Atmos. Meas. Tech. Discuss., 7, C4995–C5005, 2015
www.atmos-meas-tech-discuss.net/7/C4995/2015/

Atmospheric Measurement Techniques Discussions

Interactive comment on "Bayesian cloud detection for MERIS, AATSR, and their combination" *by* A. Hollstein et al.

A. Hollstein et al.

andre.hollstein@fu-berlin.de

Received and published: 27 February 2015

Dear Andi Walther ,

we wish to express our thanks for you work in reviewing this paper. Your comments are very much appreciated and certainly allowed us to improve the quality of the paper.

Please find specific replies to your comments in the text below

With kind regards, A.Hollstein et al.

Referee Comment: >The introduced approach is highly statistically and is based on artificial truth data. While this paper >shows interesting methods, the approach includes many risks. Improvements could include a more >physical-based selection of features, building separate Bayesian classifiers for individual surface types >to account for highly variable background impact.

Authors Reply: Yes indeed, our approach statistically. The paper deals mainly with the extension of Bayesian methods applied to cloud detection for MERIS and AATSR measurements. It is beyond the scope of the paper to fully explore this field as this is subject to ongoing research. We have introduced a clear distinction between dependent and independent schemes which use no external information (dependent) and those using external information (strongly/weakly dependent). We then developed methods and explored some areas of strongly independent methods. One point of outlining the notations was to clearly state the full range of possible approaches, while this paper deals only with a small subset of what is possible.

Methods using external data, e.g. for surface types from land use maps, would belong to a class of algorithms which were intentionally not discussed in this paper. The used truth data (some manual classified scenes, Synergy cloud mask) is clearly not sufficient to test the effectiveness of these type of algorithms. However, this is part of ongoing research within the cloud cci project and might be subject to future publications.

Referee Comment: >Specific comments: >Introduction: >Please mention whether this approach is daytime only!

Authors Reply: This information was missing and is now included in the introduction.

Referee Comment: Line30: What is the Synergy product? Please refer here to chapter 6. I would also recommend explaining chapter 6 earlier in a section after introducing.

Authors Reply: The Synergy product is explained in the following sentence: "where AATSR measurements are mapped on the MERIS swath and their mutual overlap is used.". Unfortunately, there is no official paper describing this data format. For this paper it should not be necessary to describe the co-location (based on the BEAM toolbox) process in more detail. The Synergy Cloud mask, which is shortly described in section 6 is independent from the synergy data set and uses synergy data merely as input.

Referee Comment: Section2 Bayesian inference for cloud masking: You state that by using a background probability of 0.5 you avoid circular arguments for building climatological time series because the result will otherwise eventually shift to the climatological value. However, this is also the case by using this value of 0.5. The cloudiness will then shift to 50% cloudiness instead to a climatological a-priori value.

Authors Reply: The effect of the choice of P(C), especially on level 3 data (monthly means,...) should be investigated and communicated to potential end users. We updated the discussion of the choice of P(C) accordingly. Investigating the effect of more general background probabilities (e.g. time or position dependent) is beyond the scope of this paper, but is surely interesting to look into.

Referee Comment: Section 3: There are good reasons to use external data for cloud masking. Firstly, cloud masks are usually based on contrast between measured property and an assumed clear-sky value. Estimating the assumed clear-sky value requires auxiliary data, such as surface reflectivity, surface temperature or several atmospheric profiles.

Authors Reply: We do not indicate that external data should not be used. We outline the possible range of methods (classical/naive vs. strongly/weakly and independent/independent) and then use the class of independent Bayesian schemes to demonstrate our approach. With the available truth data (some manually classified scenes, Synergy cloud mask), it was not feasible to go beyond this scheme at this point of research. Only with the appropriate set of truth data, it is possible the evaluate the added value of dependent schemes (in our case for MERIS and AATSR). We updated our discussion in the conclusions to make this point more clear.

Referee Comment: Secondly, the underlying surface have an impact on the measured signal itself due to simple radiative transfer considerations. Thin cloud signals include

high amount of surface and atmosphere impact. The same cloud will lead to different results over different surface types. You may see this not only in the global pattern of skill score, but also in the global maps of cloudiness itself.

Authors Reply: This is one of the main the reasons why the set of truth data should be large and representative, e.g. to cover similar clouds under various surface and illumination conditions. As discussed in the paper, the large volume of available truth data from the synergy cloud mask was chosen such that the data is distributed representative in space and time. We made the representative selection of data more clear in the first paragraph of section 7.1.

Referee Comment: For most cases it is not possible to closure with a sufficient accuracy from a reflectance value in a visible channel to a probability of cloudiness of a pixel. The location may be over a bright desert, snow or a dark ocean. The reflectance is also highly sensitive to viewing geometry, which can be very different.

Authors Reply: When using MERIS and AATSR, we can use visible and thermal channels and are thus not limited to reflectance values alone. The effect of viewing geometry is implicitly included when the artificial truth data sample is prepared.

Referee Comment: You may calculate a cloud probability from your truth data set for each feature set, but this likely tells you more about the regional and geometrical distribution of these truth data.

Authors Reply: This is the reason why we chose to select the truth data such that it is evenly distributed in space and time. If only data from desert regions were included, and applied globally, one could see the effects you mention. This is not what we did or described in the paper.

Referee Comment: With your approach of strongly independent feature, you may also come into trouble particular for climate purposes. To give a simple example: Assuming we have in reality no trends in cloudiness over a decade. Also assuming that one major feature is reflectance in a visible channel, which is true in reality. You will have being built a strongly independent background joint probability, which among others separates cloud and cloud free according the reflectance in this channel (the brighter the more likely a cloud). If the surface type, and thus also the surface reflectivity changes, the visible reflectance will also change, and thus you will "detect“ an artificial trend in cloud cover. This trend will be stronger for thin clouds, because surface impact is much higher. Examples for surface changes are urbanization and the increase of forest areas, this is not negligible.

Authors Reply: This question is related to the stability of our algorithm, or any algorithm in general. Also you iterate on the point of not using external information. First, the long term stability of such algorithms, and any other algorithm, must be monitored. A Bayesian scheme is no exception here, but such is any other algorithm. This problem of monitoring and long term stability is a very different subject and beyond the scope of this paper. When generating level 3 data from the FAME-C products, this question must be thoroughly analyzed.

Referee Comment: This question is related to the "accuracy vs. stability“ dilemma for generating climate data sets. Task is to find the right balance. It is of course not well applicable to use highest accurate surface value changing every week with varying accuracy. However, ignoring surface impact at all lowers accuracy heavily.

Authors Reply: We do not propose to not use external information nor do we ignore the impact of the surface (please the answers to your points above). There are very good reasons to use and not use external information and we decided, for the available artificial truth data, it is the best choice to explore the application of Bayesian classification using the independent approach.

Referee Comment: Please, add a description, which measured property you use for each channel (reflection, radiance or brightness temperature).

Authors Reply: We use the data, as it is from the synergy product. Stored are integer

values with given offset to convert to reflectance, radiance, or brightness temperate when applicable. The unit, and scale of the feature is of no importance to the estimation of the background probability. For this reason, we left a description of the units out on purpose. Since channels can be freely combined to features, possible units can easily become unusual, e.g. Temperature-Reflectance.

Referee Comment: Please, explain how you build the pseudo-channel features. (Example: How is 12um x0.55 um defined?). I can identify 40 bins for a feature from Figures 3 and 4. How do you define the range? Is there any stretching (Gamma stretching etc..) . This could be important for the "x“ pseudo-channels. Can you explain how a channel 442 um x 412 um can provide information about cloudiness? (see Table 2)

Authors Reply: We explain in the text (second paragraph of Section 4, “Construction of Feature Sets”) that we consider all basic arithmetic combinations (plus,minus,times,division) and the index function dx(a,b)=(a-b)/(a+b) to construct features. The table entries simply state which combination was used. The histograms were computed using the standard routines from numpy (python, histogrammdd) with no applied pre-processing. All performed data proceeding steps are discussed in the paper. The “x” denoted the symbol for times. The shown channels in Table 1 to 3 were selected by their success in reproducing the results of the synergy cloud mask, as we discuss in section 7.1 “Reproduction of Existing Algorithms” which indicates that the connection to cloudiness is implicit. However, this artificial channels combines two channels which information about cloud and surface reflectivity which can be used for classification. One should note that for the classical and naïve Bayessian approach, always the combination of all channels is used for classification. This means that the contribution of each feature (channel) must not be obvious and interacts with all other features. With this statistical approach, we do not intend, and don’t need to, explain the specific contribution of each feature. However, if one would want to explain the effect of a single feature, one can simply look at skill score maps for specif regions or globally for a classifier with and without this particular feature. This explains the contribution of this feature to the reduced set of features. Using a classifier with only one feature is likely not sufficient. We updated the discussion of Table 1 to three to highlight this more clearly.

Referee Comment: The random search of feature incorporates risks, which may come from unwanted correlation. To give an example from a different field: You may find a high correlation of measured radiance in a window channel to water vapor, even there is no direct impact of water vapor to the signal. Reason is that Sea surface temperature is correlated to atmospheric water vapor. Correlation is high, but the retrieval will fail if dry air is advected over Warm Ocean.

Authors Reply: This again is one of the reasons why we chose to make sure that the truth data is sufficiently distributed. In addition, for the tests which are discussed in section 7.1, data from the year 2007 was used to estimate the background probabilities while data from 2008 was used to compute the skill scores. This approach helps to ensure that effects which you mention do not occur. We made this more clear in the discussion of section 7.1.

Referee Comment: Please discuss the risks of such a non-physical (you say non-educational or statistical) approach.

Authors Reply: Please see also our comment to your previous point. We clearly state in the paper that we use the best results which were found by the search. The risks which you mention are minimized by the approach of separating input and test data. We assume that this approach is sufficient. In addition, we don't propose that the random search is the only way of finding good sets of features. If one finds a much better solution, or strongly hesitates to use a particular set of features, one is free to chose another. If one accepts the proposed metric (optimizing the skill score, separate input and test data) to measure to quality of a certain feature set, then the random search imposes no additional risks. The only risk it to stop the search premature, but this is discussed in the text.

Referee Comment: Line 245: “The experienced expert is not surprised..”: Many pseudo channels look really surprising to me. . . (But I first need clear description how they built..)

Authors Reply: Please refer to the second paragraph of section 4 for a description of how feature sets are build. Here we state that you will find traditionally used channels (11,12 microns) in the found sets of features.

Referee Comment: To Figures 3 and 4: I am wondering if really all areas of the 2d histograms were filled with data before smoothing. If not, why are the areas slightly reddish and not white (or masked out) which would mirror identical probability of cloudy and cloud-free?

Authors Reply: The original, sparsely filled histograms are not shown in the paper. With 40*40 bins one have individual 1600 bins to fill with only 1000 data samples. This means that not all bins were nonzero before smoothing. One interesting fact from both figures is that even the broad, slightly redish, feature from the left panel can be reproduced with this little data. But please note, as stated in the text, we optimized the smoothing parameter such that the representation of the original histograms is best.

Referee Comment: Why did the areas around [35,5] decrease after Gaussian smoothing?

Authors Reply: This is clearly one of the properties of the original data which was adequately not captured by the small sub sample used. This is one of the reasons why the skill score for these smoothed cases is in general smaller. Please see Figure 8 and 9 for a clear analysis of this effect.

Referee Comment: Here, I'd wished to interpret the results, but you didn't give an explanation how you build dx(442nm,12um). (see section 4 comment)

Authors Reply: Please refer to the second paragraph of section 4 for a description of how feature sets are build. dx(a,b) is defined as index function with: dx(a,b)=(a-b)/(a+b).

Referee Comment: The examples in the images separate after smoothing very well both features in cloudy and non-cloudy regimes. Thus, also a naive Bayesian approach would lead to similar results with less preparation and computational effort. Could you show a different example to illustrate the advantage of the joint (classical) approach?!

Authors Reply: The point of Figures 3 and 4 is to show how with initially insufficient amount of data on can still proceed to use the classical Bayesian approach without assuming, or carefully constructing, that the used features are indeed independent. This is not a point against using the naive approach. The new method gives potential users an additional tool for construction classifiers to find the optimum choice for the problem. The shown separation in the Figures does not imply that the naïve approach would lead to similar results and there is no relation of this to the applied smoothing. This nice separation might only imply that this problem can also be solved using support vector machines (SVMs) or other classification algorithms. When comparing the results from Table 4 and 5, one can see that the classical approach (for this application) gave better results even with fewer features. This result will of course depend on the available channels and the desired application.

Referee Comment: The right images only consist on 1000 measurements. How did you pick the samples out of billions of pixels from many different scenes, surface types, seasons, cloud types, etc..? Is this a sufficient number to build a cloud mask for all the different types?

Authors Reply: The cases were randomly selected from the data which was used for the left panel. This information was missing and we updated the text. As discussed in the text, these figures serve as extreme cases, which nevertheless show how well this approach operates. We make this more clear in the text (4th paragraph of section 5). Figure 8 and 9 clearly show how the global skill score depends on the amount of data used. Weather this is sufficient depends clearly on the application.

Referee Comment: This section should be in the beginning of this paper because it is needed for under-standing of some points in the earlier sections. I would also recommend extending the explanation of the Synergy cloud mask, because the citation does not seem to be a peer-review paper.

Authors Reply: Before submitting this paper, we discussed the placement of this section within the paper. There are good points for placing this material earlier and also for its actual placement. We prefer to leave it where is if for several reasons. There we describe a source of truth data which is used in the next section, so it makes sense to have this material close to where it is needed. Earlier in the paper we discuss general approaches and techniques and later some possible applications. We don't benefits from having this description of this particular data source in the general part of the paper. This cloud mask was used within Cloud CCI and FAME-C and the referred paper it the best source up to date for this algorithms. Since we have no common co author for these paper, it might be safer to refer to the paper than trying to summarize it here. We also extended its description in Section 6

Referee Comment: Please also mention the processing time difference between Synergy cloud mask and the Bayesian approach to defend the need of a faster retrieval.

Authors Reply: Our implementation of the described algorithms are based on python, multiprocessing, separation of computing and input and output (I/O), and efficient usage of numpy. The processing time is largely dominated by I/O time and in comparison the computing time is almost negligible. We updated the discussion to include these information's.

Referee Comment: Please explain which of the cloud masks do you intent to select for CCI?

Authors Reply: The selection of a final set of features for Cloud CCI will be determined in the near future. We updated the text to highlight this fact.

Referee Comment: Line 419: " ... can be used to reproduce...cloud mask ..“: The skill score for all examples seems to be low in comparison to other cloud masks. Figure 6 shows HSS of less of 0.6 for large parts of Asia and North America. This means an approx. POD of about 75% this is much too low for a cloud mask. This is even more the case if one considers that the skill score here is computed for a comparison of results from the same retrieval as the truth data. Please discuss this! Please provide also POD and the other measures for a better interpretation.

Authors Reply: As stated above, data from 2007 were used to estimate the background probabilities and data from 2008 is used to evaluate the algorithm. The skill score is relative to the synergy cloud mask such that comparison with actual cloudiness is implicit. As we discuss in the paper, the results clearly motivate to produce a better set of truth data for MERIS and AATSR. This might only be possible from manual classification since no active instrument such as a LIDAR is available for the full data range.

Referee Comment: If you have a feature set what has no skill to distinguish between dust and clouds, when you can correct the Bayesian coefficients as often as you want. You cannot solve this problem statistically.

Authors Reply: This is trivially true. The point behind the discussion is to have truth data with adequate representativeness to realistically evaluative a set of features. Assuming a particular feature set would be sufficient, but the truth data doesn’t clearly separate clouds and dust. Then, the feature can not be used to distinguish dust and clouds, but the cause is the initial data set and not the feature set. This is the case which we clearly discuss in the paper.

Interactive comment on Atmos. Meas. Tech. Discuss., 7, 11045, 2014.