Comments:

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

Response: The detailed theoretical introduction was meant for readers which are not familiar with satellite cloud analysis.

-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.

Response: Seasonal variability of surface reflectance and emissivity indeed affects cloud-free radiances, therefore in the PCM it is handled by derivation of different LUTs for different seasons (as the Referee has mentioned). However, from the perspective of Authors more important is the diurnal variability and BRDF induced changes of the reflectance which can be as pronounced as the entire annual cycle. To account for this in the PCM particular image acquisition geometry has its own position in the LUTs along with the relative azimuth and sensor zenith dimensions. This allows comparing /retrieving probability from the training scenes with similar angular conditions to the analyzed image, which mitigates (to some degree) the BRDF effects. Regarding, the emissivity the diurnal cycle can be also comparable to the annual one for example after the rain event over the bare soil areas. To conclude the variability of the clear-sky reflectance and emissivity is inevitable and it can be only partially handled by the utilization of annual-depended input data sets.

-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.

Response: Indeed Authors could see significant improvement especially regarding the cloud/snow misclassification.

-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).

Response: The Referee's remark is completely true therefore Authors on purpose called this process “cloud shadow estimation” and not “cloud shadow detection”. The main aim of this analysis is to provide the enduser with a crude estimation where cloud shadow could be based on the simple geometrical relationships. This information can be useful for the derivation of clear-sky composites but

it is up to the enduser if she/he takes it into account knowing the significant uncertainty related to it. For more accurate cloud shadow detection technique Authors refers to the study of Simpson et. al (1998). To clarify this issue in the text the following sentence was added (page 8468, line 9): “Its main aim is to provide a rough approximation of the cloud shadow location which could be useful for the derivation of the clear-sky composites. However, if more accurate cloud shadow mask is required please refer to the study of Simpson and Stitt (1998).”

-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.

Response: The above comment was applicable to the earlier draft of the manuscript, however in the current version of the paper this issue has been removed.

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?

Response: This remark was implemented in the text.

-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).

Response: This remark was implemented in the text.

-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.

Response: This remark was implemented in the text. The following text was added:” Some approaches (Dybbroe et al., 2005a) utilise reflectance in the channel 0.6 μm with and without the SZA normalisation, claiming that the latter one is useful for cloud detection at extremely high SZA (> 86) when..”

-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.

Response: This remark was implemented in the text.

- 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.

Response: To clarify this issue the corresponding fragment of text has been changed to:
"The size of the LUT is a limiting factor therefore to reduce its dimensionality the Invariant Coordinate System (ICS) transformation is applied (Nordhausen et al., 2008; Tyler et al., 2009). It utilises the Principal Component Analyses (PCA) (Mardia and JM, 1979) and two scatter matrices in order to construct independent components which do not rely on a distribution mean. The first scatter matrix is a regular covariance matrix used to standarise data while the second one is a matrix of the fourth moment (kurtosis) which describes data rotation within the PCA. The eigenvalue–eigenvector decomposition is performed on one matrix in a relation to the other one which results in an affine invariant co-ordinate system for the multivariate observations. The matrices are derived on the basis of a randomly selected winter satellite scene with vast snow cover and utilised throughout the rest of transformations. To save computation time the ICS technique is performed only for the daytime data to combine reflectances with the thermal contrast between SKT and the 10.8 μm BT. It is applied selectively to pixels with probable cloud contamination (high thermal contrast) which fulfil specific criteria. These restriction are meant to improve ice cloud detection over snow and broken cloud discrimination where spectral information is ambiguous but thermal contrast with surface is significant. In this way the reflectances of areas with small thermal contrast, which may be related to climate model inaccuracy (and not to presence of clouds), remain unchanged. In this respect ICS transformation over water bodies ....."

- 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.

Response: Indeed utilisation of the term “threshold” in the paper can be misleading. Therefore, it was replaced by the terms: “binning value” or “interval”.

- 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).

Response: The atmospheric absorption is surely a major aspect that anybody dealing with remote sensing data has to account for. However, this could be achieved in variety of ways. The easiest way is to take the TCWV (Total Column Water Vapor) estimates originating from climate model and to relate them to the dynamic threshold values. More sophisticated methods (Minnis et al. 2008) integrate the WV estimates at different climate model levels to diminish the influence of high sensor viewing angles where due to elongated atmospheric path it cannot be any more approximated with the total column assumption. These approaches were proven to provide an enhanced results in terms of cloud mask accuracy, however they rely on climate model performance which adds another source of uncertainty to the final classification. This means that in terms of climatological studies the observed trend line in cloud amount may originate either from the real climatic signal or from artificial sources (amongst which the performance of utilized climate model is one of the factors). Regardless this well-know fact, contemporary cloud mask algorithms incorporates more and more modeled data which adds another sources of uncertainty, complicates the analysis and computational demand (which indeed in some cases leads to enhanced results). On the contrary the idea behind the PCM method is to make a better use of available spectral information before going for any ancillary data sets (in fact the climate data are an optional input). In this respect the Referee has stated:“Essentially, I think you have succeeded in doing this but there is one major exception: How do you account for atmospheric absorption effects?”. In case of atmospheric water vapor absorption we follow similar path as Saunders and Kriebel (1988) who parametrized its influence as a function of viewing angle and brightness temperatures relations. These variables are also present in the PCM LUTs which diminishes the influence of the WV on the PCM cloud detection accuracy to the point that it is in good agreement (on average) with the sophisticated PPS algorithm, where the TCWV estimates are compulsory input data. Nevertheless, the day-to-day differences between PCM and PPS cloud masks, applicability of the proposed method to tropical regions and the possible ingestion of the TCWV data are still to be analysed.
To conclude the acquired results confirmed that there is a great potential for simplification of the cloud

mask algorithms without substantial loss or even with small gain of the classification accuracy.

- 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).

Response: The following paragraph was added in the conclusion section:
"The proposed methodology features a unique set of merits such as:
- Derivation of classification probability between clear-sky/cloudy, clear-sky/snow and snow/cloudy conditions. This suppresses the usage of two different algorithms separately suited for cloud and snow detection. Moreover, such a triple-class probability gives more flexibility in terms of derivation of binary product where different probability thresholds (more or less conservative) can be applied.
- Utilisation of all available spectral and ancillary information in a single step to retrieve probability estimates from the multidimensional LUT. Such an approach resolves the problem of interpretation of divergent test results occurring in the decision-tree methods.
- Specification of multiple, irregularly distributed binning values which resolves the problem of selection of a single threshold value which should feature the highest discrimination skills for the instantaneous image acquisition conditions.
- Robust derivation of algorithm parametrisation based on a training dataset composed of collocated satellite measurements and clear-sky/snow/cloud classification originating from another algorithm (PPS in this study), another sensor (MOD10A1 in this study) or ground observations (e.g. SYNOP).
- Simple algorithm design which allows easy inclusion/exclusion of features by adding/removing dimensions in the LUT.
- Straightforward portability to other sensors which requires only the availability of spectral channels analogous to the AVHRR instrument and collocated training dataset."

- 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.

Response: After consideration of the Referee comment the term "supervised classification" was rephrased (removed) because the applied correction was more related to the PPS/MOD10A1 image editing performed by an analyst in the graphical software. Therefore, this was a subjective decision of the analyst who applied the correction (value recoding) only to cases which were obviously wrong in the training dataset. In the situations mentioned by the Referee where the spectral signature (or image texture) did not allow unambiguous discrimination between clear-sky/cloud/snow conditions the correction was not applied. More objective and systematic way to perform such a correction would be to enhance the PPS/MOD10A1 algorithms which is beyond the scope of this study.

- 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.

Response: The probability computation presented by the Referee is correct, however hardly ever this situation occurs where the clear-sky, snow, cloudy classes exist within the same bin of the multidimensional LUT. This is well visible in the Figure 2 which presents only the 2 dimensional space. When more dimensions are analyzed (like in the PCM) the separability between these 3 classes is even better and the Equation 3.2 reveals the true probability value. This simplification was made to reduce the size of the LUT which contains only one coded probability value per bin. Following the Referee suggestion to provide two values (cloud and snow probabilities) for every bin would double the size of the LUT which is the back-bone of the PCM method. Nevertheless, this suggestion is interesting and Authors will try to implement it in the future versions of the PCM algorithm.

- 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.

Response: The term "data quantity" was replaced by the "data frequency".

- 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.

Response: The Referee's explanation is only applicable to the PPS comparison while in case of the MODIS cloud mask (MOD35) this sharp feature exists as well within the same range. Moreover, the 7 K offset does not explain the significant drop of values within the 0 - 5 K range which is then followed by raise of values within the 5 - 10 K range. In the Authors opinion it is more plausible that this feature is related to the local radiance variation analysis than to the single threshold value within the PPS algorithm.

- 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.

Response: The day-time cloud frequency reported by the PPS over Africa (and also partially over Spain) reveals strange spatial patterns with values exceeding 90 % which surely does not correspond to the arid climate of these areas. On the contrary PCM statistics are more plausible and they are consistent with the study of Meerkotter et al. (2004) entitled: "A 14-year European Cloud Climatology from NOAA/AVHRR data in comparison to surface observations" (Figure 1). Regarding the cloud cover over central and south-eastern Spain indeed PPS to some extend misses and creates artificial clouds (which was visible during the training dataset correction by the analyst). Nevertheless, to confidently assess which cloud mask (PPS or PCM) is more correct over Iberian peninsula more extensive analysis will be performed.

- 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).

Response: Please refer to the next response.

- 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.

Response: Regarding the NOAA/CALIPSO collocations the most of doubts were already covered by the Referee in the extensive comments. Just for clarification in case of NOAA16 the data from 2011 were used when the orbital drift was already substantial. Regarding, the loss of received scenes compared to the theoretically possible ones this is more the issue of the HRPT receiving station which has to be within the satellite swath (at least in the case of our station) in order to record AVHRR data. Therefore, we do not receive all satellite scenes available for the European region which is most likely the reason why there was no collocations with NOAA17. In the manuscript the following text was added:

“Out of the three selected NOAA platforms only the one labelled with number 18 was used for the comparison. In case of NOAA16 in the year 2011 the orbital drift was substantial thus no collocations were possible with TERRA/AQUA. For the NOAA17 match-ups should occur for high latitudes, however those scenes were beyond the range of receiving station antenna located in Bern (Switzerland).”

- 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.

Response: In this paragraph the PPS and MODIS cloud masks were used for comparisons without implicit stating which one is the closest to the “truth”. Moreover, the fact that one method reports more clouds than another one does not imply that the first one is more accurate (in fact this can be quite opposite). To asses this the POD and FAR for both clear-sky and cloudy conditions are needed. In case of the PPS the misclassifications over water occur for warm, low stratus clouds during the twilight conditions (at least this issue was observed by the analyst).

Editorial remarks:

- Page 8450, line 7: Change “safe” to “save”.

Response: Remark implemented.

- Page 8453, section “Reflectance tests in the 1.6 & 3.7 μm bands”, line 20: Please explain NDSI (never defined previously).

Response: Remark implemented.

- 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).

Response: Remark implemented.

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.”

Response: Remark implemented.

Page 8445, line 10: Minor change of sentence to “Over barren or sparsely vegetated areas such as deserts, the strong diurnal surface temperature cycle.....”

Response: Remark implemented.

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”.

Response: Remark implemented.

Page 8457, line 19: replace “principals” with “principles”.

Response: Remark implemented.

Page 8470, line 7: You should write “verify the agreement” rather than “verify the difference”, right?

Response: Remark implemented.

Page 8472, line 21: Change to “Another problematic region is....”.

Response: Remark implemented.

Page 8474, line 6: Change PPS to PCM (Check! There is no results for PPS data in Figure 12!).

Response: Remark implemented.

Page 8476, line 26: Write “Their shape corresponds very well to other....”

Response: Remark implemented.

Page 8480, line 1: Change “the the” to just “the”.

Response: Remark implemented.