H2Open Journal

IWA PUBLISHING

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

Decision tree-based reduction of bias in monthly IMERG satellite precipitation dataset over India

Shushobhit Chaudhary and C. T. Dhanya*

Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India

*Corresponding author. E-mail: dhanya@civil.iitd.ac.in

Abstract

Decision trees are ideally suited for handling huge datasets and modelling non-linear relationships between different variables. Given the relationship between precipitation and bias may be very complex and non-linear, bias-correction of satellite precipitation is a challenge. We examine the applicability of Classification and Regression tree (CART) for bias-correction of the Integrated Multi-satellite Retrievals for Global Precipitation Mission (IMERG) precipitation dataset over India. The gauge-based 0.25° gridded precipitation dataset from India Meteorological Department is considered as the reference. The CART model is trained (2001–2011) and tested (2012–2016) over each 0.25° grids. The training dataset is subjected to 10-fold cross-validation and optimization of the minimum size of leaf node (one of the hyper-parameter). Efficiency of the CART model is evaluated using performance metrics like R^2, RMSE and MAB over the whole of India and different climate and elevation zones in India. CART model is observed to be highly effective in capturing the bias during the training (average $R^2 = 0.77$) and testing (average $R^2 = 0.66$) period. Significant improvement in average monthly MAB (−6.3 to 29.2%) and RMSE (8.7–37.3%) was obtained post bias-correction by CART. Better performance of CART model was observed when compared to two widely adopted bias-correction techniques.

Key words: CART, GPM, IMERG, precipitation bias reduction, regression tree, satellite rainfall

Highlights

- CART is effective in modelling the relationship between satellite precipitation bias and observed precipitation.
- Significant improvement in average monthly MAB and RMSE was obtained post bias-correction by CART.
- CART bias correction algorithm improved the correlation and reduced the MAB and RMSE when compared to linear scaling and CDF matching techniques.
- Though the CART algorithm is found to be robust, over hilly and mountainous regions, the CART model should be cautiously adopted, however.

INTRODUCTION

Recent decades have witnessed a surge in the usage of satellite-based precipitation products, for a variety of hydro-climatic applications and disaster management (Hong *et al.* 2006; Sawunyama & Hughes 2008; Behrangi *et al.* 2014; Koriche & Rientjes 2016). However, the applicability of

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

satellite-based precipitation for deriving reliable hydrological and climatic predictions is still limited, owing to the presence of bias in these measurements. The bias in satellite precipitation estimates may be owing to sampling uncertainty or error in the retrieval algorithm of the satellite estimation process (Hossain & Huffman 2008; Gebregiorgis *et al.* 2017; Prakash 2019). With the advent of multiple satellites for measuring precipitation, the uncertainty among these satellite-based precipitation estimates has also increased (Hossain & Huffman 2008; Abera *et al.* 2016). To minimize the uncertainty among the existing satellite precipitation product, different merged or blended satellite-based precipitation products are proposed, wherein information from multiple sources are combined to provide a high-quality satellite rainfall product (Beck *et al.* 2017; Awange *et al.* 2019; Bhuiyan *et al.* 2019).

The multi-satellite precipitation estimate from the recent Global Precipitation Measurement Mission (GPM), i.e., Integrated Multi-satellite Retrievals for GPM (IMERG) integrates information of various multi-satellite retrievals from NASA Tropical Rainfall Measurement Mission Multi-Satellite Precipitation Analysis (TMPA; inter-satellite calibration and gauge adjustment); NOAA Climate Prediction Center (CPC) morphing technique (CMORPH; Lagrangian time interpolation); Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) – Climate Classification System (Microwave calibrated Infra-Red); and the NASA Precipitation Processing System (input data assembly and processing) (Huffman *et al.* 2014). IMERG satellite precipitation has been available since June 2000 onwards at a very high spatial resolution of $0.1° \times 0.1°$ and multiple temporal resolutions (half-hourly, 3 hourly, one day, 7-day and monthly). IMERG has been extensively compared and validated using ground-based gauges and radar datasets over different regions of the world and its errors and biases are well studied (Tan *et al.* 2016, 2017; Tang *et al.* 2016; Asong *et al.* 2017; Prakash *et al.* 2018). The error or bias in IMERG dataset can be reduced by adopting a variety of statistical techniques like linear monthly scaling, quantile mapping based on an empirical distribution, etc., which are frequently used for bias correction of satellite estimates (Abera *et al.* 2016; Worqlul *et al.* 2018; Chaudhary & Dhanya 2019). Most of the bias correction algorithms applied over satellite precipitation are either parametric or based on the assumptions of linear relationships between predictor (satellite rainfall) and response (bias) (Habib *et al.* 2014; Abera *et al.* 2016; Bhatti *et al.* 2016; Hashemi *et al.* 2017; Worqlul *et al.* 2018). However, the relationship of bias and precipitation may be very complex and non-linear (Gebregiorgis & Hossain 2015). Moreover, the non-linear relationship between the precipitation and bias/error may also be affected by various factors such as landuse/landcover, soil-cover, elevation, climate zones, etc. (Gebregiorgis & Hossain 2012, 2015; Tang & Hossain 2012), which makes it difficult to capture using the usual statistical techniques.

Various machine learning algorithms such as artificial neural networks, k-nearest neighbour, support vector machines, classification and regression trees, etc., have been successfully used in various remote sensing and hydrological applications to effectively model the non-linear associations between multiple predictors and the response variables (Waheed *et al.* 2006; Rasouli *et al.* 2012; Erdal & Karakurt 2013; Tao *et al.* 2016; Mannan *et al.* 2018; Petty & Dhingra 2018; Bhuiyan *et al.* 2019). Moreover, machine learning algorithms are also effective in handling multi-dimensional and complex data (Tan & Beklioglu 2006; Kühnlein *et al.* 2014; Yu *et al.* 2014). In recent years, Classification and Regression tree (CART) based machine learning techniques have been widely adopted in different aspects of hydrology. Abraham *et al.* (2020) adopted CART for the classification of soil into hydrologic groups. Veettil & Mishra (2020) employed the CART for estimating the thresholds associated with climate, catchment, and morphological variables that may potentially influence the hydrological drought characteristic. Pekel (2020) applied the regression tree to estimate soil moisture considering different parameters like air temperature, time, relative humidity, and soil temperature. The applicability of CART for bias correction of satellite-based precipitation estimates remains unexplored till now on both global and regional scales. Given, the potential of CART in handling multi-dimensional and complex data, it would be essential to explore the applicability of CART for the reduction of bias in satellite precipitation.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

In the present study, we explore the possibility of applying Classification and Regression Tree (CART) for possible bias reduction of IMERG precipitation dataset over India. The specific objectives of this study are – (i) to train and test the CART model (wherein, satellite rainfall is the predictor variable and observed rainfall is the response variable) grid-wise over the whole of India for bias reduction of satellite rainfall; (ii) to analyze the performance of CART model over different climate and elevation zones in India; and (iii) to compare the performance of CART-based bias correction techniques with widely adopted and established bias correction techniques. Such an analysis will serve to improve the reliability of IMERG precipitation for applications in water resource and hydrological applications over India.

STUDY AREA

The land region of the Indian sub-continent covering from 66°E to 100°E and from 6°N to 39°N is chosen as the study area. India is a country with varied geophysical and climatic features. The northern part of India is surrounded by the mountainous range of Himalayas, while the southern part is surrounded by sea/oceans (Figure 1, panel a). Western Ghats, located near to the south-western coast receives heavy rainfall. The rainfall during the monsoon months of June, July, August and September, account for more than 75% of the annual rainfall of the country and is highly significant for the agriculture-dominated country. India has seven main Koppen–Geiger climate zones showing different climatic characteristics (Peel *et al.* 2007) (Figure 1, panel b). A brief description about different Koppen–Geiger climate zones over India is shown in Table 1.

DATA USED

Integrated multi-satellite retrievals for GPM (IMERG)

The Integrated Multi-satellite Retrievals for GPM (IMERG) is a U.S. multi-satellite rainfall product from the Global Precipitation Mission (GPM) team. IMERG algorithm estimates precipitation from

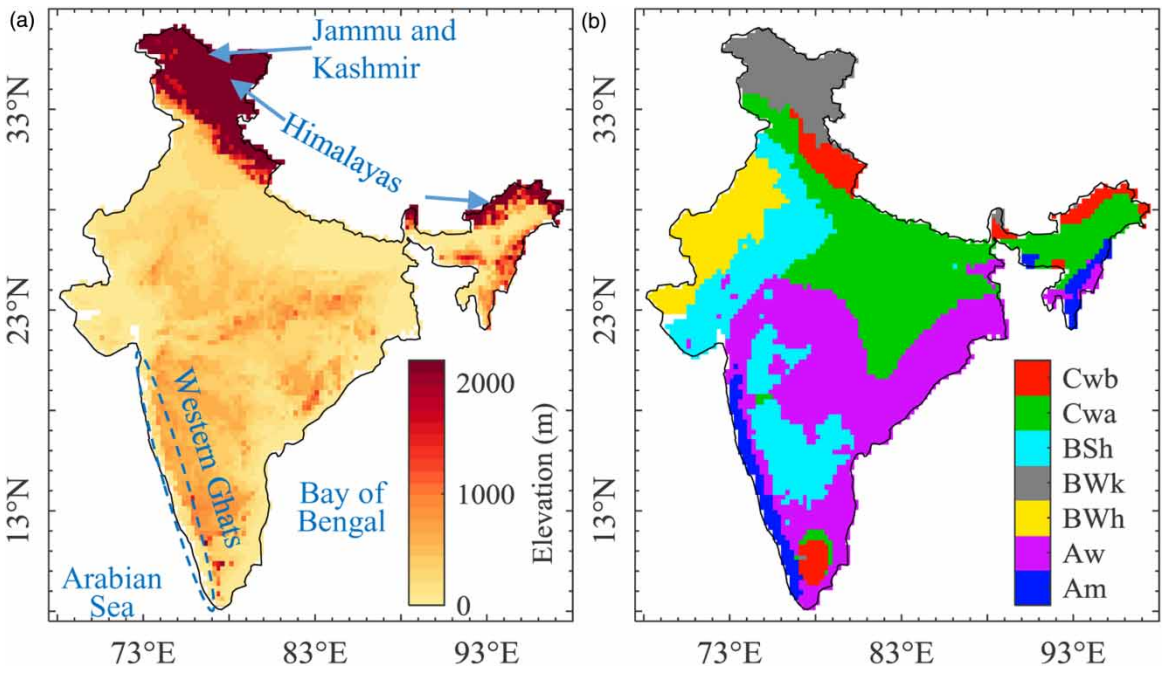

Figure 1 | (a) Digital Elevation Model (DEM) of study region (India) obtained from Shuttle Radar Topography Mission (SRTM) data (*Source*: https://earthexplorer.usgs.gov/). (b) Updated Koppen–Geiger climate zones over India (Peel *et al.* 2007).

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

Table 1 | Details of updated Koppen–Geiger climate zones over India

Climate Group	Climate Zone Code	Climate Zone Description	Location
A (Tropical)	*Am*	Tropical Monsoonal	Windward side of Western Ghats and a few parts of North-East India
	Aw	Tropical Savannah	peninsular India
B (Arid)	*BWh*	Arid Hot Desert	North-western desert regions
	BSh	Arid Steepe Hot	Leeward side of Western Ghats and semi-arid region in eastern India
	BWk	Arid Desert Cold	Himalayan regions
C (Temperate)	*Cwa*	Temperate Dry Winter Hot Summer	Gangetic plain region and North-East India
	Cwb	Temperate Dry Winter Warm Summer	Lower Himalayan ranges and few parts of Northern North-East India

Source: Peel *et al.* (2007).

various satellites in the GPM constellation, viz., Tropical Rainfall Measurement Mission Multi-Satellite Precipitation Analysis (TMPA), Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks- Climate Data Record (PERSIANN), Climate Prediction Center Morphing method (CMORPH), etc. (Hou *et al.* 2014; Huffman *et al.* 2014). GPM is the successor of TRMM and has three noteworthy improvements. Firstly, it has higher global coverage due to the increased orbital inclination from 35° to 65°. Secondly, it can measure very light precipitation (<0.5 mm/hour; Hou *et al.* 2014) owing to increased sensitivity of precipitation radar sensor due to the inclusion of two additional frequencies. Thirdly, solid form of precipitation (snow) can also be measured due to the inclusion of high-frequency channels in the passive microwave sensor imager. The IMERG data is available at a very high spatial resolution of $0.1° \times 0.1°$ and multiple temporal resolutions (half-hourly, 3-hourly, one-day, 7-day and monthly). Version 6 and Level 3 monthly IMERG rainfall data (Research/Final Run) available since June 2000 is used in the present study (https://gpm.nasa.gov/data-access/downloads/gpm). The default units of precipitation in the IMERG datasets is mm/hr, which is transformed to mm/day by accumulating over the entire day (multiplying by 24).

India Meteorological Department (IMD) rainfall

The $0.25° \times 0.25°$ gridded gauge-based daily rainfall dataset developed by India Meteorological Department (IMD) (Pai *et al.* 2014) is considered as the reference dataset (http://imdpune.gov.in/Clim_Pred_LRF_New/Grided_Data_Download.html#), for the estimation of bias. The IMD dataset is constructed by incorporating daily rainfall records information of approximately 6,955 rain gauge stations covering India. However, the availability of gauge station data varied vividly throughout the years. On average, gauge data from 2,600 stations is available per year for construction of data (Pai *et al.* 2014). Information from all the rain gauge data was subjected to various quality checks and then interpolated to $0.25° \times 0.25°$ grids using the Inverse Distance Weighted (IDW) scheme. Orographic dependence of rainfall is accurately captured by the IMD rainfall dataset. For instance, heavy rainfall is observed on the windward side of Western Ghats while low rainfall is observed on the leeward side of Western Ghats (Pai *et al.* 2014; Chaudhary *et al.* 2017). It is noteworthy that the gauge density is low over the hilly regions of northern India including the Himalayas and Jammu and Kashmir, therefore, IMD rainfall is less reliable over those regions.

The $0.1° \times 0.1°$ gridded IMERG dataset was resampled to the $0.25° \times 0.25°$ gridded IMD reference dataset using nearest-neighbour interpolation algorithm for further analysis.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

METHODOLOGY

Bias is defined as the deviation in the magnitude of satellite rainfall from gauge observed rainfall as shown in Equation (1).

$$Bias = P_s - P_g \tag{1}$$

where P_s is satellite (here, IMERG) precipitation; P_g is gauge (here, IMD) precipitation

Additionally, we have adopted the Willmott scheme to decompose the total mean square error (MSE) into its systematic and random component (Willmott 1981; AghaKouchak *et al.* 2012; Prakash *et al.* 2015). The total, systematic and random MSE is derived using Equation (2).

$$\frac{1}{n} \times \left(\sum_{i=1}^{n} (P_s - P_g)^2 \right) = \frac{1}{n} \times \left(\sum_{i=1}^{n} (P_s^* - P_g)^2 \right) + \frac{1}{n} \times \left(\sum_{i=1}^{n} (P_s - P_s^*)^2 \right) \tag{2}$$

where P_s^* is defined using a linear regression error model $P_s^* = a \times P_s + b$, where a being the slope and b being the intercept.

The left-hand side of Equation (2) represents the total MSE which is decomposed into systematic MSE (middle term in Equation (2)) and random MSE (rightmost term in Equation (2)).

Classification and regression tree (CART)

Classification and regression trees (CART), also called decision trees are data structures that are developed by recursively partitioning the input space, and defining a local model in each of the partitioned regions of input space (Murphy 2012). The decision trees are graphically represented with a primary root node at the top, subsequently leading to branches (internal nodes) and finally ending in outer nodes called leaves (terminal or decision nodes). A sample regression decision tree is shown in Figure 2, panel a.

The CART tree is progressed at each internal node by splitting the instance space into two or more subspaces according to a certain discrete function of the input attributes values. To understand the CART algorithm, let $x = x_1, \ldots\ldots x_n$ be the explanatory (predictor, or features; satellite rainfall in the present study) variables and $y = y_1, y_2 \ldots\ldots y_n$ be the target response or dependent (predictand or class; corrected satellite rainfall or observed rainfall in the present study) variables. Here n is the number of predictor and predictand variables (the total number of months in this study). The process of building a CART model involves partitioning the input space into distinct and non-overlapping regions and then predict the responses in each region using the mean of all the response variables in that region. In CART, we divide the predictor space (x) i.e., $x_1, x_2 \ldots\ldots x_n$ into J distinct and non-overlapping regions – … These regions can take any shape and are formed such that they maximize the reduction in the mean squared error (MSE) over all the splitting candidates. To partition the predictor space into J regions, the top-down recursive binary splitting approach is adopted. The splitting initially begins at the top of the tree, where two regions of partitioned, and then successively the above-partitioned regions are further split and the process continues till the optimization criteria are met (MATLAB 2019; Pekel 2020).

To understand the splitting procedure at any node, let us select a node j that is subjected to binary recursive partitioning. The node j has p predictor variables, i.e., $x_1, x_2 \ldots\ldots x_p$. The MSE of the responses in node j, i.e., $y_1, y_2 \ldots\ldots y_p$ is computed using Equation (3).

$$MSE_j = \sum_{i=1 \text{ to } p}^{i \in R_j} (y_i - \bar{y}_i)^2 \tag{3}$$

where $\bar{y}_i$ is the mean response for the predictor variables within the j^{th} box.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

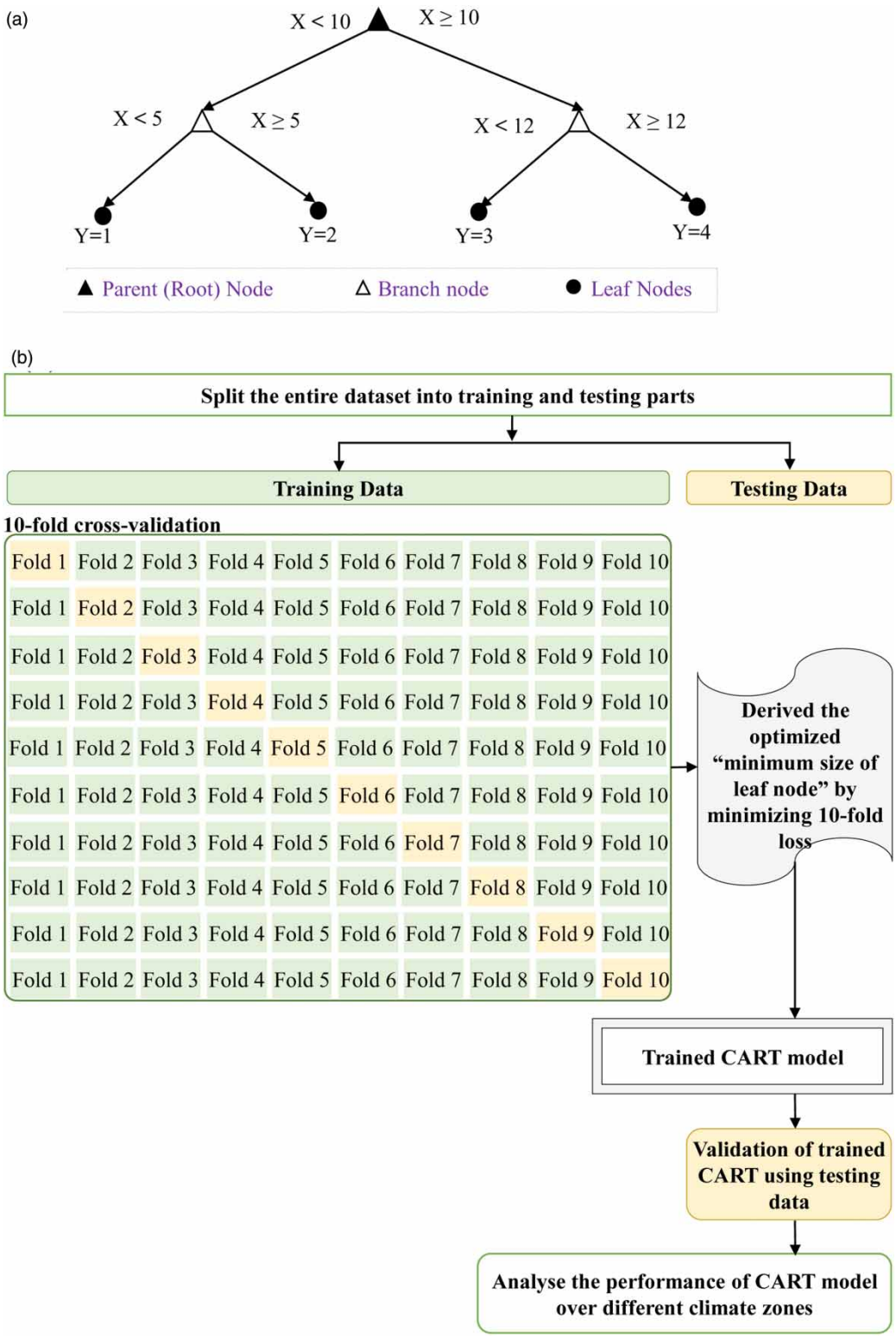

Figure 2 | (a) Illustration of a simple regression decision tree; (b) Methodology adopted for training and testing CART model in the present study.

The objective is to identify a threshold (th_j) of a split of x_j at node j such that two partitioned regions exist, i.e., $\{x|x_j < th_j\}$ (left region) and $\{x|x_j \geq th_j\}$ (right region), where th_j is selected such that it leads to maximizing the reduction in MSE (ΔI) over all the splitting threshold candidates, as shown in

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

Equation (4).

$$\Delta I = MSE_j - \{MSE_j\}_{\{x|x_j<th_j\}} - \{MSE_j\}_{\{x|x_j\geq th_j\}} \quad (4)$$

Now this process is recursively repeated until the reduction in MSE is maximized in all the regions (R_1, $R_2 \ldots\ldots R_J$). The splitting of nodes in subsequent nodes is also governed by controlling different hyper-parameters of a decision tree. The maximum number of splits of branch nodes (decision splits), minimum number observations in the branch node (size of parent node) and minimum number of observations in the leaf (size of leaf node) are the generally controlled hyper-parameters of the tree.

To implement CART, we use the statistical classification and regression tree feature available in the Statistics and Machine Learning Toolbox in MATLAB (MATLAB 2019).

Training and testing of CART

The dataset is split into two distinct parts for training and testing of developing the CART model. In this study, 2001–2011 is considered as the training period (132 monthly datasets for training) and 2012–2016 is considered as the testing period (60 monthly datasets for testing). The training dataset is subjected to 10-fold cross-validation (Waheed *et al.* 2006; Bhuiyan *et al.* 2019; Pekel 2020) as shown in Figure 2, panel b. In cross-validation, the training dataset is split into 10 random parts. Out of these 10 parts, one of the parts is retained as the validation data and the remaining 9 data parts are used to train the tree. The cross-validation process is repeated 10 times, till each of the 10 parts serve as the validation data at least once, which yields 10 different decision trees. Cross-validation while training the model leads to improved accuracy of the resulting tree, when it is tested on new data.

In this study, we train the decision trees, by optimizing only one of the hyper-parameter – minimum number observations in the leaf (size of leaf node) at each grid over India. Minimum leaf size puts a limit to split a node when the number of observations in one of the subsequent child nodes is lower than the minimum leaf size. The size of leaf node controls the depth of the decision tree and thus aids in preventing a tree from overfitting the data. We vary the size of leaf node from 0 to 100 (approximately 50% of data size) and train the regression tree for each of the above leaf node size using 10-fold cross-validation. The 10-fold cross-validation loss, which is the total MSE of all the folds is estimated for each of the tree leaf size. Finally, the leaf node size which gives the minimum 10-fold loss is selected as the optimum leaf node size of the training decision tree.

The training of the CART model is performed at each grid of India and the trained CART is subjected over the independent validation period. The performance of the CART model during training and testing is also verified using the four performance metrics – Root Mean Square Error (RMSE), Coefficient of Determination (R^2), adjusted R^2 and Mean Absolute Bias (MAB) as shown in Table 2.

Table 2 | Details of performance metrics used in the present study

Model performance measures	Equation	Range	Ideal Value
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)}$	[0, ∞)	0
Coefficient of Determination (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(y_i - \bar{x}_i)^2}$	(−∞, 1]	1
Mean Absolute Bias (MAB)	$MAB = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - x_i\rvert$	[0, ∞)	0

n is the total number of values; y is the predicted variable; x is the observed variable

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

RESULTS AND DISCUSSION

Spatial variation of rainfall and bias over India during different seasons

The spatial distribution of 16 years (2001–2016) mean monthly rainfall over India from the observed IMD and satellite-based IMERG dataset during the annual period, winter season (January, February; JF), summer season (March, April, May; MAM), monsoon season (June, July, August, September; JJAS) and post-monsoon season (October, November, December; OND) are shown in Figure 3. The IMERG dataset is effective in capturing the mean annual and seasonal spatial rainfall patterns of the reference IMD dataset. The heavy rainfall over the Western Ghats and North East India during the monsoon season and southern peninsular region during the post-monsoon season is well captured by the IMERG dataset (Figure 3, panels d, e, i, j). Although spatial rainfall pattern is well represented, slight deviations in intensities of IMERG rainfall with respect to IMD rainfall are observed in the majority of grids over India. IMERG over-estimates the all-India mean rainfall as obtained from IMD, especially during the monsoon period. On an annual scale, 3.1 mm/day of all-India mean rainfall is observed in IMERG slightly higher than all-India mean IMD rainfall 3.0 of mm/day (Figure 3, panels a, f). The spatial variation of bias intensity, i.e., deviation in the magnitude of IMERG from IMD (IMERG rainfall – IMD rainfall), is estimated for the annual period and different seasons in Figure 3, panels k–o. High magnitude of bias is evident over the Western Ghats, North-East India and northern state of Jammu and Kashmir. Over the windward side of Western Ghats, a considerably high dry bias (under-estimation) is observed in the IMERG dataset on contrary to the

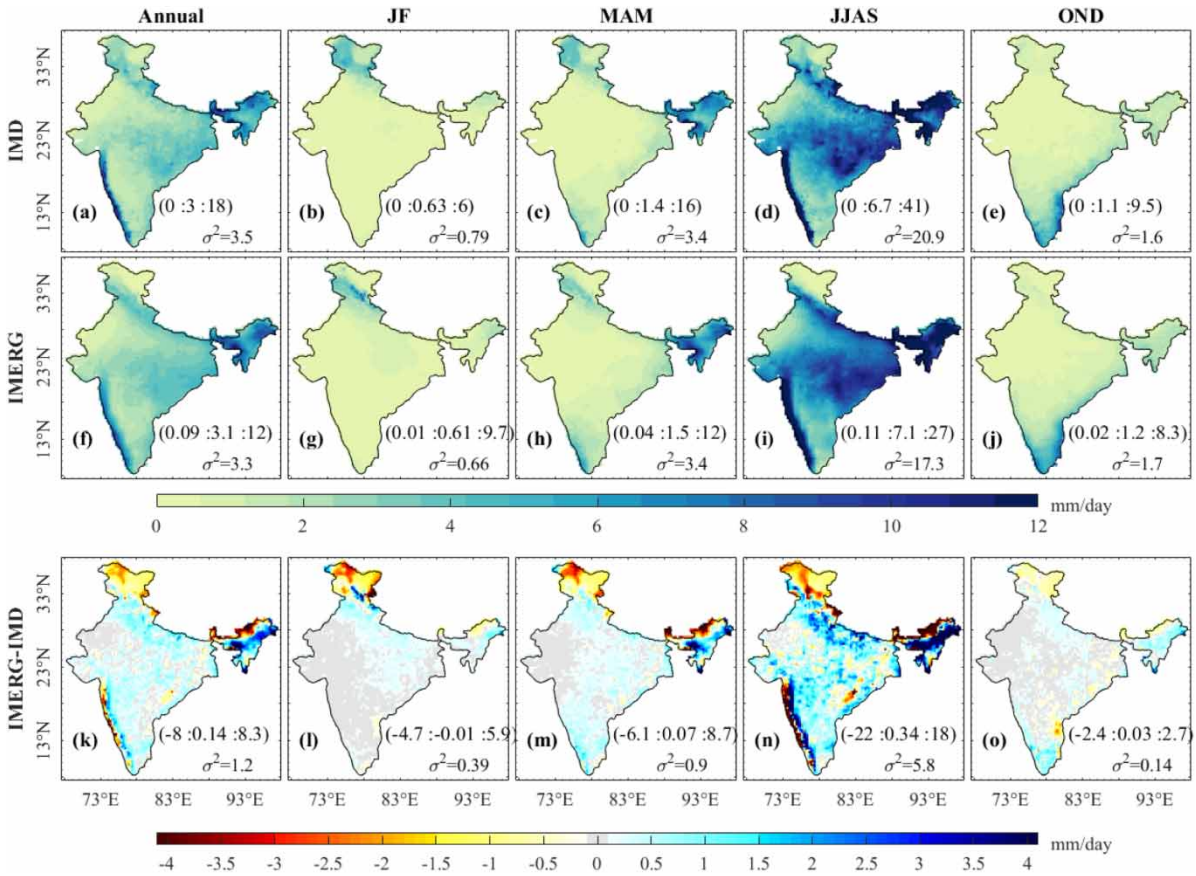

Figure 3 | Spatial variation of (a–e) IMD rainfall, (f–j) IMERG rainfall and (k–o) bias (IMERG – IMD) in rainfall during the annual, winter (JF), summer (MAM), monsoon (JJAS) and post-monsoon (OND) season over India during 16 years duration (2001–2016). All-India minimum, mean and maximum values are given in parentheses (minimum: mean: maximum). Spatial variance (σ^2) is also provided. All units are in mm/day.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

leeward side where significantly high wet bias (over-estimation) is observed (Figure 3, panels k, n). Consistent dry bias is observed in the northern state of Jammu and Kashmir in IMERG dataset over all the seasons. The IMD rainfall is also highly uncertain over these regions as it is highly mountainous and has significantly less gauge network station density (Prakash *et al.* 2019). The spread and intensity of bias are significantly higher during the monsoon season, as evident from Figure 3, panel n.

Further, the MSE in satellite-based IMERG rainfall is decomposed into systematic and random components over India as shown in Figure 4. Larger MSE is generally observed over high rainfall regimes (like Western Ghats and North-East India) and foothills of Himalayas. The spatial variations in systematic and random error components over India expressed in percentages of total MSE are shown in Figure 4(b) and 4(c). Larger systematic error observed over these mountainous zones of North-East India and Jammu and Kashmir can be either attributed to higher uncertainty of satellite rainfall over mountainous zones or uncertainty in observed rainfall due to lack in the required number of gauge station to capture rainfall variability in these zones (Prakash *et al.* 2015). Overall, a higher percentage of random MSE compared to systematic error is observed over India. Systematic error can potentially be corrected using bias correction schemes, whereas random error cannot be possibly reduced but can be statistically quantified (AghaKouchak *et al.* 2012; Prakash *et al.* 2015).

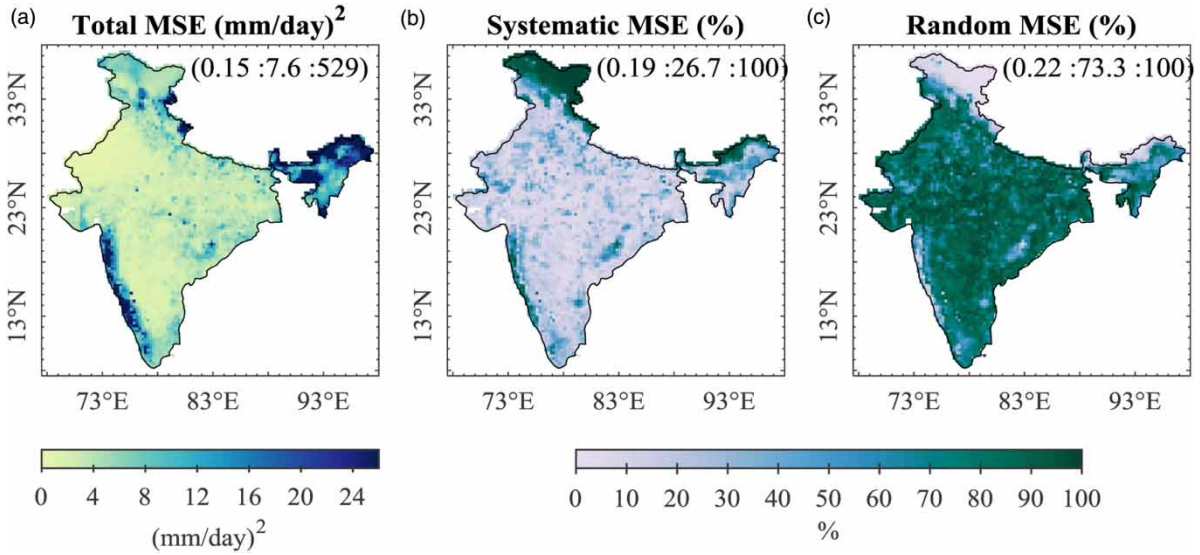

Figure 4 | Spatial variation of (a) Total MSE (b) Systematic MSE and (c) Random MSE in IMERG rainfall with reference to IMD rainfall over India during 16 years duration (2001–2016). All-India minimum, mean and maximum values are given in parentheses (minimum: mean: maximum).

Training and testing of CART over India

The optimum leaf size of the decision tree is estimated at each grid over India by minimizing the total 10-fold cross-validation MSE during the training period. Figure 5(a) shows the variation of total cross-validation loss of 10-fold with different minimum leaf size obtained while training the CART tree on a grid selected (75°E, 15.75°N) over Western Ghats in India. The minimum leaf size of 12, which observes the minimum total 10-folds cross-validation loss of 1.25 (mm/day)^2, is selected as the calibrated model parameter for the tree. The classification tree developed at the selected grid by using the optimized leaf size can be graphically visualized in Figure 5(b). Here, 'r' denotes the satellite rainfall and the values at node indicate the CART corrected satellite rainfall. At the parent node, the value of the rainfall threshold split is observed to be 2.8 mm/day. If the rainfall is greater than 2.8 mm/day,

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

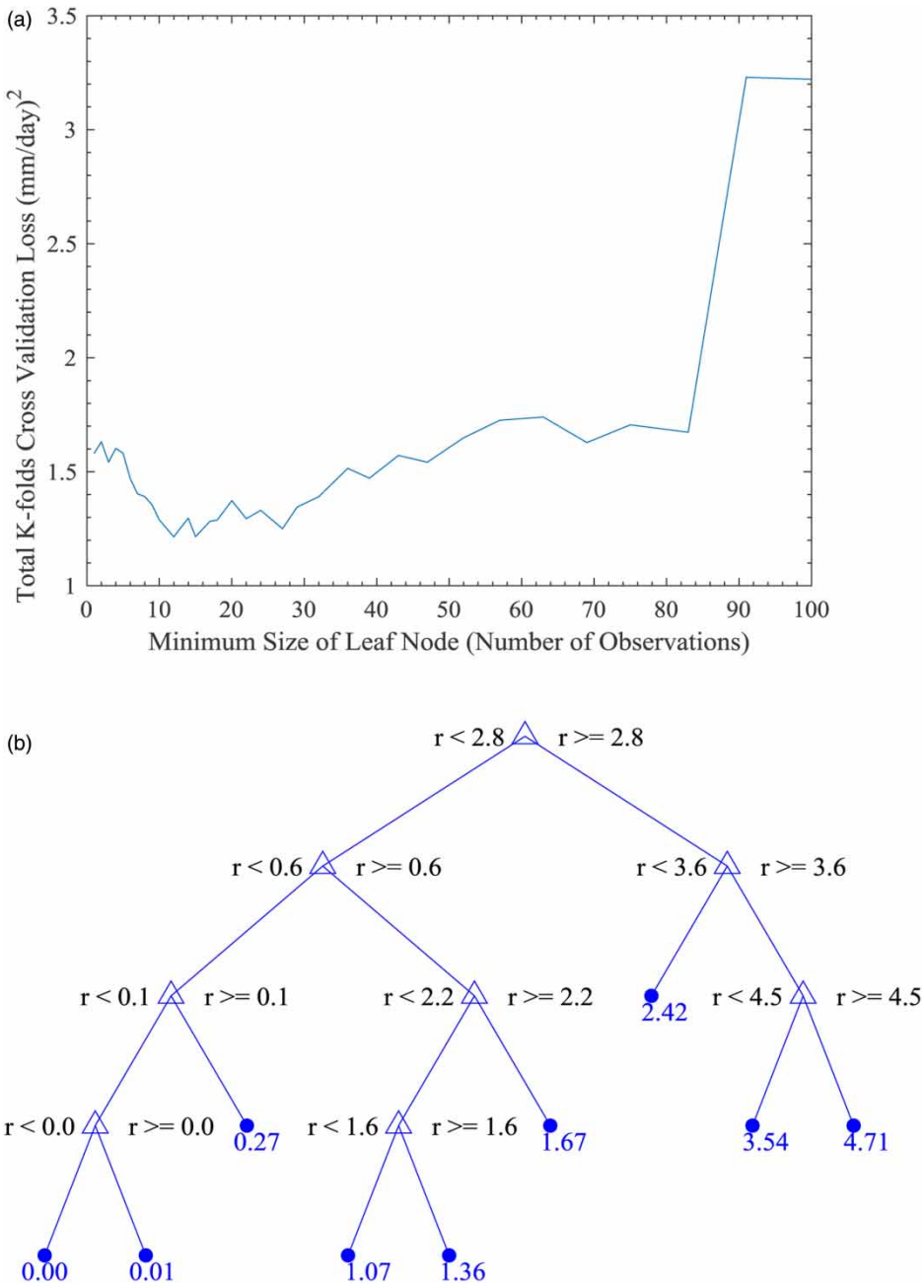

Figure 5 | (a) Variation of total cross-validation loss of 10-folds with different minimum leaf size obtained while training the CART tree; (b) CART decision tree obtained after training the data over a grid selected (75 °E, 15.75 °N) in India. 'r' indicates IMERG rainfall and values at the leaf (in dark blue) indicate the corrected IMERG rainfall.

further if rainfall is greater than 3.6 mm/day, further it is greater than 4.5 mm/day, the corrected satellite rainfall is 4.71 mm/day. Similarly, we can generate rules using the other branches of tree as shown in Figure 5(b).

The optimized minimum leaf size obtained during the training of the CART model at each grid over India is shown in Figure 6, panel a. Higher number (>30) of observations at the leaf nodes (minimum leaf size) were required to train the model in hilly and mountainous regions, where elevation is greater than 3,000 m. On an average minimum, 10 observations were required at the leaf nodes of the CART tree over India. Figure 6, panel b shows the rainfall threshold split at the root node of the CART tree trained over India. Higher (lower) rainfall thresholds were observed at heavy (low) rainfall regions, indicating the possible dependence of root node threshold split on the rainfall pattern over India.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

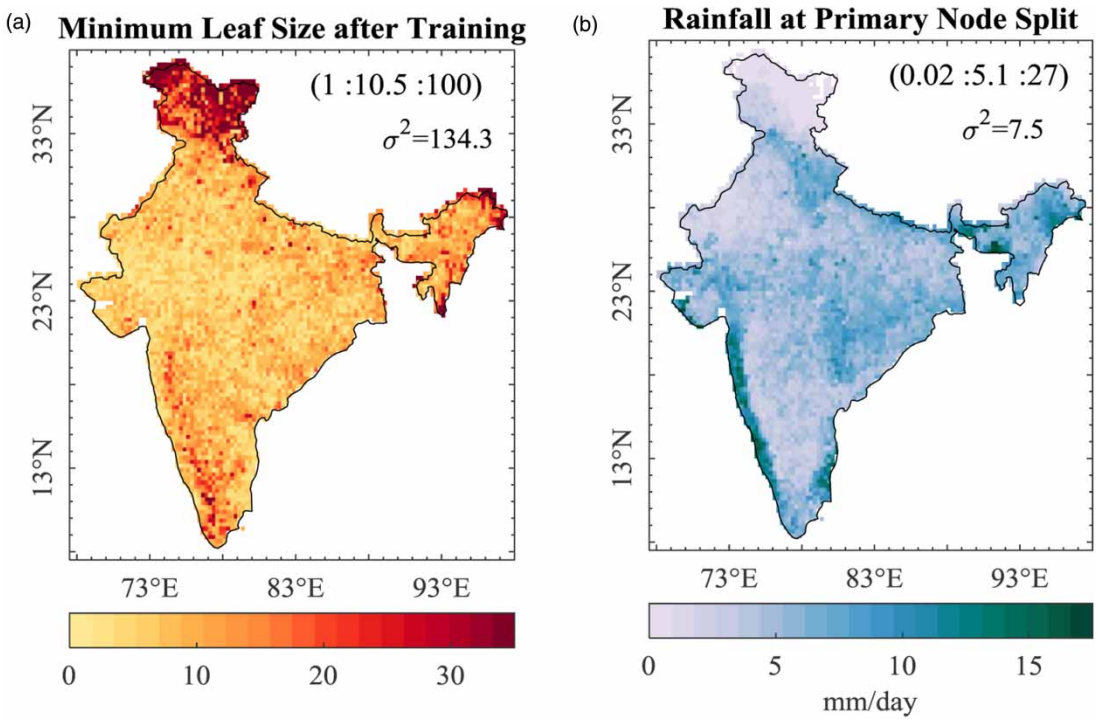

Figure 6 | Spatial variation of (a) minimum leaf size of decision trees obtained during training of CART model and (b) rainfall magnitude obtained at parent node split of CART model developed over the annual period of 2001–2011. All India mean value is indicated in parentheses. All-India minimum, mean and maximum values are given in parentheses (minimum: mean: maximum). Spatial variance (σ^2) is also provided.

The performance of the CART model during the training and testing period is analyzed over India using four performance metrics – Coefficient of Determination (R^2), adjusted R^2, Root Mean Square Error (RMSE) and Mean Absolute Bias (MAB). Figure 7, panel a and b shows the spatial distribution of R^2 during the training and the testing period. Higher R^2 is observed in most of the grids in India except for the states of Jammu and Kashmir, leeward side of Western Ghats and few grids in the southernmost corner of India. The all-India average R^2 of 0.77 and 0.66 is observed during the training and testing period respectively. Higher RMSE error was observed over the leeward side of Western Ghats and hilly regions of North East India (Figure 7, panel c and d). Similarly, high magnitude adjusted-R^2 were observed in both the training and testing period as shown in Figure S1, panel a and b. Lower magnitude of MAB, i.e., all-India average of 0.95 mm/day during the training period and 1.2 mm/day during the testing period were also observed (Figure S1, panel c and d). The above results show that the performance of the CART model is satisfactory and can be used over India for bias correction of monthly satellite estimates. However, over hilly and mountainous regions, the CART model should be cautiously adopted.

Performance of CART over different climate zones of India

The performance of the CART algorithm is evaluated over different climate zones of India. Figure 8 shows box plots that describe the variation of different R^2 and RMSE values in 7 different climate zones of India during the training and testing periods. Least R^2 (<0.4) is observed in the *BWk* climate zone during the training and testing periods. Similarly, moderately lower R^2 values, however, with high variability in R^2 (0.3–0.8) are observed in *Cwb* climate zone. Lower magnitude and higher variability in R^2 observed over *BWk* and *Cwb* climate zone can be attributed to either low reliability of reference dataset (owing to low rain-gauge density in these climatic zones) or possible complexity in capturing the satellite and reference rainfall relationships at very high

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

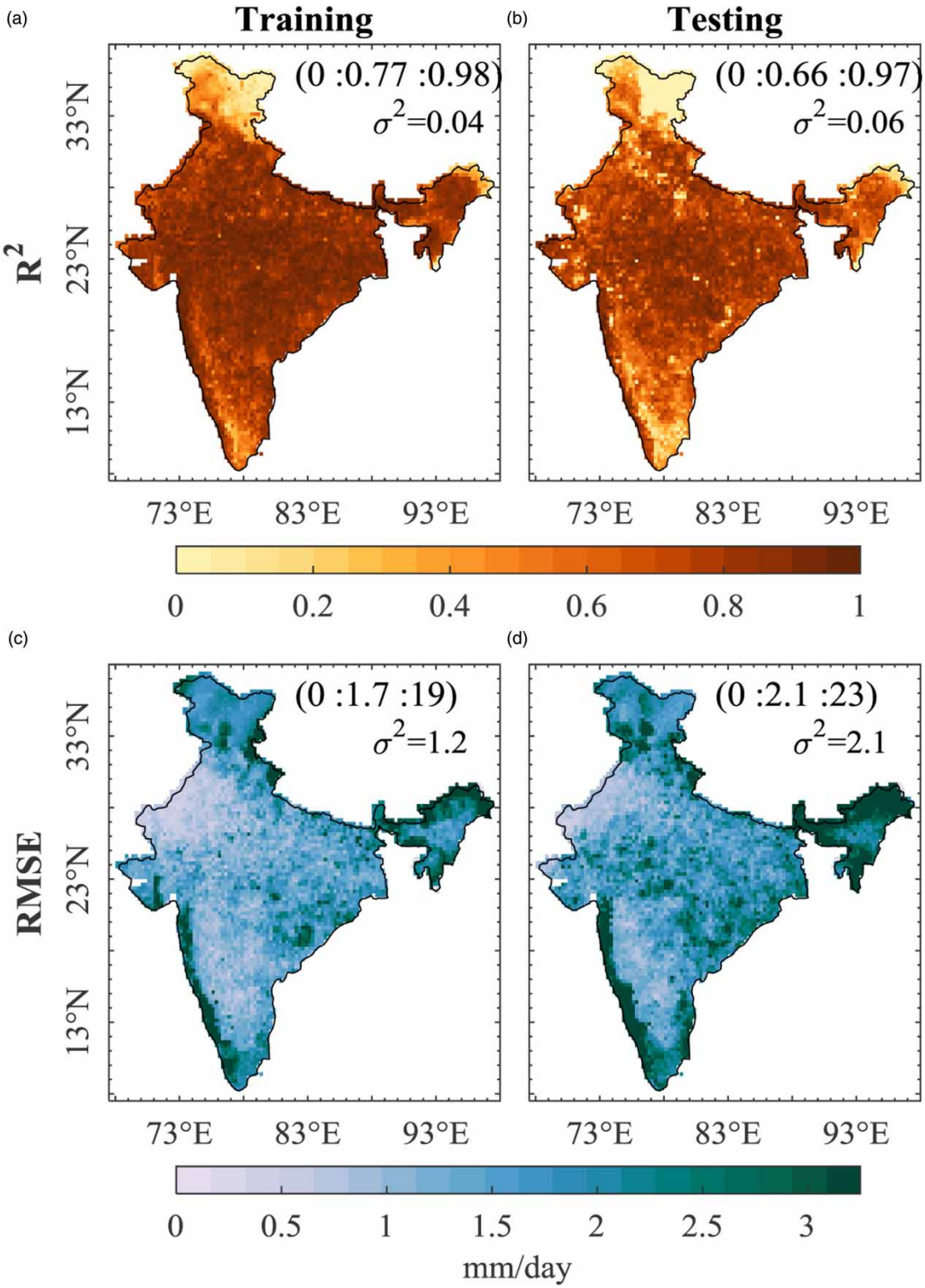

Figure 7 | Spatial variation of (a–b) Coefficient of Determination (R^2) and (c–d) Root Mean Square Error (RMSE) of the trained CART model during the training period (2001–2011) and testing period (2012–2016) over India. All-India minimum, mean and maximum values are given in parentheses (minimum: mean: maximum). Spatial variance (σ^2) is also provided. RMSE units are in mm/day.

altitudes. Over the remaining climate zone, the performance of CART as per R^2 is significantly good and less uncertain. High magnitude and variability of RMSE values are found in *Am* and *Cwb* climate zones.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

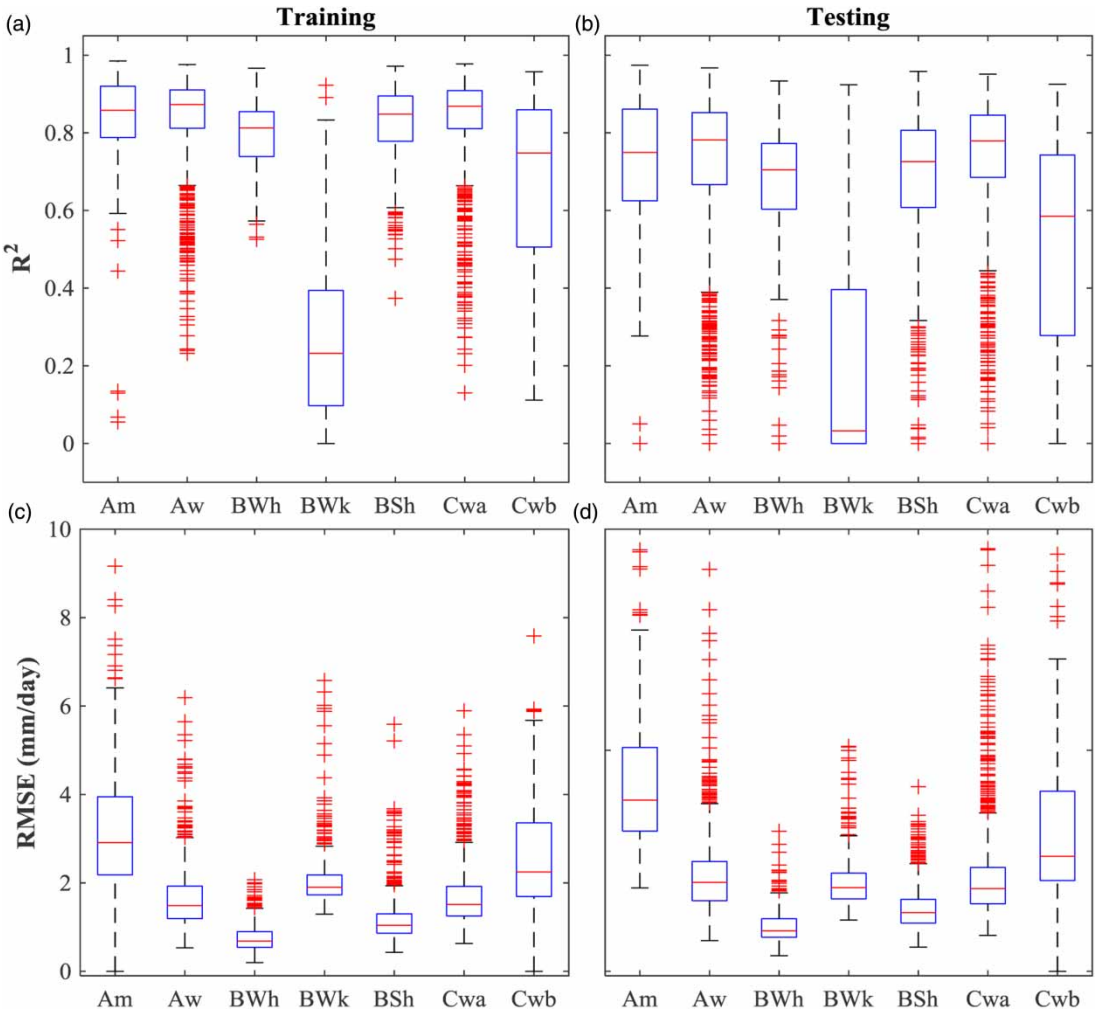

Figure 8 | Box plots showing the variation of (a–b) Coefficient of Determination (R^2) and (c–d) Root Mean Square Error (RMSE) of the trained CART model within different Koppen climate zones over India during the training period (2001–2011) and testing period (2012–2016).

Improvement in IMERG precipitation post application of CART

To quantify the improvement in the IMERG monthly satellite rainfall post-application of CART, we calculated the average monthly MAB and RMSE during the validation period. Table 3 shows the average monthly MAB and RMSE obtained before bias correction (between Original IMERG and IMD) and post bias-correction (CART corrected IMERG and IMD). The percentage of improvement in MAB and RMSE obtained was also quantified as shown in Table 3. High magnitudes of MAB (1.44–2.9) and RMSE (2.26–4.76) are observed during the monsoon season (June, July, August and September). All the months, except December, show a reduction in MAB post bias correction by CART. In December, CART was not able to reduce the absolute bias in the original IMERG dataset as an increase in MAB by 6.3% was observed. However, a significant reduction (17.1%) in RMSE was observed during December. Overall, CART was efficient in reducing the all-India average monthly MAB and RMSE in the IMERG dataset.

The performance of bias correction algorithm was further analysed over different climate and elevation zones over India. Figure 9 shows the variation of monthly MAB in original and CART corrected IMERG dataset spatially averaged over entire India (panel a) and climate zones (panel b–h) during the testing period (2012–2016). CART effectively reduced the all-India MAB over the majority of months. Although the total bias was not removed completely, a significant improvement in MAB ranging from 5.3 to 29.2% was observed. The ability of CART algorithm to reduce the MAB in IMERG dataset varied with varying

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

Table 3 | Average monthly MAB and RMSE obtained between (i) Original IMERG and IMD and (ii) CART corrected IMERG and IMD for all months during the validation period

	MAB			RMSE		
Months	Original (mm/day)	After CART Correction (mm/day)	Improvement (%)	Original (mm/day)	After CART Correction (mm/day)	Improvement (%)
January	0.46	0.43	5.3	1.04	0.95	8.7
February	0.69	0.53	22.2	1.45	1.12	22.8
March	0.78	0.62	19.6	1.58	1.22	22.5
April	0.98	0.69	29.2	2.02	1.27	37.3
May	1.05	0.83	20.3	1.91	1.45	24.0
June	2.06	1.53	25.6	3.82	2.90	24.2
July	2.90	2.20	24.1	4.76	3.63	23.8
August	2.55	1.93	24.3	4.02	3.11	22.7
September	1.98	1.44	26.9	3.17	2.26	28.8
October	1.00	0.85	14.4	1.70	1.33	22.0
November	0.63	0.54	13.8	1.48	1.13	23.8
December	0.35	0.37	**−6.3**	0.83	0.68	17.1

climate zones. Highest magnitude of MAB was observed in *Am* climate zone which receives heavy rainfall during the monsoon season. CART based bias correction was able to dampen the higher magnitudes of MAB in original data over all the climate zones except *BWk*. Over the *BWk* zone, bias correction with CART algorithm escalated the lower magnitudes of MAB, when compared to the original dataset. Moreover, *Am* zone observed minimum improvement in MAB. The RMSE in the original IMERG dataset was also significantly reduced post-application of CART as shown in Figure S2.

Figure 10 shows the variation of all-India and elevation zones averaged monthly MAB in original IMERG and CART bias-corrected IMERG during the testing period (2012–2016) over India. Seven elevation zones as shown in Table S1 were formulated for evaluating the performance of CART based bias-correction algorithm. Low MAB values were observed in very low elevation zone (<500 m) as well as very high elevation zones (>5,000 m). The performance of CART algorithm was relatively better when elevation was low and deteriorated with increasing elevation. An increase in the magnitude of MAB was observed post bias correction by CART, especially during the months with low MAB in the original IMERG data.

Comparison of CART-based bias correction scheme with other widely adopted schemes

Two widely adopted bias correction schemes – Linear scaling (LS) and Equidistant Cumulative Distribution Matching (EDCDF) are selected for comparison with the proposed CART-based bias correction technique. Detailed formulation of LS and EDCDF technique is given in Table S2. Figure 11 shows the spatial variation of average MAB and RMSE obtained from the original IMERG dataset, CART bias-corrected IMERG, LS corrected and EDCDF matching scheme during the testing period (2012–2016) over India. Over most of the grids over India, least MAB and RMSE were observed in CART-based bias correction followed by EDCDF scheme and LS scheme. Overall, least all-India average MAB of 0.94 mm/day was observed in CART, while the other two techniques, i.e., EDCDF matching and LS corrected method observed the all-India average MAB of 1.1 mm/day and 1.3 mm/day. Figure 12(a) shows the monthly variation of all-India average MAB obtained after CART based bias correction, LS bias correction and EDCDF matching method. CART based method and EDCDF matching method performed best in reducing the all-India average MAB, however, for initial months the performance of CART was much better among the two. Similar inferences were drawn when the bias correction schemes were compared based

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

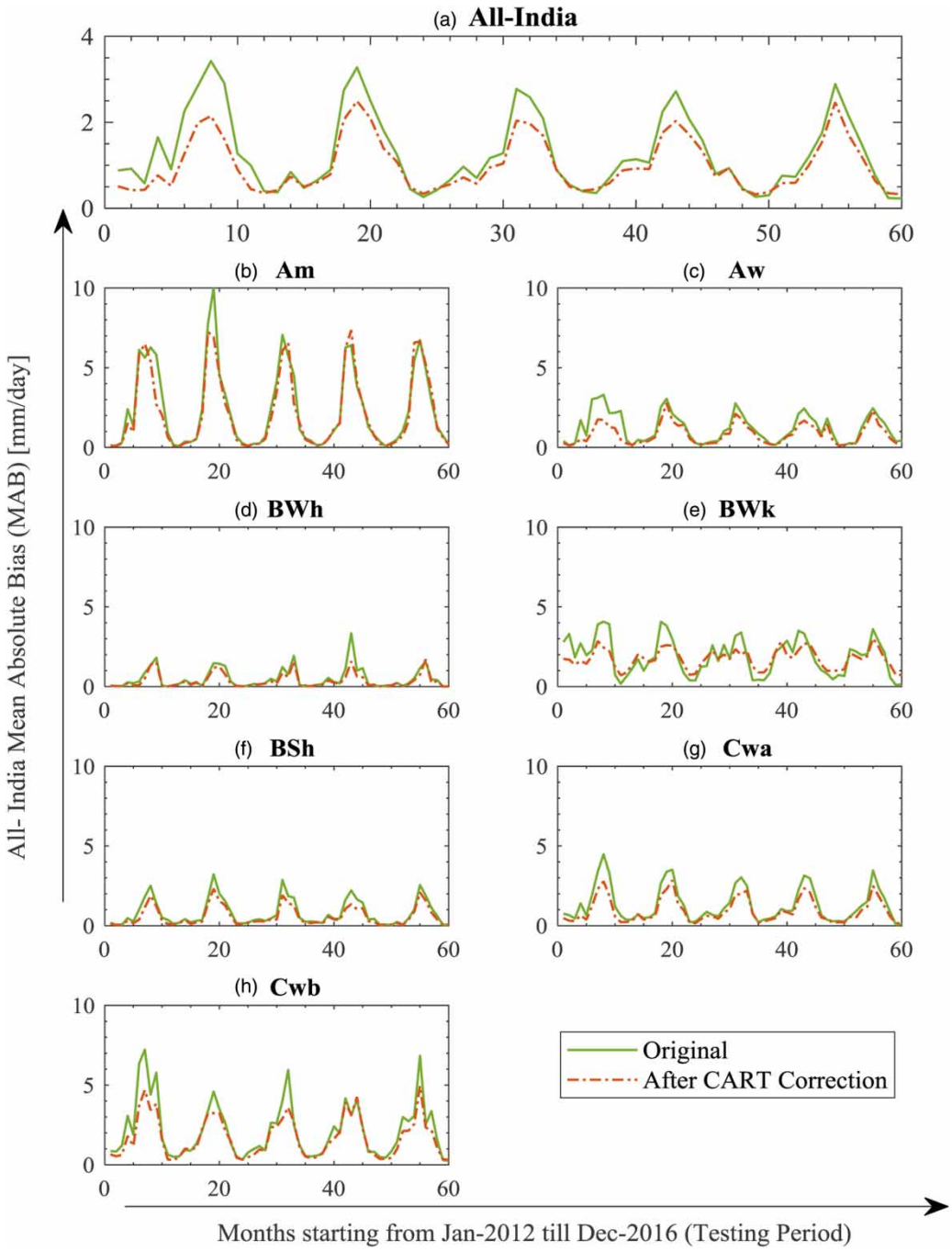

Figure 9 | Monthly variation of Mean Absolute Bias (MAB) in original IMERG and CART bias-corrected IMERG averaged over (a) entire India and (b–h) different climate zone during the testing period (2012–2016) over India.

on their average all-India RMSE values (Figure S3). Better performance of CART model over the other two methods can also be observed in Figure 12(b), wherein CART corrected IMERG dataset shows the highest correlation with IMD reference dataset during the testing period (2012–2016).

CONCLUSIONS

In the present study, we examine the applicability of Classification and Regression tree-based machine learning algorithm for possible bias reduction of the IMERG precipitation dataset over India. The major conclusions observed in the present study are:

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

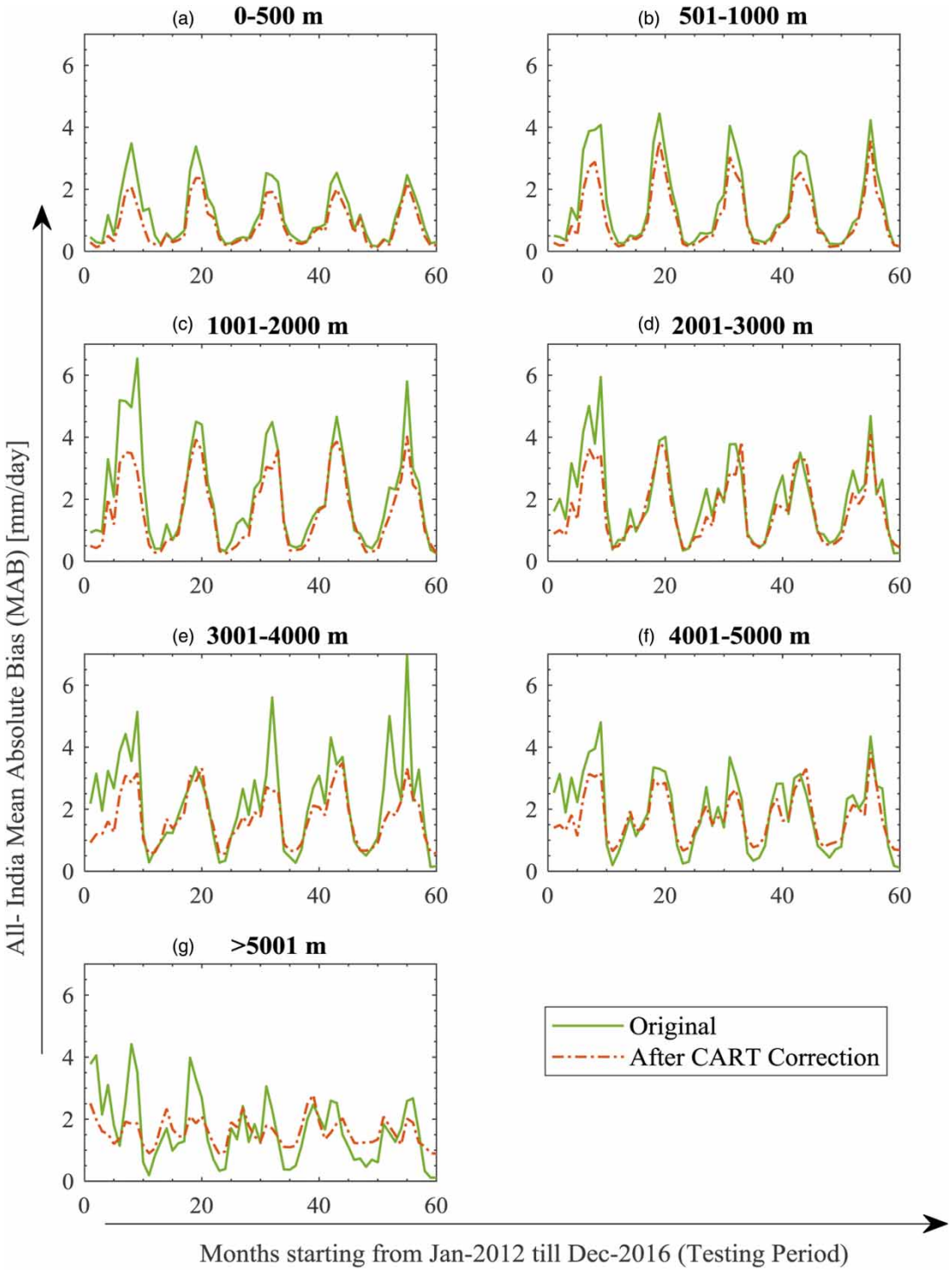

Figure 10 | Monthly variation of Mean Absolute Bias (MAB) in original IMERG and CART bias-corrected IMERG averaged over different (a–g) elevation zone during the testing period (2012–2016) over India.

- The CART model is highly effective in capturing the bias during the training (average $R^2 = 0.77$) and testing (average $R^2 = 0.66$) period. High R^2 (>0.75) and very low variability is observed in all climate zones except for *BWk* and *Cwb* zone.
- Significant improvement in average monthly MAB (−6.3 to 29.2%) and RMSE (8.7–37.3%) was obtained post-application of CART based bias-correction.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

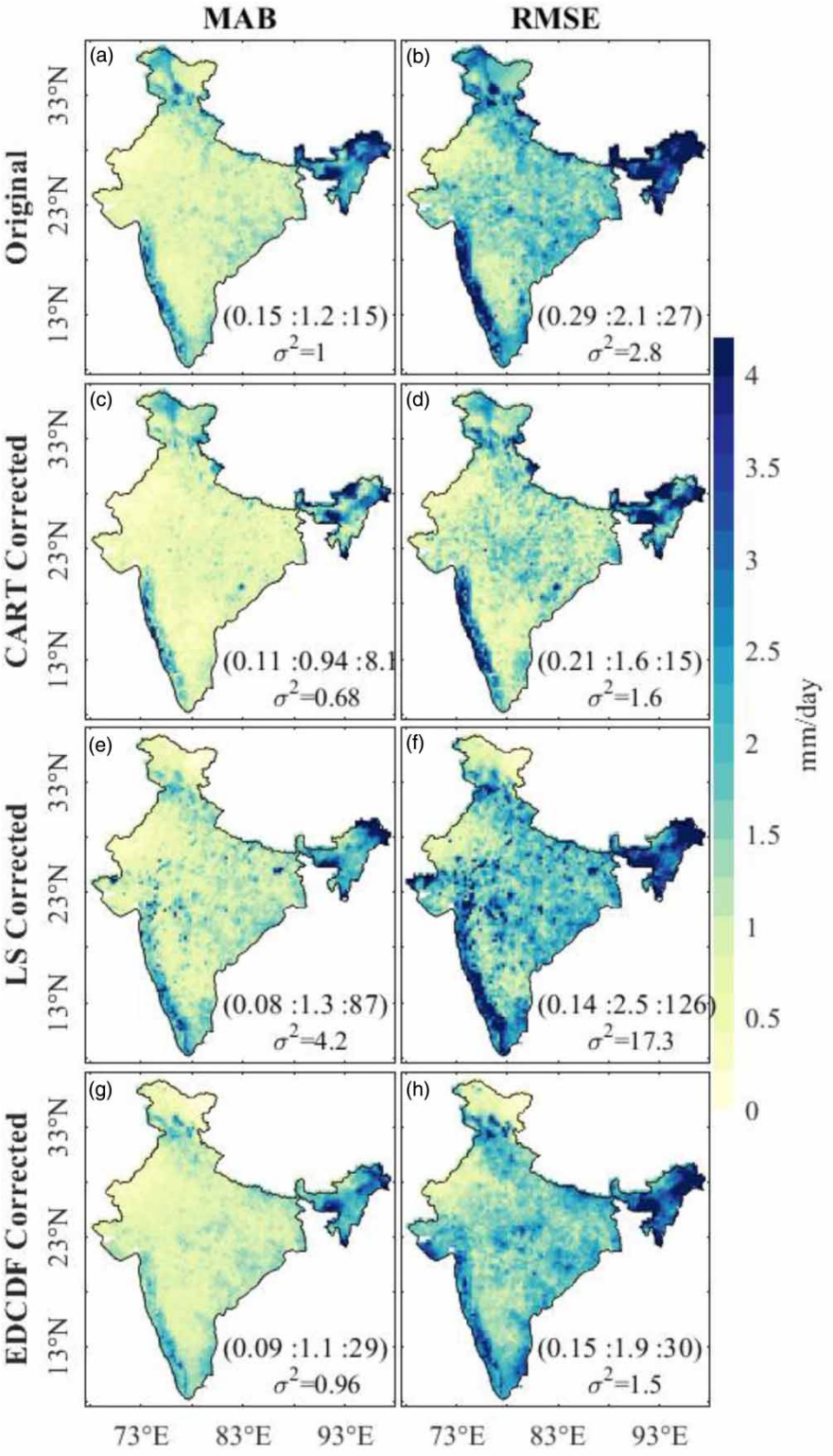

Figure 11 | Spatial variation of MAB (left column) and RMSE (right column) in the (a, b) original IMERG (c, d) CART bias-corrected IMERG (e, f) Linear Scaling (LS) corrected and (g, h) equidistant CDF (EDCDF) matching scheme averaged during the testing period (2012–2016) over India.

- CART based bias correction was able to dampen the higher magnitudes of MAB and RMSE in original data over all the climate zones except *BWk*, where the application of CART algorithm escalated the lower magnitudes of MAB and RMSE, when compared to the original dataset.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

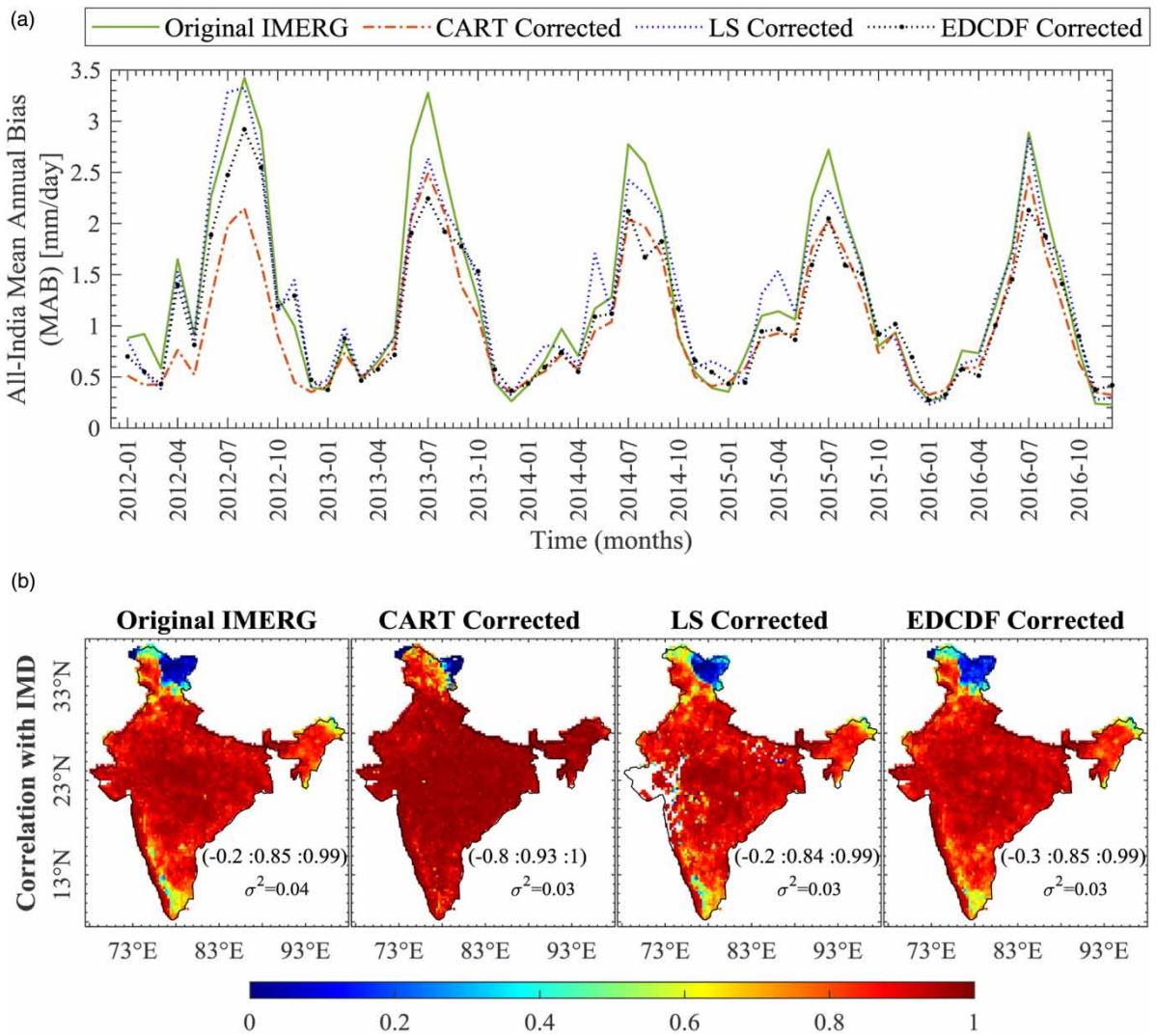

Figure 12 | (a) Monthly variation of Mean Absolute Bias (MAB) in the original IMERG, CART corrected, LS corrected and EDCDF corrected (b) Temporal correlation of original IMERG, CART corrected, LS corrected and EDCDF corrected datasets with IMD (reference) dataset during the testing period (2012–2016) over India.

- The performance of CART algorithm was relatively better when elevation was low and deteriorated with increasing elevation.
- Better performance of CART model over Linear scaling (LS) and Equidistant Cumulative Distribution Matching (EDCDF) was observed. CART corrected IMERG dataset depicted the highest correlation with IMD reference and least average MAB and RMSE dataset during the testing period.

It is important to mention here some of the limitations of the present study. Firstly, we assumed that the reference gauge-based gridded dataset from IMD to be error-free, which may not be true always. Although errors may exist in the gauge-based datasets, their magnitude would be much smaller than those in the satellite-based datasets. For further improvement in bias reduction, station data available from Doppler weather radar or automatic weather stations can be used as a reference in the future. Since over mountainous regions of India, IMD rainfall data is uncertain due to less gauge network density, in the future, a study can be conducted to compare the IMERG datasets directly with gauge station values for these regions. Secondly, we have trained the CART model on 11 years of monthly data, i.e., 132 values which are somewhat less for a data greedy algorithm like CART. Nonetheless, satisfactory fitting of the CART model is observed in the present study which aids in bias

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

correction of satellite data. The present study, nevertheless, opens the possibility of applying such a machine learning algorithm for bias correction of the daily or hourly precipitation dataset.

ACKNOWLEDGEMENTS

The authors would like to express sincere gratitude to the Indian Institute of Technology Delhi, India for supporting this work. The authors would also like to thank the (1) Indian Meteorological Department for providing the IMD daily gridded precipitation data and (2) NASA Goddard Earth Sciences Data and Information Services Center (https://disc.gsfc.nasa.gov/) for providing IMERG datasets, and making it available for research. The editor and the three anonymous reviewers are gratefully acknowledged for their valuable comments on our manuscript.

DATA AVAILABILITY STATEMENT

Datasets used in this study is available from India Meteorological Department, from the URL http://www.imdpune.gov.in/Clim_Pred_LRF_New/Grided_Data_Download.html and NASA Goddard Earth Sciences Data and Information Services Center, from the URL https://disc.gsfc.nasa.gov/.

REFERENCES

Abera, W., Brocca, L. & Rigon, R. 2016 Comparative evaluation of different satellite rainfall estimation products and bias correction in the Upper Blue Nile (UBN) basin. *Atmospheric Research* **178**, 471–483.

Abraham, S., Huynh, C. & Vu, H. 2020 Classification of soils into hydrologic groups using machine learning. *Data* **5** (1), 2.

AghaKouchak, A., Mehran, A., Norouzi, H. & Behrangi, A. 2012 Systematic and random error components in satellite precipitation data sets. *Geophysical Research Letters* **39** (9), L09406.

Asong, Z. E., Razavi, S., Wheater, H. S. & Wong, J. S. 2017 Evaluation of integrated multisatellite retrievals for GPM (IMERG) over southern Canada against ground precipitation observations: a preliminary assessment. *Journal of Hydrometeorology* **18** (4), 1033–1050.

Awange, J. L., Hu, K. X. & Khaki, M. 2019 The newly merged satellite remotely sensed, gauge and reanalysis-based multi-source weighted-ensemble precipitation: evaluation over Australia and Africa (1981–2016). *Science of the Total Environment* **670**, 448–465.

Beck, H. E., Van Dijk, A. I., Levizzani, V., Schellekens, J., Gonzalez Miralles, D., Martens, B. & De Roo, A. 2017 MSWEP: 3-hourly 0.25 global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences* **21** (1), 589–615.

Behrangi, A., Andreadis, K., Fisher, J. B., Turk, F. J., Granger, S., Painter, T. & Das, N. 2014 Satellite-based precipitation estimation and its application for streamflow prediction over mountainous western US basins. *Journal of Applied Meteorology and Climatology* **53** (12), 2823–2842.

Bhatti, H. A., Rientjes, T., Haile, A. T., Habib, E. & Verhoef, W. 2016 Evaluation of bias correction method for satellite-based rainfall data. *Sensors* **16** (6), 884.

Bhuiyan, E. M. A., Nikolopoulos, E. I. & Anagnostou, E. N. 2019 Machine learning–based blending of satellite and reanalysis precipitation datasets: a multiregional tropical complex terrain evaluation. *Journal of Hydrometeorology* **20** (11), 2147–2161.

Chaudhary, S. & Dhanya, C. T. 2019 Investigating the performance of bias correction algorithms on satellite-based precipitation estimates. In remote sensing for agriculture, ecosystems, and hydrology XXI. *International Society for Optics and Photonics* **11149**, 111490Z.

Chaudhary, S., Dhanya, C. T. & Vinnarasi, R. 2017 Dry and wet spell variability during monsoon in gauge-based gridded daily precipitation datasets over India. *Journal of Hydrology* **546**, 204–218.

Erdal, H. I. & Karakurt, O. 2013 Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. *Journal of Hydrology* **477**, 119–128.

Gebregiorgis, A. S. & Hossain, F. 2012 Understanding the dependence of satellite rainfall uncertainty on topography and climate for hydrologic model simulation. *IEEE Transactions on Geoscience and Remote Sensing* **51** (1), 704–718.

Gebregiorgis, A. S. & Hossain, F. 2015 How well can we estimate error variance of satellite precipitation data around the world? *Atmospheric Research* **154**, 39–59.

Gebregiorgis, A. S., Kirstetter, P. E., Hong, Y. E., Carr, N. J., Gourley, J. J., Petersen, W. & Zheng, Y. 2017 Understanding overland multisensor satellite precipitation error in TMPA-RT products. *Journal of Hydrometeorology* **18** (2), 285–306.

Habib, E., Haile, A. T., Sazib, N., Zhang, Y. & Rientjes, T. 2014 Effect of bias correction of satellite-rainfall estimates on runoff simulations at the source of the Upper Blue Nile. *Remote Sensing* **6** (7), 6688–6708.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.124

Hashemi, H., Nordin, M., Lakshmi, V., Huffman, G. J. & Knight, R. 2017 Bias correction of long-term satellite monthly precipitation product (TRMM 3B43) over the conterminous United States. *Journal of Hydrometeorology* **18** (9), 2491–2509.

Hong, Y., Adler, R. & Huffman, G. 2006 Evaluation of the potential of NASA multi-satellite precipitation analysis in global landslide hazard assessment. *Geophysical Research Letters* **33** (22), L22402.

Hossain, F. & Huffman, G. J. 2008 Investigating error metrics for satellite rainfall data at hydrologically relevant scales. *Journal of Hydrometeorology* **9** (3), 563–575.

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., Oki, R., Nakamura, K. & Iguchi, T. 2014 The global precipitation measurement mission. *Bulletin of the American Meteorological Society* **95** (5), 701–722.

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R. & Xie, P. 2014 *NASA Global Precipitation Measurement (GPM) Integrated Multi-SatellitE Retrievals for GPM (IMERG)*. Algorithm Theoretical Basis Document (ATBD), Version 4.4, NASA, 30.

Koriche, S. A. & Rientjes, T. H. 2016 Application of satellite products and hydrological modelling for flood early warning. *Physics and Chemistry of the Earth, Parts A/B/C* **93**, 12–23.

Kühnlein, M., Appelhans, T., Thies, B. & Nauss, T. 2014 Improving the accuracy of rainfall rates from optical satellite sensors with machine learning – a random forests-based approach applied to MSG SEVIRI. *Remote Sensing of Environment* **141**, 129–143.

Mannan, A., Chaudhary, S., Dhanya, C. T. & Swamy, A. K. 2018 Regionalization of rainfall characteristics in India incorporating climatic variables and using self-organizing maps. *ISH Journal of Hydraulic Engineering* **24** (2), 147–156.

MATLAB 2019 *Statistics and Machine Learning Toolbox™ User's Guide*. MathWorks. Available from: https://in.mathworks.com/help/stats/.

Murphy, K. P. 2012 *Machine Learning: A Probabilistic Perspective*. MIT press, Cambridge, MA.

Pai, D. S., Sridhar, L., Rajeevan, M., Sreejith, O. P., Satbhai, N. S. & Mukhopadhyay, B. 2014 Development of a new high spatial resolution (0.25 × 0.25) long period (1901–2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region. *Mausam* **65** (1), 1–18.

Peel, M. C., Finlayson, B. L. & McMahon, T. A. 2007 Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences Discussions* **4** (2), 439–473.

Pekel, E. 2020 Estimation of soil moisture using decision tree regression. *Theoretical and Applied Climatology* **139** (3), 1111–1119.

Petty, T. R. & Dhingra, P. 2018 Streamflow hydrology estimate using machine learning (SHEM). *JAWRA Journal of the American Water Resources Association* **54** (1), 55–68.

Prakash, S. 2019 Performance assessment of CHIRPS, MSWEP, SM2RAIN-CCI, and TMPA precipitation products across India. *Journal of Hydrology* **571**, 50–59.

Prakash, S., Mitra, A. K., AghaKouchak, A. & Pai, D. S. 2015 Error characterization of TRMM Multisatellite Precipitation Analysis (TMPA-3b42) products over India for different seasons. *Journal of Hydrology* **529**, 1302–1312.

Prakash, S., Mitra, A. K., AghaKouchak, A., Liu, Z., Norouzi, H. & Pai, D. S. 2018 A preliminary assessment of GPM-based multi-satellite precipitation estimates over a monsoon dominated region. *Journal of Hydrology* **556**, 865–876.

Prakash, S., Seshadri, A., Srinivasan, J. & Pai, D. S. 2019 A new parameter to assess impact of rain gauge density on uncertainty in the estimate of monthly rainfall over India. *Journal of Hydrometeorology* **20** (5), 821–832.

Rasouli, K., Hsieh, W. W. & Cannon, A. J. 2012 Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology* **414**, 284–293.

Sawunyama, T. & Hughes, D. A. 2008 Application of satellite-derived rainfall estimates to extend water resource simulation modelling in South Africa. *Water Sa* **34** (1), 1–10.

Tan, C. O. & Beklioglu, M. 2006 Modeling complex nonlinear responses of shallow lakes to fish and hydrology using artificial neural networks. *Ecological Modelling* **196** (1–2), 183–194.

Tan, J., Petersen, W. A. & Tokay, A. 2016 A novel approach to identify sources of errors in IMERG for GPM ground validation. *Journal of Hydrometeorology* **17** (9), 2477–2491.

Tan, J., Petersen, W. A., Kirstetter, P. E. & Tian, Y. 2017 Performance of IMERG as a function of spatiotemporal scale. *Journal of Hydrometeorology* **18** (2), 307–319.

Tang, L. & Hossain, F. 2012 Investigating the similarity of satellite rainfall error metrics as a function of Köppen climate classification. *Atmospheric Research* **104**, 182–192.

Tang, G., Ma, Y., Long, D., Zhong, L. & Hong, Y. 2016 Evaluation of GPM Day-1 IMERG and TMPA version-7 legacy products over Mainland China at multiple spatiotemporal scales. *Journal of Hydrology* **533**, 152–167.

Tao, Y., Gao, X., Hsu, K., Sorooshian, S. & Ihler, A. 2016 A deep neural network modeling framework to reduce bias in satellite precipitation products. *Journal of Hydrometeorology* **17** (3), 931–945.

Veettil, A. V. & Mishra, A. K. 2020 Multiscale hydrological drought analysis: role of climate, catchment and morphological variables and associated thresholds. *Journal of Hydrology* **582**, 124533.

Waheed, T., Bonnell, R. B., Prasher, S. O. & Paulet, E. 2006 Measuring performance in precision agriculture: CART – a decision tree approach. *Agricultural Water Management* **84** (1–2), 173–185.

Willmott, C. J. 1981 On the validation of models. *Physical Geography* **2** (2), 184–194.

Worqlul, A. W., Ayana, E. K., Maathuis, B. H., MacAlister, C., Philpot, W. D., Leyton, J. M. O. & Steenhuis, T. S. 2018 Performance of bias corrected MPEG rainfall estimate for rainfall-runoff simulation in the upper Blue Nile Basin, Ethiopia. *Journal of Hydrology* **556**, 1182–1191.

Yu, J., Qin, X., Larsen, O. & Chua, L. H. C. 2014 Comparison between response surface models and artificial neural networks in hydrologic forecasting. *Journal of Hydrologic Engineering* **19** (3), 473–481.

First received 1 April 2020; accepted in revised form 11 July 2020. Available online 4 August 2020