Cumulonimbus cloud prediction based on machine learning approach using radiosonde data in Surabaya, Indonesia

R M Putra¹, E Fibriantika¹,², H Herawati¹, Y Kusumayanti¹, E Afriani¹, A Hidayanti¹, A Wiwijana¹, W A Swastiko¹ and D Andrianto¹

¹Center for Aeronautical Meteorology, Agency for Meteorology, Climatology, and Geophysics, Jl. Angkasa I No.2 Kemayoran, Central Jakarta, Jakarta, 10720, Indonesia
²Corresponding author: eka.fibriantika@bmkg.go.id

Abstract. Increase in frequency and strength of cumulonimbus is one of the impacts of climate change. The presence of cumulonimbus usually causes extreme weather. Cumulonimbus can produce heavy rainfall, tornadoes, turbulence, and other extreme weather events. Upper air conditions have a great effect on the process of cloud growth. Radiosonde observations can be used to predict the presence of cumulonimbus in the short-term period of weather forecast. This study aimed to predict the occurrence of cumulonimbus using radiosonde data based on the machine learning approach. In this study, indices data from upper-air observation were used. The model prediction of radiosonde data was trained using machine learning to predict the presence of cumulonimbus. Based on data processing results, the prediction of cumulonimbus events using radiosonde indices data is good enough when implemented in new test data. The influence of the Convective Available Potential Energy (CAPE) index in the predictor index predicts cumulonimbus. Machine learning model can predict cumulonimbus incidence by 80% in one month testing period when adding the CAPE index. Meanwhile, when not using CAPE, cumulonimbus events’ predicted results only reach 72% of events. The false alarm rate when adding CAPE was 17% and without CAPE was 21%. Based on these results, it can be concluded that the prediction of cumulonimbus cloud events using radiosonde data based on the machine learning approach is sufficiently reliable to be used.

1. Introduction
Based on IPCC usage, climate change refers to any change in climate over time, whether due to natural variability or human activity [1]. A few impacts of climate change include the increase of weather phenomenon frequency and weather phenomenon that could disrupt aviation such as low-level wind shear, lightning strikes, turbulence, icing and more intense cumulonimbus (CB) [2]. Cumulonimbus can produce extreme weather events, such as thunderstorms and turbulence [3]. Previous studies in predicting cumulonimbus have been carried out using stability index WRF-ARW Model output and the Rough Set and Artificial Immune Algorithm [4][5].

Radiosonde is a telemetry instrument carried into the atmosphere, usually by balloons [6]. Radiosonde observation data that provides some weather parameters are used to understand the cloud vertical structure [7]. Radiosonde data processing could also result in instability indices. Few studies related to instability indices based on radiosonde data have been done [8][9]. Instability indices are very closely related to atmospheric stability conditions which have a great role in the cloud formation process.
Machine learning plays a central role in artificial intelligence [10]. Weather forecasting is one of the most challenging areas for machine learning [11]. Machine learning effectively is used for meteorological analysis, where there are sufficient data set of relevant features for cases with known outcomes [12]. Weather prediction and modeling using machine learning approach have been used on several weather parameters [11][12][13].

The usage of instability indices on predicting cumulonimbus using neural network method in Cengkareng has been done [14]. However, this research uses too many instability indices [14]. Meanwhile, this study focused on using instability indices that have been used operationally in BMKG namely K Index (KI), Showalter Index (SI) and Lifted Index (LI). Total Totals Index (TT) and Convective Available Potential Energy (CAPE) have a significant effect on the growth of convective clouds in Indonesia [8][9]. Hence, the usage of these 5 indices in this study is expected to increase the cumulonimbus prediction model's accuracy with a smaller number of indices than previous research [14]. Juanda Surabaya International Airport is one of the largest airports in Indonesia. The hail incident caused by cumulonimbus occurred in Surabaya on 12 January 2017 [15]. In January, the growth of cumulonimbus in Surabaya increased the intensity of vertical wind shear during that month [16]. Vertical wind shear is very dangerous to aviation safety, therefore it is necessary to develop a cumulonimbus prediction model in Surabaya. This study aimed to predict the occurrence of cumulonimbus clouds using radiosonde data based on the machine learning approach in Surabaya. In this study, the effect of using CAPE on the prediction model accuracy was also investigated.

2. Data and methods

Radiosonde data observed at Juanda Surabaya Meteorological Station from 2018-2019 were used in this study. The observation was conducted twice a day at 00.00 UTC and 12.00 UTC. Instability index analysed in this study included SI [17], LI [18], KI [19], TT [20] and CAPE [21]. These indices data can be downloaded at http://weather.uwyo.edu/upperair/sounding.html using WMO ID of Juanda Surabaya Meteorological Station, 96935.

Cumulonimbus data obtained from hourly surface synoptic observation, based on manned observations by looking directly to the cloud conditions. Cumulonimbus data used in this study were observation data for 12 hours after radiosonde observation.

In this study, the Pattern Recognition Neural Network method was used because it can be trained to classify the input data based on their target classes [22]. The pattern recognition was made to classify a category, where the system's target category was determined by the system (supervised category). The target data for pattern recognition networks should consist of vectors of all zero values except for a 1 in element i, where i is the class they are to represent. In this study, the determination of the category of any given in the NN model was the presence and absence of cumulonimbus clouds. According to the output target of this study, this method was whether there is cumulonimbus during 12 hours after radiosonde observation.

In the neural network model, 3 layers consist of an input layer in the form radiosonde indices, a hidden layer consisting of 15 neurons [23] and an output layer that is the condition for the presence or the absence of cumulonimbus during the next 12 hours interval.

Two architecture design conditions were used in this study. The first design use 5 instability indices (SI, LI, KI, TT and CAPE) as input data. While the second design only used 4 instability indices (SI, LI, KI and TT) as input data. Cumulonimbus data was used as an output for both designs.

The model running process was split into two processes namely training process and testing process. The training data used to create the prediction model were radiosonde data and cumulonimbus from January 2018 to November 2019. During training, the condition where cumulonimbus present is scored 1 as the target, while condition, where there is no cumulonimbus present, is scored 0. The training process was used for both architecture designs.

After the training process was completed, the testing process was carried out to the model that had been formed. Radiosonde data in December 2019 at 00.00 and 12.00 UTC were used in the testing process. Output target resulted by model was a value between 0 and 1. The testing result with a value
less than 0.5 are rounded to 0, which means there is no cumulonimbus. Meanwhile the value from 0.5 to 1 was rounded to 1, which implies cumulonimbus at the next 12 hours. Then the rounded testing output was compared with cumulonimbus observation data so that the percentage accuracy of the model was obtained. The testing process was used for both architecture designs.

The verification method used in this study was dichotomy verification using 2 x 2 contingency table. In the verification, model data were compared to actual observation data and then looks for the value of accuracy, probability of detection and false alarm rate [24]. The accuracy value index shows model validity in predicting cumulonimbus and validity in predicting conditions where there is no cumulonimbus. The probability of detection (POD) value is the frequency of cumulonimbus events that can be generated by the designed neural network model. This index is used to determine the model ability to predict event of actual cumulonimbus. Higher POD value indicates that actual cumulonimbus events can be detected properly. Furthermore, the False Alarm Rate (FAR) is an event when the designed model shows cumulonimbus conditions, but there is no cumulonimbus during the next 12 hours in the observation. The FAR value is the simulation error value, which indicates that cumulonimbus condition does not occur.

3. Results and discussion

3.1. Upper air’s characteristic in Surabaya

The appearance of cumulonimbus clouds can be detected by several radiosonde indices, namely SI, LI, KI, TT and CAPE. In representing the index, the Box plot method is used, a technique of describing the data in five numerical summaries of the collection to visualize the spread and slop [25].

![Figure 1. Distribution of SI in Surabaya Period January 2018 - December 2019 at 00.00 UTC (left) and 12.00 UTC (right) based on radiosonde data processing.](image)

SI is used to forecast thunderstorms based on the humidity and temperature values at the 850 mb layer and the temperature at the level of 500 mb [17]. Based on Figure 1, SI tends to have a low level when entering the rainy season, especially in DJF. This is in accordance with a previous study that the mean value in DJF of SI in Surabaya is 0.49 [26]. Meanwhile, from August to November, SI tends to be higher. The Intensity of SI show the potential for thunderstorm events. The lower value of SI, the greater the potential for thunderstorms [17].

LI is a modification of the Showalter index with differences in determining the level where the parcel is lifted [18]. At 00.00 UTC, the LI characteristics in January - May and November - December were negative. Meanwhile, in August - October were positive. The negative LI condition in DJF in Surabaya is aligned with a previous study [26]. A negative Lifted index indicates warmer air, which is less stable and has a potential for thunderstorms, while a positive Lifted Index value shows that the air tends to be cooler and more stable [18].

KI was formulated to help forecast continental summertime air mass thunderstorm potential [19]. Based on Figure 3, KI has a characteristic that is inversely proportional to the SI and LI. KI at 00.00 UTC and 12.00 UTC have a distribution that tends to be lower during the dry season, especially from
July to September and tends to increase in the rainy season (DJF). The average value of KI in DJF in Surabaya is 33.82 [26].

**Figure 2.** Distribution of the LI in Surabaya from January 2018 - December 2019 at 00.00 UTC (left) and 12.00 UTC (right) based on radiosonde data processing.

**Figure 3.** Distribution of the KI in Surabaya from January 2018 - December 2019 at 00.00 UTC (left) and 12.00 UTC (right) based on radiosonde data processing.

**Figure 4.** Distribution of the TT in Surabaya from January 2018 - December 2019 at 00.00 UTC (left) and 12.00 UTC (right) based on radiosonde data processing.

TT was an index formulated based on the Vertical Total Index and the Cross Totals Index [20]. Based on Figure 4, the patterns and characteristics of TT are similar to KI. When entering the rainy season, specifically in DJF, TT tends to be higher than during the dry season, from June to September. The mean value of TT of DJF in Surabaya is 43.11 [26]. The higher value of the Total Total Index, the higher the potential for thunderstorms. In general, the value is higher if the humidity conditions are high in the subsoil to a level of 850 mb [20].

Up shear or down shear propagation speed of a cumulonimbus cell is determined as a CAPE function [21]. Based on Figure 5, it can be seen that in July - October, the distribution of CAPE tends to be lower than in another period. From December to April, it can be seen that the distribution of CAPE was relatively higher because this period already enters in the rainy season. The average CAPE value in Surabaya was 609 J [9]. Vertical buoyancy of the rising air parcels is a good indicator to
determine the level of atmospheric instability. When the level of convection is strong, the CAPE value will increase and the potential for convective cloud growth is greater [21].

Figure 5. Distribution of the CAPE in Surabaya for the period January 2018 - December 2019 at 00.00 UTC (left) and 12.00 UTC (right) based on radiosonde data processing.

3.2. Modelling CB prediction using neural network

After the neural network model was built based on training data for January 2018 - November 2019, a model testing was carried out during December 2019. Table 1 and Table 2 show the verification of the neural network model compared to field observation data regarding the presence of Cumulonimbus.

| Result                  | CB Observation | CB Observation |
|-------------------------|----------------|----------------|
| CB Prediction without CAPE | Yes            | 33             |
|                         | No             | 9              |

Based on the cumulonimbus prediction test results during 12 hours after radiosonde observation, cumulonimbus prediction accuracy had a large number of events. Based on the overall data test, the failure rate of cumulonimbus event prediction when not using the CAPE index as one of the input parameters were 5 times out of 38 cumulonimbus events (13.2%). Meanwhile, the model prediction that showed an error indicates no cumulonimbus observation when the model predicts cumulonimbus clouds, which is 9 times out of 38 cumulonimbus events (23.7%).

| Result                  | CB Observation | CB Observation |
|-------------------------|----------------|----------------|
| CB Prediction Include CAPE | Yes            | 35             |
|                         | No             | 7              |

Besides, when the model was simulated using all indices, including CAPE, there is a significant increase in the total success in predicting cumulonimbus' appearance. Based on the test data, there were 35 cumulonimbus events predicted by the model, so that the total presence of cumulonimbus which cannot be predicted by the model in the case study of data testing, was 3 times out of 38 cumulonimbus events (7.9%). Thus, there is a decrease in errors from the model on predicting cumulonimbus clouds' appearance when there are no clouds to 7 times from 38 cumulonimbus events (18%).

The ANN model, which had been created before, will be tested in case study during December 2019. The verification was carried out using a contingency table as describes in tables 1 and 2. Figure
6 shows the level of the ANN model's ability to represent cumulonimbus events in Surabaya area during December 2019.

![Figure 6](image)

**Figure 6.** Skill Score result based on Artificial Neural Network Model which had been created.

Based on Figure 6, when the model was simulated without using the CAPE, the accuracy value reaches 72%. Meanwhile, when the CAPE was added into the training and testing model, the results increase the accuracy value to 80% of the total test data. This accuracy value is quite significant compared to previous study that used 15 variables as input models with an accuracy 82.5% [14]. The success rate in detecting cumulonimbus's appearance is indicated by the score from the Probability of Detection (POD). Based on these results, the POD value also tends to increase when the model uses all indices as discussed in this study compared to only a few indices. Then, the model's overestimate error value decreased significantly from 21% to 17% when the CAPE is added into the training or testing model.

4. **Conclusion**

Cumulonimbus prediction based on a machine learning approach using radiosonde data was successfully carried out with an accuracy level of up to 72% on models without using CAPE and 80% when using CAPE. A similar condition occurred in the POD value, when using 5 indices as input, it reached 92%. Meanwhile, when using 4 inputs, the POD value decreased to 87%. The false alarm rate had decreased significantly from 21% to 17% when adding the CAPE in the training or testing model. This shows that predicting CB clouds' presence using the machine learning approach has good predictive quality, especially when using input data in the form of SI, LI, KI, TT and CAPE. It can be concluded that the prediction of cumulonimbus cloud events using radiosonde data based on the machine learning approach is sufficiently reliable to be used.

**References**

[1] IPCC 2007 Summary for Policymakers In: *Climate Change 2007: The Physical Science basis*, p. 104.

[2] Puempel H and Williams P D 2016 The impacts of climate change on aviation: scientific challenges and adaptation pathways *ICAO Environmental Report 2016: On Board A Sustainable Future*, pp. 205–07.

[3] Storer L N, Williams P D and Gill P G 2019 Aviation Turbulence: Dynamics, Forecasting, and Response to Climate Change *Pure Appl. Geophys.* 176 2081–95.

[4] Sani P P and Ismanto H 2015 Prediksi Awan Cumulonimbus Menggunakan Indeks Stabilitas
Keluaran Model Wrf – Arw Di Bima *Jurnal Meteorologi Klimatologi dan Geofisika* 1 1–8.

[5] Wang Q and Fan W 2010 Cumulonimbus forecasting based on rough set and artificial immune algorithm, *2010 6th International Conference on Natural Computation ICNC 2010* 6 2856–60.

[6] Szuster P 2017 Data fixing algorithm in radiosonde monitoring process *J. Telecommun. Inf. Technol.* 1 81–8.

[7] Shastri N and Pathak K K 2019 New Cloud Detection Index (CDI) for Forecasting the Extreme Rain Events *Adv. Remote Sens.* 8 30–9.

[8] Budiarti M, Muslim M and Ilhamsyah Y 2012 Studi Indeks Stabilitas Udara Terhadap Prediksi Kejadian Badai Guntur (Thunderstorm) Di Wilayah Stasiun Meteorologi Cengkareng Banten *J. Meteorol. dan Geofis.* 13 110–17.

[9] Syaifullah M D 2018 Analisis Kondisi Udara Atas Wilayah Indonesia dengan Radiosonde *J. Meteorol. dan Geofis.* 14 19–27.

[10] Carbonell J G 1989 Introduction: Paradigms for machine learning *Artificial Intelligence*.

[11] Dai H 1996 Machine Learning of Weather Forecasting Rules from Large Meteorological Data Bases *Adv. Atmos. Sci.* 13 471–88.

[12] Tag P M and Peak J E 1996 Machine Learning of Maritime Fog Forecast Rules *J. Appl. Meteorol.* 35 714–24.

[13] Foresti L, Pozdnoukhov A, Tuia D and Kanevski M 2010 Extreme Precipitation Modelling Using Geostatistics and Machine Learning Algorithms *Geostatistics for Environmental Applications, Quantitative Geology and Geostatistics* 41–52.

[14] Putra A C and Lursinsap C 2014 Cumulonimbus Prediction Using Artificial Neural Network Back Propagation with Radiosome Indeces *Natl. Remote Sens. Semin.* 96 7549–51.

[15] N I Kadik T Dewi and Munandar M A 2017 Analisis Kondisi Atmosfer Pada Kejadian Hujan Es Di Surabaya ( Studi Kasus 12 Januari 2017 ) *Semin. Nas. Fis. dan Apl* 22–27.

[16] Zakir A and Hakim O S 2017 Analisis Vertical Wind Shear Dan Buoyancy Terhadap Meteorologi Juanda Surabaya *J. Meteorol. Klimatologi dan Geofis* 4 1–8.

[17] Showalter A K 1953 A Stability Index for Thunderstorm Forecasting *Bull. Am. Meteorol. Soc.* 34 250–52.

[18] Galway J G 1956 The Lifted Index as a Predictor of Latent Instability *Bull. Am. Meteorol. Soc.* 37 528–29.

[19] George J J 2014 Weather Forecasting for Aeronautics *Academic Press*.

[20] Miller R C 1975 Notes on analysis and severe storm forecasting procedures of the air force global weather central (Vol. 200). AWS.

[21] Moncrieff M W and Miller M J 1976 The dynamics and simulation of tropical cumulonimbus and squall lines *Q. J. R. Meteorol. Soc.* 102 373–94.

[22] Al-allaf O N A and Abdalkader S A 2013 Pattern Recognition Neural Network for Improving the Performance of Iris Recognition System 4 661–7.

[23] Putra R M, Saputro A H, Arazak L and Kharisma S 2019 Automatic detection of volcanic ash from Himawari-8 satellite using artificial neural network *AIP Conf. Proc.* 2202.

[24] Doswell III C A, Davies Jones R and Keller D L 1990 On summary measures of skill in rare event forecasting based on contingency tables *Weather and Forecasting* 5 576-85.

[25] McGill R, Tukey J W and Larsen W A 1978 Variations of box plots *Am. Stat.* 32 12–6.

[26] Fibriantika E and Alhaqq R I 2018 Profil Vertikal Atmosfer Selama Aktifitas Siklon Tropis Cempaka dan Dahlia *J. Meteorol. dan Geofis.* 19 49–58.