1.23

COMPARISON BETWEEN BAYESIAN TECHNIQUES AND SELF-ORGANIZING FEATURE MAP FOR TMI MEASUREMENT CLASSIFICATION : APPLICATION TO RAIN RATE RETRIEVAL

C. Mallet, N. Viltard, C. Klapisz

CETP/UVSQ, 10-12 Av. de l'Europe, 78140 Velizy, France
Tel: (+33) 1 39 25. 48. 84,: fax: (+33) 1 39 25. 47 78,
email: cecile.mallet@cetp.ipsl.fr

Abstract

This study focuses on the retrieval of liquid precipitation from the synergetic use of both TRMM Microwave Imager (TMI) and Precipitation Radar (PR). Rain rates are retrieved from TMI measurements with two different statistical approaches developed on the same database. This database is made of measured data obtained through co-localisation of the TMI brightness temperatures and the rain rates from PR. The main advantage of such a database is its representativeness which is theoretically perfect.
An emerging statistical tool for remote sensing is the Self Organizing Map (SOM). Algorithms proposed by Kohonen have been used to obtain this map. SOM maps allow the projection of measured brightness temperatures into a two dimensional space. The obtained map is topologically ordered, that is to say: the spatial location of each neuron corresponds to a particular domain or feature of input data. SOM algorithm is thus applied to TMI measurements and the neurons have been labeled with corresponding PR rain rate. The labeled map is used for rain rate retrieval.
A bayesian classifier is also used for comparison. It provides with a retrieved rain rate by calculating a weighted average of the rain rates stored in the database. The weighting is proportional to the distance in the brightness temperature space between the stored and the measured vector. A series of parameters are tuned because they are extremely difficult to estimate. This makes the classifier somehow supervised, as opposed to the SOM.
The two algorithms will be compare in terms of general performances (bias and error variances) and also in terms of topology of the rain field on a case study.

1 Introduction

Rain is an important parameter of the atmosphere because it relates to the energy cycle. Rain measurement is essential for improving the understanding of a wide variety of processes such as climate change, hydrology, meteorology, or for the improvement of weather forcasting by means of data assimilation. On the other hand, only satellite measurements are available on a global basis over ocean. The Visible/IR sensing from space does not permit direct measurement of rain quantities, but are more useful for cloud coverage estimation. The advantge of using the microwave portion of the spectrum is that microwave

Table 1: Pixel sizes as a function of frequency

Channel	10 GHz	19 GHz	21 GHz	37 GHz	85 GHz
Resolution (km)	63x37	30x18	23x18	16x9	7x5

radiation penetrates clouds and interracts strongly with liquid hydrometeores. The TRMM satellite is the first mission to carry a passive (TMI radiometer) and an active (PR radar) microwave instruments together.

PR provides vertical profile of reflectivity at 14 GHz. The TMI measures brightness temperatures at different frequencies : 10.65, 19.35, 21.3, 37.0, 85.5 GHz, polarized both vertically and horizontally but for the 21 GHz channel which is only polarized vertically. TMI swatch is about 600 km large while the PR one is about 200 km. The radiometer is thus particularly useful to obtain global rainfall maps. However, numerous difficulties exist to retrieve the surface rain rate from the radiometric brightness temperatures. In fact, radiometric measurements are representative of a measure of the total column of atmosphere within the field of view. The physical basis is that the radiation incident upon a radiometer antenna is a frequency dependent complex function of the vertical profiles of: atmospheric temperature, pressure, water vapor, and cloud liquid water, liquid and ice precipitation. Other parameters, such as the size of hydrometeor droplets and characteristics of the earth surface emission and reflectivity, must also be considered. The difficulties generally encountered to perform surface rain retrieval are due to the complexity and the non-linearity of this relationship. Moreover, the presence of hiden parameters (such as ocean surface characteristics or temperature profiles) increases the difficulty. Multi-frequency measurements are thus useful to take into account the contribution of these different components. However, the variation of the spatial resolution of the different channels (see Table 1) introduces noise in using all the channel to retrieve surface rain corresponding to a particular resolution.

Most retrieval methods needs the use of a database of brightness temperatures associated with corresponding surface rain to overcome this highly underdetermined problem. The database will contain the implicit physics that links the vertical profiles of precipitation with the "only" 9 measurements. This can be done through simulation in computing the brightness temperature associated with possible atmospheric profiles (e.g. Kummerow *et al.*, 1996 and Olson *et al.*, 1996). The main difficulty is to obtain a wide and representative set of atmospheric profiles with an horizontal resolution that allows the simulation of the variation of the spatial resolution of the different channels ; and also with knowledge concerning the microphysic of the differents hydrometeor (raindrop size distribution, melting layer,...) to allow the simulation of the higher frequency channel.

In this study, our main objective is to compare two different retrieval methods based on classification algorithms. We thus decided to develop the two methods on the same database, using the specificity of the TRMM satellite. In fact the database is obtain from co-located observation of surface rain rate and brightness temperatures. The surface rain rate is measured with PR reflectivity profiles and brightness temperature are measured with TMI. Studies have

proven the consistency between PR and TMI measurements (e.g. Viltard *et al.*, 2000). The main disadvantage of this approach is that the quality of the obtained retrieval method is highly dependent of the quality of surface rain rate deduced from PR measurements.

2 Characteristics of the data base

The TMI and the PR instruments are observing the earth surface under very different geometry. The PR is a cross-track scanning radar ($\pm$17 degrees off-nadir), leading to an almost- regular grid at the Earth's surface with a pixel roughly every 4.5 km in both cross- and along-track direction. On the center of PR swath, the radar bins reach the earth surface but on the edges of the swath, the minimum altitude of a measurement is about 1.5 km due to contamination by ground clutter of the lower bins. The TMI is a conical scanning instrument with a constant incident angle at the surface (52.8 degrees for the beam centrers) with the pixel size dependng dramatically on the frequency. The distance between two pixel centers in the same scan is 9.1 km at the low frequencies while the distance between two scans is about 13.9 km. Only the PR pixels from the center of the swath are kept, minimizing the risks of erroneous rain rates due to poorly corrected surface echoes.

The remaining brightness temperature vectors are affected with the PR-estimated surface rain rate, averaged over a circular area of 12.5 km radius surrounding the TMI pixels center. Whithin such a circle there are about 20 PR pixels. This average represents approximately the resolution of the 37 GHz channel and corresponds approximately to the spatial resolution of the Gprof or 2A12 algorithms (Kummerow and Giglio, 1996 and Kummerow *et al.*, 2001).

The PR is an attenuated radar that requires a complex processing accounting also for beamfilling errors and drop-size distribution hypotheses. In this study we use a $k - R$ relationship presented in Ferreira *et al.* (2001) and very close to the standard algorithm known as 2A25 (Iguchi *et al.*, 2000). A quite complete description of the TRMM instruments and standard algorithms can be found in Kummerow *et al.* (2000).

The resulting database is made of about 60,000 rain rates and their associated brightness temperature vectors. These profiles are distributed all over the tropical belt covered by TRMM ($\pm$35 degrees latitude). They are collected over a winter (February) and a Summer (May) month. Contamination by land and/or coast is filtered. One might notice that the PR sensitivity threshold is about 17 dBZ, which means the radar cannot detect rain below $\approx$0.1 mm.hr^{-1}. This has no impact on our study but should be considered if one wanted to estimate the retrieval error to some ground truth. Also, because the rain at 37 GHz-pixel resolution is the average of about 20 PR pixels, the database contains, indeed, rainrates below 0.1 mm.hr^{-1}.

3 Description of the retrieval methods

3.1 SOM-based Technique (SBT)

A co-located data base is used to train a Self Organizing Map (SOM) as proposed by Kohonen (1982). Each data used is a nine-dimensional vector made of the nine brightness temperatures. A representation of the whole data set (100,000 vectors) with 2D map of 25x25 neurons is generated. Each of the 625 neurons corresponds to a cluster of neighbor data of the input space. The number of data retained in each cluster depends on the density of the considered cluster.

In fact SOM of Kohonen is a vector quantization algorithm in which a set of unlabelled data vectors in a n-dimensional space (here n is equal to 9) is represented by a set of n-dimensional reference vectors which are linked according to a topological order. This order is defined by the map whose cells are discrete points of a reduced 2-dimensional space. Kohonen algorithm is a non-linear projection of the discrete space of the cells (the map) into the n-dimentional data space in which each cell corresponds to a reference vector. The main characteristics of this algorithm is to preserve the topological order, two adjacent cells are mapped into two close reference vectors. This can be view as a vector quantization with a neighboring constraint.

The obtain map can be used in classification tasks, combining supervised and unsupervised learning. This is done by labeling each cell of the map using labels. In our case we have a large amount of 100,000 labelled data, they allow us to determine a consistent classifier. Each cell of the map C_i being labeled using the mean rain rate R_i computed on the set of training data assigned to this cell. After the labeling process we can observe a set of cells C_i having the same label R_i; this corresponds to identical surface rain observed in different meteorological situation, which leads to different brightness temperature measurements because of heterogeneities, microphysics, ... The labeled map can thus be used as a retrieval method. Each pattern (composed of the nine brightness temperatures at different frequency and polarization) is assigned to its nearest cell C_j of the map and takes the label R_j of this cell C_j. So the labeled map can be used as a "hard" classifier.

This method present the advantage of being very simple and rapidly implemented. Only the first step (the training step) is time consuming. The labeling step and the use of the obtained map as a retrieval algorithm is very fast. Moreover, recent developments of probabilistic SOM, which approximates the density distribution of the data with a mixture of normal distribution can be used to improve this method Yacoub *et al.* (2000) in obtaining a more accurate and soft classifier, it will be thus possible to propose different rain rate with their related probabilities.

3.2 Bayes-Monte Carlo approach (BMC)

This retrieval algorithm is based on the Bayes theorem as extensively described in Kummerow *et al.* (1996, 2001) and Olson *et al.* (1996). The probability that a given rain profile W is associated with a set of measured brightness

temperatures T_{Obs} can be expressed as:

$$E(W) = \frac{1}{A} \int\int\int \ldots \int W_j \exp\left[-0.5 \quad \left(T_{Obs} - T_{Dbase}(W_j)\right)^T (O+M)^{-1} \left(T_{Obs} - T_{Dbase}(W_j)\right)\right] dW \tag{1}$$

where T_{Dbase} is the brightness temperature vector set in the database associated with the rain profile W_j, and O is the observation error covariance matrix and M is the error covariance matrix of the PR rain rates. A is a normalization factor defined as:

$$A = \int\int\int \ldots \int \exp\left[-0.5 \quad \left(T_{Obs} - T_{Dbase}(W_j)\right)^T (O+M)^{-1} \left(T_{Obs} - T_{Dbase}(W_j)\right)\right] dW \tag{2}$$

Finally, this expression is reduced using the Monte-Carlo method where the integral in 1 is evaluated over a large number of realizations of the retrieval parameters. It is also assumed that the O and M are diagoonal, *i.e* the errors are uncorrelated. Hence:

$$W = \sum_{k \in Dbase} \frac{W_k}{N} \exp\left[\sum_{i=1,9} -0.5 \frac{1}{\sigma_i^2} \left(T_{Obsi} - T_{Dbasei}\right)^2\right] \tag{3}$$

N is the re-written form of A, easily deduced from (2). The real function that is minimized actually uses the polarization signal to minimize the surface and thus the beamfilling effect. Then, 3 becomes:

$$W = \sum_{k \in Dbase} \frac{W_k}{N} \exp\left[\sum_{i=1,5} -0.5 \frac{1}{\sigma_i^2} \left(\Delta T_{Obsi} - \Delta T_{Dbasei}\right)^2\right] \tag{4}$$

where $\Delta T_{Obsi} = TVertical_{Obsi} - THorizontal_{Obsi}$ and $i = 1$ stands for 10 GH z channel, $i = 2$ stands for 19 GHz channel, etc ... For the specific case of 21 GHz that exist only in the vertical polarization, we simply have $\Delta T_{Obs3} = TVertical_{Obs3}$. The BMC algorithm for instance is sensitive to the σ_i in 4 which are set empirically. The method works somehow as a classifier, close to the SBT in essence which is the reason why a comparison is interesting.

4 Results

Hereafter, two types of comparisons are performed. The first one summarizes the statistical properties of the two algorithms, the second one displays a real rain field. This allows to compare the objective performances, but also the final appearance of the rain field and its spatial coherence.

4.1 Statistical comparison on a wide test data base

The database for this first series of test is generated from profiles also collected in May 2001 and February 1998. It contains about 20800 rain rates collected between ±39 degrees. The distribution is natural with a strong representation of

Table 2: Statistics for the Neural network on the test database

Rain Range ($mm.hr^{-1}$)	Error Bias ($mm.hr^{-1}$)	Error Standard Deviation ($mm.hr^{-1}$)	RMS of the Relative Error (%)
0.0 - 0.5	0.44	0.73	1476.
0.5 - 1.0	0.57	1.30	200.
1.0 - 2.0	0.78	1.87	144.
2.0 - 3.0	1.06	2.56	113.
3.0 - 4.0	1.00	2.87	88.
4.0 - 5.0	0.70	3.20	74.
5.0 - 6.0	0.13	3.42	63.
6.0 - 8.0	-0.56	4.07	60.
8.0 - 10.0	-1.59	4.71	55.
10.0- 15.0	-3.54	5.08	51.
15.0- 20.0	-6.40	7.26	56.
20.0- 35.0	-9.58	9.70	56.
TOTAL	0.41	2.44	983.

Table 3: Statistics for the Bayesian method the test database

Rain Range ($mm.hr^{-1}$)	Error Bias ($mm.hr^{-1}$)	Error Standard Deviation ($mm.hr^{-1}$)	RMS of the Relative Error (%)
0.0 - 0.5	0.42	0.38	202.
0.5 - 1.0	0.35	0.52	85.
1.0 - 2.0	0.36	0.81	61.
2.0 - 3.0	0.29	1.18	49.
3.0 - 4.0	0.18	1.60	46.
4.0 - 5.0	-0.02	1.97	44.
5.0 - 6.0	-0.21	2.45	45.
6.0 - 8.0	-0.36	3.69	54.
8.0 - 10.0	-1.20	4.91	56.
10.0- 15.0	-2.44	6.30	56.
15.0- 20.0	-6.71	7.20	58.
20.0- 35.0	-6.58	10.81	57.
TOTAL	0.048	2.23	-NA-

low to moderate rain rates (below 5 mm.hr^{-1}). For both methods, the retrieval is performed on this test database and the results are summarized in Table 2 for the SBT and in 3 for the BMC.

We can see from these two tables that the results are very close in terms of performances, either for the bias or for the error standard deviation. At this stage, the BMC is actually a more advanced version and thus its parameters are more optimized which might eventually give some better results on some of the cases. Both algorithms are indeed overestimating the rain at the low end of the spectrum and underestimating at the high end. The RMS of the relative errors might seem high but they maximize error accounting for both the bias and standard deviation for each class of rain.

4.2 Comparison on a Hurricane case

At this point, it is important to check the horizontal texture or the retrieved rain field. An excellent score in terms of bias might exhibit a very poor structure of the rain field which is accpetable for climatological studies at large scale, but unacceptable for instantaneous retrievals.

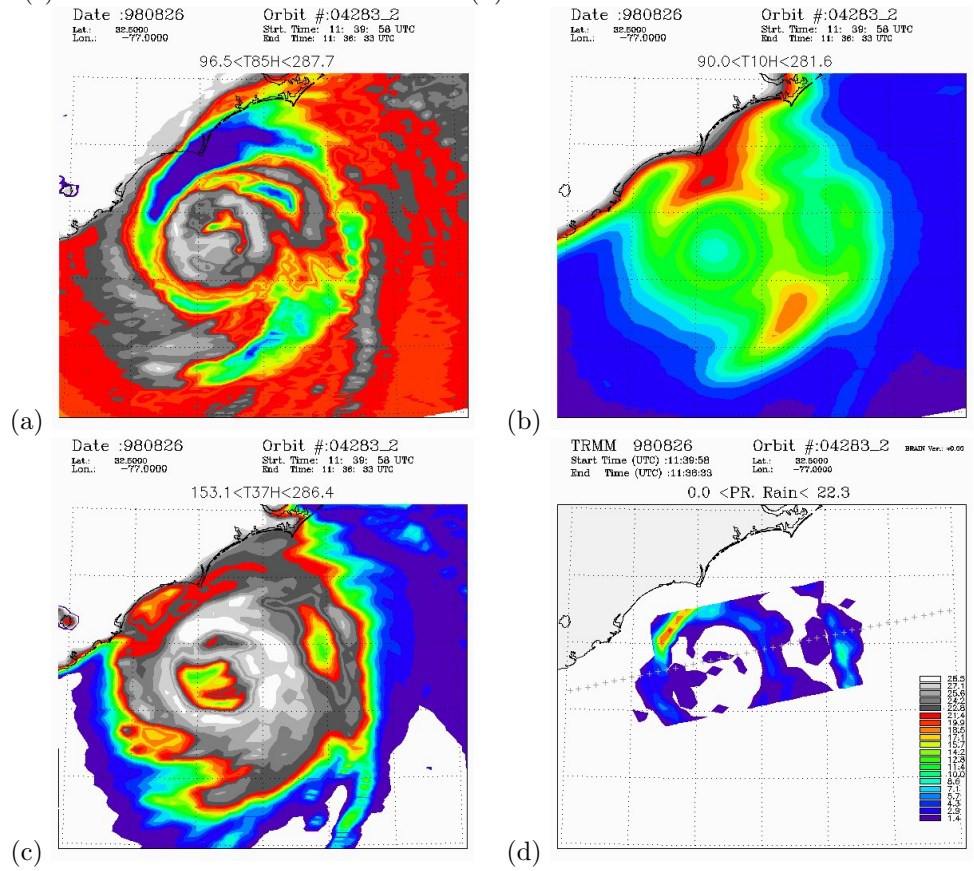
Hurricanes have a major human and economical impact and thus their forecast in terms of trajectory and evolution is very critical. They are known to be poorly represented in global forecast models and they are difficult to observe because they start over oceanic regions most of the time. This is where spaceborne sensors are of interest and particularly passive microwave radiometers. On the other hand, hurricane are complex situation where the very strong surface wind make it really difficult to develop a sea surface model for emissivity. Furthermore, the rain field is made of structures that can be at the very limit of spaceborne passive radiometer resolution.

Hurricane Bonnie was observed on 26 August 99 near the US coasts. It was a moderate hurricane with a wide eyewall at landfall. A strong asymetry in the rain intensity can be noticed on the North-Western quadrant, under the influence of the coast.

Fig. 1 shows the brighntess temperatures at 85H (85 GHz, Horizontal polarization), at 10H and at 37H, where the blue and green colors represents the coldest temperatures and the red to dark to light grey represent respectively the warmer temperatures. It presents also the rain as seen by the PR but averaged at the 37 GHz resolution, using the same color convention for the rain rates (green to red to light grey represents increasing rates). The 85H exhibits mostly the cold signature of scattering by precipitation ice in the convective regions over a warm background. 10H and 37H represent the warm emissivity signal of rain over a cold ocean background, with eventually for higher rain rates the scattering of dense ice (at 37H).

One can also notice the large differences in terms of structure resolution between the different channels and how they are correlated to the rain rate observed by the radar. One will then understand that the effects of resolution differences will add up to the effects of non-linearity between the variables to

Figure 1: Brightness temperature at 85 GHz H (a), 10 GHz H (b) and 37 GHz H (c) and PR rain at 37 GHz resolution (d)

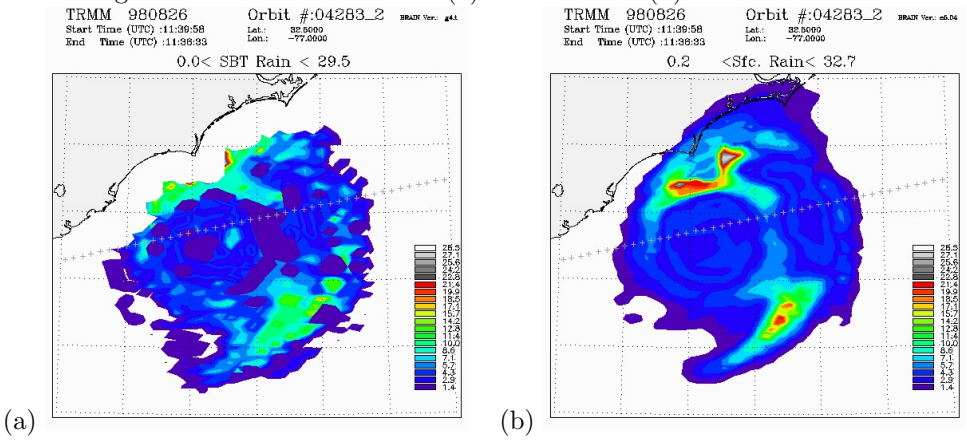


retrieve and the brightness temperature.

The two retrieved rain fields are presented Fig. 2 (a) for the SBT and (b) for the BMC. One can notice that both rain field are coherent spatially with the large scale structures of the storm quite properly retrieved. The SBT sets to 0 mm.hr^{-1} values that are unproperly learned in the training phase. These are represented by the dark blue spots spread in the middle of light blue and light greens. Indeed, a few cases of hurricanes are set in the learning database but they certainly are a minority both in terms of rain intensity and in terms of background oceanic emissivity. This is may be were the SBT classifier is showing its pros and cons at once: its inability to extrapolate results that were not learned. On the other hand, the BMC shows that it can always find a solution even if this solution might be questionable in terms of realism.

The bias with radar (which is our reference here) over the whole domain are respectively 1.65 mm.hr^{-1} for the SBT and 1.25 mm.hr^{-1} for the BMC. These

Figure 2: surface rain rate for (a) the SBT and (b) the BMC



(a) (b)

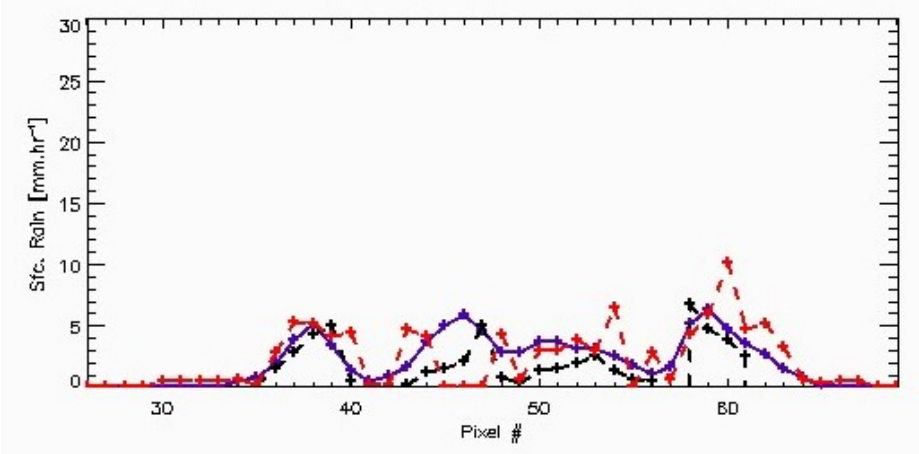
values are higher than those over the 20800 profiles, probably because of the difficulty of the hurricane case with its special condition. Notice also that the SBT was not trained to retrieve rain over land and/or coast and this is why the upper left part of the hurricane is not showed on Fig. 2(a).

Finally Fig. 3 is a comparison of a transect along the line made of grey crosses on Fig. 1(d). This transect allows us to see the variations of the rain field along the nadir line. We can see that the agreement between the two estimators and the reference is extremely variable, with sometimes substansial overestimations and sometimes excellent agreement. It is also interesting to see that it seems for some of the local maxima that there is a spatial shift which might be due to the two instruments geometry difference.

5 Conclusion

This study is trying to demonstrate the equivalence of Self-Organizing Map Based Techniques (SBT) and Bayes-Monte Carlo approaches (BMC). Although the two algorithms presented here are still under developpement, they exhibit results which are very close in terms of bias and structure: both underestimate the high rain rates while overestimating the lower ones. These two algorithms were trained on the same data base built from co-located TRMM Precipitation Radar and Microwave Imager, getting rid of the usual representativeness problem. Doing so, we also got rid of the many problems related to the direct radiative transfer simulation particularly in the ice phase. On the other hand, we have no or a very few control on the physics of the problem and we have to assume that the radar is providing us with a “good” estimate of the rain.

Figure 3: Comparison of rain rates retrieved along a transect. Blue line stands for BMC, red line stand s for SBT and black line is the radar reference



Some substansial improvements are going on right now for both algorithms. For the neural network approach, the database used was kept as is, which means a natural distribution of rain was used, emphasizing the lower rain rates. A specific work on the histogram of both the rain rates and the brightness temperature will be done. The Bayesian technique suffers from its uncertainties in adjusting the weightings or variance/co-variance error matrices. They are now adjusted empirically instead of adapted to the specific database in use. This will be implemented in the future.

For the two algorithims some more research will also be performed in trying to reduce the high variance of the brightness temperatures relative to rain rate by using scattering and emissivity indexes to get rid of hidden parameters such as background temperatures of surface emissivity biases.

6 Bibliography

Ferreira F.and P. Amayenc, 1999: Impact of adjusting rain relations on rain profiling from the TRMM precipitation radar, *Proc. 29th Conference on radar Meteorology*, Montreal, Quebec, Canada, Amer. Meteor. Soc.,643-646 (12-16 July 1999)

Kummerow, C., W. S. Olson and L. Giglio, 1996: A simplified scheme for obtaining precipitation and vertical hydrometeor profiles from passive microwave sensors. *IEEE. Trans. Geosci. Remote Sensing*, **34**, 1213-1232

-------, J. Simpson, O. Thiele, A.T.C. Chang, E. Stocker, R.F. Adler, A. Hou, R. Kakar, F. Wentz, P. Ashcroft, T. Kozu, Y. Hong, K. Okamoto, T. Igushi, H. Kuroiwa, E. Im, Z. Haddad, G. Huffman, T. Krishnamurti. B. Ferrier, W.S. Olson, E. Zipser, E. A. Smith, T.T. Wilheit, G. North, K.

Nakamura, 2000: The Status of the Tropical Rainfall Measuring Mission (TRMM) after 2 Years in Orbit, *J. Appl. Meteor.*, **39**, 1965-1982

Kummerow, C., Y. Hong, W. S. Olson, S. Yang, R. F. Adler, J. McCollum, R. Ferraro, G. Petty, B.-B. Shin T. T. Wilheit, 2001: The evolution of the Goddard profiling algorithm (GPROF) for rainfall estimation from passive microwave sensors, *J. Appl. Meteor.*, **39**, 1801-1820

Iguchi, T., T. Kozu, R. Meneghini, J. Awaka, and K. Okamoto, 2000: Rain profiling algorithm for the TRMM precipitation radar, *J. Appl. Meteor.*, **39**, 2038-2052

Ferreira, F, 2001: Exploitation des données du radar de TRMM pour l'estimation de la pluie depuis l'espace, Thèse de doctorat de l'Université Denis Diderot (Paris 7), pp235

Olson, W. S., C. D. Kummerow, G. M. Heymsfield and L. Giglio, 1996: A method for combined passive-active microwave retrievals of clouds and precipitation profiles. *J. Appl. Meteor.*, **35**, 1763-1789

Viltard N., C. D. Kummerow, W. S. Olson and Y. Hong, 2000: Combined Use of the Radar and the Radiometer of TRMM to Estimate the Influence of Drop Size Distribution on Rain Retrievals, *J. Appl. Meteor.*, **39**, 2103-2114

Yacoub, M., D. Frayssinet, F. Badran, and S. Thiria, 2000: Classification based on Expert knowledge propagation using Probabilistic Self-Organizing Map: application to geophysics. In *Data Analysis: scientific modeling and practical application*, Springer-Verlag (Studies in classification, data Analysis, and knowledge organization)