Comparison of Different Wind Time Series Simulation Methods

By

Shiyao Wu

Dr. Dalia Patiño-Echeverri, Advisor

Acknowledgments

The author thanks her advisor, Dalia Patiño-Echeverri, for her time advising on the project, for giving feedbacks, and for providing new perspectives. She also thanks PhD students, Rubenka and Ali for their generous help.

Contents

Executive Summary

The assessment of power system reliability under increasing penetration of wind power requires long-term wind data that is not available or does not exist and hence must be simulated. In this research, autoregressive models (AR) ranging from 1st order to 12th order and Markov-switching autoregressive models (MS-AR) ranging from MS(2)-AR(2) to MS(5)-AR(5) are used for simulation of wind speed data that has the same stochastic characteristics as a 10-minutes simulated wind speed time-series provided by NREL for years 2004 and 2005. Simulation results are compared between models, across different seasons, and different data lengths.

The report has six sections: introduction, research objective, theoretical framework, methodologies, results, and conclusion. The first section provides background information. It starts with current status of wind energy in the US and around the world, and then provides an overview of previous research on wind simulation models studied in this project. The second section covers the research objective of this project and general introduction of the structure and content of this report.

The third section covers the theoretical framework. AR model, MS-AR model and relevant concepts are introduced in this section. To explain how the models work and clearly describe the characteristics of models, examples are given to help clarify the simulation mechanism of each model.

The fourth section provides information on methodologies. Firstly, data source, characteristics and processing methods are covered. Then methodologies on how simulations are generated are discussed.

The fifth and sixth sections focus on the results and conclusions of this project. In the fifth section, we start by introducing the metrics of model performance, then results and observations are discussed in the rest part of the section. In the end, conclusions are drawn in the sixth section.

The comparison results can be summarized as follows:

- The Markov Chain component in MS-AR models further improves the ACF performance as well as the performance in terms of PDF.
- Increasing number of state in the Markov Chain can significantly improve the performance of MS-AR models.
- MS-AR models are more tolerant to input data, which is a result of their distributional versatility.
- Although MS-AR models are better than AR models in many ways, we cannot deny that AR models are more efficient as their simplicity and time saving characteristic can to some extent offset less perfect performance.

1. Introduction

Wind energy is considered one of the most promising and fastest growing alternative energy resources in the electric power system due to its competitive cost, clean generating process and non-exhaustible nature. The growing concern on the climate and environmental problems leads to rapid growth in wind in recent years in order to reduce GHG emissions all around the world. “The Chinese market posted a 25% growth in 2013; the Brazilian industry is set to install nearly 4 GW in 2014; Mexico has set that country on course for a ~2GW/year market for the next 10 years” [1]. Within the United States, according to the American Wind Energy Association

(AWEA), “U.S. wind energy provides enough electricity to power the equivalent of over 18 million homes. Iowa and South Dakota produced more than 25% of their electricity from wind in 2013, with a total of nine states above 12% and 17 states at more than 5%. Wind energy provided 10.6% of the electricity in 2014 on the main power system in Texas, ERCOT, and that figure is expected to reach 15-20% by 2017” [2]. The Clean Power Plan proposed by EPA, which aims at reducing carbon emission from the power sector by 30% from 2005 levels by 2030, together with the Renewable Portfolio Standard (RPS), which places an obligation on electric suppliers to increase production of energy from renewable resources, jointly motivate higher demand for renewable energy. “Most of this renewable energy will come from wind as other renewable resources are not suitable for bulk power generation” [3].

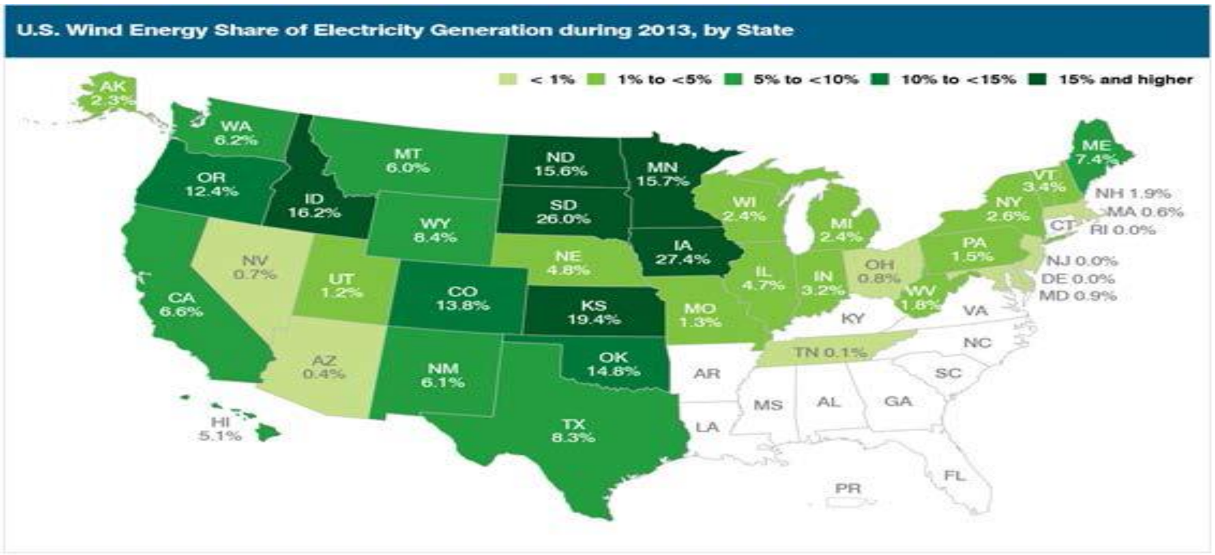

Fig. 1. U.S. wind energy share of electricity generation during 2013, by state.

However, the fact that wind power penetration will continue to increase triggers rising concerns over the reliability of the electric power system. Wind power is not dispatch-able and exhibits both time-variability and uncertainty. Being non dispatch-able, wind power can only contribute to meet electrical demand when the wind is blowing. Time-variability makes wind power incapable of generating stable electricity to the power grid to meet base electrical load. Finally, uncertainty in wind power production means that it is necessary to schedule conventional resources as power generation reserves to offset any deviations in actual wind production from its forecast.

The assessment of the impacts that increased generation capacity of wind power will have on the reliability of power systems requires long term simulations of system operations. Because long-term, high-resolution time series of wind power are not available, researchers often use synthetic time series.

There is a bulk of research on time series wind speed simulation models, owing to their ability to preserve the chronological variability and stochastic nature of the wind [4]. Two of the most commonly used wind speed simulation models are Monte Carlo Markov Chain (MCMC) models and linear Auto-Regressive (AR) models.

The MCMC method is dependent on the past state of the observation and a transition probability matrix between the states. This method is widely applied to the generation of synthetic wind speed time series as it can capture the stochastic nature of wind. It is concluded in [5] that a second-order Markov chain model can slightly improve the wind speed behavior relative to a first-order Markov model. However, in application for time resolutions of less than an hour MCMC often fails to replicate the autocorrelation function (ACF) and probability density function (PDF) of the original time series [6, 7]. Limitations regarding the accuracy of this method, such as the imperfect replication of ACF and PDF, are caused by the intrinsic nature of the Markov process [5]. In order to improve the performance of the model in terms of ACF and PDF, it is suggested to separately apply a MCMC model to monthly data to include the seasonal and monthly variation of wind [6, 8]. The multi-regime MCMC models, which divide the dataset into different sub-regime data according to their diurnal and seasonal patterns and fits MCMC models to each sub-regime, generate better results than general MCMC models. For example, in [6], the 2nd order or higher multi-regime models with a percentile-based discretization of the state-space improved ACF replication and the representation of diurnal and seasonal patterns. However, the performance of PDF and ramp distribution does not improve significantly.

The AR model is dependent on past observations and a random term. The use of linear AR models to analyze empirical time series has a long history. These models are widely used for the prediction of economic trends, signal processing, and natural phenomenon. In [9], different AR models are applied to subsets of data with different day types to include hourly, daily, seasonal and diurnal patterns in the wind time series. This improves the performance of ACF replication. When compared to MCMC models, AR models perfectly replicate the ACF; however, they do not produce satisfactory results in terms of PDF replication [10]. Moreover, AR models should be applied to stationary series. Since the nature of wind time series is a non-stationary, non-Gaussian, random process [4], it is necessary to transform the data before applying AR models to wind time series.

This Master's Project seeks to explore the potential benefits of using Markov-Switching Auto-Regressive (MS-AR) methods for the generation of synthetic time series of wind power accounting for both the stochastic nature and time dependency of wind power. To this effect an MS-AR model is developed. MS-AR models integrate both MCMC and AR methods into one. This is a generalization of Hidden Markov Chain (HMM) and AR models that includes different AR models to represents the evolution of the process at different periods of time and switch the AR models following a transition probability controlled by an HMM [11].

In [11], the advantages and limitations of MS-AR models for wind power are summarized. The MS-AR models have the ability to include statistical properties of data with diverse time scales. Moreover, due to their distributional stochasticity, data pre-processing is no longer necessary. However, the models fail to simulate the lowest part of the distribution and sometimes generate negative wind speed values.

2. Research Objective and Structure

As each of the methods introduced above has its advantages and limitations, we should be careful with choosing the appropriate method for wind time series simulation. For the reason that wind data varies significantly at different locations, the best simulation method and the use of parameters for different wind sites will be different. As a result, it is important to compare the performance of each method using the same wind time series (i.e. corresponding to the same time period and same location). The objective of this master's project is to provide a comprehensive comparison of performance of the three methods introduced above using wind time series data from NREL to see if the comparison results correspond with those of previous work that used wind data from other parts of the world.

Models are introduced in details in Section 3. Results and observations will be discussed in Section 4. Then in Section 5, conclusion will be drawn and further concerns will be discussed.

3. Theoretical Framework

3.1. Autoregressive Model

An AR process is a time series process where the value of a series at a time period is a function of its values at previous time periods plus an error term. This process is characterized by a parameter, p, which is the order of the function. An AR (p) model can be defined as:

$$y_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + \varepsilon_t \qquad (1)$$

where $y_t$represents the output value at time t, $a_1, a_2, a_3 \dots, a_p$ are the parameters of the model, a_0 is the constant, and ε_t is an error term (independently and identically distributed as random draw from a normal distribution).

For example, assume the relationship between wind speed and its previous value follows a first-order autoregressive process, AR(1), which means that the current wind speed is a function of its value at a lag of one time period. In this case, y_t represents the wind speed at current time period, t, then y_t can be calculated using the following equation:

$$y_t = a_0 + a_1 y_{t-1} + \varepsilon_t \tag{2}$$

That is, y_t is a function of some portion of y_{t-1} plus an error term. This nature of the relationship can also be expressed as follows：

$$(1 - \varphi_1 L) y_t = c + \varepsilon_t \tag{3}$$

where φ_1 is the portion of the wind speed at time t-1 carried over to the wind speed at time t. L is the lag operator: $\mathrm{L}y_t = y_{t-1}$. More than one period of time lag can be expressed using the powers of the lag operator: $L^2 y_t = y_{t-2}$

If, however, the current wind speed is jointly determined by wind speed at previous time intervals, y_t would be represented by:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t \tag{4}$$

Or

$$(1 - \varphi_1 L + \varphi_2 L^2) y_t = c + \varepsilon_t \tag{5}$$

This relationship is a second-order autoregressive relationship, designated as AR (2).

AR models require that input data must be stationary. A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time[12], which provide reliable predictor when we try to estimate future behavior.

Different tests can be used to assess the stationarity of a time series. The Augmented Dickey-Fuller (ADF) test is a test for unit root in a time series. It consists of model tests for trend stationary, drift and autoregressive. In an ADF test, the null hypothesis of a unit root is assessed using the following model:

$$y_t = \mathrm{c} + \delta t + \phi y_{t-1} + \beta_1 \Delta y_{t-1} + \cdots + \beta_p \Delta y_{t-p} + \varepsilon_t$$

where Δ is the differencing operator, $\Delta y_t = y_t - y_{t-1}$, p is the number of lagged difference terms.

The unit root null hypothesis is: $H_0\colon \phi = 1$. The alternative hypothesis is: $H_a\colon \phi < 1$. The model with $\delta = 0$ has no trend component, and the model with $\mathrm{c} = 0$ and $\delta = 0$ has no drift or trend. Result of h=1 indicates rejection of null hypothesis of a unit root in favor of alternative hypothesis. However, if the test fails to reject the null hypothesis (h=0), it fails to reject the possibility of a unit root[13].

The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test is a test for stationarity. It provides a straightforward test for null hypothesis of trend stationary against the alternative hypothesis of a unit root. It assumes the following model:

$$y_t = c_t + \delta t + u_t$$

$$c_t = c_{t-1} + \varepsilon_t$$

where u_t is a stationary process and ε_t is an error term that is independently and identically distributed as random draw from a normal distribution with mean of 0 (m=0) and variance (σ^2). The null hypothesis of the KPSS test is: $H_0\colon \sigma^2 = 0$, indicating stationarity. The alternative hypothesis is: $H_a\colon \sigma^2 > 0$. If the test fails to reject the null hypothesis of trend stationarity, it fails to reject the trend stationarity[14]. KPSS test is a good complement test for ADF test.

3.2. Markov Switching Autoregressive Model

3.2.1. Markov Chain

The Markov Chain referred to in this document is a discrete-time, discrete state Markov process. It is a system with a series of random variables (states) that transient through one another in a stochastic manner. This process is characterized by a transition probability matrix, which represents the probability of transitioning from one state to another. For example, assume different ranges of wind speed were assigned to 3 groups: Low, Medium and High. Low state represents wind speed between 0-10 m/s, Medium state represents wind speed between 11-20 m/s, and High represents wind speed between 21-30 m/s. Further assume that at time t, wind speed falls in the range of Low wind state. However, as wind is stochastic, wind speed may switch to Medium or High at the next time interval, t+1. It is also possible that wind speed remains in Low. As a result, a transition probability matrix is needed to express the statistical probability of transition from one state to another at time lag. Fig. 2 can help illustrate this process:

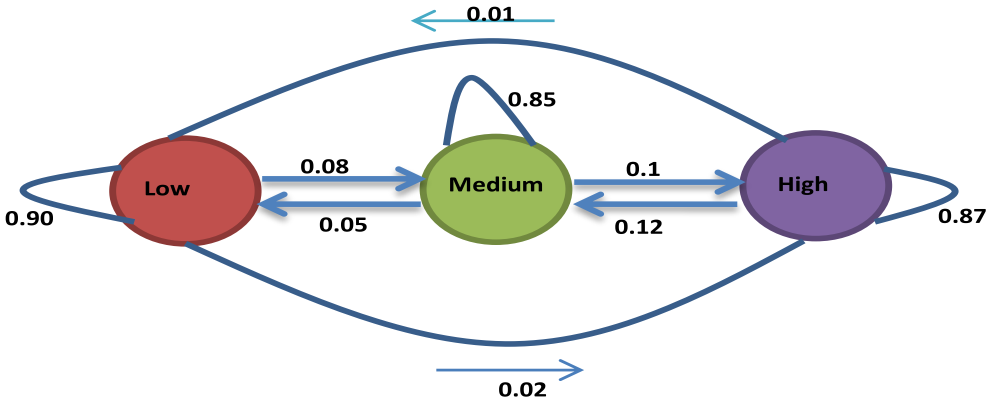

Fig. 2. Three-state Markov chain process

The following table summarizes all the probabilities in the graph above.

Table 1 Transition Probability Table

	Low	Medium	High
Low	0.90	0.08	0.02
Medium	0.05	0.85	0.1
High	0.01	0.12	0.87

The process illustrated above is a first-order Markov chain, i.e. the transition from one state to the next state is independent of previous states, and the next state depends only on the current state regardless of how the system proceeded to current state.

To express this process in a statistical way, we firstly define the series of states in a Markov chain as S_1, S_2, S_3 .... Assuming the total number of state is N, then the conditional probability of moving to s_j at time t+1, given current state of s_i at time t could be denoted as:

$$\mathrm{P}\left(S_{t+1} = s_j \middle| S_t = s_i\right), 1 \leq \mathrm{i}, \mathrm{j} \leq \mathrm{N} \qquad (6)$$

For a first-order Markov chain, the transition process is independent of previous states, so we also have:

$$\mathrm{P}\big(S_{t+1} = s_j \big| S_t = s_i\big) = \mathrm{P}\big(S_{t+1} = s_j \big| S_1 = s_1, S_2 = s_2, S_3 = s_3 \dots S_t = s_i\big) \tag{7}$$

Generally, with an nth order Markov chain, we have:

$$\mathrm{P}\big(S_{t+1} = s_j \big| S_t \dots S_{t-n}\big) = \mathrm{P}\big(S_{t+1} = s_j \big| S_1, S_2, S_3 \dots S_t\big) \tag{8}$$

Let's continue with the previous example. As the current wind speed is within the range of group 1, we have $S_t = s_1$. According to Table 1, the conditional probability of moving from $S_t = s_1$ to $S_{t+1} = s_2$ is:

$$P_{12}(S_{t+1} = s_2 | S_t = s_1) = 0.08$$

All the probabilities in the transition matrix can be denoted in this way.

For general form, the transition probability matrix can be presented as follows:

$$\mathrm{A} = \begin{bmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ij} \end{bmatrix}, a_{ij} \geq 0, 1 \leq \mathrm{i}, \mathrm{j} \leq \mathrm{N} \tag{9}$$

where a_{ij} is the conditional probability of moving from state i, or s_i, to state j, or s_j, and N is the total number of states. Note that the sum of probability in each row must be equal to 1.

Higher order Markov chains remember more previous states when transitioning form a current state to then next and can generally lead to more accurate models and simulation results.

3.2.2. Hidden Markov Model

In a Hidden Markov Model (HMM), the system also follows the Markov process described above. The characteristic that distinguishes an HMM from a general Markov chain is that the states in HMM are unobservable.

In other words, in the previous example, the different ranges of wind speed observed are not the real sequence of states that build up the Markov chain. Instead, the range of wind speed is determined by the weather. In this case, different weather types are the hidden states in a Markov chain that finally determine the observable states (ranges of wind speed observed). For further illustration, two more components are added: the observable state at time t (i.e. the wind speed), denoted as O_t, and another matrix that represents the transition probability from hidden states to observable states (i.e. transition probability from weather type to wind speed).

Assume there are a total number of 3 types of weather: sunny (1), cloudy (2), and rainy (3). Then the conditional probability of moving from sunny weather (type 1) to cloudy weather (type 2) should be:

$$\mathrm{P}(S_{t+1} = s_2 | S_t = s_1) = a_{12}$$

The conditional probability of observing wind speed (O_t) within the range of group 1 at time t given the weather is sunny can be expressed as:

$$P(O_t = o_1 | S_t = s_1) = b_{11}$$

The conditional probability of observing wind speed within the range of group 2 at time t+1given the weather is cloudy can be expressed as:

$$\mathrm{P}(O_{t+1} = o_1 | S_{t+1} = s_2) = b_{21}$$

Then, the probability transition matrix between hidden and observable states can be presented as:

$$B = \begin{bmatrix} b_{11} & \cdots & b_{1k} \\ \vdots & \ddots & \vdots \\ b_{j1} & \cdots & b_{jk} \end{bmatrix}, b_{ij} \geq 0, 1 \leq \mathrm{j} \leq \mathrm{N}, 1 \leq \mathrm{k} \leq \mathrm{M} \qquad (10)$$

where b_{jk} is the conditional probability of getting an observable output (wind speed) of o_k given the state (weather type) at the same time interval is s_j. N is the total number of hidden states (i.e. weather types) and M is the total number of observable states (wind speed ranges).

With the two transition probability matrixes, we can calculate the probability of certain event.

Assume the transition probability matrix between different weather types is:

$$\mathrm{A} = \begin{matrix} 0.93 & 0.06 & 0.01 \\ 0.07 & 0.89 & 0.04 \\ 0.02 & 0.07 & 0.91 \end{matrix}$$

and the transition probability matrix between weather type (hidden state) and wind speed range (observable state, classified in the previous example) is:

$$B = \begin{matrix} 0.78 & 0.20 & 0.02 \\ 0.10 & 0.70 & 0.20 \\ 0.05 & 0.20 & 0.75 \end{matrix}$$

Fig. 3 illustrates this process:

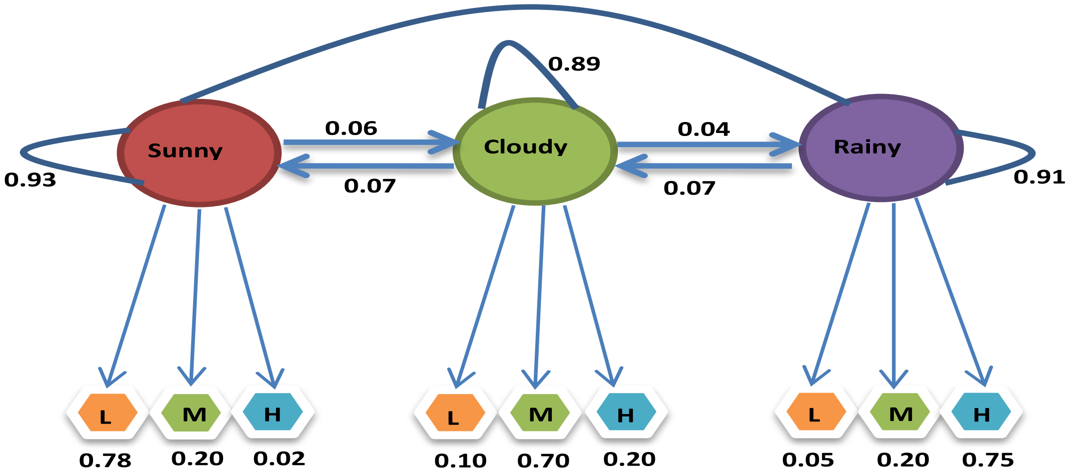

Fig. 3. Hidden Markov Model

Assume further that the length of the time series is 2 time intervals, the initial state $S_0 = s_1$. What is the probability of the process ending at $S_2 = s_3$ and a series of wind speed output of $O_1 = o_3, O_2 = o_1$?

We can easily find out that there are only 3 possible paths:

$$Path\ 1: s_1 \to s_1 \to s_3$$

$$Path\ 2: s_1 \to s_2 \to s_3$$

$$Path\ 3: s_1 \to s_3 \to s_3$$

Then the probability of each of the paths can be calculated as follows:

$$P_1 = \mathrm{P}(S_1 = s_1|S_0 = s_1) * \mathrm{P}(O_1 = o_3|S_1 = s_1) * \mathrm{P}(S_2 = s_3|S_1 = s_1) * \mathrm{P}(O_2 = o_1|S_2 = s_3)$$

$$= a_{11} * b_{13} * a_{13} * b_{31} = 0.93 * 0.02 * 0.01 * 0.05 = 0.000009$$

$$P_2 = \mathrm{P}(S_1 = s_2|S_0 = s_1) * \mathrm{P}(O_1 = o_3|S_1 = s_2) * \mathrm{P}(S_2 = s_3|S_1 = s_2) * \mathrm{P}(O_2 = o_1|S_2 = s_3)$$

$$= a_{12} * b_{23} * a_{23} * b_{31} = 0.06 * 0.20 * 0.04 * 0.05 = 0.000024$$

$$P_3 = \mathrm{P}(S_1 = s_3|S_0 = s_1) * \mathrm{P}(O_1 = o_3|S_1 = s_3) * \mathrm{P}(S_2 = s_3|S_1 = s_3) * \mathrm{P}(O_2 = o_1|S_2 = s_3)$$

$$= a_{13} * b_{33} * a_{33} * b_{31} = 0.01 * 0.75 * 0.91 * 0.05 = 0.000341$$

As a result, the probability of this event is $\mathrm{P} = P_1 + P_2 + P_3 = 0.000374$

Note that there are three underlying assumptions for HMM:

Assumption 1: the Markov process underlies the HMM follows a first-order Markov chain. This relationship can be presented using equation (7).

Assumption 2: the transition of state is independent of time, i.e. the relationship of transition probability at different time interval meets the following equations:

$$P\left(S_{t+1} = s_j \middle| S_{t+1} = s_i\right) = P\left(S_t = s_j \middle| S_t = s_j\right) \quad (11)$$

$$P\left(O_{t+1} = o_k \middle| S_{t+1} = s_j\right) = P\left(O_t = o_k \middle| S_t = s_j\right) \quad (12)$$

where k is the number of observable states.

Assumption 3: the observable state at time t is only dependent on the hidden state at time t indicating a relationship in equation (13):

$$\mathrm{P}(O_t = o_k|S_t = s_i) = \mathrm{P}(O_t = o_t|O_1 = o_1, O_2 = o_2 \ldots O_{t-1} = o_{k-1}; S_1 = s_1, S_2 = s_2 \ldots S_t = s_i) \quad (13)$$

3.2.3. Markov-Switching Autoregressive Model

A MS-AR process, as is mentioned in section 1, is a generalization of Hidden Markov Model (HMM) and AR models. This process is characterized by two components: S_t and Y_t. Y_t represents the observable states at time t and S_t represents the hidden state at time t. Here we assume that the hidden weather type follows a first order Markov Chain process. As a result, a MS(m)-AR(p) model, which means that the model includes an autoregressive process with an order of p and Markov chain process with m states, can then be interpreted as below:

The conditional distribution of S_t is a 1st order, m states Markov chain process. The value of S_t depends on the values of S_{t-1}.

The distribution of Y_t conditional on S_t is a pth order autoregressive process. The value of Y_t depends on the values of $Y_{t-1}, Y_{t-2}, Y_{t-3} \dots Y_{t-p}$ and S_t. So Y_t can be expressed as:

$$Y_t = a_0^{S_t} + \sum_{i=1}^{p} a_i^{S_t} Y_{t-i} + \delta^{S_t} \varepsilon_t \tag{14}$$

where $a_1^{S_t}, a_2^{S_t}, a_3^{S_t} \dots a_{t-p}^{S_t}$ are the coefficients of the autoregressive process given the state of S_t. $a_0^{S_t}$ is a constant given the state of S_t, ε_t is a sequence of error terms and δ^{S_t} is the standard deviation of the error sequence given the state of S_t. We can see that this equation is in a similar form of equation (1)

For example, In an MS (m)-AR (0) model, there are m states in the Markov Chain and the observable states only depend on the hidden state at the same interval, which is equivalent to an HMM.

For an MS (2)-AR (1) model, there are two states in the Markov Chain and the output observation at time t, Y_t is determined by both the hidden state at the same time period and the output observation at time t-1, Y_{t-1}. This relationship can be expressed using the equation similar to equation (2):

$$Y_t = a_0^{S_t} + a_1^{S_t} Y_{t-1} + \delta^{S_t} \varepsilon_t \tag{15}$$

The stream graph in Fig. 4 illustrates this process:

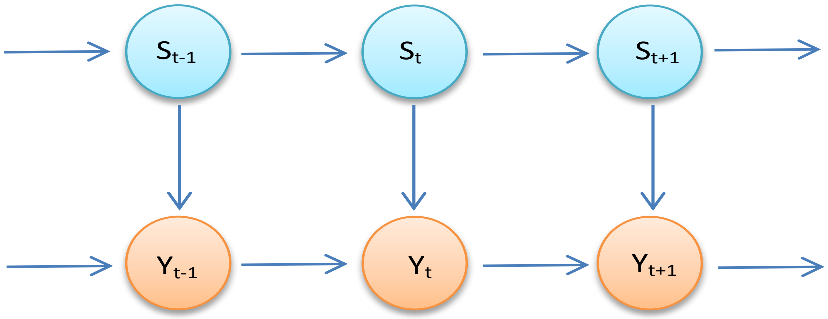

Fig. 4.MS (1)-AR (1) process

4. Method

4.1. Data

The wind speed data used in this research were obtained from NREL at 80 meters, with resolution of 10 minutes, from 2004 to 2005. The site chosen is in the middle part of Texas (Latitude: 35.77, Longitude: -100.94). The reason for choosing wind speed data instead if wind power output in this research is that wind speed data is not affected by transmission constraints or power system operator decisions and can reflect the nature relationship between weather type and wind speed. As MS-AR model introduced a latent variable that represents weather type, using wind speed data is more appropriate.

For comparison purposes, the two-year data are blocked by season (eight seasonal groups). In total, ten input groups are created (Table 2).

Table 2 Input groups

One-year Groups	
2004	2005
Seasonal Groups	
04 Spring	05 Spring
04 Summer	05 Summer
04 Fall	05 Fall
04 Winter	05 Winter

4.2. Data pre-processing

As most of the wind time series data are non-stationary, data pre-processing may be necessary. According to the Augmented Dickey-Fuller test, which has a null hypothesis of unit root, for all data groups, the null hypothesis is rejected in favor of alternative hypothesis (trend stationary, autoregressive, and autoregressive plus drift). The KPSS test shows that for all data groups, null hypothesis of stationarity is rejected. The results from the two tests indicate that data pre-processing is necessary.

Common data transformations are: power transformation, logarithm transformation, root square transformation, etc. According to previous study on hourly wind speeds [15], the diurnal non-stationarity in hourly wind speed data can be removed by the following transformation:

1. Power transformation is applied to adjust for non-Gaussian distribution of the series.

$$y_t' = (y_t)^{\,m} \tag{16}$$

2. The hourly expected wind speeds are subtracted from the results of power transformation. The differences are then divided by the hourly standard deviations.

$$y_t^* = [y_t' - \mu_t]/\sigma_t \tag{17}$$

μ_t represents the hourly expected wind speeds, σ_t represents the hourly standard deviations. After the simulation, the data can then be transformed back again.

According to the Box-Jenkins Method, non-stationary series can achieve stationarity by successively differencing the data. First and second order difference can be expressed as follows:

First order differencing:

$$y_t' = y_t - y_{t-1} \tag{18}$$

Second order differencing:

$$y_t'' = y_t' - y_{t-1}' \tag{19}$$

For AR model, data differencing is applied to all data groups. The first-order differenced data are then used as inputs to the models. Fig.5 shows the original wind speed data. Clear trends, drift and random walk can be identified from the series. Fig.6 shows the data after the first order differentiation. Note that the non-stationary components (i.e. trend, drift and random walk) are removed from the series. For the MS-AR model, the pre-processing of the data is generally not necessary due to the stochastic nature of the MS-AR model; however, initial transformation on the data may improve the performance of the model in terms of the negative simulated wind speed [11]. As a result, data with and without transformation are both needed for fitting MS-AR model. Differences in modeling results will be discussed in section 5.

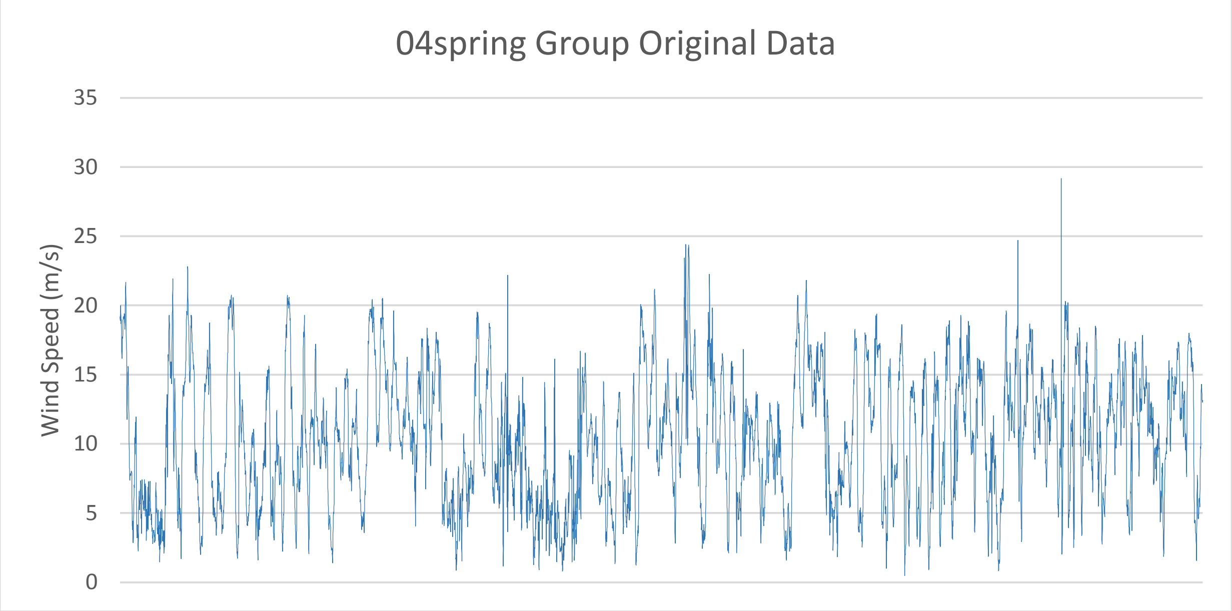

Fig. 5. Original wind speed data

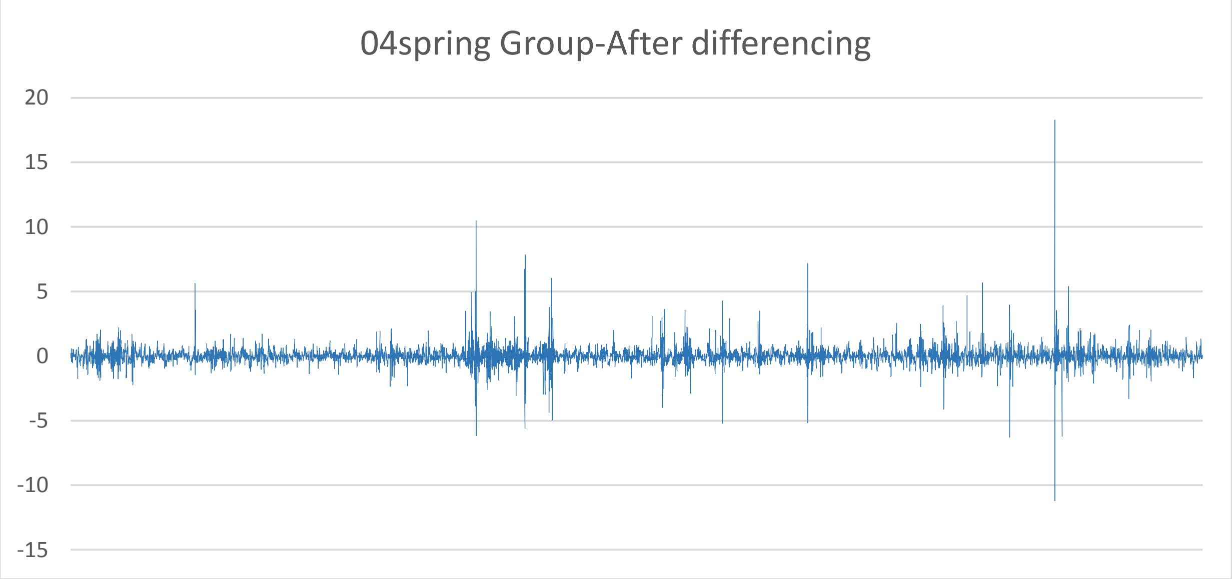

Fig. 6. Wind speed data after 1st order differencing

4.3. Simulation

In this study different models with various number of parameters are fitted to the input data. For AR model, 12 AR models ranging from 1st order to 12th order are fitted to each data group. After fitting model, 100 simulations are generated for each AR model. For MS-AR model, 16 MS-AR models ranging from MS(2)-AR(2) to MS(5)-AR(5) are fitted to each data group. 100 simulations are then generated for each MS-AR models.

5. Results and Discussion

5.1. Metrics of Model Performance

The performance of the models is measured by the average root mean square error (RMSE) of autocorrelation function (ACF), probability density function (PDF), and ramp distribution. This is calculated by taking the average value of RMSEs of 100 simulations for each model. Lower value of average RMSE indicates better model performance.

ACF, PDF and ramp distribution are very important attributes of data. ACF represents the correlation of values at different time periods in a series. It is expressed as a standardization of the autocovariance function (ACV), which shows the covariance in a series between one observation and another observation in the same series k lags away[16]. ACF can be calculated using the formulas below:

$$ACV(k) = \sum_{t=1}^{n-k}(y_t - \mu_t)(y_{t-k} - \mu_{t-k})/(n-k) \tag{20}$$

$$ACF(k) = \frac{ACV(k)}{\sigma_t \sigma_{t-k}} \tag{21}$$

If y_t is a stationary process, then the mean μ and variance $\sigma^2$are time independent, which gives us:

$$ACF(k) = \frac{E[(y_t - \mu)(y_{t-k} - \mu)]}{\sigma^2} \tag{22}$$

PDF describes the relative likelihood for a variable to take on given value. Ramp distribution represents changes in the value of the series from a time period to next time period. The performance of a simulation model will be based on how well it replicates these data attributes.

5.2. Autoregressive Models

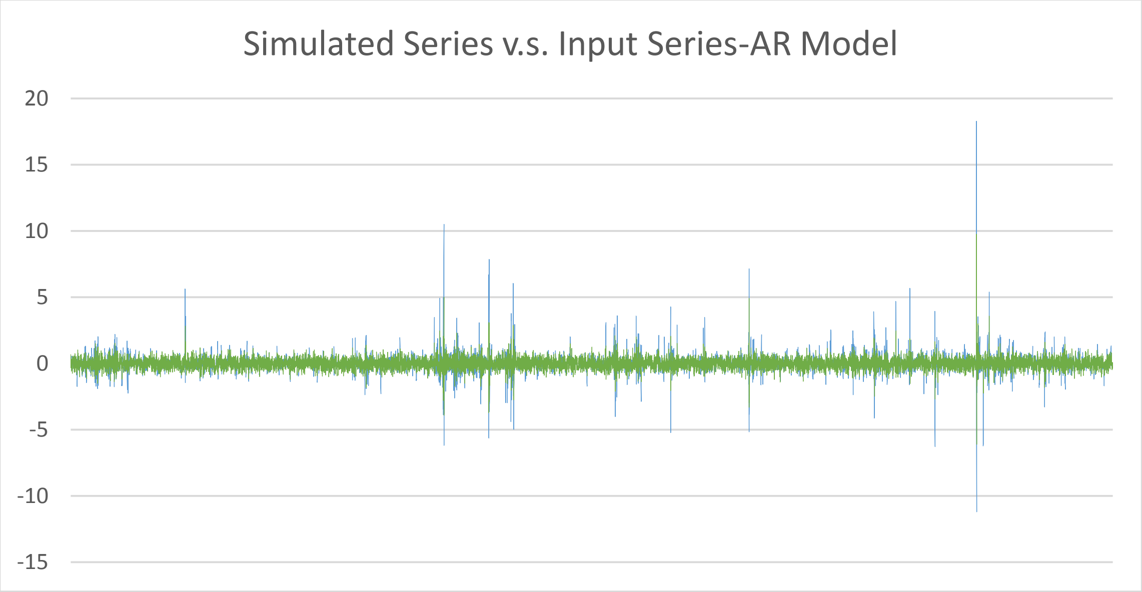

Fig. 7. Simulation result of AR model

Fig.7 shows one of the simulation results of the AR model. The green line represents the simulated data. As can be seen, the simulated series is centered at 0 with an obvious regressive pattern. In general, the autoregressive model has the ability to replicate the regressive pattern of the input series. It performs very well in terms of ACF, but not as good in terms of PDF.

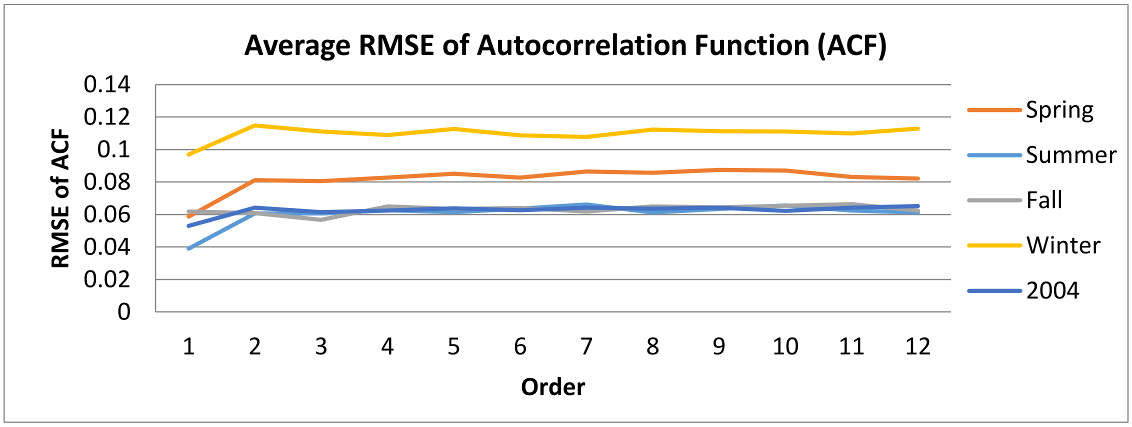

Fig. 8. Average RMSEs of AR models in terms of ACF

Fig. 8 shows the results for RMSEs in terms of ACF. For 04spring, 04summer, 04winter and 2004 groups, the 1st order AR model performs best. For 04fall group, the third order AR model performs best (average RMSE of 0.057). However, this value is not significantly smaller than that of a 1st order AR model (average RMSE of 0.062). Increasing order of model does not necessarily improve the performance of the model.

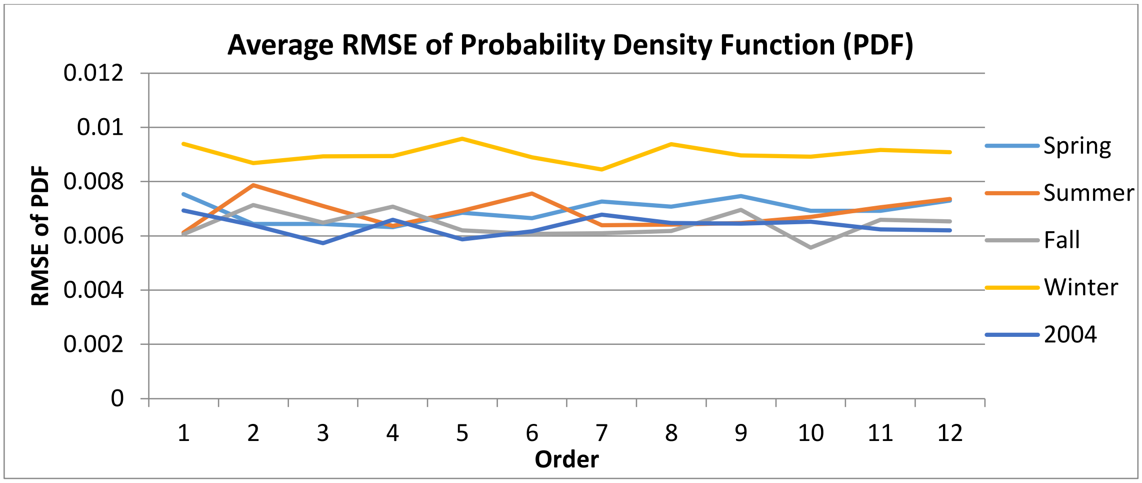

Fig. 9. Average RMSEs of AR models in terms of PDF

According to Fig. 9 the average RMSEs of PDF vary a lot with different orders. It is difficult to identify a best model or a trend in the improvement of the model performance.

However, according to Bayesian Information Criterion (BIC), the best model that can be applied to all data groups is 1st order AR model. BIC is a model selection criterion and model with the lowest BIC value is selected. When fitting models, it is possible to increase the number of parameters to increase the likelihood of the model. This, however, may result in over-fitting. An over-fitted model will not be able to estimate future behavior in an appropriate way. BIC solves this problem by heavily penalizing the models with higher complexity given the same performance. In the end, the model with the lowest BIC value should be selected.

It is worth noting that average RMSEs of 04winter group are particularly higher than that of other groups. The reason may be related to the stationarity of input data. According to [16], there are two types of stationarity, weak stationarity and strong stationarity. A weakly stationary series is a series whose mean and variance are constant over time. The autocovariance only depends on the number of time lags. However, a strongly stationary series is a series that meets the requirement of weak stationarity, and is also normally distributed. Further analysis shows that after first order differencing, the non-stationary components are removed from the data; however, the series is not normally distributed. For other groups, the series after first order differencing follows normal distribution. Results can be seen in Table 3. As AR model is very sensitive to input data, weak stationarity may lead to inaccurate simulation. As a result, the average RMSE between simulated data and input data is greater for 04winter data group. 2005 data groups show the same results.

Table 3 Result of normal distribution test: h value of 0 indicates normal distribution, h value of 1 indicates rejection of normal distribution.

Data groups	04spring	04summer	04fall	04winter	2004
H value	0	0	0	1	0
Data groups	05spring	05summer	05fall	05winter	2005
H value	0	0	0	1	0

5.3. Markov-Switching Autoregressive Models

Fig.11 shows the simulated wind speed time series using the MS-AR model. The simulated series (red) also shows a regressive pattern but has more jumps and better captures the characteristics of the input data.

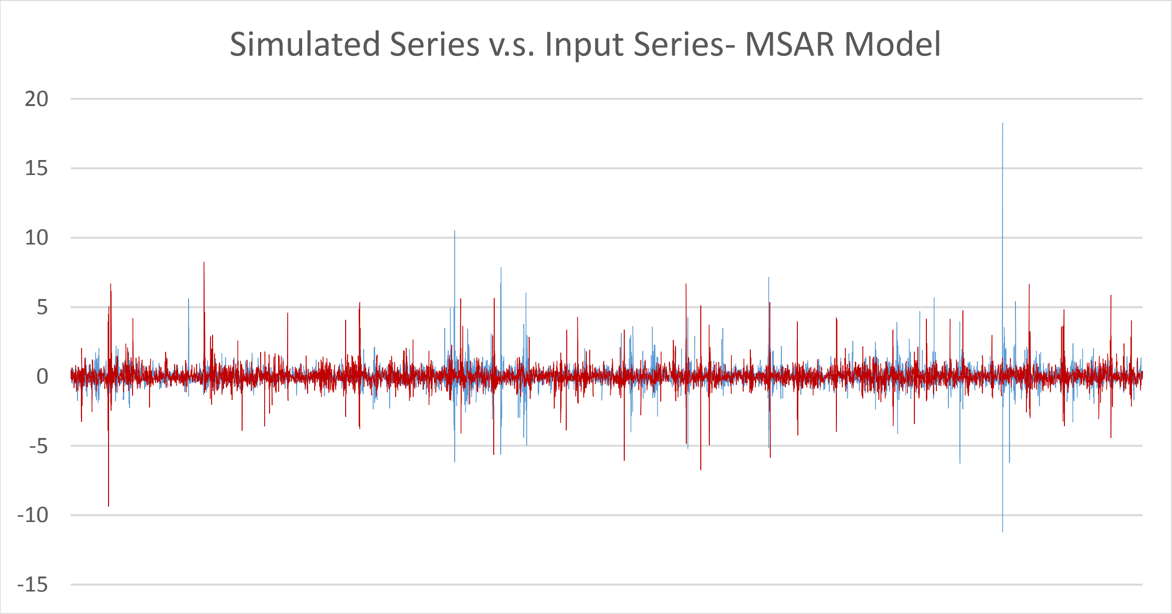

Fig. 10. Simulation result of MS-AR model

5.3.1. Differenced data

Fig.11 and Fig.12 show the results of MS-AR model in terms of ACF and PDF respectively. The average RMSEs in terms of both ACF and PDF decrease with increasing number of states in the Markov Chain, indicating that the model performance increases with increasing number of states in the Markov Chain.

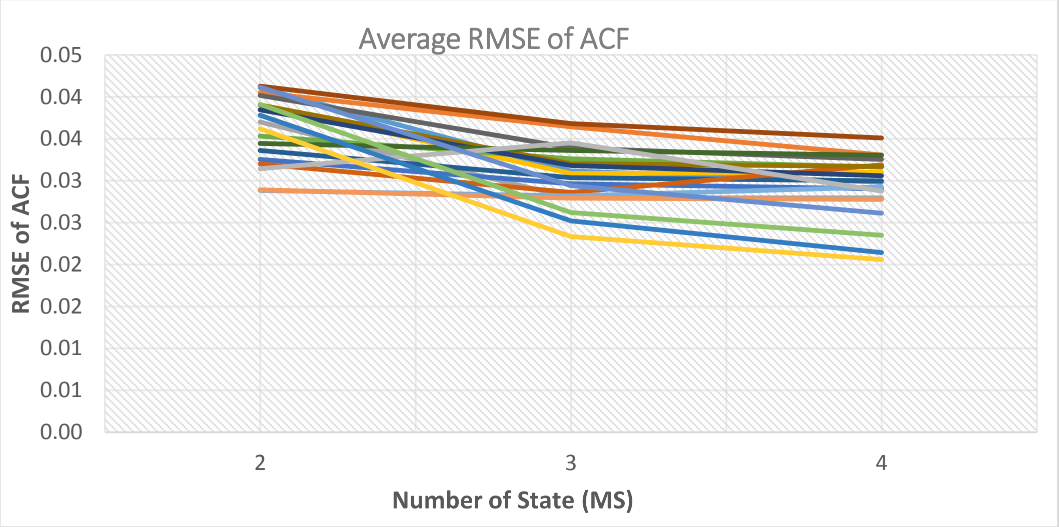

Fig. 11. Average RMSEs of MS-AR models in terms of ACF

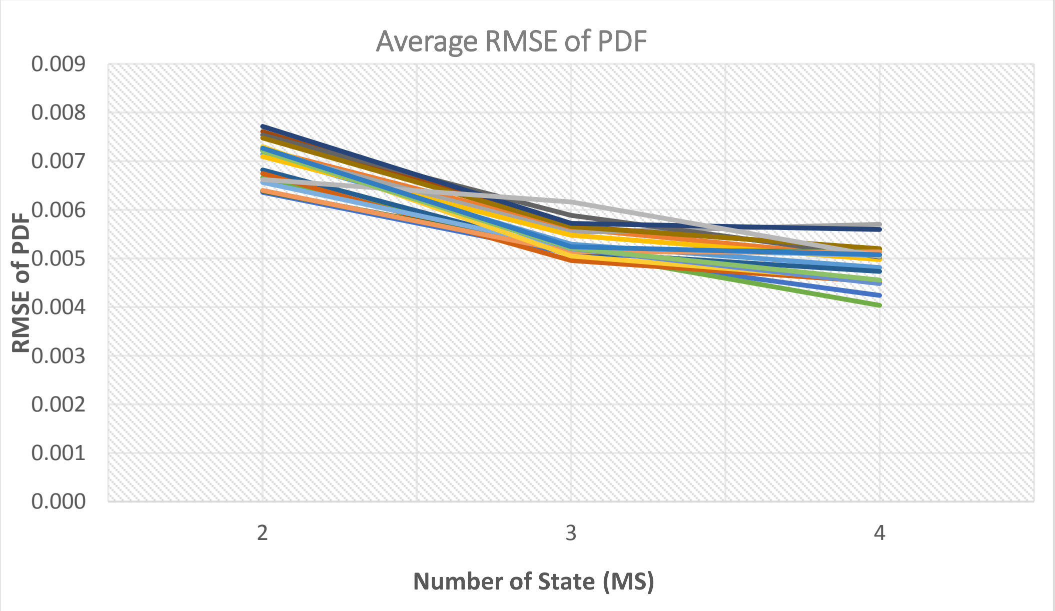

Fig. 12. Average RMSEs of MS-AR models in terms of PDF

5.3.2. Original data

Original 04spring data are also used as input to MS-AR model. Results show obvious improvement in terms of ACF, PDF and ramp distribution with increasing number of state in the Markov Chain and the increasing order in autoregressive component (Table 4). However, there are some limitations in using original data in this research. Firstly, as is mentioned previously, although no data pre-processing is required for MS-AR model, using the original series directly as input data leads to simulation results that include negative wind-speed values which are meaningless. Secondly, as MS-AR model works by switching between states in the Markov Chain and autoregressive components using two transition probability matrixes, it requires large amount of computation. Large, non-stationary data with increasing number of parameters in the model will make more difficult model convergence. As the model used in this research is a simple version model, increasing model complexity may result in parameters' estimation from very limited data, which may in turn lead reduce model performance. Also, it will significantly increase the computation time. Moreover, models that have more than 5 states in the Markov Chain and autoregressive component higher than 5th order are not able to generate meaningful results. Thirdly, for models with less than 4 states and 4th order, the results of ACF and ramp distribution are comparable to that of AR model. The performance of MS-AR model in terms of ACF and ramp distribution is better than that of AR model (Table 4 & 5). However, the results of PDF of MS-AR model and AR model are in different order of magnitude, which may result from different distribution or different order of magnitude of input data.

Table 4 Average RMSEs of MS-AR models in terms of ACF, PDF, and ramp distribution with 04spring original data as input

RMSE ACF				
	Order	2	3	4
State	2	0.0390	0.0248	0.0278
	3	0.0267	0.0186	0.0231
	4	0.0148	0.0125	0.0220
RMSE PDF				

State	Order	2	3	4
	2	0.0222	0.0187	0.0253
	3	0.0140	0.0113	0.0155
	4	0.0137	0.0104	0.0125
RMSE Ramp				
	Order	2	3	4
State	2	0.8772	0.8856	0.8733
	3	0.8635	0.8539	0.8625
	4	0.8565	0.8427	0.8218

Table 5 Average RMSEs of AR models in terms of ACF, PDF, and ramp distribution with o4spring differenced data as input

04 Spring	
RMSE ACF	0.058
RMSE PDF	0.007
RMSE Ramp	0.870

5.4. Comparison of AR and MS-AR Models

For comparison purposes, the difference between average RMSEs of AR model and that of MS-AR model in terms of ACF and PDF are calculated and plotted. Fig.13 shows the results of difference in ACF. The difference values are all positive, indicating that the average RMSEs of AR model are greater than that of MS-AR model. This means that MS-AR model outperforms AR model in terms of ACF. Moreover, the difference increases with increasing number of state in the Markov Chain. As can be seen in Fig.14, the difference in PDF shows the same increasing trend. However, when there are only two states in the Markov Chain, some difference values are negative. Groups with the negative values are 04summer, 04fall and 2004 groups. As is mentioned earlier, the performance of AR model significantly varies with the input data quality. The negative values are the result of better performance of AR model with these groups. All difference values become positive as the number of state in the Markov Chain increases from 2 to 3. As the number of state continues to increase, the difference becomes greater. Moreover, as Markov Chain with 2 states is rarely used for simulation in real life, we can conclude that generally the performance of MS-AR model can also outperform AR model in terms of PDF.

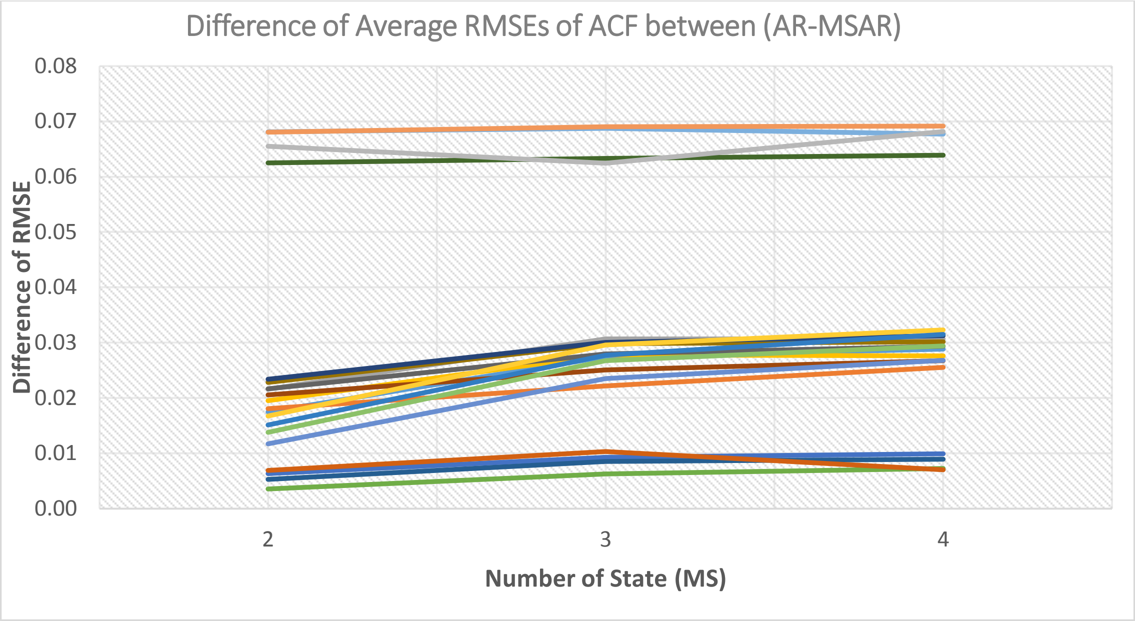

Fig. 13. Difference between Average RMSEs of MS-AR models in terms of ACF

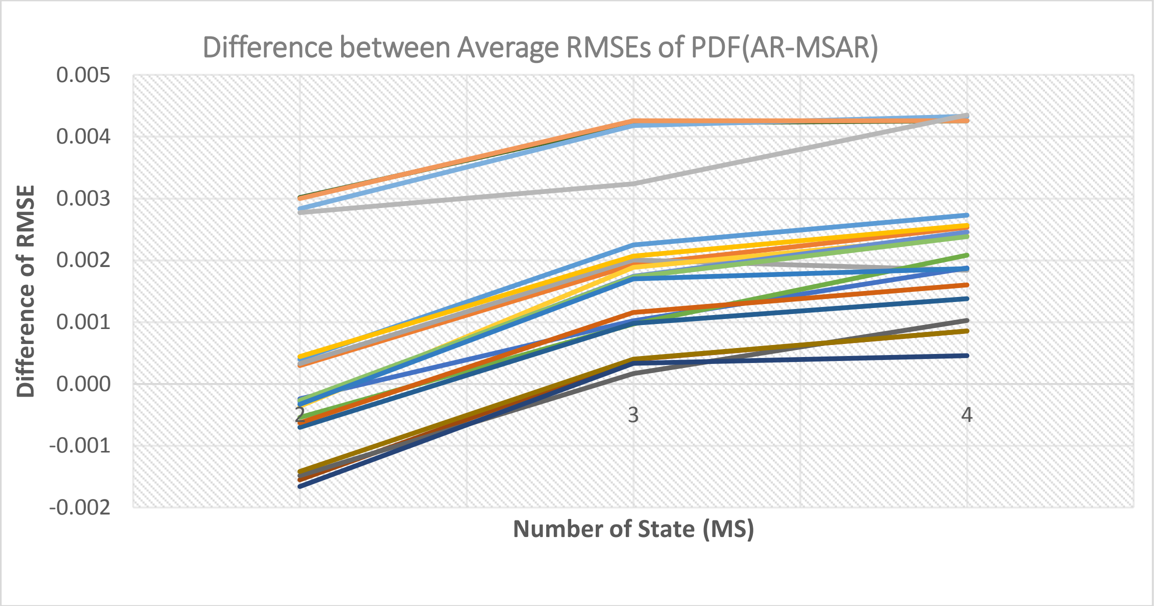

Fig. 14. Difference between Average RMSEs of MS-AR models in terms of PDF

In Fig.13 and Fig.14, 04winter group also stands out. This results from particularly higher average RMSEs of AR model. However, as is shown in Fig.11 and Fig.12, the performance of MS-AR model is not affected by the weak stationarity of 04winter group. This is because of the MS-AR models' ability to describe the marginal distribution of the time series[11]. With the model’s distributional versatility, achieving strict stationarity or stationarity in general is not necessary. This finally leads to the greater difference value for 04winter group.

6. Conclusions

This master’s project investigates the performance of two wind simulation models, AR model and MS-AR model, using wind speed data from NREL. From this application we can draw the following conclusions:

1. MS-AR models outperform AR models in terms of both ACF and PDF. AR models are known to be able to perfectly replicate ACF, but perform not as good in terms of PDF. The Markov Chain component in MS-AR models further improves the ACF performance as well as the performance in terms of PDF.
2. In this research, the effect of increasing model order in the autoregressive component in MS-AR models is unclear. However, the increasing number of state in the Markov Chain can significantly improve the performance of MS-AR models.
3. MS-AR models are more tolerant to input data, which is a result of their distributional versatility. Although all input data groups are 1st order differenced in this research because of comparison purpose and some model limitation, in general no data pre-processing is needed. However, for wind simulation specifically, negative simulated data caused by not transforming the original data may be problematic.
4. Although MS-AR models are better than AR models in many ways, we cannot deny that AR models are more efficient as their simplicity and time saving characteristic can to some extent offsets less perfect performance.
5. According to the results of ramp distribution, the average RMSEs of AR model are lower than that of MS-AR models in terms of ramp distribution, which leads to negative difference values. However, the difference is decreasing with increasing number of state in the Markov

Chain and the difference values are close to zero with 4 states in the model. It is probable that with more state in the Markov Chain, MS-AR models will finally outperform AR models. In this research, however, MS-AR models with more than 5 states in the Markov Chain are not used because of the limitations of time and the simple version model. As a result, ramp distribution is not discussed here.

7. References

1. *Global Wind Energy Outlook 2014*. 2014, Global Wind Energy Council.
2. *AWEA Reliability White Paper*. 2015, American Wind Energy Association. p. 5.
3. Karki R., H.P., Billinton R., *A simplified wind power generation model for reliability evaluation.* IEEE, 2006. **21**: p. 533-540.
4. A., A., *Evaluating wind power generating capacity adequacy using MCMC time series model*. 2014, University of Waterloo. p. 80.
5. Shamshad A., B., M., Wanhussin W., Majid T., Sanusi S., *First and second order Markov chain models for synthetic generation of wind speed time series.* Energy, 2005. **30**(5): p. 693-708.
6. Denaxas E. A. , B.R., Patino-Echeverri D., Pitsianis N. *SynTiSe: A Modified Multi-Regime MCMC approach for Generation of Wind Power Synthetic Time Series*. in *IEEE Systems Conference 2015*. 2015. Vancouver.
7. *SynTiSe Software*. 2015 [cited 2015 April 23rd]; Available from: Available at sites.nicholas.duke.edu/daliapatinoecheverri/files/2015/SynTiSe_Folder.gz.
8. Wu T., A.X., Lin W., Wen J., Luo W., *Markov Chain Monte Carlo method for the modeling of wind power time series.* Innovative Smart Grid Technologies-Asia (ISGT Asia), 2012: p. 1-6.
9. Suomalainen K., S., C.A., Ferrão P., Connors, S., *Synthetic wind speed scenarios including diurnal effects: Implications for wind power dimensioning.* Energy, 2011.
10. Blanchard M., D.G., *Generation of autocorrelated wind speeds for wind energy conversion system studies.* Solar Energy, 1983. **33**: p. 571-579.
11. Ailliot P., M.V., *Markov-switching autoregressive models for wind time series.* Environmental Modelling & Software, 2012.
12. University., D. *Stationarity and Diffrencing*. Available from: http://people.duke.edu/~rnau/411diff.htm.
13. MathWorks. *Documentation: adftest*. Available from: http://www.mathworks.com/help/econ/adftest.html.
14. MathWorks. *Documentation: kpsstest*. Available from: http://www.mathworks.com/help/econ/kpsstest.html?searchHighlight=kpss.

15. Brown B.G., K.R.W., Murphy A.H., *Time series models to simulate and forecast wind speed and wind power.* Journal of Climate and Applied Meteorology, 1984. **23**.

16. Yaffee R., M.M., *Introduction to Time Series Analysis and Forecasting*. 2000.