Fractional Dynamics of PMU Data

Laith Shalalfeh, *Member, IEEE,* Paul Bogdan, *Member, IEEE,* Edmond Jonckheere, *Life Fellow, IEEE*

Abstract—Novel dynamics are emerging in the power system due to the new Smart Grid (SG) environment. The high sampling rate of the Phasor Measurement Units (PMUs) enables them to capture the dynamic fluctuations in the power system measurements. Understanding the statistical and dynamic characteristics of the PMU data requires advanced data analytics techniques capable of performing accurate modeling of the power system variables (voltage, frequency, phase angle, and rate of change of frequency (ROCOF)). In this paper, we provide evidence of the non-stationarity and fractality of PMU data collected from Europe. We adopt the Autoregressive Fractionally Integrated Moving Average (ARFIMA) models with non-integer differencing parameter to model the short-range and long-range correlations in the PMU data. Furthermore, the goodness-of-fit of the ARFIMA model is confirmed by analyzing the correlation and independence of the model residuals. Anomaly detection is among the promising applications of the PMU ARFIMA models. It is shown that the 2012 Indian blackout is accompanied by a change point in the differencing parameter opening the road to event (anomaly) detection by ARFIMA monitoring.

Index Terms—PMU Data, Long-Range Memory, ARFIMA Models, Anomaly Detection, Fractional Dynamics, Smart Grid

I. Introduction

THE SG is a modernized grid that overcomes the challenges and issues in the conventional power grid. Several challenges have arisen from the higher penetration of renewable energy resources and the increasing number of electric vehicles. Therefore, wide-area monitoring, protection, and control systems will have an important role in the future SG by developing a reliable, secure, and efficient operation.

Wide-area monitoring systems collect real-time measurements from all over the power grid via advanced sensing devices, such as PMUs. That enables more accurate monitoring of the grid state in real time. The PMU data is collected at higher sampling rate, 30-120 samples/s, which exposes the fast dynamic events and the contingencies in the power grid.

Understanding the statistical characteristics of PMU data is of great importance due to several applications in power system studies. The authors of [1] show the existence of self-organized criticality in blackout data. In [2], the authors show that the autocorrelation and variance of frequency time series increase as the power system approaches instability. The authors of [3] provide evidence for an increase in the Hurst exponent of real frequency data collected from the Indian grid before approaching the 2012 blackout [4]. This increase could provide early-warning of catastrophic events in the power system. A deeper aim of the present paper is to provide stronger theoretical relevance of the Hurst exponent

L. Shalalfeh is with the Department of Energy Engineering, German Jordanian University, Amman 11180, Jordan.

P. Bogdan and E. Jonckheere are with Dept. of Electrical and Computer Engineering, Univ. of Southern California, Los Angeles, California 90089.

phenomena by corroborating them with data-driven models of the PMU time series, with the objective of transforming the promising observations of [3][5] to detection of anomalies with controllable false alarm rates in an autonomous SG.

In [6], we have shown that the PMU data (voltage magnitude, frequency, and phase angle) is not random and possesses long-range memory with scaling exponent $(\alpha^\star)$ higher than the one of short-memory data $(\alpha^\star = 0.5)$. The long-range dependence in the PMU data was evaluated using Detrended Fluctuation Analysis (DFA) [7] by calculating the scaling exponents of several data sets from the synchrophasor network in Texas. Modeling of long-memory data requires ARFIMA models [8] that can capture both the short- and long-range memories. Furthermore, the ARFIMA modeling of the power loads was suggested in [9], consistently with the multi-fractality of such signals.

In [10], we investigated the fractality of PMU data by calculating the three fractal parameters: scaling exponent $(\alpha^\star)$, power exponent (β), and differencing parameter (d). The calculated differencing parameters from large data set of PMU data had non-integer mean values, so the ARFIMA model was adopted as the best model describing the short and long memories. The selection of the best model was based on two information criteria: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

In this paper, we extend our work on the ARFIMA modeling [10] by first providing more accurate identification and estimation of the PMU ARFIMA models, subject to a precise significance level. In the identification phase, we estimate the differencing parameter and the autoregressive and moving average coefficients. Then, in the estimation phase, we use these values as initial guess for the Whittle estimator to find the parameters of the best ARFIMA model. Secondly, we show that ARFIMA models are well suited for PMU data by showing that their α-stable residuals are uncorrelated and independent. At the end, we derive the ARFIMA models of the Indian frequency toward the 2012 blackout. Our results show an increase in the differencing parameter that could provide an early warning to the proximity to a blackout.

The paper is organized as follows: we study the stationarity and fractality of PMU data in Sec. II and Sec. III, respectively. In Sec. IV, we find the best ARFIMA models to fit the PMU data. Analyzing the residuals of the ARFIMA models is carried out in Sec. V. In Sec. VI, we exploit the ARFIMA models to anticipate power blackouts. Sec. VII is the conclusion.

II. Stationarity of PMU Data

The new SG deployment, its renewable energy generation and smart loads, are accompanied by novel stochastic and dynamic behaviors. The resulting dynamics of the power
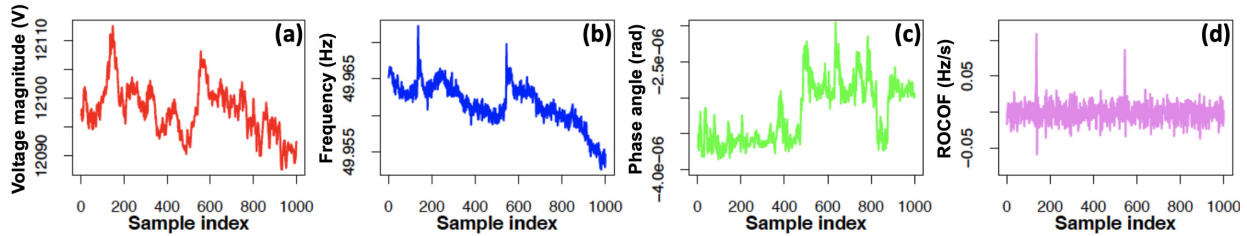
Figure 1. PMU data collected from the EPFL campus grid in 2014: (a) Voltage magnitude (b) Frequency (c) Angle (d) ROCOF

system lead to fluctuations in the PMU data under both normal and emergency conditions. Evidently, the PMUs are capable of capturing the fast dynamic states in the power grid. The stationarity of the data should be rigorously determined in advance of modeling the PMU data as some statistical methods are not suitable for non-stationary data.

In this section, we first provide a description of the PMU data and the power grid from which the data was collected. Then, we evaluate the data stationarity using the unit root tests.

A. Overview of PMU Data

Typical PMUs provide measurements for the following variables in the power system: voltage (V), current (I), frequency (f), ROCOF, active power (P), and reactive power (Q). The measured voltages and currents are represented using the phasor format, which consists of magnitude and phase angle. In this paper, we use data collected from the Ecole Polytechnique Fédérale de Lausanne (EPFL) campus grid as part of their real-time state estimation project [11]. The rated voltage magnitude (line-line) and frequency of the EPFL campus grid are 20 kV and 50 Hz, respectively. Several PMUs were installed throughout the campus grid to collect the data at sampling rate of 50 samples/s. We focus our analysis on a large data set of voltage magnitude (V), frequency (f), phase angle (θ), and ROCOF. The data set consists of 160,000 time series, each having 1,000 sample points, of the four variables collected from the campus grid in 2014 [12]. In Figs. 1 (a)-(d), we show 1000-sample time series of voltage magnitude (red), frequency (blue), phase angle (green), and ROCOF (magenta).

B. Stationarity

In PMU data, we can have a glimpse at the stationarity or the lack thereof by calculating their autocorrelation functions (ACFs). As shown in Figs. 2 (a)-(d), the ACFs of the voltage, frequency, and angle show a slow hyperbolic decay compared to the exponentially decaying one for the ROCOF. The slow decay of the ACF could be a sign of lack of stationarity.

More formally, we test the stationarity of PMU measurements using the Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) unit root tests. Such tests can classify a time series as either stationary or not, based on the existence of unit root in the autoregressive polynomial of the time series.

1) ADF test: The ADF test classifies the time series as stationary or not using hypothesis testing. The null hypothesis (H_0) is that the time series is non-stationary and a unit root exists. The alternative hypothesis (H_1) is that the time series is stationary.

We conduct the ADF test on 160,000 time series, each having 1,000 sample points, of PMU data (V, f, θ, and ROCOF). Using the command "adf.test" in R software [13], we calculated the p-value for each time series to determine its stationarity. The percentages of time series with p-values above 0.01 (accept the null hypothesis) and time series with p-values below or equal to 0.01 (reject the null hypothesis) are shown in Table I.

The first column (p-value$>$ 0.01) under the ADF test shows that we cannot reject the null hypothesis for the voltage (81.33%), frequency (96.86%), and angle (45.88%) time series. Not rejecting the null hypothesis in the ADF test indicates existence of a unit root and non-stationarity of these time series. However, we can reject the null hypothesis for all the ROCOF time series with p-values below or equal to 0.01. That indicates the stationarity of the ROCOF time series.

2) KPSS test: In contrast to the ADF test, the KPSS test considers that the null hypothesis (H_0) represents the absence of a unit root and the alternative hypothesis (H_1) represents the presence of a unit root.

Similarly, we applied the KPSS test on 160,000 time series, each having 1,000 sample points, of PMU data. We used the command "KPSS.test" in R software to determine the stationarity by calculating the p-values. The percentages of time series with p-values above 0.01 (accept the null hypothesis (H_0)) and time series with p-values below or equal to 0.01 (reject the null hypothesis (H_0)) are shown in Table I.

The second column (p-value$\leq$ 0.01) under the KPSS test shows that the PMU data (V, f, and θ) are non-stationary. The percentages of non-stationary PMU data (V, f, and θ) are 93.62%, 99.70%, and 78.63%, respectively. On the other hand, the ROCOF time series are mostly stationary (p-value$>$ 0.01) with a percentage of 91.18%.

Table I. Percentages of stationary (2nd and 3rd columns) and non-stationary (1st and 4th columns) time series

PMU data	ADF		KPSS	
	$p > 0.01$	$p \leq 0.01$	$p > 0.01$	$p \leq 0.01$
$Voltage$	81.33%	18.67%	6.38%	93.62%
$Frequency$	96.86%	3.14%	0.30%	99.70%
$Angle$	45.88%	54.12%	21.37%	78.63%
$ROCOF$	0%	100%	91.18%	8.82%

III. FRACTALITY OF PMU DATA

Fractal time series have the unique characteristics of exhibiting a slow (non-exponential) decay of the ACF, heavy-tailed probability density function (PDF), and power spectral density function in the form $1/f^{\beta}$. The slow decay of the ACF indicates a long-range memory (dependence) in the time
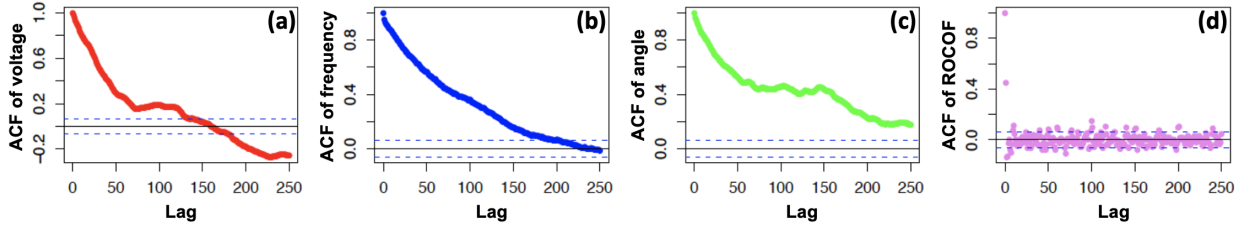
Figure 2. ACFs of PMU data: (a) Voltage magnitude (b) Frequency (c) Angle (d) ROCOF

series, characterized by persistent correlation between the time series samples as the lag increases.

In PMU data, the measurement errors are local in time and intermittent whereas the grid dynamics have long-range characteristics at multiple time scales. The existence of long-range correlations originates from the dynamic interactions between the different elements in the power system. The increasing penetration of renewable energy sources reduces the overall inertia and negatively influences the dynamics of the power grid. Therefore, severe frequency deviations and slower recovery times are expected in low-inertia grids. We believe that the persistent frequency deviations contribute to the long-range correlations.

The modeling of long-range correlated data differs from the short-range correlated data. Given that the short- versus long-range correlation might not be clear, it is crucial to quantify the fractality of PMU data via several statistical methods. We use these methods to calculate three fractal parameters: scaling exponent $(\alpha^\star)$, differencing parameter (d), and power exponent (β). The relationships among these parameters for non-stationary time series are shown in Eq. (1),

$$d = \alpha^\star - 0.5 = \beta/2. \tag{1}$$

In the bulk of this section, we estimate the three parameters of the PMU data using three corroborating methods: (1) DFA method [7], (2) Geweke and Porter-Hudak (GPH) method [14], and (3) Power Spectral Density (PSD) method [15].

Table II. The means and standard deviations of the scaling exponents $(\alpha^\star)$, the differencing parameters (d), and the power exponents (β)

Parameter	Voltage	Frequency	Angle	ROCOF
$\alpha^\star$	1.18 ± 0.18	1.58 ± 0.21	1 ± 0.27	0.41 ± 0.13
d	0.86 ± 0.17	1 ± 0.14	0.63 ± 0.26	0.03 ± 0.19
β	1.75 ± 0.29	1.9 ± 0.21	1.44 ± 0.37	-0.11 ± 0.24

A. DFA method

Scaling (Hurst) exponent is a measure of long-range memory in the time series. It is related to the rate of decay of the ACF. It has been shown that ACF does not provide an accurate estimation of long-range memory, especially in short-length and non-stationary time series [16].

The DFA was developed as a reliable tool to quantify the correlations in non-stationary time series. The DFA measures the average variance of detrended signal over different window sizes. The slope of the linear regression line in double-logarithmic coordinates represents the scaling exponent of the

time series. The method was first introduced in 1994 [7] to study the long-range dependence of DNA nucleotides.

We applied the DFA method on the PMU data (V, f, θ, and ROCOF) to calculate the scaling exponent, as shown in Table II. The voltage magnitude, frequency, phase angle, and ROCOF time series have average scaling exponents of 1.18, 1.58, 1, and 0.41, respectively.

For a non-stationary time series, the scaling exponent $(\alpha^\star)$ falls between 1 and 2. The time series has a long-range correlation not following the power law if it has scaling exponent higher than 1. Based on the results in Table II, most of the voltage and frequency time series are non-stationary and long-range correlated with scaling exponents higher than 1 $(\alpha^\star \geq 1)$. Furthermore, the angle time series have scaling exponents distributed between 0.5 and 1.5. That means also the angle time series have long-range correlation, but it could be either stationary $(\alpha^\star < 1)$ or non-stationary $(\alpha^\star > 1)$. Finally, the ROCOF time series are stationary $(\alpha^\star < 1)$ and short-range correlated with scaling exponents very close to the white noise $(\alpha^\star = 0.5)$.

B. GPH method

The ARIMA models can be generalized to ARFIMA models by allowing the differencing parameter (d) to be non-integer. The ARFIMA models are able to model the short-range and long-range correlations in the time series. By comparing the long-correlated time series to the ARFIMA model, we can derive a relationship between the scaling exponent $(\alpha^\star)$ of the time series and the differencing parameter (d) in the ARFIMA.

The GPH method [14] is a semi-parametric method to estimate the differencing parameter (d). The method does not assume any knowledge of the short-range memory component in the non-stationary time series. It estimates the differencing parameter using linear regression of the log periodogram. The periodogram of any time series, $m(n)$, with N samples is defined as

$$I_N(\omega_k) = \frac{1}{2\pi N} \left| \sum_{n=0}^{N-1} m(n) e^{-i\omega_k n} \right|^2, \tag{2}$$

where ω_k represents the kth Fourier frequency, $2\pi k/N$. On the other hand, the spectral density of any weakly-stationary time series, $m(n)$, with long-range memory is $f(\lambda) = |2\sin(\lambda/2)|^{-2d} f^*(\lambda)$.

$f^*(\lambda)$ is the spectral density of the short-range memory component of the time series, $m(n)$. λ is the frequency. By comparing the logarithm of the periodogram and the logarithm of the spectral density at low frequencies, the estimation of the

differencing parameter ($\hat{d}$) is performed by linear regression of $\log(I_N)$ on $-2\log|2\sin(\lambda/2)|$ at low frequencies. We calculate the differencing parameter of PMU data using the command "fdGPH" from the package "fracdiff" in R software.

As shown in Table II, the differencing parameters of the PMU data have mean values between 0.5 and 1, except the ROCOF that has an average differencing parameter approximately equal to 0. Using Eq. (1), the results in Table II confirm that the PMU data (V, f, and θ) are non-stationarity ($d > 0.5$) and long-range correlated ($d > 0$). On the other hand, the ROCOF time series are stationary ($d < 0.5$) and short-range correlated ($d = 0$) with differencing parameters distributed around 0.

C. PSD method

The PSD of any signal quantifies the power at different frequency components in the signal. The uncorrelated time series or white noise has equal power at all frequencies with power exponent (β) of 0. On the other hand, the long-correlated time series has linear PSD on the log-log scale with non-zero power exponent (β).

The PSD method [15] estimates the PSD exponent of non-stationary time series after some modifications to improve the accuracy of the PSD estimation. These modifications include detrending the data using bridge detrending and estimating the power exponent after excluding the high frequency component of the PSD. Using the R code in [17], we estimate the power exponent (β) of PMU data, as shown in Table II.

The power exponents of the PMU data (V, f, and θ) are between 1 and 2. It is clear from the power exponent values that the signals are not random ($\beta = 0$) and possess long-range memory. These exponents show that the PSD of the PMU data falls between pink noise ($\beta = 1$) and brown noise ($\beta = 2$). The ROCOF time series have an average power exponent of -0.11 that is similar to the white noise ($\beta = 0$). Based on the fractal parameters, we conclude that most of the ROCOF time series are stationary and short-range correlated.

The stationarity and short-range correlation in the ROCOF data could be a result of calculating the ROCOF data samples locally in time compared to the other PMU signals. That can justify the ROCOF missing the long-range dependence in their models. Since the best statistical model of the ROCOF is the Autoregressive Moving Average (ARMA) model ($d = 0$), we focus on the ARFIMA models of the other PMU data variables (V, f, and θ) for the rest of the paper.

IV. ARFIMA MODELS OF PMU DATA

We aim in this section to find the best models that fit the PMU data. The selected models should be capable of handling the non-stationarity and fractality of PMU data, as shown in Sec. II and Sec. III, respectively. It is known that the classical models (Autoregressive (AR), Moving-average (MA), ARMA, ARIMA) are suitable for modeling stationary and non-stationary *short-range* correlated time series. Therefore, we adopt the so-called fractional ARIMA or ARFIMA models for PMU data as they are capable of capturing the non-stationarity and long-range memory properties.

ARFIMA was introduced by Granger and Joyeux in 1980 [8]. This model is a generalization of the ARIMA model (d is integer) developed by Box and Jenkins [18] in the sense that the differencing parameter (d) could have fractional (non-integer) values. The ARFIMA models are capable of characterizing the long-range memory in the data by applying the fractional differencing on the time series.

A. ARFIMA Modeling

Let X_t be a zero-mean time series with long-range memory. The ARFIMA(p, d, q) model of X_t is defined in Eq. (3),

$$\Phi_p(B)\Delta^d X_t = \Theta_q(B)\epsilon_t. \tag{3}$$

B is the backshift operator and d is the differencing parameter. $\Phi_p(B)$ is the $p-$order autoregressive polynomial $(1 - \phi_1 B - ... - \phi_p B^p)$. $\Theta_q(B)$ is the $q-$order moving average polynomial $(1 + \theta_1 B + ... + \theta_q B^q)$.

The innovations or residuals, ϵ_t, are i.i.d random variables. They are uncorrelated with zero mean. The ARFIMA model is well defined for α-stable innovations ($0 < \alpha < 2$) with infinite variance [19] and Gaussian innovations ($\alpha = 2$) with finite variance [8][20].

The term Δ^d is the fractional difference operator, $(1 - B)^d$, with non-integer differencing parameter (d). The fractional differencing can be defined as an infinite "binomial" expansion,

$$\Delta^d = (1 - B)^d = \sum_{k=0}^{\infty} \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}(-B)^k, \tag{4}$$

where $\Gamma(\cdot)$ is the Gamma function ($\int\limits_0^{\infty} s^{z-1}e^{-s}ds$).

Assuming that the polynomials, $\Phi(z)$ and $\Theta(z)$, have no common roots and the autoregressive polynomial, $\Phi(z)$, does not have roots in the closed unit disk, the ARFIMA model can be solved for the time series (X_t) as

$$X_t = \Phi^{-1}(B)\Theta(B)(1 - B)^{-d}\epsilon_t = \sum_{k=0}^{\infty} u_k \epsilon_{t-k}, \tag{5}$$

where the coefficients, u_k's, are resulting from the power series expansion of $\Phi^{-1}(z)\Theta(z)(1 - z)^{-d}$. In [19], it has been shown that the series of the ARFIMA model with α-stable innovations, as shown in Eq. (5), converges *almost surely* when

$$-\infty < d < 1 - \frac{1}{\alpha}. \tag{6}$$

That means the ARFIMA model with α-stable innovations ($0 < \alpha \le 2$) is defined for $d < 1 - 1/\alpha$.

B. Model Identification

Throughout this section, we perform the identification of the ARFIMA models on the three time series shown in Figs. 1 (a)-(c) as a representative sample of the PMU data. Since ARFIMA modeling is defined for stationary time series ($-0.5 < d < 0.5$), we should first differentiate the PMU
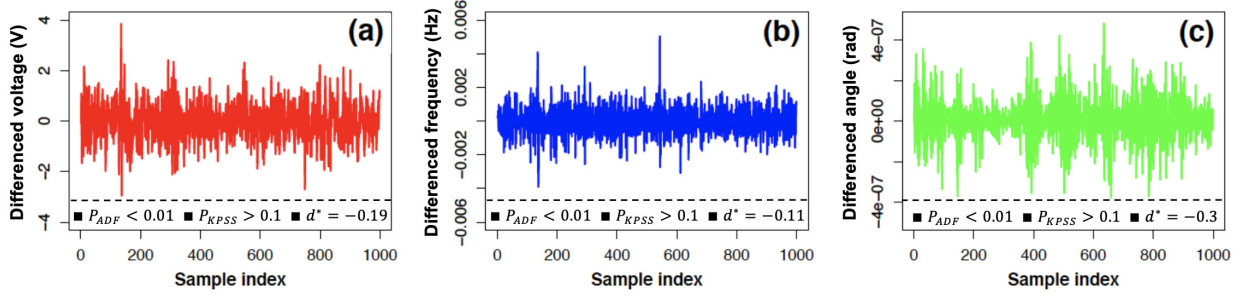
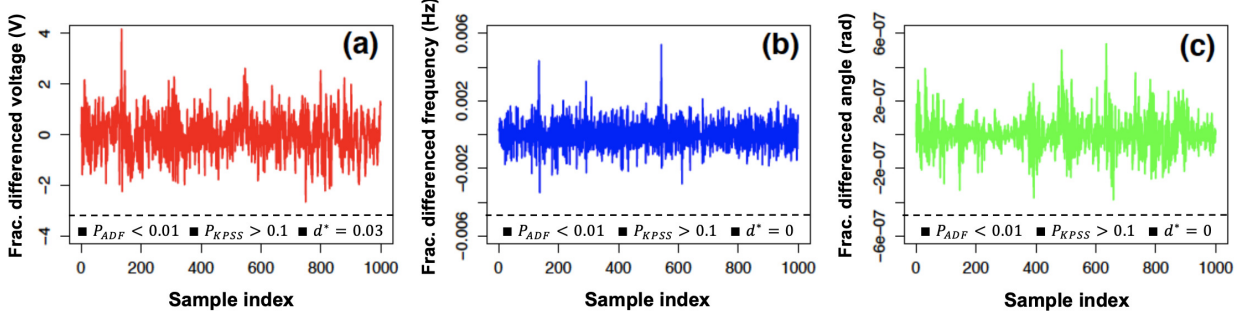Figure 3. First-differenced time series (Y_t) of PMU data: (a) Voltage magnitude (b) Frequency (c) Angle

Figure 4. Fractionally-differenced time series (Z_t) of PMU data: (a) Voltage magnitude (b) Frequency (c) Angle

time series (X_t) to remove any non-stationarity. The first-differenced time series will be confirmed to be a stationary time series,

$$Y_t = X_t - X_{t-1} = (1 - B)X_t. \qquad (7)$$

After the first differencing, the three resulting time series of PMU data are shown in Figs. 3 (a)-(c). Now, applying the ADF and KPSS tests on the three first-differenced time series confirms the stationarity of these series. At significance level of 0.01, the p-values of the two tests after differencing are shown in the legends of Figs. 3 (a)-(c).

Then, the ARFIMA modeling of the PMU data will be carried out on the first-differenced time series. The ARFIMA model of the time series (Y_t) becomes

$$\Phi(B)(1 - B)^{d^*} Y_t = \Theta(B)\epsilon_t, \qquad (8)$$

where $d^* = d - 1$ and the time series (Y_t) is stationary.

As expanded upon in this paper, the difficulty is to compute the d^* parameter. Once the latter is computed, the ARFIMA reduces to the classical Box-Jenkins ARMA modeling of $Z_t = (1 - B)^{d^*} Y_t$. So, the fractional differencing of the time series (Y_t) generates a time series (Z_t) possessing short-range memory.

Using Eq. (4), we fractionally differentiate the three first-differenced time series (Y_t's) of PMU data. The fractional differencing is applied on each differenced time series based on its own differencing parameter (d^*). The three fractionally differenced time series (Z_t's) are shown in Figs. 4 (a)-(c). The differencing parameters before and after fractional differencing are shown in the legends of Fig. 3 and Fig. 4. It is clear that the fractionally differenced time series (Z_t's) have short-range memory with differencing parameters very close to 0.

Therefore, the three time series (Z_t's) of PMU data can now be modeled using the ARMA model,

$$\Phi(B)Z_t = \Theta(B)\epsilon_t. \qquad (9)$$

Finally, we fit the time series, Z_t, using ARMA models with several combinations of autoregressive and moving average polynomials. We cover all the combinations of p and q between 0 and 5, like $(0, d, 0), (1, d, 0), (0, d, 1), ...(5, d, 5)$. Then, we use the two information criteria (AIC and BIC) to compare the different models and choose the best fit. We did not consider p and q higher than 5 because both criteria have higher penalty term for larger number of coefficients.

The best ARMA models of the voltage, frequency, and phase angle are ARMA$(0, 1)$, ARMA$(0, 2)$, and ARMA$(1, 2)$, respectively. The differencing parameter ($d^* + 1$) and the ARMA model parameters constitute initial estimation of the ARFIMA model. The first estimations of the ARFIMA models for the three time series of PMU data are shown in Table III.

Table III. Initial estimation of the ARFIMA models

Data	Model	(Φ_1, Φ_2)	(Θ_1, Θ_2)
Voltage	ARFIMA $(0, 0.81, 1)$	$(-, -)$	$(+0.55, -)$
Frequency	ARFIMA $(0, 0.89, 2)$	$(-, -)$	$(-0.38, -0.17)$
Angle	ARFIMA $(1, 0.70, 2)$	$(+0.61, -)$	$(-0.68, +0.20)$

The various steps and algorithm, as shown in Algorithm 1, for the identification of the ARFIMA model of a time series (X_t) are summarized below:

(1) Test the stationarity of the time series (X_t) via the ADF and KPSS tests. If the time series (X_t) is stationary, we can skip step (2) and go directly to step (3). In this case, the time series (Y_t) is equal to the time series (X_t).

(2) Differentiate the time series (X_t) to generate the stationary time series ($Y_t = X_t - X_{t-1}$).

(3) Estimate the differencing parameter (d^*) of the series (Y_t) in the ARFIMA model ($\Phi(B)(1 - B)^{d^*} Y_t = \Theta(B)\epsilon_t$).

Algorithm 1: Identification of the ARFIMA models

Input: X_t
Output: $model, d, aic, bic$
1   set $aic = \infty$, $bic = \infty$;
2   get P_{ADF} and P_{KPSS} of X_t;   // as shown in Sec. II B
3   **if** $P_{ADF} \leq 0.01$ *and* $P_{KPSS} \geq 0.01$ **then**
4 | $Y_t \leftarrow X_t$;
5   **else**
6 | $Y_t \leftarrow X_t - X_{t-1}$;
7   get d^* from Y_t;   // as explained in Sec. III
8   $d \leftarrow d^* + 1$;
9   $Z_t \leftarrow (1 - B)^{d^*} Y_t$;
10   **for** $p = 0$ *to* 5 **do**
11 | **for** $q = 0$ *to* 5 **do**
12 | | $model^{temp} \leftarrow$ fit $(Z_t, \text{ARMA}(p,q))$;
13 | | $aic^{temp} \leftarrow$ AIC (model);
14 | | $bic^{temp} \leftarrow$ BIC (model);
15 | | **if** $aic^{temp} < aic$ *and* $bic^{temp} < bic$ **then**
16 | | | $model \leftarrow model^{temp}$;
17 | | | $aic \leftarrow aic^{temp}$;
18 | | | $bic \leftarrow bic^{temp}$;

(4) Apply the fractional differencing on the time series (Y_t) to generate the time series ($Z_t = (1-B)^{d^*} Y_t$) with short-range memory only. The time series (Z_t) can be modeled via ARMA models ($\Phi(B)Z_t = \Theta(B)\epsilon_t$).

(5) Estimate the order and parameters of the best ARMA model based on AIC and BIC criteria.

C. Model Estimation

The estimation of the ARFIMA parameters can be conducted using several estimation techniques in either the time domain or the frequency domain [21]. One of the reliable estimators of Gaussian and stable ARFIMA models is Whittle approximate maximum likelihood estimation (MLE) [22][23]. The Whittle estimation of the ARFIMA parameters can be achieved by minimizing

$$Q(\hat{\zeta}) = \sum_{j=1}^{m} \frac{I(\lambda_j)}{f(\lambda_j, \hat{\zeta})}, \tag{10}$$

where $I(\lambda_j)$ is the spectral density function of the time series. The $f(\lambda_j, \hat{\zeta})$ function is the spectral density function of the model at frequency (λ_j). The λ_j's are the Fourier frequencies, $2\pi j/m$. $\hat{\zeta}$ is the vector of unknown parameters ($d^*, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q$).

We start with the parameters of the ARFIMA models in Table III as initial values for the Whittle estimator. The best ARFIMA models of the three PMU time series based on the Whittle estimation are shown in Table IV.

Table IV. Best ARFIMA models of the three time series

Data	ARFIMA Model
Voltage	$(1-B)^{0.80}V_t = (1+0.56B)\epsilon_t$
Frequency	$(1-B)^{0.92}F_t = (1-0.40B-0.17B^2)\epsilon_t$
Angle	$(1-0.61B)(1-B)^{0.70}T_t = (1-0.68B+0.20B^2)\epsilon_t$

V. RESIDUALS OF ARFIMA MODEL

Evaluating the goodness-of-fit of the PMU ARFIMA models requires analyzing the residuals ($\hat{\epsilon}_t$) of these models. Toward that end, we first fit the distribution of the residuals to α-stable distribution and calculate its parameters. This is a very crucial step to be performed because the ARFIMA model is well-defined only for α-stable innovations (residuals), including the Gaussian case ($\alpha = 2$). Moreover, a good ARFIMA model of the data should result in uncorrelated and independent residuals to ensure all the information in the data is represented. We use Ljung-Box test to check the independence of the residuals.

A. Residuals of the ARFIMA Model

The residuals ($\hat{\epsilon}_t$) or innovations of the ARFIMA model can be estimated using

$$\hat{\epsilon}_t = \Phi(B)\Delta^d \Theta^{-1}(B) X_t. \tag{11}$$

We calculate the residuals of the ARFIMA models of voltage (ARFIMA$(0, d, 1)$), frequency (ARFIMA$(0, d, 2)$), and phase angle (ARFIMA$(1, d, 2)$). The resulting residuals of the three models of PMU data are shown in Figs. 5 (a)-(c). Because $d > 1 - 1/\alpha$, expressing X_t as a series in $\hat{\epsilon}_{t-j}$ would result in lack of almost sure absolute convergence [19, Sec. 7.13]. To go around this difficulty, we utilize Eq. (8) to express Y_t as a series in ϵ_{t-j}. Now the relevant differencing parameter is d^*, and since $d^* < 1 - 1/\alpha$, the representation of Y_t as a series in ϵ_{t-j} now converges absolutely almost surely.

It is worth fitting the residuals of the PMU ARFIMA models to α-stable distribution and calculating its parameters (α, β, γ, and δ) [19]. The histograms of the ARFIMA model residuals and their best α-stable fit based on Koutrouvelis regression method [24] are shown in Figs. 6 (a)-(c). To test the normality of the model residuals, we generate the quantile-quantile (Q-Q) plots for the residuals of the ARFIMA models, as shown in Figs. 7 (a)-(c).

The residuals of the voltage ARFIMA model follow a Gaussian distribution ($\alpha \approx 2.0$). For the voltage time series, the ARFIMA model is defined because its differencing parameter, $d_v^* = -0.2$, is less than 0.5.

The residuals of the ARFIMA models of the frequency and angle time series follow α-stable distribution with α equals to 1.80 and 1.48, respectively. The differencing parameter (d_f^*) of the frequency time series is equal to -0.08 and smaller than $1 - (1.8)^{-1}$. Also, the differencing parameter (d_t^*) of the angle time series is equal to -0.3 and smaller than $1 - (1.48)^{-1}$. Satisfying the condition in Eq. (6) validates the models of Y_t.

The ARFIMA model of the voltage, $\Delta^d V_t = \Phi^{-1}(B)\Theta(B)\epsilon_t$, is driven by the residual noise (ϵ_t), and the question arises as to what the residual noise really is, physically rather than statistically speaking. The linearized fluctuating power flow equations on the other hand yield the V_t fluctuation as a function of the fluctuation of the power injected. In view of the stochastic modeling of the fluctuating power of wind farms [25], it is fair to conjecture that near wind farm, where penetration of the renewables is observed, $\Phi^{-1}(B)\Theta(B)\epsilon_t$ may contain the colored noise
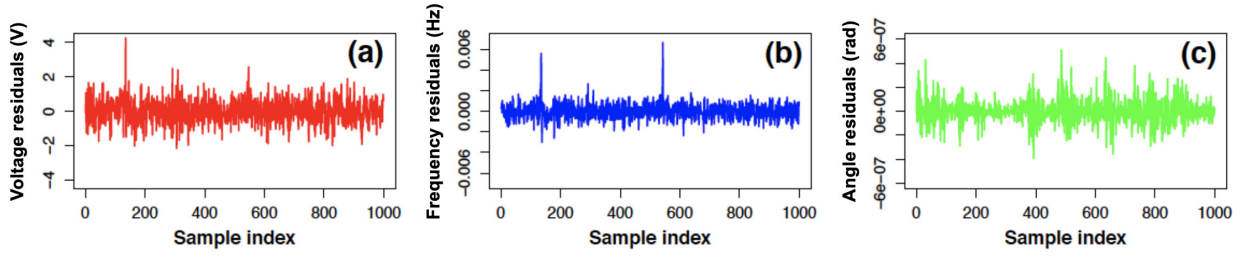
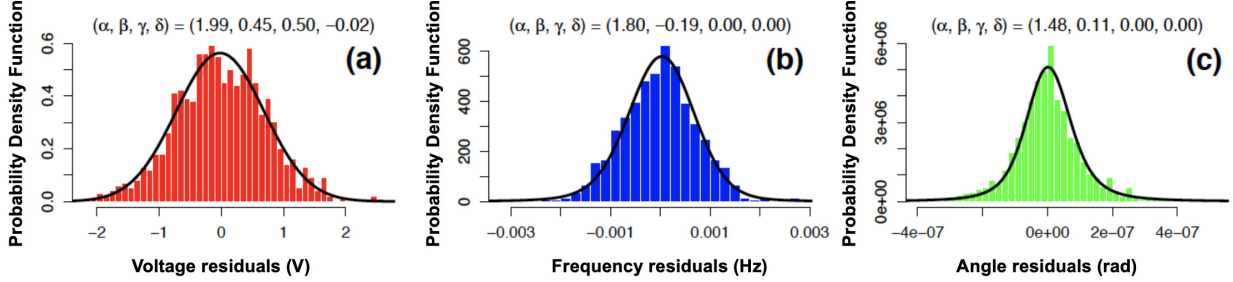Figure 5. Residuals of ARFIMA models of PMU data: (a) Voltage magnitude (b) Frequency (c) Angle

Figure 6. Sample density functions of the residuals of the ARFIMA models: (a) Voltage magnitude (b) Frequency (c) Angle
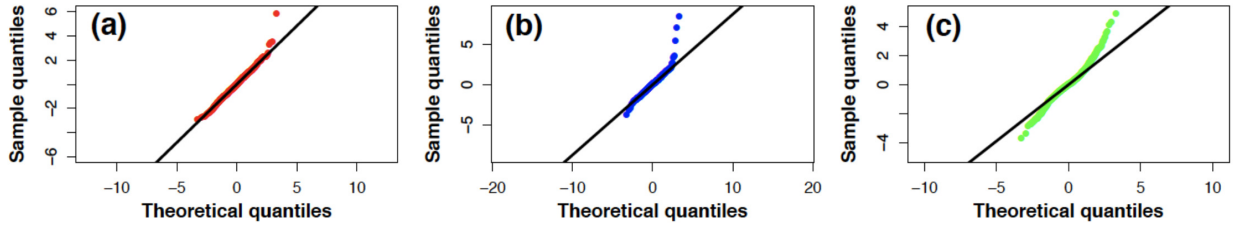

Figure 7. Q-Q plots of the residuals of the ARFIMA models: (a) Voltage magnitude (b) Frequency (c) Angle

fluctuating power generated by wind farms. The observed near Gauss property of the residuals, hence the power fluctuations, corroborates the observation made in [25] that such fluctuations at the output of a wind farm are Gaussian.

B. Correlation of the ARFIMA Residuals

Any significant dependence in the residuals of a fitted model can be diagnosed by calculating the ACF of the residuals. So, we calculate the autocorrelations ($\hat{r}_k$) of the residuals ($\hat{\epsilon}_t$) at different lags (k) using the formula,

$$\hat{r}_k = \frac{\sum_{t=k+1}^{n} \hat{\epsilon}_t \hat{\epsilon}_{t-k}}{\sum_{t=1}^{n} \hat{\epsilon}_t^2}. \tag{12}$$

The ACFs of the residuals from the three ARFIMA models of the PMU data are shown in Figs. 8 (a)-(c). Inspecting graphically the first 50 autocorrelation lags shows that most of the autocorrelations are within the 95% confidence band ($\pm 1.96/\sqrt{n}$), except a small number of outliers. So, we can conclude that we have independent and uncorrelated residuals.

A more formal way of analyzing the dependence in the residuals is through the Ljung-Box test [26]. The test was introduced by Ljung and Box in 1978 as a modified version of the Box and Pierce test [27]. The test provides joint instead of individual testing of the adequacy of the fitted model. The Ljung and Box test is based on the statistic,

$$Q(\hat{r}) = n(n+2) \sum_{k=1}^{m} (n-k)^{-1} \hat{r}_k^2, \tag{13}$$

where n is the size of the residuals series and m is the number of lags. The statistic $Q(\hat{r})$ would asymptotically follow a Chi-squared (χ^2) distribution with $m - (p+q)$ degrees of freedom under the assumption that the residuals ($\hat{r}_k$) are normally distributed. The test tends to be more strict in case of non-Gaussian and heavy tailed distributions.

The null hypothesis (H_0) is that residuals are independent and there is no lack of fit. On the other hand, the alternative hypothesis (H_1) indicates dependence among residuals and a lack of fit. We can reject the null hypothesis of the test at h degrees of freedom, and significance level, 0.05, if $Q(\hat{r}) > \chi^2_{0.95,h}$, where $\chi^2_{0.95,h}$ is the 0.95-quantile of the χ^2-distribution at h degrees of freedom. The p-value of the Ljung-Box test represents the probability of having Q higher than the calculated, $Q(\hat{r})$, in the corresponding χ^2-distribution. At significance level of 0.05, we reject the null hypothesis (H_0) if the p-value < 0.05 and we fail to reject the null hypothesis (H_0) if the p-value > 0.05.

As shown in Figs. 9 (a)-(c), the p-values of the residuals of the three ARFIMA models are higher than the significance level of 0.05 (dotted line), so we cannot reject the Null hypothesis (H_0). That means we do not have evidence that the residuals of the selected ARFIMA models are autocorrelated.

VI. ANOMALY DETECTION VIA ARFIMA MODELS

Anomalies in PMU data could be related to data quality issues or physical events. Several data mining and machine learning algorithms have been proposed to detect and classify
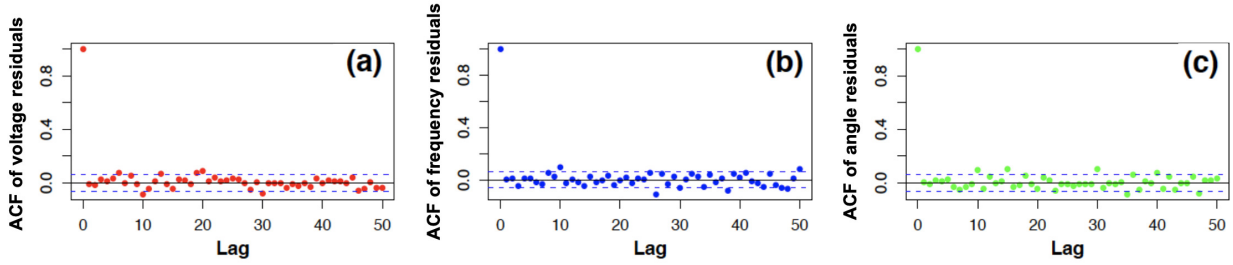
Figure 8. ACFs of the residuals of the ARFIMA models: (a) Voltage magnitude (b) Frequency (c) Angle
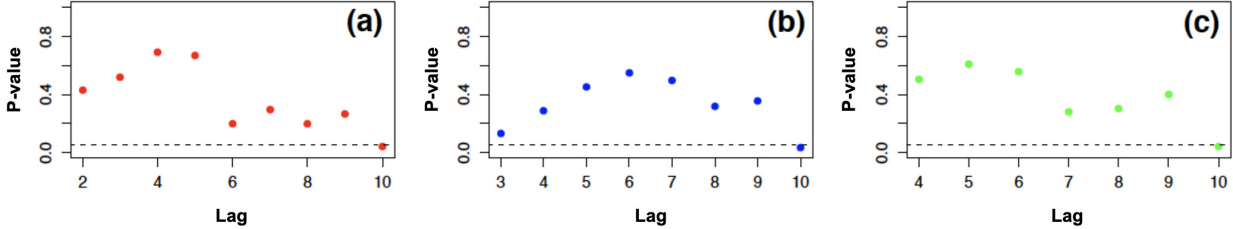

Figure 9. P-values of Ljung-Box statistic of the ARFIMA residuals: (a) Voltage magnitude (b) Frequency (c) Angle

these anomalies, such as ensemble learning [28], convolutional neural network [29], principal component analysis [30]. These algorithms have been proven to detect anomalies at different response times and accuracy rates.

In this section, we propose a novel method capable of detecting the hidden changes in the power grid before critical transitions, like power blackouts. Using the 2012 Indian blackout, we show that these changes can be detected in the system frequency several minutes ahead of the blackout.

A. Effect of Data Quality Issues on ARFIMA Modeling

Here, we discuss the impact of real-world data quality issues (data corrupted by noise, packet drop, bad data) on the fractal characteristics of PMU data. As any small deviation in the estimated differencing parameter results in inaccurate data models, we devote this section to study the effect of these issues on the estimated differencing parameter.

1) Fractal behavior of data with induced noise: The PMU data contains measurement errors and noise induced by the PMU device or the communication channel. It has been shown that the measurement errors in PMU data behave like a gaussian white noise ($d = 0$) [31]. We aim to examine the effect of the white noise on the differencing parameters of PMU data for different signal-to-noise ratios (SNRs).

At each SNR, we calculate the differencing parameters of 100 PMU time series (induced by noise) for each variable (V, f, and θ). Then, we plot the distribution of the relative errors in the differencing parameter versus the corresponding SNR, as shown in Fig. 10 (a). In [31], the authors show that the SNR of different PMU variables from field data is approximately 45 dB. Based on our results, at a SNR of 45 dB, the means of the relative errors for the voltage, frequency, and phase angle are 0.12%, 0.11%, and 0.15%, respectively. It is clear that the white noise has a minimal effect on the estimated differencing parameters for SNR $\geq$ 20 dB.

2) Fractal behavior after data loss: As part of the wide area monitoring system in the SG, some PMU data loss as a result of packet drop and loss (User Datagram Protocol

(UDP) protocol) or congestion followed by retransmission and delay (Transmission Control Protocol/Internet Protocol (TCPIP) protocol) can be expected.

We evaluate the effect of different percentages of missing data samples on the differencing parameters of PMU data. At each percentage, we calculate the differencing parameters of 100 PMU time series for each variable (V, f, and θ). Then, we plot the distribution of the relative errors in the differencing parameter versus the corresponding missing data percentage, as shown in Fig. 10 (b). Practically, the average percentage of missing samples in PMU data is between $5 - 10\%$ [32]. At a percentage of 10%, the means of the relative errors for the voltage, frequency, and phase angle are 5.47%, 4.82%, and 9.45%, respectively. It seems that the relative errors increase gradually as the data loss percentage increases. Up to a data loss percentage of 50%, the median relative error in the differencing parameter is approximately 12%.

3) Fractal behavior of bad data: The distinction between a disturbance and bad data is of great importance to help in making informed decisions. In Table II, we have calculated the fractal parameters for a large set of PMU data collected under normal operating conditions. The statistics of these parameters could provide thresholds for the expected fractal behavior of the PMU signals. Any persistent deviation from these limits should be labeled as a potential bad data. On the other hand, an abrupt change in the fractal behavior could provide an early warning for a major transition.

Under the assumption of Gaussian bad data, we distribute the bad data samples randomly in the PMU time series. For different percentages of bad data samples, we calculate the differencing parameters of 100 PMU time series for each variable (V, f, and θ). Then, we plot the distribution of the relative errors in the differencing parameter versus the corresponding bad data percentage, as shown in Fig. 10 (c). The expected percentage of bad PMU data samples in real power grid is below 4% [33]. At a percentage of 4%, the means of the relative errors for the voltage, frequency, and phase angle are 5.38%, 5.89%, and 6.31%, respectively.

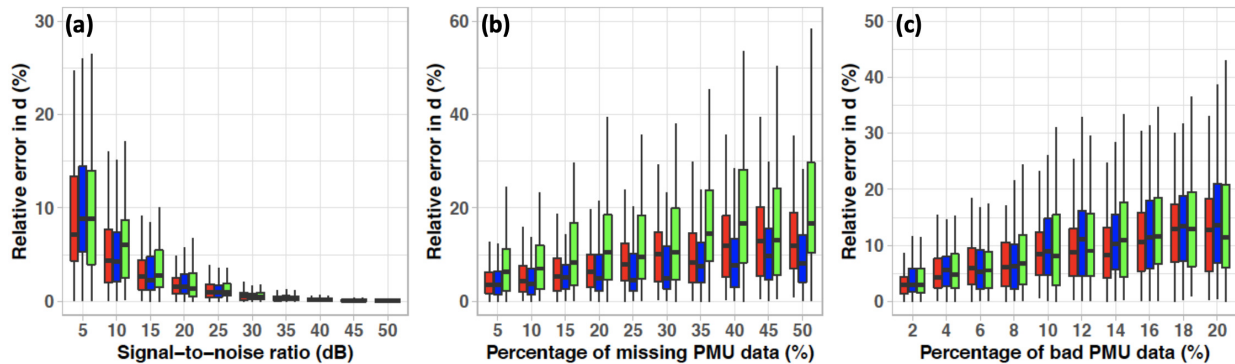Figure 10. Relative error in the differencing parameters due to: (a) Induced noise (b) Missing data (c) Bad data

B. Detection of the 2012 Indian Blackout

In complex systems, the critical slowing down [2] is one of the most important indicators to the proximity to a critical transition. The critical slowing down could lead to certain changes in the characteristics of system fluctuations, including increased lag-1 autocorrelation (ϕ_1) [34] and increased scaling exponent ($\alpha^\star$) [3]. To verify the existence of the critical slowing down in the power grid, we can monitor the changes in the ARFIMA parameters (d and ϕ_1) of the data.

One of the largest power blackouts in the history is the 2012 Indian blackout [35]. The blackout had two events occurring over two consecutive days and affected more than 600 millions people. The recorded data of the system frequency is shown in Fig. 11 (a). Our proposed early warning proceeds from the derivation of accurate ARFIMA models of the frequency signal toward the Indian blackout. In ARFIMA modeling, the Whittle approximate MLE has been proven to outperform other estimation methods in terms of bias and errors [36].

As shown in Fig. 11 (a), we estimate the ARFIMA models of the system frequency inside 16.67-minute sliding window with 1-minute shift toward the 2012 Indian blackout. The ARFIMA parameters (d and ϕ_1) inside the sliding window are shown in Figs. 11 (b)-(c). By examining the plots of the differencing and lag-1 autoregressive parameters, we notice an increase in their values toward the blackout with the increase more noticeable on the differencing parameter. The increase is starting around 11-12 minutes before the blackout. The mean of the differencing parameter is shifting from 1.02 to 1.23, and the mean of the lag-1 autoregressive parameter is shifting from 0.2 to 0.31. This increase could provide an early warning of the proximity to a major transition or blackout. The increase in the ARFIMA parameters (d and ϕ_1) of the frequency could be a sign of a critical slowing down in the power grid.

Currently, the PMUs play a critical role in the power grid operation. Extracting the useful information embedded in the massive amount of PMU data is a challenge, especially in real-time applications. Therefore, our statistical models serve as anomaly detection tool that can identify these anomalies and warn the operators to take the proper actions.

VII. Conclusion

The starting point of this paper has been evidence of non-stationarity in PMU data using unit root tests, ADF and KPSS. We then followed up with different methods (DFA, GPH, and PSD) that capitalize on non-stationarity to compute the fractal parameters, showing existence of long-range memory in the PMU data. The estimated fractal parameters ($\alpha^\star$, d, and β) showed consistency among the three methods. Since most of the PMU data have long-range memory, we have used the ARFIMA models to reproduce the PMU data, and its parameters were estimated using the Whittle estimator. The goodness-of-fit was analyzed through testing the autocorrelation and independence in the model residuals.

Practical grid applications already emerge, and will be subjects of further investigations–most importantly, change point detection of ARFIMA parameters to detect abnormal events with the specific property that the detection delay is minimized given an upper bound on the false alarm rate. Last but not least, ARFIMA models of the voltage (V) and the angle (θ) are data driven substitutes for the power flow equations in a symbiotic approach that remains to be developed.

References

[1] B. Carreras, D. Newman, I. Dobson, and A. B. Poole. Evidence for self-organized criticality in a time series of electric power system blackouts. *IEEE Transactions on Circuits and Systems*, 51(9), September 2004.

[2] E. Cotilla-Sanchez, P. D. H. Hines, and C. M. Danforth. Predicting critical transitions from time series synchrophasor data. *IEEE Transactions on Smart Grid*, 3(4):1832–1840, Dec 2012.

[3] L. Shalalfeh, P. Bogdan, and E. Jonckheere. Kendall's tau of frequency hurst exponent as blackout proximity margin. In *2016 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 466–471, Nov 2016.

[4] AS Bakshi et al. Report of the enquiry committee on grid disturbance in northern region on 30th july 2012 and in northern, eastern & north-eastern region on 31st july 2012. *New Delhi, India*, 2012.

[5] J. Sia, Edmond Jonckheere, Laith Shalalfeh, and Paul Bogdan. PMU change point detection of imminent voltage collapse and stealthy attacks. Dec 2018.

[6] L. Shalalfeh, P. Bogdan, and E. Jonckheere. Evidence of long-range dependence in power grid. In *2016 IEEE Power and Energy Society General Meeting (PESGM)*, pages 1–5, July 2016.
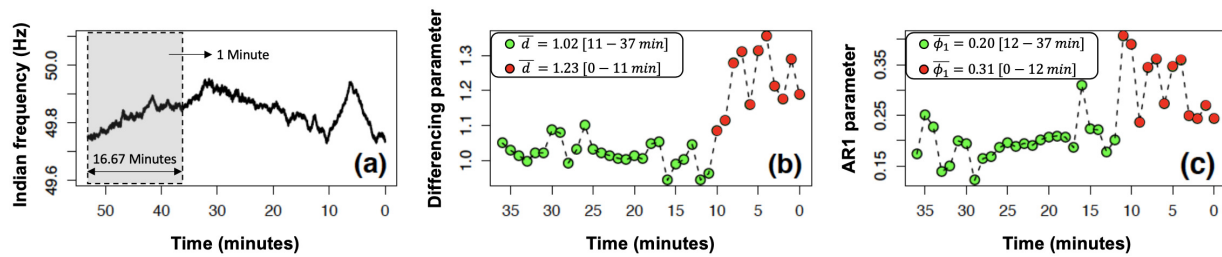
Figure 11. ARFIMA modeling before the Indian blackout: (a) Frequency (b) Differencing parameter (c) Lag-1 AR parameter

[7] C.-K. Peng et al. Mosaic organization of DNA nucleotides. *Phys Rev E*, 49(2), 1994.

[8] C. W. J. Granger and R. Joyeux. An introduction to long-range time series models and fractional differencing. *Journal of Time Series Analysis*, 1:15–30, 1980.

[9] H. Lavicka and J. Kracik. Fluctuation analysis of electric power loads in Europe: correlation multifractality vs. distribution function multifractality. June 2017.

[10] L. Shalalfeh, P. Bogdan, and E. Jonckheere. Modeling of PMU data using arfima models. In *2018 Clemson University Power Systems Conference (PSC)*, pages 1–6, Sep. 2018.

[11] M. Pignati et al. Real-time state estimation of the EPFL-campus medium-voltage grid by using PMUs. In *2015 IEEE PES Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, Feb 2015.

[12] PMU data from EPFL. http://nanotera-stg2.epfl.ch/. [Online; accessed 12-April-2020].

[13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[14] John Geweke and Susan Porter-Hudak. The estimation and application of long memory time series models. *Journal of Time Series Analysis*, 4(4):221–238, 1983.

[15] A. Eke et al. Physiological time series: distinguishing fractal noises from motions. *Pflugers Archives - Eur J Physiol*, 439, 2000.

[16] A. V. Coronado and P. Carpena. Size effects on correlation measures. *J Biol Phys*, 31:121–33, 01 2005.

[17] E. Stroe-Kunold, T. Stadnytska, J. Werner, and S. Braun. Estimating long-range dependence in time series: an evaluation of estimators implemented in R. *Behavior Research Methods*, 41:909–923, 2009.

[18] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

[19] G. Samorodnitsky and M. Taqqu. *Stable Non-Gaussina Random Processes–Stochastic Models with Infinite Variances*. Chapman & Hall, CRC, 1994.

[20] J. R. M. Hosking. Fractional differencing. *Biometrika*, 68:165–176, 1981.

[21] Jan Beran. *Statistics for long memory processes*. Monographs on statistics and applied probability;61. Chapman and Hall, 1994.

[22] Robert Fox and Murad S. Taqqu. Large-sample properties of parameter estimates for strongly dependent stationary gaussian time series. *Ann. Statist.*, 14(2), 06 1986.

[23] L. Giraitis and D. Surgailis. A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotic normality of whittle's estimate. *Probability Theory and Related Fields*, 1990.

[24] Ioannis A. Koutrouvelis. Regression-type estimation of the parameters of stable laws. *Journal of the American Statistical Association*, 75(372):918–928, 1980.

[25] Joaquin Mur-Amada and Jesús Sallán-Arasanz. Power fluctuations in a wind farm compared to a single turbine. In Gesche Krause, editor, *From Turbine to Wind Farms*, chapter 6. IntechOpen, Rijeka, 2011.

[26] G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2), 1978.

[27] G. E. P. Box and D. A. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332):1509–1526, 1970.

[28] M. Zhou et al. Ensemble-based algorithm for synchrophasor data anomaly detection. *IEEE Transactions on Smart Grid*, 10(3):2979–2988, 2019.

[29] S. Basumallik, R. Ma, and S. Eftekharnejad. Packet-data anomaly detection in pmu-based state estimator using convolutional neural network. *International Journal of Electrical Power & Energy Systems*, 107:690–702, 2019.

[30] A. A. Imayakumar, A. Dubey, and A. Bose. Anomaly detection for primary distribution system measurements using principal component analysis. In *2020 IEEE Texas Power and Energy Conference (TPEC)*, pages 1–6, 2020.

[31] M. Brown, M. Biswal, S. Brahma, S. J. Ranade, and H. Cao. Characterizing and quantifying noise in pmu data. In *2016 IEEE Power and Energy Society General Meeting (PESGM)*, pages 1–5, 2016.

[32] M. Wang. Data quality management of synchrophasor data in power systems by exploiting low-dimensional models. In *2017 51st Annual Conference on Information Sciences and Systems (CISS)*, pages 1–2, 2017.

[33] I. Idehen, W. Jang, and T. Overbye. Pmu data feature considerations for realistic, synthetic data generation. In *2019 North American Power Symposium (NAPS)*, pages 1–6, Oct 2019.

[34] Anthony R. Ives. Measuring resilience in stochastic systems. *Ecological Monographs*, 65(2):217–233, 1995.

[35] L. L. Lai, H. T. Zhang, C. S. Lai, F. Y. Xu, and S. Mishra. Investigation on july 2012 indian blackout. In *2013 International Conference on Machine Learning and Cybernetics*, volume 1, pages 92–97, 2013.

[36] J. Smith, N. Taylor, and S. Yadav. Comparing the bias and misspecification in ARFIMA models. *Journal of Time Series Analysis*, 18(5):507–527, September 1997.