Review of AMTD paper by J. P. Musial et al. entitled

”Probabilistic approach to cloud and snow detection on AVHRR imagery"

General impression

This manuscript presents a novel approach to cloud masking (PCM) based on the estimation of cloud and snow probabilities for a very large set of possible conditions inferred from original image radiances and various ancillary datasets. The latter include e.g. NWP-analysed skin temperatures, various land use classes, different illumination and viewing conditions, and the time of observations. An interesting feature of the method is the use of some Principal Component transformations of some features to compress the information content.
The method is trained using external datasets, i.e., the NWCSAF/PPS cloud mask and MODIS-derived snow information. Validation results show good performance that is equal or better than reference datasets (mainly achieved through some supervised enhancement of the basic input).
The descriptions are generally of good quality and based on sound scientific reasoning. Nevertheless, a quite large number of issues still need some further consideration and some clarifying discussions. One of the more serious issues is why the PCM method is not explicitly taking into account atmospheric absorption effects caused by atmospheric water vapor. There are also numerous editorial changes to make for improving the text.
All issues (further specified below) need to be addressed before approval.

Comments:

- Excellent review of current methods and especially the use of different image features in section 2!

- Page 8463, land cover use: This is an alternative approach to account for surface-induced variability in cloud-free radiances from Earth surfaces. Most other methods are going towards using a more physical-based way of accounting for this through the use of seasonally or even monthly varying (mostly MODIS-retrieved) surface reflectivities and surface emissivities. The latter are also affected by soil moisture effects that could be used in addition to modify this input. Your version of handling this is to use a static land use database (although maybe the best available dataset so far) and to update your statistics for different times of day and for four different seasons. It would be interesting in the future to see a more systematic inter-comparison of these two approaches. Both approaches have their weaknesses in that they are essentially not able to compensate for the large inter-annual, seasonal and even day-to-day variation which exists. This is especially critical for climate monitoring applications spanning over several decades.

- The time interpolation of NWP SKT information and the modification of SKT by using detailed topography information from DEM are indeed important steps taken for improving classification quality. Good.

- Unfortunately I would not give you good chances of getting the cloud shadow estimation very accurate. This has mainly to do with the fact that only a small fraction of all clouds have BTs in 11 µm which are close to the true cloud top temperature. If not trying to correct for this you will generally find far too low cloud top heights and thus too short distances with shadows. I think that this is actually visible already in your example (compare Figure 5 a and c where the estimated shadow appear to be smaller than the observed one).

- Page 8469, section 3.7, general question: You have claimed (somewhere in the text) that the speed of computation is much faster than for traditional multispectral thresholding schemes. In fact, I doubt it. The reason is that the actual multispectral thresholding process can be executed very fast (if programmed in C and Python) and mostly in less than a minute for an ordinary HRPT scene received locally. Because of the need to search in the large LUT I am not convinced that PCM is significantly faster than this. In addition, what takes a lot of time is the preparation of image features and ancillary data. This takes normally much more time than the actual testing of thresholds. Since this is dominating the processing time I think that you cannot claim processing speed as the big advantage of the PCM method. Only if comparing with true (not naïve) Bayesian classifiers processing speed could still be an issue.

Questions and critical remarks:

- Page 8449, line 18: You probably mean “latter” when you write “former” here, right? I mean, a scheme that takes into account atmospheric effects should be more robust, don’t you think?

- Page 8458, line 13: To use the word “retrieve” here is not entirely correct. I would suggest the word “approximate” since this is really a rough approximation that is only valid for optically thick clouds (with true blackbody appearance).

- Page 8453, paragraph “Reflectance tests in the 0.6 & 0.8 µm bands”, line 9: Your description of the use of the factor Cosine of Sun Zenith Angle is not entirely correct. You express it as something that is nice to do. Actually, it is absolutely necessary for getting the true reflectance correct for any horizontal surface (e.g. satellite FOV) on Earth. This then takes into account that the amount of radiation reaching a surface will decrease with Solar Zenith Angle. If you don’t do it you will have reflectances that always decrease with increasing Solar Zenith angles. The problem encountered at very high SZAs (close to 90 degrees) is that you risk to divide your reflectance with something that approaches zero which will result in extremely high reflectances (e.g. for clouds being illuminated on their sides). The actions taken by Dybbroe et al (2005) was just for still allowing the use of visible radiances for cloud screening even at these very high SZAs.

- Page 8455, lines 12-14: The statement is generally true over land surfaces but not over ice free ocean. For the latter, this feature is of vital importance at night and one of the reasons why we get more clouds over water surfaces than over land surfaces. It is simply a consequence of that clouds are more easily identified over a warm and comparably homogenous surface than over a cold and often inhomogeneous one. This artificial bias in cloud climatologies over land and ocean is probably something we have to live with.

- Page 8459, line 16: The use of the ICS transformation is actually one of the more important new features of this methodology compared to other methods. As such, it should have deserved a more prominent place in the descriptions, in my view. Preferably, you should have written a separate paper on this which you could have referred to. For example, now we are just given some facts on how it was finally implemented without any information about why just those restrictions mentioned a few sentences further down have been imposed. At least add a short discussion on this.

- Page 8462, line 1: You write that "multiple, irregularly distributed threshold values per feature" resolves the problem of single threshold estimation by multi-spectral thresholding methods. I have a problem with how you use the word "threshold" here (and in many other places in the text). What you refer to is the actual binning size/distribution of your feature values and not thresholds in the meaning it is used in multi-spectral schemes. So, I would actually advise you to try to use another word here for avoiding confusion. You are not using multiple thresholds in the sense of traditional thresholding schemes. Instead you subdivide your feature values in discrete categories (which could be quite numerous), each of which will be given different cloud probabilities (depending also on all other features). In that way you allow a much more flexible way of finding your cloud mask which could indeed mean that clouds can be found not only above a certain threshold for a feature value but in several sub-sections of your feature space. This is good but try to avoid describing your bin boundaries/sizes as being thresholds.

- General, pages 8457-8462: Your PCM scheme is very simple in the sense that it is only estimating the cloud frequency (cloud pdf) in multi-dimensional bins. Thus, it is more or less an empirical method. With such a method, the quality of final products will surely be sensitive to what input parameters/features you have chosen to define your multi-spectral domain. Most important is that you try to cover all the variability that exists. Essentially, I think you have succeeded in doing this but there is one major exception: How do you account for atmospheric absorption effects? This is a central question for all satellite remote sensing applications. You have indeed mentioned the importance of atmospheric absorption effects in the general description of which image features that normally have been used so far. There are several features that e.g. are sensitive to the effects of atmospheric water vapour. This is especially the third spectral feature but also the SKT-10.8 µm BT difference used for the first and second enhanced features. You do have a dependence on viewing angle and azimuth angles (which takes care of some of these effects) but not on the day-to-day variation of the total moisture content in the atmosphere (more than what is given on average when looking at different seasons). This is likely to create quite some noise in your results, thus probably lowering or smoothing out the estimated cloud probabilities. You may have some implicit compensation for this through the use of PPS cloud masks (which you are piggy back riding on in the training) which have taken these effects into account. I think that if you had used also total integrated moisture as an additional feature, your results could have been even better. Here, your scheme is actually inferior to most other schemes. This issue becomes much more important if you try to apply your method in warmer and more moist climate regimes (e.g. in the Tropics).

- General comment: The most critical part of this manuscript is to motivate why this methodology should be superior to the methods which have been used for training (i.e., PPS and MODIS retrievals). I think this is missing to some extent. The output of probability estimates instead of fixed cloud masks is certainly one such aspect but the question on why overall results should be improved is not really discussed. Achieved improvements in terms of improved scores in the validation exercises are also quite modest. I would like to see some more arguments here (and I think there are some).

- Another critical part of this work is the supervised enhancement of the PPS cloud masks that were used for the training (and which probably explains why validation scores are slightly better than for PPS). It is said that some obvious errors in the PPS cloud masks were removed. Since these were introduced by subjective methods it means that an additional uncertainty has also been introduced. A few more words on how the supervised training was performed are recommended. For example, was it just depending on one (analyst) person's opinion or was it supported by further more objective tools? I don't say that it is difficult to identify misclassifications but since misclassifications are normally due to spectral signatures being very close to each other it is not obvious that the human eye is always able to tell the truth.

- Page 8465, step 3, line 20: I do not understand why you have to select the two most frequent categories (of the three cloud, snow, cloud free) in the definition of the probability. Isn't this to make too much violence to the true (observed) class frequencies? I mean, originally you did estimate frequencies for all three categories. If now discarding one of them, it means that you throw away important information. Or did I misinterpret this? Tell me what happens for a bin where original frequencies for clouds, snow and clear-sky are 32 %, 34 % and 34 %, respectively? If I interpret Equation 3.2 and the text descriptions properly this would yield P=34 % probability of clear-sky implying 66 % probability of snow and zero % probability for clouds. Correct? If not, you have to describe the process better. If I interpreted it correctly it means that you are not really providing correct probability estimates. In the example above it means that you will significantly underestimate the cloud probability. Why couldn't you have split your method in providing two separate output items: 1. Output the cloud probability 2. Output the snow probability. From this, you would then easily be able to calculate the remaining clear-sky probability. Some users would only need 1 while others may need to use both. Please comment.

- Page 8496, Figure 7: I am sorry but the description in the caption of Figure 7 made it very difficult to understand these sub-panels. You write that the "data quantity is presented as grey-shaded histograms". It would have been better to write "data frequency" or "number of cases". The term "data quantity" could be misinterpreted as the PCM-PPS difference (honestly, I did it). For a long time I really did not understand what the black points and red curve meant. Please modify.

- Page 8496, Figure 7: As clearly documented in available ATBD and in other documents, PPS uses generally (except for mountainous areas) a temperature offset or threshold of 7 K in the BT 11 μm test against SKT. This is most likely explaining why you get this sharp feature in the temperature difference interval 5-10 K. The explanation referring to a different use of the texture test over ocean surfaces is probably not responsible for this, or, at least only to a small extent. My interpretation is that the probabilistic estimate will smear out results on both sides of this threshold value causing both underestimations and overestimations when you compare again to PPS values. That's why results are jumping like this in this interval.

- Page 8472, lines 13-18, discussion in last paragraph: Again, referring to a previous item, how can you be sure that the PPS features over Spain and over northern Africa are artificial? And, if so, how do you explain the PCM maximum in cloud cover in central and south-eastern Spain in Figure 9? This is not seen in PPS results. Do you mean that PPS both misses and creates artificial clouds over Spain? It's quite confusing.

- Page 8473, lines 14-16: There is something strange with the discussion of the matching/inter-comparison of results of NOAA, AQUA and TERRA orbits/scenes. The indisputable truth is that the three satellites NOAA-16, NOAA-18 and AQUA were placed in afternoon orbits (i.e., local solar time when passing the equator close to 1:30) meaning that they will practically follow the same orbit (i.e., being orbiting in the same orbital plane). Thus, scenes from these three platforms should be easy to inter-compare. The "only" problem with NOAA-16 is that it has drifted away from its original orbit. However, there should still be quite some overlap in some of the overpasses between NOAA-16 and AQUA to perform AVHRR-MODIS inter-comparisons, in my opinion. At least, the statement "…only the one labelled with 18 has a sufficiently close orbit to the TERRA and AQUA satellites…" is remarkable since it is very clear that TERRA is a morning satellite and should be much farther away from the AQUA orbit than the corresponding NOAA-16 orbit. Or did you lose those NOAA-16 cases because of HRPT conflicting reasons (e.g. NOAA-16 overpasses came too close in time to NOAA-18 overpasses)? It seems you have some loss of received scenes compared to the theoretically possible ones. Please explain better what you mean.
  (I realise, however, that the orbital drift has been substantial for NOAA-16 so perhaps it is enough that you just confirm that this is the problem).

- In the same sense, NOAA-17 and TERRA are both in morning orbits (although not in exactly the same orbital plane) and should for the same reason allow inter-comparisons of AVHRR and MODIS results. The only place where scenes from morning and afternoon orbits can be inter-compared in a reasonable way is in the most northerly part of the covered area (close to 70 degrees latitude). My question is: What happened to overlapping scenes between NOAA-17 and TERRA? You should have data also for those cases. Please comment this.

- Page 8480, lines 2-4: It is concluded that PCM gives less clouds over ocean surfaces compared to PPS. But please remember that there are actually indications from various validation exercises (e.g. based on CALIPSO-CALIOP) that PPS still misses a substantial amount of clouds over ocean. Thus, be a bit careful in the discussion here, especially with regards to what we can consider as the truth.

Editorial remarks:

- Page 8450, line 7: Change "safe" to "save".
- Page 8453, section "Reflectance tests in the 1.6 & 3.7 µm bands", line 20: Please explain NDSI (never defined previously).
- Page 8454, lines 9-10: Maybe a bit unfortunate formulations here. To say that this test is only used at night and at the same time write that it scatters radiation more effectively at this wavelength is not consistent (since there is no sunlight to reflect during night). Write that the test is used both during day and during night but that the appearance is very different day and night for especially water clouds (reflecting a lot

during day at 3.7 micron and therefore not being blackbodies leading to a colder appearance at night in this channel).

- Page 8445, line 2, first sentence in section “Temperature difference….”: Rephrase this sentence to “The spectral region around 10.8 µm is only slightly affected by the absorption by atmospheric gases (a region we normally call an atmospheric window), thus, it approximates well the surface temperature, at least in regions well outside the Tropics.”
- Page 8445, line 10: Minor change of sentence to “Over barren or sparsely vegetated areas such as deserts, the strong diurnal surface temperature cycle……”
- Page 8445, line 18: Be careful how you use the word “texture”. Texture is a very general term (meaning a lot of things) but here we just look at the local variation of radiances within a certain pixel window. Skip the word “texture” and replace it with the “local radiance variation”.
- Page 8457, line 19: replace “principals” with “principles”.
- Page 8470, line 7: You should write “verify the agreement” rather than “verify the difference”, right?
- Page 8472, line 21: Change to “Another problematic region is….”.
- Page 8474, line 6: Change PPS to PCM (Check! There is no results for PPS data in Figure 12!).
- Page 8476, line 26: Write “Their shape corresponds very well to other….”
- Page 8480, line 1: Change “the the” to just “the”.