A statistical scheme to forecast the daily lightning threat over southern Africa using the Unified Model

Morné Gijben[a,b], Liesl L. Dyson[b], Mattheus T. Loots[c]

[a]South African Weather Service, 442 Rigel Avenue South, Erasmusrand, Pretoria, 0181, South Africa, Private Bag X097, Pretoria 0001, South Africa, morne.gijben@weathersa.co.za

[b]Department of Geography, Geoinformatics and Meteorology, University of Pretoria, cnr Lynnwood Road and Roper Street, Hatfield, Pretoria, 0002, South Africa, Private Bag X20, Hatfield 0028, South Africa, liesl.dyson@up.ac.za

[c]Department of Statistics, University of Pretoria, cnr Lynnwood Road and Roper Street, Hatfield, Pretoria, 0002, South Africa, Private Bag X20, Hatfield 0028, South Africa, theodor.loots@up.ac.za

Corresponding author: Morné Gijben (morne.gijben@weathersa.co.za)

Abstract

Cloud-to-ground lightning data from the Southern Africa Lightning Detection Network and numerical weather prediction model parameters from the Unified Model are used to develop a lightning threat index (LTI) for South Africa. The aim is to predict lightning for austral summer days (September to February) by means of a statistical approach. The austral summer months are divided into spring and summer seasons and analysed separately. Stepwise logistic regression techniques are used to select the most appropriate model parameters to predict lightning. These parameters are then utilized in a rare-event logistic regression analysis to produce equations for the LTI that predicts the probability of the occurrence of lightning. Results show that LTI forecasts have a high sensitivity and specificity for spring and summer. The LTI is less reliable during spring, since it over-forecasts the occurrence of lightning. However, during summer, the LTI forecast is reliable, only slightly over-forecasting lightning activity. The LTI produces sharp forecasts during spring and summer. These results show that the LTI will be useful early in the morning in areas where lightning can be expected during the day.

Keywords

Lightning

Forecasting

Rare-event logistic regression

Numerical weather prediction

1. Introduction

Severe thunderstorms are a major concern to the weather community and the public due to their ability to cause death, injury and damage (Lang et al., 2004). Lightning, tornadoes, strong wind, heavy rainfall and hail are some of the phenomena associated with severe thunderstorms (Kohn et al., 2011). Lightning alone poses a severe threat, since it can cause injury or death to humans and animals (Blumenthal et al., 2012), damage to infrastructures (Lynn and Yair, 2010), and can be a hazard to various industries like aviation and forestry (Price, 2013). It is estimated from satellite observations that about 39-49 lightning flashes occur around the globe every second (Christian et al., 2003). This equates to more than 1.4 billion flashes a day. Lightning is one of the leading causes of death from natural disasters. It causes approximately 24 000 deaths and 240 000 injuries annually around the globe (Blumenthal et al., 2012).

In South Africa, the annual mortality rate due to lightning is estimated to be between 1.5 (in urban areas) and 8.8 (in rural areas) people per million of the population (Blumenthal et al., 2012; Holle, 2008). These statistics are based on published data (Blumenthal et al., 2012), but it is likely to be an underestimate of the actual mortality rate, since lightning deaths are often not reported, especially in rural areas (Trengrove and Jandrell, 2011). Bhavika (2007) stated that the number of lightning deaths in South Africa is about four times higher than the global average. South Africa is a country that consists of mainly two rainfall seasons, austral summer and winter

rainfall seasons. The central to northern interior of the country falls within the summer rainfall region and receives most of its rainfall from convective thunderstorms (De Coning and Poolman, 2011; Kruger, 2007; Landman et al., 2012; Tyson, 1986; Dyson et al., 2015). Most of the rainfall in the winter rainfall regions of the south-western and coastal parts of the country originates from either stratiform clouds or shallow convective clouds that form from the on-shore flow by ridging high-pressure systems and cold fronts (De Coning and Poolman, 2011). Consequently, these systems are associated with low lightning activity (Figure 1d). On the other hand, the summer rainfall area of South Africa is extremely vulnerable to cloud-to-ground (CG) lightning, which occurs predominantly during spring (Figure 1a) and summer (Figure 1b). The lightning activity decreases during autumn (Figure 1c). Since most lightning activity occurs during spring and especially summer (Figure 1a,b), these were the seasons of interest in this study.

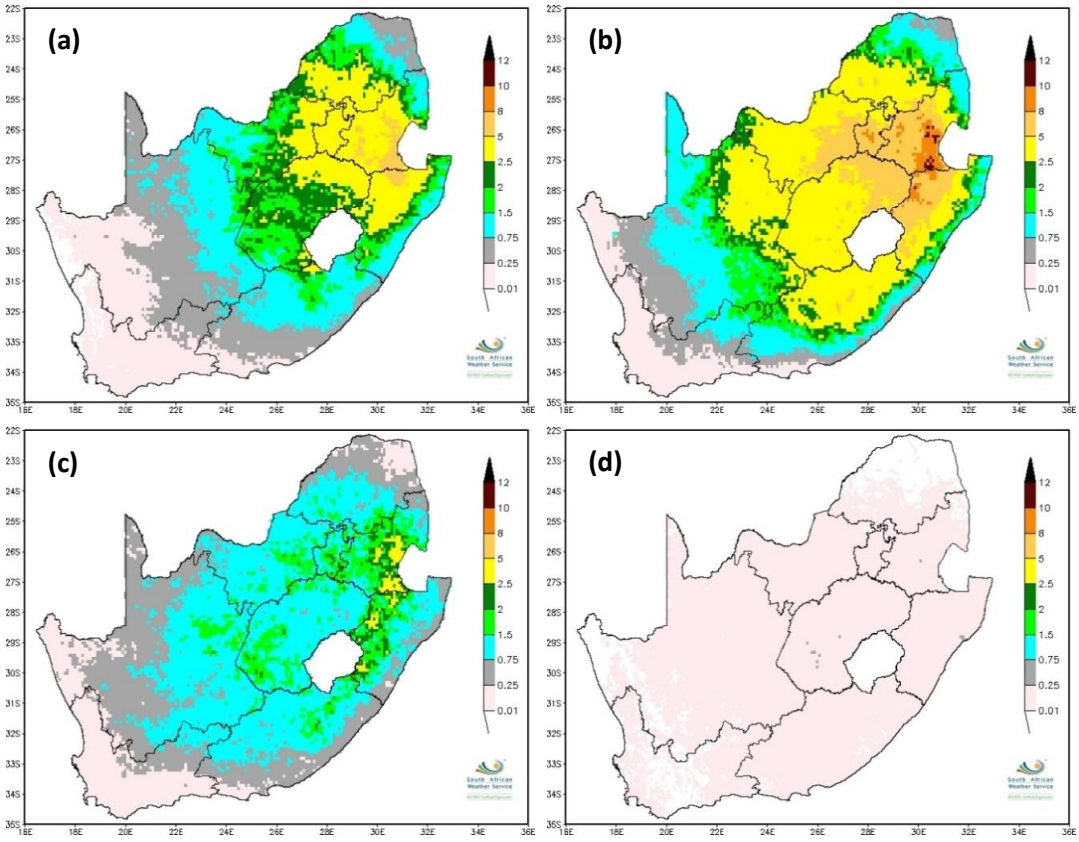

Fig. 1. The distribution of CG lightning ground flash densities (flashes per square kilometre per season) over South Africa for (a) September to November, (b) December to February, (c) March to May, and (d) June to August, during a nine-year period from 2006 to 2014.

Due to the hazardous nature of lightning, there is a need for prediction techniques to ensure the protection of people and property (McCaul et al., 2009; Lynn and Yair, 2010). Forecasting thunderstorms remains a challenge due to their small spatial and temporal scales, as well as uncertainty in the processes that govern thunderstorm development (Rajeevan et al., 2012). To predict lightning from a thunderstorm poses an even bigger challenge, since the processes that

govern the electrification of a thundercloud are still poorly understood (Shafer and Fuelberg, 2008). Many techniques have been developed to forecast lightning, ranging from nowcasting (0 to 2 hours ahead) and very short-range (2 to 12 hours ahead) up to short-range (12 to 72 hours ahead) forecasting time scales. Past studies have utilised lightning data from lightning detection networks (LDN), parameters from atmospheric soundings and numerical weather prediction (NWP) models to aid with lightning forecasts.

LDN are capable of detecting lightning strokes in real-time and the data measured by these networks could be utilised to aid in the nowcasting of thunderstorms. Many of these LDN networks however are designed to detect only CG lightning, and it has been shown that inconsistent relationships exist between CG lightning trends and thunderstorm nowcasting (Schultz et al, 2011). Total lightning sensors that can detect CG and cloud lightning have been found to be useful in the nowcasting of CG lightning strikes since lightning in the clouds mostly precedes CG lightning on the ground. MacGorman et al. (2011) showed that cloud lightning can precede the first CG lightning flash by up to an hour. This shows that total lightning sensors can be used for the nowcasting of thunderstorms or lightning but not for short-range forecasts. In South Africa the LDN detects mostly CG lightning which makes it less useful for nowcasting purposes.

Statistical techniques have been used extensively to aid in the prediction of thunderstorms and lightning (Shafer and Fuelberg, 2008). These techniques often rely on the connections between lightning occurrence and the parameters of the pre-storm environment (Rajeevan et al., 2012; McCaul et al., 2009). Many examples of such lightning prediction schemes exist (Livingston et al., 1996; Mazany et al., 2002; Benson, 2005; Lambert et al., 2005; Shafer and Fuelberg, 2006). Parameters are often derived from atmospheric soundings to predict lightning (Shafer and Fuelberg, 2008) however soundings in South Africa are typically only performed twice daily and at a limited amount of locations (de Coning et al., 2011). As a result, morning soundings are typically used to predict thunderstorms or lightning later in the day, which may result in inaccurate forecasts due to changes in atmospheric conditions later in the day or the site-specific sounding not being able to represent a large forecast domain (Shafer and Fuelberg, 2008). Due to the lack of soundings performed in South Africa, lightning cannot be forecasted using this approach.

With the advent of NWP models, many of the forecasting schemes started focusing on utilizing data from NWP models to predict thunderstorms. The latest NWP models provide accurate forecasts with a high spatial and temporal resolution (Shafer and Fuelberg, 2008), which results in the parameters related to lightning formation to be available over large domains for several hours ahead (McCaul et al., 2009). Parameters of the pre-storm environment usually derived from soundings can now be obtained from NWP models. The parameters can be obtained for the entire country and on an hourly basis for the next few days. Statistical prediction schemes that forecast the threat of lightning by relying on connections between lightning occurrence and parameters of the pre-storm environment has also been developed by making use of NWP data (Reap, 1994 ; Burrows et al., 2005 ; Bothwell, 2008 ; Shafer and Fuelberg, 2008 ; Rajeevan et al., 2012). These models were developed for specific regions and NWP models and the techniques used underestimated the occurrence of lightning. Whenever NWP data is utilised, users must always remember that NWP is a model representation of reality and the output of the model will only be as good as the performance of the model calculations for a specific day. Nevertheless, when the output of NWP models are used to developed a new statistical model

with data over a sufficiently long period, the new model should learn the typical errors and biases of the NWP model and consider it in the calculations.

This study, which is inspired by the work of Frisbie et al. (2009), presents a NWP model-based statistical lightning prediction scheme to forecast the probability of CG lightning over southern Africa for the austral summer months (September to February). Logistic regression techniques are used to select the most appropriate predictors from the Unified Model (UM) to include in the lightning prediction scheme. A rare-event logistic regression technique was chosen to develop equations that predict the probability of at least one lightning stroke per grid box between 07:00 and 22:00 UTC. The selection of predictors and the development of the equations are derived from CG lightning and NWP data from the austral summer days of 2011/12 and 2012/13. The scheme is validated over the independent austral summer days of 2013/14. The statistical scheme introduced in this study is named the lightning threat index (LTI) and aims to provide improved forecast guidance of lightning over southern Africa.

2. Data

2.1. Study area and period

The study domain covers the entire area of Lesotho, South Africa and Swaziland, as well as small areas of Botswana, Mozambique, Namibia and Zimbabwe and the surrounding oceans (Figure 2). The grey areas were excluded due to a reduction in the accuracy of lightning data in those areas. The domain was divided into a 0.5° x 0.5° grid that was used throughout this study.

This study focuses on the daily 15-hour period between 07:00 and 22:00 UTC. Most thunderstorms occur during the afternoon and evening in South Africa (De Coning et al., 2011), with a smaller frequency in the morning (Rouault et al., 2013). The 07:00 UTC starting time was selected to correspond with the availability of NWP data in an operational environment.

The UM prognosis for the austral summer days of 2011/12 and 2012/13 is used to select the most appropriate parameters to predict lightning (Section 3.4) and train the LTI statistical model (Section 3.5). Moreover, the austral summer was divided into spring (September to November (SON)) and summer (December to February (DJF)). A different LTI is developed for each of these seasons respectively. In early summer, the atmospheric circulation is generally extra-tropical with a conditionally unstable atmosphere over the summer rainfall areas, while in late summer the circulation is mostly tropical with a convectively unstable atmosphere (Dyson et al., 2015). Different model parameters and thresholds are needed to describe the atmospheric conditions during spring and summer, which necessitated two separate LTIs. The spring and summer LTIs were verified for the austral summer days of 2013/14.

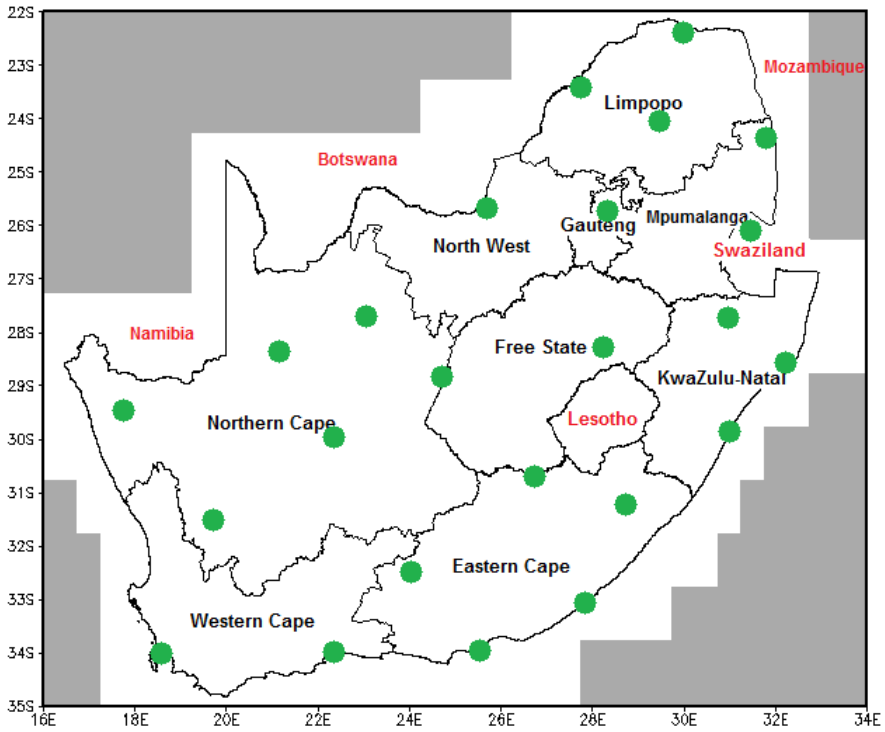

Fig. 2. The study domain over southern Africa with the location of lightning sensors in green. The grey areas were not considered in the analysis.

2.2. Lightning data

CG lightning data from the Southern Africa Lightning Detection Network (SALDN) of the South African Weather Service (SAWS) is utilized in this study. The SALDN became operational towards the end of 2005 and underwent a series of upgrades throughout the years. From 2011, the network consisted of 24 Vaisala CG lightning sensors (Gijben, 2012) as indicated by the green dots on Figure 2. The SALDN can detect lightning with a location accuracy of ~0.5 km and an estimated detection efficiency of ~90% over most of South Africa (Gijben, 2012). This means that the SALDN can detect at least 90% of all CG flashes and position them within 0.5 km.

Daily lightning data for 06:30 to 21:30 UTC is assigned to each of the 0.5° x 0.5° grid boxes over the study domain from where the number of CG strokes is counted. The NWP model data is available hourly and was considered representative of the lightning occurrences in the 30-minute interval before and after each NWP time step. Lightning is the predictand (observed) in this study, and when at least one lightning stroke occurred in a grid box during the study period, a value of 1 was assigned to the predictand. If no lightning strokes occurred in a grid box, the predictand was given a value of 0. This means that the study attempts to predict lightning occurrence (1) and non-occurrence (0) and does not attempt to predict the amount of lightning.

2.3. The Unified Model

The UM is the NWP model that was developed at the United Kingdom Meteorological Office (UKMO) (Davies et al., 2005). It is a fully compressible and non-hydrostatic model that follows the terrain and resolves many layers in the atmosphere with height-based vertical coordinates (Davies et al., 2005). The UKMO runs the UM on various resolutions, but also offers a global model that runs four times daily on a 40 km-resolution (Landman et al., 2012). The SAWS has been running a local version of the UM since 2006. Different versions of the UM run with different configurations (parametrization schemes, horizontal resolutions and with or without data assimilation) at the SAWS and uses initial and boundary conditions from the global model from the UKMO (Landman et al., 2012). In the study, data from the main operational model at the SAWS was utilized. This version of the model has a 12 km horizontal resolution and runs once daily to produce hourly forecasts on 38 vertical levels for 48 hours ahead. No data assimilation is available in this model and it produces a forecast for the entire southern Africa. The domain is bounded by 0° to 44°S and 10°W to 56°E.

As was mentioned in Section 2.1, the domain of South Africa was divided into a 0.5˚ x 0.5˚ grid that was used throughout the study. This grid is much coarser than the UM grid, but due to the long periods considered and the intensive calculations performed, the coarser grid was utilized to save on computation time. The 0.11˚ X 0.1112˚ resolution grid of the UM would have meant that all the calculations would have taken about 22 times longer than using the coarser grid. This work had to be done on a research server, and the increase in computation time would have made it unfeasible. As a result, the model output from the UM was re-gridded to a 0.5˚ x 0.5˚ grid before the calculations of the LTI was performed. A higher resolution would provide more detail in the LTI especially with the smaller forecast scales of thunderstorms. The LTI equations determined in this research will be directly applied to the higher resolution model output in the operational environment where supercomputers are available. Verification of this product in an operational environment is commencing.

Hourly output from the UM forecast is utilized from 07:00 to 22:00 UTC daily. The most favourable value of every parameter is identified per grid box within this 15-hour period and assigned to the parameter value of that specific day. For example, if the lowest value of the lifted index -7 occurred in a grid box at 12:00 UTC, then -7 is assigned as the lifted index value for that specific 15-hour forecast.

In total, 25 parameters were calculated from the UM data as possible predictors (forecast) of lightning. Six main types of parameters were considered in this study. They are convective available potential temperature (CAPE), precipitable water (PW), relative humidity (RH), lifted index (LI), lapse rates of equivalent potential temperature (Θ_e) and air temperature (T). Different variations from these six groups of parameters were considered. The 25 predictors were selected because they are useful in either lightning prediction studies or thunderstorm/rainfall development (see Table 1). Table 1 provides a summary of the candidate predictors used in this paper. Complete discussions of the 25 predictors are not provided in this paper, but references to authors who have utilized these parameters for lightning or thunderstorm prediction are listed in Table 1. The abbreviation of each predictor, a description and reference to other studies that utilized these predictors are included. SON indicates those parameters identified as the most appropriate to predict lightning in September to October and DJF in December to February.

Table 1. A list of considered NWP model predictors.

Abbreviation	Description (units)	Reference
$muCAPE_{0,3\ km}$	Largest CAPE obtained when each parcel between the surface and 3 km above ground level (AGL) is lifted from the level with the highest Θ_e (J kg^{-1})	Frisbie et al. (2009); Groenemeijer and Van Delden (2007)
$muCAPE_{1,6km}^{\mathbf{SON,\ DJF}}$	Largest CAPE obtained when each parcel between 1 km and 6 km AGL is lifted from the level with the highest Θ_e (J kg^{-1})	Frisbie et al. (2013)
$muCAPE_{low,300}$	Largest CAPE obtained when each parcel between the surface and lowest 300 hPa AGL is lifted from the level with the highest Θ_e (J kg^{-1})	Craven and Brooks (2004)
CAPE	CAPE obtained when a parcel is lifted from the surface (J kg^{-1})	Zepka et al. (2014)
Θ_e	Surface equivalent potential energy (Θ_e) (K)	Livingston et al. (1996); Zepka et al. (2014)
$\Theta_e\Gamma_{600}$	Θ_e lapse rate ($\Theta_e\Gamma$) at 600 hPa (K)	Frisbie et al. (2009)
$\Theta_e\Gamma_{850,400}^{\mathbf{DJF}}$	$\Theta_e\Gamma$ between 850 and 400 hPa (K)	Dyson et al. (2015)
$\Theta_e\Gamma_{850,500}$	$\Theta_e\Gamma$ between 850 and 500 hPa (K)	Dyson et al. (2015)
$\Theta_e\Gamma_{1,6\ km}$	$\Theta_e\Gamma$ between 1 and 6 km AGL (K)	Frisbie et al. (2009)
$\Theta_e\Gamma_{m10,m20}$	$\Theta_e\Gamma$ between -10˚C and -20˚C levels (K)	Frisbie et al. (2013)
$\Theta_e\Gamma_{700,500}^{\mathbf{SON}}$	$\Theta_e\Gamma$ between 700 and 500 hPa (K)	Zepka et al. (2013)
$SLI^{\mathbf{SON,DJF}}$	Surface LI when lifting the parcel from the surface to 500 hPa (°C)	Garreaud et al. (2014); Haklander and Van Delden (2003)
BLI	The best lifted index (BLI) obtained from the most unstable LI when each parcel is lifted between the surface and 700 hPa (°C)	Frisbie et al. (2013) Shafer and Fuelberg (2008)
$PW_{850,300}^{\mathbf{SON,DJF}}$	Total PW between 850 and 300 hPa (cm)	Dyson et al. (2015)
$PW_{700,400}$	Total PW between 700 and 400 hPa (cm)	Burrows et al. (2005)
$PW_{surf,100}$	Total PW between surface and 100 hPa (cm)	Burrows et al. (2005); Shafer and Fuelberg (2008)
RH_{m10}	RH at the -10˚C level ($RH = e/e_s$) (%)	Frisbie et al. (2009)

Abbreviation	Description (units)	Reference
$RH_{m12,m18}$	Mean RH in the -12˚C to -18˚C levels (%)	Frisbie et al. (2013)
$aveRH_{3,6\ km}^{DJF}$	Mean RH in 3 to 6 km AGL (%)	Frisbie et al. (2013)
$maxRH_{3,6\ km}$	Maximum RH in 3 to 6 km AGL (%)	Frisbie et al. (2013)
$minRH_{3,6km}^{SON}$	Minimum RH in 3 to 6 km AGL (%)	Frisbie et al. (2013)
T_{1p5m}	T at 1.5 meters above the ground (K)	Mazany et al. (2002)
T_{700}	T at the 700 hPa level (K)	Burrows et al. (2005)
$T_{850,700}^{SON,DJF}$	Mean T of pressure levels from 850 and 700 hPa (K)	Dyson et al. (2015)
$T_{500,300}$	Mean T of pressure levels from 500 and 300 hPa (K)	Dyson et al. (2015)

3. Model development

3.1. Selection of predictors

Not all of the 25 predictors listed in Table 1 were added to the final LTI model. There are six main groups of candidate predictors (CAPE, PW, RH, LI, Θ_e and T). Each of these groups consists of different variations of the predictors. The goal was to select the best preforming lightning predictor parameter from each of the six groups. One parameter from each group was selected. In order to achieve this goal, stepwise logistic regression techniques with Statistical Analysis System (SAS) and R software were utilized to select the most appropriate parameters.

R software was used to perform a stepwise binary logistic regression analysis (R Development Core Team, 2015) in order to select the most appropriate predictors from the six main groups to forecast the occurrence of lightning. The Akaike Information Criterion (AIC) is calculated using the R functions. The predictor with the lowest AIC value was selected to be the most appropriate for predicting the occurrence of lightning (Chaurasia and Harel, 2012; Posada and Buckley, 2004; Snipes and Taylor, 2014). This process was repeated with a full (backwards and forwards) stepwise logistic regression using Firth's Penalized Likelihood method in the SAS software (Firth, 1993; SAS Institute Inc., 2010) in order to confirm that the most appropriate parameter was identified for each parameter group. The same parameter was identified with both methods in all instances.

3.2. Statistical model

Binary logistic regression techniques are used to develop the new LTI. Logistic regression is often used to predict the probability of an event by means of a set of predictors (Kiezun et al., 2009) and it can be expressed by Equation 1, where p_i is the probability of the event as a function of m independent variables X, when i ranges from 1 to m. The regression coefficients,

$\hat{\alpha}$ and $\hat{\beta}$, are estimated from the dataset by means of the maximum likelihood method (Guns and Vanacker, 2012; Kleinbaum and Klein, 2010).

$$p_i = \frac{1}{1+\exp[-(\hat{\alpha}+\sum \hat{\beta}_i X_i)]} \qquad \text{i = 1 to m} \qquad (1)$$

In the development of the LTI by means of the logistic regression technique, the binary outcome of lightning occurrence was the dependent variable, while the six selected model predictors were the independent variables. Initial tests in the development of the LTI by means of ordinary logistic regression resulted in very low probabilities of lightning occurrence. King and Zeng (2001) showed that ordinary logistic regression (Equation 1) often underestimates the probabilities of rare events (Guns and Vanacker, 2012). This underestimation is due to the logistic regression favouring the larger amount of non-events (0's) compared to the smaller amount of events (1's) when developing a model. King and Zeng (2001) states that rare events in a dataset are classified as dozens to thousands of times more non-events compared to events, while Yap et al. (2014) considers a rare event to be when the events make up 5% or less of the data. In the datasets considered in this study, the SON dataset had approximately 20 times more non-events than events. The DJF dataset had approximately 34 times more non-events than events. In both datasets, the non-events made up less than 5% of the data.

Building on the work of King and Zeng (2001), Guns and Vanacker (2012) and Imai et al. (2009), the LTI is developed as follows:

1. Take all the events (1's or lightning occurrences) in the dataset and select a random sample of non-events (0's or no lightning occurrences) with equal size from the data.
2. Run the "Zelig" package in R to perform a rare-event logistic regression with the bias correction and addition of the correction term to the estimated probabilities.
3. Repeat the previous two steps 1 000 times by selecting a new sample of random non-events (0's or no lightning occurrences). The random samples of non-events are taken with repetition where the non-events of the previous sample are added back to the dataset and have the chance to be chosen again.

The 1 000 models produced by the procedure above were combined by averaging their intercept term and regression coefficients. This process is similar to the bootstrap aggregating technique that aims to improve any instability found in the estimation of the regression output (Kotsiantis et al., 2006). The average of the intercept terms and regression coefficients was added to Equation 1. Separate equations for SON and DJF were developed with this approach.

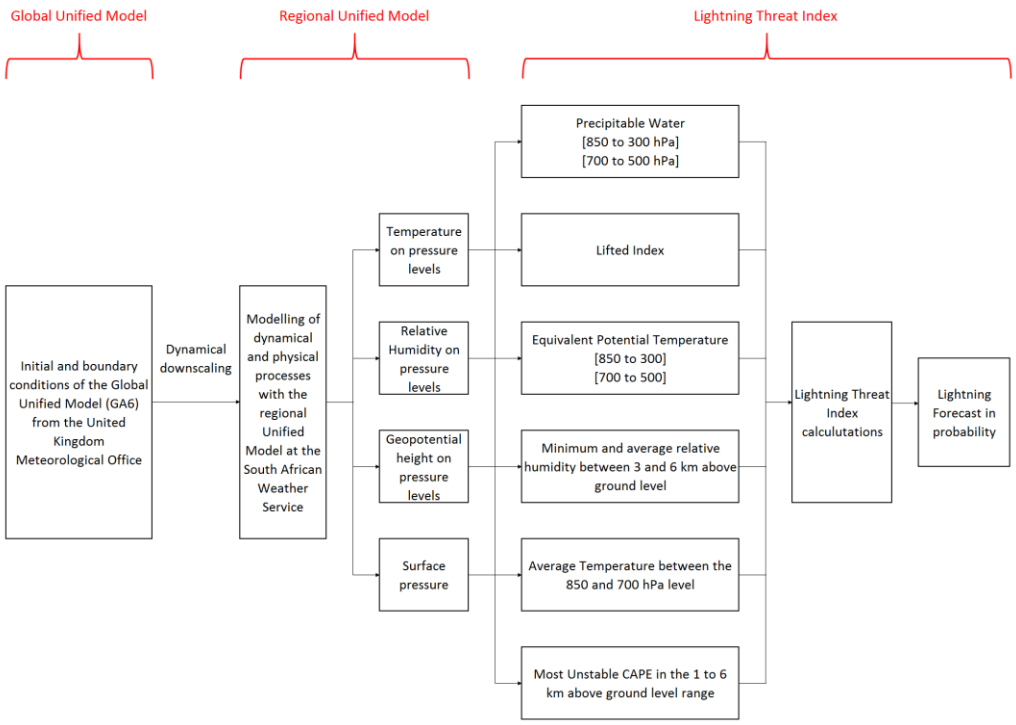

Fig. 3. The steps involved to produce the LTI forecast from the UM output.

Figure 3 shows a flow diagram of the steps to generate the LTI forecast. The first step in the process is to download the initial and boundary conditions from the Global UM, which is operational at the UKMO. This information is prepared for input into the UM at SAWS by means of dynamical downscaling. Once the initial and boundary conditions are given to the UM at SAWS, the modelling of dynamical and physical processes are performed. The UM produce a list of output parameters, from which temperature, relative humidity and geopotential height on all the pressure levels, as well as surface pressure are utilised. With these output parameters from the UM, the CAPE, LI, PW, RH, Θ_e and air temperature needed for the LTI can be calculated. The final output of the LTI provides a probability forecast of lightning occurrence.

3.3. Independent evaluations

The LTI forecast output gives the probability of lightning occurrences, and standard probabilistic evaluation techniques are considered. The Receiver Operating Characteristic (ROC) curve, reliability diagram and sharpness plot were produced with independent lightning and NWP model parameters datasets for SON of 2013 and DJF of 2013/14.

ROC curves are used to compare the sensitivity and specificity of a forecast over the range of all possible values (Florkowski, 2008). Sensitivity is the ability of a forecast to predict events, while specificity is the forecast's ability to predict the non-events (Robin et al., 2011). The "plot.roc" function in the R-package "pROC" was utilized to plot the ROC curves and area-under-the-curve (AUC) values for the validation of the LTI (Robin et al., 2011; R Development Core Team, 2015).

A reliability diagram is often used to determine the reliability of a probabilistic forecast by showing how well forecasted probabilities correspond to their observed frequency of occurrence (Weisheimer and Palmer, 2014). As such, the reliability diagram is created by plotting the observed relative frequencies against forecast probabilities, where the forecast probabilities are divided into bins (Bröcker and Smith, 2007). In this study, the "verify" and "reliability.plot" functions in the R package "verification" was utilized to plot the reliability diagrams for the validation of the LTI (NCAR Research Applications Laboratory, 2014; R Development Core Team, 2015).

The sharpness of a forecast is a measure of how forecast probabilities vary and is often presented on a sharpness diagram or sharpness histogram, which displays the relative frequencies of occurrence for probability intervals (bins). Sharpness diagrams often accompany reliability plots (Callado et al., 2013). In this study, the "verify" and "reliability.plot" functions in the R package "verification" was modified to plot the sharpness diagrams separately from the reliability plots for the validation of the LTI (NCAR Research Applications Laboratory, 2014; R Development Core Team, 2015).

4. Results and discussion

4.1. Selection of predictors

The approach discussed in Section 3.1 was used to select the six top-performing NWP model parameters from the 25 candidate predictors (Table 1). A stepwise logistic regression analysis with SAS and R software was performed to select the most appropriate predictor from each main group of parameters. However, due to the large amount of output produced by the regression analysis, the results will not be shown in this paper. The final six predictors selected from the regression output, which predicts the occurrence of lightning the best during SON and DJF, are shown in Table 1.

With the exception of Θ_e lapse rates and RH, the same parameters were identified for both seasons. The min$RH_{3,6\ km}$ and $\Theta_e\Gamma_{700,500}$ performed the best in SON, while the ave$RH_{3,6\ km}$ and $\Theta_e\Gamma_{850,400}$ performed the best in DJF. The CAPE, Θ_e lapse rates and LI predictors provide information on the updrafts that supply a thundercloud with the necessary hydrometeors in the cloud's charge separation zone (Murugavel et al., 2014; Singh and O'Gormon, 2015; Bright et al., 2005; Madhulatha et al., 2013; Houston and Wilhelmson, 2012; Cummings and Pickering, 2013; Huntrieser et al., 2011; Kuo, 1966). The moisture needed to create favourable conditions for thunderstorm development and the hydrometeors for lightning formation is provided by the RH and PW predictors (Burrows et al., 2005; Duplika and Reuter, 2006; Berdeklis and List, 2001; Xiong et al., 2006). Surface heating, which is responsible for the convective processes that result in atmospheric instabilities, is represented by the temperature predictor (Bharatdwaj, 2006; Price, 2013; Williams, 1992, 1994, 2009; Reeve and Toumi, 1999; Markson and Price, 1999; Price and Asfur, 2006; Markson, 2007). The six predictors for SON and DJF can be utilized in the development of the LTI.

4.2. Statistical model

The dataset for SON consisted of 137 864 observations, of which 3 945 were lightning events and 133 919 were non-events. The rare-event logistic analysis described in Section 3.4 was applied to this data set using the six identified parameters (Table 1). Five of the six predictors were significant (p-value less than 0.05 and Z-value not between -1.96 and +1.96). The $muCAPE_{1,6\ km}$ was not significant and was removed from the analysis. The rare-event logistic regression procedure was repeated again for SON, but this time without $muCAPE_{1,6\ km}$. The rare-event logistic analysis was also performed for DJF where 128 562 observations were made. Some 6 118 of these observations were events and 122 444 were non-events. In DJF, all six parameters were significant (see Table 2). SON was constructed without using $muCAPE_{1,6\ km.}$ Table 2 also shows the regression coefficients (Coef), standard error on Coef (SE), odds ratio (Odds), maximum parameter value (MPV) and measure of parameter importance (MPI) for the intercept term and the six model parameters. The MPV is the largest value of a parameter (or smallest for parameters where a negative value is important) in the dataset. The MPI is the MPV value multiplied by the regression coefficient (Coef) and it is a measure to determine the most important variables in the regression analysis (Guns and Vanacker, 2012; Vanwalleghem et al., 2008). The intercept term of the regression analysis does not have a MPV and MPI value. All of the predictors have regression coefficients that are significant with p-values < 0.05 and absolute values of $z \geq 1.96$ (not shown).

Table 2. Output from the rare-event logistic regression for SON and DJF in brackets.

	Coef	**SE**	**Odds**	**MPV**	**MPI**
Intercept	81.597771 (96.373973)	1.8112 (1.8712)	6.612E+35 (1.985E+42)		
$muCAPE_{1,6\ km}$	(-0.000362)	(2.378E-05)	(0.9996)	(6757.06)	(-2.45)
$PW_{850,300}$	1.858302 (1.649851)	0.0536 (0.0349)	6.4196 (5.2081)	3.90 (4.80)	7.25 (7.92)
SLI	-0.301027 (-0.274410)	0.0070 (0.0089)	0.7401 (0.7600)	-17.04 (-13.63)	5.13 (3.74)
$\Theta e\Gamma_{700,500}$ **($\Theta e\Gamma_{850,400}$)**	-0.199863 (-0.253844)	0.0059 (0.0039)	0.8189 (0.7758)	-20.83 (-33.60)	4.16 (8.53)
$minRH_{3,6\ km}$ **($aveRH_{3,6\ km}$)**	0.021953 (0.022844)	0.0008 (0.0007)	1.0222 (1.0231)	104.10 (105.60)	2.29 (2.41)
$T_{850,700}$	-0.308811	0.0065	0.7343	298.81	-92.28

	(-0.364683)	(0.0066)	(0.6944)	(299.80)	(-109.33)

All the parameters for SON and DJF have very low SE values, which mean that the model fits the data well. In both seasons, $T_{850,700}$ plays the principal role in the regression model, as can be seen from the high MPI values. This confirms the importance of surface heat as a trigger for the development of thunderstorms (Bharatdwaj, 2006). The relative importance of the other parameters varies according to season. In SON, $PW_{850,300}$ is the second-most important parameter, followed by the SLI, $\Theta_e\Gamma_{700,500}$ and $minRH_{3,6\ km}$. In DJF, $\Theta_e\Gamma_{850,400}$ is the second-most important parameter in the model, followed by the $PW_{850,300}$, SLI, $muCAPE_{1,6\ km}$ and $aveRH_{3,6\ km}$.

The intercept term and the regression coefficients listed in Table 2 were added to Equation 1 to produce a new LTI for SON and DJF respectively. The LTI for SON (DJF) is given by Equation 2 (3) and provides the probability that lightning will occur (values between 0 and 1).

$$LTI_{SON} = \frac{1}{1 + \exp[-[\hat{\alpha} + \widehat{\beta_1}(PW_{850,300}) + \widehat{\beta_2}(SLI) + \widehat{\beta_3}(\theta_e\Gamma_{700,500}) + \widehat{\beta_4}(minRH_{3,6km}) + \widehat{\beta_5}(T_{850,700}))]} \quad (2)$$

where: $\hat{\alpha} = 81.597771$ $\widehat{\beta_1} = 1.858302$ $\widehat{\beta_2} = -0.301027$ $\widehat{\beta_3} = -0.199863$
$\widehat{\beta_4} = 0.021953$ $\widehat{\beta_5} = -0.308811$

$$LTI_{DJF} = \frac{1}{1 + \exp[-[\hat{\alpha} + \widehat{\beta_1}(PW_{850,300}) + \widehat{\beta_2}(SLI) + \widehat{\beta_3}(\theta_e\Gamma_{700,500}) + \widehat{\beta_4}(minRH_{3,6km}) + \widehat{\beta_5}(T_{850,700}))]} \quad (3)$$

where: $\hat{\alpha} = 96.373973$ $\widehat{\beta_1} = -0.000362$ $\widehat{\beta_2} = 1.649851$ $\widehat{\beta_3} = -0.274410$
$\widehat{\beta_4} = -0.253844$ $\widehat{\beta_5} = 0.022844$ $\widehat{\beta_6} = -0.364683$

4.3. Independent verification

During SON, (Figure 4a) and DJF (Figure 4b), the ROC curves approach the top left corner of the diagram and fall above the no-skill diagonal line. The curves show that the LTI has a high sensitivity (or high hit rate) and high specificity (or low false alarm rate) across all the possible probability ranges. The high sensitivity indicates that the LTI correctly predicts the lightning events, while the high specificity shows that the LTI correctly predicts the lightning non-events. One can conclude from the ROC curves in Figure 4a and Figure 4b that the LTI discriminates well between lightning occurrences and non-occurrences and that the LTI forecasts are very accurate. The ROC curves in Figure 4a and Figure 4b are also accompanied by an AUC value that represents the overall performance of the LTI. During SON, the AUC value was 0.927 and

during DJF, it was 0.899. Since an AUC value of 1.0 represents a perfect forecast and a value of ≤ 0.5 represents a worthless forecast (Fawcett, 2006), the LTI performed well, as the AUC values are close to 1.0. The AUC was slightly higher for SON than for DJF.

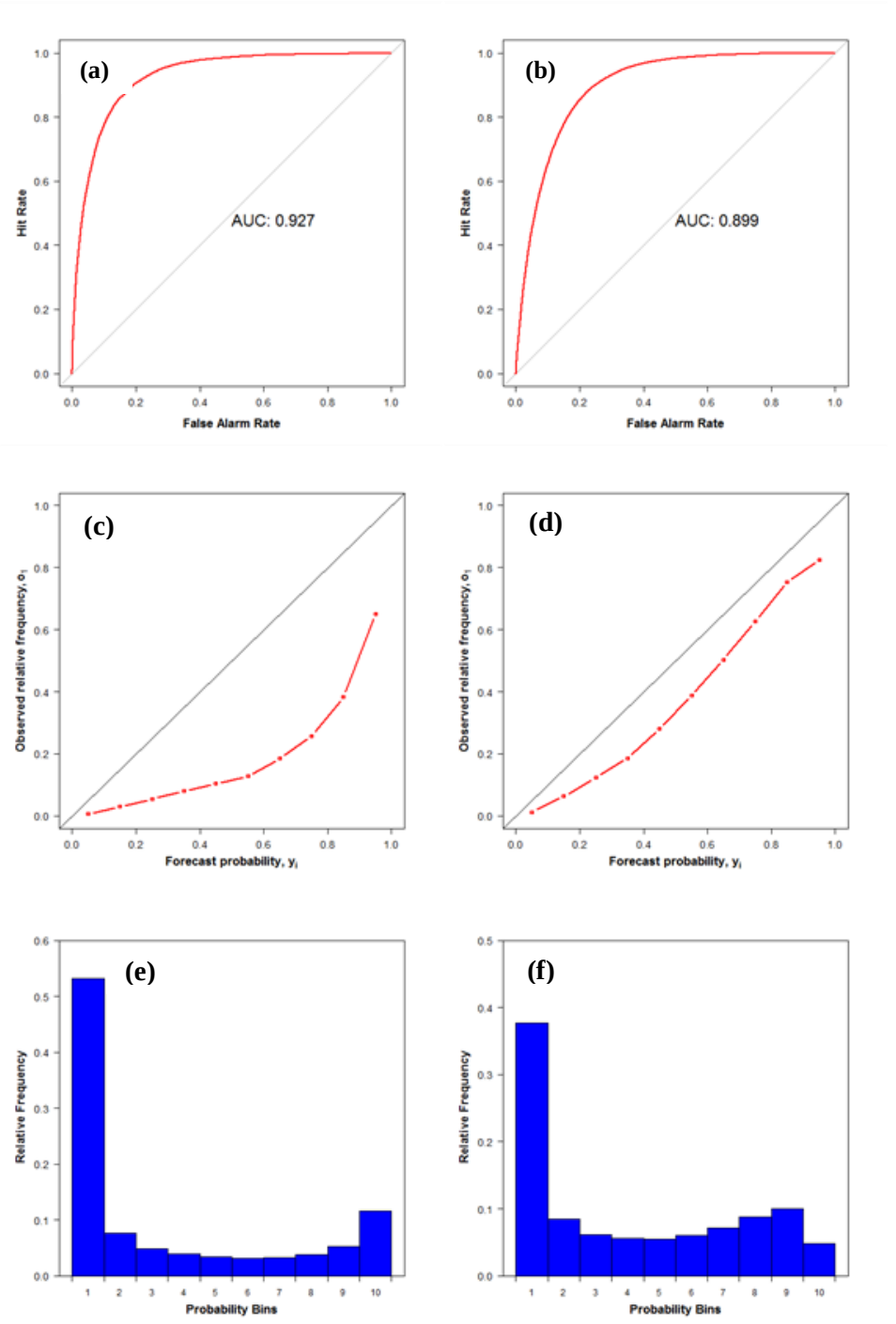

Fig. 4. The ROC curves, together with the AUC values for SON (a) and DJF (b), the reliability diagrams during SON (c) and DJF (d) and sharpness diagrams during SON (e) and DJF (f) for the UM LTI forecasts against lightning observations. The SON months of 2013 and DJF months of 2013/14 were utilized.

During SON (Figure 4c), the LTI over-forecasts the observed frequency of lightning occurrence. This is evident from the curve falling under the diagonal line. For the first bin, the forecast is reliable, but it becomes increasingly more unreliable towards the 8th bin. It then starts moving back to the diagonal line. The reliability diagram for DJF (Figure 4d) presents much better than that of SON. The LTI only slightly over-forecasts the observed frequency of lightning occurrence that is evident from the curve falling just under the diagonal line. The forecast starts out to be reliable in the first probability bin, moves away slightly from the diagonal line up to the 6th probability bin, from where it gradually moves back to the diagonal line. This shows that the LTI forecast is reliable during DJF and much more reliable than the LTI forecasts for SON.

The LTI forecasts have good sharpness during SON (Figure 4e) and DJF (Figure 4f) since the sharpness diagrams have a U-shaped distribution. The most forecasts are made in the first probability bin for both SON and DJF, from where they decrease to the sixth probability bin for SON and the fifth bin for DJF. From here, it increases again. During DJF, more forecasts are made in the ninth probability bin compared to the tenth.

5. Case examples

The independent verifications (Section 4.3) indicated that the LTI is skilful in predicting lightning for both SON and DJF over South Africa. Two case examples are included here to provide an example of how the operational LTI appears and to illustrate its performance on a daily basis. These case examples were randomly chosen based on days where significant amounts of lightning occurred over South Africa.

5.1. 20 December 2013

On 20 December 2013, lightning occurred over large areas of southern and north-western South Africa (Figure 5b). Most of the lightning occurred over the Northern Cape and Eastern Cape and extended to the surrounding territories (See Figure 2 for location map). The LTI probability forecast is shown in Figure 5a. The highest lightning probability is >90% over parts of Northern Cape and Eastern Cape. Lightning was also predicted for large areas of Botswana and Namibia, which could not be verified due to the inaccuracy of the SALDN over these countries. Most of the lightning over South Africa occurred in the areas where the LTI probability exceeded 60%. Some lightning was observed in areas where the probability was less than 60% (North West and Gauteng). In these areas, the number of lightning stokes was relatively low (< 15). No lightning materialized over the western Free State, where high lightning probability was predicted. Over Eastern Cape and North West, the area of maximum lightning prediction was slightly misplaced. Nevertheless, the lightning probability forecast provided good indications of the actual occurrence of lightning over South Africa on this day.

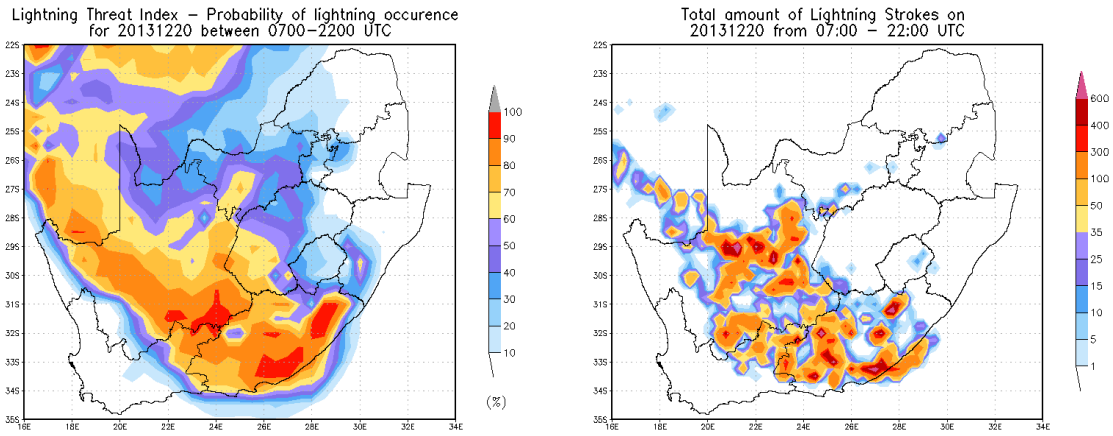

Fig. 5. The LTI forecast (a) and occurrence of lightning (b) for 20 December 2013 between 07:00 and 22:00 UTC.

5.2. 15 January 2014

On 15 January 2014, the LTI forecasted lightning to occur over a large area of the central interior of South Africa, extending into Namibia (Figure 6a). When the LTI forecast is compared with the observed lightning (Figure 6b), one can see that the LTI performed well on this day. Most of the lightning activity was observed in the areas where the probabilities ≥ 60%. The remaining lightning activity occurred in areas with lower probabilities, especially when the probabilities were ≥ 40%. Some areas over Mpumalanga and KwaZulu-Natal were over-forecasted, while the small isolated storms over the southern parts of the Northern Cape were missed.

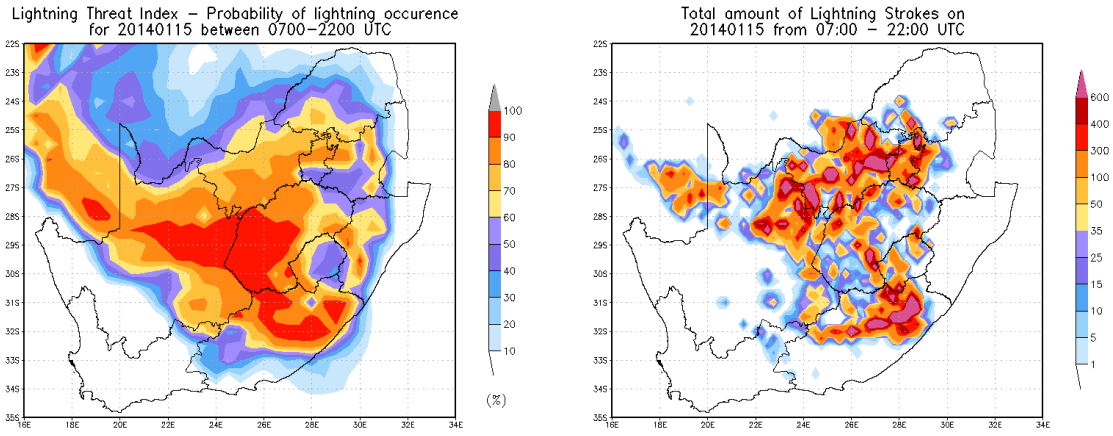

Fig. 6. The LTI forecast (a) and occurrence of lightning (b) for 15 January 2014 between 07:00 and 22:00 UTC.

6. Summary and Conclusions

CG lightning data and UM data for austral summer days during 2011/12 and 2012/13 were utilized to develop a statistical scheme to predict the daily occurrence of lightning over southern Africa for the period 07:00 to 22:00 UTC. The austral summer months were divided into two seasons, spring (September to November) and summer (December to February), since the atmospheric conditions are different between the seasons.

Before the LTI could be developed, it was necessary to identify the most appropriate NWP model parameters capable of predicting the occurrence of lightning over southern Africa. Stepwise logistic regression techniques were used to identify six parameters from a list of 25 candidate predictors. One parameter from the six groups, CAPE, LI, PW, RH, Θ_e, and air temperature was selected. The top-performing predictors for SON and DJF were the $muCAPE_{1,6\ km}$, SLI, $PW_{850,300}$, and $T_{850,700}$. In addition to these predictors, the $minRH_{3\_6km}$ ($aveRH_{3\_6km}$) and $\Theta_e\Gamma_{850,400}$ ($\Theta_e\Gamma_{700,500}$) were identified for SON (DJF). These predictors were used in the development of the new LTI.

The new LTI was developed by means of a rare-event logistic regression technique. Since there is typically a larger number of non-events (or no lightning occurrences) compared to events (lightning occurrences), the rare-event logistic regression technique was utilized so that the regression procedure does not favour the larger number of non-events that result in the output of low probabilities. A separate LTI was created for SON and DJF and gives the probability that at least one lightning stroke can be expected between 07:00 and 22:00 UTC at any particular grid point.

A probabilistic evaluation over the independent period of the 2013 SON and 2013/14 DJF showed that the LTI forecasts have a high sensitivity and specificity for both SON and DJF. The LTI is not so reliable during SON, since it over-forecasts the occurrence of lightning quite significantly, but during DJF, the LTI forecast is reliable, only slightly over-forecasting lightning activity. Lastly, the results also show that the LTI produces sharp forecasts during both SON and DJF. The reason for the LTI being more reliable during DJF can be due to the nature of convection, which develops in different atmospheric conditions during the 2 seasons. In early summer, the atmospheric circulation is generally extra-tropical with a conditionally unstable atmosphere. Convection is often surface heat driven and develops from favourable local conditions. In late summer, the circulation is near tropical with a convectively unstable atmosphere and convection results from large-scale synoptic circulation systems (Dyson et al., 2015).

In this paper, a LTI was developed for South African conditions. In this paper, a LTI was developed for South African conditions. South Africa has its own unique geographical and synoptic circulation patterns, which requires a unique product for predicting lightning. Most of the interior of South Africa rises to 1500 m and more above mean sea level resulting in significant modification of the thermodynamics of the atmosphere. The amount of moisture available for convective development is much less than areas located close to sea level. (Dyson et al., 2015). Lightning prediction parameters identified elsewhere in the world can therefore not be applied directly to South Africa and bespoke methods need to be developed. Note how for both seasons CAPE was identified is being the least important parameter for predicting lightning while low level temperature lapse rates and moisture were identifies as the most important

parameters. In fact, it was found that CAPE was not significant in the LTI regression model (p-value > 0.05) during SON and was not used during the spring season. This deviates from findings elsewhere in the world where CAPE is often considered important (Burrows et al., 2005; Livingston et al., 1996; Shafer and Fuelberg, 2006, 2008; Zepka et al., 2014). These results emphasize the importance of identifying parameters specific to local conditions. Furthermore, different NWP models differ in dynamics, physics and initial conditions. Statistical lightning models therefore needs to be trained for specific NWP models as was done in this research. The LTI was developed with the operational NWP employed at SAWS and this provides a unique insight into the behaviour of this model for the identification of convection and lightning.

This paper contributes to the development of statistical models in predicting rare events by proposing a rare-event binary logistic regression approach to produce a probability based lightning forecast. This eliminates the model favouring the large amount of non-events (no lightning) that is often found with ordinary binary logistic regression techniques. Moreover, previous studies have focused on using multiple linear regression and ordinary binary logistic regression techniques to develop a model. The multiple linear regression techniques have been found to over-estimate the occurrence of lightning while ordinary binary logistic regression under-estimates the occurrence of lightning.

The LTI will be a useful tool to operational weather forecasters or sectors interested in lightning forecasts, to provide guidance early in the morning on the areas where lightning can be expected during the day. It can ultimately contribute to society by aiding with timely warnings of lightning or thunderstorms to protect humans, animals and property. Users of the product should however keep in mind that the lightning forecast is only a model of reality and its performance depends on the accuracy of the NWP model for a particular day. The evaluations of the LTI shows that in most cases the model performs well and will be a valuable additional tool that can be used by forecasters and various sectors of society.

Acknowledgments and Data

Thank you to the South African Weather Service for supplying the valuable lightning and Unified Model data utilized in this study.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Benson, R.P., 2005. Predicting Lightning Strikes for the Enhancement of Fire Weather Forecasts. Preprints, 6th Symposium on Fire and Forest Meteorology, Canmore, Alberta. Amer. Meteor. Soc. (Available online at: <https://ams.confex.com/ams/6FireJoint/techprogram/paper_97679.htm>).

Berdeklis, P., List, R., 2001. The Ice Crystal–Graupel Collision Charging Mechanism of Thunderstorm Electrification. J. Atmos. Sci. 58, 2751–2770. http://dx.doi.org/10.1175/1520-0469(2001)058<2751:TICGCC>2.0.CO;2.

Bharatdwaj, K., 2006. Physical Geography: Hydrosphere. Discovery Publishing House, New Delhi, India (356 pp.).

Bhavika, B., 2007. The influence of terrain elevation on lightning density in South Africa (MSc Thesis). University of Johannesburg, South Africa.

Blumenthal, R.E., Trengrove, E., Jandrell, I.R., Saayman, G., 2012. Lightning medicine in South Africa. South African Med. J. 102, 625–626. http://dx.doi.org/10.7196/SAMJ.5219.

Bright, D.R., Wandishin, M.S., Jewell, R.E., Weiss, S.J., 2005. A physically based parameter for lightning prediction and its calibration in ensemble forecasts. Preprints, Conf. on Meteo. Appl. of Lightning Data. San Diego, CA, Amer. Meteor. Soc., 4.3. (Available online at: <http://ams.confex.com/ams/pdfpapers/84173.pdf>).

Bröcker, J., Smith, L.A., 2007. Increasing the Reliability of Reliability Diagrams. Weather. Forecast. 22, 651–661. http://dx.doi.org/10.1175/WAF993.1.

Burrows, W.R., Price, C., Wilson, L.J., 2005. Warm Season Lightning Probability Prediction for Canada and the Northern United States. Weather. Forecast. 20, 971–988. http://dx.doi.org/10.1175/WAF895.1.

Callado, A., Escribà, P., García-Moya, J.A., Montero, J., Santos, C., Santos-Muñoz, D., Simarro, J., 2013. Ensemble Forecasting. Clim. Chang. Reg. Responses. Dr Pallav Ray (Ed.), InTech, http://dx.doi.org/10.5772/55699.

Chaurasia, A., Harel, O., 2012. Using AIC in multiple linear regression framework with multiply imputed data. Heal. Serv. Outcomes Res. Methodol. 12, 219–233. http://dx.doi.org/10.1007/s10742-012-0088-8.

Christian, H.J., Blakeslee, R.J., Boccippio, D.J., Boeck, W.L., Buechler, D.E., Driscoll, K.T., Goodman, S.J., Hall, J.M., Koshak, W.J., Mach, D.M., Stewart, M.F., 2003. Global frequency and distribution of lightning as observed from space by the Optical Transient Detector. Journal of Geophysical Research: Atmospheres, 108(D1), pp. ACL 4-1–ACL 4-15. http://dx.doi.org/10.1029/2002JD002347.

Craven, J.P., Brooks, H.E., 2004. Baseline climatology of sounding derived parameters associated with deep moist convection. Natl. Wea. Dig. 28, 13–24.

Cummings, K.A., Pickering, K.E., 2013. Lightning Flash Rate and Chemistry Simulation of Tropical Island Convection Using a Cloud-resolved Model (M.S. Thesis). University of Maryland, USA.

Davies, T., Cullen, M.J.P., Malcolm, A.J., Mawson, M.H., Staniforth, A., White, A.A., Wood, N., 2005. A new dynamical core for the Met Office's global and regional modelling of the atmosphere. Q. J. R. Meteorol. Soc. 131, 1759–1782. http://dx.doi.org/10.1256/qj.04.101.

de Coning, E., Poolman, E., 2011. South African Weather Service operational satellite based precipitation estimation technique: applications and improvements. Hydrol. Earth Syst. Sci. 15, 1131–1145. http://dx.doi.org/10.5194/hess-15-1131-2011.

de Coning, E., Koenig, M., Olivier, J., 2011. The combined instability index: a new very-short range convection forecasting technique for southern Africa. Meteorol. Appl. 18, 421–439. http://dx.doi.org/10.1002/met.234.

Dupilka, M.L., Reuter, G.W., 2006. Forecasting Tornadic Thunderstorm Potential in Alberta Using Environmental Sounding Data. Part II: Helicity, Precipitable Water, and Storm Convergence. Weather. Forecast. 21, 336–346. http://dx.doi.org/10.1175/WAF922.1.

Dyson, L.L., van Heerden, J., Sumner, P.D., 2015. A baseline climatology of sounding-derived parameters associated with heavy rainfall over Gauteng, South Africa. Int. J. Climatol. 35, 114–127. http://dx.doi.org/10.1002/joc.3967.

Fawcett, T., 2006. An Introduction to ROC Analysis. Pattern Recogn. Lett. 27, 861–874. http://dx.doi.org/10.1016/j.patrec.2005.10.010.

Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80, 27–38. http://dx.doi.org/10.1093/biomet/80.1.27.

Florkowski, C.M., 2008. Sensitivity, Specificity, Receiver-Operating Characteristic (ROC) Curves and Likelihood ratios: Communicating the Performance of Diagnostic Tests. Clin. Biochem. Rev. 29 (Suppl 1), S83-87.

Frisbie, P.R., Colton, J.D., Pringle, J.R., Daniels, J.A., Ramey Jr., J.D., Meyers, M.P., 2009. Lightning Prediction by WFO Grand Junction using Model Data and Graphical Forecast Editor Smart Tools. Preprints, 4th Conf. on Meteo. Appl. of Lightning Data, Phoenix, AZ. Amer. Meteor. Soc. (Available online at: <https://ams.confex.com/ams/89annual/techprogram/paper_149101.htm>).

Frisbie, P., Colton, J., Pringle, J., Daniels, J., Meyers, M., 2013. A forecasting methodology that uses moisture parameters to pinpoint locations of potential lightning, report, Central Region Technical Attachment Number 13-01, Grand Junction, CO, NOAA/National Weather Service. NOAA/National Weather Service, Grand Junction, CO.

Garreaud, R.D., Nicora, M.G., Bürgesser, R.E., Ávila, E.E., 2014. Lightning in Western Patagonia, J. Geophys. Res.: Atmospheres, 119, 4471–4485. https://doi.org/10.1002/2013JD021160.

Gijben, M., 2012. The lightning climatology of South Africa. S. Afr. J. Sci. 108, 44–53. https://doi.org/10.4102/sajs.v108i3/4.740.

Groenemeijer, P.H., van Delden, A., 2007. Sounding-derived parameters associated with large hail and tornadoes in the Netherlands. Eur. Conf. Sev. Storms 83, 473–487. http://dx.doi.org/10.1016/j.atmosres.2005.08.006.

Guns, M., Vanacker, V., 2012. Logistic regression applied to natural hazards: rare event logistic regression with replications. Nat. Hazards Earth Syst. Sci. 12, 1937–1947. https://doi.org/10.5194/nhess-12-1937-2012.

Haklander, A.J., Van Delden, A., 2003. Thunderstorm predictors and their forecast skill for the Netherlands. Atmos. Res. 67, 273–299. http://dx.doi.org/10.1016/S0169-8095(03)00056-5.

Holle, R.L., 2008. Annual rates of lightning fatalities by country. Preprints, 20th International Lightning Detection Conference, Tucson, AZ (Available online at: < http://www.vaisala.com/Vaisala%20Documents/Scientific%20papers/Annual_rates_of_lightning_fatalities_by_country.pdf >).

Houston, A.L., Wilhelmson, R.B., 2012. The Impact of Airmass Boundaries on the Propagation of Deep Convection: A Modeling-Based Study in a High-CAPE, Low-Shear Environment. Mon. Weather Rev. 140, 167–183. http://dx.doi.org/10.1175/MWR-D-10-05033.1.

Hunt, H.G.P., Nixon, K.J., Jandrell, I.R., 2014. Establishing a methodology to investigate LDN median error ellipses used as corroborating evidence for a lightning event at a specific geographic location. Electric Power Systems Research. 113, 104–114. http://dx.doi.org/10.1016/j.epsr.2014.02.033.

Huntrieser, H., Schlager, H., Lichtenstern, M., Stock, P., Hamburger, T., Höller, H., Schmidt, K., Betz, H.-D., Ulanovsky, A., Ravegnani, F., 2011. Mesoscale convective systems observed during AMMA and their impact on the NO_x and O_3 budget over West Africa. Atmos. Chem. Phys. 11, 2503–2536. http://dx.doi.org/10.5194/acp-11-2503-2011.

Imai, K., King, G., Lau, O., 2009. Zelig: Everyone's statistical software. R Package. Version 3(5) (Available online at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.118.7412&rep=rep1&type=pdf>).

Kiezun, A., Lee, I.T.A., Shomron, N., 2009. Evaluation of optimization techniques for variable selection in logistic regression applied to diagnosis of myocardial infarction. Bioinformation 3, 311–313.

King, G., Zeng, L., 2001. Logistic Regression in Rare Events Data. Polit. Anal. 9, 137–163.

Kleinbaum, D.G., Klein, M., 2010. Logistic Regression: A Self-Learning Text. third ed., Springer, New York (702 pp.).

Kohn, M., Galanti, E., Price, C., Lagouvardos, K., Kotroni, V., 2011. Nowcasting thunderstorms in the Mediterranean region using lightning data. Atmos. Res. 100, 489–502. http://dx.doi.org/10.1016/j.atmosres.2010.08.010.

Kotsiantis, S.B., Kanellopoulos, D., Zaharakis, I.D., 2006. Bagged averaging of regression models, artificial intelligence applications and innovations. 3rd IFIP Conference on Artificial Intelligence Applications and Innovations. Athens, Greece. http://dx.doi.org/10.1007/0-387-34224-9_7.

Kruger, A.C., 2007. Climate of South Africa, Precipitation. Report WS47, 1–41, South African Weather Service, Pretoria, South Africa.

Kuo, H.L., 1966. On the Dynamics of Convective Atmospheric Vortices. J. Atmos. Sci. 23, 25–42. http://dx.doi.org/10.1175/1520-0469(1966)023<0025:OTDOCA>2.0.CO;2.

Lambert, W., Wheeler, M., Roeder, W., 2005. Objective lightning forecasting at Kennedy space center and Cape Canaveral air force station using cloud-to-ground lightning surveillance system data. Preprints, Conference on Meteorological Applications of Lightning Data, San Diego, CA. Amer. Meteor. Soc. (Available online at: <https://ams.confex.com/ams/pdfpapers/85944.pdf >).

Landman, S., Engelbrecht, F.A., Engelbrecht, C.J., Dyson, L.L., Landman, W.A., 2012. A short-range weather prediction system for South Africa based on a multi-model approach. Water SA 38, 765–774.

Lang, T.J., Miller, L.J., Weisman, M., Rutledge, S.A., Barker, L.J., Bringi, V.N., Chandrasekar, V., Detwiler, A., Doesken, N., Helsdon, J., Knight, C., Krehbiel, P., Lyons, W.A., Macgorman, D., Rasmussen, E., Rison, W., Rust, W.D., Thomas, R.J., 2004. The Severe Thunderstorm Electrification and Precipitation Study. Bull. Am. Meteorol. Soc. 85, 1107–1125. http://dx.doi.org/10.1175/BAMS-85-8-1107.

Livingston, E.S., Nielsen-Gammon, J., Orville, R.E., 1996. A Climatology, Synoptic Assessment, and Thermodynamic Evaluation for Cloud-to-Ground Lightning in Georgia: A study for the 1996 Summer Olympics. Bull. Am. Meteorol. Soc. 77, 1483–1495. http://dx.doi.org/10.1175/1520-0477(1996)077<1483:ACSAAT>2.0.CO;2.

Lynn, B., Yair, Y., 2010. Prediction of lightning flash density with the WRF model. Adv. Geosci. 23, 11–16. http://dx.doi.org/10.5194/adgeo-23-11-2010.

MacGorman, D.R., Apostolakopoulos, I.R., Lund, N.R., Demetriades, N.W., Murphy, M.J. and Krehbiel, P.R., 2011. The Timing of Cloud-to-Ground Lightning Relative to Total Lightning Activity. Monthly Weather Review, 139, 3871–3886. http://dx.doi.org/10.1175/MWR-D-11-00047.1.

Madhulatha, A., Rajeevan, M., Venkat Ratnam, M., Bhate, J., Naidu, C.V, 2013. Nowcasting severe convective activity over southeast India using ground-based microwave radiometer observations. J. Geophys. Res. Atmos. 118, 1–13. http://dx.doi.org/10.1029/2012JD018174.

Markson, R., 2007. The Global Circuit Intensity: Its Measurement and Variation over the Last 50 Years. Bull. Am. Meteorol. Soc. 88, 223–241. http://dx.doi.org/10.1175/BAMS-88-2-223.

Markson, R., Price, C., 1999. Ionospheric potential as a proxy index for global temperature. Atmos. Res. 51, 309–314. http://dx.doi.org/10.1016/S0169-8095(99)00015-0.

Mazany, R.A., Businger, S., Gutman, S.I., Roeder, W., 2002. A Lightning Prediction Index that Utilizes GPS Integrated Precipitable Water Vapor. Weather. Forecast. 17, 1034–1047. http://dx.doi.org/10.1175/1520-0434(2002)017<1034:ALPITU>2.0.CO;2.

McCaul, E.W., Goodman, S.J., LaCasse, K.M., Cecil, D.J., 2009. Forecasting Lightning Threat Using Cloud-Resolving Model Simulations. Weather. Forecast. 24, 709–729. http://dx.doi.org/10.1175/2008WAF2222152.1.

Murugavel, P., Pawar, S.D., Gopalakrishan, V., 2014. Climatology of lightning over Indian region and its relationship with convective available potential energy. Int. J. Climatol. 34, 3179–3187. http://dx.doi.org/10.1002/joc.3901.

National Center for Atmospheric Research (NCAR) Research Applications Laboratory, 2014. Verification: Weather Forecast Verification Utilities, R package version 1.41. (Available at http://CRAN.R-project.org/package=verification).

Posada, D., Buckley, T.R., 2004. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53, 793–808. http://dx.doi.org/10.1080/10635150490522304.

Price, C., 1993. Global surface temperatures and the atmospheric electrical circuit. Geophys. Res. Lett. 20, 1363–1366. http://dx.doi.org/10.1029/93GL01774.

Price, C., Asfur, M., 2006. Can lightning observations be used as an indicator of upper-tropospheric water vapor variability? Bull. Am. Meteorol. Soc. 87, 291–298. http://dx.doi.org/10.1175/BAMS-87-3-291.

Price, C.G., 2013. Lightning Applications in Weather and Climate Research. Surv. Geophys. 34, 755–767. http://dx.doi.org/10.1007/s10712-012-9218-7.

R Development Core Team, 2015. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria.

Rajeevan, M., Madhulatha, A., Rajasekhar, M., Bhate, J., Kesarkar, A., Rao, B.V.A., 2012. Development of a perfect prognosis probabilistic model for prediction of lightning over south-east India. J. Earth Syst. Sci. 121, 355–371. http://dx.doi.org/10.1007/s12040-012-0173-y.

Reeve, N., Toumi, R., 1999. Lightning activity as an indicator of climate change. Q. J. R. Meteorol. Soc. 125, 893–903. http://dx.doi.org/10.1002/qj.49712555507.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 1–8. http://dx.doi.org/10.1186/1471-2105-12-77.

Rouault, M., Roy, S.S., Balling, R.C., 2013. The diurnal cycle of rainfall in South Africa in the austral summer. Int. J. Climatol. 33, 770–777, http://dx.doi.org/10.1002/joc.3451.

SAS Institute Inc., 2010. SAS/STAT 9.22 User's Guide, SAS Institute Inc., Cary, NC.

Shafer, P.E., Fuelberg, H.E., 2006. A Statistical Procedure to Forecast Warm Season Lightning over Portions of the Florida Peninsula. Weather. Forecast. 21, 851–868. http://dx.doi.org/10.1175/WAF954.1.

Shafer, P.E., Fuelberg, H.E., 2008. A Perfect Prognosis Scheme for Forecasting Warm-Season Lightning over Florida. Mon. Weather Rev. 136, 1817–1846. http://dx.doi.org/10.1175/2007MWR2222.1.

Schultz, C.J., Petersen, W.A., Carey, L.D., 2011. Lightning and Severe Weather: A Comparison between Total and Cloud-to-Ground Lightning Trends. Weather. Forecast. 26, 744–755. http://dx.doi.org/10.1175/WAF-D-10-05026.1.

Singh, M.S., O'Gorman, P.A., 2015. Increases in moist-convective updraught velocities with warming in radiative-convective equilibrium. Q. J. R. Meteorol. Soc. 141, 2828–2838. http://dx.doi.org/10.1002/qj.2567.

Snipes, M., Taylor, D.C., 2014. Model selection and Akaike Information Criteria: An example from wine ratings and prices. Wine Econ. Policy 3, 3–9. http://dx.doi.org/10.1016/j.wep.2014.03.001.

Trengrove, E., Jandrell, I.R., 2011. Strategies for understanding lightning myths and beliefs. Int. J. Res. Rev. Appl. Sci. 7, 287–294.

Tyson, P.D., 1986. Climatic change and variability in southern Africa. Oxford University Press, Cape Town (220 pp.).

Vanwalleghem, T., van Den Eeckhaut, M., Poesen, J., Govers, G., Deckers, J., 2008. Spatial analysis of factors controlling the presence of closed depressions and gullies under forest:

Application of rare event logistic regression. Geomorphology 95, 504–517. http://dx.doi.org/10.1016/j.geomorph.2007.07.003.

Weisheimer, A., Palmer, T.N., 2014. On the reliability of seasonal climate forecasts. J. R. Soc. Interface 11. http://dx.doi.org/10.1098/rsif.2013.1162.

Williams, E.R., 1992. The Schumann Resonance: A Global Tropical Thermometer. Science 256, 1184–1187. http://dx.doi.org/10.1126/science.256.5060.1184.

Williams, E.R., 1994. Global Circuit Response to Seasonal Variations in Global Surface Air Temperature. Mon. Weather Rev. 122, 1917–1929. http://dx.doi.org/10.1175/1520-0493(1994)122<1917:GCRTSV>2.0.CO;2.

Williams, E.R., 2009. The global electrical circuit: A review. Atmos. Res. 91, 140–152. http://dx.doi.org/10.1016/j.atmosres.2008.05.018.

Xiong, Y.J., Qie, X.S., Zhou, Y.J., Yuan, T., Zhang, T.L., 2006. Regional Responses of Lightning Activities to Relative Humidity of the Surface. Chinese J. Geophys. 49, 311–318. http://dx.doi.org/10.1002/cjg2.840.

Yap, B.W., Rani, K.A., Rahman, H.A.A., Fong, S., Khairudin, Z., Abdullah, N.N., 2014. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. Proceedings, First International Conference on Advanced Data and Information Engineering (DaEng-2013). 285, 13–22. http://dx.doi.org/10.1007/978-981-4585-18-7_2.

Zepka, G.S., Pinto Jr., O., Saraiva, A.C. V, 2014. Lightning forecasting in southeastern Brazil using the WRF model. Atmos. Res. 135–136, 344–362. http://dx.doi.org/10.1016/j.atmosres.2013.01.008.