# An Enhancement to Time Series Model for Short-Term Forecasting of Multi-Regional Electricity Load

Dinh Duong Le

***Abstract*—** *Load forecasting plays a key role in managing and operating power systems in an efficient, reliable, and secure way. Load forecasting estimates future consumption based on available data and information. Among several types of load forecasting, the article focuses on the short-term forecast. In addition to the temporal correlation, loads at different areas connected to a power system may be correlated to each other. Both temporal and spatial correlations need to be taken into account in forecasting strategies. This paper presents a methodology for short-term forecasting of multi-regional electricity load based on a combination of univariate time series model, principal component analysis and pre-processing techniques. The obtained results are then compared to the results from multivariate time series models as well as actual measurement data showing good performance of the proposed method. The model is tested on real load data from multiple regions in Da Nang, Vietnam.*

***Keywords*— Short-term forecast, time series model, multi-regional electricity load.**

## 1. INTRODUCTION

Load forecasting is one of the very important areas for managing and operating power systems in an efficient and reliable way. Load forecasting can assist in determining the operating plan, generation plan and investment direction to develop the system in the future. Load forecasting can be divided into different categories depending on various ways of classification. According to future time horizons, load forecasting can be classified into three categories [1]: short-term (from a few minutes to a few hours or a day), medium-term (a few days to a few weeks) and long-term (a few months to a year or more) forecasts. In particular, while long-term load forecasting can support electric utility companies in determination of future needs for system expansion, purchases of devices, staff hiring, etc., the medium-term forecasting is usually used for scheduling fuel supplies, unit maintenance and so on. Compared to the medium and long-term forecasts, short-term load forecast plays an important role in the operation of the power system and this is of interest in this article.

Currently, there are many short-term load forecasting methods. They can be classified into four categories [1]: Statistical techniques, artificial intelligent techniques, knowledge based expert systems, hybrid techniques. Statistical techniques are suitable for short-term forecasting. With the development of artificial intelligent techniques, many methods for forecasting load have been developed. These forecasting techniques usually require several types of data for model training (including load data and data of environmental processes such as solar radiation, temperatures, etc.). Hybrid techniques are the combination of different methods to take advantage of each individual one. However, this method is often more complicated and the combination should be implemented effectively to promote the effectness of each individual method. In general, each method has its own advantages and disadvantages, depending on purpose of using forecasting results, the specific input data collected, the characteristics of the load, etc., that determine the choice of the most appropriate method.

In this paper, the popular short-term load forecasting models (belonging to statistical technique category), i.e., time series models, are considered. Time series models [2] are convenient for modeling and managing to be suitable for the input data collected. Univariate time series models include Auto-Regressive (AR), Moving Average (MA), Auto-Regressive Moving Average (ARMA), Auto-Regressive Integrated Moving Average (ARIMA). In order to analyze and forecast many time series, multivariate time series models are used. One of the most common multivariate time series models used in short-term load forecast is the Vector Auto-Regressive (VAR) model [3]. In addition to the temporal correlation, multivariate time series can take into account cross-correlation. It is worth noting that when analyzing and calculating the power system using data from loads in different areas connected to the system, these loads need to be forecasted using multivariate models. In that case, the obtained results are of practical usefulness. Another point needs to be considered is that with large number of time series, it is very complicated to build multivariate time series models and may give poor results. In addition, actual data collected, such as load measurement data, often have daily and seasonal characteristics, so the data series need to be pre-processed before developing multivariate time series model for the data.

In order to build a time series model that can be used to forecast for multi-regional loads in the power system

Dinh Duong Le is with Faculty of Electrical Engineering, The University of Danang – University of Science and Technology, 54 Nguyen Luong Bang St., Danang city, Vietnam.

*Corresponding author: Dinh Duong Le; Phone: +84-905320755; E-mail: lddduong@dut.udn.vn.

that meets the above-mentioned requirements, in this paper, a short-term load forecasting approach based on a combination of univariate time series model, principal component analysis [4, 5] and pre-processing techniques is developed. The obtained results are then compared to the results from a multivariate time series model (i.e., VAR) as well as actual measurement data showing good performance of the proposed method.

In Section 2, the fundamental background of time series models, principal component analysis and pre-processing techniques are presented. The proposed methodology is described in Section 3; Section 4, we discuss the results obtained on load data from multiple regions in Da Nang, Vietnam; discussions on assessing the resulting scenarios are also given. Section 5 concludes the paper.

## 2. FUNDAMENTAL BACKGROUND

### *Univariate Time Series Model*

The typical linear model for a stationary time series is ARMA [2] that includes two parts, i.e., AR and MA. It is usually referred to as the ARMA($p$, $q$) model, where $p$ is the order of the AR part and $q$ is the order of the MA part. An ARMA($p$, $q$) model of a stochastic process {X(t)} can be mathematically represented as

$$X(t) = a_0 + a_1 X(t-1) + ... + a_p X(t-p) + \varepsilon(t) + b_1 \varepsilon(t-1) + ... + b_q \varepsilon(t-q) \quad (1)$$

where, $a_0$ is a constant; $a_1$, $a_2$, . . ., $a_p$ and $b_1$, $b_2$, . . ., $b_q$ are the parameters of AR and MA, respectively. The stochastic process {ε(t)} is called a white noise process [2].

When $q$ = 0, the ARMA($p$, $q$) model becomes an AR($p$) model. Similarly, when $p$ = 0, it becomes an MA($q$) model. An AR model expresses a time series as a linear combination of its past values. In AR($p$) model, the order of $p$ tells how many lagged past values are included in the model. MA model is a series of linear combination of noise process with different lagged terms.

Box–Jenkins [2] proposed a procedure for building a univariate time series model and this procedure is used in this paper. It is worth noting that stationarity is a required condition in building an ARMA model. However, this condition may not always hold with real time series data. In such a case, ARIMA($p$, $d$, $q$) model can be used. It is a generalization of an ARMA model. In the model, $d$ is the degree of differencing that is the number of times the data have had past values subtracted. For a non-stationary time series, it may be necessary to difference the data only one time or sometimes two times to obtain a stationary time series. However, there is another option that is applying pre-processing techniques to transform a non-stationary time series into a stationary one, allowing the use of an ARMA model. The second way is realized in this article. The stationarity of a time series data can be assessed by a statistical test such as Augmented Dickey–Fuller (ADF) test and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test [6, 7].

The order $p$ and $q$ of an ARMA model can be chosen on the basis of Auto-Correlation Function (ACF) for MA component and Partial Auto-Correlation Function (PACF) for AR components [2], as well as several popular information criteria like BIC (Bayesian Information Criterion), AIC (Akaike Information Criterion), etc.

Parameters of an ARMA model can be easily estimated by using Ordinary Least Square (OLS) or Maximum-Likelihood Estimation (MLE).

### *Multivariate Time Series Model*

Among various multivariate time series models, Vector Auto-Regressive (VAR) is one of the most common ones and it is used in this paper.

For two time series $\{X_1(t)\}$ and $\{X_2(t)\}$, VAR model can be expressed as [3]:

$$\begin{aligned} X_1(t) &= c_1 + \sum_{i=1}^{p} \beta_i X_1(t-i) + \sum_{i=1}^{p} \gamma_i X_2(t-i) + \varepsilon_1(t) \\ X_2(t) &= c_2 + \sum_{i=1}^{p} \varphi_i X_1(t-i) + \sum_{i=1}^{p} \theta_i X_2(t-i) + \varepsilon_2(t) \end{aligned} \quad (2)$$

where, $a_0$ is a constant; $a_1$, $a_2$, . . ., $a_p$ and $b_1$, $b_2$, . . ., $b_q$ are the parameters of AR and MA, respectively. The stochastic process $\{\varepsilon(t)\}$ is called a white noise process [2].

where, $c_1$ and $c_2$ are constants; $\beta$, $\gamma$, $\varphi$, $\theta$ are parameters of the model; $p$ is the lag; $\{\varepsilon_1(t)\}$ and $\{\varepsilon_2(t)\}$ are white noise processes. For $n$ time series, the model is similarly written.

Similarly to ARMA model, after obtaining multivariate stationary time series data, VAR model can be built using techniques and criteria like those for ARMA.

### *Principal Component Analysis*

Principal component analysis (PCA) performs an orthogonal transformation on data of correlated variables to obtain data of uncorrelated variables called principal components (PCs). It can be used to reduce a large set of variables to a small set without losing significant information in the original set.

Suppose matrix $\mathbf{X}$ (size n×m) contains the original data (rows of $\mathbf{X}$ correspond to observations and columns correspond to variables, e.g., load at each bus in the power system), PCA is performed on the dataset step by step as follows:

(1) Center the data (by subtracting the mean of each variable) to obtain centered matrix $\mathbf{X}_c$;

(2) Form covariance matrix and compute its eigenvalues ($\lambda_i$, i = 1, 2, ..., m) and corresponding eigenvectors ($\mathbf{e}_i$, i = 1, 2, ..., m);

(3) Sort eigenvalues in descending order: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$;

(4) Construct the projection matrix: $\mathbf{E} = [\mathbf{e}_1\ \mathbf{e}_2\ ...\ \mathbf{e}_m]$;

(5) Transform the dataset $\mathbf{X}_c$ via $\mathbf{E}$ to obtain a n×m matrix $\mathbf{Y} = (\mathbf{E}^T \mathbf{X}_c^T)^T$, where each column of $\mathbf{Y}$ is called a PC.

The variance of the i$^{th}$ PC (i = 1, 2, ..., $m$) is equal to

the eigenvalue $\lambda_i$ associated with that PC. The first column of **Y** (the first PC) corresponding to the largest eigenvalue $\lambda_1$ is the most important component, which contains most of information in the original dataset **X**, followed by the second component, and so on.

The contribution of the $i^{th}$ PC to total variance of the data can be calculated as:

$$\alpha_i = \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j} \quad (3)$$

If the first k eigenvectors that correspond to the k (k << m) largest eigenvalues are selected, we could obtain a reduced matrix $\mathbf{Y}_k = (\mathbf{E}_k^T \mathbf{X}_c^T)^T$, where $\mathbf{E_k} = [\mathbf{e}_1\ \mathbf{e}_2\ ...\ \mathbf{e}_k]$;

The cumulative contribution of the first k PCs is:

$$\Sigma_k = \sum_{i=1}^{k} \alpha_i \quad (4)$$

Consequently, PCA projects a data in matrix **X** (size n×m) into lower dimension subspace (size n×k) by picking up a few numbers of components (i.e., k) with the largest variances.

***Data Pre-processing Techniques***

As discussed, to transform data from a non-stationary time series into a stationary one, we could detect and remove possible trend, daily and seasonal patterns in the data. Then, they are normalized by useful tools in Matlab.

The stationarity of a time series data can be assessed by a statistical test such as Augmented Dickey–Fuller (ADF) test and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test.

## 3. METHODOLOGY

The proposed method is a combination of univariate time series model, principal component analysis and pre-processing techniques. It is implemented as follows:

- *Step 1:* Perform pre-processing techniques to obtain stationary time series data. Stationarity is assessed by either ADF test or KPSS test;
- *Step 2:* Use PCA to transform multivariate time series data with cross-correlation (spatial correlation between loads) into PCs (each PC is uncorrelated with other PCs);
- *Step 3:* Choose number of the first PCs (corresponding to the largest eigenvalues) which contain most of information (usually around 80% at least in practice) in the original dataset.
- *Step 4:* Build univariate stationary time series model for each selected PC time series;
- *Step 5:* Use the model obtained for each PC to forecast for a pre-defined future time horizon;
- *Step 6:* Back-transform data forecasted for PCs and add back the items removed in the pre-processing step to remain characteristics of the observed data.

## 4. CASE STUDY

The proposed method is tested on hourly load data collected from 16 substations in Da Nang city (Vietnam), provided by Da Nang Power Company. Data from the whole month of January, 2018 are used to build the forecasting models, while data for the first week of February, 2018 are used to test the obtained results. VAR model is also applied to forecast and its results are compared with the results obtained by the proposed model.

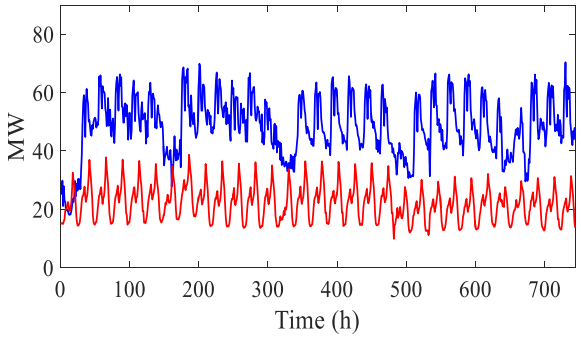


**Fig. 1. Measured Data of Load 1 and Load 3.**

For illustration, measured data of Load 1 and Load 3 are plotted in Fig.1. As can be seen from the figure, data are clearly non-stationary with daily and seasonal (weekly) patterns that should be removed before building stationary time series models.

Data from 16 substations are strongly correlated with correlation coefficients range from - 0.43 to + 0.91. Fig.2 and Fig.3 show scatter plot for Load 2 and Load 4 (correlation coefficient: 0.53) and Load 6 and Load 7 (correlation coefficient: 0.91), respectively.

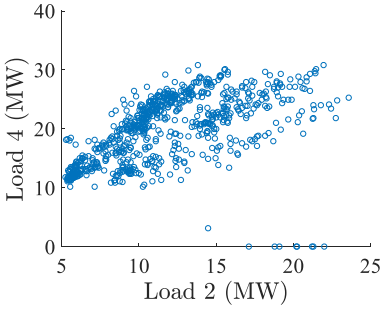


**Fig. 2. Scatter Plot for Load 2 and Load 4.**

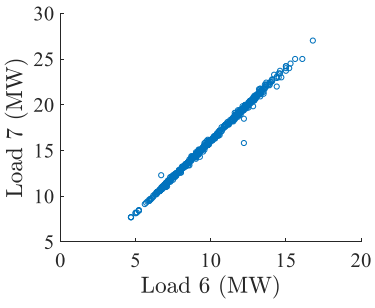


**Fig.3. Scatter Plot for Load 6 and Load 7.**

In order to choose the number of PCs, Scree plot can be used. Fig.4 shows Scree plot in which we can choose the order of eigenvalue (corresponding to number of PCs) at the break point in the graph. In this case, we choose five first PCs that explain 83% of total information in the original data. Therefore, instead of building a multivariate model for 16 correlated time series, we build five univariate models for five PC time series. The latter is usually easier to perform than the former. This is one of important advantages of the proposed model.

Fig.5 describes the first seven resulting PC time series. It is clearly that PC time series are quite different in terms of magnitudes. The first PC contains the largest percentage of variance in the data set with 38%; the second PC the second largest percentage with 17%, and so on.

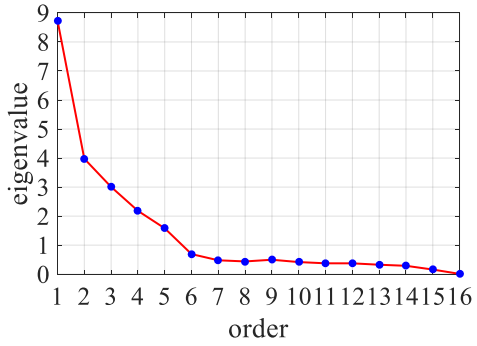


**Fig.4. Scree Plot.**

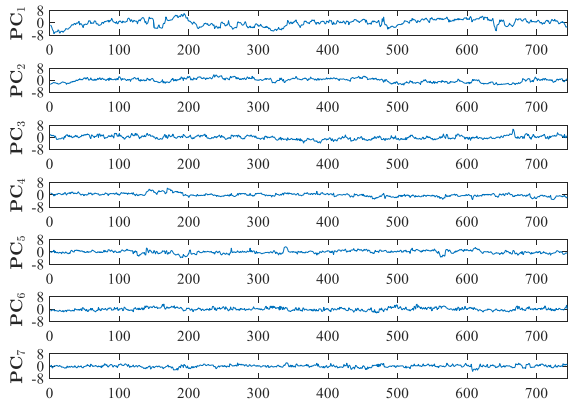


**Fig. 5. Plotting for the First Seven Resulting PC Time Series.**

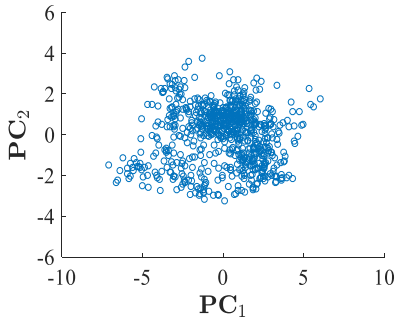


**Fig.6. Scatter Plot for $PC_1$ and $PC_2$.**

PC time series are uncorrelated as shown, for example, in Fig.6 for Scatter Plot for $PC_1$ and $PC_2$. We build five univariate time series models for the five selected PCs. We follow the Box–Jenkins procedure. In order to determine the order $p$ and $q$ of an ARMA model we use ACF and PACF plots. Looking at ACF and PACF plots for $PC_1$ time series as in Fig.7 and Fig.8, the data clearly follow an AR process: the PACF displays a sharp cutoff while the ACF decays gradually. The lag at which the PACF cuts off is the indicated order of AR. In this case, $p$ = 2. Parameters of AR model are: $a_1$ = -0.9296; $a_2$ = 0.0044. Residual test for time series model built for $PC_1$ is performed by plotting ACF of the residual. Looking at Fig.9, it is clear that the process has an ACF of zero at all lags except a value of unity at lag zero, so the process is completely uncorrelated. It is a white noise process indicating that the model built for $PC_1$ is true [2].

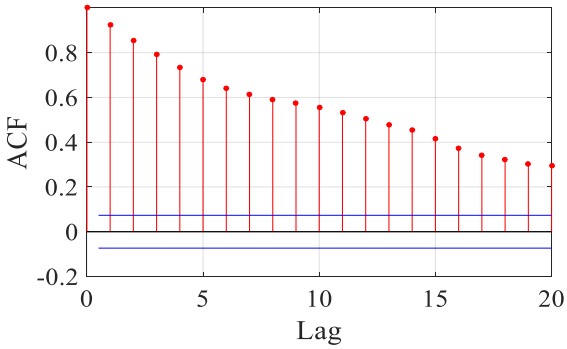


**Fig. 7. ACF Plot for $PC_1$ Time Series.**

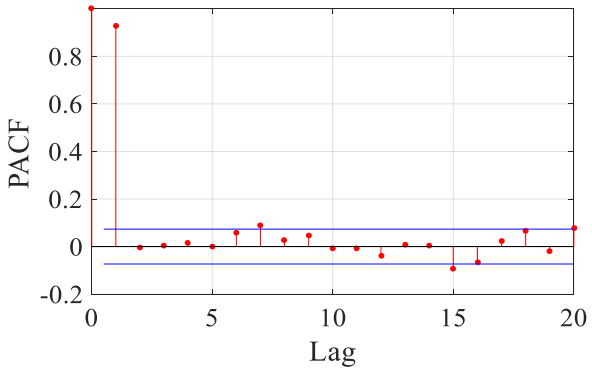


**Fig. 8. PACF Plot for $PC_1$ Time Series.**

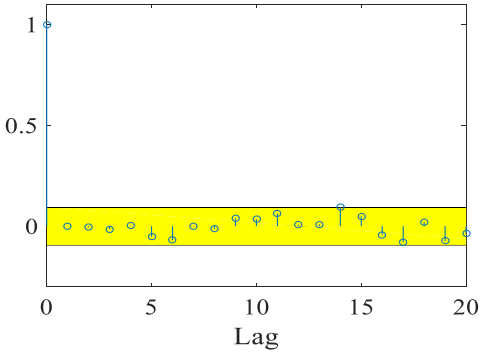


**Fig.9. Residual Test for Time Series Model Built for $PC_1$.**

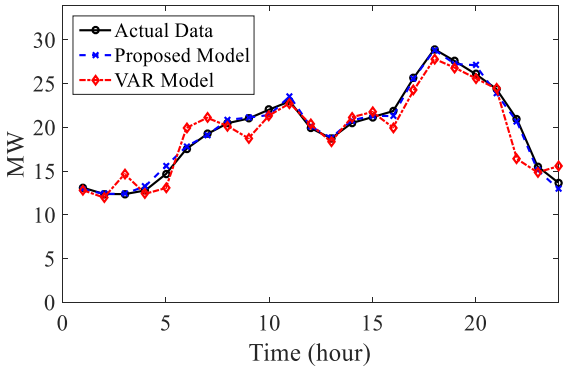


**Fig. 10. Forecast Results for Load 3.**

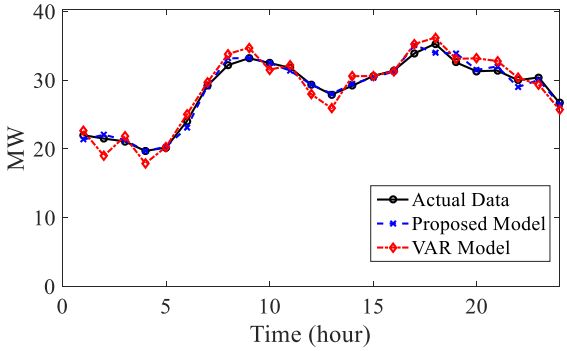


**Fig. 11. Forecast Results for Load 10.**

**Table 1. Comparison of MAE**

| Load | MAE (MW) Proposed Model | MAE (MW) VAR Model |
|---|---|---|
| Load 1 | 1.11 | 1.50 |
| Load 2 | 0.34 | 0.52 |
| Load 3 | 0.52 | 1.22 |
| Load 4 | 0.41 | 0.64 |
| Load 5 | 0.61 | 0.92 |
| Load 6 | 0.24 | 0.39 |
| Load 7 | 0.30 | 0.51 |
| Load 8 | 0.45 | 0.72 |
| Load 9 | 0.31 | 0.63 |
| Load 10 | 0.50 | 1.12 |
| Load 11 | 0.72 | 0.98 |
| Load 12 | 0.45 | 0.67 |
| Load 13 | 0.73 | 1.00 |
| Load 14 | 0.21 | 0.43 |
| Load 15 | 0.30 | 0.48 |
| Load 16 | 0.31 | 0.73 |

We use the model obtained for five PCs to forecast for a pre-defined future time horizon (24 hours in this paper). Then, we back-transform data forecasted for PCs and add back the items removed in the pre-processing step. We also build and run a VAR model for the data obtaining after using the pre-processing techniques for the original data. All the obtained results are compared as shown in Fig.10, Fig.11, Fig.12, for example, for Load 3, Load 10, Load 16, respectively. MAEs (Mean Absolute Errors) [8, 9] are also calculated for all loads: They range from 0.21 to 1.11 MW. They clearly show that the proposed method can give very accurate results and better than VAR model.

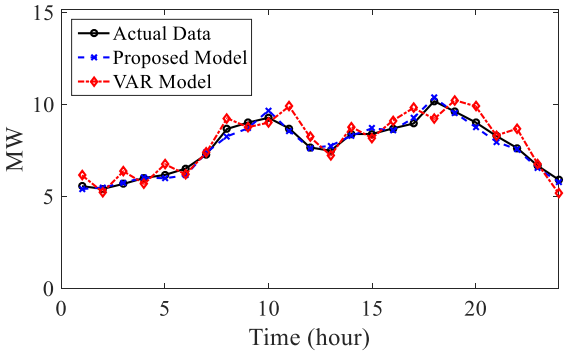


**Fig. 12. Forecast Results for Load 16.**

## 5. CONCLUSIONS

Load forecasting plays an important role in managing and operating power systems in an efficient and reliable way. In power systems, when analyzing and calculating the system in the future time horizon, forecasting data from loads in different areas connected to the system are required. These loads need to be forecasted usually using multivariate models. In that case, the obtained results are of practical usefulness. However, building a multivariate model for a large number of time series are very complicated.

This paper develops a methodology for short-term forecasting of multi-regional electricity load based on a combination of univariate time series model, principal component analysis and pre-processing techniques. The proposed model is tested on real load data from multiple regions in Da Nang, Vietnam, showing good performance.

## REFERENCES

[1] Srivastava, A.K., Pandey, A.S., and Singh, D. 2016. Short-Term Load Forecasting Methods: A Review. In *Proceedings of International Conference on Emerging Trends in Electrical, Electronics and Sustainable Energy Systems (ICETEESES–16)*. Uattar Pradesh, India, 11-12 March,130–138.

[2] Box, G. E. P. and Jenkins, G.M. 1976. Time Series Analysis: Forecasting and Control, San Francisco, CA: Holden Day.

[3] Stock, J.H., and Watson, M.W. 2001. Vector Autoregressions. *Journal of Economic Perspectives*, 15, 101–115.

[4] Jolliffe, I.T. 2002. Principal Component Analysis. New York, NY, USA: Springer.

[5] Jackson, J.E. 1991. A Users Guide to Principal Component Analysis. Hoboken, NJ, USA: Wiley.

[6] Cromwell, J.B., Labys, W.C., and Terraza, M. 1994. Univariate Tests for Time Series Models. Newbury Park, CA, USA: Sage.

[7] Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., and Shin. Y. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. J.Econometrics, vol. 54, 159–178.

[8] Shaker, H., Zareipour, H., and Wood, D. 2013. On error measures in wind forecasting evaluations. In *Proceedings of 26th IEEE Canadian Conference of Electrical and Computer Engineering (CCECE).* Regina, Canada, 5-8 May.

[9] Willmott, C.J. and Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*. 30: 79–82.