A Report

On

Statistical Analysis and Forecasting of Wind Speed Data (Inter-States)

By

Group 1

Aditya Garg	2019A8PS0309P
Agneya Bhardwaj	2019A8PS0297P
Harsh Solanki	2019A1PS0670P
Josh Wadhwa	2019A1PS0824P
Sagar Jain	2019A2PS0902P
Saisreya NV	2018B4TS1154P
Surya Rathi	2019A7PS0128P
Yash Gupta	2019A7PS1138P

Prepared in Partial Fulfillment of the Course

Math F432 Applied Statistical Methods

Birla Institute of Technology and Science, Pilani

December, 2021

Introduction

In the recent United Nations Climate Change Conference, better known as COP26, there had been a lot of pressure on India to eliminate its dependency on fossil fuels - coal in this case - for electric power supply [1]. Not only does the use of thermal energy based on coal consumption present the problem of air pollution and global warming, but it is also limited in supply; we, therefore, can not possibly hope to rely on the present conventional method of power generation. Hydroelectric and nuclear power, as alternatives, present the problem of mass forest destruction for dam construction and hazardous radioactive pollution respectively.

Our search for better sources of energy, therefore, takes us to options like solar and wind energies. One of the reasons why this is already not taking place at a large scale in India is because of the difficulty in its implementation due to its supposed seasonal and geographical irregularities. Wind energy, however, is governed by natural weather and climatic conditions, and can not be deterministically set for a future period - much unlike our consumption of fossil fuels, where a thermal power plant works within prescribed bounds. Such nature of wind energy warrants thorough statistical analysis to be of practical use to energy policy makers and power grid regulators.

The present report endeavours to perform a statistical study of wind energy and the various environmental factors associated with it. Specifically, we focus on four major Indian states: Andhra Pradesh, Madhya Pradesh, Rajasthan, and Tamil Nadu. Considering wind speed as a proxy for wind energy generation rate, we investigate its correlation with other parameters - temperature, pressure, solar irradiance, and the like - and present a descriptive statistical analysis of the same. We perform a search for a distribution to fit the available data with a view to make future analysis easier and systematic.

Time-series analysis on the wind speed data to try to find any trend or seasonality in the data. Finally, we used the statistical techniques AR, MA, ARMA, ARIMA, and SARIMA to forecast weekly wind speed.

1. Dataset

The given dataset contains hourly data of various environmental factors that might be related to wind energy for the Indian states of Rajasthan, Madhya Pradesh, Andhra Pradesh and Tamil Nadu over the period 2000 to 2014.

In every entry, it had the following attributes:

Attribute	What does it mean?
Timestamp of Measurement	The time of recording the environmental observations as given by date and time. [2]
DNI and Clearsky DNI	Direct Normal Irradiance: solar irradiance measured at the surface of the Earth at a given location with a surface element perpendicular to the Sun. [2]
DHI and Clearsky DHI	Diffuse Horizontal Irradiance: radiation at the Earth's surface from light scattered by the atmosphere. [2]
GHI and Clearsky GHI	Global Horizontal Irradiance: total irradiance from the sun on a horizontal surface on Earth. $GHI = DHI + DNI\,cos(z)$ [2]
Solar Zenith Angle	The angle between the sun's rays and the vertical direction. [2]
Temperature	The temperature, in degree celsius, at the time of measurement.
Pressure	The atmospheric pressure, in millibar, at the time of measurement.
Relative Humidity	The amount of water vapour present in air expressed as a percentage of the amount needed for saturation at the same temperature.
Dew Point	The temperature the air needs to be cooled to at a constant pressure to obtain a relative humidity of 100%. [3]
Wind Speed	The rate at which air is moving over the area in m/s.

The data is in the form of a time series, where a time series is defined as a sequence of periodically recorded observations of a variable.

Throughout our discussion, we will be measuring the quality of predictions using the metric MAPE, or the Mean Absolute Percentage Error, which is defined as follows:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Ft - Xt}{Xt}\right|$$

where n is the number of observations, F_t is the forecast at time t, and X_t is the actual observation at time t.

2. Correlation

Correlation is a statistical measure that expresses the extent to which two variables are linearly related. We describe correlations with a unit-free measure called the correlation coefficient which ranges from -1 to +1 and is denoted by r. The closer r is to zero, the weaker the linear relationship. Positive r values indicate a positive correlation, where the values of both variables tend to increase together. Negative r values indicate a negative correlation, where the values of one variable tend to increase when the values of the other variable decrease.

Mathematically, correlation between two variables x and y can be given as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

For the given data we obtained a feature-feature correlation map for each state using Excel's data analysis tool-pack. For the sake of simplicity, only the correlation map of Andhra Pradesh is shown here. Other states which have similar results have been shown in the appendix.

	DHI	DNI	GHI	Clearsky DHI	Clearsky DNI	Clearsky GHI	Dew Point	Temperature	Pressure	Relative Humidity	Zenith Angle	Wind Speed
DHI	1	0.71962	0.87491	0.93709	0.94296	0.94209	-0.0146	0.68862	-0.1256	-0.6203	-0.8739	0.009
DNI	0.71962	1	0.94005	0.87227	0.8701	0.86297	-0.1254	0.66154	0.01174	-0.6402	-0.7838	-0.1655
GHI	0.87491	0.94005	1	0.94592	0.94202	0.97521	-0.0777	0.73999	-0.0773	-0.6792	-0.8675	-0.0873
Clearsky DHI	0.93709	0.87227	0.94592	1	0.96443	0.95852	-0.087	0.76105	-0.0739	-0.7023	-0.9025	-0.1013
Clearsky DNI	0.94296	0.8701	0.94202	0.96443	1	0.97091	-0.0325	0.69242	-0.084	-0.6424	-0.9072	-0.0367
Clearsky GHI	0.94209	0.86297	0.97521	0.95852	0.97091	1	-0.0394	0.73504	-0.1156	-0.6643	-0.8947	-0.0311
Dew Point	-0.0146	-0.1254	-0.0777	-0.087	-0.0325	-0.0394	1	-0.0338	-0.2988	0.50569	-0.0055	0.09011
Temperature	0.68862	0.66154	0.73999	0.76105	0.69242	0.73504	-0.0338	1	-0.3771	-0.8435	-0.7164	0.0083
Pressure	-0.1256	0.01174	-0.0773	-0.0739	-0.084	-0.1156	-0.2988	-0.3771	1	0.15991	0.14749	-0.4639
Relative Humidity	-0.6203	-0.6402	-0.6792	-0.7023	-0.6424	-0.6643	0.50569	-0.8435	0.15991	1	0.63502	0.00555
Zenith Angle	-0.8739	-0.7838	-0.8675	-0.9025	-0.9072	-0.8947	-0.0055	-0.7164	0.14749	0.63502	1	0.01807
Wind Speed	0.009	-0.1655	-0.0873	-0.1013	-0.0367	-0.0311	0.09011	0.0083	-0.4639	0.00555	0.01807	1

Fig 1: Correlation heat map for Andhra Pradesh wind energy data

3. Best-Fit Distributions for Wind Speed Data

The objective of the exercise was to find a "good" fit for the underlying population distribution of the wind speed data for the various states. Since normal distribution is a widely used distribution, applicable to a huge spectrum of situations, we decided to first test our data for normality. If the normal distribution does not perform well, we decided to search for a suitable distribution among commonly used continuous distributions.

3.1 Tests for Normality

We performed tests to check if our data fits the normality assumption by using two common statistical tests: the Shapiro-Wilk test and the D'Agostino's K^2 test.

Null Hypothesis H_0	The wind speed population has a normal probability distribution.
Alternative Hypothesis H_a	The wind speed population does **NOT** have a normal probability distribution.
Level of Significance $\alpha = 0.05$	Reject H_0 if $p - value \leq \alpha$

3.1.1 Visual inspection

Histograms for the wind speed data were plotted for each state. Even a cursory glance at the visual representation reveals that the data have a significant amount of skew associated with them, making the assumption of normality unlikely.

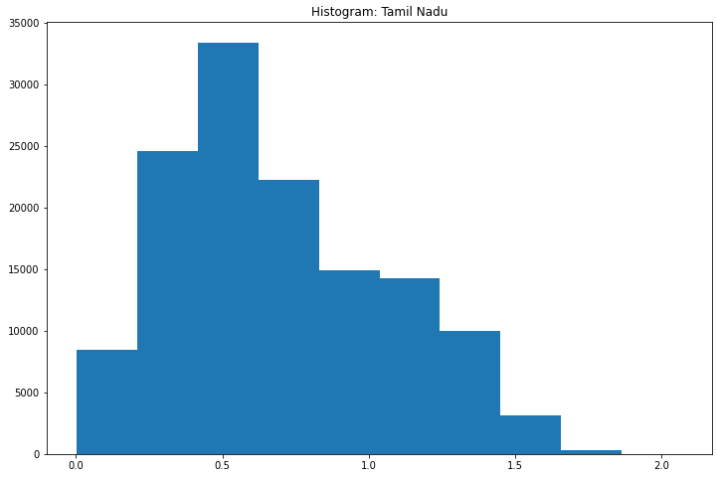


Figure 3.1: Histogram of wind speeds for Tamil Nadu.
Notice that the data is somewhat right skewed.

3.1.2 Shapiro-Wilk Test

The library function `scipy.stats.shapiro()` was used to test the data for each state for normality. The null hypothesis - of the data coming from normal distribution - was *rejected* for each state. The Shapiro–Wilk test, however, is generally considered a more appropriate method for small sample sizes ($n \leq 50$) [4]. Hence, we decided to perform further tests to verify the results.

3.1.3 D'Agostino's K^2 Test

The D'Agostino's K^2 test measures the departure from normality of the given data by measuring its sample skewness g_1 and sample kurtosis g_2. The K^2 test statistic, which may be translated to a $p - value$, is estimated as $K^2 = Z_1(g_1)^2 + Z_2(g_2)^2$, where $Z_1$and Z_2 are appropriate transformations employed in the test.

The test was implemented with `scipy.stats.normaltest()`, and the assumption of normality was *rejected* for each state.

3.2 Finding the Best-Fit Distributions

Since the assumption of normality did not hold for any of the four states, we proceeded to search for a suitable distribution to fit the data from a list of plausible candidates: *Beta, Gamma, Rayleigh, Normal, Logistic (Sech-squared), Weibull Minimum Extreme Value, Weibull Maximum Extreme Value, Lognormal, Chi, Chi-squared, Exponential, Inverse Gamma, Log Gamma, Log Laplace*, and *Cosine* distributions.

Each of the above distributions was fitted to the wind speed data for the particular state, and the arguments, location, and scale for the pdf were calculated. The *sum of squared errors (SSE)* and the *negative log likelihood function (NNLF)* - as available through `scipy.stats.rv_continuous.nnlf()` - were calculated to estimate the error in fitting the data. The distributions were ranked in the increasing order of the magnitude of *NNLF*, and the pdfs of the first three distributions for each state were plotted with the data to visualize the fit.

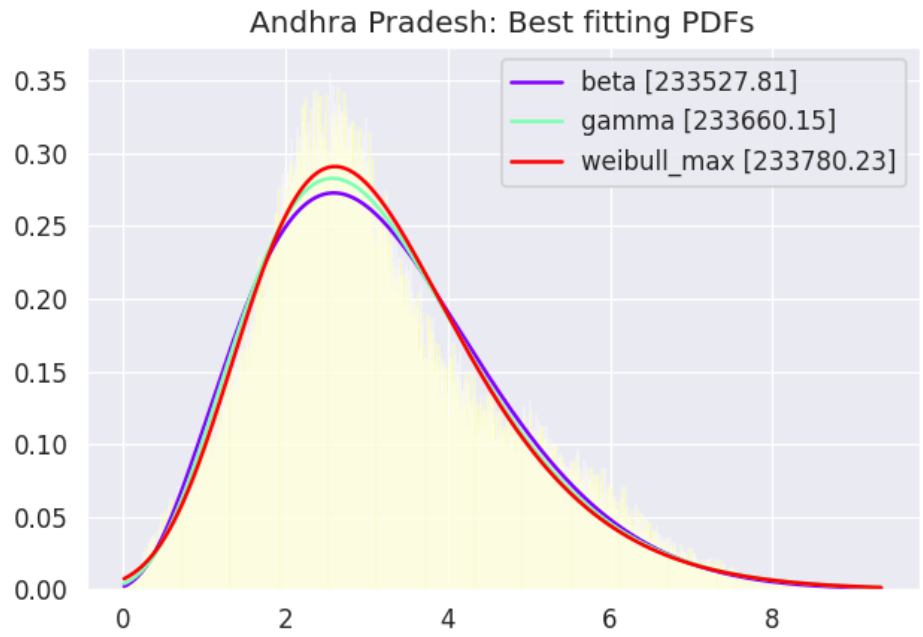


Figure 3.2: The top 3 best fitting distributions for Andhra Pradesh.

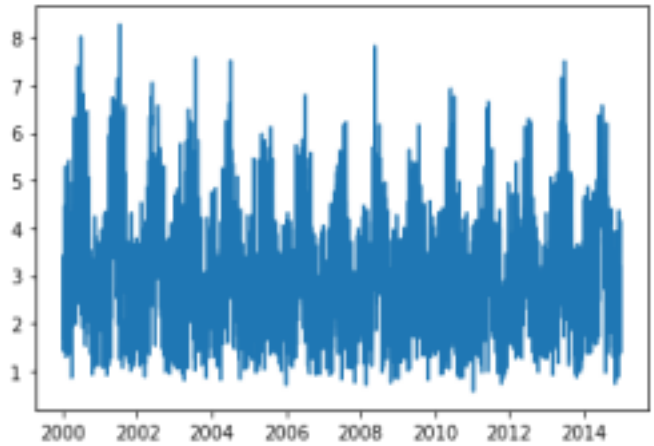
The results of the search are summarized in table 6.1 below.

Table 3.1: States and the corresponding best three fitting distributions for wind speeds.

State	Best-fit Distr. (1st)	Best-fit Distr. (2nd)	Best-fit Distr. (3rd)
Tamil Nadu	beta	chi	weibull min
Andhra Pradesh	beta	gamma	weibull max
Madhya Pradesh	weibull min	chi	beta
Rajasthan	weibull min	chi	beta

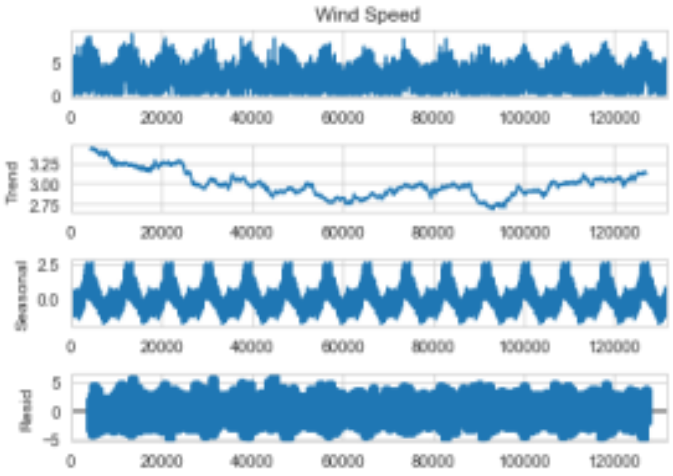
It is interesting to note how the two coastal states - Andhra Pradesh and Tamil Nadu - have the same best fitting distribution - beta - and how the landlocked states - Rajasthan and Madhya Pradesh - too share the same weibull min as the best fitting distribution.

4. Time Series Decomposition



To begin our Time Series Analysis, we plotted the given time series data. A plot of the hourly data for Rajasthan is as shown. We observed 15 peaks for the 15-year data, which led us to assume yearly seasonality (modelled after real seasons). Also, we did not see a noticeable variation of the data mean over time, leading us to the choice of additive model for the time series decomposition.

We decomposed the time series into trend, seasonality and residual components using the additive model, illustrated as follows for the state of Rajasthan.



- The uniform seasonality, no uniform variation in the trend component, and low residuals gave us justification for the choice of the additive decomposition model.
- Moreover, the variations in Wind Speed for each of the states can be explained through natural seasonal wind patterns in each location. For instance, the observed seasonal peak around May-June each year in the data for Rajasthan can be explained by the Loo, which hits a peak around that time as well.
- Finally, the data looks to be roughly stationary for the time series. However, this assumption needs to be tested.

5. Stationarity Tests

5.1 ADF Test: To check the stationarity we performed the Augmented Dickey-Fuller test. The hypotheses are as follows:

H_0: The series has a unit root

H_a: The series has no unit root

```
Test Stat: -22.487712452767557
p-value: 0.0
Crit value at 1% LOS: -3.4303997953780967
Crit value at 5% LOS: -2.8615620088348286
Crit value at 10% LOS: -2.5667817145225587
```

As shown, the value of the test statistic is seen to be more negative than the 1% critical value. Thus, we reject the null hypothesis and conclude that this is a stationary series.

5.2 KPSS Test: It is always better to apply both the tests (ADF and KPSS), so that it can be ensured that the series is stationary or if it needs any kind of differencing to make it stationary. To validate the results of stationarity we performed the KPSS test.
H_0: The series is stationary
H_a: The series has a unit root (it is not stationary)

As shown, the p-value (0.01001….) is larger than the alpha which tells us that we cannot reject the null hypothesis, and thus, conclude that the data is stationary.

```
Test Stat: 1.9474099510337564
p-value: 0.01
num lags: 73
Crit value at 1% LOS: 0.739
Crit value at 5% LOS: 0.463
Crit value at 10% LOS: 0.347
```

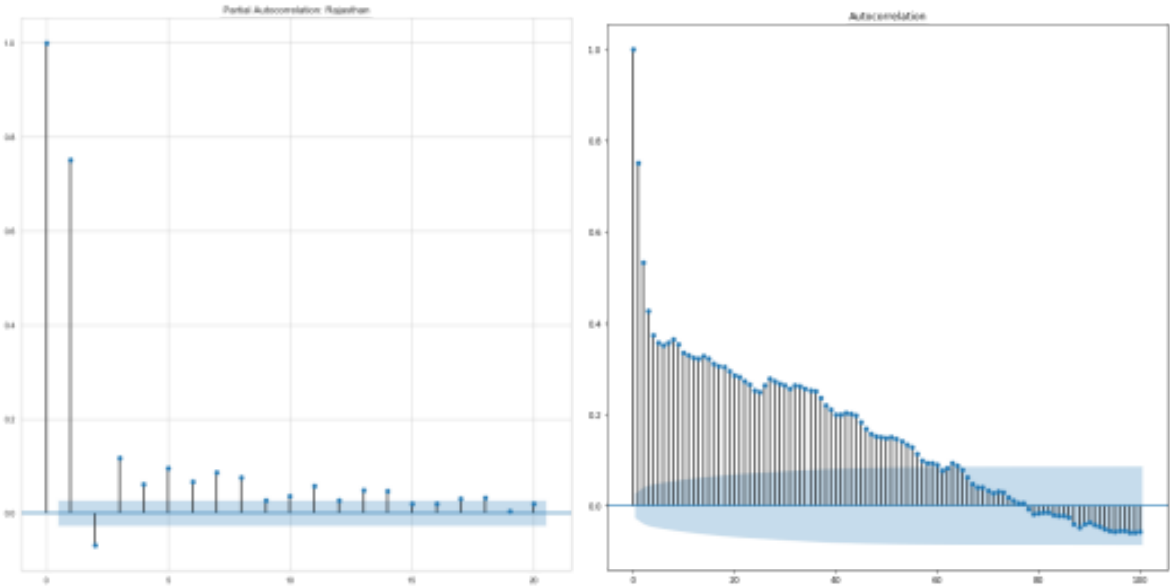
The two tests returned that the data was stationary for all four states.

6. Model Fitting and Forecasting

We performed forecasting on the time series data by fitting AR, MA, ARMA, ARIMA and SARIMA models. The first four models were used to forecast daily and weekly data, while the SARIMA model was used only to forecast weekly and monthly data for lack of computing power. The data was aggregated as a mean according to the timespan (daily/weekly/monthly). This section will analyze part of the process for the data for Rajasthan. Appendix gives the same for the state of Madhya Pradesh.

6.1 Parameter Estimation

We plotted the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots for the data – the plots for the daily Wind Speed data of Rajasthan are as shown below.



Observing the number of significant lag values before the data falls into the confidence region (marked with blue) gives us the estimates for the parameters of the MA and AR models respectively (from ACF – MA (62) model, from PACF – AR (16) model).

For the other models, the parameters were obtained through grid search in small regions of the domain (keeping the computational time in mind). For daily data of Rajasthan, the final models were found as:

• AR (16)
• MA (62) *[MA (20) was used for fitting due to computational power constraints.]*
• ARMA (9, 10)
• ARIMA (9, 1, 10)

For monthly data, the SARIMA model was found as:
• SARIMA ([2, 0, 3], [2, 0, 1], 12)

6.2 Model Fitting

All the models were fit on the data, and MAPE values were calculated using the residuals to find the best model for out-of-sample forecasting. The residual ACF plots are also shown for the ultimately best fitting models to justify the noise being 'white'. The fits are as described below:

6.2.1 AR Model

This model is a multiple regression model that forecasts the values of the target variable using a linear combination on its past values. It can be described by the following equation, where X is the forecasted variable, φ is the autoregressive operator (polynomial of order p), B is the backshift operator, and w is the white noise.

$$\phi(B)\, X_t = w_t$$

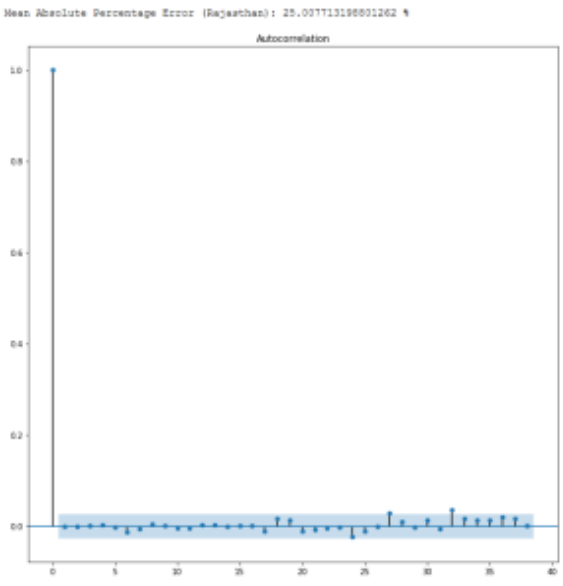
We fit the model with an order of 16, and obtained a MAPE value of 25.078%. The parameters estimated showed low p-values, indicating a good fit. The residual ACF plot also showed mostly insignificant values, indicating that the residuals can be considered white noise.

6.2.2 MA Model

This model is effectively a regression model on the past residuals/errors/noise, and uses moving averages of the past errors to forecast the new value. It can be represented by the following expression, where θ is the moving average operator (polynomial of order q).

$$X_t = \theta(B)\, w_t$$

We fit the model with an order of 20 (due to computational power limitations), and obtained a MAPE value of 25.341%. The parameters estimated showed low p-values, indicating a good fit. The residual ACF plot showed some statistically significant values, however, indicating that this model is clearly a worse fit than the AR model.



6.2.3 ARMA Model

This model essentially combines the previous two models for a stationary time series, performing regression-like fitting both ways.

Using previously described notation, it can be expressed by:

$$\phi(B)\, X_t = \theta(B)\, w_t$$

We fit the model with order (9, 10), and obtained a MAPE value of 25.007%, as well as low p-values for the feature weights, indicating a good fit. The residual ACF plot was also the best of all the models, having barely any significant values. This model would turn out to be the best fit for daily data (although still not very good as we can see from the feature weights p-values and the MAPE score).

6.2.4 ARIMA Model

This is a modification of the ARMA model that applies to non-stationary time series data as well. It converts the time series data to stationary by using differencing between present and past values (up to a given order d) as the new values for the ARMA model. It can be represented as:

$$\varphi(B)(1-B)^d X_t = \theta(B) w_t$$

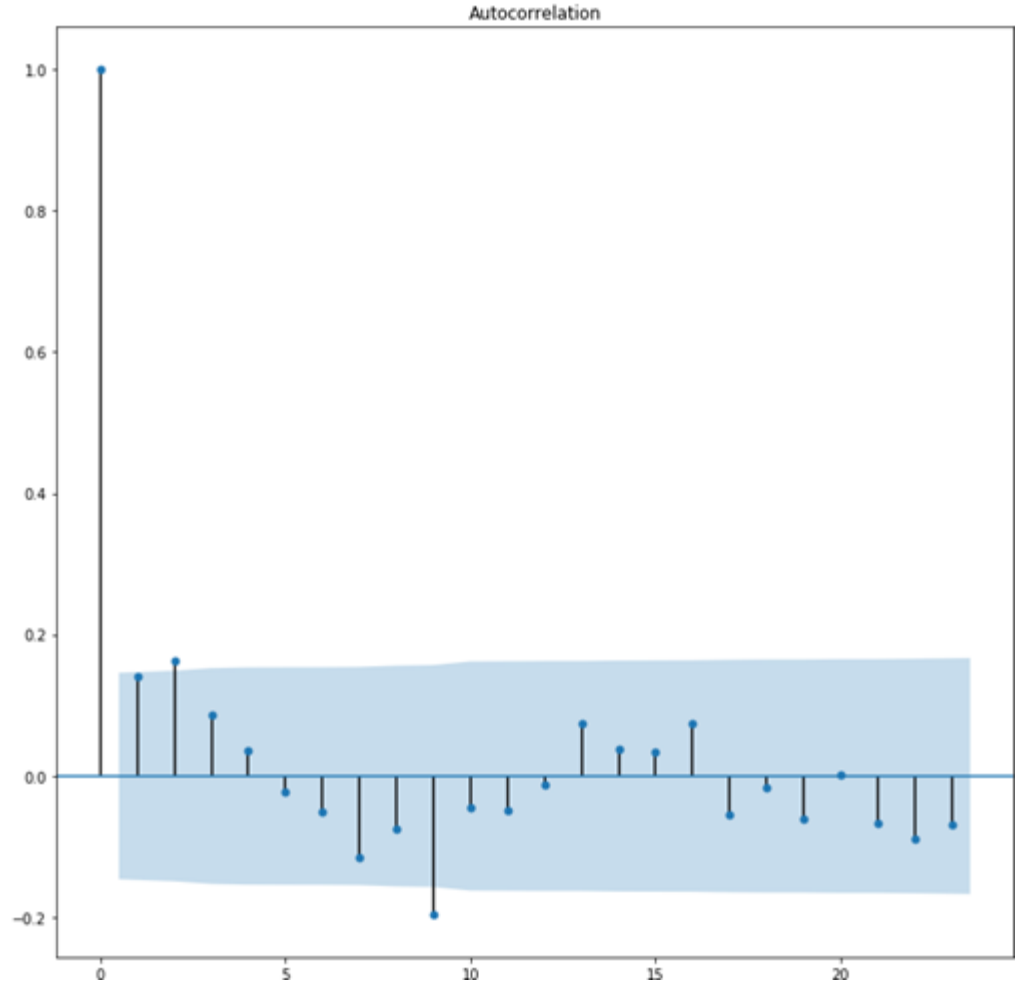
We fit the model with order (9, 1, 10), and obtained a MAPE value of 25.012%, as well as low p values for the feature weights, indicating a good fit – these were, however, more significant compared to the ARMA model. The residual ACF plot also had barely any significant values, justifying the white noise assumption of the model. The fact that this model did not perform better than the ARMA model confirms that the time series was stationary. However, a proper comparison would require use of the optimal orders as determined via the ACF and PACF plots earlier, which could not be done due to computational power limitations.

6.2.5 SARIMA Model

This model uses separate regression-like models after differencing the data at a lag equal to the seasonality of the data (approximated as yearly for our case). It can be represented using (p, d, q) – ARIMA order, (P, D, Q) – seasonal order, m – seasonality. Model expression:

$$\phi(B_m)\, \phi(B)\, (1 - B_m)^D\, (1 - B)^d\, X_t = \theta(B_m)\, \theta(B)\, w_t$$

```
Mean Absolute Percentage Error (Rajasthan): 10.167677551012869 %
```



We fit the model on monthly data, and not daily/weekly as the seasonality of 365/52 was too computationally expensive for grid search. Later we would do a weekly analysis with a more simplified pre-determined SARIMA model. For monthly data, we found the parameters as ([2, 0, 3], [2, 0, 1], 12). Fitting this model gave us extremely good results, with a MAPE of only 10.167%, and good results for both the residuals and residual ACF plot, which is shown.

7 Forecast Validation

To validate our choice of models for out-of-sample forecasting, we performed in-sample forecasting by splitting the data into training and test sets, and doing a rolling forecast on the test data. The splits were determined keeping computation times in check, and were as follows:

Train : Test Splits	AR, MA, ARMA, ARIMA	SARIMA
Daily data	90:10	N/A
Weekly data	90:10	95:5
Monthly data	N/A	70:30

The parameters for each model were also searched again in a smaller search space due to even the above models (which are already simpler than what is required for accurate forecasting) being too computationally expensive for a rolling forecast. MAPE and MSE (Mean Squared Error) were the two metrics used for evaluating the models. The results of the forecast for each model on the Wind Speed data of Rajasthan are as follows.

7.1.1 AR Model

Parameters: 9 (daily), 10 (weekly)
MAPE = 22.728% (daily), 20.298% (weekly)
MSE = 0.516 (daily), 0.479 (weekly)

```
Mean Absolute Percentage Error (Rajasthan): 22.72879148073194 %
Mean Square Error (Rajasthan): 0.516221049117893

<matplotlib.legend.Legend at 0x2208dc49198>
```

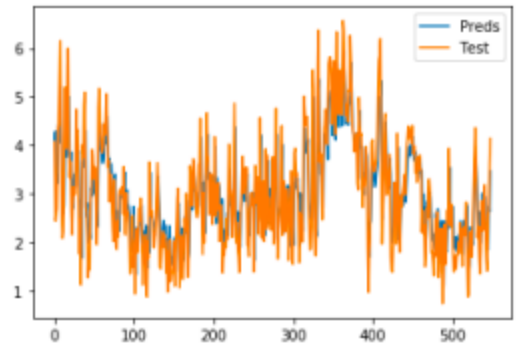


Figure: Weekly forecasting using AR model for Rajasthan.

```
Mean Absolute Percentage Error (Rajasthan): 20.298803690700883 %
Mean Square Error (Rajasthan): 0.4789770582422554

<matplotlib.legend.Legend at 0x22786563ba8>
```

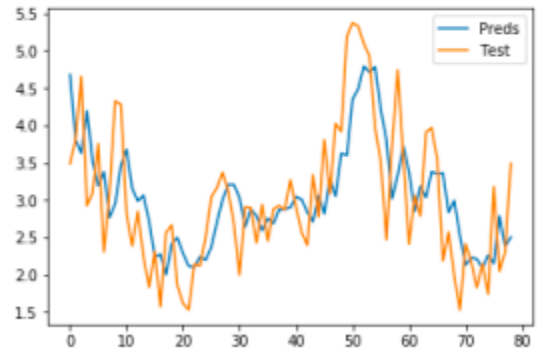


Figure: Monthly forecasting using AR model for Rajasthan.

7.1.2 MA Model

Parameters: 11 (daily), 11 (weekly)
MAPE = 23.424% (daily), 20.198% (weekly)
MSE = 0.545 (daily), 0.485 (weekly)

```
Mean Absolute Percentage Error (Rajasthan): 23.42486830037479 %
Mean Square Error (Rajasthan): 0.5450029090314916

<matplotlib.legend.Legend at 0x2208ddddd68>
```

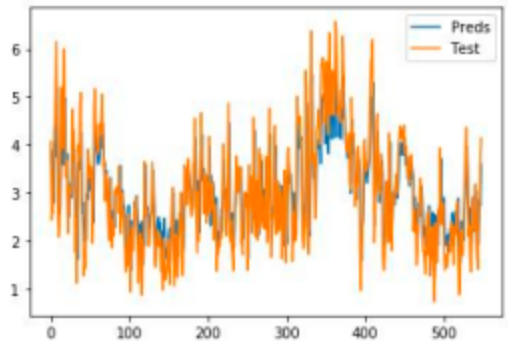


Figure: Weekly forecasting using MA model for Rajasthan.

```
Mean Absolute Percentage Error (Rajasthan): 20.19860488128218 %
Mean Square Error (Rajasthan): 0.4852124811023596

<matplotlib.legend.Legend at 0x227874c2be0>
```

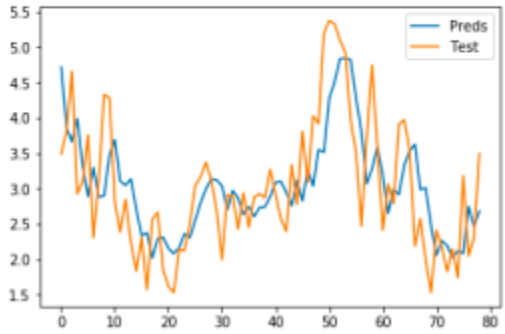


Figure: Monthly forecasting using MA model for Rajasthan.

7.1.3 ARMA Model

Parameters: 2, 3 (daily), 2, 4 (weekly)
MAPE = 22.874% (daily), 19.495% (weekly)
MSE = 0.519 (daily), 0.472 (weekly)

```
Mean Absolute Percentage Error (Rajasthan): 22.87413662095137 %
Mean Square Error (Rajasthan): 0.5191950322659858

<matplotlib.legend.Legend at 0x22092303e80>
```

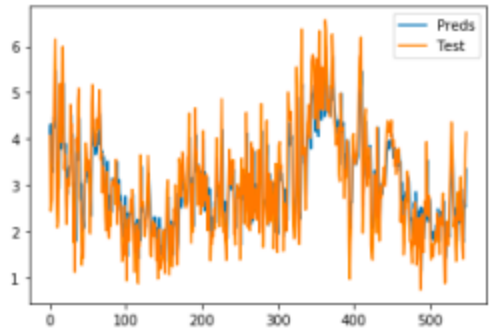


Figure: Weekly forecasting using ARMA model for Rajasthan.

```
Mean Absolute Percentage Error (Rajasthan): 19.495407904757872 %
Mean Square Error (Rajasthan): 0.4719975909768209

<matplotlib.legend.Legend at 0x227800d6d68>
```

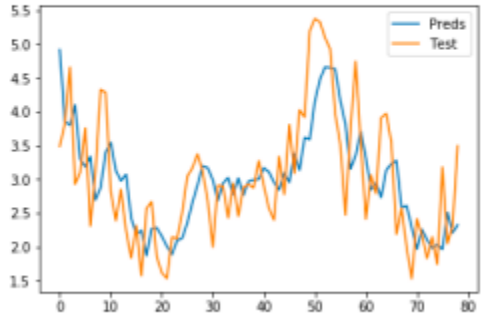


Figure: Monthly forecasting using ARMA model for Rajasthan.

7.1.4 ARIMA Model

Parameters: 2, 2, 3 (daily), 2, 2, 4 (weekly)
MAPE = 23.551% (daily), 20.749% (weekly)
MSE = 0.603 (daily), 0.505 (weekly)

```
Mean Absolute Percentage Error (Rajasthan): 23.551036184300177 %
Mean Square Error (Rajasthan): 0.6031709469152163

<matplotlib.legend.Legend at 0x2208da04320>
```

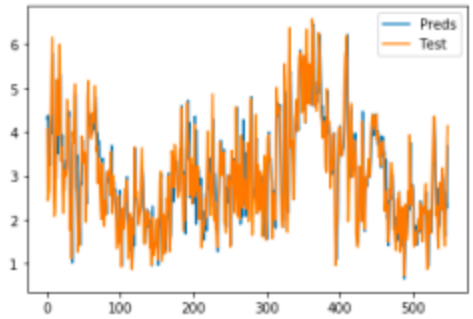


Figure: Weekly forecasting using ARIMA model for Rajasthan.

```
Mean Absolute Percentage Error (Rajasthan): 20.749269432561693 %
Mean Square Error (Rajasthan): 0.505006280939738

<matplotlib.legend.Legend at 0x227817218d0>
```

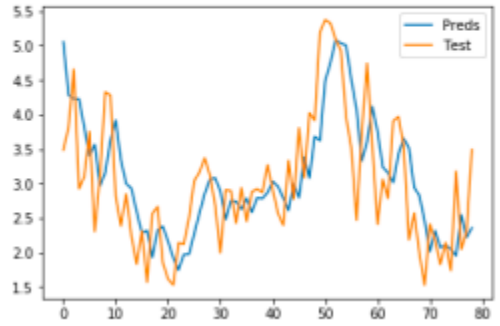


Figure: Monthly forecasting using ARIMA model for Rajasthan.

7.1.5 SARIMA Model

Parameters: ([1, 0, 1], [1, 0, 1], 52) (weekly – simplest model), ([2, 0, 3], [2, 0, 1], 12) (monthly)
MAPE = 18.958% (weekly), 10.714% (monthly)
MSE = 0.577 (weekly), 0.175 (monthly)

```
Mean Absolute Percentage Error (Rajasthan): 18.958795025742248 %
Mean Square Error (Rajasthan): 0.577061044379608

<matplotlib.legend.Legend at 0x227a0a02e10>
```

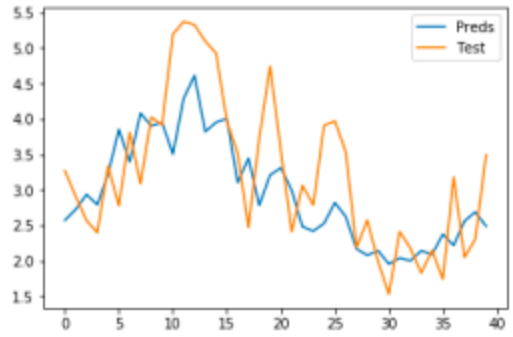


Figure: Weekly forecasting using SARIMA model for Rajasthan.

```
Mean Absolute Percentage Error (Rajasthan): 10.714485781394401 %
Mean Square Error (Rajasthan): 0.17517053075415187

<matplotlib.legend.Legend at 0x2208dc28e10>
```

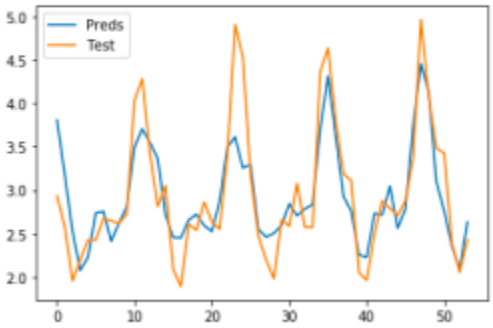


Figure: Monthly forecasting using SARIMA model for Rajasthan.

7.2 Conclusion

After performing similar analysis for all the states, the following general conclusions about the models were drawn:

Daily Data	ARMA > ARIMA > AR > MA (SARIMA not used)
Weekly Data	SARIMA > ARMA > ARIMA > AR > MA
Monthly Data	SARIMA was the best model of all.

The best model accuracies achieved during the tests for each of the states is as given in the table

below (based on MAPE scores – a similar trend was seen using MSE, however).

Table: Model accuracies as measured using MAPE.

State	Daily accuracy (%)	Weekly accuracy (%)	Monthly accuracy (%)
Madhya Pradesh	76.41	84.62	89.79
Rajasthan	77.27	81.05	89.29
Tamil Nadu	83.64	76.06	84.70
Andhra Pradesh	84.02	80.89	91.77

The inaccuracy of the fitted models reflects the complexity of the distribution the data is drawn from, which requires more complex models than the ones we were able to fit using the limited computation power of the systems used, which necessitated searching for parameters in much smaller search spaces. The discrepancy in the weekly data fits for Andhra Pradesh and Tamil Nadu (less accurate than the daily data) can also be explained by that – as the low order model fits may not be able to capture the variation in the possibly more irregular weekly data.

It is also worth noting that none of the models were actually great fits despite the accuracy calculated, as the confidence in all the calculated models gave at least a few significant p-values (although not many) for their estimated feature weights. However, they provide reasonably good accuracies for the simplicity of the models in terms of computational cost.

References

[1] D.R. Anderson et al. Statistics for Business & Economics. Cengage Learning, 2013.

[2] N. S. Khadka, "COP26: Did India betray vulnerable nations?," *BBC News*, 16-Nov-2021. [Online]. Available: https://www.bbc.com/news/world-asia-india-59286790. [Accessed: 06-Dec-2021].

[3] "Solar irradiance," *Wikipedia*, 25-Nov-2021. [Online]. Available: https://en.wikipedia.org/wiki/Solar_irradiance. [Accessed: 06-Dec-2021].

[4] N. O. A. A. US Department of Commerce, "Dew Point vs humidity," *National Weather Service*, 26-Jan-2021. [Online]. Available: https://www.weather.gov/arx/why_dewpoint_vs_humidity#:~:text=The%20dew%20point%20is%20the,water%20in%20the%20gas%20form.&text=The%20higher%20the%20dew%20point,of%20moisture%20in%20the%20air. [Accessed: 06-Dec-2021].

[5] Ghasemi A, Zahediasl S. Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. Int J Endocrinol Metab. 2012;10(2):486-9. DOI: 10.5812/ijem.3505.

[6] Skipper Seabold and Josef Perktold. "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[7] Ralph B. D'Agostino and Albert Belanger. "A Suggestion for Using Powerful and Informative Tests of Normality". In: The American Statistician 44.4 (1990), pp. 316 – 321.

[8] Wind Energy, Ministry of New and Renewable Energy, Govt. of India. url: https://www.mnre.gov.in/wind/current-status/

[9] Glossary, Wind Energy THE FACTS. url: https://www.wind-energy-the-facts.org/glossary.html

Appendix I:
Correlation Matrices

Madhya Pradesh

	DHI	DNI	GHI	Clearsky DHI	Clearsky DNI	Clearsky GHI	Dew Point	Tempera-ture	Pressure	Relative Humidity	Solar Zenith	Wind Speed
DHI	1											
DNI	0.72247	1										
GHI	0.87591	0.94145	1									
Clearsky DHI	0.95771	0.80972	0.91351	1								
Clearsky DNI	0.9429	0.84551	0.92881	0.97353	1							
Clearsky GHI	0.94406	0.8245	0.95113	0.96358	0.97322	1						
Dew Point	0.06503	-0.1484	-0.0523	0.0235	-0.012	0.02015	1					
Tempera-ture	0.51178	0.45748	0.54621	0.54121	0.50446	0.55247	0.13032	1				
Pressure	-0.1096	0.06728	-0.0428	-0.0847	-0.0366	-0.1017	-0.5055	-0.6452	1			
Relative Humidity	-0.2727	-0.4053	-0.3725	-0.3269	-0.3284	-0.3209	0.76762	-0.4684	-0.0695	1		
Zenith Angle	-0.8766	-0.73	-0.8378	-0.9052	-0.8975	-0.8901	-0.0824	-0.5785	0.17904	0.29624	1	
Wind Speed	0.04228	-0.0599	0.00651	0.03146	0.00577	0.04981	0.09239	0.33475	-0.4482	-0.1082	-0.0946	1

Tamil Nadu

	DHI	DNI	GHI	Clearsky DHI	Clearsky DNI	Clearsky GHI	Dew Point	Temperat ure	Pressure	Relative Humidity	Zenith Angle	Wind Speed
DHI	1											
DNI	0.6552095	1										
GHI	0.8350332	0.9383768	1									
Clearsky DHI	0.9337077	0.8223202	0.9207794	1								
Clearsky DNI	0.9334014	0.8330021	0.9285848	0.9891293	1							
Clearsky GHI	0.9251993	0.8330065	0.9610139	0.9661461	0.9738674	1						
Dew Point	-0.06843	-0.155777	-0.123863	-0.106958	-0.098211	-0.093805	1					
Temperat ure	0.678001	0.722965	0.7780884	0.7638547	0.7530499	0.7690548	-0.225046	1				
Pressure	-0.083877	-0.106867	-0.128755	-0.104456	-0.111062	-0.130382	-0.25958	-0.26091	1			
Relative Humidity	-0.560204	-0.619173	-0.645207	-0.635174	-0.62743	-0.629455	0.7211749	-0.801415	-0.008689	1		
Zenith Angle	-0.855594	-0.758494	-0.855478	-0.913354	-0.913477	-0.895978	0.0513414	-0.724672	0.1463416	0.5760856	1	
Wind Speed	0.0795251	0.0565085	0.0792783	0.0449804	0.0865455	0.0910865	0.141935	0.0206436	-0.380715	0.0415399	-0.105618	1

Rajasthan

	DHI	DNI	GHI	Clearsky DHI	Clearsky DNI	Clearsky GHI	Dew Point	Temperature	Pressure	Relative Humidity	Zenith Angle	Wind Speed
DHI	1											
DNI	0.8233563	1										
GHI	0.9288724	0.9445052	1									
Clearsky DHI	0.9816405	0.8558076	0.9448347	1								
Clearsky DNI	0.9122828	0.9476131	0.940262	0.9163743	1							
Clearsky GHI	0.9568675	0.9068663	0.9852615	0.9654257	0.9530105	1						
Dew Point	0.147906	-0.044296	0.080136	0.1505118	-0.018322	0.097627	1					
Temperature	0.619133	0.4681534	0.5912641	0.64237	0.4881377	0.5981509	0.6189814	1				
Pressure	-0.159938	0.0688707	-0.077545	-0.170571	0.0370746	-0.100113	-0.824563	-0.654188	1			
Relative Humidity	-0.216009	-0.32859	-0.270091	-0.226857	-0.309766	-0.252896	0.8036683	0.0926681	-0.59201	1		
Zenith Angle	-0.890437	-0.807959	-0.869949	-0.904409	-0.869655	-0.888591	-0.190437	-0.627054	0.1901987	0.1598697	1	
Wind Speed	-0.150573	-0.237597	-0.17532	-0.143565	-0.252756	-0.176556	0.3634103	0.1189166	-0.448246	0.3434737	0.1519921	1

Appendix II:
Finding the Best-Fit Distribution: Plots

II.1 Histograms for Wind Speed Data of the 4 States

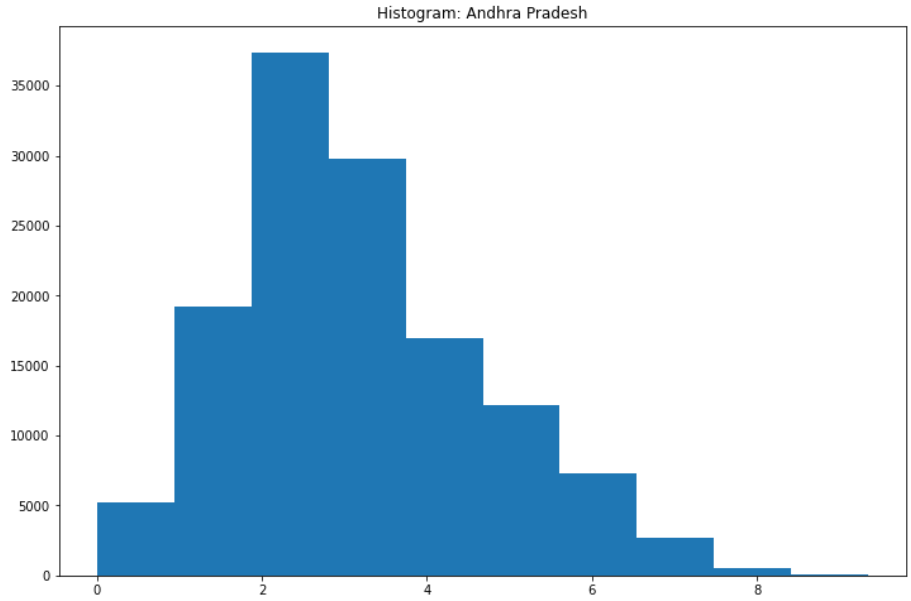


Figure I.1.1 Histogram of wind speeds for Andhra Pradesh

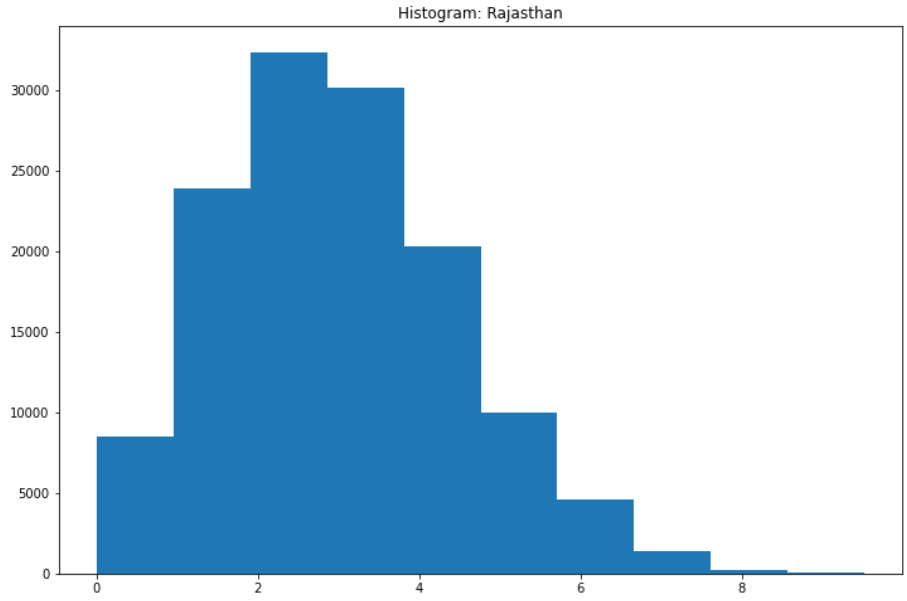


Figure I.1.2 Histogram of wind speeds for Rajasthan

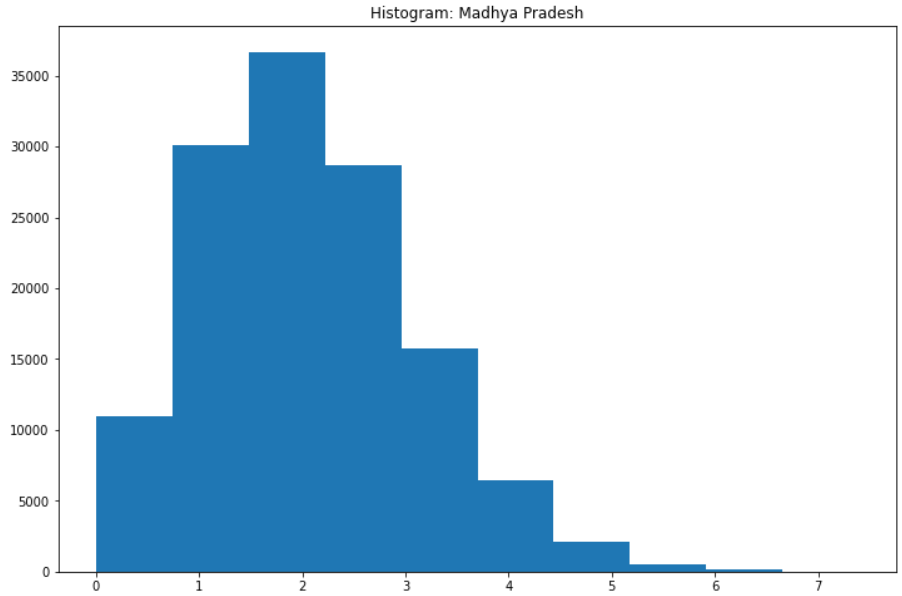


Figure I.1.3 Histogram of wind speeds for Madhya Pradesh

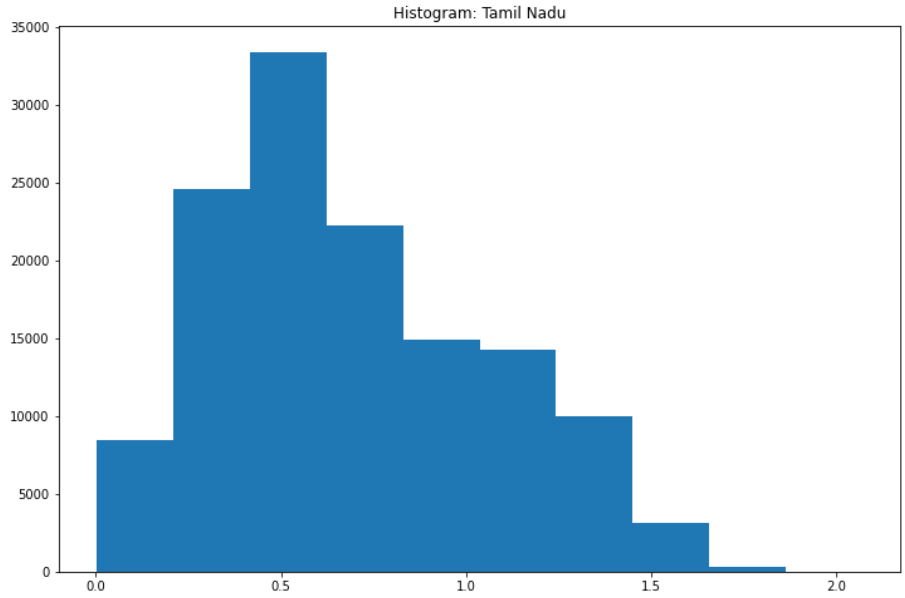


Figure I.1.4 Histogram of wind speeds for Tamil Nadu

II.2 Plotting Best-fit Distributions for Wind Speed Data of the 4 States

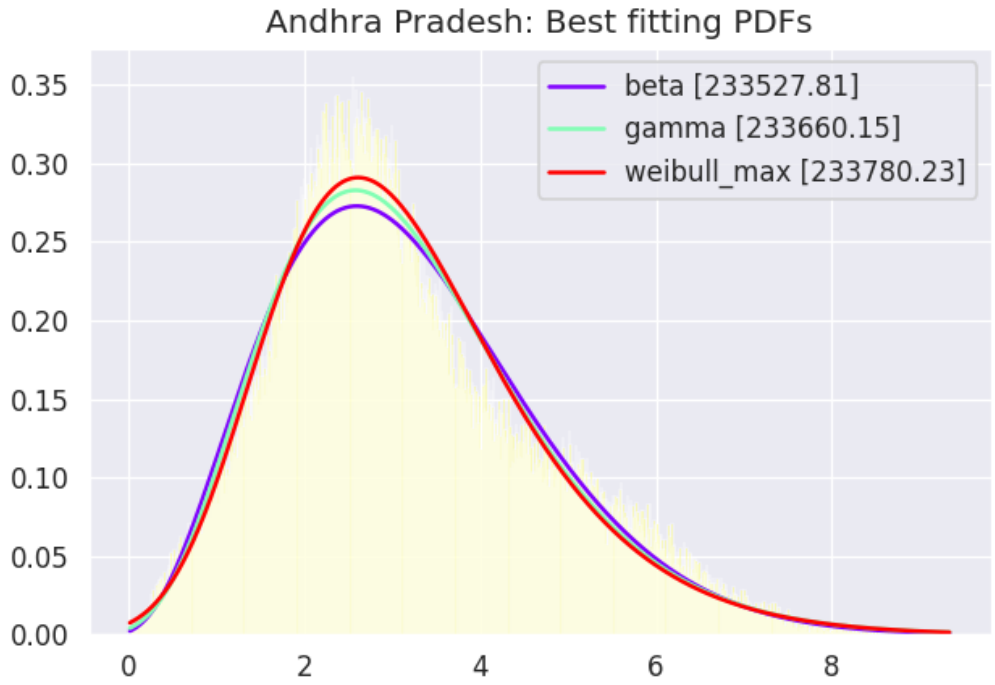


Figure I.2.1: Best-fit distributions for Andhra Pradesh.

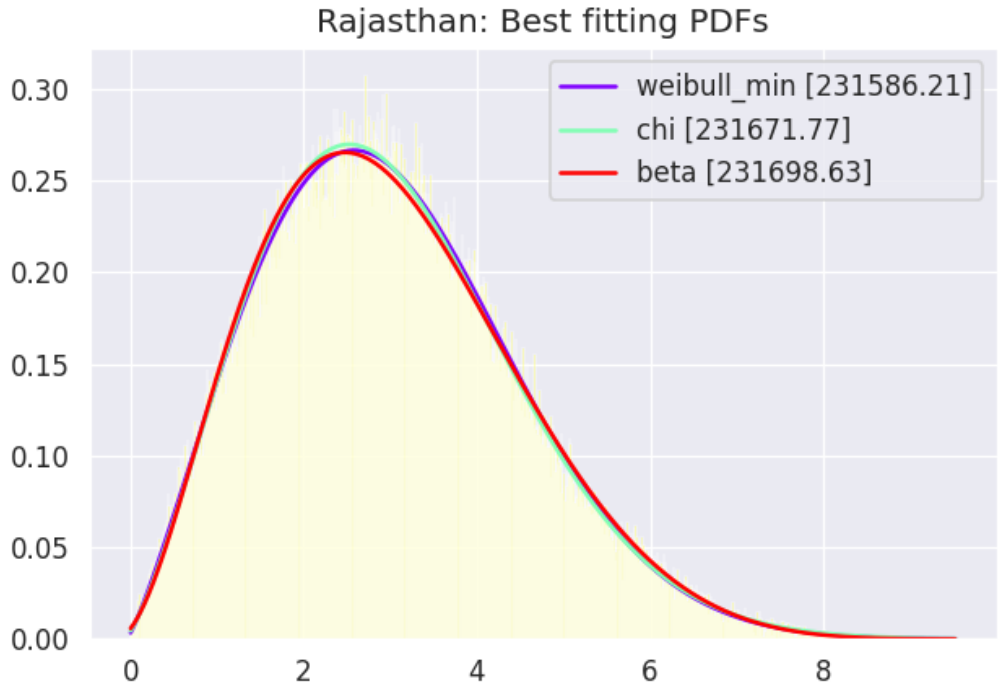


Figure I.2.2: Best-fit distributions for Rajasthan.

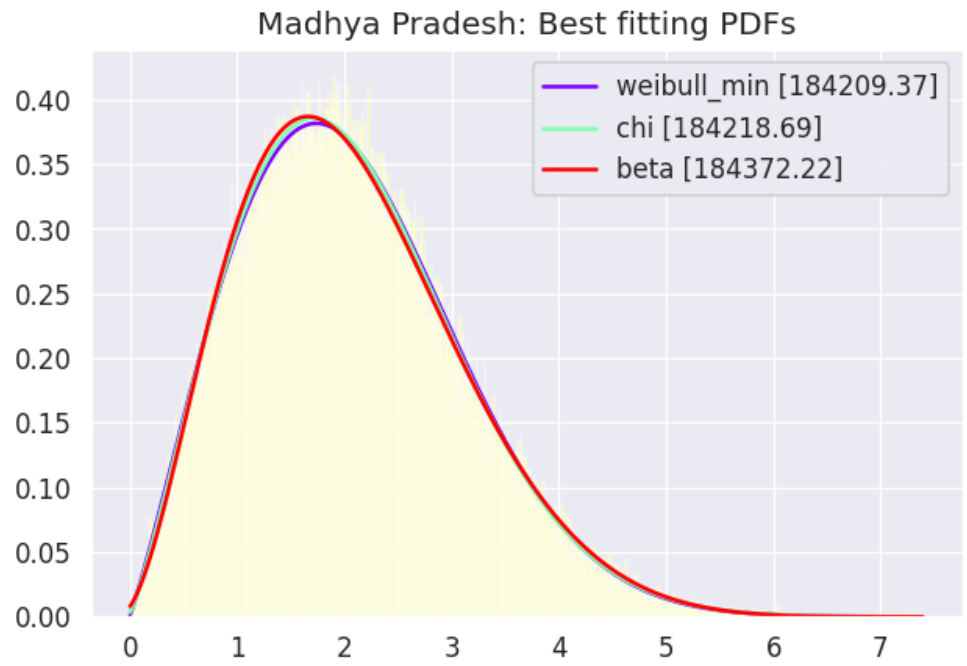


Figure I.2.3: Best-fit distributions for Madhya Pradesh.

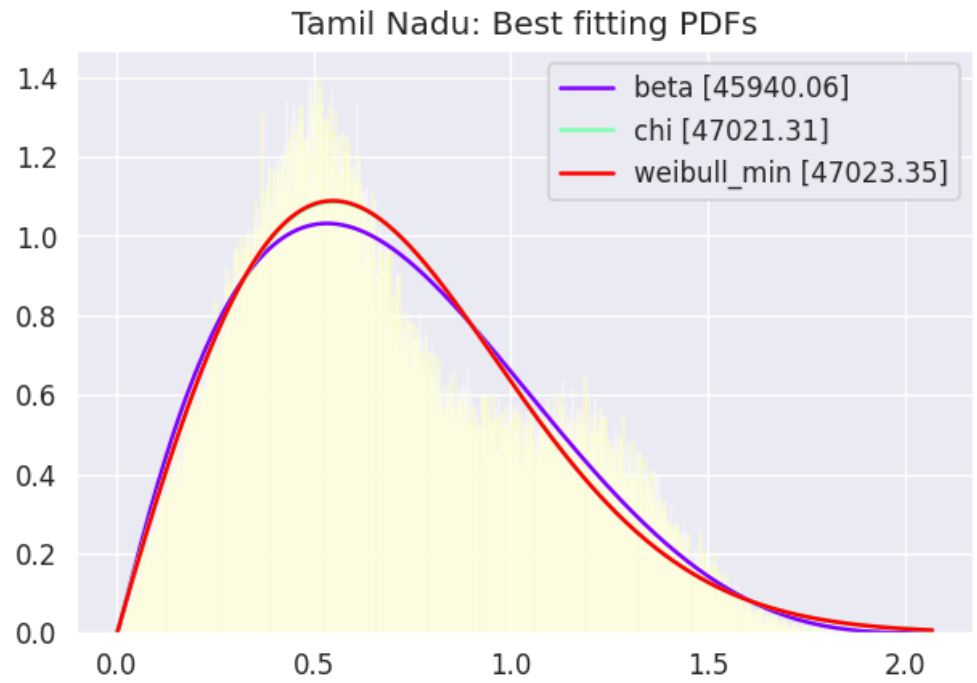


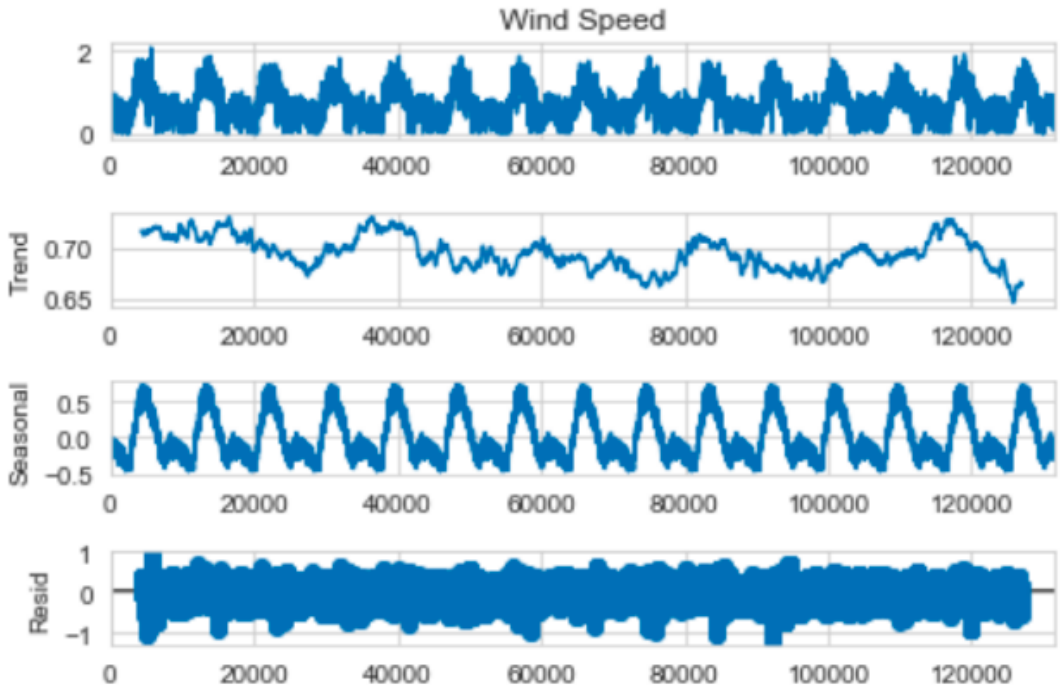
Figure I.2.4: Best-fit distributions for Tamil Nadu.

Appendix III:
Code Segment for Normality Testing

```
# setting the significance of normality tests
alpha = 0.05

# hypothesis testing
# H0: The Distribution is Normal
# Ha: The Distribution is Not Normal
def hypothesisTesting(pVal, alphaVal):
  if (pVal > alphaVal):
    print("The Distribution is Normal. Failed to reject H0.")
  else:
    print("The Distribution is Not Normal. H0 rejected.")
  print("\n")

# Shapiro-Wilk test
def shapiroWilkTest(dat):
  swStat, pvalue = st.shapiro(dat)
  print('\nShapiro-Wilk Statistic = %.3f; p = %.3f' % (swStat,
pvalue))
  hypothesisTesting(pvalue, alpha)

# D'Agnostino's K^2 test
def dagnostino(dat):
  daStat, pvalue = st.normaltest(dat)
  print('\nD\'Agnostino\'s K^2 Statistic = %.3f, p = %.3f' % (daStat,
pvalue))
  hypothesisTesting(pvalue, alpha)
```

Appendix IV:
Time Series Decomposition and Stationarity Testing

IV.1 Time Series Decomposition



IV.2 Stationarity Testing

Adapted Dickey Fuller Test

```
Test Stat: -13.419314697388598
p-value: 4.2204640597949785e-25
Crit value at 1% LOS: -3.4303997953780967
Crit value at 5% LOS: -2.8615620088348286
Crit value at 10% LOS: -2.5667817145225587


Test Stat: -21.875115048848524
p-value: 0.0
Crit value at 1% LOS: -3.4303997953780967
Crit value at 5% LOS: -2.8615620088348286
Crit value at 10% LOS: -2.5667817145225587


Test Stat: -22.487712452767557
p-value: 0.0
Crit value at 1% LOS: -3.4303997953780967
Crit value at 5% LOS: -2.8615620088348286
Crit value at 10% LOS: -2.5667817145225587


Test Stat: -11.656812627312476
p-value: 1.976389307362444e-21
Crit value at 1% LOS: -3.4303997953780967
Crit value at 5% LOS: -2.8615620088348286
Crit value at 10% LOS: -2.5667817145225587
```

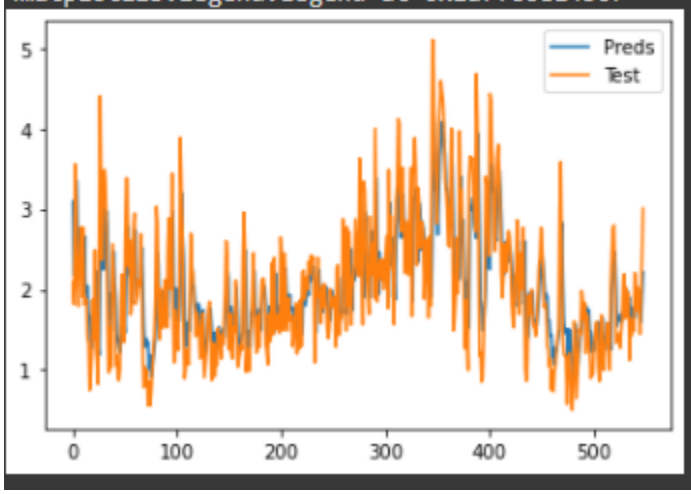
KPSS Non-parametric Test

```
Test Stat: 0.15758708105037433
p-value: 0.1
num lags: 73
Crit value at 1% LOS: 0.739
Crit value at 5% LOS: 0.463
Crit value at 10% LOS: 0.347


Test Stat: 0.3601461711453459
p-value: 0.09433354692010952
num lags: 73
Crit value at 1% LOS: 0.739
Crit value at 5% LOS: 0.463
Crit value at 10% LOS: 0.347


Test Stat: 1.9474099510337564
p-value: 0.01
num lags: 73
Crit value at 1% LOS: 0.739
Crit value at 5% LOS: 0.463
Crit value at 10% LOS: 0.347


Test Stat: 0.17282498421379994
p-value: 0.1
num lags: 73
Crit value at 1% LOS: 0.739
Crit value at 5% LOS: 0.463
Crit value at 10% LOS: 0.347
```

Appendix V:
Forecasting Plots for Madhya Pradesh

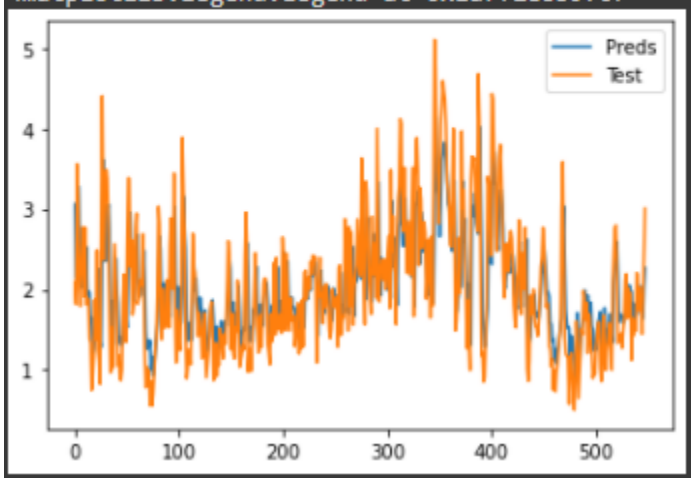
V.1 AR Model

```
Mean Absolute Percentage Error (Madhya Pradesh): 23.59244928648903 %
Mean Square Error (Madhya Pradesh): 0.3149077261661964
<matplotlib.legend.Legend at 0x1a7f86e1430>
```



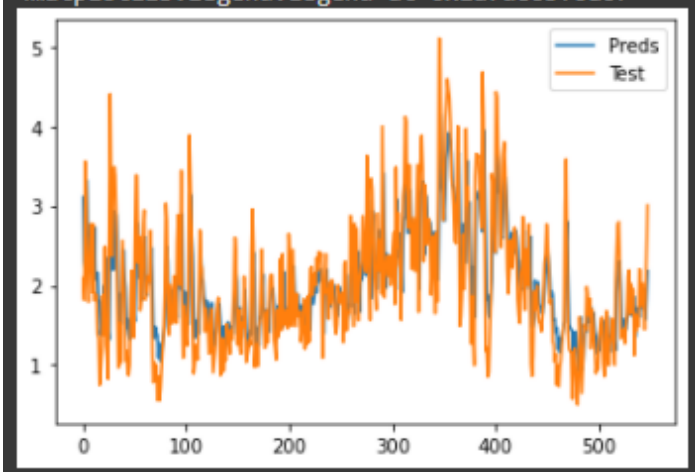
V.2 MA Model

```
Mean Absolute Percentage Error (Madhya Pradesh): 23.94610521223634 %
Mean Square Error (Madhya Pradesh): 0.31843242772896324
<matplotlib.legend.Legend at 0x1a7f2e83970>
```



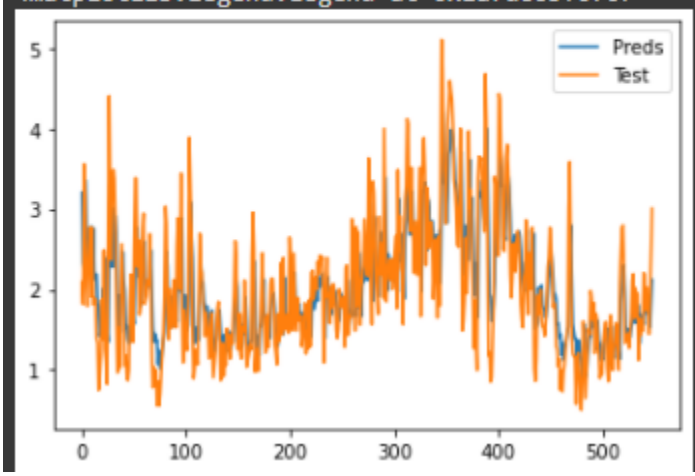
V.3 ARMA Model

```
Mean Absolute Percentage Error (Madhya Pradesh): 23.676007458349865 %
Mean Square Error (Madhya Pradesh): 0.3087118863199755
<matplotlib.legend.Legend at 0x1a7d863f6a0>
```



V.4 ARIMA Model

```
Mean Absolute Percentage Error (Madhya Pradesh): 23.633782936464396 %
Mean Square Error (Madhya Pradesh): 0.310796890772641
<matplotlib.legend.Legend at 0x1a7d863f070>
```



V.5 SARIMA Model

```
Mean Absolute Percentage Error (Madhya Pradesh): 10.210075617457461 %
Mean Square Error (Madhya Pradesh): 0.09625995321104693
<matplotlib.legend.Legend at 0x1a7f3328460>
```