climate

MDPI

Article

The Role of Instability Indices in Forecasting Thunderstorm and Non-Thunderstorm Days across Six Cities in India

Kopal Arora [1,*], Kamaljit Ray [1], Suresh Ram [1] and Rajeev Mehajan [2]

[1] Ministry of Earth Sciences, Delhi 110059, India
[2] Science and Engineering Research Board, Department of Science and Technology, New Delhi 110070, India
* Correspondence: kopal.arorasharma@hiscox.com; Tel.: +44-744-507-5732

Abstract: Thunderstorms are one of the most damaging natural hazards demanding in-depth understanding and prediction. These convective systems form in an unstable environment which is quantitatively expressed in terms of instability indices. These indices are studied over six locations across the Indian landmass in an attempt to predict thunderstorm activity on any given day. A combination of multiple regression, logistic regression, and range analysis provides new insight into the prediction of these storms. A supervised machine learning-based logistic regression model is developed in this study for thunderstorm prediction over Patna and can be further extended for operational forecasting of Thunderstorms over the region. Critical thresholds for the instability indices are determined over the considered locations providing valuable insight into the domain of Thunderstorm prediction

Keywords: thunderstorms; multiple linear regression; logistic regression; instability indices; thunderstorm forecasting; supervised machine learning

Citation: Arora, K.; Ray, K.; Ram, S.; Mehajan, R. The Role of Instability Indices in Forecasting Thunderstorm and Non-Thunderstorm Days across Six Cities in India. *Climate* **2023**, *11*, 14. https://doi.org/10.3390/cli11010014

Academic Editor: Nir Y. Krakauer

Received: 1 November 2022
Revised: 27 December 2022
Accepted: 28 December 2022
Published: 4 January 2023

1. Introduction

Thunderstorms are natural hazards that cause extensive damage to life and property. Thus, a profound understanding of their occurrence is of high interest to the communities facing these. Apart from this, the topic is of high interest to the weather and climate community in particular and risk assessment communities in general.

Thunderstorms are produced by cumulonimbus clouds (CB) and are often accompanied by lightning, squalls, hail, and/or blowing dust. They often develop in the presence of synoptic weather systems. Several studies have analyzed various synoptic systems and revealed three essential requirements for thunderstorm development. These requirements are (1) the presence of instability in the atmosphere, (2) Moisture inclusion in the lower tropospheric levels, and (3) the lifting mechanism needed to release the potential instability and generate convection [1]. Such requirements are mathematically expressed in terms of thermodynamic indices or instability indices. Instability indices represent the potential for convection using mathematical calculations based on temperature and moisture data at different pressure levels. Many Studies have analyzed different instability indices at various places and have suggested threshold values for the prediction of thunderstorms.

1.1. Earlier Studies

The eastern and northeastern parts of India are severely affected by Thunderstorms during pre-monsoon season which are also known as Norwesters [2,3]. These locations include Assam, Orissa, Jharkhand, Bihar, West Bengal, and other parts of northeastern states. Thunderstorms over these regions are called Kal-Baisakhi. Focusing on such thunderstorms, many studies have attempted to predict their occurrence over different regions using suitable indices. One of the oldest studies on thunderstorms and the associated indices developed a method for forecasting monsoon thunderstorm/dust storm (DS) activity

over New Delhi region by using. Showalter Index (SI) and convective condensation level (CCL) [4]. A year later a study suggested the empirical threshold value of the Total Totals Index (TTI) as 48 for Thunderstorm occurrence [5]. Another study based on 10 years (1971-80) data from Delhi and Jodhpur has found that the higher TTI values offer more favorable conditions for the occurrence of thunderstorms from March to June [6]. This study also found that dew point at 850 hPa, i.e., lower level moisture incursion plays a crucial role in the occurrence of thunderstorms. The computed values of the SI and Lifted Index (LI) during summer monsoon months based on the data of seven years (1958–60, 1963–66) show that LI is a better predictor than SI over Delhi [7]. Considering the thunderstorm activity over Lucknow it was found that the SI of −4 or less, mean relative humidity of 45% or more, below the level of 850 Pa and dew point higher than the climatological normal, presents a supportive situation for thunderstorm activity [8]. Later a thunderstorm intelligence prediction system (TIPS) was developed using the principles associated with convection using the 1200 UTC sounding, for a particular station to provide a 12 h forecast [9]. Considering various thermodynamic parameters, for instance, CAPE (Convective available potential energy) Total perceptible water (TPW) in combination with conventional charts to study thunderstorm occurrence, CAPE and TPW profiles provide vital clues for its formation and that changes in the atmosphere are at times available 6–12 h before the thunderstorm occurrence. This study was carried over Delhi, and Jodhpur [10]. Another study done over Jodhpur again found that a combination of critical threshold values of the instability indices- SI, LI, Cross Total Index (CTI), Vertical Totals (VT), Jefferson's modified index, George K-Index along with lifting (whether available or not) gives a good indication of thunderstorm occurrence [11]. To forecast thunderstorms over Delhi, two objective methods were developed [12] using 15 different types of instability indices. The first method is graphical whereas the second one uses a multiple regression approach. The second method uses 9 predictors to forecast thunderstorms in probabilistic terms. The study concluded that the multiple regression approach is more suitable for the operational forecasting of thunderstorm occurrence or non-occurrence over Delhi. Similarly, an attempt was made to develop an expert system for thunderstorm forecasting in pre-monsoon season over northwest India and suggested that exceedance of critical value of four parameters viz SI (>0 to <−9), equivalent potential temperatures (θ_e) at 850 hPa (>340 °K), a meridional component of wind at 850 hPa (>10 knots) and dew point at 850 hPa (>13 °C) can cause thunderstorm [13]. Such studies indicate the importance of instability indices in predicting thunderstorms over a few focused regions across North India, for instance, Delhi and Jodhpur. Despite establishing the importance of using instability indices for thunderstorm prediction, these studies cover limited regions. A lot of work was done by scientists in Germany [14], Colorado [15,16], the Netherlands [17], etc. However, in recent years it is mostly thunderstorm forecasts based on parameters derived from NWP mesoscale models. Prediction of rainfall and cloud properties associated with thunderstorms is a very useful tool for identifying the thunderstorm probable zones in advance but for Nowcasting, the use of thermodynamic stability indices is still more skillful. Thus, this paper attempts to assess and explore the importance of the instability indices across six cities of India thereby filling in the missing gap in the literature in terms of spatial coverage. Another aspect this study offers is thunderstorm forecasting using machine learning algorithms over all the considered cities. This advanced approach provides fresh insight into thunderstorm prediction and simplistic equations to estimate thunderstorm probability.

1.2. Present Study

Data from the multi-institute program on Severe Thunderstorm Observations and Regional Modelling (STORM) was used for this paper. The program was launched to popularize and create general awareness of thunderstorms in public through an e-forum for discussions among scientists, teachers, students, and laymen. Later this program was upgraded and established as the SAARC-STORM program jointly undertaken by 8 South Asian countries under the South Asian Association for Regional Cooperation (SAARC).

The SAARC STORM program [18–20] also complements the Severe Weather Forecast Demonstration Project (SWFDP) of WMO. The data generated under this program has generated large-scale interest in fueling research among the scientific community and broadening the perspectives of operational meteorologists and researchers [21].

Most studies have been carried out in Northwest, Eastern, and Northeastern India while the research in the other regions is very limited. The present study thus considers the widespread data from the SAARC STORM project [21,22]. The project covers the comprehensive dataset over major cities Bangaluru, Delhi, Panjim, Jodhpur, Patiala, and Patna located across India (Figure 1). The present study is based on synoptic and upper air observations (Radiosonde/Radiowind ascent) at 00 UTC of these stations for the period 2013–2015 for the summer months of April, May, and June. An attempt has been made to develop thunderstorm forecasting models throughout the locations across India based on thermodynamic indices. A major advantage of using 00 UTC data is that the successful model developed using it will provide a longer time frame to issue warming and take action.

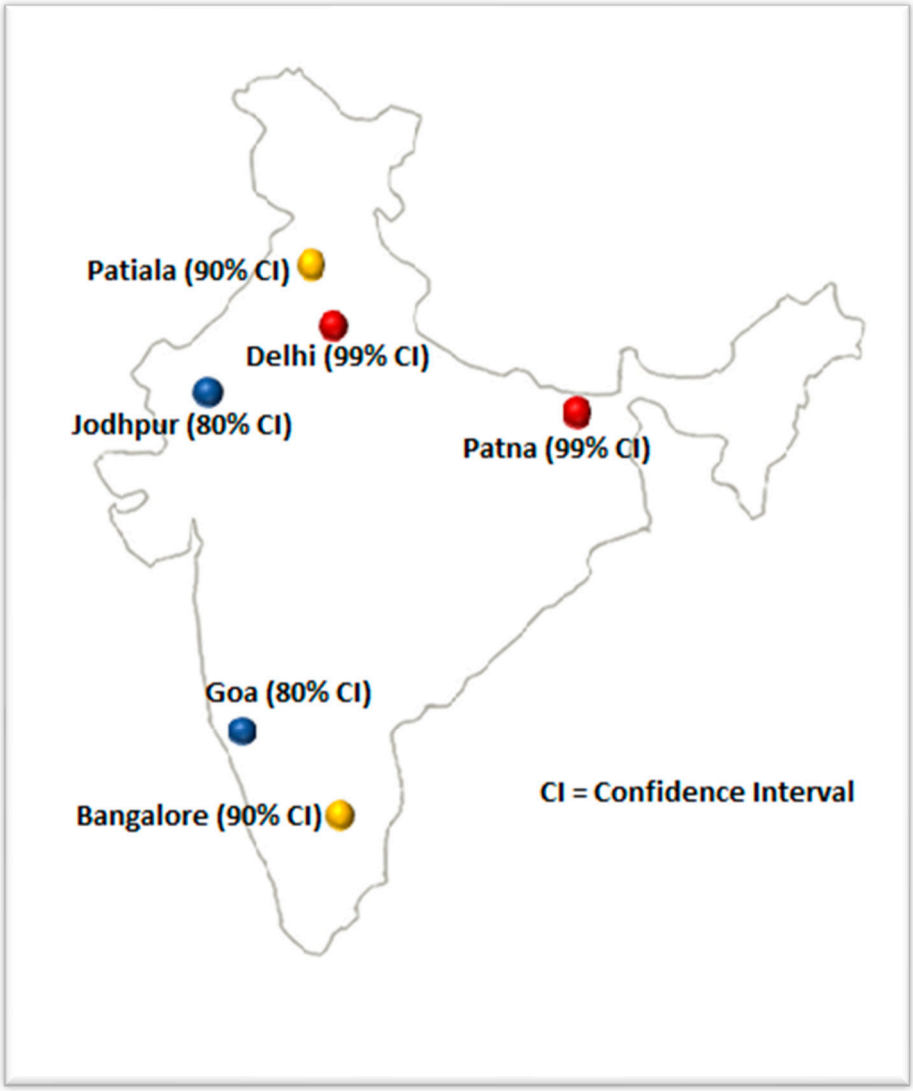

Figure 1. Considered locations and their corresponding confidence interval for the, Thunderstorm (TS), TS-Day prediction.

Based on the above data set the study aims to (1) explore the role of various instability indices in determining thunderstorm occurrence/non-occurrence, (2) figure out which index/indices act more prominently in determining thunderstorm occurrence, and (3) to

develop a model using multiple and logistic regression as applicable. The logistic method used here uses a supervised machine-learning algorithm. This advanced approach imparts fresh insight into the subject.

There are two types of datasets used in this study. First, the observation-based thunderstorm data and second, instability indices over the same locations of the sounding data. The observation-based thunderstorm occurrence and non-occurrence data have been obtained from SAARC project archives and the information about the indices has been taken from the radiosonde sounding database of the University of Wyoming [23].

The next section of this paper discusses "Data and Methodology" which is followed by the "Results and Discussion" section and finally, the findings are reported and discussed in the "Conclusions" section.

2. Data and Methodology

2.1. Data

2.1.1. SAARC Project Data

To understand how various indices help in the Nowcast of thunderstorm, the observations-based data in six cities across India are considered for this study. The thunderstorm data used in this study was collected as part of the SAARC STORM project. The program aimed to build an operational early warning system for catastrophic Thunderstorms over different parts of India. thunderstorm database is a comprehensive dataset including various parts of India, Nepal, Bhutan, and Bangladesh under the SAARC STORM Project of the Ministry of Earth Sciences [21,22]. During the project campaign, the thunderstorm development was recorded regularly, and a bulletin was issued twice daily. These datasets are from the Doppler weather radar observations at the considered locations.

2.1.2. Atmospheric Sounding

Atmospheric sounding [23], also known as upper air profiling, quantifies the vertical properties of the atmospheric column. The columnar quantities of temperature, wind speed, and direction are measured at various pressure heights. Based on these soundings (00:00 UTC), thermodynamical indices are computed as explained in the methodology section further. The sounding indices [23] considered here are *K-Index*, *SI*, *CAPE*, Convective Inhibition (*CIN*), *LI*, and *TTI*.

Both the datasets, SAARC STORM Project and the sounding indices are considered for each day for 3 years 2013, 2014 and 2015 during the thunderstorm prominent months of India, April, May and June.

Following are the indices used in the study.

Lifted Index (*LI*)

The *LI* is used to assess low-level parcel instability in the troposphere. The *LI* is computed as the difference between the observed environmental temperature at 500 mb (*Tenv* | 500) and that of the parcel temperature (*Tparcel* | 500) at the same level. *LI* is calculated according to the following relation.

$$LI = Tenv, 500 - Tparcel, 500$$

Thus, a negative *LI,* represents unstable troposphere and more positively buoyant parcel acceleration from the Planetary Boundary Layer (PBL).

Advantages: *LI* is relatively easy to determine using the skew T chart as it relies only on three sounding inputs namely- temperature and dew point of the boundary layer and the environmental temperature at 500 hPa.

Limitations: *LI* only estimates instability in one level of the troposphere, unlike *CAPE* which assesses instability in the entire troposphere and is most relevant in the barotropic troposphere and fails to serve the purpose when shallow polar air mass moves in PBL and for forecasting precipitation during winters. *LI* is not useful for situations like dry layers

and or inversions. The index does not consider vertical wind shear which is a vital element in a severe convective environment.

Showalter Index (*SI*)

SI is a measure of thunderstorm potential and severity. The *SI* is useful when a shallow cool layer of air below 850 hPa conceals greater convective potential. The *SI* is similar to *LI* except that *LI* considers just the lowest 1000 hPa layer, whereas *SI* considers parcel lifting from 850 hPa to 500 hPa. At 500 hPa the parcel temperature is subtracted from the ambient (sounding) temperature and is given by the following relation.

$$SI = T500 - TL$$

where *TL* is the parcel temperature (in °C) which is lifted from 850 hPa to 500 hPa dry adiabatically. The negative *SI* values represent instability and thus a higher likelihood of convective events like thunderstorms.

Advantages: *SI* is relatively easy to compute and is thus often employed to study environmental instability.

Limitation: If the top of the moist layer falls below 850 hPa, *SI* under-represents the instability. The index is useful at locations with low elevations (~1000 hPa) and fails to effectively represent instability at high elevations and does not consider vertical wind shear which affects the storm potential.

K-Index

K-Index also known as George's index is a measure of the convective potential. The index is a combination of Vertical Totals (*VT*) and lower tropospheric moisture characteristics. *VT* is a representative of the lapse rate between 850 hPa and 500 hPa while the moisture parameters are the 850 hPa dew point and 700 hPa dew point depression ($T_{dd700} = T_{700} - T_{d700}$).

K-index, *KI*, is the sum of *VT* and T_{dd}, i.e.,

$$KI = (T850 - T500) + (Td850 - Tdd700)$$

The index is specifically useful for identifying convective and heavy rain-producing environments. The index takes into account the vertical distribution of both moisture and temperature and does not require a skew-T diagram. A higher value of the K-Index indicates a higher potential for convection and thence thunderstorm activity.

It is a useful tool to diagnose the thunderstorm potential. It does not require a Skew-T diagram and the index computation is solely based on the vertical distribution of the temperature and moisture.

K-Index may not pick up a capping inversion that prevents thunderstorm from developing and cannot be used to determine the severity of thunderstorms. Even when moisture is lacking, *VT* could be very high thus contributing to a high index value. In such cases, the index will be unrealistically too unstable. The index is most suitable for flat to low elevation areas and does not work for high elevations and changes seasonally and with the location. Thus, the index is more suitable for forecasting within a deep layer of maritime tropical air and not in a differential advection situation where an elevated mixed layer advects over maritime tropical air.

Convective Available Potential Energy (*CAPE*)

CAPE is an index that is indicative of instability through the depth of the atmosphere. The index thus quantifies how strong updrafts will be if a convective system develops. On a skew-T diagram, it represents the positive area on the skew-T sounding. The positive area is the one where the parcel's (theoretical) temperature is greater than that of environmental temperature at each pressure level in the troposphere. The parcel's theoretical temperature is the lapse rate a parcel would acquire if it is raised from the lower PBL. *CAPE* exists between the conditionally unstable layer of the troposphere, the free convective layer, and

the equilibrium level,. Thus, *CAPE* is given by integrating the vertical local buoyancy of a parcel between these two layers and is given by,

$$CAPE = \int_{z=hf}^{z=he} g\left(\frac{Tv, parcel - Tv, environment}{Tv, environment}\right) dz$$

where, *hf* = height of the level of free convection, *he* is the height of the equilibrium level (neutral buoyancy), *Tv*, *parcel* represents the virtual temperature of the specific parcel (K), *Tv*, *the environment* is the virtual temperature of the environment and *g* represents acceleration due to gravity.

Any positive value of *CAPE* (>0) represents atmospheric instability and the possibility of thunderstorm development. Integration is the work done by the buoyant force- work is done against gravity and thus represents the excess energy that can become kinetic energy. Thus, the higher the *CAPE*, the higher will be the possibility of convection and thunderstorm development

Even if *CAPE* is high but the low-level capping inversion is not broken, the storm would not occur. Additionally, *CAPE* magnitude could rise and fall rapidly over time and space.

Convective Inhibition (*CIN*)

CIN represents the amount of energy that will prevent an air parcel from rising from the surface to the level of free convection.

CIN is calculated by using the measurement of temperature and pressure from weather balloons. For a parcel lifted from the *surface* to the *level of free convection (LFC)* with virtual temperature, *Tv*, *parcel*, and the environmental virtual temperature of *Tv*, *env*,

$$CAPE = \int_{Surface}^{LFC} g\left(\frac{Tv, parcel - Tv, environment}{Tv, environment}\right) dz$$

CIN is also referred to as negative buoyant energy A low *CIN* is associated with a higher possibility of the development of thunderstorm because *CIN* hinders or even inhibits thunderstorm development. *CIN* can be weakened by daytime heating, from lifting associated with low-level convergence storm-generated outflows, upper-level divergence, and other lifting mechanisms.

On a skew T diagram, *CIN* represents the negative area, i.e., parcel is cooler than the surrounding. Thus, *CIN* is a practically significant indicator of how much an updraft is suppressed and gives valuable information about the thunderstorm potential when used in conjunction with *CAPE*.

The index is mainly applicable in barotropic environments or the warm regions of mid-latitude storms. The index is limited to PBL-based convection only and is meaningless if there is no *CAPE*, i.e., *CAPE* should be positive to break the inversion cap. Thus, *CIN* is meaningful when considered along with *CAPE*.

Totals Totals Index (*TTI*)

TTI is a severe weather index and is used to *assess storm strength*. The index is a combination of Vertical Totals (*VT*) and Cross Totals *(CT*) and is thus defined as the sum of the two indices viz- *VT* and *CT*.

$$VT = T850 - T500$$

i.e., *VT* is the temperature difference between 850 mb and 500 mb while *CT* is the 850 mb dew point minus the 500 mb temperature, i.e.,

$$CT = Td850 - T500$$

Adding the two gives,

$$TTI = VT + CT = T850 - Td850 - 2 * T500$$

A higher *TTI* value signifies a higher potential for thunderstorm occurrence. The index is comprehensive as it captures the vertical and cross total of the environment and works best for flat areas in low to moderate elevation. However, the index does not assess wind shear and *CAPE* directly, which is a storm-significant parameter and may not pick up capping inversion that prevents storms from developing.

2.2. Methodology

2.2.1. Multiple Linear Regression

Multiple linear regression is an algebraic equation with each term either a constant or a product of a constant and a variable. Multiple linear regressions equation is expressed as,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

where $\beta_1, \beta_2, \ldots, \beta_n$ represents the least square estimate and β_0 the constant.

Variables here are the seven instability indices namely, *TTI*, *KI*, *CAPE*, *CIN*, *LI*, and *SI*, and '*y*' is the predictand, thunderstorm day, obtained by a linear combination of the predictors.

The value of predictand, thunderstorm day, is '1' if thunderstorm occurs on any given day and is '0' if it does not.

2.2.2. Logistic Regression and Supervised Machine Learning

Logistic regression is a statistical classification method used to fit a regression model when the response variable is binary (0 = No, 1 = Yes). The method produces an estimate of the probability of an event occurring. For $p\,(Z) > 0.5$, thunderstorm occurrence on any given day, based on the values of the instability indices provided, is highly likely. Similarly, for $p\,(Z) < 0.5$, the occurrence of thunderstorm is highly unlikely.

The method uses a logistic function called S-shaped curve or the sigmoid function. This function provides a real value number between 0 and 1. The logistic function is as follows:

$$P(Z) = \frac{1}{(1 + e^{-Z})}$$

where '*e*' is the exponential function, '*Z*' represents a linear boundary function (a line) that separates the input into two categories. Thus, '*Z*', also known as the *logit* function, is of the form, βx or $(ax + b)$, where x represents the input variable, '*a*' denotes the coefficient and '*b*' the bias. Thus, the equation can be re-written as,

$$P(x) = \frac{1}{\left(1 + e^{-(ax+b)}\right)} = \frac{1}{\left(1 + e^{-\beta x}\right)}$$

Rearranging and taking the log of the above equation gives:

$$\log\left(\frac{P(x)}{1 + P(x)}\right) = \beta x$$

The logistic regression model is trained by fitting the best value of coefficient and bias to the decision boundary function, *Z* which is done by a maximum likelihood estimation

method. This technique computes the parameters by finding the parameter value that maximizes the likelihood of making a given observation given the parameter.

The data were collected using Microsoft Excel (Microsoft Corporation) and were parsed using an analytic solver add-in (Front line systems). The objective here was to minimize the sum of the log-likelihood values by varying the coefficient values (β: a_0, a_1, ...). For the non-linear optimization, GRG non-linear algorithm is utilized here. The GRG non-linear algorithm is provided with a small arbitrary value to begin the training. The algorithm makes small changes in these values. This step gradually varies the coefficient values until it minimizes the log-loss and optimizes the coefficient values of the logit (Z).

The observed dataset is divided into a training set and a test set. The training set comprises 80% of the total dataset and the test set considers the remaining 20% of the data which is used for the model validation.

2.2.3. Significance Testing

Analysis of Variance (ANOVA) was applied to determine whether the predictor variables (instability indices) have an association with the predictand (thunderstorm occurrence). The confidence level considered in the study was 95%. The method splits the sum of squares into two components, a residual sum of squares and a regression sum of squares and the mathematical expression is:

$$\sum_i (y_\mathrm{i} - \overline{y})^2 = \sum_i (y_\mathrm{i} - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \overline{y})^2$$

where, $\hat{y}_\mathrm{i}$ represent the value of y_i predicted from the regression equation and $\overline{y}$ is the sample mean of 'y'.

F-test signifies whether the linear model provides a better fit to the data than a model without the variables. Thus, according to the F-test used for the hypothesis testing, the null hypothesis (H_0) assumes a model in which no independent variables fit the data. On the other hand, the alternate hypothesis (H_a) supports that the model fits the data better than the intercept-only model.

Comparing the p-value for the F-test to the significance level of 0.05 (95% confidence interval) denotes the confidence in the model. If the p-value is less than the significance value (0.05), H_0 is rejected to conclude that the regression model fits the data better than the model with no variables included.

2.2.4. Methodology/Algorithm

(1) Collect and import the data
(2) Transform all variables into numeric values ex- thunderstorm days = 1 and non-thunderstorm days = 0
(3) Clean the data and remove correlated independent variables using correlation matrix/heatmap, etc.
(4) Split the data into training and test set. In our case, the training and the test set are chosen randomly. The training set comprises 80% of the total data used and the test set uses the remaining 20%, which is then used for the model evaluation. To ensure that both the groups, training and test, are reasonably chosen, the means of the sets are checked before creating the classifier/model.
(5) Create the model
(6) Evaluated the model. We have used the confusion matrix to help quantify model precision, accuracy and recall.

2.2.5. Confusion Matrix

As a performance measure of our machine learning classification, we use the confusion matrix for the two outcomes- Thunderstorm occurrence (1) and non-occurrence (0).

The table has four different combinations of predicted and actual values. It is a useful measure for accuracy, recall, precision and specificity.

Precision: The precision represents the ability of the classifier/model to not label a sample as a positive if it is negative. It is given as the ratio of the true positive and the sum of the true positive and false positive.

Precision = True positive/(True positive + False positive)

Recall: It represents the ability of the classifier to find all the positive samples. It is given by the ratio of true positives and the sum of true positives and false negatives.

Recall = True positive/(True positive + False negative)

Specificity: It is given by the ratio of the true negative and the sum of the true negative and false positive.

Specificity = True Negative/(True Negative + False positive)

F-beta score: It is the weighted harmonic mean of the precision and recall. The best value is 1 and the worst is 0. The score weighs recall more than precision by a factor of beta. In our case, both, recall and precision are important. This is because we consider both, thunderstorm occurrence and non-occurrence prediction to be important and thus, we have used beta = 1, i.e., both recall and precision are equally weighted.

3. Results and Discussion

The convective indices, *SI*, *LI*, *KI*, *TTI*, *CAPE*, and *CIN* were analyzed over the six locations- Bangalore, Delhi, Goa, Jodhpur, Patiala, and Patna for a period of three-year from 2013 to 2015 during the thunderstorm prone, summer months. The first section aims to understand dominant indices over each of the 6 locations, using regression analysis. Considering the binary nature of the thunderstorm dataset used, an attempt is made to forecast the thunderstorm occurrence on any given day using logistic regression. Finally, the role of indices is investigated using range analysis.

3.1. Bangalore

The convective indices, *SI*, *LI*, *KI*, *CTI*, *TTI*, *CAPE*, and *CIN* were analyzed in Bangalore. Out of all the considered indices, *TTI*, *LI*, and *SI* turn out to be the dominant factor in predicting thunderstorm activity over Bangalore. Further linear regression of the indices indicates a higher contribution by *TTI* and *LI* indices as compared to *SI*. The linear dependence of the thunderstorm occurrence on any given day for a given *TTI* and *LI* values is given by,

$$TS\ day = F(TTI, LI) = -0.92 + 0.03 * TTI + 0.03 * LI$$

The above linear relationship, however, is not significantly above the considered threshold of 95% significance level and thus, the *LI* index, with the high *p*-value was eliminated. The new relationship between thunderstorm activity and *TTI* thus comes out to be,

$$TS\ day = -0.4 + 0.02 * TTI \quad (1)$$

The relationship is significant above 95% significance level when tested with F-test of ANOVA analysis. It should be noted that the r^2 value is small, representing only 3% of the total variability in thunderstorm activity by the relation.

Thus, to capture the thunderstorm characteristics better, we chose a non-linear logistic regression approach.

As explained in the data section, the thunderstorm day data (dependent variables) is dichotomous (binary) in nature and thus logistic regression is an appropriate method to explain the relationship between the independent and dependent variables. The logistic regression-based non-linear statistical model to operationally forecast thunderstorm events on any given day using 00:00 UTC data is thus,

$$P(Z) = exp\ (Z)/(1 + exp(Z)) \quad (2)$$

where $Z = 9.6 + 0.22 \times TTI$

The non-linear model (2). The relationship is useful in predicting thunderstorm activity in only a handful of cases. A clear model to predict thunderstorm operationally could not

be determined using the 00:00 UTC data in this analysis. This could be due to the following reasons; (1) the observational sites providing thunderstorm data are sparsely located and could have missed any event occurrence (2) limitation in data: Data in the region had only 24% thunderstorm cases which were not sufficient enough to predict a reliable logistic model.

Range analysis shows that most of the thunderstorms occur between the *TTI* range 37.1 and 57 (Figure 2). Both thunderstorm and non- thunderstorm days are quasi-Gaussian in nature being symmetrical around the previously stated *TTI* range. Thunderstorm and non- thunderstorm days were analyzed in different *TTI* ranges over Bangalore. No thunderstorm exists on the days when *TTI* < 37 °C and most thunderstorm days (30%) occur when *TTI* lies between the range, of 47.1 to 57. Accordingly, there lies a 30% probability of thunderstorm occurrence over Bangalore when *TTI* ranges between 47.1 to 57 °C (Figure 2). The probability is computed by dividing the number of thunderstorm days in the *TTI* range by the sum of both, thunderstorm and non-thunderstorm days in the range. The results obtained from this range analysis also provide a probability (~23%) when *TTI* over Bangalore is between the range, of 37.1 to 47 °C. These results are comparable to the known values of *TTI* indicating a potential for thunderstorm occurrence. The thunderstorm is thus highly likely (~52%) over Bangalore when *TTI* > 37 °C. A rule of thumb here can thus be that thunderstorm is highly probable if *TTI* is between the given range, 37.63 to 50.22 °C. No thunderstorm occur for *TTI* < 37 °C.

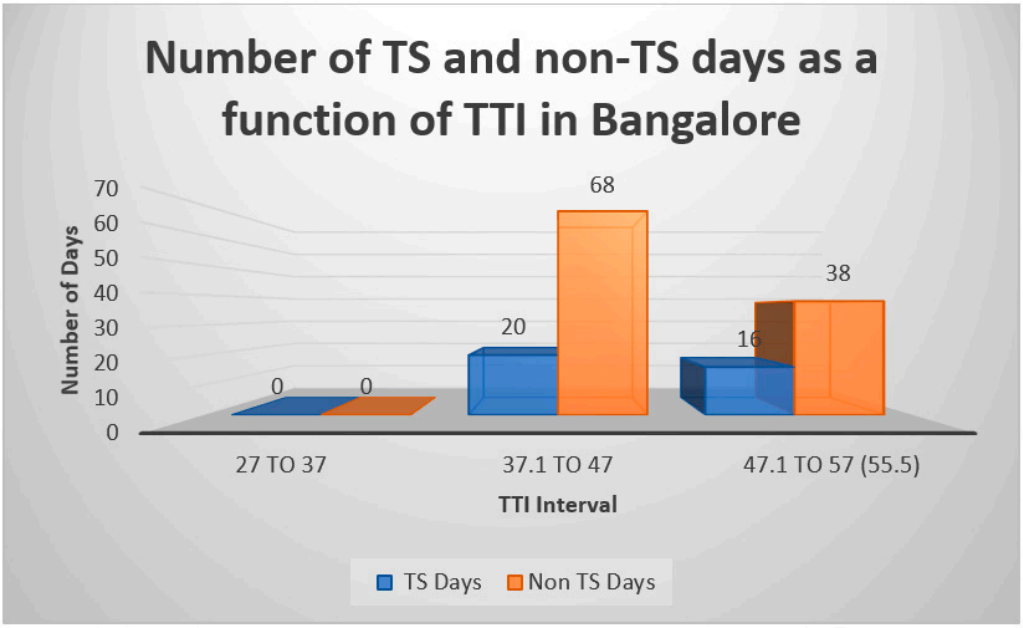

Figure 2. Distribution of thunderstorm and Non-thunderstorm days for given TTI values at 00:00 UTC values over Bangalore.

3.2. Delhi

SI, *LI*, *KI*, *CTI*, *TTI*, *CAPE* and *CIN* are the thermodynamic convective indices analyzed over the Delhi region. A 2-Way ANOVA revealed that *TTI* and *SI* are the significant contributors of thunderstorm activity over Delhi (3a). The result is significant at $p = 0.01$ level. A positive correlation was found between the indices however, only 4% of the total variation in the thunderstorm could be explained by the linear relation,

$$TSday = 1.41 - 0.02 * TTI - 0.05 * SI \quad (3a)$$

Considering this aspect, the non-linear approach was considered next in the analysis. Further analysis of *TTI* and *SI* indices using a logistic regression (3) approach provides us with the following probabilistic model,

$$P(Z) = exp\ (Z)/(1 + exp(Z))$$

where,

$$Z = -0.7 - 0.02 * \mathrm{TTI} - 0.08 * SI \quad (3b)$$

This nonlinear logistic regression based model is useful in predicting very few thunderstorm cases correctly. Consequently, no clear nonlinear relation was found between the thunderstorm occurrence on any given day using a convective index at 00:00 UTC over Delhi. The *range-based analysis* of the *TTI* index however provides more information. Most thunderstorms occurred when *TTI* was in the range of 37 to 50.22. Results (Figure 3) indicate that No thunderstorm occurred for *TTI* < 5 °C units and *SI* < −23 °C. The dominant range of SI where thunderstorm occurred most is between −2.9 to 7.1 °C (Figure 3).

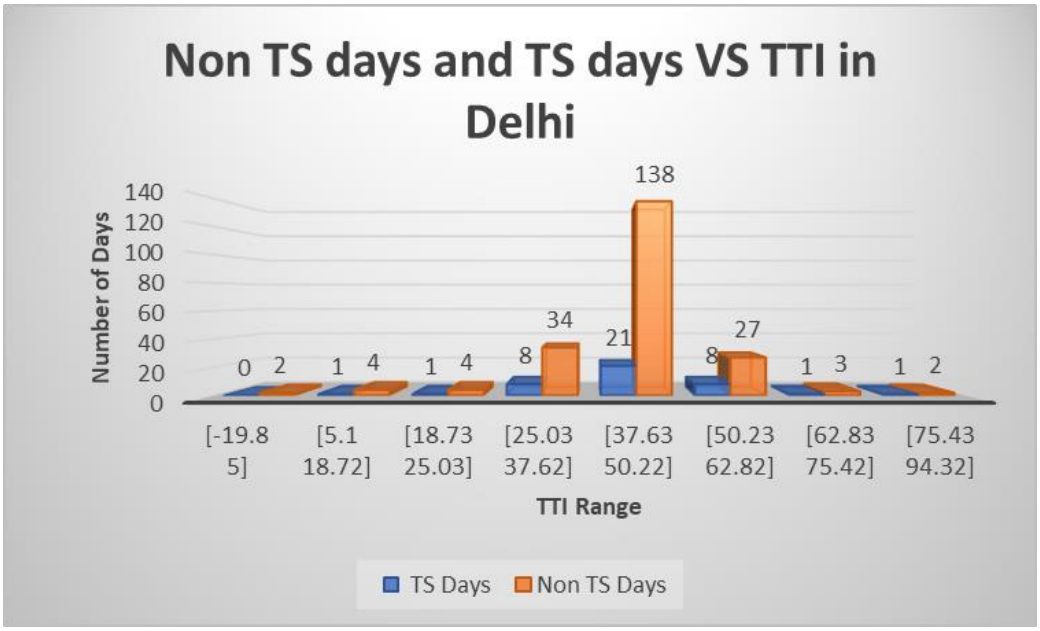

Figure 3. The figure shows the distribution of thunderstorm and non-thunderstorm days in the given *SI* range over Delhi. As can be seen here, thunderstorms are prominent between the *SI* range of 37.6 and 50.2.

3.3. Goa

To assess the predictability of a thunderstorm event on any given day using convective indices at 00:00 UTC, the following indices were used in Goa, *LI*, *CTI*, *SI*, *TTI*, *CIN*, *CAPE*, and *KI* (Figure 4). The indices and observational thunderstorm occurrence data were analyzed for the 3-year period, 2014, 2015, and 2016 for April, May, and June months which are the thunderstorm prominent months in India. The first set of analyses examined the contribution of various convective indices on a thunderstorm activity over the Goa region. Out of the considered seven indices in the region, *LI* and *KI* turned out to be the most prominent indices. The coefficient of determination indicates that 11% of the total variability in thunderstorm occurrence over Goa can be explained by the entire set of considered 6 indices together.

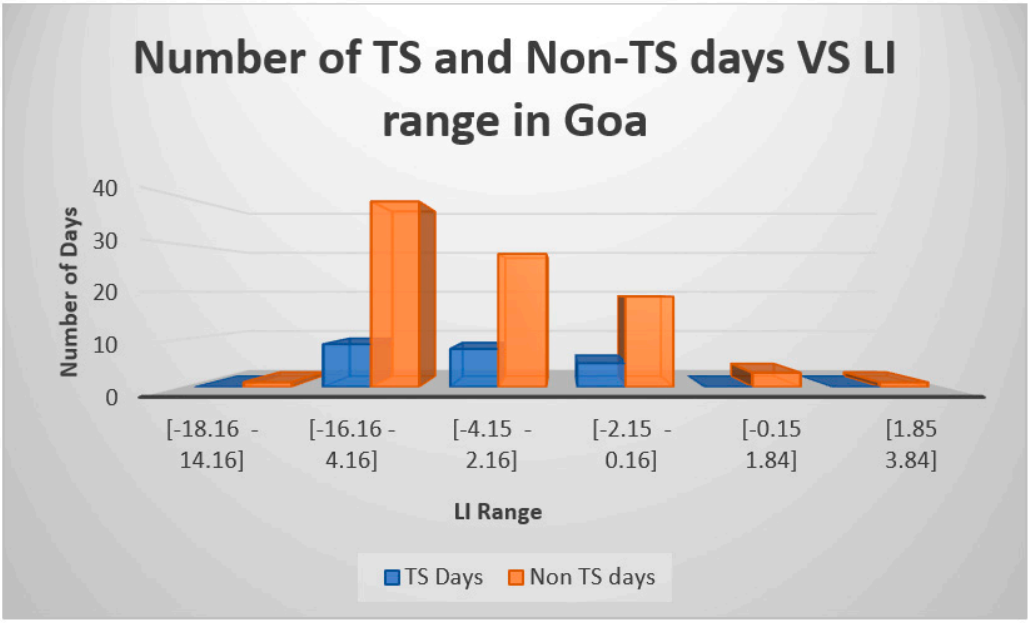

Figure 4. The distribution of thunderstorm and non-thunderstorm days in the given *SI* range over Goa. As can be seen here, Thunderstorms are more prominent between the *LI* range of −16.6 and −4.16 °C.

The dependence, however, increases to 15% when *LI* is included in the relation. The connection between thunderstorm occurrence and *LI* independently however are not significant above 95% significance level is,

$$TSday = 13.4 + 0.7 * LI \quad (4)$$

The small value of coefficients of determination indicates a lack of linear relationship between the independent (*LI*) and dependent (*TSday*) variables. Considering the result and binary nature of the thunderstorm dataset, the next analysis of the data uses logistic regression on independent (*LI*) and dependent (thunderstorm event) data. Using logistic regression (5, which is a more pragmatic approach for analyzing binary data, leads to the following probabilistic model.

$$P(Z) = exp\ (Z)/(1 + exp(Z))$$

where,

$$Z = -1.4 + 0.02 * LI \quad (5)$$

thunderstorm and non-thunderstorm days were studied in various *LI* ranges. shows an overall spread of thunderstorm and non-thunderstorm days at different *LI* intervals. A closer inspection of the figure indicates that no thunderstorm occurs when $LI < -0.16$ °C. Non-thunderstorm day occurrence is about 62% for the *LI* ranging between −4.15 and −0.16 while most of the thunderstorm days were observed in the LI range of 16.6 and −4.16 °C (Figure 4).

3.4. Jodhpur

The occurrence of thunderstorm days over Jodhpur was analyzed using *LI*, *TTI*, *SI*, *CAPE*, *CIN*, and *KI*. The 3 years of data spanning 2014, 2015, and 2016's April, May, and June months were quantitatively analyzed to predict thunderstorm occurrence on any given day using convective indices at 00:00 UTC. The regression method was first used to determine the factors that prominently influence thunderstorm occurrence. Out of the considered thermodynamic indices, *LI*, *SI*, and *CIN* turn out to be the most effective indices

determining thunderstorm occurrence on any given day over the Jodhpur region. The result is significant at the p = 0.01 level. The linear relation (6) is thus given as,

$$TSday = -0.22 + 0.02 * KI + 0.02 * SI - 0.0004 * CIN \quad (6)$$

However, the coefficient of determination indicates that only 10% of the total thunderstorm variability is explained by this linear relation.

Thus, we further explored the relationship between the dependent and independent variables to determine the variable that most significantly impacted thunderstorm occurrence. Further analysis reveals that *KI* alone accounts for thunderstorm occurrence on a given day over Jodhpur and the relation comes out to be

$$TS\ day = 6.1 + 0.1 * KI$$

The relation is significant only above the 84% significance level which indicates the presence of a non-linear relation between the dependent and independent variables over Jodhpur.

Considering the dichotomous nature of the dependent dataset, the logistic regression method was applied which is thus a more pragmatic approach here (7). The probabilistic model based on the logistic regression method is,

$$P(Z) = exp\ (Z)/(1 + exp(Z)) \quad (7)$$

where,

$$Z = -6.1 + 0.16 * KI$$

This Relationship (7) correctly determined thunderstorm day for 30% of the total cases when the probability, p (Z), is greater than 50%. Thus, we finally employed a range of analyses to determine the thunderstorm day for practical applications. The range analysis depicts no thunderstorms were observed for $KI < -2$ °C (Figure 5). thunderstorm days are highly likely for $KI > 22$ °C. The probability of thunderstorm occurrence on any given day with *KI* between 22.1 and 44 °C is over 43% (Figure 5).

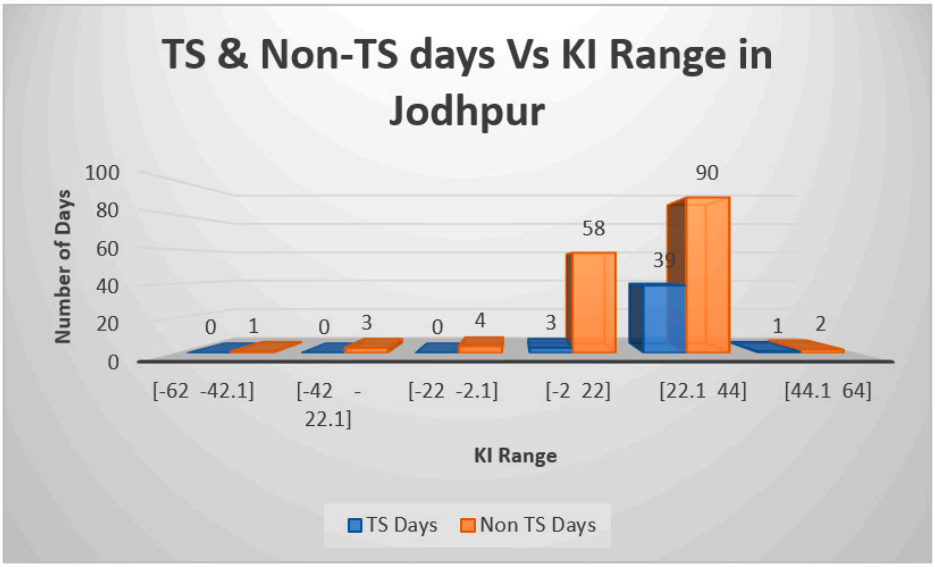

Figure 5. The distribution of thunderstorm and nonthunderstorm days in the given *KI* range over Jodhpur.

3.5. Patiala

Heating leads to the convection and production of thunderstorms over Patiala. The convective indices considered here are *TTI*, *KI*, *CIN*, *CAPE*, *CTI*, *LI*, and *SI* during the

summer months of April, May, and June of the years 2013, 2014, and 2015. Linear regression was used to examine the role of the thermodynamic indices in predicting a thunderstorm on any given. Amongst all the considered indices, *KI* significantly contributes positively to predicting a thunderstorm occurrence, and the linear Relation (8) between the two is given by,

$$TSday = 4.6 + 0.14 * KI \quad (8)$$

This relation is significant above the 90% significance level and 19% of the total variability in the dependent variable (*TSday*) is explained by the independent variable (*KI*).

The small value of the coefficient of determination in the previous analysis (8 indicates a possibility of a non-linear relation which is explored further using logistic regression method, because the dependent variable (*TSday*) is dichotomous. A day with thunderstorm is marked as 1, irrespective of the number of thunderstorms that occurred on the day, and 0 for the days which did not encounter any thunderstorm event. The log regression model (9). to estimate the probability that a given data entry belongs to is,

$$P(Z) = exp\ (Z)/(1 + exp(Z))$$

where,

$$Z = -2.9 + 0.3 * KI \quad (9)$$

The model provides a crude idea about thunderstorm occurrence and during the training and validation never leads to a 100% probability of thunderstorm occurrence on any given value of *KI* mainly because of several non-thunderstorm days even for the high value of *KI* in association with small *SI* values and high inhibition values quantified as *CIN*. Another reason for this could be because the location of the data centers, and providing the indices values are farther away from the region in Patiala where thunderstorm occurred. The spatial mismatch between the model reliability could have led to the limitation of this approach. However, it still provides a rough estimation of thunderstorm occurrence on a given day. To estimate the range of *KI* values contributing most to thunderstorm days, *KI* is spread at an interval of 62 units. No thunderstorm was observed for $KI < -22$ (Figure 6). A significant amount, 70.2%, of the thunderstorms occurs during the *KI* ranging between 19.1 and 43.1 °C wherein *KI* between, 25.1 and 31.1 °C comes out to be the most probable range demonstrating thunderstorm cases (Figure 6).

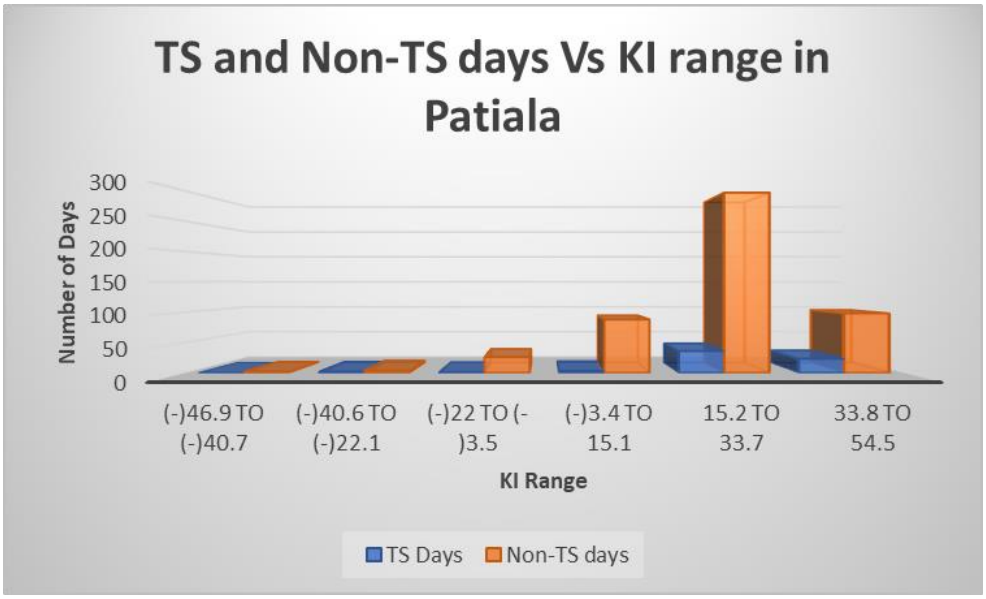

Figure 6. The distribution of thunderstorm (TS) and non-thunderstorm days in the given *KI* range over Patiala.

3.6. Patna

Six thermodynamic indices were considered over Patna. The indices were analyzed for three years for the thunderstorm prominent months of April, May, and June. Regression analysis revealed *KI* to be the index that contributes most to thunderstorm occurrence on any given day in Patna. The relation between thunderstorm occurrence and a thermodynamic index using 1-way Anova test indicates a relation significant above 95% significance level. The relation over Patna is given as follows:

$$TSday = -4.2 + 1.1 * KI$$

Considering the binary nature of thunderstorm day dataset, the logistic regression method was applied for further analysis. The logistic equation representing thunderstorm occurrence in Patna is,

$$P(Z) = exp\ (Z)/(1 + exp(Z))$$

where,

$$Z = -6.6 + 0.16 * KI \quad (10)$$

This logistic model (10 helps predict thunderstorm occurrence on any given day with a good probability. For p (Z) value > 30%, thunderstorm occurrence is highly likely. Simple model verification indicates correct thunderstorm prediction above 95% and non-thunderstorm day for p (Z) < 30% with a good probability of 95% and above. This is a significant outcome in Patna that could be well applied to estimate as thunderstorm day or a non-thunderstorm day using a threshold of 30%. A day is estimated to see a thunderstorm activity using the 00:00 UTC data over Patna if p (Z) > 30% and no-thunderstorm day for p (Z) < =30%. The total number of days, thunderstorms plus non-thunderstorms days, is 83. The number of thunderstorm days is 20 and the non-thunderstorm days is 63.

It can thus be concluded from Table 1, that the model specificity is 0.99, the sensitivity is 0.99 and the model accuracy is 82%.

Table 1. The confusion matrix for the classification model in Patna is as follows.

	Precision	**Recall**	**F1-Score**
0 (No-thunderstorm occurrence)	0.99	0.79	0.88
1 (Thunderstorm occurrence)	0.5	0.99	0.67
Accuracy		0.82	

The model validation using the observational data, Table 2, shows that thunderstorms might be likely when p (Z) < 30%, and almost certainly unlikely and when p (Z) > 30%.

Table 2. The contingency table for the validation of the classification model in Patna.

Forecast \ **Observed**	**0**	**1**
0	11	0
1 (Thunderstorm occurrence)	3	3

Authors generally advise interpreting the results with caution and suggest scope for further research with a longer dataset.

3.7. Thunderstorm Prediction Using Ranges

The thunderstorm and non-thunderstorm days were inspected in various *KI* ranges. Figure 7 represents an overall distribution of non-thunderstorm and thunderstorm days over a different *KI* ranges. Closer analysis shows that thunderstorm potential is highest (%) between 38.1 and 45.3 °C range. No thunderstorms were observed for *KI* < 31.8 °C (Figure 7).

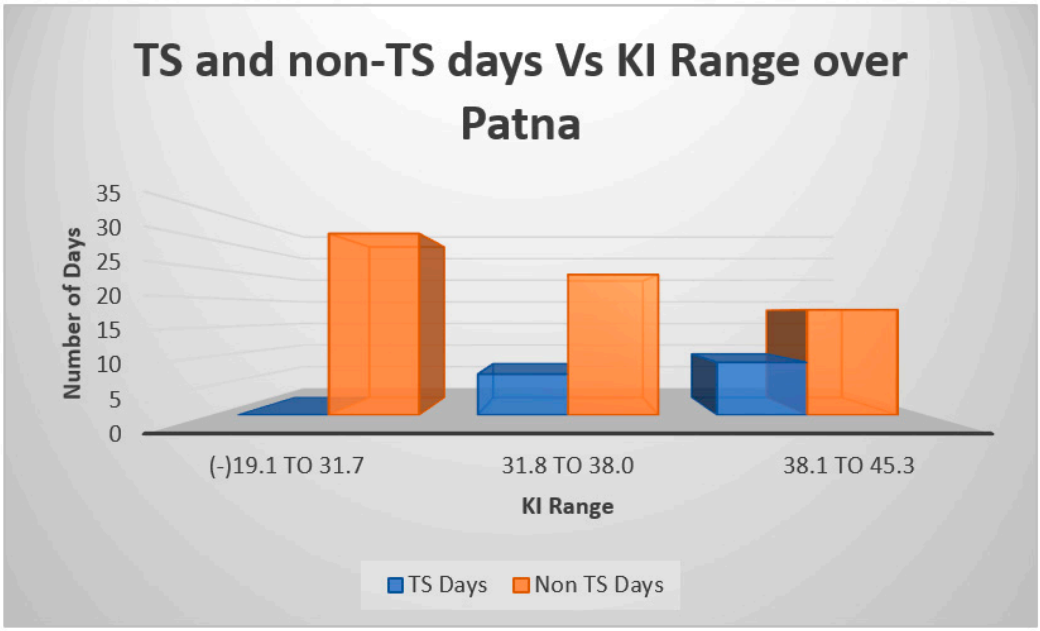

Figure 7. Distribution of thunderstorm and non-thunderstorm days in the given *KI* range over Patna.

Conclusively, we reject the Null hypothesis in the 99% confidence interval for the Delhi and Patna region. The null hypothesis being, H0: no relationship exists between the thunderstorm day forecast and instability indices considered. It indicates a highly significant statistical relationship between the TS-Day estimation and the instability indices considered in each of the locations as given by Equations (3) and (10) for the Delhi and Patna region, respectively. The relationship was statistically significant in the 90% confidence interval over Bangalore and Patiala (Equations (1) and (9)). while the relationship was feeble (significant only at 80% confidence interval) over Goa and Jodhpur (Equations (5) and (7)).

4. Conclusions

Bangalore: The current study found that *TTI* is the primary convective index out of the considered seven indices that predominantly help in determining thunderstorm occurrence over Bangalore. No thunderstorm was observed on the days when $TTI < 37$. On the other hand, thunderstorm over Bangalore is 52% likely when the value of TTI on any given day ranges between *37.1 and 47* °C. Therefore, thunderstorm over Bangalore can be considered highly probable when $TTI > 37$ °C.

Delhi: The results of this study indicate that over Delhi, TTI is the dominant index, followed closely by SI. No thunderstorm was found to occur when $TTI < 5$ °C. Interestingly, thunderstorm day was observed even on the days when TTI was low (TTI < 18) unlike that in Bangalore where no-thunderstorm was observed when $TTI < 37$ °C.

Goa: LI was found to explain 15% of the total variability in the overall number of thunderstorm days during the summer months over Goa. No thunderstorm was noticed when $LI < -0.16$ °C. However, these values suggest that a weak link exists between a thunderstorm day forecast and LI values.

Jodhpur: The result of this research suggests KI as the vital index that serves the thunderstorm forecasting over Jodhpur. No thunderstorm was observed for $KI < -2$ °C. thunderstorm days increase with KI values. However, on a day with $KI > 44$ °C, *thunderstorm was not observed because of the high value of capping inversion which prevents thunderstorm development.*

Patiala: Over Patiala, KI turns out to be the most relevant index in determining thunderstorm occurrence. The likelihood of thunderstorm occurrence on any given day

increases with KI values. No thunderstorm was observed over Patiala for $KI < 15$ °C. On most days (>70%) a thunderstorm was found to occur for KI between *19.1 and 43.1* °C.

Patna: A strong relationship between KI values and thunderstorm occurrence was found over Patna. The logistic method-based supervised machine learning approach provisioned a crucial relationship to forecast a thunderstorm occurrence on a given day. According to the model, when p (Z) is less than 30%, a thunderstorm is most certainly unlikely and might be likely when p (Z) is greater than 30% over Patna.

These findings have important implications for developing a thunderstorm forecasting system over various locations in India. The study suggests vital linear and non-linear connections between the major convective indices and thunderstorm occurrence. One of the most significant findings to emerge from this study is the operationally applicable thunderstorm forecasting model, for Patna.

The findings of this study thus have several important implications for operational Weather forecasters in India. However, with a limited sample size used here, the results should be interpreted with caution. We recommend further research with a larger dataset to tackle this issue.

Author Contributions: Conceptualization: K.A. and K.R.; methodology, K.A.; software, K.A.; validation, K.A.; formal analysis, K.A. and K.R.; investigation, K.A.; resources, K.R.; data curation, K.A.; writing—K.A. and K.R.; writing—review and editing, K.A., K.R., S.R. and R.M.; visualization, K.A.; supervision, K.A. and K.R.; project administration, K.A. and K.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by the Ministry of Earth Sciences, Delhi, India. This study used the sounding data from the website of the University of Wyoming's Department of Atmospheric Science and we would like to acknowledge it. The authors thank the anonymous reviewers and the academic editor for their comments, which have helped in improving the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Khole, M.; Biswas, H. Role of total-totals stability index in forecasting of thunderstorm/nonthunderstorm days over Kolkata during pre-monsoon season. *Mausam* **2007**, *58*, 369–374. [CrossRef]
2. Desai, B. Mechanism of Nor'wester of Bengal. *Mausam* **1950**, *1*, 74–76. [CrossRef]
3. Kessler, E. Thunderstorm morphology and dynamics. In *National and Atmospheric Administration, Environmental Research Laboratories*; U.S. Department of Commerce: Washington, DC, USA, 1982; Volume 2.
4. Kumar, S. An objective method of forecasting pre-monsoon thunderstorm/duststorm activity over Delhi and neighbourhood. *Mausam* **1972**, *23*, 45–50. [CrossRef]
5. Srinivasan, V.R.; Ramamurthy, K.; Nene, Y.K. *Discussion of Typical Synoptic Weather Situation: Summer-Norwesters and Andhis; and Large Scale Convective Activity over Peninsula and Central Parts of the Country, India Meteorological Department Forecasting Manual, Part 3*; Delhi, India, 1973.
6. Asoilal. Forecasting of thunderstorm around Delhi and Jodhpur. *Mausam* **1989**, *40*, 267–268.
7. Seshadri, N.; Jain, P.S. Study of role of stability index in forecasting thunder squall. *Mausam* **1989**, *40*, 101–106. [CrossRef]
8. Lal, R. Forecasting of severe convective activity overLucknow in premonsoon season. *Mausam* **1990**, *41*, 455–458. [CrossRef]
9. Lee, R.; Passner, J.E. The development and verification of TIPS: An expert system to forecast thunderstormoccurrence. *Weather Forecast.* **1993**, *8*, 271–280. [CrossRef]
10. Sahu, J. A study on the thermodynamics conditions of atmosphere for forecasting thunderstorms over northwest India. *Vatavaran* **1996**, *19*, 27–35.
11. Devrani, A.M. A forecasting tool for predicting pre-monsoon thunderstorm/dustorm over Jodhpur. *Vatavaran* **1997**, *21*.
12. Ravi, N.; Mohanty, U.; Madan, O.; Paliwal, R. Forecasting of the thunderstorm in the pre-monsoon season at Delhi. *Meteorol. Appl.* **1999**, *6*, 29–38. [CrossRef]
13. Dhavan, V.; Tyagi, A.; Bansal, M. Forecasting in pre-monsoon season over northwest India. *Mausam* **2008**, *59*, 433–444.
14. Kunz, M.; Sander, J.; Kottmeier, C. Recent trends of thunderstorm and hailstorm frequency and their relation to atmospheric characteristics in southwest Germany. *RMetS* **2009**, *29*, 2283–2297. [CrossRef]

15. Schultz, P. Relationships of several stability indices to convective weather events in northeast Colorado. *Weather Forecast.* **1989**, *4*, 73–80. [CrossRef]
16. Modahl, A.C. Synoptic parameters as discriminators between hailfall and less significant convective activity in northeast Colorado. *J. Appl. Meteor.* **1979**, *18*, 671–681. [CrossRef]
17. Haklander, A.; Delden, A. Thunderstorm predictors and their forecast skill for The Netherlands. *Atmos. Res.* **2003**, *67*, 273–299. [CrossRef]
18. Ray, K.; Bandopadhyay, B.; Sen, B.; Sharma, P.; Warsi, A.; Mohapatra, M.; Yadav, B.; Debnath, G.; Stella, S.; Das, S.; et al. *Report on Pre-Monsoon Season 2013 Thunderstorms over India (SAARC STORM Project-2013)*; ESSO Document Number: ESSO/IMD/NWFC/SR/01(2013)/1/Scientific Report; IMD (ESSO): Delhi, India, 2014.
19. Ray, K.; Bandopadhyay, B.; Sen, B.; Sharma, P.; Warsi, A.; Mohapatra, M.; Yadav, B.; Debnath, G.; Stella, S.; Das, S.; et al. *Thunderstorms 2014—A Report (SAARC STORM Project-2014)*; No. ESSO/IMD/SMRC STORM Project-2014/01(2014)/03; Nowcast Unit, India Meteorological Department: New Delhi, India, 2015.
20. Ray, K.; Bandopadhyay, B.; Sen, B.; Sharma, P.; Warsi, A.; Mohapatra, M.; Yadav, P.; Debnath, G.; Stella, S.; Das, S.; et al. *Pre-Monsoon Thunderstorms 2015: A Report*; 10.13140/RG.2.1.2663.5285; IMD (ESSO): Delhi, India, 2016.
21. Pradip, S.; Kamaljit, R.; Bikram, S. Monitoring Convective Activity over India During Pre-Monsoon Season-2013 under the SAARC STORM Project. Delhi, India. *Vayumandal* **2017**, *42*, 98–116.
22. Ray, K.A.; Kannan, B.; Sharma, P.; Sen, B.; Warsi, A.H. Severe Thunderstorm Activities over India during SAARC STORM Project 2014-15: Study Based on Rada. *Vayumandal* **2017**, *43*, 33–49.
23. Wyoming. 2020. Available online: https://weather.uwyo.edu/upperair/sounding.html (accessed on 20 September 2020).