Quantification and Forecasting of Cumulonimbus (Cb) Clouds Direction, Nebulosity and Occurrence With Autoregression Using 2018-2020 Dataset From Yaounde-nsimalen

Mbucksek Blaise Ringnwi
University of Maroua

DAÏKA Augustin (augustindaika@yahoo.fr)
University of Maroua

TSEDEPNOU Rodrigue
University of Maroua

Bon Firmin André
University of Maroua

Kossoumna Libaa Natali
University of Maroua

Research Article

Keywords: quantification of cumulonimbus cloud direction, Forecasting of Cumulonimbus, SARIMA Model, Autoregressive Neural Network (ANN)

Posted Date: January 5th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1174494/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
QUANTIFICATION AND FORECASTING OF CUMULONIMBUS (Cb) CLOUDS DIRECTION, NEBULOSITY AND OCCURRENCE WITH AUTOREGRESSION USING 2018-2020 DATASET FROM YAOUNDE-NSIMALEN

Mbucksek Blaise Ringnwi¹, Daïka Augustin¹, ², TSEDEPNOU Rodrigue³, BON Firmin André¹ and KOSSOUMNA LIBAA Natali¹

¹ Department of Meteorology and Climatology, National Advanced School of Engineering, University of Maroua, Maroua, Cameroon
² Laboratory of Earth’s Atmosphere Physics, Department of Physics, Faculty of Science, University of Yaoundé 1, Yaoundé, Cameroon
³ Agency for Aerial Navigation Safety in Africa and Madagascar of Yaounde-Nsimalen, Yaoundé, Cameroon

Corresponding author: augustindaika@yahoo.fr ; blaisembucksek@gmail.com
ABSTRACT

This work reports the quantification and forecasting of Cumulonimbus (Cb) clouds direction, nebulosity and occurrence with auto regression using 2018-2020 dataset from Yaoundé –Nsimalen of Cameroon. Data collected for October 2018-2020 consisted of 2232 hourly observations. Codes were written to automatically align, multi-find and replace data points in Excel to facilitate treating big datasets. The approach included quantification of direction generating time series from data and determination of model orders using the correlogram. The coefficients of the SARIMA model were determined using Yule-Walker equations in matrix form, the Augmented Dickey Fuller test (ADF) was used to check for stationarity assumption, Portmanteau test to check for white noise in residuals and Shapiro-Wilk test to check normality assumptions. After writing several algorithms to test different models, an Autoregressive Neural Network (ANN) was fitted and used to predict the parameters over window of 2 weeks. Autocorrelation Function (ACF) shows no correlation between residuals, with $p \leq 0.05$, fitting the stationarity assumption. Average performance is 80%. A stationary wavelike occurrence of the direction has been observed, with East as the most frequented sector. Forecast of Cb parameters is important in effective air traffic management, creating situational awareness and could serve as reference for future research. The method of decomposition could be made applicable in future research to quantify/forecast cloud directions.

Keywords: quantification of cumulonimbus cloud direction, Forecasting of Cumulonimbus, SARIMA Model, Autoregressive Neural Network (ANN)
I-INTRODUCTION

Cumulonimbus (Cb) clouds are the most important clouds in the study of various natural phenomena and in everyday life (Bhat and Kumar, 2015; LeMone and Zipser, 1980; Thierry Lombry, 2005; Wmo, 2017, Putra et al., 2021; Rais et al., 2020). We can cite, among others, agriculture, economy, telecommunications, aviation safety and recreation. Indeed, they play a key role in the water and energy cycles and are often accompanied by lightning, visibility and thunderstorms (Hale, 2018, 2012; Mbane, 2009, 2015). They are sometimes accompanied by severe weather conditions causing turbulence, structural damage and safety problems for air navigation (Daïka et al., 2021). These cumulonimbus clouds are continuously monitored in air navigation because of their dangerous and adverse effects they present. Although Cb is often included in the Terminal Area Forecast (TAF) section of aeronautical messages transmitted, forecasts on the Cb parameters are never included in any messages, not even in flight take-off forecast. However, Cb is a major determinant in flight schedules and delays, as it influences the flight trajectory and time of flight take-off. Thus, if Cb were included in flight take-off forecasts, then this would reduce uncertainty and risk in air traffic management and increase safety in air navigation.

The sudden occurrence of Cb often leads to flight delays, rescheduling, and change in flight trajectories. Such courses of action present:

(i) Economic disadvantages due to increased energy consumption and time wastage.

(ii) Uncertainty and psychological constraints in pilots and air traffic controllers. This leads to poor decision making and less efficient air traffic management.

Despite these undesirable issues resulting from Cb on air navigation, very limited or few studies have tried to analyze and forecast the occurrence of Cb. In addition, no recent study has been done to forecast the nebulosity and direction of occurrence of the cloud.

Also, it is important to note that data on direction and nebulosity are either qualitative or categorical, making descriptive analysis possible, but quantitative analysis of variables such as Cb direction (N, SE, SE/NE) has been rendered difficult.

This work seeks to:

(a) Employ a simple but novel approach of quantification (decomposition) to quantify the variables.

(b) Use auto regression models to train the dataset (October 2018-2020)

(c) Fit the best model and use it to provide roll-forward forecasting of Cb parameters, beginning October 2021.

(d) Provide potential necessary elements for a more effective air traffic management by using the model to forecast future outcomes.
The solutions proposed here could be helpful in an effective air traffic management, increasing efficiency, reducing uncertainty and economic disadvantages.

The method of quantification used here could be made applicable in combination with vector auto-regression (recommended at the end of the work) to have better results in any future research on the subject.

The research findings in this work are based on the claim that air traffic management and improvement of air safety could be made possible by providing forecasts for Cb parameters. Forecasts of Cb parameters improves situational awareness of pilots and enhances their decision making process. Also, a prior knowledge about the future occurrence of Cb on flight trajectories could inform air traffic controllers to better manage air traffic and navigation during Cb occurrence. An empirical analogy was made on the harmonic or rhythmic occurrence of Cb over Yaounde-Nsimalen.

II- DATA ACQUISITION, PREPROCESSING AND VISUALIZATION

The data relating to the two variables was obtained from records of meteorological data at Nsimalen. Such hourly records were found in the observation booklet and had been stored on the PROGIMET. Each hourly observation was then checked for the presence or absence of Cb. When Cb was present, the numerical value corresponding to the nebulosity of Cb was recorded for the variable, Nebulosity. A value of zero was assigned to the variable when hourly observations showed an absence of Cb.

The direction of the cloud had been recorded in a qualitative or categorical form, as in “N”, or “N/NE” as shown in Table 1 below. Thus, time series analysis on the direction of the cloud necessitated quantification, the conversion of the qualitative variable into a quantitative variable (numerical expression).

Table 1: The first 6 hourly observations of CB-direction for the first week of October 2018

| Day/Time | 0 | 1 | 2 | 3 | 4 | 5 |
|----------|---|---|---|---|---|---|
| 1        | NE/SW/NW | NE/SW/NW | NE/SW/NW | NE/SW/NW | W  | W  |
| 2        | E/S/N     | E/S/N     | E/S/N     | E/S/N     | E/S/N | E/NW |
| 3        |           |           |           |           |     |     |
| 4        |           | NE/NW/N   | NE/NW/N   | NE/NW/N   | NE/NW/N | NE/NW/N |
| 5        |           |           |           |           |     |     |
| 6        | E/S       | E/S       | E/S       | E/S       | E/S |     |
| 7        |           |           |           |           |     |     |
Since the nebulosity had already been collected in a quantitative form, it necessitated no quantification. Therefore, we will mostly deal in the following with the pre-processing of the qualitative variable, direction of cumulonimbus clouds.

Figures 1-2 show the time evolution of the direction when binary representation has been used.

In the Figure 1, we notice that the values evolve as binaries or square signals between 0 and 1, discarding the need for normalization. This is done for the sectors N, S, NE and E. In this figure 2, we notice that the direction of Cb with respect to the other 4 principal components SE, SW, W, and NW are shown here.

The dataset corresponding to each variable (Nebulosity and direction) was partitioned into two: the training set and validation set. The goal of data partitioning is to use the training set to train the statistical model and fit it to the given dataset. Partitioning is important because after training of the model with the dataset, the validation set can now be used to test or validate the accuracy of the model and test model performance. Later on, the results obtained from the validation would tell whether or not the model is fit, whether the model should be modified by some exponential smoothing method or some log transform method, or whether the model should be replaced by a better performing model.

Figure 1: Time series showing the evolution of occurrence of decomposed direction.
The partition was such that:

Length of Training set per variable = 80% (Length of dataset per variable)

\[
\text{Length of Validation set per variable} = \text{Length of dataset per variable} - \text{Length of Training set per variable}
\]

Thus, the length of the training set for each variable = 2233 x (80/100) = 1786.

For the validation set, we have:

\[
\text{Length of Validation set per variable} = \text{Length of dataset per variable} - \text{Length of Training set per variable}
\]

The length of the validation set per variable = 2233 – 1786 = 447.

The partitioned dataset for each of the nine variables is shown in Figures 3, 4, 5 and 6. Figure 3 shows the partitioning of the nebulosity dataset, while Figures 4, 5 and 6 show partitioning of the direction. The red plots represent the training set, while the green ones represent the validation set. The linear line along the plot of the dataset is the trend line that varies as a function of the level of the series. Figure 3 shows the entire dataset in red and the partitioned dataset. In red, we have the training set, and in blue, we have the validation set. The black horizontal line is the trend line. In these figures 4, 5 and 6, the red lines display the training set (1st 1786 observations). The blue lines show the validation set (for the next 446 observations), and the black line is the trend line which shows the tendency of evolution of the dataset.
Figure 3: Partitioning of nebulosity dataset into a training and validation set.

Figure 4: Partitioning of the dataset for the principal components; N, NE, and E into training and validation set.

Figure 5: Partitioning of the dataset for the principal categories SW, W, and NW into the training and validation set.
The approach taken in this work was to partition the dataset of each variable into 2 sets: the training set and the validation set. The dataset for the nebulosity variable consisted of 2233 data points, while for the direction variable we have $8 \times 2233$ data points (2232 data points for each category). The higher number of data points for the direction was accounted for by the decomposition of direction into 8 independent variables.

An algorithm was first written and used to test several ARIMA and SARIMA models of different orders. The following tests were done:

- The Augmented Dickey Fuller test (ADF) was done to check for the stationarity condition of each variable.
- The LJung Box (Box Pierce) test was used to check for the presence of white noise in the residuals.
- Shapiro-Wilk test for normality was made to check if the residuals (errors) had a normal distribution.

The optimal models were then selected based on the results of these statistical tests. As will be seen, other tests were performed, and the work made use of several other criteria to choose an optimal model.

Parameters of the model for each variable were estimated using the autocorrelation and partial autocorrelation function plots (correlogram) to determine the orders of the autoregressive process respectively.

The coefficients of the mathematical equations that model each process were estimated using the Yule Walker equations. Finally, a Neural Network Auto-regression model was appropriated to the analysis of the direction and nebulosity when ARIMA models didn’t provide a better fit.

A roll-forward forecast was adopted with a forecast window of 2 weeks (336 hours). A simple comparison between forecasted value and actual value was used to show the predictive ability of the model.
Since the range for the direction variables was [0, 1], normalization was not necessary, while for Nebulosity, normalization was not considered an option, due to presence of zero values.

### III.1- Methods and Models in Time series analysis

The methods used in time series analysis are data-driven methods, econometric methods and extrapolation methods (Galit and Kenneth, 2016; Cryer and Kung-Sik Chan, 2008). This work focuses on the application of a data-driven method to time series analysis using autoregression. Linear regression or logistic regression methods are used depending on the nature of the fit and the variable. When a time series is data-driven, auto-regression models (AR) could be appropriated to the dataset. Equation (1) is an AR process of order $p$, and noise $Z_t$ (Brockwell and Davis, 2016; MONTGOMERY et al., 2015):

$$X_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_p Z_t$$  \hspace{1cm} (1)

Where $f(I) = Z_t$, $Z_t \equiv (0,\sigma^2)$

When $X_t$ is expressed as a function of $Z_t$ and some linear coefficient, $\theta_i$, $i \in \{1, 2, \ldots, q\}$, as in (2), $X_t$ is a moving average process (MA) of order $q$ (Wayne et al., 2017).

$$Z_t = \theta_0 + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q X_{t-q} + X_t$$  \hspace{1cm} (2)

Where $\theta_0 = -\mu$ and $\mu$ is the mean.

The basic models encountered and widely used in time series analysis include AR(p) models, describing auto-regressive processes of order $p$, obeying (1) and MA(q) models, describing moving average processes of order $q$, obeying (2). Based on these two models, we could have ARMA models, ARMA integrated models (ARIMA), seasonal integrated models (SARIMA), GLM, GARCH, ARCH, ARUMA, and even ANN, Auto-regressive Neural Network models (Galit and Kenneth, 2016; Brockwell and Davis, 2016; Wayne et al., 2017).

Differencing, simple exponential smoothing (SES), or Holt-winter’s smoothing could be used to adjust or properly fit our dataset.

### III.2- Parameter Estimation and model identification

Parameters of the various models were estimated using the following methods: Yule walker equations, Autocorrelation and partial autocorrelation plots and Algorithmic determination.
### III.2.1 Parameter Estimation

For an MA(q) process, the order, q, of the moving average process is often given as the number of significant spikes that show up in the ACF plot. This may be true for a weakly stationary time series. However, if the ACF plot is sinusoidal, then the significant spikes indicate the order of the AR(p) process. The PACF would then do for the MA(q) process, and vice versa.

From the AR process defined in (1), we can develop the Yule Walker Equations. By taking expectations on both sides of equation (1):

$$E(X_t) = E(\beta_0) + E(\beta_1 X_{t-1}) + E(\beta_2 X_{t-2}) + ... + E(\beta_p X_{t-p}) + E(\beta_0 Z_t)$$  \hspace{1cm} (3)

From (3), since the AR process is considered stationary with a constant mean:

$$E(X_t) = E(X_{t-1}) = E(X_{t-2}) = \mu$$ and $$E(Z_t) = 0$$, recall $$Z_t \equiv (0, \sigma^2)$$.

Equation (3) becomes:

$$\mu = \beta_0 + \beta_1 \mu + \beta_2 \mu + ... + \beta_p \mu$$  \hspace{1cm} (4)

Subtracting (4) on both sides of (1):

$$(X_t - \mu) = \beta_1 (X_{t-1} - \mu) + \beta_2 (X_{t-2} - \mu) + ... + \beta_p (X_{t-p} - \mu) + \beta_0 Z_t$$  \hspace{1cm} (5)

Considering that $$(X_t - \mu) = \tilde{X}_t$$

Equation (5) becomes:

$$\tilde{X}_t = \beta_1 \tilde{X}_{t-1} + \beta_2 \tilde{X}_{t-2} + ... + \beta_p \tilde{X}_{t-p} + \beta_0 Z_t$$  \hspace{1cm} (6)

Since $$\tilde{X}_t$$ is a dummy variable, we can consider (6) as equivalent to:

$$X_t = \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_p X_{t-p} + \beta_0 Z_t$$  \hspace{1cm} (7)

By multiplying (7) by $$X_{t-1}$$ and taking the expectation on both sides, we have:

$$E(X_t, X_{t-1}) = \beta_1 E(X_{t-1}, X_{t-1}) + \beta_2 E(X_{t-2}, X_{t-1}) + ... + \beta_p E(X_{t-p}, X_{t-1}) + \beta_0 E(X_{t-1}, Z_t)$$  \hspace{1cm} (8)

$$E(X_t, X_{t-1}) = \beta_1 E(X_{t-1}, X_{t-1}) + \beta_2 E(X_{t-2}, X_{t-1}) + ... + \beta_p E(X_{t-p}, X_{t-1})$$  \hspace{1cm} (9)

Since $$E(\beta_0 X_{t-1} Z_t) = 0$$
For a lag k process, (8) is:

\[ X_{t+k} = \beta_1 X_{t+k-1} + \beta_2 X_{t+k-2} + \ldots + \beta_p X_{t+k-p} + \beta_0 X_{t-1} + Z_{t+k} \]  

(10)

Thus, (9) becomes:

\[ E(X_{t+k}, X_{t-1}) = \beta_1 E(X_{t+k-1}, X_{t-1}) + \beta_2 E(X_{t+k-2}, X_{t-1}) + \ldots + \beta_p E(X_{t+k-p}, X_{t-1}) \]  

(11)

But \( E(X_{t+k}, X_{t-1}) = E(X_t, X_{t-1}) \), for all stationary process and \( \rho_k = E(X_{t+k}, X_{t-1}) \)

Equation (11) becomes:

\[ \rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \ldots + \phi_p \rho_{k-p} \]  

(12)

The relation (12) is the Yule walker equation, which is used to estimate the coefficients of the model, \( \phi_1, \phi_2, \ldots, \phi_p \).

Note that \( \rho_0 = 1 \), since the autocorrelation function at lag 1 is always 1.

Where \( k = 1, 2, 3, 4, 5 \ldots p \). Substituting the values of \( k \):

\[ \begin{align*}
\rho_1 &= \phi_1 + \phi_2 \rho_1 + K + \phi_p \rho_{1-p} \\
\rho_2 &= \phi_1 \rho_1 + \phi_2 + K + \phi_p \rho_{2-p} \\
\rho_3 &= \phi_1 \rho_2 + \phi_2 \rho_1 + K + \phi_p \rho_{3-p} \\
&\vdots \\
\rho_p &= \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + K + \phi_p \rho_{p-p} 
\end{align*} \]  

(13)

Equation (13) in matrix form gives:

\[
\begin{pmatrix}
\rho_1 \\
\rho_2 \\
\vdots \\
\rho_p
\end{pmatrix} =
\begin{bmatrix}
1 & \rho_1 & L & \rho_{1-p} \\
M & O & M & M \\
M & O & M & M \\
M & O & M & M \\
M & O & M & M
\end{bmatrix}
\begin{pmatrix}
\phi_1 \\
\phi_2 \\
\vdots \\
\phi_p
\end{pmatrix}
\]  

(14)

Thus, the general form of the Yule Walker equation is \( r = M \times R \) where \( R \) is the column matrix of the coefficients, and \( r \), the column matrix of the covariance function. The coefficients are then estimated by an inverse matrix operation of the form: \( R = M^{-1} \times r \).

From the Yule walker estimates, the coefficients of the two experimented ARIMA models of the nebulosity are:
1) Arima(3, 0, 1); coefficients are 1.6051, -0.5955, -0.0714, -0.8062

Governing equation:

\[ X_t = 1.6051X_{t-1} - 0.5955X_{t-2} - 0.0714X_{t-3} - 0.8062Z_t \]  

2) Arima(5, 0, 1); coefficients are 1.4912, -0.5274, -0.0333, 0.0327, -0.0592, -0.6963

Governing equation:

\[ X_t = 1.4912X_{t-1} - 0.5274X_{t-2} - 0.0333X_{t-3} + 0.0327X_{t-4} - 0.0592X_{t-5} - 0.6963Z_t \]

III.2.2 - Model Identification

The models that were then selected in the next phase were chosen based on a given set of criterion. The relation (17) displays the Akaike Information Criteria (AIC) is defined as a function of the maximum likelihood estimator and the orders of the MA (q) and AR (p) processes respectively:

\[ AIC = -\ln(MLE) + (p + q + 1) \]  

The Bayesian Information Criteria (BIC) is a similar estimator as the AIC; however, the BIC applies more punishment to a complex model (model of higher order).

The parsimony principle was used to choose a simpler model (model of lower order) to a complex model as a result of performance and O (n). Equation (18) states the parsimony principle:

\[ p + d + q \leq 6 \]

III.3-Model fitting and residual analysis

The smoothing method finally used here was simple exponential smoothing (SES). Finally, the performance of the model was assessed. As it shall be seen, the accuracy of the model and the residual analysis results are the main reasons why an autoregressive neural network was preferred.

The residuals of each variable correspond to the forecast errors during the validation period. The residuals were computed using the R software.

IV- RESULTS AND DISCUSSION

IV.1- Forecasts of time series

IV.1.1 Nebulosity

The forecasts of the nebulosity for the first 400 hours of October (first 2 weeks) are shown in Figure 7. The blue lines show the forecasts, while the black lines show the dataset. The forecasts are based on an artificial
autoregressive neural network with 33 input nodes, 25 hidden nodes, 876 weights or coefficients and 1 output node.

This model was chosen after trying a previous 33-17-1 ANN model. The Activation function used in this case is the linear activation (instead of the logit). The reason is because of feasibility performance. In this figure 7, Black curves show the actual dataset, green curves show the fitting of the training dataset and red lines show fitting of the validation set. The superposed blue lines are forecasts of the nebulosity.

Figure 7: Forecasts of the nebulosity based on an optimally chosen NNAR.

IV.1.2 Forecasted Sector

The forecasted sector or direction of the Cb obtained from the analysis and forecasting of the dataset is shown in Figures 8, 9 and 10, respectively, for the 8 principal categories. In this figure 8, Red lines are test results from the validation process; the blue lines are the final predicted values for the next 2 weeks. In this figure 9, the black lines show the actual dataset, the red lines, the behavior of the model during the validation period and the blue lines, the forecasts of the categories for the next 2 weeks. In this Figure 10, the black lines show the actual dataset, the red lines are the forecasts during the validation period, and the blue lines, the final forecasts after testing for the next two weeks. Table 2 summarizes the entire CB forecast for 23 hourly observations on October 1, 2021.
Figure 8: Forecasts of the Cb sector for the first 4 categories (N, NE, E, and SE).

Figure 9: Forecasts of the Cb sector for the categories S, SW, and W.
Figure 10: Forecasts of the NW direction, using the NNA model.

Table 2: CB forecasts for 1st 23 hours of October 1, 2021

| Time | Occurrence | Direction | Nebulosity |
|------|------------|-----------|------------|
| 1    | CB         | NO CB     | FEW        |
| 2    | CB         | N/E       | FEW        |
| 3    | CB         | N/E       | FEW        |
| 4    | NO CB      | N/E       | NO CB      |
| 5    | NO CB      | E         | NO CB      |
| 6    | NO CB      | E         | NO CB      |
| 7    | NO CB      | NO CB     | NO CB      |
| 8    | NO CB      | NW        | NO CB      |
| 9    | NO CB      | NW        | NO CB      |
| 10   | CB         | NW        | FEW        |
| 11   | CB         | NW        | FEW        |
| 12   | CB         | NO CB     | FEW        |
| 13   | CB         | SE        | FEW        |
| 14   | CB         | SE/SW     | FEW        |
| 15   | CB         | SE/SW     | FEW        |
| 16   | NO CB      | SE        | NO CB      |
| 17   | NO CB      | SE        | NO CB      |
| 18   | CB         | NO CB     | FEW        |
| 19   | CB         | NO CB     | FEW        |
| 20   | CB         | N         | FEW        |
| 21   | CB         | N         | FEW        |
| 22   | CB         | N/E       | FEW        |
| 23   | CB         | N/E       | FEW        |
IV.2-Model Evaluation

IV.2.1-Results from residuals

Figure 11 displays the distribution of the forecast errors. The forecast errors have a turkosis or skewness inclined to the left, which shows that the distribution of the forecast errors does not meet exactly with the normality assumption for forecast errors. Thus, while this model was used and found as a better fit compared to antecedent arima models that were tested, the Portmanteau test for residuals (Box Pierce test) shows that this normality is not met.

The high frequency at 0 shows a deviation from the normal distribution of residual errors. This could possibly mean that the model is over-fitted to the dataset.

In this figure 11, Forecast errors are in blue while the red curve represents the actual normal distribution. Notice the high frequency at 0, presenting a skewness or turquosis.

Figure 11: Histogram showing the frequencies of forecast errors.

Figure 12 gives the plot of the residuals (forecast errors) of the nebulosity as a function of time at the top, while the 2 bottom plots show the autocorrelation functions of the residual errors and the distribution of the forecast errors. The ACF shows clearly that there is no autocorrelation of residuals. Thus, the residuals are completely generated from a random process. The autocorrelation values are all found to be below the p=0.05 level. This could be confirmed with the Shapiro-Wilk test for residuals that checks the fit of models against residuals as null hypothesis.

The bottom right plot shows the distribution of forecast errors, which has a normal distribution. The normal distribution for the NNAR (33, 17) model shows that the model is quite fit to use for forecasting. Top figure displays time evolution of forecast errors; bottom left, the autocorrelation for first 33 lags, and bottom right, the distribution of forecast errors.
Again, the previous neural model that was tested for our dataset could be modeled differently by applying log transformation to the absolute value of residuals to get a normal distribution of forecast errors. Figure 13 shows the absolute log-transformation, which shows an approximation to a normal distribution. Red lines show the normal distribution. A log-transform have been applied to the data in Figure 11 (in blue) to get a distribution closer to the normal than in the former.

IV.2.2-Accuracy Measures and Predictive Performance

The accuracy of the final model was tested by checking the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE), and the Root Mean Squared Error (RMSE). These are all predictors of the degree of fitness of each model. A better predictor such as the AIC would take into account the MLE. However, the former estimators are estimators of the goodness of fit of the model.
For the Nebulosity dataset, Figure 14 gives the accuracy in terms of performance of the model fitted to the training set.

From computation, the goodness of fit estimators obtained for the ANN process includes:

\[ \text{MAE} = 0; \text{MAPE} = 0; \text{RMSE} = 0 \]

While these values seem very ideal, this could possibly be an indicator of model over fitting. This seems to be the case because, graphically, the model seems to work very well on the training data and poorly on the validation set.

**Figure 14**: Testing the accuracy of an over-fitted ANN.

In this figure 14, we observe the high fit during the training period, but a poor fit for the validation period. Clearly, it was necessary to choose a better model that fits both validation period and training period.

To evaluate the predictive performance, each forecasted value has been compared against the actual value in the validation period. Table 3 summarizes this process.

**Table 3**: Test forecast for CB using validation set

| TRUE_OCCUR | VALIDX_OCCUR | TRUE_NEB | VALIDX_NEB | TRUE_DIR | VALIDX_DIR |
|------------|--------------|----------|------------|----------|------------|
| NO CB      | NO CB        | NO CB    | NO CB      | NO CB    | NO CB      |
| NO CB      | NO CB        | NO CB    | NO CB      | NO CB    | NO CB      |
| NO CB      | NO CB        | NO CB    | NO CB      | NO CB    | NO CB      |
| NO CB      | NO CB        | NO CB    | NO CB      | NO CB    | NO CB      |
| CB         | NO CB        | FEW      | NO CB      | N        | NO CB      |
| CB         | CB           | FEW      | FEW        | E/N      | NO CB      |
| CB         | CB           | SCT      | FEW        | TS       | NE         |
| CB         | CB           | SCT      | FEW        | TS       | SE         |
| CB         | CB           | SCT      | FEW        | TS       | SE/SW      |
| CB         | CB           | SCT      | FEW        | S/NW     | NE/SW      |
| CB         | CB           | SCT      | FEW        | S/NW     | NE/SE/SW   |
| CB         | CB           | FEW      | FEW        | NW       | NE/SE/SW/W |
| CB         | CB           | FEW      | FEW        | NW       | NE/SW/W    |
scores  35/42  29/42
accuracy  83.33%  69%

IV.3-Accuracy measures

The accuracy and performance of each dataset has been computed. The predictive performance shows the percentage of correct predictions out of a hundred. The accuracy measures for the forecasts of CB directions is given in table 4. Table 5 shows the accuracy measures of the forecasts for CB nebulosity.

Table 4: Accuracy and performance of CB direction forecasts

| Measure                        | N  | NE | E  | SE | S  | SW | W  | NW |
|--------------------------------|----|----|----|----|----|----|----|----|
| N_True/100                      | 88 | 75 | 69 | 77 | 79 | 81 | 79 | 67 |
| Predictive Performance (in %)   | 88 | 75 | 69 | 77 | 79 | 81 | 79 | 67 |
| Av_Performance (in %)           |    |    |    |    |    |    |    | 77 |
Table 5: Accuracy and performance of CB Nebulosity forecasts

| Variable                  | Performance (in %) |
|---------------------------|--------------------|
| Exact_Neb                 | 60                 |
| Accuracy (Exact Category) | 69                 |
| Correct_Occurrence        | 83                 |
| Predictive Performance    | 83                 |

CONCLUSION

This study consisted of analyzing the time series of Cb over Yaoundé-Nsimalen of Cameroon for 2018-2020 October, to forecast the occurrence, Nebulosity and direction of Cb for October 2021. It aimed to provide a new approach to quantification of qualitative data (decomposition and indexation to quantify Cb direction), forecasts of Cb direction and nebulosity based on a statistical model and a clear idea about the future nature of Cb (thereby making traffic controllers and pilot aware and ready ie improvement efficiency). Recent studies seek to predict simply the occurrence from MTSAT data and others, the forecasts generated here prove to be well performant within narrow limits. The accuracy and predictive performance lies at 77% and 83% for direction and nebulosity respectively. That is relatively as good as a study carried with 75% of accuracy.

The findings in this study could be crucial in strategic flight planning, improvement of situational awareness of pilots via provision of forecast of Cb parameters before take-off and an improvement in safety related to the risks resulting from Cb.

Data decompositions and direction quantification could be a useful approach in future research that incorporate vector auto-regression methods to produce a more powerful forecast.

Acknowledgements:

The authors thank the editor and anonymous reviewers for their careful review, important comments, and helpful suggestions that will serve to improve this article.

Conflict of Interest

The Authors declare that no conflict of interests exist in this paper.

Funding Statement

This study was funded by the authors.
Author's Contribution

Mbucksek Blaise Ringwi: Investigation, Data curation, Visualization, Writing and Methodology; Daïka Augustin: Investigation, Visualization, Writing - original draft and Formal analysis; TSEDEPNOU Rodrigue: Conceptualization, Methodology, Visualization and Software; BON Firmin André and KOSSOUMNA LIBAA Natali: Validation, Writing, Supervision, review and editing.

Availability of data and material:

We have the materials, all the raw and processed data and result products. We can provide the processed data and documents if it is required.

Code availability:

The tools used in this work included Excel software for pre-treatment of data, and R software for statistical analysis and modeling.

Ethics approval:

Not applicable to this paper because there was no potential conflict of interest.

Consent to participate:

Not applicable to this paper

Consent for publication:

The submission of this paper has been approved by all relevant authors and is the independent work of the authors. The latters have seen and agreed to the submitted version of the paper. It has not been submitted and not published or accepted for publication and is not under consideration for publication in another journal or book.

REFERENCES

Bhat G S, and Kumar S (2015) Vertical structure of cumulonimbus towers and intense convective clouds over the South Asian region during the summer monsoon season. J. Geophys. Res. Atmos., 120, 1710–1722, doi: 10.1002/2014JD022552.

Brockwell PJ and Davis R A (2016) Introduction to Time Series and Forecasting. 2nd Ed. Springer, pp 443.

Cryer J D and Kung-Sik Chan (2008) Time Series Analysis with Applications in R. 2nd Ed. Springer, pp 500.
Daïka A, Igri MP, Mounabbahou MO and Mbane BC (2021) Analysis of weather Conditions for Aerial Security and Optimization in Maroua-Salak. Atmospheric and Climate Sciences, Accepted

Galit Shmueli and Kenneth CL Jr (2016) Practical Time Series Forecasting with R: A Hands-On Guide, 2nd Edition. A X E L R O D S C H N A LL P U B L I S H E R S.

Hale RB (2018) Rare, anomalous non-transient lights associated with cumulonimbus clouds, Weather, 73 (7): 215-216.

Hale RB (2012) A ray or streamer of cloud emerging from a cumulonimbus. Weather 67: 306.

LeMone MA and Zipser E J (1980) Cumulonimbus vertical velocity events in GATE: Part I. Diameter, intensity, and mass flux, J. Atmos. Sci., 37, 2444–2457.

Mbane B C (2009) Vertical profiles of winds and electric fields triggered by tropical storms - under the hydrodynamic concept of air particle. International Journal of Physical Sciences, 4 (4), 242-246.

Mbane BC (2015) Earth’s Atmosphere Dynamic Balance Meteorology. Scientific Research Publishing Inc., 110 pp.

MONTGOMERY DC, JENNINGS CL and KULAHCI M (2015) Introduction to time series analysis and forecasting, 2nd ed. Wiley Series in Probability and Statistics, pp. 671.

Putra R M, Fibriantika E, Herawati H, Kusumayanti Y, Afriani E, Hidayanti A, Wiujiana A, Swastiko WA and Andrianto D (2021) Cumulonimbus cloud prediction based on machine learning approach using radiosonde data in Surabaya. Indonesia. IOP Conference Series: Earth and Environmental Science, 724(1), 012047. https://doi.org/10.1088/1755-1315/724/1/012047

Rais AF, Setiawan F, Yunita R, Meinovelia E, Soenardi S, Fadli M and Wijayanto B (2020) Prediction of cumulonimbus (Cb) cloud based on integrated forecast system (ifs) of european medium-range weather forecast (ecmwf) in the flight information region (fir) of jakarta and ujung pandang. Jurnal Sains and Teknologi Modifikasi Cuaca, 21(2). https://doi.org/10.29122/jstmc.v21i2.4100

Thierry Lombry (2005) Cumulus, Cumulonimbus, Futura. Retrieved from https://www.futura-sciences.com/planete/dossiers/climatologie-nuages-tout-genre-differencier-545/page/9/

Wayne AW, Gray H L and Elliott A C (2017) Applied time series analysis, with R. 2nd Ed. Boca Raton: Taylor & Francis, pp 752

WMO (2017) Species: International Cloud Atlas. International Cloud Atlas. Retrieved from https://cloudatlas.wmo.int/en/species-cumulonimbus-cb.html.