Short-term Prediction of Lightning in Southern Africa using Autoregressive Machine Learning Techniques

Yaseen Essa
School of Computer Science and Applied Mathematics,
The University of the Witwaterstand,
Johannesburg, South Africa
YaseenEssa@essamail.co.za

Hugh G.P. Hunt
The Johannesburg Lightning Research Laboratory,
School of Electrical and Information Engineering,
The University of the Witwaterstand,
Johannesburg, South Africa
Hugh.Hunt@wits.ac.za

Ritesh Ajoodha
School of Computer Science and Applied Mathematics,
The University of the Witwaterstand,
Johannesburg, South Africa
Ritesh.Ajoodha@wits.ac.za

***Abstract*—Lightning is responsible for both human and economic loss but its prediction remains challenging. We seek to find a lightning prediction model in South Africa that uses historical lightning-flash data only. This type of prediction model is cost-effective, easy to interpret and may be used for real-time forecasting. We evaluated and compared three popular time-series machine learning techniques on their ability to predict the number of Cloud-to-ground lightning flashes in South Africa for three-hours ahead. These models are the Auto Regressive (AR), Auto Regressive Integrated Moving Average (ARIMA) and the Long-Short-Term-Memory Recurrent Neural Network (LSTM) models. We used historical lightning data from the South African Lightning Detection Network during 2018. Our prediction model parameters were AR(lag=8), ARIMA (AR lag=8, integrate=0, MA lag=2) and LSTM (2x50 layers, activation=ReLU, optimizer=adam) and models were minimized for Root Mean Square Error but evaluated based on Mean Absolute Percentage Error (MAPE). We used a 70%/30% Train-test split. The AR and ARIMA models performed comparably with a MAPE of 15312 and 15080 respectively. The LSTM Model outperformed considerably with a MAPE of 3705. Although the LSTM model outperformed, predictions errors in absolute terms were still high. This paper highlights the usefulness of non-parametric predictions models for lightning prediction.**

***Index Terms*—lightning forecast, univariate, Long-Short-Term-Memory Recurrent Neural Network, weather forecasting, ARIMA, autoregressive**

I. Introduction

Lightning is responsible for both human and economic loss. It is estimated that 264-people on average die from lighting strikes in South Africa every year [4]. Further, lightning is responsible for about twenty-percent of all outages of electrical distribution in South Africa [5]. Knowing when lightning will occur, will help reduce human loss and assist in planning for expected lightning damage.

Lightning is an electrostatic discharge that results in a spectacular display of electromagnetic radiation and a pressure-wave called thunder [1]. There are between 30-to-100 lightning strokes every second on earth [2]. Seventy-five-percent of these lightning strikes originates and ends in the clouds; this is termed cloud-to-cloud lightning [3]. The remaining twenty-five-percent of lightning originates from the cloud and discharges to the ground; this is called Cloud-to-Ground (CG) lightning.

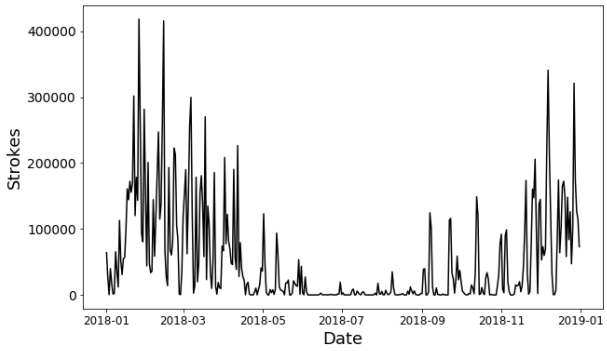

Fig. 1. Number of Lightning Strokes per day in South Africa during 2018.

Although lightning is familiar and well-researched, its prediction remains challenging. Recent academic studies have focused on using Numerical Weather Prediction (NWP) data to build a weather model for lightning forecasting. NWP models attempts to use mathematical models of the atmosphere and oceans to predict the weather based on current weather conditions. NWP model data is the culmination of sophisticated and complicated weather modeling often using supercomputers that relies on weather stations for data.

A lightning prediction model based only on actual historical lightning would be advantageous. A model based only on historical time-series would be cost-effective and can be used for real-time forecasting. Three common machine learning models for univariate time-series models are the Auto Regressive (AR), Auto Regressive Integrated Moving Average (ARIMA) and Long-Short-Term-Memory Recurrent-Neural-Network (LSTM) models.

The AR model is a linear predictive modeling technique. The model predicts future values based on past values of the same series by using the AR parameters as coefficients [14]. The ARIMA model also uses this concept but builds on it by including moving average error and integrated components.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

The moving average component indicates that the regression error is a linear combination of error terms whose values occurred contemporaneously and at various times in the past [10]. Moving average component removes non-determinism or random movements from a time series. The integrated component assists to make the data stationary if required.

The LSTM RNN Model is a type of artificial recurrent neural network (RNN). Unlike standard feedforward neural networks, RNN has feedback connections. LSTM models have a unit called a cell that is composed on an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing and making predictions based on time series data since there can be lags of unknown duration between important events in a time series.

Our aim is to compare and evaluate the short-term (three-hours) lightning predictive ability (number of flashes) of the AutoRegressive model, ARIMA model and the LSTM Recurrent Neural Network model using historical CG lightning flash data only. Doing this will lower the cost of lightning prediction and allow for real-time lightning prediction. Time-series univariate analysis have been applied in various datasets; although their accrual is limited, they are easy to implement and often provide a decent approximation for predictions.

Our study will contribute to current literature. This is the first time a LSTM RNN model has been applied to historical lightning data events from the SALDN dataset to evaluate its effectiveness against other common autoregressive techniques. This study will set the foundations of a more accurate lightning forecast model that will also incorporate weather data variables [15].

II. Methodology

Dataset Historical Cloud-to-ground Lightning Data in year 2018. The South African Weather Service established and maintains the South African Lightning Detection Network that records mainly CG lightning strikes [11]. The network currently consists of 24 Vaisala lightning sensors. The SALDN can detect lightning with a location accuracy of approximately 0.5km and an estimated detection efficiency of 90% over most of South Africa [12]. Our dataset had just under 20 million lightning observations that was grouped for every thee-hourly. Data was scaled using with a feature rand between 0 and 1.

Train/test Split. We trained the models using 70% of data and predicted the remaining 30%. For the year, this corresponds to training dates between 1 Jan 2018 to 12 Sep 2018, and testing dates between 13 Sep 2018 and 31 Dec 2018. All negative test predictions were considered as zero, as it is impossible to have a negative number of lightning strikes.

Test for stationery. The AR and ARIMA models require the historical time-series lightning data to be stationary in nature. We used the Augmented Dickey-Fuller test to test for stationary data. The test results indicate a test statistic of t = 5.96 with a p-value of $2.03 \times 10^{-}7$ (p <0.05). This indicates that we can reject the null hypothesis that the data in non-stationary.

A. Machine Learning Models

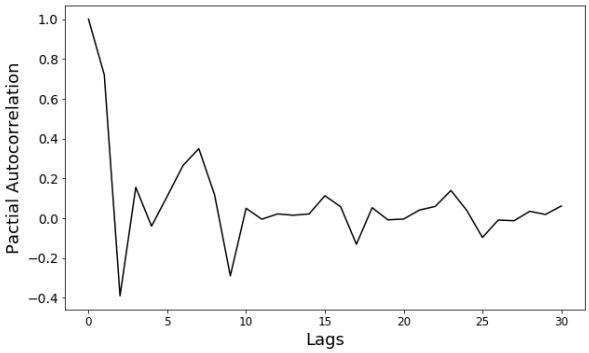

Fig. 2. Partical AutoCorrelation Function with SALDN Historical Lightning Data. The graph indicates a lag value of 8 is inclusive.

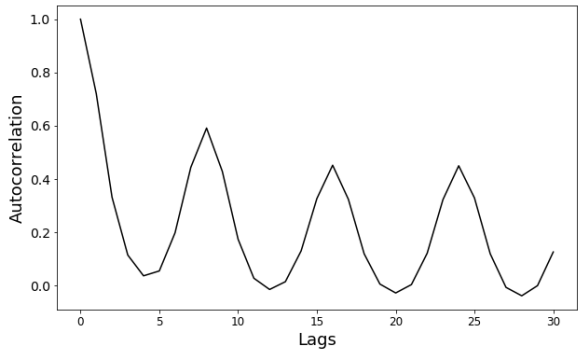

Fig. 3. AutoCorrelation Function with SALDN Historical Lightning Data. The graph indicates a lag value of 1 is optimal.

AutoRegressive Model. The general equation for the AR model is as follows:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-1} + \epsilon_t,$$

where p indicates number of lag steps, ϕ_i are the parameters of the model, c is a constant and ε_t is white noise. We used a lag value based on the Partial-Auto-Correlation-Function 2, which corresponds to a lag value of eight. The AR model from the Scikit-learn library of Anaconda Python v2019.10 was used. The Ordinary Least Square (OLS)/MSE error was minimized.

ARIMA Model. ARIMA is a popular class of autoregressive models that builds on the AR. model. In addition to AR component, the ARIMA model makes data stationary and also incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Given a time series data X_t where t is an integer index and the X_t are real numbers, This defines an ARIMA(p,d,q) process with drift $\frac{\delta}{1-\sum \phi_i}$,

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t = \delta + \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t,$$

where L is the lag operator, the α_i are the parameters of the autoregressive part of the model, the θ_i are the parameters of the moving average part and the ε_t are error terms.

Our ARIMA model is optimized with an AR value of 8, Integrated value of 0, and MA value of 2. The AR parameter is taken from Fig. 2. The MA value of 1 is taken from the Auto-Correlation-Function of Fig. 3. An integrated value of 0 is used as the data is stationary. The ARIMA model from the Scikit-learn library of Anaconda Python v2019.10 was used. The Ordinary Least Square (OLS)/MSE error was minimized.

LSTM Recurrent Neural Network Model. Long Short-Term Memory Recurrent Neural Network. As mentioned earlier, LSTM is a special kind of RNN with additional features to filter recurrent data that is well-suited for time-series data forecasting. Our network has two dense network layer of fifty units followed by one dense layer with an activation function of rectified Linear Unit (ReLU) to reduce negative forecast values. We ran the model for 200 epochs; Fig 4 indicates that 200 epochs are sufficient minimize loss. The LSTM Keras module version 2.3.1 was used for LSTM RNN Network.

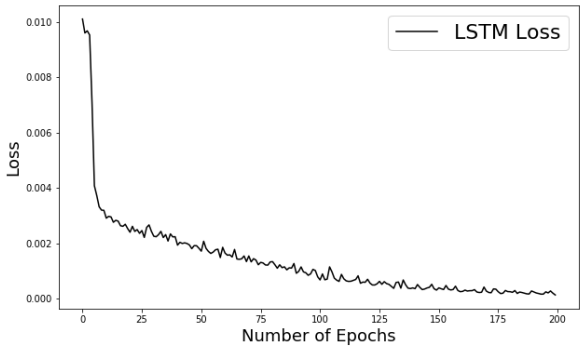

Fig. 4. Loss Function of LSTM RNN Model.

Limitations. Our data only processed one year of historical lighting data; this data should have seasonality which will improve prediction power.

III. Results

A graphical summary of the predicted values versus the actual data is shown in Figures 5, 6, and 7. The AR and ARIMA models performed comparably similar with MAPE values of 15312 and 15080 respectively. The RMSE values were 8579 and 8301 respectively I. The LSTM model had MAPE and RMSE values of 3705 and 9426, indicating it had outperformed AR and ARIMA considerably.

TABLE I
Prediction Model Error Values

Model	MAPE Value	RMSE Value
AR(p)	15312	8579
ARIMA	15080	8301
LSTM	3705	9426

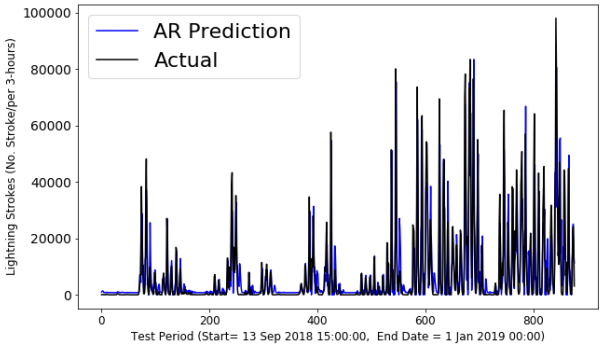

Fig. 5. AR Prediction Model Results vs. Actual Flashes.

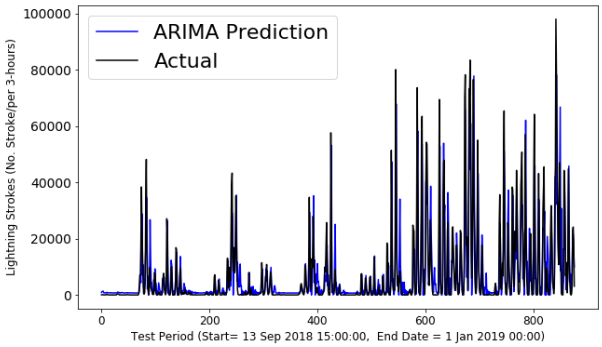

Fig. 6. ARIMA Prediction Results vs. Actual Flashes.

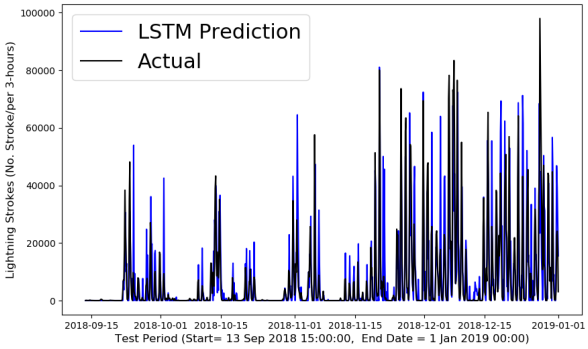

Fig. 7. LSTM RNN Prediction Results vs. Actual Flashes.

A graphical summary of the predicted values versus the actual data is shown in Figures 5, 6, and 7. The AR and ARIMA models performed comparably similar with MAPE values of 15312 and 15080 respectively. The RMSE values were 8579 and 8301 respectively I. The LSTM model had MAPE and RMSE values of 3705 and 9426, indicating it had outperformed AR and ARIMA considerably.

TABLE II
RESULTS OF AR COEFFICIENT AND INTERCEPT

Lag	AR Coefficients
AR Lag 1	0.9479
AR Lag 2	-0.5274
AR Lag 3	0.2505
AR Lag 4	-0.1218
AR Lag 5	0.0615
AR Lag 6	-0.127
AR Lag 7	0.3206
AR Lag 8	0.0702
Intercept	0.0054

TABLE III
RESULTS OF ARIMA COEFFICIENT AND INTERCEPT

Lag	ARIMA Coefficients
AR Lag 1	0.2063
AR Lag 2	-0.0816
AR Lag 3	0.1331
AR Lag 4	-0.0736
AR Lag 5	-0.0054
AR Lag 6	-0.019
AR Lag 7	0.0936
AR Lag 8	0.4647
MA Lag 1	0.7489
MA Lag 2	0.3323
Intercept	0.0456

Table II shows that the first lag value has the highest correlation compared to other lag steps for the AR model. Table III indicate the Moving Average Lag 1 values have the highest correlations actual data.

IV. DISCUSSION

In this study, we evaluated the AutoRegressive, ARIMA and LSTM models to forecast the number of lightning strikes for a period of 3-hours into the future. The AR and ARIMA performed comparably, and the LSTM considerably outperformed all models based on MAPE values (Table I).

Our study is not directly comparable with recent lightning prediction studies. The most recent study to investigate same-day lightning prediction in South Africa is by [11]. In this study, the authors found an AUC ratio of 90% using a stepwise logistic regression mode. But the study predicted the occurrence of a lightning strike occurring rather than the quantitative number of lightning strikes.

Numerous studies have found that LSTM models outperform AR and ARIMA models in forecasting using univariate time-series data [6], [7]. Even with weather series data [8], [9]. Reference [6] found the average reduction in error rates obtained by LSTM was between 84 - 87% when compared to ARIMA indicating the superiority of LSTM to ARIMA for various time-series data. Reference [8] found LSTM models performed about 20% better than ARIMA to predict wind-speed based on MAPE.

LSTM RNN models have several advantages over linear regression models but does not provide explanatory ability. Firstly, LSTM models are able to perform more complex functions than regression models. Secondly, they are able to analyses data with less restrictions such as the stationary requirement of data in regression models. The fact that LSTM models outperform autoregressive models indicates that lightning is non-parametric in nature and involves numerous dependencies. A limitation of neural network models is that they do not provide an understanding of how the results arises.

Recommendation for future studies. We believe that LSTM prediction accuracy can improve if we used a larger dataset to incorporate seasonal trends. This should increase predictive accuracy as lightning is seasonal. If we want an more accurate but data-intensive model, Numerical Weather Prediction parameters can also incorporated into the LSTM model. To gain a better understanding on the factors that influence lightning density, we suggest using a non-parametric machine learning techniques,

V. CONCLUSION

In this study, we predicted the number of lightning strikes for a three-hour period within South Africa. We found that the LSTM RNN model significantly outperforms AR and ARIMA models based on MAPE values. But all models still have relatively high error rates. This study indicates that lightning is better predicted using non-parametric techniques to due its nature. Future models may focus on using non-parametric modelling to better understand and predict lighting.

ACKNOWLEDGMENT

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant number: 121835 and 98244). The authors would also like to thank The South African Weather Service, specifically Morné Gijben and Andrew van der Merwe, for the use of the SALDN data.

The support of the DSI-NICIS National e-Science Postgraduate Teaching and Training Platform (NEPTTP) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NEPTTP

REFERENCES

[1] M. A. Uman, "The Lightning Discharge," ISSN. Academic Press, ISBN: 9780080959818, 1987, pp. 10—18.
[2] J. R. Dwyer, and M. A. Uman, "The physics of lightning," In: Physics Reports 534.4, ISSN: 0370-1573, pp. 147—241, 2014.
[3] V.A. Rakov and M.A. Uman, "Lightning: Physics and Effects," Cambridge University Press, ISBN: 9780521583275, pp. 15—20, 2003.
[4] R. L. Holle, and M. A. Cooper, "Lightning Fatalities in Africa From 2010-2017," In: 2018 34th International Conference on Lightning Protection (ICLP), pp. 1—4, 2018.
[5] T. B, Andersen, and C. J. Dalgaard, "Power outages and economic growth in Africa," In: Energy Economics 38, ISSN: 0140-9883, pp. 19—23, 2013.

[6] S. Siami-Namini, N. Tavakoli, and A. Siami-Namin, “A comparison of ARIMA and LSTM in forecasting time series,“ 2018 17th IEEE International Conference on Machine Learning and Applications, 2018.
[7] E. S. Karakoyun, and A. O. Cibikdiken, “Comparison of arima time series model and lstm deep learning algorithm for bitcoin price forecasting,“ The 13th multidisciplinary academic conference in Prague, 2018.
[8] Q. Cao, B. T. Ewing, and M. A. Thompson, “Forecasting wind speed with recurrent neural networks,“ European Journal of Operational Research, Volume 221, Issue 1, pp 148—154, 2012.
[9] Z. Pala, and R. Atici, “Forecasting Sunspot Time Series Using Deep Learning Methods,“ Solar Physics, 294(5), 50, 2019.
[10] G. E. P Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, “Time Series Analysis: Forecasting and Control,“ 5th Edition, John Wiley & Sons, pp 55—69, 2016
[11] M. Gijben, L. Dyson, and M. Loots, “A statistical scheme to forecast the daily lightning threat over southern Africa using the Unified Model,“ Atmospheric Research, vol. 194, pp 78—88, 2017.
[12] M. Gijben, “The lightning climatology of South Africa,“ South African Journal of Science, vol. 108, pp 44—53, 2012.
[13] V. Gandhi, “Brain-Computer Interfacing for Assistive Robotics Electroencephalograms,“ ISBN-13: 978-0128015438, Elsevier, pp 7—63, 2015.
[14] V. Gandhi, “Brain-Computer Interfacing for Assistive Robotics Electroencephalograms,“ ISBN-13: 978-0128015438, Elsevier, pp 7—63, 2015.
[15] Y. Yaseen, R. Ajoodha, and H. G. P. Hunt, ”A LSTM Recurrent Neural Network for Lightning Flash Prediction within Southern Africa using Historical Time-series Data,” 6th IEEE International Conference on Sustainable Technology and Engineering, 2020.