Indian Journal of Geo Marine Sciences
Vol. 51 (02), February 2022, pp. 117-125

Precipitation event detection based on air temperature over the Equatorial Indian Ocean

R V Shesu[a], M Ravichandran[b,d], K Suprit[a,e], E P Rama Rao[a] & B Venkateswara Rao[c]

[a]Indian National Centre for Ocean Information Services, Ministry of Earth Sciences, Hyderabad – 500 090, India

[b]National Centre for Polar and Ocean Research, Ministry of Earth Sciences, Goa – 403 804, India

[c]Jawaharlal Nehru Technological University, Hyderabad – 500 085, India

[d]Ministry of Earth Sciences, New Delhi – 110 003, India

[e]India Meteorological Department, Ministry of Earth Sciences, Port Blair – 744 106, India

*[E-Mail: venkat@incois.gov.in]

Received 20 September 2020; revised 22 February 2022

Air temperature (AT) and precipitation observations obtained from RAMA (Research Moored Array for African-Asian-Australian Monsoon Analysis and prediction) buoy at 0° N, 90° E from July 2009 to June 2017 are used to identify rainfall events. Based on the Random forest method, which consists of classification and regression based on decision trees, an algorithm is developed to identify the rainfall events from the change in AT data with high accuracy. During the study period, a total of 22461 abrupt drops in air-temperature events were identified by the algorithm. Around 75 % of these events were used to train and develop the clustering algorithm, and the rest of the events were used for validation with the precipitation data available from the buoy. The algorithm can identify more than 94 % of rain events accurately when the classification is binary. When the rain events are classified similar to the India Meteorological Department's classification, the algorithm is still able to identify the rain events; however, the performance degrades to ~ 84 % accuracy.

[**Keywords**: Air temperature, Classification, Equatorial Indian Ocean, Precipitation, Random forest]

Introduction

Precipitation plays a very important role in modulating weather and climate across the globe and has been designated as an essential climate variable[1,2]. Over the oceans, precipitation modulates Sea Surface Temperature (SST) and near-surface air temperature (AT), affecting the heat and salt budgets. Precipitation also affects the long wave radiation budget through the associated cloudiness. Precipitation is a key variable that influences various modes of variability, such as the Madden-Julian Oscillation (MJO), and the Indian Ocean Dipole (IOD) mode oscillations.

However, over the oceans, precipitation measurements are very sparse[3]. In the National Centers for Environmental Prediction – National Center for Atmospheric Research (NCEP-NCAR) reanalysis dataset, which is a widely used synthesis of observations model reanalysed data, precipitation is designated as a grade C variable and has to be used with caution[4]. Recent advances in remote-sensing precipitation measurements and algorithms have partially addressed this issue with the development of various merged remote sensing precipitation data products[5,6] while merging them with the available rain-gauge data from islands or coastal stations.

Over the years, merging data products have been a challenging task given the heterogeneous nature of the merged products. The NCEP-NCAR reanalysis datasets and the newer reanalysis gridded datasets can be explored for association mining across temperature and precipitation. However, it is vital to realize that errors in the model solutions can be dumped into the precipitation variable since the models do not conserve water, etc. In addition, while advances in remote sensing of precipitation have certainly occurred since the 1980s, for this very reason the data record is not easy to use. Early satellite rainfall estimates based mainly on IR are difficult to tie to more modern satellite products based on IR calibrated using microwave imagers and, even more important, space borne radars. So, *in-situ* rainfall is still essential for validation, even in the satellite era, and gauge data from islands are still used to create merged products.

Rain events are an external forcing on the air-sea interface; in the tropical oceans, they cool the near-surface layer due to the difference in temperature between falling rain drops (colder) and the underlying

SST (warmer). Various studies[7-9] have documented the net cooling effect of precipitation events on the SST and the overlying marine boundary layer. Building on these previous studies, here we explore near-surface AT and precipitation data to study the cooling events associated with the precipitation. It should be noted that AT datasets are simple to measure and easier to check the quality of data. A natural way of progression is to hypothesize that, can we use AT data to make an algorithm to detect rainfall events?

A reliable method for *in-situ* precipitation measurements is the self-siphoning rain gauges on board the moored buoys. However, they too suffer from measurement uncertainties[10]. A major challenge for obtaining precipitation from self-siphoning rain gauges deployed on open ocean buoys is maintaining the instruments under such harsh conditions. Hence, identifying the precipitation events is crucial. In the absence of rain data, the identification of rain events from AT can provide an independent source of information.

Using available long-term high-resolution moored buoy, a methodology is proposed to detect the peak rainfall events and their associated high-temporal variability in the Equatorial Indian Ocean. Rainfall and AT data are obtained from Research Moored Array for African-Asian-Australian Monsoon Analysis and Prediction (RAMA; McPhaden *et al.*[11]) buoy at 0° N, 90° E (location marked as the blue circle in Fig. 1) from July 2009 to June 2018. The

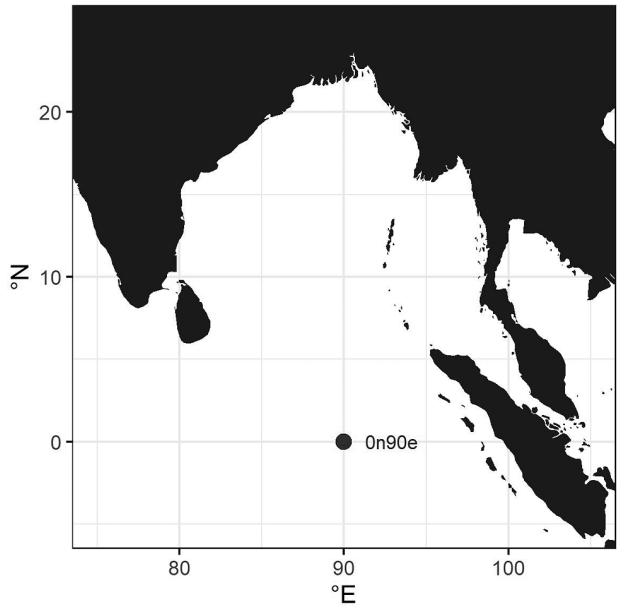

Fig. 1 — RAMA buoy (0n90e) location marked as a blue filled circle at 0° N, 90 °E in the Eastern Equatorial Indian Ocean

RAMA buoy measures real-time meteorological and oceanographic parameters, namely SST, AT, precipitation, Wind Speed (WS) and Wind Direction (WD) every 10 min. It is noted that the parameters are sampled every second, with 10 min averages transmitted to the shore station at the end of the 1 h. The Research Moored Array for African-Asian-Australian Monsoon Analysis and prediction (RAMA) buoy data are extensively used to understand air-sea fluxes and thermohaline structure in the equatorial Indian Ocean region and validation of satellite-derived products and models.

High resolution AT data obtained from the RAMA data buoy is used to develop a methodology to identify rainfall events. This algorithm is applied to the 10 min AT and precipitation data as the training to identify peak events. Some of the precipitation data were kept aside for validating the results. It will be shown in the later paragraphs that the algorithm can detect rain events from the AT data (rate of temperature drop) with high accuracy.

Extremes (maxima or minima) in the time series can affect the overall distribution of the data and can change the statistics. It is difficult to identify extremes as data outliers or some inherent forcing mechanisms that drive the parameters. The detection of peaks in time series data[12-14] has been applied in many research areas and applications, ranging from signal processing to quality control procedures.

After finding peaks in the AT drop rate, the time series is partitioned into "events" which later serve as input to the decision tree algorithm to predict rainfall events.

Random forest is used to predict events using classification and regression technique[15,16]. Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes of individual trees. Random forest increases the predictive power of the algorithm and prevents overfitting. Random forest is an ensemble-bagging algorithm designed to achieve low prediction error. It reduces the variance of the individual decision trees by randomly selecting trees and then either averaging them or picking the class that gets the most votes.

Classification accuracy alone can be misleading if there are an unequal number of observations in each class or if there are more than two classes in the dataset. A confusion matrix summarizes the performance of a classification algorithm. Calculating

a confusion matrix provides information on the idea of what the classification model is getting right and the type of errors it is making. Finding the peaks in AT drop rate using peak segment detection and using this information for the prediction of precipitation classes using a classification scheme is the central motivation of this study.

Materials and Methods

Data

AT and rainfall sampled at 10 min from RAMA buoy are available from 2009 to 2017 (Fig. 1). However, the time-series data of AT and rainfall is not continuous because of gaps. The gaps in the time series occur due to the absence of the buoy deployment and sensor failure. AT is measured by a Pt-100 RTD (Resistance Temperature Recorder) of Rotronic Instrument with a resolution of ±0.2 °C and rain by R.M. Young Self-Siphoning capacitance rain gauge with a resolution of ±0.4 mm/h. Accuracies and details of both of the above sensors are presented in the McPhaden *et al.*[11]. Both the datasets have gone through standard quality control checks, and only the data that passed through the prescribed checks were used in this study.

Eastern Equatorial Indian Ocean (EEIO) is a warm pool region and one of the climatically sensitive regions, which undergoes various time scales of variability, such as mesoscale eddies, intra-seasonal disturbances forced by the atmosphere, inter-annual variations associated with Indian Ocean Dipole (IOD) and the El Niño/Southern Oscillation (ENSO). Hence, the location is chosen for this study, apart from the data availability. The large variability in AT and precipitation is seen in Figure 2. These two datasets appear to be poorly correlated, with a correlation coefficient of -0.28 (Fig. 3). However, on examining the events using the highest resolution data (sampled every 10 min), the relation between the two variables is more evident. Figure 4 shows the occurrence of rain events along with AT time series. It depicts that whenever rain events occur, the AT drop events follow them. It is well documented and known that rain events in tropical regions cool the boundary layer and decrease the AT as the temperature of falling rain is generally lower[7]. However, there are other factors also that can cool the ambient air such as advection of cold air masses, local temperature tendency due to heat and radiation forcing. However, in this study, we aim to identify the number of cooling events

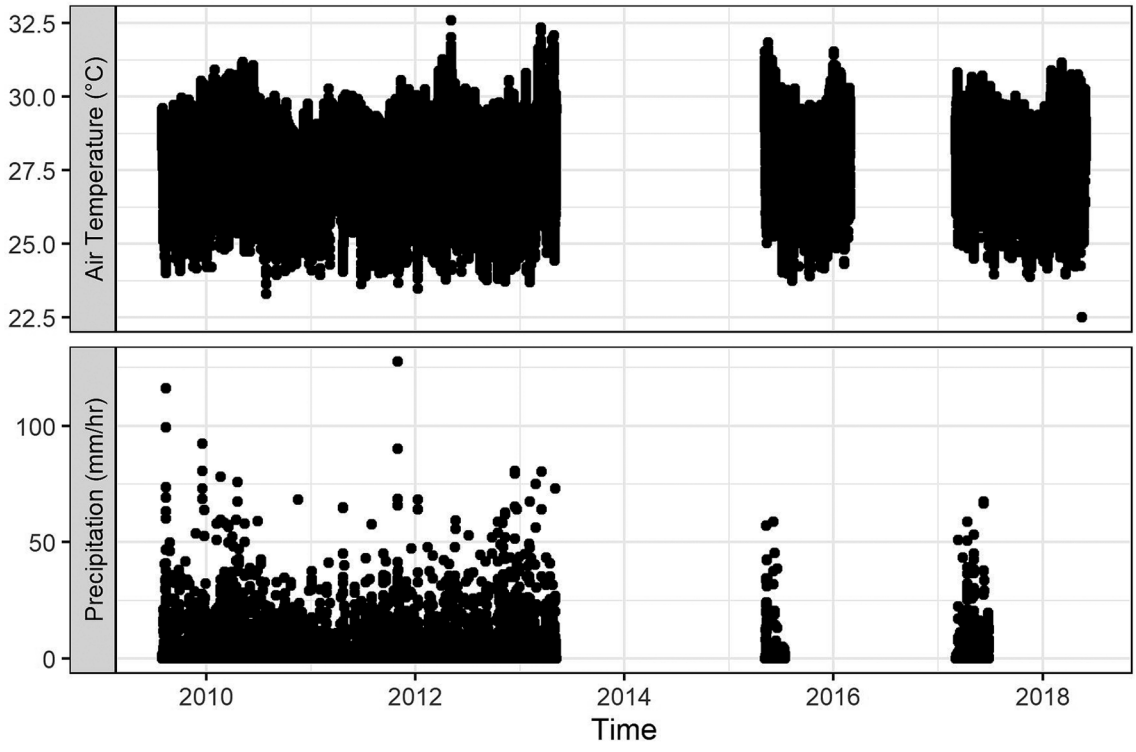

Fig. 2 — Time series of air temperature (AT; °C; top-panel) and precipitation (rain; mm/h; bottom panel) observations at the buoy location 0° N, 90° E

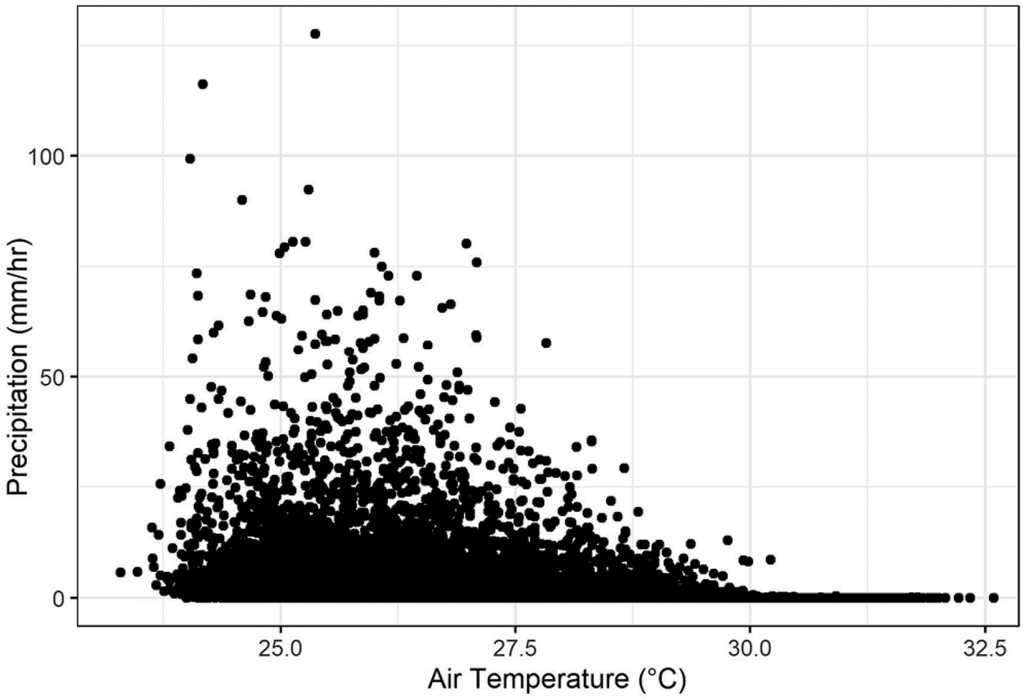

Fig. 3 — Scatter plot of AT (°C; on the x-axis) and precipitation (mm/h; on the y-axis) during the observation period. Correlation coefficient was -0.28

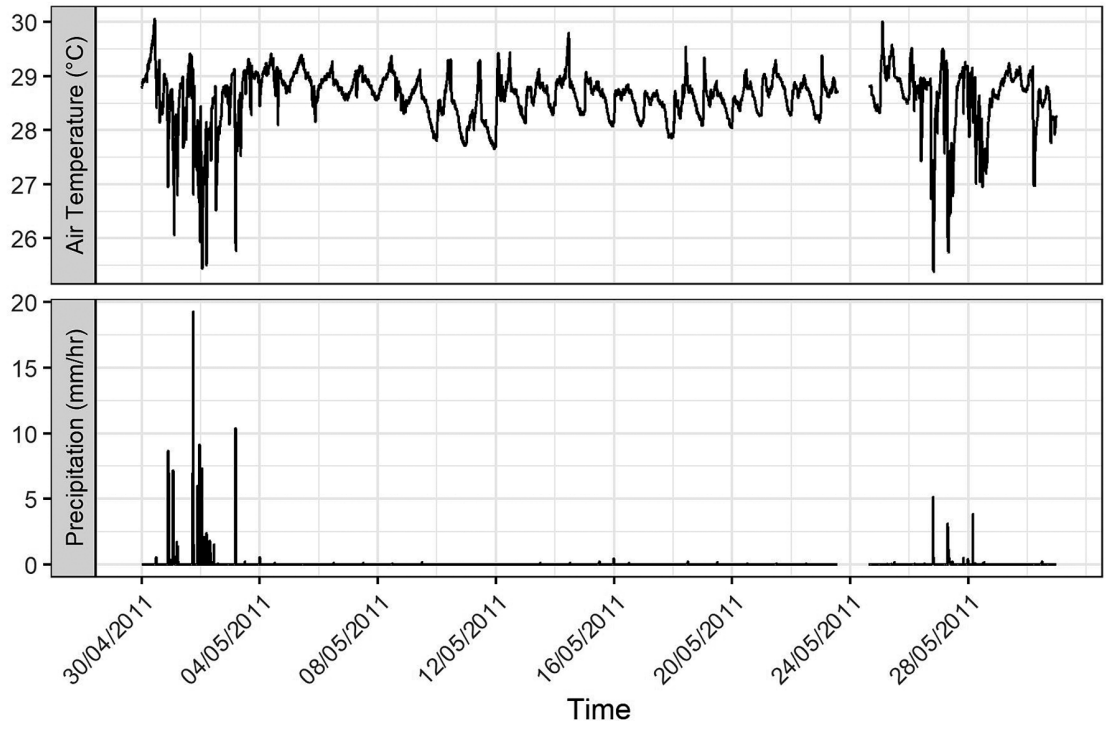

Fig. 4 — Time-series of AT (°C; top panel), and precipitation (mm/h; bottom panel) during a spell of rain 1 – 31 May 2011

associated with the rainfall and use the information in a predictive manner. The identification of these cooling events as reflected in AT drop will be a significant goal in identifying the rain events. A data-mining methodology is shown, which identifies the events that may lead to the precipitation events.

Methodology

Segmentation of time series

Segmentation is a process of dividing the time series into smaller data records, where the probability distribution of the time series shows significant change. A univariate time series *T* of *n* number of values defined as $T = \{(t_1, x_1), (t_2, x_2), \ldots, (t_n, x_n)\}$, where the *i*th observation is represented as pair of values $(t_1, x_1), t_i$ refers to the time at which observation is recorded and x_i refers to the observed variable, which may be AT or precipitation or any other parameter. Here, the x_i refers to AT.

We assume that a significant AT drop occurs only during a precipitation event (Fig. 4). The rain events can be identified from the AT drop events: the rain event starts from the rapid drop of the AT and the rainfall event ends when the AT reaches the mean state. For the above time series, another time-series variable is introduced, which is called as the maximum rate of drop (R) and defined as $R = \{(t_1, y_1), (t_2, y_2), \ldots, (t_n, y_n)\}$,

$$\text{Where, } y_k = \begin{cases} 0, k = 1 \\ max\left(\frac{x_p - x_k}{t_k - t_p}\right), 1 \leq p < k \leq n \end{cases}$$

The local minima in time series *R* are the timestamps *v* at which the $x_{v-1} > x_v > x_{v+1}$. Based on these assumptions the events are defined as the subset of time series *R*; from the valley prior and next to the segments where the values of y_i cross some threshold. The event segments of *R* can be defined as $E = \{(t_s, y_s), \ldots, (t_e, y_e)\}$,

$$\text{Where, } y_k = \begin{cases} <\propto, if k = s \text{ or } k = e \\ \geq\propto otherwise \end{cases}$$

Here, $\propto$ is a threshold, it may be a user-defined value or any suitably calculated value. In this study, $\propto$ is taken as 0.003.

Decision tree for classification and prediction

The second component of the proposed method is the classification of the segments which cross the threshold value. For this, a well-defined classification tree methodology based on multivariate analysis is utilized. Decision trees are supervised non-parametric machine learning algorithms which are used for both classification and regression problems. The ensemble method to generate a classification model called Random forest is used for the classification of these segments[15]. Random forest invokes two important machine-learning ideas: bagging and random feature selection. Bagging is a simple and powerful ensemble method that uses a bootstrap procedure for high-variance machine learning, especially decision trees to improve the accuracy and stability of the algorithm and reduces the variance to avoid over-fitting. Bagging combines the predictions from many machine learning algorithms to make more accurate predictions compared to a single model. Thus, bagging is a special case of the model averaging method. Random forest is an improvement over the bagged decision trees, as the resulting predictions from all the sub-trees have less correlation. To select the optimal split point, the learning algorithm looks through all the variables and values. Random forest randomly selects a subset of features to split at each point when growing a tree. In this study, the Random forest R package[17] is used for implementing the classification tree.

The performance of classification models is evaluated using the confusion matrix. There are only four possible outcomes for any event: (a) True Positive (TP) meaning the positive event is classified as positive; (b) False Negative (FN) meaning the positive event is classified as negative; (c) True Negative (TN) meaning the negative event is classified as negative; (d) False Positive (FP) meaning the negative event is classified as positive (Table 1). This confusion matrix (Table 1) can be used for the calculation of a variety of statistics for establishing the accuracy of the model.

The accuracy of the model is defined as the ratio of true positives and negatives to the total number of events ((TP+TN)/(TP+TN+FP+FN)) and is a measure of how often the classifier is correct. Misclassification rate, which is also an error rate, can be defined as the ratio of false positives and negatives to the total number of events ((FP+FN)/(TP+TN+FP+FN)) and is a measure of how often classification goes wrong[18]. True positive rate, which is also called sensitivity or recall, can be defined as the ratio of true positives to the sum of TP and FN (TP/(TP+FN)) and it is a measure of how often the classifier is predicting true as true. Similarly, a False positive rate, can be defined

Table 1 — Confusion matrix

		Actual	
		Event	No Event
Predicted	Event	TP	FP
	No Event	FN	TN

as the ratio of false positives to the sum of TN and FP (FP/(TN+FP)) and is a measure of how often the classifier is predicting false as true. True negative rate or specificity is the ratio of true negatives to the sum of TN and FP (TN/(TN+FP)) and is a measure of how often the classifier is predicting false as false. Precision can be defined as the ratio of true positives to the total number of positive predicated events (TP/(TP+FP)) and is a measure of how often the classifier is correct when predicting as true. Prevalence can be defined as the ratio of the sum of TP and FN to the total number of events ((TP+FN) / (TP+TN+FP+FN)); how often the actual true condition occurs in our sample. Finally, F-Measure which is a weighted harmonic mean of the test's precision and recall, can be defined as ((2*Recall*Precision)/(Recall+Precision)) and is a measure of the test's accuracy. These statistics are further discussed in the results section.

Results and Discussion

The period from July 2009 to June 2017 is considered for the analysis, where both AT and precipitation data are available. The minor gaps were filled using a median filter for both parameters and considered for further analysis without smoothing the data. Figure 5 shows the derived maximum drop rate series along with AT and rain for the period 08 – 14 August 2009. However, note that the whole time series of the maximum drop rate is utilized for the further segmentation of data and analysis.

The threshold to mark events through the peak window was derived by taking the mean of max-drop rate after removing the outlier values, which were beyond the limit of three standard deviations from the mean of maximum drop rate (mean value is ≈ 0.003 °C/min). Figure 6 shows the segmented time series (for the same duration as in Fig. 5), where each segment starts with the valley before the defined threshold and ends up with the valley after the same threshold. The algorithm was able to identify and delineate a total of 22461 segments for classification. For classification, 75 % of the segments were used for training data and 25 % for validation of the results.

Figure 7 represents the lag between the time it takes to reach max precipitation and the timestamp where the AT reaches the minimum level. The time lag is maximum at the first time steps (10 min). The

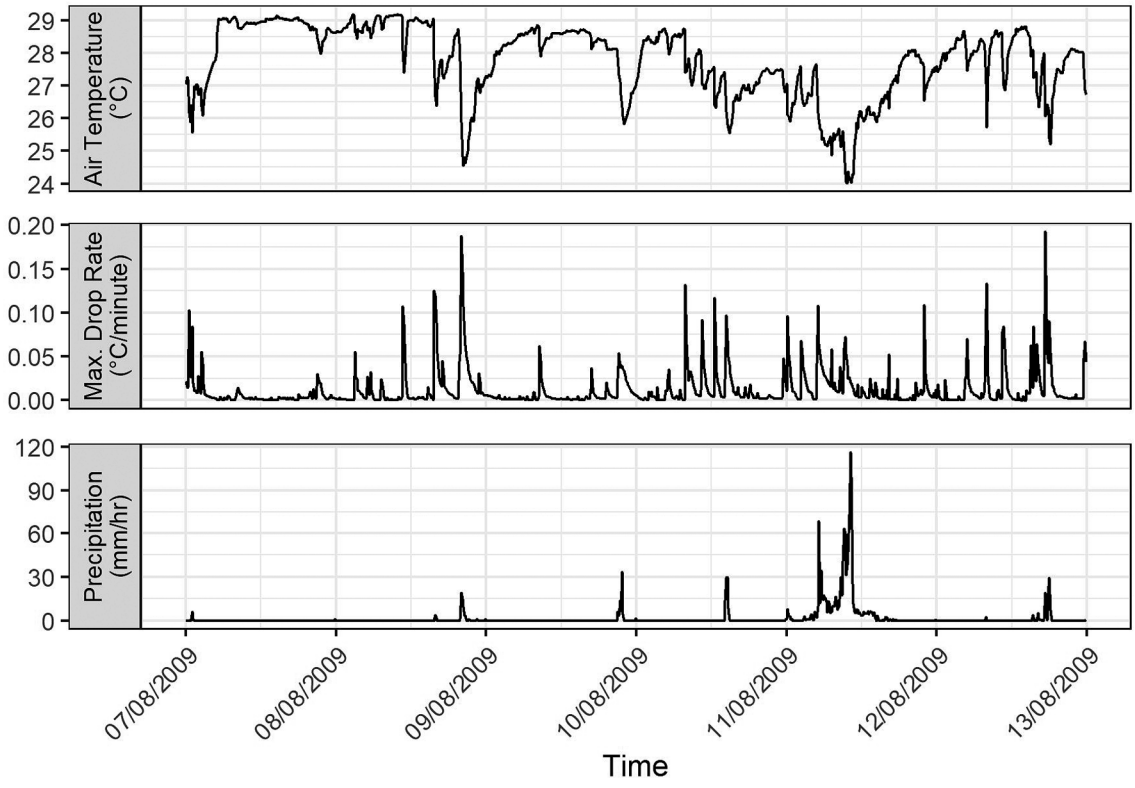

Fig. 5 — Time-series of AT (°C; top panel), Maximum AT Drop Rate (°C/min; middle panel) and rain (mm/h; bottom) panel during a spell of rain 7 – 13 August 2009)

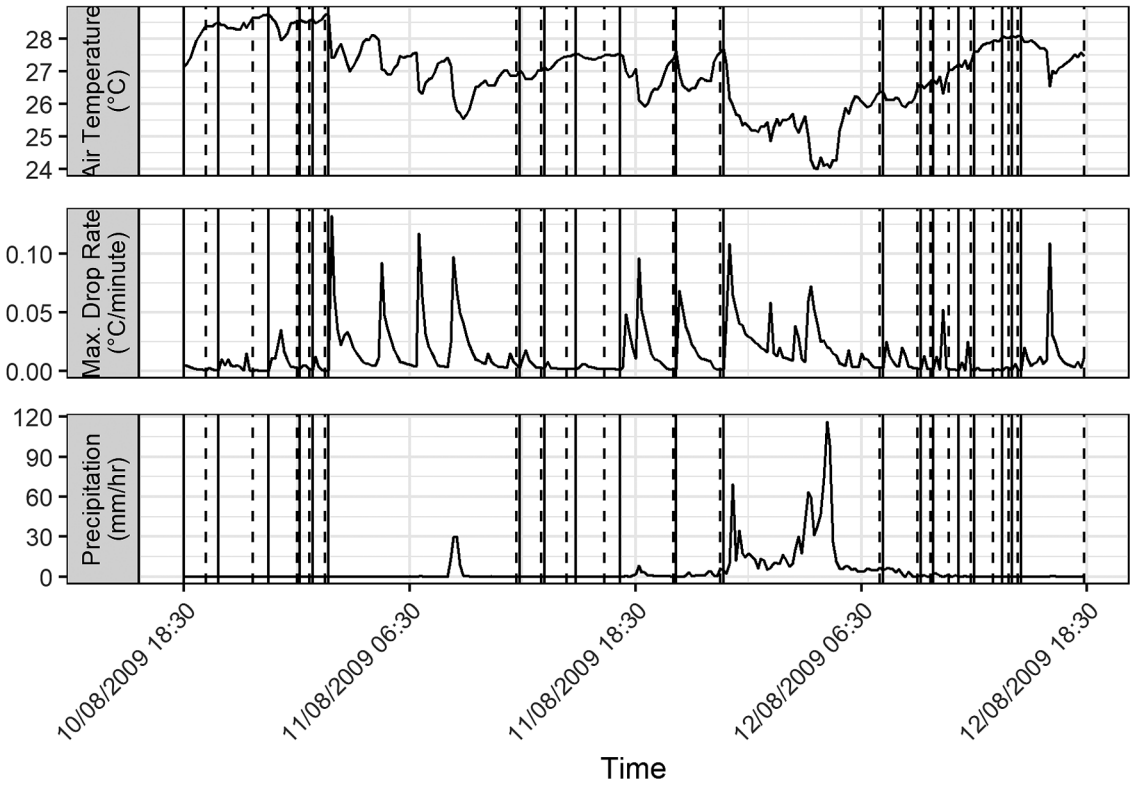

Fig. 6 — Same as Fig. 5, with segments overlaid as black vertical lines (the region between the solid line and dashed line)

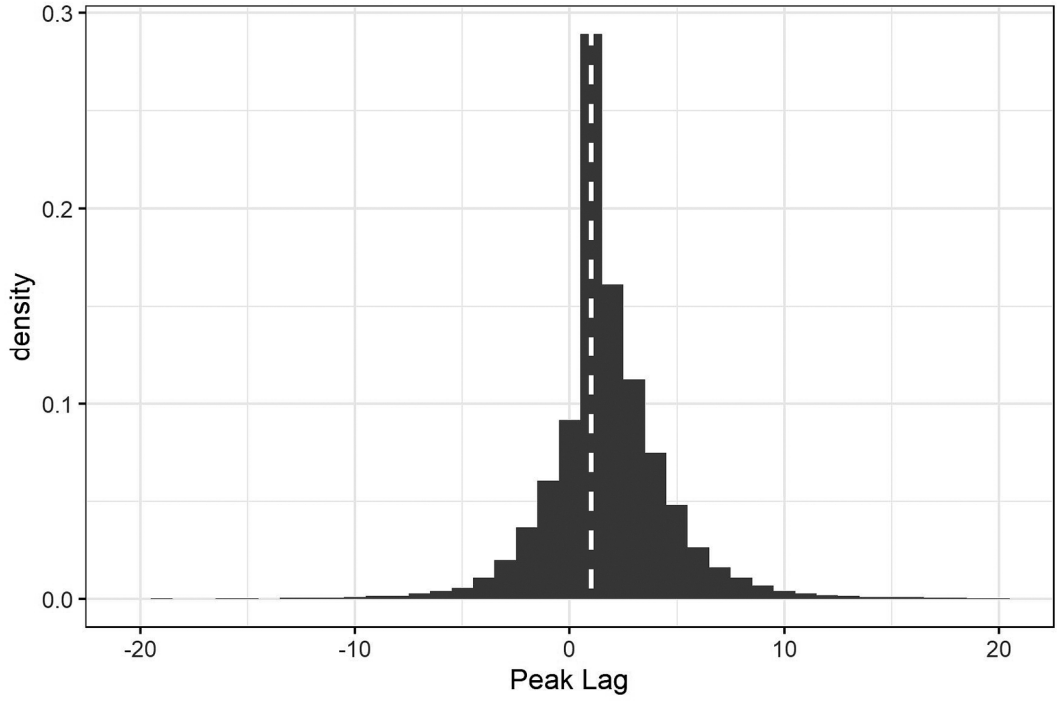

Fig. 7 — Histogram of peak lag (in steps of 10 min) between maximum precipitation and minimum air temperature drop period over the data record length

segmented data is then classified with the Random forest classification tree. For each segment, the variables measured are the length of each segment (length), minimum air temperature of segment (minAT), the maximum drop rate of air temperature in segment (maxDropRate) and total air temperature drop in segment (totalDrop). These variables are next used for the generation of a classification tree for the prediction of different classes of maximum precipitation in a given segment (maxPrecipitation). First training data are taken as input for the Random forest and the output classifier was used for validation of test data. Based on the maximum precipitation rate, two cases are considered which are discussed in the next subsections.

Case 1

In this case, the maximum precipitation rate is categorized into two categories: Negligible (< 2 mm/h) and Rainy (> 2 mm/h). It is a binary classification of the precipitation events: whether rain has occurred or not and the cutoff considered for rain is 2 mm/h. Table 2 depicts the confusion matrix generated from the classification process having the two classes.

In this classification, around 94 % of events were identified accurately (Table 3). As per the objective of present study, the true positive rate for class "Rainy" is also very high (94 %), which shows that the classifier is capable of identifying most of the rainy events from the time series of AT. Also, it shows that in test data only 18 events were predicted wrongly as non-rainy events, the false positive rate is around 6 %. Among these false-positive events, only one event was of significant precipitation event (maximum value was 12.6 mm/h) and for the rest of the events, the mean maximum precipitation value was around 4.2 mm/h.

Case 2

To account for the small scale variations in rainfall intensity and amount, the maximum precipitation rate is categorized into multiple categories based on prevalent standard meteorological conventions (Table 4). In this case, a multiple classification model of the maximum precipitation events is created. Table 5 represents the confusion matrix generated from the classification process using the same Random forest method but now having multiple classes. Here, the overall accuracy is around 83 %. The performance degradation is mainly due to the unbalanced classification, where the number of samples in negligible class is more dominating than in other cases. If the results are seen as multiple classification problems, negligible conditions are predicted mostly below heavy rains and torrential conditions are not predicted below heavy rains.

Table 2 — Confusion matrix (Case 1)

		Actual	
		Rainy	Negligible
Predicted	Rainy	266	346
	Negligible	18	4985

Table 3 — Accuracy and other statistics for Case 1

Accuracy	94 %
Misclassification rate (Error rate)	06 %
True positive rate (Sensitivity or Recall)	94 %
False positive rate	06 %
True negative rate (Specificity)	94 %
Precision	43 %
Prevalence	05 %
F-measure	59 %

Table 4 — Classification of precipitation events (Case 2)

Precipitation type	Range
Negligible	≤ 0.5 mm/h
Light	(0.5 – 2) mm/h
Moderate	(2 – 15) mm/h
Heavy	(15 – 30) mm/h
Very	(30 – 60) mm/h
Torrential	≥ 60 mm/h

Table 5 — Confusion matrix for classification (Case 2)

		Actual					
		Negligible	Light	Moderate	Heavy	Very	Torrential
Predicted	Negligible	4494	20	8	0	0	0
	Light	629	63	40	5	0	0
	Moderate	63	46	77	17	13	0
	Heavy	3	7	30	15	10	5
	Very	1	4	20	10	6	1
	Torrential	0	1	6	6	12	3

Conclusion

It is shown that the AT drop events coincide with the precipitation (rainfall) events. By identifying AT drop events and employing a classification algorithm, a methodology is developed to identify rain events. The developed methodology is based on defining peak windows to detect the maximum AT drop events in time series data. During the study period, a total of 22461 maximum AT drop events were singled out, and out of these events, more than 94 % of the events were correctly clustered into corresponding rain events. It was also observed that the effectiveness of this algorithm is more statistically significant for Case 1 where the events are classified as rainy and negligible-rain events when compared with Case 2 where the events were classified in 6 different groups (Negligible, Light, Moderate, Heavy, Very Heavy and Torrential). In Case 2 of multiple classifications, results are not as accurate (83 % overall accuracy) when compared with Case 1 (94 % accuracy). The degradation of the performance is due to the further division of rain events based on the rainfall intensity, which creates a large number of events for comparison. The methodology developed here gives the quantitative framework to identify the precipitation events due to the temperature drop data for large-scale events. The study showed that the temperature drop and length of the event are the leading indicators of rainfall. This data mining methodology improves the functionality of detecting sudden changes and trends in oceanic data using the algorithm as an automated procedure, which can be used in quality control checks on large datasets.

Acknowledgements

The encouragement provided by the Director, INCOIS is gratefully acknowledged. RAMA buoy data used in this study are freely available from the Pacific Marine Environmental Laboratory (PMEL), a laboratory in the National Oceanic and Atmospheric Administration Office (https://www.pmel.noaa.gov/gtmba/pmel-theme/indian-ocean-rama). Analysis and graphics are generated using R, an open-source software. This is INCOIS contribution number 460.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

RVS designed the experiments, analysed the data and wrote the manuscript. MR proposed the idea of the study and also contributed to writing the manuscript. KS contributed to the data analysis and writing of the manuscript. EPR & BVR contributed to writing the manuscript.

References

1 Bojinski S, Verstraete M, Peterson T C, Richter C, Simmons A, *et al.*, The concept of essential climate variables in support of climate research, applications, and policy, *Bull Am Meteorol Soc*, 95 (9) (2014) 1431-1443.

2 Hollmann R, Merchant C J, Saunders R, Downy C, Buchwitz M, *et al.*, The ESA climate change initiative: Satellite data records for essential climate variables, *Bull Am Meteorol Soc*, 94 (10) (2013) 1541-1552.

3 Serra Y L, Precipitation measurements from the tropical moored array: A review and look ahead, *Q J R Meteorol Soc*, 144 (2018) 221-234.

4 Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, *et al.*, The NCEP/NCAR 40-year reanalysis project, *Bull Am Meteorol Soc*, 77 (3) (1996) 437-472.

5 Huffman G J, Bolvin D T, Nelkin E J, Wolff D B, Adler R F, *et al.*, The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales, *J Hydrometeorol*, 8 (1) (2007) 38-55.

6 Xie P & Arkin P A, Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs, *Bull Am Meteorol Soc*, 78 (11) (1997) 2539-2558.

7 Katsaros K & Buettner K J, Influence of rainfall on temperature and salinity of the ocean surface, *J Appl Meteorol*, 8 (1) (1969) 15-18.

8 Schlössel P, Soloviev A V & Emery W J, Cool and freshwater skin of the ocean during rainfall, *Bound-Layer Meteorol*, 82 (3) (1997) 439-474.

9 Gosnell R, Fairall C & Webster P, The sensible heat of rainfall in the tropical ocean, *J Geophys Res Oceans*, 100 (C9) (1995) 18437-18442.

10 Serra Y L, A'hearn P, Freitag H P & McPhaden M J, Atlas self-siphoning rain gauge error estimates, *J Atmos Ocean Technol*, 18 (12) (2001) 1989-2002.

11 McPhaden M J, Meyers G, Ando K, Masumoto Y, Murty V, *et al.*, Rama: the research moored array for African-Asian-Australian monsoon analysis and prediction, *Bull Am Meteorol Soc*, 90 (4) (2009) 459-480.

12 Palshikar G, Simple algorithms for peak detection in time-series, In: *Proc 1st Int Conf Advanced Data Analysis, Business Analytics and Intelligence*, Vol. 122, 2009, pp. 14.

13 Jean-Paul, *Stack overflow*, Peak signal detection in realtime time series data https://stackoverflow.com/questions/22583391/peak-signal-detection-in-realtime-timeseries-data (2014).

14 Schneider R, *Survey of peaks/valleys identification in time series*, (Department of Informatics, University of Zurich, Switzerland), 2011, pp. 12.

15 Breiman L, Random forests, *Mach Learn*, 45 (1) (2001) 5-32.

16 Zakariah M, Classification of large datasets using Random Forest algorithm in various applications: Survey, *IJJEIT*, 4 (3) (2014) 189-198.

17 Liaw A & Wiener M, Classification and regression by Random Forest, *R news*, 2 (3) (2002) 18-22.

18 Visa S, Ramsay B, Ralescu A L & Van Der Knaap E, Confusion matrix-based feature selection, *MAICS*, 710 (2011) 120-127.