**Tornado Forecast Using Mathematical Analysis**

**By**

**Abraham Olu**

**TTU.**

**TABLE OF CONTENT**

Contents

[1.0 INTRODUCTION 3](#_Toc26880968)

[1.1 DATASET 4](#_Toc26880969)

[2.0 IMPLEMENTATION 5](#_Toc26880970)

[2.1 Time-series Data With Time-dependent Statistic Features 5](#_Toc26880971)

[2.2 Checking Randomness Using Dickey-Fuller Method 8](#_Toc26880972)

[2.3 Autocorrelation and Correlogram of Dataset 8](#_Toc26880973)

[2.4 Differencing Method to Detrend the Time-series 10](#_Toc26880974)

[2.5 ARIMA Model Data Fit 11](#_Toc26880975)

[3.0 RESULT 13](#_Toc26880976)

[4.0 CONCLUSION 14](#_Toc26880977)

[REFERENCE 15](#_Toc26880978)

## 1.0 INTRODUCTION

Analysis of almost all physical occurrence or phenomenon has shown that it can be argued there are often predicates on which the phenomenon being observed is based[1](#_REFERENCE) and these, more often than not can be modelled mathematically[2, 3](#_REFERENCE) . For example, an occurrence, *y* is dependent on a number of independent variables, . So, event *y* can be ascertained or at least calculated almost accurately from the observations of variables belonging to set *.* Furthermore, the advent of computer-aided mathematical design is shown not only to help in modelling the physical occurrences but also in generating models of new phenomena[2](#_REFERENCE).

When considering a situation where some scientific or physical phenomenon is to be modelled e.g. the cloud formation and air movement that causes lightning strikes and thunderstorms, or the market indicators that point where a stock price is headed, but without the luxury of observing the dependencies or variables. In scenarios such as these, mathematical analysis of available data can help reveal the relationship function that can be inferred as a pattern[12](#_REFERENCE). For example, the simple analysis of an event by finding the event intercept on a graph of available data, yields the dependent variable, *y,*  from the line equation in (0.1). Importantly, however, a more focused mathematical analysis can reveal a hidden relationship in a complex event’s occurrence [4](#_REFERENCE) and then using a probabilistic approach, predict future occurrences of such event[5](#_REFERENCE).

)

In the same vein, we examine in this paper if tornado, a natural weather phenomenon which was previously considered, together with other natural weather events, to be unforecastable be forecasted probabilistically from mathematical analysis alone? We examine this for a number of reasons, for example, in the United States of America alone, the government budget for the National Weather Service(NWS) is around 1 billion USD mark. Interestingly, tornadoes and such other violent weather were considered unpredictable up until 1960, causing massive damage to property and even loss of lives. Until the first weather satellites were launched, bringing about the possibility of monitoring weather formation and predicting natural weather phenomenon, saving countless lives in the process.

In the USA, the state of Texas statistically has the highest frequency of tornadoes annually, singlehandedly averaging around 150 per year[6](#_REFERENCE). Numerical forecast of tornadoes or other severe weather patterns proposed in this paper can be considered in some ways quite nascent and academic corpus in this area comparatively sparse. Also, the focus of most of these researches have mainly been to use the ‘*rawinsonde method’* observation to determine the probability of severe weather conditions[7](#_REFERENCE). And the NWS’s best forecast tool, remains the tornado and severe weather thunderstorm watches; error sensitive sensors and stations for monitoring weather development. Inasmuch as there have been some incursions into numerical forecasts here and there, as reported in[8](#_REFERENCE), there is no active ongoing research in this area. The need for the use of numerical analysis based on curated data in order to forecast has been stated clearly in[11](#_REFERENCE); importantly it can help cut down the government budget of incessant monitoring, and yield comparative or better yet, more reliable predictions.

The intent of this paper, therefore, is to then examine the question as stated above; can mathematical analysis of curated dataset be used to forecast the number of tornadoes? And investigate the answer to this question. Equipped with the tornado dataset in Texas from 1950–2018, we intend to forecast the yearly number of tornadoes in Texas over the period, 2013-2031, comparing the predicted overlap years with actual data. Forecasts generally have uncertainties, and it is common practice to have an uncertainty or stability index alongside the probability value, therefore the prediction error in this body of work is similarly investigated. In the absence of predictive features in the NWS maintained tornado dataset, we attempt to use an Auto-Regressive model to generate a forward forecast of the target variable based solely on previous observations.

The hypothesis constructed for the available dataset to match in order to ascertain the plausibility of the paper objective and predict results that are mathematically acceptable at all, as reported in this paper include the following stipulations;

* I.I.D data; Investigate that the tornado data is composed of random events.
* Time variant; investigate that the analysis reveals a time series with changing mean and variance..
* Investigate the correlation between the data point in the series, including discrete white noise, partial correlation and sub-trends in the curated dataset.
* If stipulations above; detrend the time series and then investigate the residuals for sub-correlation. Ensure the time stationarity of the dataset.
* Fit data to a precise model of ARIMA, and make forward predictions, investigating accuracy and errors.

### 1.1 DATASET

The tornado dataset used here is a publicly available dataset obtained from and maintained by the National Weather Service[6](#_REFERENCE). NWS has tried to maintain a public database of every recorded tornado, hail and other extreme weather occurrences in every state in the USA from 1950 to 2018. Using eyewitness sighting, from storm chasers, populace or reliable source, and chiefly weather stations. We have carefully curated the dataset for tornadoes recorded in Texas from 1950-2018, the latest available record. Worth mentioning is the fact that the NWS reports that the national deployment of the NWS-WSR-88D weather monitoring units from 1990, made weather event recording much more accurate than in prior years.

The decision to however include available data from 1950 to the present date is simply down to mathematical analysis needs. From a mathematical stand point, all of the data allows for as much accurate sampling and hence modelling and hence prediction as possible.

## 2.0 IMPLEMENTATION

The Random Walk is a stochastic forecast technique of a succession of steps dependent on some interval, e.g. time, it started as far back as 1905 by Karl Pearson to model the probability or random distribution in a rain forest[9](#_REFERENCE) . Or distance from a starting position, k, after *n-1* walks[10](#_REFERENCE). For the first stipulation, the dataset is investigated to show if it is an independent identically distributed(i.i.d) set, that is, can be classified as a distribution with random probability and without memory or dependence. This stipulation is important in most mathematical analysis, an i.i.d distribution is easier to model as the process is more readily modelled probabilistically as compared to a non-random distribution. While we would like to establish a correlation, it is only serially, and the randomness of this data will show if it is memoryless independent variables. While randomness cannot be proved with certainty, figure 1 shows a graph of the dataset with a non-equiprobable distribution of the data. There is an upward trend, then a downward trend toward the end of the dataset.

A screenshot of a computer

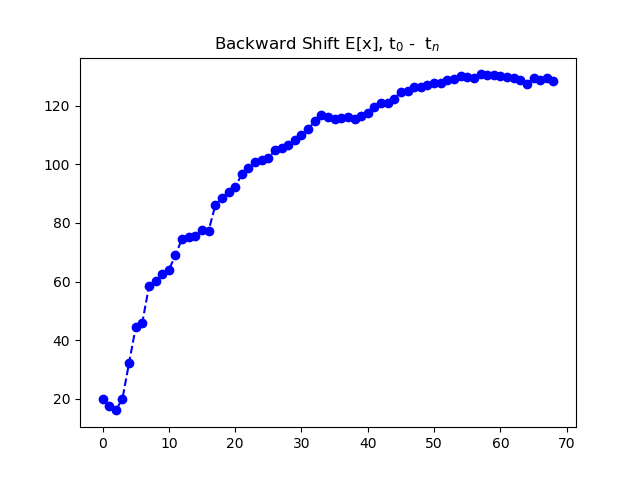
Description automatically generated

*Figure 1: Tornado Data, 1950-2018*

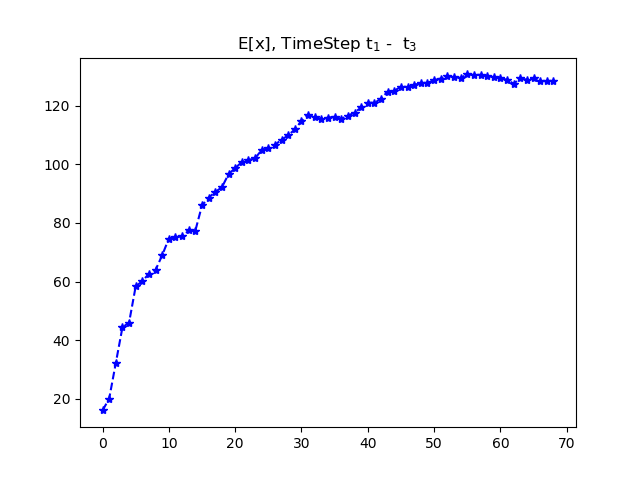
### 2.1 Time-series Data With Time-dependent Statistic Features

Determining the validity of the stipulation of a moving trend in the data. A number of moving point average calculations are computed, including the Backward shift moving average from as the starting point and single time-step increments to , as shown in figure 2. The other computed moving averages are 3, 4 and 6 moving time-step to analyze the mean and then also to find the expected value deviation, discrete white noise. In figures 3, 4 and 5 the graph of the moving averages over 3, 4 and 6 steps are shown. The mean is obvious to be non-stationary, a constantly increasing upward tick is observed in all the mean plots.

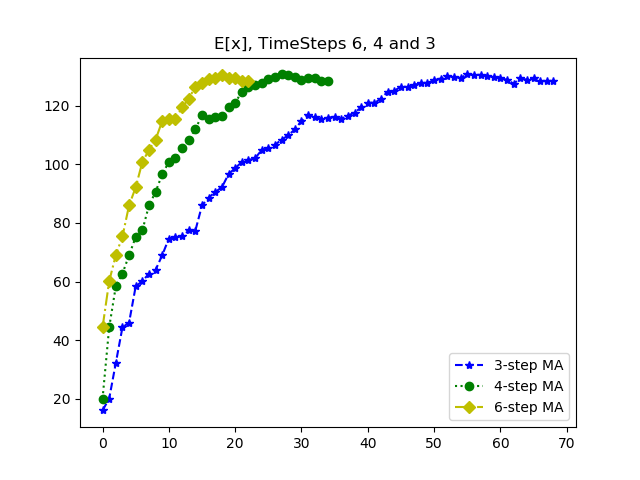
The initial observations, therefore, matches the randomness hypothesis. The expectation from equation (2) used in making the plot, shows that there is expected time dependency in the dataset. Therefore, the next stipulation that the mean is clearly a function of time infers the data is time-variant. Equation (1) shows the 3 time-step moving average function, computed by taking the first 3 indexed elements and calculating the mean, and increasing by steps of 1 and samples of 3 over the whole dataset. Similar computations for time step 4 and 6 were performed. The observation from all moving mean() calculations is similar and supports the stipulation of the time-series moving mean. The trend in the Backward Shift method establishes the relationship of mean() with time(t) across the data as shown in the figure 2 below. Figures 3-4 similarly show a plot with the obvious mean-trend with time relation. Graph plot with the mean on the ordinate and the number of samples taken on the abscissa.



*Figure 2: Trend of Mean in Tornado Dataset*



*Figure 3 Moving Average, 2 time-step*



*Figure 4 Moving Average, 3 time-step*

Additionally, in a separate computation to see a change in observations, the mean was computed with a random sampling of elements to generate a result that is unbiased. However, the same trend as with the graphs above was similarly observed.

### 2.2 Checking Randomness Using Dickey-Fuller Method

In this setup, in order to try and develop an Auto-Regression(AR) model for the dataset obtained. First, the statistical tool, the Augmented Dickey-Fuller(ADF) test with a null hypothesis that claims; “the time series is non-stationary, or solving the characteristic equation yields at least a unit root” is used to numerically ascertain randomness and non-stationarity. The equation of the ADF implementation is given in Equations (3 and 4), and the results of the Dickey-Fuller test as shown in table 1 confirms that the set is a non-stationary series, with the observation of a unit root. The Null test is formulated with a significance value of 5% and is accepted if the p-value is greater than the 5% critical value, or rejected if the p-value is lower than the critical value.

As observed in the results of Dickey-Fuller test and contrasted with the formulated Null hypothesis check, the p-value in the result is 0.872, which is greater than the 5% critical value of -3.01, calculated from solving for deltasubn from Equation (3). Similarly, the test statistic value of -0.5 is greater than the either of the 1%, 5% or 10% critical values, and this also confirms the assertion of the Null hypothesis and the original stipulations, and the null hypothesis is accepted. The Null hypothesis holds, but now moving on, we cautiously stipulate that the dataset is correlated. An Auto-Regression model of the first-order will confirm that the dataset has a simple linear relation with the previous datapoint. However, a quick model fit and the data couldn’t be modelled by a first-order AR.

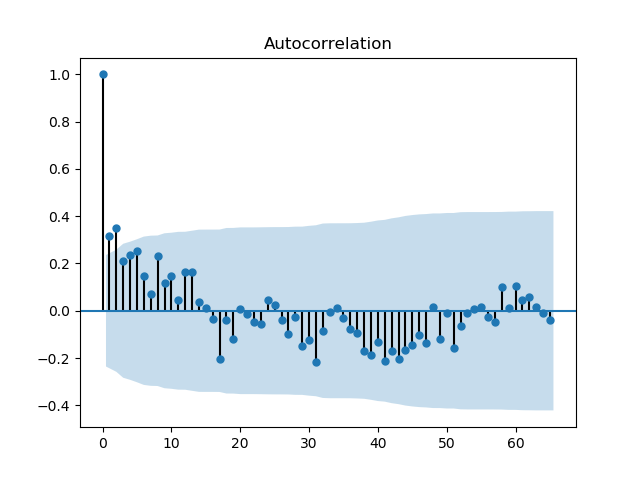
|  |  |
| --- | --- |
| ADF Test Statistic | -0.597193 |
| P-Value | 0.871615 |
| # Lags Used | 7.000000 |
| # Observations Used | 21.000000 |
| Critical Value (1%) | -3.788386 |
| Critical Value (5%) | -3.013098 |
| Critical Value (10%) | -2.646397 |

*Table 1: Python Augmented Dicky Fuller Test*

### 2.3 Autocorrelation and Correlogram of Dataset

The correlation function given in Equation (5) is used to find the serial-correlation of the dataset and the plot is shown in figure 6. Analyzing the confidence band fit of the lag model, and it is seen that other than point 1, we have subsequent lag points exceeding the 5% level at points equal 2 and 3. This is not unexpected for the confidence level selected, 95%, and as much as 4 lags are permissible to exceed the confidence bound in the data correlogram. This lag points can be explained away in the model as being caused by a number of reasons, such as the pdf or cmf of the distribution. Barring such explanation, however, we stipulate one, that there exists a serial correlation in the dataset if weak or strong, two, there is sub-trend in the dataset causing observed sub-lags .

Furthermore, as this observation could also be as a result of sub-correlation in the dataset, the residual or discrete white noise in the dataset is computed using the ‘Differencing method’ and a graph of the ‘Partial-Correlation’ is plotted using bias correction formula. As stipulated, the correlogram shows there is a sub-trend in the residual which is not explained away with the ACF. We point out here, that not this sub-trend is not further explored beyond necessary in this scope but rather simply, a more suitable model, the Auto-Regression Moving Average model(ARMA) is chosen to model the dataset. Another point was that there was an unexpected, impossible observation in the PACF plot, around the halfway point. We observed correlation points in the residual plot at lags 33 and 34 that was higher than +1 or -1, going as high as +2.0 and -1.5. The cause was discovered to be that residual data was first assumed to be biased, and the bias correction of the elements in the correlation calculation led to this error in the residual PACF plot.

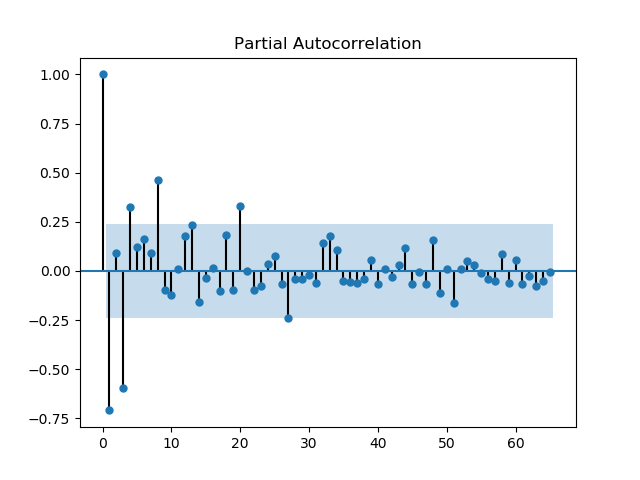


*Figure 5 Correlogram of Tornado Dataset*

### 2.4 Differencing Method to Detrend the Time-series

To remove the time dependence and make the data stationary we use the ‘Differencing method’ to detrend the datapoint. The stationarity is generated not only to analyze the statistic feature of the data without time dependence but also, so as to find the best fit for the data using the time-stationary data derivative. The Differencing technique is implemented by starting at a time, t and computing the datapoint at t as the difference between the datapoint at t and the previous datapoint at t-1, as can be seen in Equation (7), this is called the residual or white noise in the dataset.

As noted earlier, an assumption that the detrended dataset was biased as been shown to be false. Therefore, we concluded the residual data is indeed unbiased, and observation of the partial correlation plot of the detrended dataset without taking bias correction indicates a strong correlation between series element of the residual dataset. There are lags observable at points between 1 – 20, but the lag falls off smoothly afterwards, supporting this conclusion hypothesis. However, such conclusions are yet still cautiously avoided at this time for 2 main reasons; one, we intend to use a mathematical model that would take into consideration the correlation left in the differenced data. Two, the sub-correlation in the dataset is not investigated further and neither is the earlier stipulation of a weak correlation between the datapoints disproved. Hence, the stipulation of the weak correlation is taken to lend itself to the overall methodology in this paper and is therefore accepted, and the data concluded to be serially correlated with a max-lag taken to be 20.



*Figure 6 PACF of 1st order Differenced Dataset*

### 2.5 ARIMA Model Data Fit

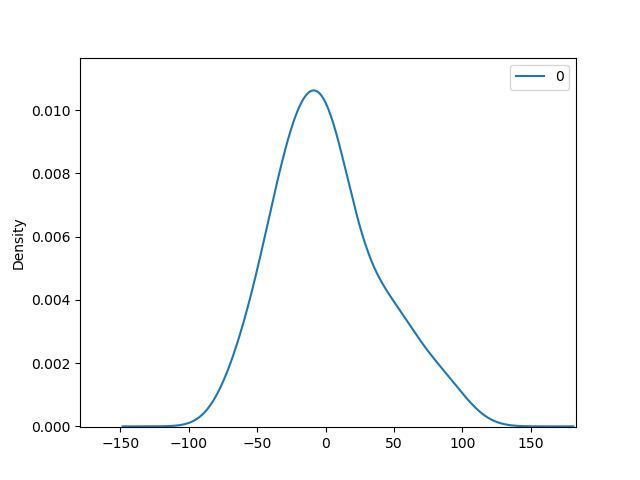
The decision of what regression model to use in fitting the tornado dataset was based on a number of considerations; first, the observed lags in the correlogram, the differencing order, and the moving average window. A number of models were tried to get the best fit, minimum error and near accurate predictions with the test samples. Two of these models are reported here; ARMA order 1 model, which is the less accurate of the two, and the more precise Auto-Regression Integration Moving Average (ARIMA) , reported next.

After the residual of the dataset was found with a differencing order of one, the observation of the residual graph plot shows a model that is not stationary, as already observed in the PACF plot in figure 6. A distribution plot of the residual is implemented, and simple observation shows it is a Gaussian distribution with a minimally biased mean() of approx. 0.9 as shown in figure 8 below. The ARIMA model was fit with a lag order of 20, chosen from the lag observed at k=20 in the PACF plot, a differencing order of 1 which we used in calculating the residuals, and a moving average window of zero.

A screenshot of a cell phone

Description automatically generated

*Figure 7 Residual Data Plot*



*Figure 8 Residual Data Density Distribution*

The model is trained with approx. 91% or 63 points in the data and tested to forecast the remaining 6 points in the dataset. Afterwards, the trained model is made to predict the data for an additional 13 steps. A condensed summary of the algorithm model fit with the data includes; Log Likelihood error of -343.8909, Akaike Corr. Number of 731.619, and standard deviation as observed in figure 8. Importantly, however, the root mean square error of every other model tested was much greater than the tolerance which we have arbitrarily set at 15. For example, the ARMA model discussed later has a mean square error of 4564.9 and root mean square error of 67, the ARIMA model discussed here, however, has the lowest root mean square error we have developed till date, approximately 35. The outcome of the model forecast compared to the test samples, and the 13 forward step predictions are shown in the result section.

## 

## 3.0 RESULT

In this paper, an Auto-Regression method not dissimilar to Random Walk forecast is used to implement future step predictions of a stipulated random occurrence, tornadoes in Texas. It is pertinent to point out that the first 6 points predicted by the model is from 2012-2018 which is compared to the available dataset value to test for accuracy. The mean square error calculated for the model’s prediction is approximately 1230. Afterwards, the remaining 13 time-points, are out of step or forward step annual tornado predictions from the year 2019 to 2031. The results from most other model fit shown in this work are done so here, simply here for comparison and are less accurate than the ARIMA model chosen. The ARIMA model gives us the least mean square error and also the closest observations to the available 2013-2018 data. However, this model is still being fine-tuned for this prediction for best accuracy without overfitting the model, the plot in figure 9 shows the comparison of the expected and predicted numbers.

*Figure 9 Expected vs Prediction Plot*A screenshot of a map

Description automatically generated

In conclusion, the NWS provided data for Tornado in 2019 approximately was 170 confirmed occurrences. However, the ARIMA forecast taken in November, 2019 was approximately 130(131.5).

## 

## 4.0 CONCLUSION

It has been shown in this paper that the current expensive method of weather observation for forecast might have a more analytical and less expensive alternative. This technique as shown in the result displays a significant margin of error from actual expected values in the provided data. Albeit, with a more fitting to find the best possible prediction range, for as high an accuracy score as possible, this technique can replace preexisting ones.

Some of the thought processes and considerations on the methodology includes, as an example, consideration of the dataset as a Bayesian model, where it is assumed the available data is an exact model for the future data forecast and the challenge is to then try and calculate the statistic features of the model. However, what was observed was a replicate forecast, in which there was no observable trend in the time series. We can’t claim the data actually models a forward step forecast if it is based only on a Bayesian probability. Therefore, we chose to think of the problem as a non-Bayesian challenge, where we try to ascertain present statistic feature in the data, then generalize the model from the forward steps or forecast in the time series. This model made more sense, as we do not assume what the distribution of future samples would be, and only tried to find the best model fit.

Further research extension formulation with a little thought experiment and using the methodology of seasonal data variation shows that a step can be taken downwards in the time series rather than forward. That is, we can implement monthly tornado data forecast based on the monthly data of tornado recorded over the years. This hypothesis formulated on the basis that analysis of a particular month over a large number of samples can reveal a statistic that can be used in forecasting data for future months. It is our belief that this can drastically increase the time warning for weather occurrence and save a huge amount on weather monitoring as outlined in this paper.

## REFERENCE

1. Baker, A. (2005). Are there genuine mathematical explanations of physical phenomena?. *Mind*, *114*(454), 223-238.
2. Nemirovsky, R., Kaput, J., & Roschelle, J. (1998). Enlarging mathematical activity from modeling phenomena to generating phenomena. In *Proc. of 22nd Annual Meeting of the Int. Group for the Psychology of Mathematics*.
3. Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1984), 20110553.
4. Devlin, K. (1996). *Mathematics: The science of patterns: The search for order in life, mind and the universe*. Macmillan.
5. Steen, L. A. (1988). The science of patterns. *Science*, *240*(4852), 611-616.
6. NOAA, National Weather Service Tornado [Dataset](https://www.spc.noaa.gov/wcm/)
7. Doswell III, C. A., Weiss, S. J., & Johns, R. H. (1993). Tornado forecasting: A review. *The Tornado: Its Structure, Dynamics, Prediction, and Hazards, Geophys. Monogr*, *79*, 557-571.
8. Hamill, T. M., & Church, A. T. (2000). Conditional probabilities of significant tornadoes from RUC-2 forecasts. *Weather and forecasting*, *15*(4), 461-475.
9. Hughes, B. D. (1996). Random walks and random environments.
10. Carazza, B. (1977). The history of the random-walk problem: considerations on the interdisciplinarity in modern physics. *La Rivista del Nuovo Cimento (1971-1977)*, *7*(3), 419-427.
11. Stensrud, D. J., Xue, M., Wicker, L. J., Kelleher, K. E., Foster, M. P., Schaefer, J. T., ... & Tuell, J. P. (2009). Convective-scale warn-on-forecast system: A vision for 2020. *Bulletin of the American Meteorological Society*, *90*(10), 1487-1500.
12. Ribeiro, R.C.M., Quadros, T.A., Saldarriaga, J.J., Júnior, S., de Almeida, J.F. and Marques, G.T., Forecasting incidence of tuberculosis cases in Brazil based on various univariate time-series models.
13. Tornado Forecast in Texas Using Mathematical Analysis - Random Walk [Github Repository](https://github.com/EmagIm/TexasTornado-StochasticForecast)