Islam, T., Srivastava, P., Dai, Q., Gupta, M., & Zhuo, L. (2015). Rain Rate Retrieval Algorithm for Conical-Scanning Microwave Imagers Aided by Random Forest, RReliefF, and Multivariate Adaptive Regression Splines (RAMARS). IEEE Sensors Journal, 15(4), 2186-2193. 10.1109/JSEN.2014.2372814

Rain Rate Retrieval Algorithm for Conical-Scanning Microwave Imagers Aided By Random Forest, RReliefF and Multivariate Adaptive Regression Splines (RAMARS)

Tanvir Islam[*,1,2,3], Prashant K. Srivastava[3,4,5], Miguel A. Rico-Ramirez[3], Qiang Dai[3], Dawei Han[3], Manika Gupta[4,6], Lu Zhuo[3]

ABSTRACT

This study proposes a rain rate retrieval algorithm for conical-scanning microwave imagers (RAMARS), as an alternative of the NASA Goddard Profiling (GPROF) algorithm, that does not rely on any *a priory* information. The fundamental basis of the RAMARS follows the concept of the GPROF algorithm, which means, being consistent with the TRMM PR rain rate observations, but independent of any auxiliary information. The RAMARS is built upon the combination of *state of the art* machine learning and regression techniques, comprising of Random Forest algorithm, RReliefF, and Multivariate Adaptive Regression Splines. The RAMARS is applicable to both over ocean and land as well as coast surface terrains. It has been demonstrated that, when comparing with the TRMM PR observations, the performance of the RAMARS algorithm is comparable to the 2A12 GPROF algorithm. Furthermore, the RAMARS has been applied to two cyclonic cases, hurricane Sandy in 2012 and cyclone Mahasen in 2013, showing very good capability to reproduce the structure and intensity of the cyclone fields. The RAMARS is highly flexible, thanks to its four processing components, making it extremely suitable for use to other passive microwave imagers in the global precipitation measurement (GPM) constellation.

Keywords: brightness temperature (TB); passive microwave (PMW); precipitation estimation; precipitation radar; global precipitation measurement (GPM); constellation; radiometer; hurricane;

1. INTRODUCTION

There have been on-going research efforts to improve the satellite precipitation estimate using passive microwave (PMW) radiometers for a few decades. The algorithms and validation results related to passive microwave estimate of precipitation can be found in [1], [2], [3], [4], and others. Among many proposed algorithms, the Goddard Profiling Algorithm (GPROF) is well-known and being used as an operational algorithm for many PMW sensors, including the TMI [5-7]. Currently, the GPROF algorithm is based on a "look-up" table that is constructed rom observed TRMM radar and radiometer measurements with some ancillary atmospheric information for the rain adjustment [7, 8]. A Bayesian methodology is used to invert the measured brightness temperatures (TBs) to rain rate information based on this look-up table that matches the observed rain and TB profiles to those stored in the look-up database.

Although, the use of ancillary information makes the GPROF algorithm robust and more physical, the drawback is that, the real-time application of GPROF satellite precipitation estimation becomes very limited. In one end, feeding the GPROF derived precipitation information to a numerical weather prediction (NWP) model for forecasting purpose gets trickier, since the ancillary NWP information is already a part of the GPROF retrieval. In other ends, it takes quite a while to obtain the ancillary information from an NWP model before applying the GPROF to the radiance measurements. This makes it unsuitable for real time application.

In this study, we propose a rain rate retrieval algorithm for conical PMW imagers that use three data mining techniques- the random forest algorithm, the RReliefF, and the Multivariate Adaptive Regression Spline (hereinafter the algorithm is named as RAMARS). The RAMARS shares a common thought with the GPROF algorithm, that is, the microwave imagers based retrieval being consistent with the PR retrieval. This is logical, as the PR has better capability of providing rainfall measurement, and can be considered as "reference" for the imager estimate of rainfall. Furthermore, the RAMARS is specifically designed to provide rainfall estimate at high resolution, which is the PR's resolution. Nonetheless, the RAMARS is independent of using any NWP or ancillary information, as opposed to the case for GPROF that uses ancillary information. The RAMARS is applicable to all surface terrains (ocean, land, and coast).

2. DATA DESCRIPTION

2.1. TMI calibrated brightness temperatures

The TMI is a conical-scanning passive microwave imager operating at nine channels, at five frequencies, with a constant

[1] NOAA/NESDIS Center for Satellite Applications and Research, College Park, MD, USA
[2] Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA
[3] Department of Civil Engineering, University of Bristol, Bristol, UK
[4] NASA Goddard Space Flight Center, Greenbelt, MD, USA
[5] Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA
[6] Universities Space Research Association, Columbia, MD, USA

incident angle of 52.8 degrees. Except for the water vapour absorption band channel at 21 GHz, each frequency has one vertically (V) and one horizontally (H) polarized channels. Nevertheless, the footprint size varies depending on the frequencies. The higher-frequency channels have smaller footprint sizes as compared to the lower frequency channels. For example, the instantaneous field of view (IFOV) for the 85.5 GHz channel is 7 x 5 km, whereas for 10.65 GHz, the IFOV is 63 x 37 km. The TMI's swath width is limited to 760 km due to the low orbital altitude of the TRMM satellite.

In the current study, the TMI calibrated brightness temperatures are taken from TRMM 1B11 data product. In 1B11, the radiometer counts are converted to antenna temperatures by applying a linear relationship. Further, the antenna temperatures are corrected for cross-polarization and spill over to produce brightness temperatures.

2.2. PR near surface rain rate

The PR is a cross-track scanning radar that scans ±17º off nadir at intervals of 0.35º. Such geometry projects an almost-regular grid on the earth's surface with a horizontal footprint of about 5 km and a vertical resolution of 250 m at nadir. The PR operates at a frequency of 13.8 GHZ (2.17 cm wavelength).

The PR surface rain rate is obtained by converting the reflectivity factor measured by the PR to rain rate, taking into account certain drop size distribution assumptions. Prior to the conversion, the measured reflectivity factor is corrected for attenuation following a hybrid method based on the Hitschfeld-Bordan method and the surface reference technique [9]. Some other factors related to surface echoes, non-uniform beam filling (NUBF), and the identification of the phase state (i.e., water, mixed or ice) are also considered. The derived rain rate is stored in the TRMM 2A25 product. In the present study, this 2A25 (V7) data product is used to obtain the PR near-surface rain rate, while the near surface rain rate is defined as the rain rate at the lowest range bin in the clutter free ranges.

2.3. GPROF near surface rain rate

In order to compare the RAMARS rain rate retrieval with the GPROF algorithm, the GPROF produced near surface rain rate is used, which is from the TRMM 2A12 V7 product. As mentioned earlier, the GPROF uses a Bayesian inversion methodology to produce instantaneous rain rate by matching the observed brightness temperatures to PR measurements. More detailed information about the GPROF algorithm can be found in Kummerow *et al.* [8].

3. ALGORITHM BASIS

3.1. Random forest

The random forest is an ensemble learning algorithm that combines the ideas of "bootstrap aggregating" [10] and "random subspace method" [11] to construct randomized decision trees with controlled variation, introduced by Breiman [12]. According to the theory of random forest algorithm, for a collection of classifiers $h_1(\mathbf{x})$, $h_2(\mathbf{x})$, . . . , $h_K(\mathbf{x})$, and with the training set at random from sampled random vector Y, $\mathbf{X}$, the margin function is termed as:

$$mg(\mathbf{X},Y)=av_k I(h_k(\mathbf{X})=Y)-\max_{j\neq Y} av_k I(h_k(\mathbf{X})=j) \qquad (1)$$

where, $I(.)$ represents the indicator function. This margin function measures the extent to which the fraction of correct classifications exceeds the fraction of the most voted incorrect classifications. The generalization error is given as:

$$PE^{*}=P_{X,Y}(mg(\mathbf{X},Y)<0) \qquad (2)$$

where, the probability is over the space $\mathbf{X}$, Y. This depends upon the strength of the individual weak learners in the forest and the correlation between them. By definition, in random forests,

$$h_k(\mathbf{X})=h(\mathbf{X},\Theta_k) \qquad (3)$$

Therefore, the margin function for a random forest would be:

$$mr(\mathbf{X},Y)=P_{\Theta}(h(\mathbf{X},\Theta)=Y)-\max_{j\neq Y} P_{\Theta}(h(\mathbf{X},\Theta)=j) \qquad (4)$$

And the expected strength of the classifiers in a random forest is:

$$s=E_{X,Y}mr(\mathbf{X},Y) \qquad (5)$$

The fundamental idea of the random forest is that at each tree split, a random sample of m features is drawn, and only those m features are considered for splitting, where $m = \sqrt{N}$, N being the total number of features. For each tree grown on a bootstrap sample, the "out-of-bag" strength is monitored. The forest is then re-defined based on this "out-of-bag" strength by de-correlating the irrelevant trees.

3.2. RReliefF

The RReliefF, also known as a regression version of ReliefF, is a feature selection algorithm that provides information about quality of attributes [13]. Theoretically, let $W[A]$ is the quality of attribute A, which is an approximation of the following Bayes rule:

$$W[A]=\frac{P_{diffC|diffA}P_{diffA}}{P_{diffC}}-\frac{(1-P_{diffC|diffA})P_{diffA}}{1-P_{diffC}} \qquad (6)$$

where, P_{diffA} and P_{diffC}, and $P_{diffC|diffA}$ are defined as so that $W[A]$ can directly be evaluated using the probability of the predicted values of two instances being different:

P_{diffA} = P(different value of A | nearest instances) (7)

P_{diffC} = P(different prediction | nearest instances) (8)

$P_{diffC|diffA}$ = P(different prediction | (different value of A and nearest instances) (9)

The key idea of the RReliefF is to estimate the quality of attributes according to how well their values distinguish between instances that are near to each other.

3.3. MARS

The multivariate adaptive regression splines (MARS) is an adaptive approach for multivariate nonparametric regression, introduced by Friedman [14]. The fundamental basis of the MARS approach is that it does not make any assumption about the underlying functional relationship between the response and predictor variables. As an alternative, it constructs the relationship through the use of basis functions coming from the datasets, in turn, partitions the input space into regions, having regression equation for each region. It is able to automatically model the non-linearities as well as can interact between the predictor variables.

For the sake of explanation, let y be the single response variable (reference rain rate in our case) which depends on n predictor variables $x = (x_1, x_2, \ldots, x_n)$ comprising of an M number of samples $x_m = (x_{1m}, x_{2m}, \ldots \ldots, x_{nm})$. Therefore,

$$y_m = f(x_m) + \varepsilon_m \tag{10}$$

where, $f(.)$ is assumed smooth in $E^{(n)}$ and $[\varepsilon_m]$ are mean zero random variables. The primary objective is to identify a rational approximation of $f(.)$ over the predictor domain.

Friedman [14] proposed the MARS algorithm, a new way to approximate the multivariate function taking the subbasis from a n-variate complete spline basis tensor product in the form of two-sided truncated power basis functions:

$$\left(\pm(x-t)\right)_+^q \tag{11}$$

where, knot t is the knot site selected from the observed values of corresponding components and q represents the order of the spline approximation. The jth basis function is expressed as:

$$T_j(x) = \prod_{k=1}^{K_j} \left[s_{k,j} \left(x_{v(k,j)} - t_{k,j} \right) \right]_+^q \tag{12}$$

where, K_j represents the interaction level in the basis function T_j, s_{kj} accepts two values (-1, +1), $v(k,j)$ labels the predictor variable associated with the corresponding level of T_j, and t_{kj} is a knot location for $x_{v(k,j)}$. In order to produce a set of basis functions, two-stage procedure, the forward stepwise addition and backward stepwise deletion are adopted. In forward stage, the procedure starts with only the constant function:

$$T_0(x) = 1 \tag{13}$$

Following Jth iteration, there are $2J$+1 basis functions:

$$\left[T_j(x)\right]_0^{2,J} \tag{14}$$

Subsequently, the J+1 iteration adds two new basis functions:

$$\begin{aligned} T_{2J+1(x,l,v,t)} &= T_t(x)\left[+(x_v - t)\right]_+^q \\ T_{2J+2(x,l,v,t)} &= T_t(x)\left[-(x_v - t)\right]_+^q \end{aligned} \tag{15}$$

In this way, a large model is constructed with J_{max} tensor product basis functions, that typically overfits the data. Therefore, a backward deletion algorithm is applied in order to achieve optimal functions by the help of generalized cross validation criterion (GCV):

$$GCV(J) = \frac{1}{M} \sum_{m=1}^{M} \left[y_m - \hat{f}_J(x_m) \right]^2 \Bigg/ \left[1 - \frac{C(J)}{M} \right]^2 \tag{16}$$

The two-stage procedure produces a model in the form of:

$$\hat{f}(x) = a_0 + \sum_{j=1}^{J} a_j \prod_{k=1}^{K_j} \left[s_{kj} \left(x_{v(k,j)} - t_{kj} \right) \right]_+^q \tag{17}$$

where, the coefficients a_j are computed by minimizing the residual sum-of-squares by standard linear regression.

4. ALGORITHM DEVELOPMENT

A flowchart illustrating the components of the RAMARS algorithm is shown in Figure 1. The RAMARS is comprised of four components: a pre-processing component for data preparation, the random forest component for rain/no rain screening, RRreliefF component for selecting the important features, and finally MARS component for retrieving the rain rates in a quantitative manner. However, note that the RRreliefF component is used offline only once, to identify the best possible features to feed into the MARS model. This RRreliefF component is somewhat useful, especially to reduce the computing powers and adapting the RAMARS for new sensors, principally for future use. The basic idea of the RAMARS is to retrieve rain rate in a robust manner, taking only the important attributes sensitive to hydrometeors depending on surface type.

4.1. Pre-processing component

Since the low frequency channels have the larger instantaneous field of view in comparison with the high frequency channels, it is important to bring all the channel information to a single domain. This is done in the pre-processing component. That means, the pre-processing component is primarily responsible for gridding the TBs from all available channels to a particular designated resolution. Currently, the TMI TBs are interpolated to the high resolution PR grid, which is around 5 km, by employing a triangle based linear interpolation algorithm:

$$TBs_{PR} = f(Lat_{TMI}, Lon_{TMI}, TBs_{TMI}, Lat_{PR}, Lon_{PR}) \tag{18}$$

where, TBs_{TMI} is the brightness temperatures at TMI footprint for a particular channel, Lat_{TMI} and Lon_{TMI} are the latitudes and longitudes for the corresponding channel's measurements, Lat_{PR} and Lon_{PR} are the latitudes and longitudes for the PR's measurements, and TBs_{PR} is the brightness temperature interpolated at PR footprint. The algorithm fits a surface of the form $TBs_{TMI} = f(Lat_{TMI}, Lon_{TMI})$ to the data (Lat_{TMI}, Lon_{TMI}, TBs_{TMI}) and interpolates the surface at the points specified by (Lat_{PR}, Lon_{PR}) to produce TBs_{PR}. The reason for gridding the information in latitude-longitude space rather than pixel scan position is in accounting the indirect variations in the relative pixel position connected to the satellite altitudes.

After gridding, the TBs are then used to compute the necessary indices. In this study, the following indices are computed from the TBs, and included in the input features [15]:

$$PCT_{37} = 2.20TB_{37V} - 1.20TB_{37H} \quad (19)$$

$$PCT_{85} = 1.82TB_{85V} - 0.82TB_{85H} \quad (20)$$

where, the PCT37 and PCT85 are defined as polarization corrected temperature at 37 and 85 GHz, respectively,
and,

$$SI_{85} = TB_{e(85V)} - TB_{85\mathrm{V}} \quad (21)$$

$$SI_{37} = TB_{e(37V)} - TB_{37V} \quad (22)$$

where, SI_{85} indicates the scattering index at 85 GHz, SI_{37} indicates the scattering index at 37 GHz, TB_{85V} is the observed TB at 85 GHz, TB_{37V} is the observed TB at 37 GHz, $TB_{e(85V)}$ is the estimated TB_{85V} in scattering free case and $TB_{e(37V)}$ is the estimated TB_{37V} in scattering free case. The $TB_{e(85V)}$ and $TB_{e(37V)}$ are calculated as follows [16]:

Land $$TB_{e(85V)} = 451.9 - 0.44 \times TB_{19V} - 1.775 \times TB_{22V} + 0.00575 \times TB_{22V}^2 \quad (23)$$

Ocean $$TB_{e(85V)} = -174.4 + 0.72 \times TB_{19V} + 2.439 \times TB_{22V} - 0.00504 \times TB_{22V}^2 \quad (24)$$

$$TB_{e(37V)} = 62.18 + 0.773 \times TB_{19V} \quad (25)$$

where, TB_{19V} and TB_{22V} are the vertically polarized TBs for 19 GHz and 22 GHz channels respectively.

The main advantage of the above indices is their ability to decrease the background surface emissivity effects in complex surface conditions. In the radiative transfer process, radiation energy is scattered out by ice content and large raindrops. Therefore, such scattering indices could provide an indirect estimate of rainfall over complicated surface conditions. Furthermore, some of the earlier studies have reported that the different TB combinations may provide better insight of precipitation characteristics than the single channel TB information. More specifically, the polarization difference can provide scattering and the emission phenomenon along with the information of water vapour and temperature contents in a profile. You *et al.* [17] stated that the combination of TBs from 19 and 37 GHz (V19-V37) or from 21 and 37 GHz (V21-V37) could explain 10% more variance of near-surface rain rate than can the 85 GHz channel over land. As such, 72 features from the combination of TBs (only “addition” and “subtraction” operators) along with the PCTs and SIs are considered, making it a total of 85 features for the inclusion as input features in the pre-processing component (Table 1). One should note that the PCT and SI features used in this work are actually developed for the SSM/I, which had a much coarser spatial resolution than TMI. Therefore, the calculated PCT and SI features could be different than the ones, if developed for the TMI. However, in this study, we are assuming such differences are expected to be very marginal. The development of new PCT and SI features exclusively for TMI could be a subject of future work.

Another step that is done in the pre-processing component is, assigning ocean/land/coast mask in each grid. A topography database is loaded in order to accompany the surface masks, which is actually the same as the PR’s ocean/land/coast flag database.

4.2. Random forest component

The random forest component is particularly used for the screening of rain – no rain information based on the classifier developed with Breiman’s random forest algorithm. A detailed description of the approach and the validation results are well stated in our previous article [18]. However, for the sake of completeness, a brief outline of the approach is reminded here. The approach is particularly based on randomized decision trees with bootstrap aggregating associated between the TMI input features such as calibrated brightness temperatures and the TRMM PR rain/no rain information. The method is quite robust, easy to implement in the RAMARS system, and it has been shown in the previous article that it outperforms the GPROF algorithm based on various dichotomous skill scores. Overall, the accuracy reported with the random forest algorithm was around 97-98%.

4.3. RReliefF component

The primary idea of the RRreliefF component is to identify the best possible features sensitive to precipitation information depending upon the surface types. In other words, the RRreliefF is a feature selection technique that distinguishes the quality of attributes in a problem with strong dependencies between the attributes. The feature selection is a frequent term often used in artificial intelligence. The foremost benefit of the feature selection is that it reduces the number of features, allowing the inclusion of only the important features in the MARS model. In this way, model complexity of the MARS model is reduced, but without compromising the retrieval accuracy of the model. It is to be noted that, in the RAMARS, the RReliefF is run only once in offline and not used in “run time” within the RAMARS.

Based on an offline investigation with a considerable number of orbital samples from the year of 2012-2013, the RReliefF weights are plotted in Figure 2. Top 5 indices from the ranking over three different surface terrains are tabulated in Table 2. The expectation is, the emission signatures will be somehow more correlated to the rain rate over the ocean, while over land, the scattering signatures at high frequency channels will be of great importance. This has been reflected in the RReleifF ranking, which suggests, the top-ranked indices over the ocean are more associated with emission signatures than the scattering signatures, and vice versa for land surface terrain. Although, there are exceptions, for instance, the polarization difference at 85 GHz (85V-85H), is ranked the fourth over the ocean.

Figure 3 provides an example of the association between the top-ranked features and the rain rate in terms of scattergrams, for three different surfaces, taken from a few profiles. The linear fitting trend is quite evident, and this gives us the confidence of using the RReliefF ranked features to propagate into the MARS model. Note that, in this article, only these top 5 features are allowed to participate in the MARS model. However, the choice of using the top 5 features is somewhat arbitrary. Eventually, the use of top 5 features will be computationally less expensive than the use of all the features. Nevertheless, there is flexibility in the RAMARS, to fine-tune the number of features to be participated. Despite the use of only 5 features, the performance is found to be reasonable, which we will be demonstrating in Section 5.

4.4. MARS component

The MARS is the core component of the RAMARS system, which is responsible for producing quantitative rain rate information.

In the present study, according to the definition, the development of the MARS model is engaged in two phases- the forward selection and backward deletion. The maximal number of basis function is set to 21. The GCV penalty per knot is fixed as 3. The piecewise-cubic modelling is adopted. Self-interactions for the input features are not allowed (*s* = 1), and the maximum degree of interactions between the input features is set to *n* x *s*, where *n* is the number of input variables. In our case, the value of *n* is 5 (5 input features), making the interaction levels to 5. Note that, during the backward deletion phase, one least important basis function is deleted one at a time based on the GCV information, and ultimately, a final model is produced.

For the sake of sanity, we tabulate the predictive performance of the final MARS model by using 5-fold cross validation in Table 3. Again, the training is performed using a large number of orbital samples from the 2012-2013 time periods. It can be seen from the statistical measures, the model is well trained to be included in the RAMARS. The calculated correlations are in the range of 0.61 to 0.73 (GCV 13~32). The numbers of basis functions included in the model are 20, 20, and 16 for ocean, land, and coast, respectively.

5. RAMARS ASSESSMENT

In this section, we report the validation of the RAMARS algorithm taking the TRMM PR as “truth” estimate. For the sake of comparison, we also evaluate the performance of our algorithm in comparison with the GPROF 2A12. In order to do the assessment, a “considerable” number of orbital samples are taken into account, independent from the development datasets. The orbital samples are randomly chosen from the 2012-2013 time periods. In the following sections, we include the dichotomous and descriptive assessments from these datasets. Furthermore, for the sake of evaluation, the RAMARS is applied to two cyclonic cases, and also described here.

5.1. Dichotomous assessment

The dichotomous assessment, in other words, “yes-no” assessment is crucial in understanding the accurate rain prediction of an algorithm. The dichotomous assessment is done through a contingency table, built upon “yes”, “no”, frequency of occurrences. Let us consider a contingency table (Table 4), in which joint distribution of observations and predictions are shown. Based on this, a large number of dichotomous scores can be computed. In this article, we consider the following scores:

$$POD = \frac{hits}{hits + misses} \tag{26}$$

$$FAR = \frac{false.alarms}{hits + false.alarms} \tag{27}$$

$$CSI = \frac{hits}{hits + misses + false.alarms} \tag{28}$$

where, POD, FAR, and CSI represents the probability of detection, false alarm ratio, and critical success index, respectively.

In Figure 4, we construct the dichotomous scores as a function of rain rate over ocean, land, and coast surface terrains. It is worth mentioning that only those samples are considered where PR has estimated rain rate (R>0). As can be seen from the figure, the RAMARS performs reasonably well in most of the cases, especially in low rain rate spectrums. Among the dichotomous scores, the CSI is a balanced measure, taking into account both false alarms and missed cases. Nevertheless, the CSI could be somewhat sensitive to the climatology, tending to provide poorer measures for infrequent samples. This is reflected in the figure, showing an exponentially decreasing trend towards the high rain rates.

5.2. Descriptive assessment

Following the dichotomous assessment, here, we accompany the descriptive assessment of the RAMARS algorithm. Figure 5 presents the scatter diagrams of the RAMARS retrieval and the TRMM PR surface rain rate for three surface cases. The scattergrams of 2A12-PR rain rate are also included in the figure. The performance is measured using four statistical metrics, which are- correlation coefficient (Corr), bias (Bias), fraction standard error (FSE), and root mean squared error (RMSE). It is evident that the RAMARS algorithm agrees better with the PR estimate than that of the TMI 2A12 GPROF algorithm. This is true over all three surface types. The correlation coefficients for the RAMARS algorithm are found as 0.48 (Bias -0.01), 0.49 (Bias -0.30), and 0.42 (Bias -0.12), respectively, over ocean, land, and coast surface terrains. In contrary for the 2A12 GPROF, the correlation coefficients are calculated as 0.44, 0.45, and 0.42 over ocean, land, and coast, respectively. The other two statistical measures, the FSE and RMSE, are also in favour of the RAMARS algorithm.

5.3. Case studies (Sandy and Mahasen)

The hurricanes/cyclones cover a large range of rain structures and intensities; therefore, they are very useful to validate the performance of an algorithm. For the sake of illustrating the rain structure field, the RAMARS has been applied to two recent hurricane/cyclone cases –Sandy and Mahasen.

The hurricane Sandy was the most devastating hurricane among the hurricanes taking place in the 2012 Atlantic hurricane season, but having different cyclonic structure than the conventional ones. The Sandy was started with a typical tropical cyclone blowing through the tropics, however, it transitioned into an extra-tropical cyclone by merging with a frontal system coming from the west. Thanks to the TRMM satellite, that has taken a good number of overpass events during the occasion. Both TMI and PR data were able to see the hurricane, and as such, this gives us an excellent opportunity to evaluate our algorithm to an extra-tropical cyclonic occasion. One such good overpass was on 28th October 2012 (orbit 85175), in which the RAMARS algorithm is applied to. Figure 6 provides the rain rate retrieval illustration of the event from the TMI RAMARS and PR 2A25 product. The 2A12 GPROF (both gridded and non-gridded) retrieval was also included in the comparison. As the figure shows, at this particular time, the Sandy became a Category 1 hurricane and its eyewall was modest, containing only light precipitation. However, surrounding its eyewall, the region was experiencing a high intense precipitation. Remarkably, as the figure reveals, the RAMARS algorithm is able to capture the precipitation intensity very well, in agreement with the PR 2A25. Indeed, the performance is comparable to the GPROF 2A12 outputs.

In contrary, the Mahasen was the Northern Indian Ocean tropical cyclone that hit Bangladesh on mid-May 2013, before dissipating over eastern India. A good TRMM overpass occurred on 16th May 2013 UTC 0406. Similar to the Sandy case, we illustrate the Mahasen event in Figure 7. The eye of the storm is somewhat visible, free of precipitation. A band of thunderstorms can be seen in the figure. Apparently, the RAMARS provides a good estimate of the rain rate, taking the TRMM PR as a reference.

6. CONCLUSIONS

This paper proposed a rain rate retrieval algorithm for conical-scanning microwave imagers through three different data mining techniques viz random forest, RReliefF, and MARS (RAMARS). The approach is developed for the tropical region by constructing a database based on the TMI and PR observations. It has been demonstrated that the RAMARS is likely to perform as reasonable as the TRMM PR estimate. Additional evaluation is shown on hurricane and cyclone cases, in which RAMARS is found to reproduce the structure and intensity of the precipitation field.

The fundamental advantage of the RAMARS is that it is not dependent on any NWP or auxiliary information. However, currently, the RAMARS lacks the idea of using atmospheric radiative transfer equations in the retrieval process. Since there is no use of any radiative transfer model, the proposed algorithm can be termed as empirical, not physical. However, it should be fairly straightforward to replace the observed TBs in the database by simulated TBs through a radiative transfer model. Further, by using the radiative transfer model, the RAMARS can be adapted to other sensors with very little effort, such as the AMSR2 on-board GCOMW-1 and Madras on-board Megha tropiques in the GPM constellation.

REFERENCES

[1] T. S. Biscaro and C. A. Morales, "Continental passive microwave-based rainfall estimation algorithm: Application to the Amazon Basin," *Journal of Applied Meteorology and Climatology,* vol. 47, pp. 1962-1981, Jul 2008.

[2] A. Mugnai, E. A. Smith, G. J. Tripoli, B. Bizzarri, D. Casella, S. Dietrich, *et al.*, "CDRD and PNPR satellite passive microwave precipitation retrieval algorithms: EuroTRMM/EURAINSAT origins and H-SAF operations," *Natural Hazards and Earth System Sciences,* vol. 13, pp. 887-912, 2013.

[3] T. Islam, M. A. Rico-Ramirez, D. W. Han, P. K. Srivastava, and A. M. Ishak, "Performance evaluation of the TRMM precipitation estimation using ground-based radars from the GPM validation network," *Journal of Atmospheric and Solar-Terrestrial Physics,* vol. 77, pp. 194-208, Mar 2012.

[4] T. Islam, P. K. Srivastava, M. A. Rico-Ramirez, Q. Dai, D. Han, and M. Gupta, "An exploratory investigation of an adaptive neuro fuzzy inference system (ANFIS) for estimating hydrometeors from TRMM/TMI in synergy with TRMM/PR," *Atmospheric Research,* vol. 145, pp. 57-68, Aug-Sep 2014.

[5] T. Islam, M. A. Rico-Ramirez, D. W. Han, and P. K. Srivastava, "Using S-band dual polarized radar for convective/stratiform rain indexing and the correspondence with AMSR-E GSFC profiling algorithm," *Advances in Space Research,* vol. 50, pp. 1383-1390, Nov 2012.

[6] S. Seto, T. Kubota, T. Iguchi, N. Takahashi, and T. Oki, "An Evaluation of Over-Land Rain Rate Estimates by the GSMaP and GPROF Algorithms: The Role of Lower-Frequency Channels," *Journal of the Meteorological Society of Japan,* vol. 87, pp. 183-202, Mar 2009.

[7] C. Kummerow, Y. Hong, W. S. Olson, S. Yang, R. F. Adler, J. McCollum, *et al.*, "The evolution of the Goddard profiling algorithm (GPROF) for rainfall estimation from passive microwave sensors," *Journal of Applied Meteorology,* vol. 40, pp. 1801-1820, 2001.

[8] C. D. Kummerow, S. Ringerud, J. Crook, D. Randel, and W. Berg, "An Observationally Generated A Priori Database for Microwave Rainfall Retrievals," *Journal of Atmospheric and Oceanic Technology,* vol. 28, pp. 113-130, Feb 2011.

[9] T. Iguchi, T. Kozu, J. Kwiatkowski, R. Meneghini, J. Awaka, and K. Okamoto, "Uncertainties in the Rain Profiling Algorithm for the TRMM Precipitation Radar," *Journal of the Meteorological Society of Japan,* vol. 87, pp. 1-30, Mar 2009.

[10] L. Breiman, "Bagging predictors," *Machine Learning,* vol. 24, pp. 123-140, Aug 1996.

[11] T. K. Ho, "The random subspace method for constructing decision forests," *Ieee Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, pp. 832-844, Aug 1998.

[12] L. Breiman, "Random forests," *Machine Learning,* vol. 45, pp. 5-32, Oct 2001.

[13] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning,* vol. 53, pp. 23-69, Oct-Nov 2003.

[14] J. H. Friedman, "Multivariate adaptive regression splines," *Annals of Statistics,* vol. 19, pp. 1-67, Mar 1991.

[15] R. Spencer, R. Hood, and H. Goodman, "Precipitation retrieval over land and ocean with the SSM/I- Identification and characteristics of the scattering signal," *Journal of Atmospheric and Oceanic Technology,* vol. 6, pp. 254-273, 1989.

[16] R. R. Ferraro, "Special sensor microwave imager derived global rainfall estimates for climatological applications," *Journal of Geophysical Research-Atmospheres,* vol. 102, pp. 16715-16735, Jul 1997.

[17] Y. L. You, G. S. Liu, Y. Wang, and J. Cao, "On the sensitivity of Tropical Rainfall Measuring Mission (TRMM) Microwave Imager channels to overland rainfall," *Journal of Geophysical Research-Atmospheres,* vol. 116, Jun 2011.

[18] T. Islam, M. A. Rico-Ramirez, P. K. Srivastava, and Q. Dai, "Non-parametric rain/no rain screening method for satellite-borne passive microwave radiometers at 19-85 GHz channels with the Random Forests algorithm," *International Journal of Remote Sensing*, vol. 35, pp. 3254-3267, May 2014.