Fractional Dynamics of PMU Data

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

***Abstract*—Phasor Measurement Units (PMUs) are collecting real-time measurements from the smart grid as part of the Wide-Area Monitoring System (WAMS). The increase in the number of installed PMUs at the transmission and distribution levels of the grid is accompanied by an increasingly large amount of collected data calling for advanced data analytics techniques. The statistical characteristics of the PMU data are utilized to perform accurate modeling and estimation of the power system variables (voltage, frequency, and phase angle). Based on real PMU data collected from the EPFL campus grid, we show that most of the PMU data are non-stationary based on the Augmented Dickey-Fuller and Kwiatkowski-Phillips-Schmidt-Shin tests. Then, we provide evidence for the fractality of PMU data by estimating the differencing parameter (d) in the Autoregressive Fractionally Integrated Moving Average (ARFIMA) model. Finally, it is shown that the 2012 Indian blackout is accompanied by a change point in the differencing parameter opening the road to event detection by ARFIMA monitoring.**

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

I. Introduction

SMART Grid (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 increasing number of electric vehicles. Therefore, wide-area monitoring, protection, and control systems will have an important role in the future smart grid by securing 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 Phasor Measurement Units (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, that exposes the fast dynamic events and 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 foundation of the Hurst exponent results 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 smart grid.

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 fractality 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 the 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 best ARFIMA model. Secondly, we show that ARFIMA models are well suited for PMU data because their residuals follow Gaussian and α-stable distributions [11]. At the end, we validate our ARFIMA models by showing that the residuals are uncorrelated and independent.

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. Statistical Characteristics of PMU Data

We first provide a description of the PMU data and the power grid from which that data was collected. Then, we investigate the stationarity of the data using 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

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

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

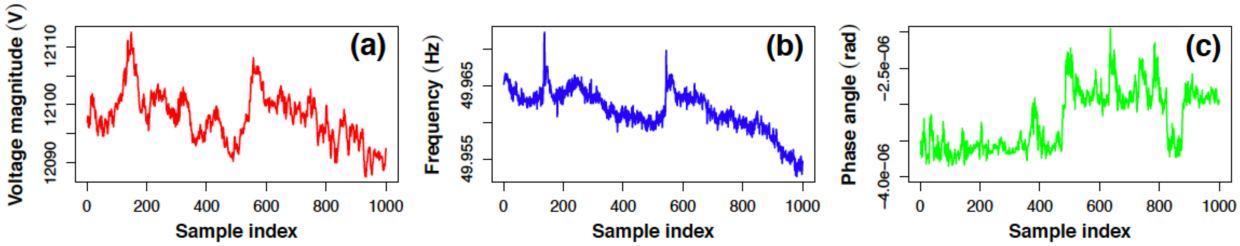

Figure 1. PMU data collected from the EPFL campus grid in 2014: (a) Voltage magnitude (b) Frequency (c) Angle

(f), 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 EPFL campus grid as part of their real-time state estimation project [12]. 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), and unwrapped phase angle (θ). The data set consists of 120,000 time series (1000 samples each) of the three variables collected from the campus grid in January, April, June, and December 2014 [13]. In Figs. 1 (a)-(c), we show 1000-sample time series of voltage magnitude (red), frequency (blue), and angle (green).

B. Stationarity

We can have a glimpse at the stationarity or the lack thereof of PMU data by calculating their autocorrelation functions (ACFs). The autocorrelation functions of the voltage magnitude (red), frequency (blue), and phase angle (green) are shown in Figs. 2 (a)-(c), respectively. The autocorrelation functions show a slow hyperbolic decay compared to the exponentially decaying one in random time series. The slow decay of the autocorrelation function could be a sign of lack of stationarity.

More formally, we test the stationarity of PMU measurements (V, f, and θ) 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 Auto-Regressive polynomial of the time series.

1) Augmented Dickey-Fuller (ADF) Test: The ADF test can classify 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 120,000 time series of PMU data (V, f, and θ), each time series contains 1000 samples (20 seconds). Using the command "adf.test" in R software, 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 (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.

2) Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test: In contrast to the ADF test, the KPSS test considers that the null hypothesis (H_0) represents the absence of unit root and the alternative hypothesis (H_1) represents the presence of unit root.

Similarly, we applied the KPSS test on 120,000 time series 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.

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%	06.38%	93.62%
Frequency	96.86%	03.14%	00.30%	99.70%
Angle	45.88%	54.12%	21.37%	78.63%

III. Fractality of PMU Data

Fractal time series have the unique characteristics of exhibiting a slow (non-exponential) decay of the autocorrelation function (ACF), heavy-tailed probability density function (PDF), and power spectral density function in the form $1/f^{\beta}$. The slow decay of the autocorrelation function indicates a long-range memory (dependence) in the time series, characterized by persistent correlation between the time series samples as the lag increases.

We quantify the fractality and long-range dependence in PMU data using the fractality parameters, scaling exponent ($\alpha^{\star}$), differencing parameter (d), and power exponent (β). The relationships among these parameters for non-stationary Gaussian 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) Detrended Fluctuation Analysis (DFA) [7], (2) Geweke and Porter-Hudak (GPH) method [14], and (3) Power Spectral Density (PSD) method [15].

A. Detrended Fluctuation Analysis (DFA)

DFA is a robust method to estimate the scaling exponent ($\alpha^{\star}$) of non-stationary time series. The method was first introduced in 1994 [7] to study the long-range dependence of DNA nucleotides. One of the main advantages of DFA method is dealing with the non-stationarity in the data using detrending.

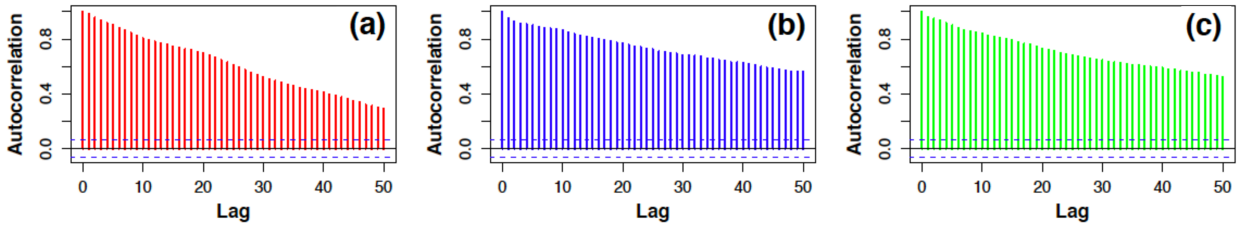

Figure 2. Autocorrelation functions of PMU data: (a) Voltage magnitude (b) Frequency (c) Angle

We applied the DFA method on the PMU data (V, f, and θ) to calculate the scaling exponent. The results show that the voltage magnitude, frequency, and phase angle time series have average scaling exponents of 1.18, 1.58, and 1.00, respectively. Most of the voltage and frequency time series are non-stationary with scaling exponents higher than 1 ($\alpha^\star \geq 1$). Additionally, the angle time series have scaling exponents distributed between 0.5 and 1.5. That means the angle time series could be either stationary ($\alpha^\star < 1$) or non-stationary ($\alpha^\star > 1$). The results are consistent with ADF and KPSS tests on the angle time series.

Moreover, the data have long-range dependence ($\alpha^\star \neq 0.5$) that is not following the power law. Using Eq. (1), the distributions of the differencing parameter (d) are shown in the first columns in Figs. 3 (a)-(c).

B. Geweke and Porter-Hudak (GPH) Method

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 of the series in non-stationary time series. It estimates the differencing parameter using linear regression of the log periodogram. The periodogram of any time series, $m(t)$, with n samples is defined as

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

where ω_k represents the k^{th} Fourier frequency, $2\pi k/n$. On the other hand, the spectral density of any weakly-stationary time series, $m(t)$, with long-range memory is

$$f(\lambda) = |\, 2\sin(\lambda/2)\,|^{-2d} f^*(\lambda). \quad (3)$$

$f^*(\lambda)$ is the spectral density of the short-range memory component of the time series, $m(t)$. 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. The differencing parameters of the PMU data have mean values between 0.5 and 1.0. Similarly, that indicates the non-stationarity ($d > 0.5$) and long-range memory ($d > 0$) of the PMU data. The distributions of the differencing parameter using GPH method and Eq. (1) are shown in the second columns in Figs. 3 (a)-(c).

C. Power Spectral Density (PSD) Method

The PSD method [15] estimates the power spectral density 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. The power exponents have mean values between pink noise ($\beta = 1.00$) and brown noise ($\beta = 2$). It is clear from the power exponent values that the signals are not random ($\beta = 0$) and possess a long-range memory. We further calculated the differencing parameters of the PMU data from the PSD method using Eq. (1). The distributions of the differencing parameter using the PSD are shown in the third column in Figs. 3 (a)-(c).

IV. ARFIMA Models of PMU Data

ARFIMA is a stochastic model, which 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 [16] in the sense that the differencing parameter (d) could have fractional (non-integer) values. The fractional ARIMA (ARFIMA) models are capable of characterizing the short-range and long-range memories in the data by applying the fractional differencing on the time series.

A. Overview of ARFIMA Model

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. (4),

$$\Phi_p(B)\Delta^d X_t = \Theta_q(B)\epsilon_t. \quad (4)$$

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 [11] and Gaussian innovations ($\alpha = 2$) with finite variance [8][17].

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,

$$\Delta^d = (1-B)^d = \sum_{k=0}^{\infty} \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)} (-B)^k, \quad (5)$$

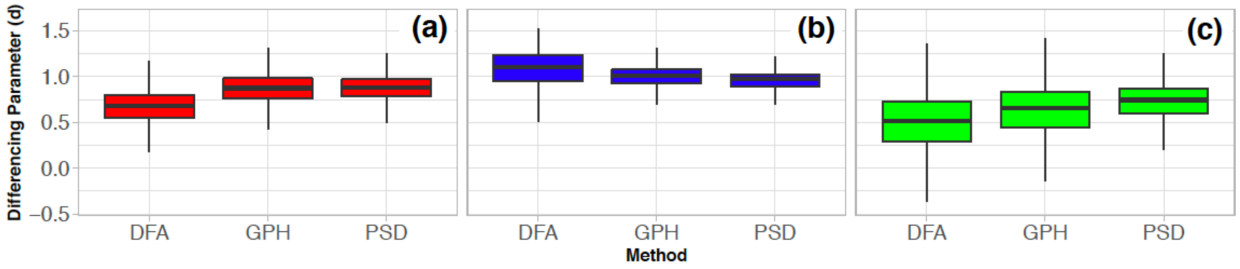

Figure 3. Differencing parameters of PMU data using DFA, GPH, and PSD: (a) Voltage magnitude (b) Frequency (c) Angle

where $\Gamma(\cdot)$ is the Gamma function ($\int_0^\infty s^{z-1}e^{-s}ds$).

Assuming that the polynomials, $\Phi(z)$ and $\Theta(z)$, have no common roots and the 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

$$X_t = \Phi^{-1}(B)\Theta(B)(1-B)^{-d}\epsilon_t = \sum_{k=0}^{\infty} u_k \epsilon_{t-k}, \qquad (6)$$

where the coefficients, u_k's, are resulting from the power series expansion of $\Phi^{-1}(z)\Theta(z)(1-z)^{-d}$. In [11], it has been shown that the series of the ARFIMA model with α-stable innovations, as shown in Eq. (6), converges *almost surely* when

$$-\infty < d < 1 - \frac{1}{\alpha}. \qquad (7)$$

That means the ARFIMA model with α-stable innovations $(0 < \alpha \leq 2)$ 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 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 time series (X_t) to remove any non-stationarity. The first-differenced time series will be converted to a stationary time series,

$$Y_t = X_t - X_{t-1} = (1-B)X_t. \qquad (8)$$

After the first differencing, the three resulting time series of PMU data are shown in Figs. 4 (a)-(c). Now, applying the ADF and KPSS tests on the three first-differenced time series shows 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. 4 (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

$$\Phi(B)(1-B)^{d^*}Y_t = \Theta(B)\epsilon_t, \qquad (9)$$

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 $(1-B)^{d^*}Y_t$. So, the fractional differencing of the time series (Y_t) generate a time series (Z_t) possessing short-range memory.

Using Eq. (5), 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. 5 (a)-(c). The differencing parameters before and after fractional differencing are shown in the legends of Fig. 4 and Fig. 5. It is clear that fractionally differenced time series (Z_t's) have short-range memory only 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,

$$\Phi(B)Z_t = \Theta(B)\epsilon_t. \qquad (10)$$

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 three time series of PMU data are shown in Table II.

Table II. 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 steps and algorithm of 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 $(\Phi(B)(1-B)^{d^*}Y_t = \Theta(B)\epsilon_t)$.

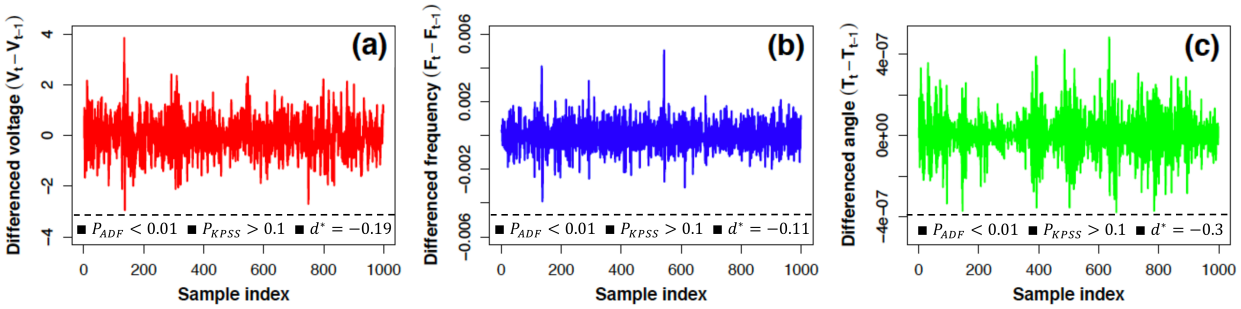

Figure 4. First-differenced time series (Y_t) of PMU data: (a) Voltage magnitude (b) Frequency (c) Angle

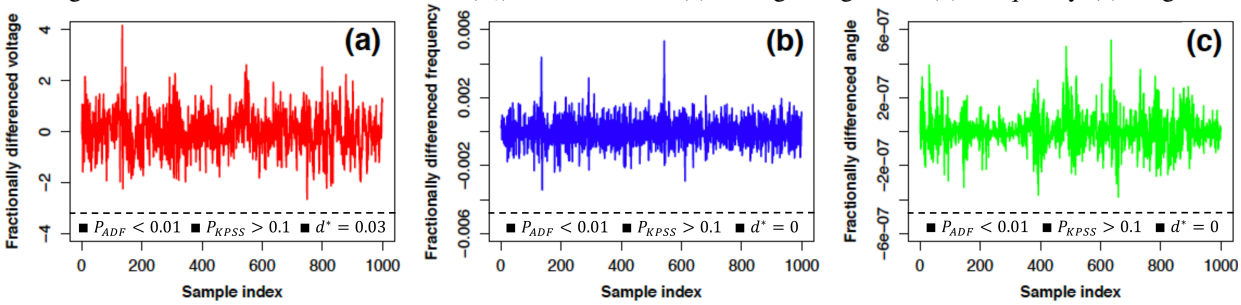

Figure 5. Fractionally-differenced time series (Z_t) of PMU data: (a) Voltage magnitude (b) Frequency (c) Angle

(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.

Algorithm 1: Identification of the ARFIMA models

Input: X_t

Output: $model$, d, aic, bic

set $aic = \infty$, $bic = \infty$;
get P_{ADF} and P_{KPSS} of X_t ; // as shown in Sec. II B
if $P_{ADF} \leq 0.01$ *and* $P_{KPSS} \geq 0.01$ **then**
  $Y_t \leftarrow X_t$;
else
  $Y_t \leftarrow X_t - X_{t-1}$;
get d^* from Y_t ; // as explained in Sec. III
$d \leftarrow d^* + 1$;
$Z_t \leftarrow (1-B)^{d^*} Y_t$;
for $p = 0$ *to* 5 **do**
  for $q = 0$ *to* 5 **do**
    $model^{temp} \leftarrow$ fit (Z_t, ARMA(p, q));
    $aic^{temp} \leftarrow$ AIC (model);
    $bic^{temp} \leftarrow$ BIC (model);
    if $aic^{temp} < aic$ *and* $bic^{temp} < bic$ **then**
      $model \leftarrow model^{temp}$;
      $aic \leftarrow aic^{temp}$;
      $bic \leftarrow bic^{temp}$;

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 [18]. One of the reliable estimators of Gaussian and stable ARFIMA models is Whittle approximate maximum liklihood estimation (MLE) [19][20]. The Whittle estimation of the ARFIMA parameters can be achieved by minimizing

$$Q(\hat{\zeta}) = \sum_{j=1}^{m} \frac{I(\lambda_j)}{f(\lambda_j, \hat{\zeta})}, \tag{11}$$

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 II 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 III.

Table III. 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. Analysis of Residuals of the ARFIMA Model

Evaluating the goodness-of-fit of the PMU ARFIMA models requires analyzing the residuals ($\hat{\epsilon}_t$) of these models. Towards 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 for α-stable innovations (residuals).

Moreover, a good ARFIMA model of the data should result in uncorrelated residuals to ensure all the information in the data is represented in the model. So, we find the autocorrelation functions of the residuals to study the autocorrelation between residual samples at different lags. Furthermore, we use Ljung-Box test to check the independence of the residuals.

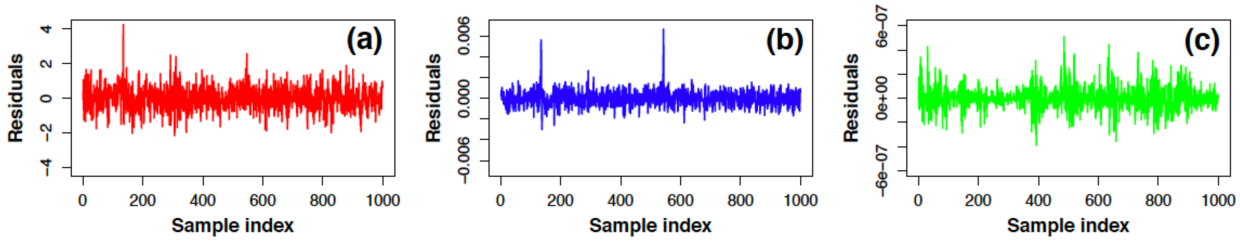

Figure 6. Residuals of ARFIMA models of PMU data: (a) Voltage magnitude (b) Frequency (c) Angle

A. Residuals of the ARFIMA Model

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

$$\hat{\epsilon}_t = \Phi(B)\Delta^d\Theta^{-1}(B)X_t. \tag{12}$$

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. 6 (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 [11, Sec. 7.13]. To go around this difficulty, we utilize Eq. (11) 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 calculate its parameters (α, β, γ, and δ) [11]. The histograms of the ARFIMA model residuals and their best α-stable fit based on Koutrouvelis regression method [21] are shown in Figs. 7 (a)-(c). To test the normality of the model residuals, we generate the Q-Q plots for the residuals of the ARFIMA models, as shown in Figs. 8 (a)-(c).

The residuals of the voltage ARFIMA model follows 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 angle time series is equal to -0.3 and smaller than $1 - (1.48)^{-1}$. Satisfying this condition validates the derived models of Y_t.

Comparing the ARFIMA modeling with the power flow equations, it transpires that the "residuals" are, physically, driving the fluctuating powers injected. The observed near Gauss property of the "residuals," hence the power fluctuations, corroborates the observation made in [22] that such fluctuations at the output of a wind farm are Gaussian.

B. Correlation of the ARFIMA Model Residuals

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

$$\hat{r}_k = \frac{\sum_{t=k+1}^{n} \hat{\epsilon}_t \hat{\epsilon}_{t-k}}{\sum_{t=1}^{n} \hat{\epsilon}_t^2}. \tag{13}$$

The autocorrelation functions of the residuals from the three ARFIMA models of the PMU data are shown in Figs. 9 (a)-(c). Inspecting graphically the first 50 individual autocorrelations 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 of the ARFIMA models.

A more formal way of analyzing the dependence in the residuals is through the Ljung-Box test [23]. The test was introduced by Ljung and Box in 1978 as a modified version of the Box and Pierce test [24]. 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,

$$Q(\hat{r}) = n(n+2)\sum_{k=1}^{m}(n-k)^{-1}\hat{r}_k^2, \tag{14}$$

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 degrees of freedom, h or $m-(p+q)$, 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. 10 (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 can not reject the Null hypothesis (H_0). That means we do not have evidence that the residuals of the selected ARFIMA are autocorrelated.

VI. Anomaly Detection via ARFIMA Models

A promising application of our proposed data models is detecting anomalies and abnormal events in the power grid. In our novel method, we aim to detect hidden changes in the power system signals by tracking the changes in the parameters of the ARFIMA models as the system is approaching a major transition.

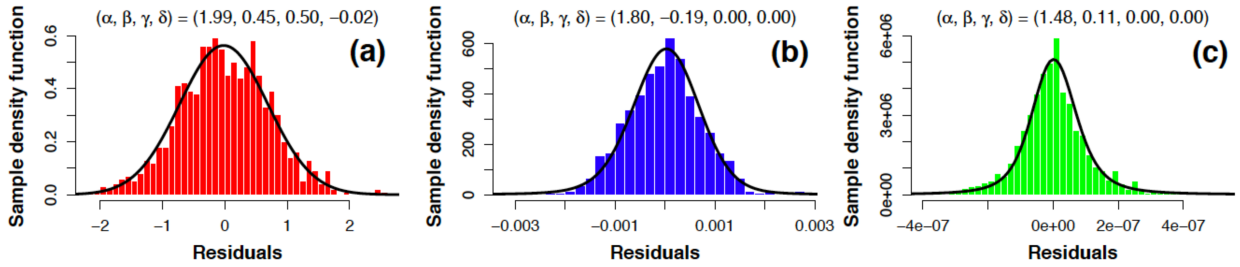

Figure 7. Sample density functions of the residuals of the ARFIMA models: (a) Voltage magnitude (b) Frequency (c) Angle

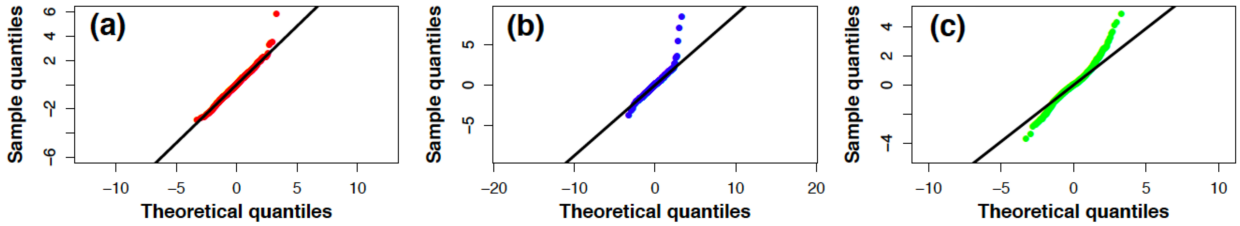

Figure 8. Quantile-quantile plots of the residuals of the ARFIMA models: (a) Voltage magnitude (b) Frequency (c) Angle

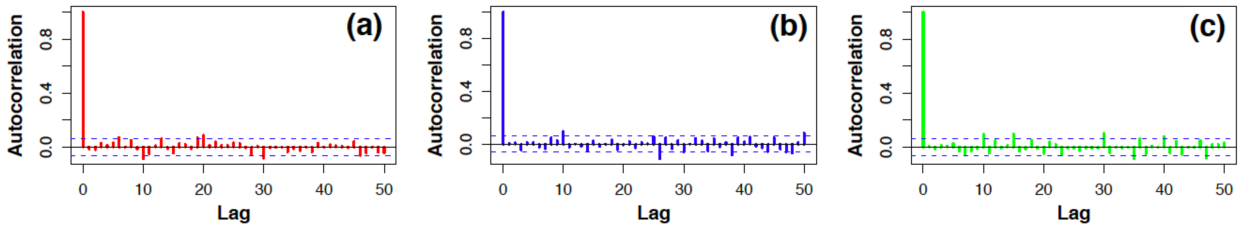

Figure 9. Autocorrelation functions of the residuals of the ARFIMA models: (a) Voltage magnitude (b) Frequency (c) Angle

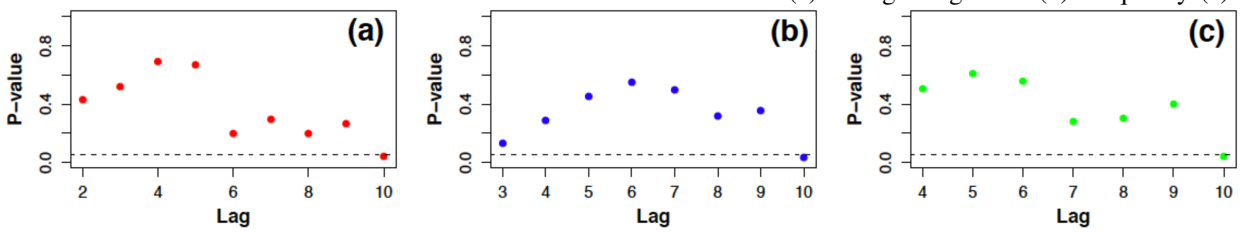

Figure 10. P-values of Ljung-Box statistic of the ARFIMA residuals: (a) Voltage magnitude (b) Frequency (c) Angle

Power grids are often susceptible to unpredictable disturbances that could lead in the worst case scenario to a major power blackout. One of the largest power blackouts in the history is the 2012 Indian blackout [25]. 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).

To track the hidden changes in frequency time series of 2012 Indian blackout, we calculate the ARFIMA models inside a sliding window toward the blackout. The size of the sliding window is 16.67 minutes (1000 samples) with a shift of 1 minute (60 samples). The model ARFIMA (1, d, 1) is used to model the frequency time series inside each window. In Figs. 11(b)-(c), we have the differencing parameter (d) and the lag-1 autoregressive parameter (ϕ_1) of the ARFIMA models inside the windows as the system is approaching the blackout.

By examining the plots of the differencing and lag-1 autoregressive parameters, we can notice a clear increase in their values toward the blackout. The increase is starting around 10 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, while driving a change point detection [5] to control False Alarm Rate, could provide an early warning to the proximity to a major transition or blackout.

We believe that the increase in the differencing and lag-1 autoregressive parameters of the frequency time series could be a sign of a critical slowing down phenomenon [26] [27] in the power grid.

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 fractality parameters, showing existence of long-range memory in the PMU data. The estimated fractality parameters (scaling exponent $\alpha^\star$, differencing parameter d, and power exponent β) are showing consistency among the different methods. Since most of the PMU data have long-range memory, we have used the ARFIMA models to model 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.

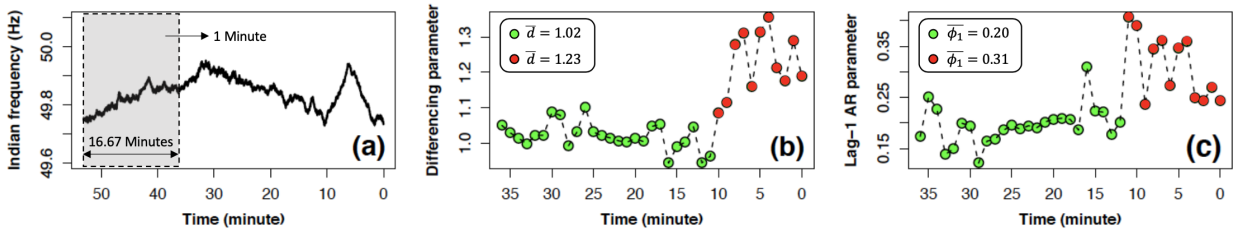

Figure 11. ARFIMA modeling before the Indian blackout: (a) Frequeny (b) Differencing parameter (c) Lag-1 AR parameter

Practical grid applications already emerge, and will be subjects of further invetigations–most importantaly, change point detection of ARFIMA parameters to detect abnormal events. As already said, ARFIMA models of the voltage (V) and the angle (θ) are dynamical 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.

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

[12] 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.

[13] PMU data from EPFL. http://nanotera-stg2.epfl.ch/. [Online; accessed 12-April-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] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

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

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

[19] 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.

[20] 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.

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

[22] 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.

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

[24] 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.

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

[26] C. Wissel. A universal law of the characteristic return time near thresholds. *Oecologia*, 65(1):101–107, 1984.

[27] E. Nes and M. Scheffer. Slow recovery from perturbations as a generic indicator of a nearby catastrophic shift. *The American naturalist*, 169:738–47, 07 2007.