 2017.
- [26] North American Electric Reliability Corporation, Atlanta, GA, USA, "Balancing and frequency control," 2011, pp. 17–40.
- [27] P. T. Boer, D. Kroese, S. Mannor, and R. Y. Rubinsteyn, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 34, no. 1, pp. 19–67, 2005.
- [28] W. Wang *et al.*, "Frequency disturbance event detection based on synchrophasors and deep learning," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3593–3605, Jul. 2020.
- [29] "PSSE-high-performance transmission planning and analysis software," 2020. [Online]. Available: <https://new.siemens.com/global/en/products/energy/services/transmission-distribution-smart-grid/consulting-and-planning/pss-software/pss-e.html>. Accessed on: Jul. 23, 2019.
- [30] X. Deng, D. Bian, D. Shi, W. Yao, L. Wu, and Y. Liu, "Impact of low data quality on disturbance triangulation application using high-density PMU measurements," *IEEE Access*, vol. 7, pp. 105054–105061, 2019.
- [31] Y. Liu *et al.*, "Recent developments of FNET/GridEye — A situational awareness tool for smart grid," *CSEE J. Power Energy Syst.*, vol. 2, no. 3, pp. 19–27, 2016.
- [32] F. W. Bernal, J. Wold, R. Concepcion, and J. Budai, "A method for correcting frequency and RoCoF estimates of power system signals with phase steps," in *Proc. North Amer. Power Symp.*, 2019, pp. 1–6.
- [33] X. Deng, H. Li, W. Yu, W. Wang, and Y. Liu, "Frequency observations and statistic analysis of worldwide main power grids using FNET/GridEye," in *Proc. IEEE Power Energy Soc. General Meeting*, 2019, pp. 1–5.
- [34] P. Wright, P. Davis, K. Johnstone, G. Rietveld, and A. J. Roscoe, "Field measurement of frequency and ROCOF in the presence of phase steps," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 9, pp. 1688–1695, Jun. 2019.
- [35] W. Wang, W. Yao, C. Chen, X. Deng, and Y. Liu, "Fast and accurate frequency response estimation for large power system disturbances using second derivative of frequency data," *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 2483–2486, May 2020.
- [36] "Batch writes," in Optimize Writes to InfluxDB, 2020. [Online]. Available: <https://v2.docs.influxdata.com/v2.0/write-data/best-practices/optimize-writes/>. Accessed on Apr. 01, 2020.
- [37] North American Electric Reliability Corporation, Atlanta, GA, USA, Frequency Response Standard Background Document, 2012.



**Weikang Wang** (Student Member, IEEE) received the B.S. degree in computer science from the School of Control and Computer Engineering, North China Electric Power University, Beijing, China, in 2016. He is currently working toward the Ph.D. degree in computer engineering in the University of Tennessee, Knoxville, TN, USA.

His current research interests include wide-area monitoring, situation awareness, big data, and machine learning.



**Chang Chen** received the B.S. degree in electrical engineering from the College of Electrical Engineering, Sichuan University, Chengdu, China, in 2015, where she is currently working toward the Ph.D. degree in electrical engineering.

She is currently a Visiting Scholar with the University of Tennessee, Knoxville, TN, USA. Her current research interests include harmonic modeling and analysis, power quality, etc.



**Wei Qiu** received the B.S. degree from Hubei University of Technology, Wuhan, China, in 2015, and the M.Sc. degree from Hunan University, Changsha, China, in 2017, both in electrical engineering. He is currently working toward the Ph.D. degree in electrical engineering from Hunan University. He has been a joint doctoral student in the University of Tennessee, Knoxville, TN, USA, since 2019.

His current research interests include power system analysis, power quality measurement, and reliability analysis of power equipment.



**Wenxuan Yao** (Member, IEEE) received the B.S. and Ph.D. degrees from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2011 and 2017, respectively, and the Ph.D. degree from the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, USA, in 2018, all in electrical engineering.

He is currently a Postdoctoral Research Associate with the Oak Ridge National Laboratory, Oak Ridge, TN, USA. His current research interests include wide-area power system monitoring, synchrophasor measurement application, embedded system development, power quality diagnosis, and big data analysis in power systems.



**Kaiqi Sun** (Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Shandong University, Jinan, China, in 2015 and 2020, respectively.

He is currently a Research Associate with the Department of Electrical Engineering and Computer Science, The University of Tennessee, Knoxville, TN, USA. His current research interests include the control strategy of HVDC system, and the application of VSC-MTDC systems for renewable energy integration and urban power grid.

urban power grid.



**Yilu Liu** (Fellow, IEEE) received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, and the M.S. and Ph.D. degrees from Ohio State University, Columbus, OH, USA, in 1986 and 1989, respectively, all in electrical engineering.

She was a Professor with Virginia Tech, where she led the effort to create the North American Power Grid Frequency Monitoring Network, which is now operated at the University of Tennessee, Knoxville (UTK), Knoxville, TN, USA, and Oak Ridge National Laboratory (ORNL), Oak Ridge, TN, USA, as GridEye. She is currently the Governor's Chair with UTK and ORNL. She is also the Deputy Director of the DOE/NSF-Cofunded Engineering Research Center CURENT. Her current research interests include power system wide-area monitoring and control, large interconnection-level dynamic simulations, electromagnetic transient analysis, and power transformer modeling and diagnosis.

Prof. Liu is elected as a member of National Academy of Engineering in 2016.