Responses to Reviewers

This document includes our responses to the reviewers' comments and suggestions for the manuscript [doi:10.5194/amt-2019-409]: "A Machine Learning-Based Cloud Detection and Thermodynamic Phase Classification Algorithm using Passive Spectral Observations".

We thank all the reviewers for their helpful suggestions and comments. We hope the revisions are found responsive and appropriate, and that the revised manuscript will be deemed acceptable for publication in the *Atmospheric Measurement Techniques*.

Our responses to the general comments and suggestions from the reviewers (Reviewer #1: Blue; Reviewer #2: Green; Reviewer #3: Orange) are listed below (response in black):

General Responses:

R1: The authors describe a machine learning (ML) based approach to first detect clouds and second to assign cloud thermodynamic phase (liquid versus ice). The ML algorithm is trained using CALIOP detected liquid and ice clouds but is limited to the most straightforward single phase and single layer cloud configurations (or multilayer with the same phase), thus mixed phase and multi-layered clouds of different phases are not included in this study. The approach is tested against existing MODIS Collection 6 (C6) and MODIS/VIIRS continuity products (both the cloud mask and cloud phase). The ML approach is shown to improve the phase characterization over the existing MODIS and MODIS/VIIRS continuity algorithms, with greater improvements over certain surface types including snow and ice. Cloud characterization efforts from satellite remote sensing platforms are increasingly utilizing ML algorithms and this paper is timely and a useful exploration of the potential of ML for passive cloud imagery characterization. Parts of the methodology are not as well detailed as need be and the results need to be placed into a broader context. After addressing the comments below and suggestions for straightforward revisions, this paper would be a nice addition to the literature.

Response: We appreciate the insightful comments from the first reviewer (R1). We also noticed that some details, in particular the training/validating dataset selection and model configurations are not well described in the original version. Therefore, in the revised version, we provided more details of the method and results. Please check the new Tables 2-5, and corresponding responses to R1.6, R1.8, and R2.18.

R2: This paper applies a machine learning (ML) approach to the problem of cloud detection and thermodynamic phase assignment from passive satellite measurements. This is potentially significant considering the challenges noted in the manuscript with the traditional methods currently being employed and the rapidly increasing interest in using ML for satellite analyses of clouds. The ML approach evaluates a number of models that are tested and evaluated using various combinations of passive sensor radiances and ancillary data products as inputs while CALIOP data are used to define the reference labels for cloud occurrence and phase. Two models are selected for evaluation, one that employs solar and infrared radiances (daytime) and one that employs only infrared radiances (all day). The view angle, latitude, longitude and the surface skin temperature were found to be the most important ancillary data needed. In addition, the models are trained for 7 surface types. The two models are found to perform reasonably well and performance metrics generally exceed the current approaches employed on MODIS and VIIRS by the MODIS Science Team. However, the significance of the results are difficult to gauge for a variety of reasons. For example, the ML and current (referred to as traditional in the manuscript) approaches are designed much differently with regards to the targeted clouds, atmospheric correction, scene type dependencies and other factors. With respect to the clouds, the ML model development excludes the most difficult clouds which are pervasive over the Earth. In particular, clouds in polluted environments, broken clouds and single-layer and multi-layer ice overlapping water clouds are screened out of the training and validation dataset. The rationale for taking this approach is not well described. Part of the evaluation of the ML method against current methods with respect to CALIPSO (figs 6-9) could perhaps be considered an apples to apples comparison in that the same pixels are being evaluated. But, considering that the ML approach was developed using a particular subset of (screened) data while the current approaches were designed for application over a much wider range of conditions is possibly unfair, and the comparison are potentially misleading. I wish the authors had taken a more globally applicable ML approach to the problem. It seems to me that at best the results suggest that ML methods can perhaps perform at least as well as the current non-ML methods and that these can be developed for application to other satellites much easier (and cheaper). Despite all of these issues, the study is a reasonable initial step, the results are clearly presented and the manuscript is grammatically clean. Therefore, I find that the manuscript could be published after some revision. In particular, I recommend that the authors clarify the rationale for the approach, clarify the significance of the results, and temper suggestions regarding the potential for ML to improve the accuracies of global cloud analyses since in my view this is not adequately demonstrated here given the heavily restricted dataset that is used.

Response: We appreciate the insightful comments from the second reviewer (R2). We agree with the major concern from R2 that the current training/validation results could be problematic or cannot represent global clouds considering a large fraction of “mixed phase”, “inhomogeneous”, or “aerosol contaminated” clouds are excluded. To address this concern and other related questions and comments, we made necessary modifications and gave more explanations in the revised manuscript and response. Please find our detailed responses below, in particular responses to R1.6, R1.8, R1.11, R1.12, R2.4, R2.17, R2.18, R2.23, and R2.27.

R3: Comments from Dr. Luca Bugliaro.

Dear Authors,

I appreciate your work very much and think that this is a valuable contribution to remote sensing. Nevertheless, I think that you also should mention two of our papers in this same journal since they use similar methods of machine learning to perform cloud detection and cloud property derivation. In particular, they also use measurements of the CALIOP lidar as a reference and collocate them with passive observations.

Could also check references therein for papers with similar topics.
Strandgren, J., Bugliaro, L., Sehnke, F., and Schröder, L.: Cirrus cloud retrieval with

MSG/SEVIRI using artificial neural networks, Atmos. Meas. Tech., 10, 3547–3573, https://doi.org/10.5194/amt-10-3547-2017, 2017.
Kox, S., Bugliaro, L., and Ostler, A.: Retrieval of cirrus cloud optical thickness and top altitude from geostationary remote sensing, Atmos. Meas. Tech., 7, 3233–3246, https://doi.org/10.5194/amt-7-3233-2014, 2014.

Response: The two papers match the topic perfectly and should be included in the reference list. We appreciate the comments and suggestions from Dr. Luca Bugliaro.

Detailed Responses

R1.1: Abstract: I found it to be a bit too detailed and meandering. Would suggest tightening it up and focusing on the main points rather than the details.

Response: Done. We removed some details about the accuracy rates for the two RF models in cloud mask and phase detections.

R1.2: Line 59: ‘having radiometric stability issues’ is colloquial and not specific enough to be useful.

Response: Done. We replaced “radiometric stability issues” with “calibration drifts”.

R1.3: Lines 79-80: There are two issues here that need to be raised and appear elsewhere. First issue, is this even true? There are many Bayesian methods in the literature that assign uncertainties as a part of the retrieval methodology. Furthermore, using the look up table methodology of MODIS C6, the reported uncertainties for the optical properties appear to be quite useful and rooted in physics. I don’t know about uncertainties regarding phase so this could be a different issue. For the cloud mask, the raw Q values are quite useful for an estimate of cloud detection uncertainty. Second issue, calling one set of algorithms ‘traditional’ is confusing at best. Machine Learning (ML) research dates back to the 1950s and outdates many satellite retrieval algorithm approaches that are currently used. Wording along the lines of “in contrast to most operational and research methods,” and similar changes elsewhere, will help make your points clearer. Then you could stick to “ML” as a separate algorithm branch.

Response: The reviewer is quite correct that quantitative uncertainty datasets now accompany the retrieval of continuous variables, e.g., MODIS cloud optical properties. And as the reviewer points out, the MODIS CLDMSK cloud detection algorithm reports a continuous “clear sky confidence” or “Q value”, ranging from 0 to 1, for each pixel. Therefore, we decided to remove this statement. We have also made additional modifications to the rest of the manuscript. For the second suggestion, we agree with the reviewer. “Traditional” could lead to unnecessary confusion. Therefore, we changed the word “traditional” to “hand-tuned” throughout the manuscript.

R1.4: Lines 138-139: Now the random forest (RF) model is mentioned and apparently it has a proven record, yet is "not traditional"? The "author classification" of algorithms needs to be reworked throughout the paper.

Response: We have changed the word "traditional" to "hand-tuned". See our previous response.

R1.5: Lines 233-234: Need to be more explicit as to what "the fix" was to the thermodynamic phase algorithm.

Response: We found that our initial phase algorithm implemented in CLDPROP Version 1.0, which is based on the MOD06 Collection 6/6.1 optical property phase algorithm with some modification, omitted a key cold cloud sanity check that led to spurious liquid cloud decisions at the edge of ice clouds. This in turn caused spuriously large liquid cloud fractions and a discontinuity in ice cloud effective radius retrieval statistics. We subsequently implemented a new cold cloud sanity check and reprocessed CLDPROP to Version 1.1. More details about this fix and its impacts can be found in the Product Version 1.1 Change Summary section (Section 1.4) of the CLDPROP User's Guide (https://atmosphere-imager.gsfc.nasa.gov/sites/default/files/ModAtmo/EOSSNPPCloudOpticalPropertyContinuityProductUserGuidev11.pdf) available on the Atmosphere Discipline Team website (https://atmosphere-imager.gsfc.nasa.gov/). However, following the second reviewer's comment, we believe this detail is irrelevant to this paper and have decided to remove this statement from Lines 233-234.

R1.6: Section 4.2: This is a long paragraph with a lot of information and should be make clearer than currently written. First recommendation: it would be very helpful to list the pixel count and relative percentages of each of the cloud/no cloud, aerosol/no aerosol, phase and cloud configuration categories that are kept for ML training or are discarded, and should be denoted clearly. It took me a while to figure out that some multilayer clouds are included but only for the same phases. How many multilayer clouds of the same phase occur relative to multiple phases? Second recommendation: say more clearly up front what is in the ML training rather than what is tossed out. Then follow with detail of what is tossed out. It is really hard to keep track of what goes into the sausage. First additional comment: why not attempt to address the ambiguous/mixed phase categories? Some advances in detection and characterization could be made with these types using ML. Do you have plans to do this in follow-on work? Second additional comment: why only use clouds with at least five consecutive labels that are the same? Doesn't this limit the number of cases greatly? Also doesn't this bias the ML training to larger-scale cloud behavior even though the classification is (presumably) done on a pixel-by-pixel basis? Small scale clouds might behave differently (with respect to phase sensitivity) than large scale clouds. Third additional comment: why are cloud structures in the ITCZ any more "complicated" than other geographical regions? What makes a cloud structure "complex"?

Response: We appreciate the very insightful comments and suggestions. Accordingly, we made necessary modifications in Section 4.2 as listed below:

- First, we added a new table (Table 2) that gives more details about the sample. In this table, it is clear how we select highly reliable datasets by using CALIOP L2 products. For all surface types, approximately 39.3% of all collocated VIIRS 750m pixels are selected for training and testing, while 1/3 of all VIIRS pixels are excluded because of aerosol contamination (e.g., column 532nm AOD > 0.05).

- Second, we reorganized the paragraph by mentioning that only aerosol-free, homogenous clear, and homogenous single-phase cloudy pixels are included in the training/validation datasets. Also, we give clear definitions of "*aerosol-free*", "*homogenous*", and "*single-phase cloud*" in the text and in Table 2.

  We should note that the performance of ML models is strongly dependent on the quality of the training dataset. In this study, the two RF models are trained and tested with simple yet highly confident samples collected from 2013 to 2016, with the expectation that the RF models will capture the key spectral features from these simple samples more efficiently. Of course, it is then not surprising that the two models perform well when comparing with CALIOP using similar simple samples from 2017. However, we note that many current operational/research-level phase algorithms, including the MYD06 and CLDPROP optical property phase (OP-Phase) algorithms considered in this study, were also tuned (often by hand) with CALIOP using data filtering strategies similar to those employed here (see, e.g., *Baum et al.*, 2012; *Marchant et al.*, 2016). The better performance of the RF models compared with the operational algorithms, even if only for these simple cases, highlights the advanced capabilities of ML approaches over human tuning to more efficiently identify and effectively utilize spectral information content.

  That said, the reviewer raises an important point regarding more complicated cloud scenes. For example, we expect that the RF models may recognize signals from both ice and liquid clouds in overlapping cases when the upper layer cloud is not optically thick in the relevant spectral channels. Of course, this is also the case for current operational phase algorithms (e.g., MYD06, CLDPROP) for which tuning/testing also did not include complicated cloud scenes. Nevertheless, we expect that the classification probabilities that are the output of the RF models can provide important information. For instance, we find that, for simple cases (i.e., homogeneous clear or single-phase cloudy), the probability distributions from the RF all-day model have strong peaks (see Figure 10 a, b, and c in the revised manuscript) close to either 0 or 1. However, for more complicated cases, such as ice over liquid cloud (panel d), the liquid and ice probabilities are more broadly distributed, indicating that the RF all-day model may recognize signals from both liquid and ice and therefore provides ambiguous results. Ambiguous liquid/ice probabilities could be used to define a third, "unknown" phase category, following MYD06 and CLDPROP convention, and also provide a useful quality assurance metric for the downstream cloud optical property retrievals. We also would like to point the reviewer to a manuscript that is relevant to the discussion here: *Marchant et al*. (2020), currently in review, gives a more detailed discussion on MYD06 multilayer cloud detection and the impact on phase detection. We have added this discussion in Section 4.4 and Section 5.

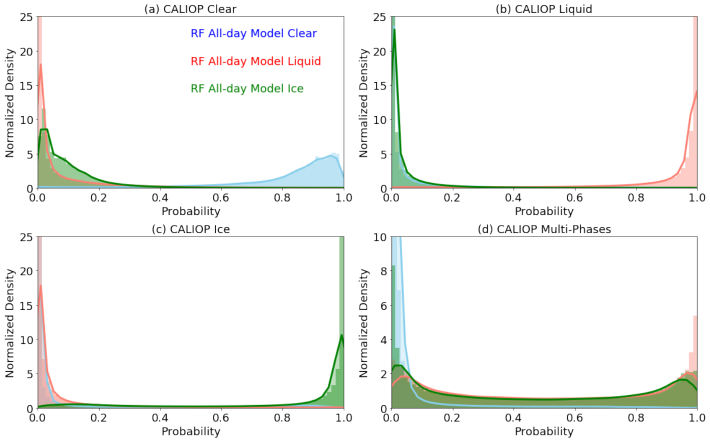


Figure: Clear, liquid, and ice probability distribution functions of the RF all-day model for four lidar pixel categories: (a) CALIOP clear, (b) CALIOP liquid water cloud, (c) CALIOP ice cloud, and (d) CALIOP multiple phases. The multiple phase pixels (d) are not used in model training/validation.

New Reference added:

Marchant, B., Platnick, S., Meyer, K., and Wind, G.: Evaluation of the Aqua MODIS Collection 6.1 multilayer cloud detection algorithm through comparisons with CloudSat CPR and CALIPSO CALIOP products, Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2019-448, in review, 2020.

- Finally, we mentioned that for some regions, such as the ITCZ, the sample selection rates are low because of the complicated cloud structures. For example, clouds always have very complicated vertical structures (such as multiple layers with difference thermodynamic phases) and strong horizontal heterogeneity due to convection. We modified our previous statement for clarity.

R1.7: Lines 346-353: Is this a description of other experiments tried that are not shown in figures or tables? Or is this paragraph part of the methodology?

Response: For the daytime model, we also tried different input combinations. Another table (Table 4) with all of the details are included in the revised version.

R1.8: Lines 401-405: It would be really helpful to report what total percentage of all pixels considered these represent. The crux of the matter: does ML greatly help for a large percentage of cloudy pixels, or does it help for a small percentage of cloudy pixels? Also, in figures 6-9 showing the true versus false positive rates, it would greatly enhance the presentation of the results by including percentages for each subpanel of the total number of pixels considered.

Response: We agree. In the cloud mask and cloud thermodynamic phase TPR-FPR plots (Figs. 6-9), we have added the total number of pixels for the corresponding surface types. Moreover, we have added the following text and a new table (Table 5) to Section 4.5.2 to demonstrate the importance of "unknown phase" category for each cloud phase product:

"*It is also important to note that the number of pixels used for cloud phase TPR-FPR comparisons in Figures 8 and 9 are different for products that have "unknown phase" categories, namely, MYD06 IR-Phase, MYD06 OP-Phase, and CLDPROP OP-Phase. As shown in Table 5, the MYD06 IR-Phase has a relatively large "unknown phase" phase fraction (15% for all surface* *types and 34% for snow/ice) in comparison to the OP-Phase products from both MYD06 and CLDPROP, which have 2~3% "unknown phase" fraction approximately*".

R1.9: Lines 418-423: There is a disconnect between this discussion and the earlier discussion on lines 308-310. How are inhomogeneous clouds being considered when earlier the authors state that they are "discarded"? These may be different issues but it is worth making clearer how inhomogeneous clouds are (or are not) considered and dealt with in this study.

Response: This comment is related to R1.6. Please see our response above.

R1.10: Lines 456-457: "A few hours" doesn't really mean anything scientifically. And without describing what is calculated and on what kind of computing platform, this also doesn't convey any information.

Response: Please see our response to R1.11.

R1.11: Lines 457-459: While not written directly in this way, reading between the lines written by the authors, one could deduce that ML approaches could render instrument calibration efforts and algorithm continuity efforts pointless and irrelevant. Will ML have the potential to address discontinuous satellite observational records by a thorough and accurate labeling of training data for a ML algorithm? I don't think this is what you intended to say, but it does raise the point – can ML methods be used in lieu of a properly calibrated and characterized satellite instrument? Same point applies to lines 467-468.

Response: For the first question, we believe that instrument calibration efforts and algorithm continuity efforts are very important. Instead, our main point is that ML approaches have the potential to streamline algorithm tuning and/or threshold selection processes that often occur in response to instrument calibration changes or when porting to other instruments. With non-ML methods, such tuning and/or threshold selection processes need to be done manually, which is a time-consuming effort. We have modified the text in response to the reviewer's comments.

"*With hand-tuned methods, adjustment is always required in the case of calibration changes, algorithm porting to another similar instrument, or changes in solar/viewing geometries and surface conditions. Manual adjustments can be time-consuming (e.g., months or years), whereas the two RF models used in this study were trained and tested for 7 surface types and using different input variables in 3 hours (on an HPC Platform using 32 Intel Xeon Gold 6126 Processors @ 2.60 GHz). More important, manual algorithm adjustment may not provide the best continuity between two instruments. For example, although the MODIS CLDPROP OP-Phase and VIIRS CLDPROP OP-Phase are designed for climate record continuity purpose, cloud thermodynamic phases from the two products are different by up to 4% for all surface pixels, and by up to 10% over surfaces covered by snow/ice (see Figure 8 light blue and light green dots). Further investigation is necessary to understand if, using ML approaches, a better climate record continuity will be achieved with a uniform training dataset.*"

For the reviewer's second question, it is likely true that a properly trained ML algorithm can still achieve a high level of skill in the presence of calibration errors if (a) calibration errors are relatively small and spectrally/spatially uncorrelated in such a way that physically-relevant signals are not masked by the errors/correlations, and (b) the instrument is radiometrically stable or radiometric changes are monitored/corrected on orbit (which gets back to our main point above). Confirmation of both assumptions requires a dedicated and robust on orbit instrument characterization effort.

R1.12: Lines 470-478: Regarding the use of CALIOP for labeling, one could make the argument that CALIOP is a distinctly different observation and should in fact see something different than a VIS/SWIR observation (e.g., MODIS and VIIRS). Doesn't CALIOP labeling essentially "force" MODIS and VIIRS to observe like a lidar even though they do not contain the same physical sensitivity to clouds as the lidar? Will differences in instrument sensitivity (e.g., CALIOP vs. VIIRS) to a given cloud ultimately lead to poorer performing ML algorithms because one is made to "look like" the other? It is an interesting question to consider. For some clouds, the lidar and passive spectrometer could provide a lot of valuable complementary information, and that is basically "thrown out" in a ML algorithm when one is forced to behave like the other.

Response: We agree with the reviewer's comment regarding different sensitivities between MODIS/VIIRS and CALIOP. This in fact is the reason why we only train the models with simple, single-phase samples for which we expect agreement between the passive and active sensors. This allows the models to learn the spectral signatures of liquid and ice clouds separately. For more complicated cases, i.e., horizontally/vertically heterogeneous and/or multilayer pixels, we then let the models make their own decisions regarding what phase makes the most radiative sense given the observations. Further discussion can be found in our response to R1.6.

R1.13: Lines 489-490: not sure what is meant by "screening process"

Response: We modified our statement to "*to check if the training dataset collection process introduces*".

R1.14: Lines 518-519: why is it more impractical to consider aerosol and cloud together?

Response: Adding complexity to the RF (or other ML) model requires more overhead, such as memory at run-time, computational resources, etc. It could be a potential (but not critical) problem when implementing in an operational algorithm production environment, where there often are limitations on such resources (e.g., caps on memory usage). That said, we decided to remove this statement because there are ways to mitigate these technical issues given sufficient resources.

R2.1: Line 23: Strongly suggest something like this: "It is shown using a conservative screening process that excludes the most challenging cloudy pixels for passive remote sensing…

Response: Done.

R2.2: Line 35: 'will' need further attention

Response: Corrected.

R2.3: Line 62: Zhou reference may need updating

Response: We removed this reference since this paper is not submitted.

R2.4: Line 79-80: This statement is too vague and possibly misleading. How is the uncertainty assessment more difficult for a cloud classification derived with the traditional methods vs the ML approach? It is true that in a Bayesian context, uncertainties in satellite retrievals associated with inversion are easy to extract, but these do not include uncertainties w.r.t ground truth data due to simplifying assumptions in the forward models and a host of other factors. Please elaborate to clarify and support your contention.

Response: We agree with the reviewer's point. Quantitative uncertainties are available for Bayesian methods, and are frequently used in retrievals of continuous variables, e.g., cloud-top height, cloud optical thickness, etc. Furthermore, in the MODIS CLDMSK cloud detection algorithm, a continuous "clear sky confidence" or "Q value", ranging from 0 to 1, is provided for each pixel. Therefore, we decided to remove this statement. Please also see our response to comment: R1.3.

R2.5: Line 195: should be Sayer et al 2017?

Response: Corrected.

R2.6: Line 221-223: not clear what you mean here.

Response: Thanks for pointing it out. We removed this statement from this paragraph.

R2.7: Line 231-234. Not sure what the relevance of this update is to the paper unless you used the older version. If this is the case, then you'll need to elaborate on the impact of the deficient version 1.0 algorithm on this study.

Response: We agree with the reviewer. We removed this statement because it is irrelevant to this paper. Please also see our response to R1.5.

R2.8: Line 249. Not sure what GOES-16/17 have to do with anything. Suggest 'which is now applied to VIIRS.'

Response: Done.

R2.9: Line 301-311: This is an important section with no rationalization for the decisions made to create the training/validation datasets. You should explain why each of these decisions were made and justified.

R2.10: Line 316: define complicated.

Response (2.9 and 2.10): Thanks for the suggestions. Both are highly relevant to comments from the first reviewer R1.6 and R1.12. We gave a very comprehensive response and made necessary modifications.

R2.11: Line 327: describe how the tuning and optimization were achieved.

Response: The remainder of Section 4.3 gives a brief introduction of the tuning and optimization. However, to make our point more clearly, we have added the following statement to the revised text: "*In this study, we tested six groups of input variables for each RF model. The set of model input variables with a relatively high accuracy score and low memory/computing requirement will be selected.*"

R2.12: Line 334: It would be useful to elaborate on possible reasons for the importance of geolocation as an input and the lack of importance for Ts. Why use Ts instead of Tclr computed at TOA? Wouldn't the latter be more consistent with the traditional approaches?

Response: As shown in Table 3, we found that both geolocation and Ts are important in the RF all-day model. $\boldsymbol{\varepsilon_s}$ is less important likely because it is correlated to surface type and geolocation. Here we use Ts instead of T_{clr} because the calculation of T_{clr} requires more input (e.g., temperature/humidity profiles), and a RT model, which introduces more uncertainty and requires more computational resources.

R2.13: Line 346: Not clear what you mean by similar tests. Consider elaborating further.

Response: We modified the "similar tests" to "similar input variable tests". For the daytime model, we also tried 6 different input combinations. We added another table (Table 4) in the revised version.

R2.14: Line 348: change to 'IR bands used in the all-day model'

Response: Corrected.

R2.15: Line 353: Consider tabulating the daytime results similar to table 2. I think this would be useful.

Response: Done.

R2.16: Line 378 and further: Figs 6-9 are fine but it would help the reader better understand the comparisons if these data could also be tabulated (unless of course you don't think that they are significant enough to further illuminate)

Response: We agree with the reviewer. To make the figures easier to understand, we have added the total number of pixels for each surface type to the corresponding plot. Moreover, we have inserted a detailed description of "unknown phase" category and a new table (Table 5) in Section 4.5.2 to demonstrate the importance of "unknown phase" category for each cloud phase product.

R2.17: Line 387-389: Is this any surprise considering that you have eliminated the most difficult clouds?

Response: As mentioned at the beginning of this section (Section 4.4), we emphasized that the comparisons (shown in Figures 6-9) are also based on “aerosol-free”, “homogeneous”, “single-phase” pixels. It is not a big surprise considering that these simple cases are used in model training and testing (see Tables 3 and 4). However, we were surprised by the performance of the RF all-day model. Although only 3 IR window bands are used, the TPR-FPR points from the RF all-day model looks much better than the current MODIS MYD06 IR-Phase, and are comparable to the OP-Phase that uses more spectral information from shortwave bands.

R2.18: Lines 406-412: the results in figures 8 and 9 are not very clear or well described. In a relative sense, which algorithms are overdetecting or underdetecting ice and water clouds and why?

Response: For cloud phase classification, we arbitrarily define ice clouds and liquid water clouds as “positive” and “negative” events, respectively. Therefore, a low TPR indicates underestimation of ice cloud fraction, while a high FPR indicates a large fraction of liquid water cloud samples are identified as ice cloud. It is found that for snow/ice and barren regions, many non-ML models have much lower accuracy rates than for ocean and grassland surfaces. Possible reasons include strong surface reflection, low surface cloud contrast, relatively less training samples and high solar zenith angles (for snow/ice surface).

To address the reviewer’s questions, we have added the following statement to Section 4.5.2:
“*A low TPR indicates underestimation of ice cloud fraction, while a high FPR indicates a large fraction of liquid water cloud samples are identified as ice cloud.*”
“*Overall, the performance of the hand-tuned algorithms decreases significantly over snow/ice or barren surfaces. For example, the TPR-FPR plot shows that over daytime snow/ice surface (Figure 8 g), the MODIS CLDPROP OP-Phase and MODIS MYD06 IR-Phase frequently predict liquid water cloud as ice cloud. Similar to the daytime plot, the MYD06 IR-Phase also shows a high FPR rate over snow/ice surface, indicating an overestimated (underestimated) ice (liquid water) cloud fraction. Possible reasons include strong surface reflection, low surface cloud contrast, relatively less training samples and high solar zenith angles. However, the two RF models work fairly well and show consistent accuracy rates across all surface types.*”

R2.19: Line 450: change to something like this “The above results indicate that for the screened data considered here, the two RF models have better and more consistent performance over different regions and surface types in comparison with the MODIS and VIIRS products suggesting the potential to improve the overall performance in more global operational applications.

Response: Done. We appreciate the reviewer’s suggestion.

R2.20: Line 457: It is good to drive home the point regarding the ease and cost savings of applying ML vs the traditional approaches which took years to develop. ‘a few hours’ seems vague tho. Consider elaborating further.

Response: Good point! We reorganized the structure of this paragraph by including necessary information on the “labor comparison” between ML and non-ML methods. Please also see our response to R1.11 for more details.

R2.21: Line 459-462: Do they really use similar input? The channel complements are different, so if this in any way affects the phase determination, then what you are saying could be unfair and misleading since the two methods were not designed for continuity.

Response: We modified the statement to “For example, although the MODIS CLDPROP OP-Phase and VIIRS CLDPROP OP-Phase are designed for climate record continuity purpose, cloud thermodynamic phases from the two products are different by up to 4% for all surface pixels, and by up to 10% over surfaces covered by snow/ice (see Figure 8 light blue and light green dots).”

R2.22: Line 465: In this section it should be emphasized again that a screened dataset is used to train and test the ML methods that excludes the more difficult pixels for passive sensor methods. While the ML methods appear to offer some advantages, the higher accuracies found here compared to the traditional approaches may not be representative of those found when applied to a more inclusive dataset.

Response: We agree with the reviewer, though we note that the traditional approaches considered in this study, particularly the MYD06 and CLDPROP OP-Phase algorithms, were themselves tuned off of CALIOP data using similar single-phase data screening (see *Marchant et al.*, 2016), and thus may also suffer degraded performance in complex scenes. In the revised version, we have added a new paragraph and a new figure to demonstrate the performance of the RF all-day model with CALIOP detected multi-phase scenes. We find that probabilities could be more informative than using a single “label”. It is obvious that for complicated samples, ice/liquid cloud probabilities from the RF model are more broadly distributed, resulting in a reduced peak at either 0 or 1. However, further investigation is required to understand how to quantitatively use these probabilities in complex cases. Please also see our response to R1.6.

R2.23: Lines 474-478: This is also vague and won’t make much sense to most readers. What is the objective for your passive determination? Consider elaborating further on the definition and applications for cloud phase (cloud top or radiative), and the relative sensitivities of passive vs active. Maybe then it would be more clear what you mean when you say that a multi-layer clouds category could help.

Response: We agree with the reviewer. In this section, our intent is to mention the limitations of using CALIOP data only for the collection of “simple” cases. Therefore, we modified this paragraph as:

“*The RF models learn spectral structures of cloud/clear pixels according to the reference labels. As a consequence, the present model performance relies heavily on the quality of CALIOP Level-2 data. It is already known that the lidar signal has limitations in detecting the bottom of an optically thick cloud or lower level clouds underneath an opaque cloud [Sassen and Cho, 1992]. Some complicated multiple-phase scenes may be misidentified as simple single-phase scenes due* *to the penetration limit of CALIOP (e.g., the uppermost ice cloud optical thickness greater than 3). Using combined CALIOP and CloudSat data as reference in the future could be a better way to improve the training/validation datasets [Marchant et al., 2020]. However, as noted in that study, CloudSat observations cannot be used without careful filtering since a multilayer scene that is radiatively indistinct from the upper level cloud layer is not necessarily consistent with multilayer detection detected from a cloud radar.*"

R2.24: Lines 489-490. The screening process almost certainly impacts the comparisons with the traditional methods which were not developed with a similar screening process. Please make sure that you address this somewhere in the manuscript.

Response: The non-ML approaches considered in this study, particularly the MYD06 and CLDPROP OP-Phase algorithms, use a similar data screening (see *Marchant et al.*, 2016), and thus may also suffer degraded performance in complex scenes. It is very hard to quantitatively estimate to what extent the screening process could impact those non-ML methods. However, in the revised version, we provided more details about the data selection strategy in Section 4.2 plus two new Tables (2 and 5).

R2.25: Line 518: why is this more impractical? It actually seems necessary.

Response: Adding complexity to the RF (or other ML) model requires more overhead, such as memory at run-time, computational resources, etc. It could be a potential (but not critical) problem when implementing in an operational algorithm production environment, where there often are limitations on such resources (e.g., caps on memory usage). That said, we decided to remove this statement because there are ways to mitigate these technical issues given sufficient resources.

R2.26: Line 534: using the collocated CALIOP products in 2017 and excluding the more difficult pixels associated with polluted, broken and mixed-phase cloud conditions.

Response: Corrected.

R2.27: Line 553: should read " : : :phase detections in a limited set of conditions.

Response: We understand the reviewer's concern. Instead of simply adding "in a limited set of conditions" here, we updated this paragraph to:

"*In this study, we have demonstrated the advantages of using ML-based (specifically, RF) models in cloud masking and thermodynamic phase detection. In contrast with hand-tuned methods, the RF models can be efficiently trained and tested for different surface types and using different input variables. Meanwhile, for aerosol-free, homogeneous samples, the two RF models show better and more consistent performance over different regions and surface types in comparison with existing VIIRS and MODIS datasets. For more complicated scenes, RF probabilities are more informative than binary mask/phase designations. However, further investigation is required to understand how to use probabilities more quantitatively.*"

R2.28: Line 555: consider changing 'a few hours' to 'considerably more efficiently' ??

Response: Done.

R2.29: Line 562 and 563: change ‘can’ to ‘could’

Response: Done.

R2.30: Line 564: Suggest adding this at the end: It remains as future work to determine how such an approach might lead to improved consistency in cloud properties derived from different satellite remote sensors.

Response: Done.

R2.31: Line 607: reformat with last name first or change reference on line 150.

Response: Done.

R2.32: Line 651: reformat with last name first or change reference on line 121

Response: Done.

R2.33: Line 829: Why is MODIS CLDPROP not shown in figure 12?

Response: For legibility reasons, we decided to limit the number of line plots in the figure. The MODIS CLDPROP curves are not included because their locations and structures are quite similar to the VIIRS products.

A Machine Learning-Based Cloud Detection and Thermodynamic Phase Classification Algorithm using Passive Spectral Observations

Chenxi Wang[1,2], Steven Platnick[2], Kerry Meyer[2], Zhibo Zhang[3], Yaping Zhou[1,2]

[1]Joint Center for Earth Systems Technology, University of Maryland Baltimore County, Baltimore, MD, USA

[2]Earth Science Division, NASA Goddard Space Flight Center, Greenbelt, MD, USA.

[3]Department of Physics, University of Maryland Baltimore County, Baltimore, MD, USA.

Deleted: [2]NASA

Abstract

We trained two Random Forest (RF) machine-learning models for cloud mask and cloud thermodynamic phase detection using spectral observations from VIIRS on Suomi NPP (SNPP). Observations from CALIOP were carefully selected to provide reference labels. The two RF models were trained for all-day and daytime-only conditions using a 4-year collocated VIIRS/CALIOP dataset from 2013 to 2016. Due to the orbit difference, the collocated CALIOP and SNPP VIIRS training samples cover a broad viewing zenith angle range, which is a great benefit to overall model performance. The all-day model uses 3 VIIRS infrared (IR) bands (8.6, 11, and 12 μm) and the daytime model uses 5 Near-IR (NIR) and Shortwave-IR (SWIR) bands (0.86, 1.24, 1.38, 1.64 and 2.25 μm) together with the 3 IR bands to detect clear, liquid water, and ice cloud pixels. Up to 7 surface types, namely, ocean/water, forest, cropland, grassland, snow/ice, barren/desert, and shrubland, were considered separately to enhance performance for both models. Detection of cloudy pixels and thermodynamic phase with the two RF models were compared against collocated CALIOP products from 2017. It is shown that, with a conservative screening process that excludes the most challenging cloudy pixels for passive remote sensing, the two RF models have high accuracy rates in comparison with the CALIOP reference for both cloud detection and thermodynamic phase. Other existing SNPP VIIRS and Aqua MODIS cloud mask and phase products are also evaluated, with results showing that the two RF models and the MODIS MYD06 optical property phase product are the top 3 algorithms with respect to lidar observations during the daytime. During the nighttime, the RF all-day model works best for both cloud detection and phase, in particular for pixels over snow/ice surfaces. The present RF models can be extended to other similar passive instruments if training samples can be collected from

Deleted: -

Deleted: It is shown that

Deleted: For cloud detection, the accuracy rates of the daytime RF model are higher than 92% for all surface types, while the accuracy rates of the all-day RF model decrease by 3~8%, depending on surface type. For cloud thermodynamic phase, both RF models agree well with CALIOP, except over barren/desert regions.

CALIOP or other lidars. However, the quality of reference labels and potential sampling issues that may impact model performance would need further attention.

1. Introduction

Detection and classification (DC) of atmospheric constituents using satellite observations is often a critical initial step in many remote sensing algorithms. For example, a prerequisite for cloud optical and microphysical property retrievals is identifying the presence of clouds, i.e., a clear/cloudy classification [*Frey et al.*, 2008; *Heidinger et al.*, 2012]. Additionally, characteristics such as cloud thermodynamic phase are needed as they can strongly impact the scattering/absorption properties of cloud droplets/particles [*Pavolonis et al.*, 2005; *Platnick et al.*, 2017]. Similarly, current operational aerosol algorithms can only retrieve aerosol optical depth (AOD) for "non-cloudy" pixels since even slight cloud contamination can result in erroneously high retrieved AOD [*Remer et al.*, 2005]. Therefore, errors in detecting and classifying atmospheric components can significantly impact downstream retrieval products and scientific analyses.

Deleted:].

There are many examples of hand-tuned DC algorithms designed for satellite instruments. For example, the Moderate Resolution Imaging Spectroradiometer (MODIS) has algorithms developed for cloud masking [*Frey et al.*, 2008; *Ackerman et al.*, 2008], cloud thermodynamic phase [*Baum et al.*, 2012; *Marchant et al.*, 2016], aerosol type [*Levy et al.*, 2013; *Sayer et al.*, 2014], and snow coverage over land surfaces [*Hall and Riggs*, 2016]. Decision trees or voting schemes involving multiple thresholds are typically used in these hand-tuned algorithms. The decision tree branches, tests, and thresholds are often determined empirically after a tedious hand tuning/testing process based on the developer's experience and access to validation datasets. Further, the branches and thresholds are often very sensitive to the specific instrument (e.g.,

Deleted: traditional

Deleted: traditional

spectral band pass, calibration, noise characteristics, view/solar geometry sampling). Therefore, an obvious weakness of these hand-tuned methods is that it is challenging and time consuming to develop algorithms across multiple instruments and to maintain performance for individual instruments that may have noticeable calibration drifts. Meanwhile, a well-designed hand-tuned method may have remarkable performance in a specific region and season yet have significant biases when applied globally and/or annually [*Cho et al.*, 2009; *Liu et al.*, 2010]. Additional complexities arise when DC problems become more non-linear across large spatial and temporal scales, and more variables need to be considered. It is difficult to develop and apply a single or a few decision trees to complicated non-linear problems that are controlled by dozens or more variables. As expected, a single decision tree can grow very deep and tend to have a highly irregular structure in order to consider a large number of features (variables) simultaneously, leading to a significant overfitting effect (i.e., an over-constrained training that makes predictions too close to the training dataset but fails to predict future observations reliably). For example, MODIS provides an all-day cloud phase product based only on infrared (IR) observations (hereafter referred to as IR-Phase [*Baum et al.*, 2012]). Although it can be expected that the tests and thresholds should vary with satellite viewing geometry [*Maddux et al.*, 2010], full consideration of viewing geometries, together with the variations of many other factors such as surface emission, geolocation, and cloud properties, is very challenging based on manual tuning. As a consequence, it is found that the liquid water and ice cloud fractions from the IR-Phase product exhibit noticeable view zenith angle (VZA) dependency (see Figure 12). This is an undesirable but unavoidable artifact since cloud phase statistics should be independent from solar/viewing geometry. Such VZA dependencies may strongly affect similar products from

Deleted: traditional methods (e.g., decision trees/voting schemes/thresholds)

Deleted: a single instrument having radiometric stability issues.…

Deleted: traditional

Deleted: ; *Zhou et al.*, 2019].

geostationary imagers because of the fixed VZA-geolocation mapping. Similar artifacts may also impact aerosol type and retrieval products [*Wu et al.*, 2016].

Deleted: instruments

Deleted: Finally, it is difficult to acquire pixel-level classification uncertainties with traditional methods.

In contrast to hand-tuned methods, Machine Learning (ML) based DC algorithms are designed to autonomously find information (e.g., patterns of spectral, spatial, and/or time series) in one or more given datasets and learn hidden signatures of different objects. An obvious advantage of ML models is that the training process is efficient and highly flexible. Manually defined thresholds or matching conditions to expected spectral patterns are no longer needed. Recently, ML models have been utilized in a wide variety of cloud/aerosol related applications, such as cloud detection [*Thampi et al.*, 2017], cirrus detection and optical property retrievals [*Kox et al.*, 2014; *Strandgren et al.*, 2017], surface-level PM2.5 concentration estimation [*Hu et al.*, 2017], and automatic ship-track detections [*Yuan et al.*, 2019]. In this paper, we developed two ML-based DC algorithms for detecting cloud and cloud thermodynamic phase for different local times (i.e., daytime and nighttime) with observations from the Visible Infrared Imaging Radiometer Suite (VIIRS) on Suomi NPP (SNPP). The ML models are trained with collocated observations from SNPP VIIRS and Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP), with CALIOP data used as the reference. In Section 2, we give a brief discussion of the ML models. Data generated for model training and validation will be introduced in Section 3. Details of the model training and evaluation are shown in Section 4. Section 5 discusses the advantages and potential limitations of the present ML models. Conclusions are given in Section 6.

Deleted: traditional

Deleted: -

2. Hand-tuned DC methods and Machine Learning Models

Deleted: Traditional

2.1 Hand-tuned DC methods

Deleted: Traditional

All DC algorithms with remote sensing observations are based on the underlying physics of the spectral, spatial, and/or temporal structures of specified objects. In hand-tuned DC algorithms,

Deleted: traditional

all the physical rules and structures have to be explicitly defined as various tests and thresholds. For example, the MODIS MOD35/MYD35 cloud mask algorithm uses more than 20 tests with visible/near-infrared (VNIR), shortwave-infrared (SWIR), and infrared (IR) observations [*Frey et al.*, 2008] that are carefully designed to consider numerous scenarios, including different surface types (e.g., ocean, land, desert, snow, etc.) and local times (day/night). Similar algorithms are designed for aerosol type and cloud thermodynamic phase classifications. As an example, Figure 1 illustrates spectral patterns of 5 typical daytime oceanic scenes (pixel types) observed by SNPP VIIRS. The spectral pattern of each of the 5 scenes, namely, clear sky, liquid water cloud, ice cloud, dust, and smoke, is averaged by using more than 1,000 pixels with the same type. It is clear that the 5 scenes are different in either reflectance ratios between a given VNIR/SWIR band and the 0.86 μm band, or brightness temperature differences (BTD) between two IR window bands (Figure 1). Consequently, such spectral features are frequently used to differentiate pixel types in DC algorithms. In addition to spectral patterns, simple methods are developed to take into account spatial information. For example, it is found that cloud reflectance usually has larger spatial variability than aerosols [*Martins et al.*, 2002] and clear sky pixels [*Platnick et al.*, 2017]. Therefore, spatial variabilities of VNIR and SWIR reflectance bands are used to differentiate clouds from non-cloudy pixels in the current MODIS clear sky restoral (CSR) algorithm [*Platnick et al.*, 2017] and Dark Target aerosol retrieval algorithm [*Levy et al.*, 2013].

Deleted: Spectral

Deleted: .

2.2 Machine learning models

Different from the hand-tuned DC methods, ML algorithms are developed to autonomously learn the hidden spectral/spatial/temporal patterns of different objects. Consequently, manually defined thresholds or matching conditions to expected patterns are no longer needed. In image recognition applications, numerous ML algorithms [e.g., *Joachims* 1998; *Breiman* 1999;

Deleted: traditional

Dietterich 2000] were developed in late 1990s for independent pixels using a single or small number of decision trees. *Ho* [1998] and many other studies have demonstrated that, although these single or small number of decision trees can always provide maximum prediction accuracies in training processes, significant overfitting effects cannot be avoided. Tremendous efforts have been made to overcome the dilemma between maintenance of prediction accuracy and avoiding overfitting. Among these, the Random Forest (RF) and Gradient Boosting (GB) algorithm [*Breiman* 1999; *Dietterich* 2000; *Friedman* 2001] provide a framework of using a large number of decision trees (ensemble) but a subset of features in each tree to achieve optimization in the performance. It has been demonstrated that the ensemble-based algorithms can largely correct mistakes made by individual trees [*Ji and Ma*, 1997; *Tumer and Ghosh*, 1996; *Latinne et al.*, 2001] and avoid overfitting [*Freund et al.*, 2001]. Currently, the RF and GB algorithms are frequently used in non-linear classification and regression problems. For example, RF models have been used in several cloud/aerosol remote sensing applications, such as differentiating cloudy from clear footprints for the Clouds and the Earth's Radiation Energy System (CERES) instrument [*Thampi et al.*, 2017], estimating surface-level PM2.5 concentrations [*Hu et al.*, 2017], and detecting low clouds with the Advanced Baseline Imager (ABI) on the recent Geostationary Operational Environmental Satellites (GOES) [*Haynes et al.*, 2019]. In our study, we also choose the RF model based on its proven record in earth science applications.

Deleted: have been

Deleted: in late 1990s.

Deleted:

Deleted: proven

Deleted: cloud using

In the RF model, a final prediction is made based on majority vote computed from probability (P_i) of each class (i^{th}):

$$P_i = \frac{w_i N_i}{\sum_{j=1}^{j=m} w_j N_j}, \tag{1}$$

where m is the total number of classes, N_i and N_j are the number of trees that predict the i^{th} and j^{th} classes, and w_i and w_j are weightings for the i^{th} and j^{th} classes, respectively. If all trees are equally weighted, w for each individual class is equal to 1. The two most important parameters for tuning the RF algorithm are the number of decision trees (N_{Tree}) and the maximum tree depth (N_{Depth}). However, an optimal definition of these two parameters is still an open question [*Latinne et al.*, 2001]. Larger N_{Tree} and N_{Depth} provides more accurate predictions at the cost of significantly increased computational resources. For many cases, larger N_{Depth} may cause overfitting effects [*Oshiro et al.*, 2012; *Scornet*, 2018]. Generally, the two parameters have to be large enough to let the decision trees have a relatively wide diversity and capture the hidden patterns. However, for practical purposes, the two parameters have to be small enough to prevent the models from overfitting and to reduce computing burden [*Latinne et al.*, 2001; *Scornet* 2018].

Deleted: classes are

Deleted: provide

Deleted: ,

Deleted: For

Deleted: however,

Deleted: cannot

Deleted: too large

In this study, we adopt a widely applied RF algorithm in the Scikit-learn Machine Learning package [*Pedregosa et al.*, 2011]. We train two RF models for object DC using SNPP VIIRS spectral observations at two observational times: an all-day RF model using three VIIRS thermal IR observations (hereafter referred to as the RF all-day model) and a daytime-only RF model that uses both VNIR/SWIR and thermal IR observations (hereafter the RF daytime model). The models are trained to detect clear sky, liquid water cloud, and ice cloud pixels with single pixel level information. Parameters of the two RF models will be tuned and tested carefully to achieve the best accuracy and to avoid the overfitting effect. Details will be discussed in Section 4.

3. Data

3.1 Reference label of pixels

Space-borne active sensors, such as CALIOP onboard CALIPSO [*Winker et al.*, 2013], the Cloud-Aerosol Transport System (CATS) [*McGill et al.*, 2015] onboard the International Space

Station (ISS), and CPR on board CloudSat [*Stephens et al.*, 2002], are frequently used to evaluate the performance of hand-tuned cloud/aerosol DC and property retrieval algorithms designed for passive sensors [*Stubenrauch et al.*, 2013; *Wang et al.*, 2019]. CALIPSO, a key member of the Afternoon Constellation of satellites (A-Train) until its exit on 13 September 2018 to join CloudSat in a lower orbit, began providing profiling observations of the atmosphere in 2006 [*Winker et al.*, 2013]. The CALIPSO lidar CALIOP operates at wavelengths of 532 nm and 1064 nm, measuring backscattering profiles at a 30-meter vertical and 333 m along-track resolution. CALIOP also measures the perpendicular and parallel signals at 532 nm, along with the depolarization ratio at 532 nm that is frequently used in cloud phase discrimination algorithms because of its strong particle shape dependence. The CALIOP Version 4 Level 2 1 km/5km Layer product is used to provide reference cloud phase labels in both model training and validation stages.

Deleted: traditional

Formatted: Font color: Auto

Deleted: Until its exit on 13 September 2018 (to join CloudSat in the C-Train),

Deleted: was

Deleted:), and

Deleted: Cloud

Deleted: will be

While the CATS lidar and the CloudSat radar CPR also provide profiling information, both have limitations that preclude their use here. CATS had a relatively short life time (from January 2015 to October 2017), and its low inclination angle (51°) orbit aboard the ISS excludes sampling of high-latitude regions [*Noel et al.*, 2018]. CloudSat CPR observes reflectivity profiles at 94-GHz, which are more sensitive to optically thicker clouds consisting of large particles but are blind to aerosols and optically thin clouds. CloudSat also has difficulty in detecting clouds near the surface due to the surface clutter effect [*Tanelli et al.*, 2008]. Therefore, only CALIOP data are used to provide reference cloud phase labels in this study.

3.2 RF model input

It should be pointed out that ML models use similar input datasets as hand-tuned methods. The input variables (features) and reference labels of the present RF models are carefully selected based on prior physical knowledge of the spectral characteristics of each object.

Deleted: traditional

VIIRS on SNPP and the NOAA-20+ series provides spectral observations from 0.4 to 12 μm at sub-kilometer spatial resolutions [*Lee et al.*, 2006]. Specifically, VIIRS has 16 moderate resolution bands (M band) and 5 higher resolution imagery bands (I band) at 750 m and 375 m nadir resolutions, respectively. The spectral capabilities of VIIRS allow for extracting abundant information on the surface and atmospheric components, such as clouds [*Ackerman et al.*, 2019] and aerosols [*Sayer et al.*, 2017]. It is also worth noting that VIIRS utilizes an on-board detector aggregation scheme that minimizes pixel size growth in the across-track direction towards swath edge [*Cao et al.*, 2013]. As an example, although the VIIRS M-bands and MODIS 1 km bands have similar nadir spatial resolutions, the VIIRS across-track pixel size increases to roughly 1.625 km at scan edge, which is much smaller than a MODIS pixel size of roughly 4.9 km at scan edge [*Justice et al*., 2011]. Another obvious advantage of using SNPP VIIRS rather than Aqua MODIS data is that, due to the CALIPSO and SNPP orbit differences, the training samples cover a broader viewing zenith angle range, which is a great benefit to overall model performance. Consequently, Level-1B M-band observations from the SNPP VIIRS are used here.

Deleted: The

Deleted: VIIRS

Deleted: 2018

Deleted: only to

Deleted: the

Ancillary data, including the surface skin temperature, spectral surface emissivity, surface types, and snow/ice coverage, are important in cloud DC related remote sensing applications [*Frey et al.*, 2008; *Wolters et al.*, 2008; *Baum et al.*, 2012] and cloud/aerosol retrievals [*Levy et al.*, 2013; *Wang et al.*, 2014; 2016a; 2016b; *Meyer et al.*, 2016; *Platnick et al.*, 2017]. The inst1_2d_asm_Nx product (version 5.12.4) from the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) [*Gelaro et al.*, 2017] is utilized to provide the hourly instantaneous surface skin temperature and 10-meter surface wind speed. The UW-Madison baseline fit land surface emissivity database [*Seemann et al.*, 2008] and the Terra/Aqua MODIS combined Land surface product (MCD12C1 [*Sulla-Menashe and Friedl* 2018]) are used to provide monthly mean land surface emissivities for the mid-wave to thermal IR bands (3.6 ~ 14.3 μm) and surface white sky albedo for the VNIR bands (0.4 ~ 2.3 μm), respectively, at a 0.05×0.05° spatial resolution. Surface types and snow/sea ice coverage data are from the International Geosphere-Biosphere Programme (IGBP) and daily Near-real-time Ice and Snow Extent (NISE) data [*Brodzik and Stewart*, 2016], respectively.

3.3 Clear and cloud phase classifications from existing VIIRS and MODIS products

Since the present RF models are trained with SNPP VIIRS observations, the first priority of this study is evaluating and comparing the trained RF models with CALIOP and the existing VIIRS cloud products. However, existing cloud mask and phase products from Aqua MODIS are still used as a reference in this work.

Deleted: Due to the viewing geometry differences between VIIRS and MODIS, it is difficult to simply attribute the mask/phase differences from the RF models and MODIS products to algorithms.

Deleted: compared

The Aqua MODIS and SNPP VIIRS CLDMSK (cloud mask) and CLDPROP (cloud top and optical properties) [*Ackerman et al.*, 2019] products represent NASA's effort to establish a long-term consistent cloud climate data record, including cloud detection and thermodynamic phase, across the MODIS and VIIRS observational records. While the CLDMSK (version 1.0) and CLDPROP (version 1.1) algorithms share heritage with the standard Collection 6.1 MODIS cloud mask (MYD35) and cloud top and optical properties (MYD06) algorithms, the algorithms use only a subset of bands common to both sensors to minimize differences in instrument spectral information content.

Deleted: both

Deleted: The initial Version 1.0 of the CLDMSK and CLDPROP products were publicly released in April 2019; CLDPROP has since been reprocessed to Version 1.1, which includes a fix to the optical property thermodynamic phase algorithm, with public release in October 2019.

The CLDMSK and MYD35 algorithms use a variety of band combinations and thresholds depending on cloud and surface types [*Frey et al.*, 2008; *Ackerman et al.*, 2008]. Meanwhile, the algorithms use different approaches for daytime (i.e., solar zenith angle less than 85°) and nighttime pixels. In the CLDMSK and MYD35 algorithms, pixels are categorized into four categories, namely confident clear, probably clear, probably cloudy, and cloudy. The CLDPROP and MYD06 algorithms separate cloudy and probably cloudy pixels into liquid water, ice, and unknown phase categories. Specifically, the MYD06 product includes two cloud phase algorithms: an IR-Phase algorithm [*Baum et al.*, 2012] that uses observations in four MODIS IR bands for daytime and nighttime phase classification (hereafter referred to as the MYD06 IR-Phase), and a daytime-only algorithm designed for the cloud optical properties retrievals [*Marchant et al.*, 2016; *Platnick et al.*, 2017] that uses VNIR/SWIR and IR observations (hereafter referred to as the MYD06 OP-Phase). A notable change for the VIIRS/MODIS CLDPROP algorithm with respect to the standard MODIS MYD06 algorithm is the replacement of the MYD06 IR-Phase by a NOAA operational algorithm originally developed for Clouds from AVHRR-Extended (CLAVR-x) [*Heidinger et al.*, 2012] and now applied to VIIRS. This algorithm is used to provide cloud top properties, including thermodynamic phase (hereafter CLDPROP CT-Phase), in the absence of the MODIS CO$_2$ IR gas absorption bands. IR bands are primarily used in the CLDPROP CT-Phase algorithm, while complementary SWIR bands are used when available. The MYD06 OP-Phase algorithm, applied to daytime pixels only, is included with only minor alteration (related to cloud top properties changes) in the VIIRS/MODIS CLDPROP product (hereafter referred to as the CLDPROP OP-Phase).

Deleted: ,

Deleted: and GOES-16/17.

Although the MYD06 and CLDPROP OP-Phase products are developed for “cloudy” and “probably cloudy” pixels from the MYD35 and CLDMSK products, a Clear Sky Restoral (CSR) algorithm [*Platnick et al.*, 2017] is implemented to remove “false cloudy” pixels from the clear-sky conservative MYD35 and CLDMSK products. Specifically, the CSR uses a set of spectral and spatial reflectance variability tests to remove dust, smoke, and strong sunglint pixels that are erroneously identified as “cloudy” or “probably cloudy” by the MYD35 and CLDMSK products

[*Platnick et al.*, 2017]. One should keep in mind that the CSR algorithm is only applied for the optical property retrievals. Thus, the MYD35 and CLDMSK, and consequently the MYD06 IR-Phase and CLDPROP CT-Phase, may have "false cloudy" pixels in comparison with CALIOP, while the impact on the MYD06 and CLDPROP OP-Phase is reduced due to the CSR algorithm. The cloud mask and thermodynamic phase products used in this study are summarized in Table 1.

4. Model training and validation

Here we discuss the training of the all-day and daytime RF models for different surface types. Both shortwave (SW) and IR observations will be used in the daytime models while only IR observations will be used in the all-day models. ML model performance is strongly dependent on the quality of training samples. In this study, the two RF models are trained and tested with simple yet highly confident samples (Section 4.2). With this training strategy, the RF models are expected to capture the key spectral features from the pure samples efficiently. As discussed in Section 4.4, we conducted a model validation that evaluates performance of the two models for simple cases. Furthermore, an analysis of probability distributions from the RF all-day model is conducted to demonstrate that the RF models have capability to recognize spectral features from more than one category when atmospheric columns are more complicated.

4.1 Surface Types

RF models are trained for different surface types, defined here by the Collection 6 (C6) MODIS annual IGBP surface type product (MCD12C1), to improve model performance over a single general model for all surface types. Although the MCD12C1 product includes up to 18 surface types, for this work we attempt to reduce the total number of surface types by combining surface types with similar spectral white sky albedos and emissivities, as suggested by *Thampi et al.* [2017]. An annual global IGBP surface type map and surface albedo data from the MODIS

MCD12C1 [*Sulla-Menashe and Friedl* 2018] and a UW-Madison monthly global land surface emissivity database [*Seemann et al.*, 2008] are used to generate the climatology of land surface white-sky albedo and IR emissivity spectra. The UW-Madison database is derived using input from the MODIS operational land surface emissivity product MOD11 [*Wan et al.,* 2004] at six wavelengths located at 3.8, 3.9, 4.0, 8.6, 11, and 12 μm. A baseline fit method is applied to fill the spectral gaps and provides a more comprehensive IR emissivity dataset at 10 wavelengths from 3.6 to 14.3 micron for global land surface with a 0.05° spatial resolution [*Seemann et al.*, 2008]. The MODIS MCD12C1 product also provides a white-sky albedo dataset at 0.47, 0.56, 0.66, 0.86, 1.24, 1.64, and 2.13 μm with a 0.05° spatial resolution [*Sulla-Menashe and Friedl* 2018]. The means and standard deviations of surface emissivity and white-sky albedo spectra are shown in Figures 2 a) and 3 a), respectively, for 16 different land surface types generated from the UW-Madison and MCD12C1 data in 2015. Land surface types with similar IR emissivity and SW white-sky albedo spectra are grouped to reduce to the total number of land surface types to 6 (forest, cropland, grassland, snow/ice, barren/desert, and shrubland), as shown in Figures 2 (b-f) and 3 (b-f). Figure 4 shows an example map of the reduced global surface type data generated from the MCD12C1 product for 2015.

Deleted:],

Deleted:],

4.2 Generating Training/Validation Datasets

Formatted: Font: Not Bold, Font color: Auto

The training and validation data are obtained from a 5-year (2013-2017) SNPP VIIRS and CALIOP collocated dataset. The collected dataset is generated with a collocation algorithm that fully considers the spatial differences between the two instruments and parallax effects, as described in *Holz et al.* [2008]. The SNPP VIIRS data include L1B calibrated reflectance and brightness temperatures, and the CALIOP data include the L2 1km/5km cloud and aerosol layer products. Although more than 332 million VIIRS 750m pixels are collocated with CALIOP

Deleted: strict

Deleted: 1 km cloud layer products. A strict three-step quality control process is applied to all collocated pixels to ensure data quality in the training process.

observations, 130.6 million of these pixels (39.3%) that include only aerosol-free, homogeneous, clear (39.1 million) or single-phase cloud (49.7 million liquid and 41.8 million ice) pixels are used in our training/validation process. Unless otherwise specified, "*aerosol-free*" is defined as those pixels having collocated CALIOP 5km column 532 nm aerosol optical depth less than 0.05, "*homogeneous*" is defined as those pixels for which the collocated CALIOP 1km and 5km products have the same pixel labels, and "*single-phase cloud*" is defined as those pixels for which the collocated CALIOP 1km and 5km products indicate the same thermodynamic phase for all identified cloud layers. More details are given in Table 2.

A strict three-step quality control process is applied to collect samples for the training/validation process. First, VIIRS 750 m pixels that are potentially contaminated by aerosol are excluded using a threshold of 0.05 column AOD at 532 nm from the CALIOP L2 5 km aerosol layer product. Second, each aerosol-free pixel is labelled by one of four categories, namely, "clear sky" and "liquid-water cloud", "ice cloud", and "ambiguous" with the CALIOP L2 1km/5km layer product. The "ambiguous" pixels, including uncertain/unknown cloud phases from CALIOP and/or overlapping objects belonging to different types (e.g., cirrus over liquid), are discarded. Third, horizontally inhomogeneous pixels, determined when the CALIOP 1km label changes within 5 consecutive VIIRS pixels, or pixels with inconsistent CALIOP 1km and 5km labels, are discarded. Figure 5 shows the global distributions of the 5-year collocated clear (first row) and cloudy pixels (second row) before and after applying the three-step quality control. Globally, 50% of all clear pixels are excluded due to contamination of broken-cloud and/or aerosol. In particular, a large fraction of clear pixels in central Africa, India, and southern China (Figure 5c) are excluded due to relatively large aerosol optical thicknesses in those regions. About 40% of global cloudy pixels (Figure 5f) are excluded due to cloud heterogeneity and aerosol contamination. The

Deleted: 1

Deleted: ,

Deleted: ,

Deleted: clouds using

Deleted: 1 km

Deleted: , where the liquid and ice categories include both single-layered cloud scenes and multi-layered clouds with the same thermodynamic phases (e.g., ice over ice, liquid over liquid). Furthermore, the pixels with

Deleted: types

Deleted: Finally

Deleted: 1 km

Deleted: total

Deleted: probably

Deleted: Regions with complicated cloud structures (e.g., minimum selection rate (~20%) can be found in some particular regions, such as the Inter Tropical Convergence Zone (ITCZ), where clouds have complicated horizontal/vertical structures due to strong convections (i.e., clouds are highly heterogeneous in both the horizontal and vertical dimensions). The remaining data are separated into a training/testing population that consists of 32.4, 41.2 and 34.9 million pixels for clear sky, liquid water cloud, and ice cloud from years 2013-2016, respectively, and a validation dataset that consists of 6.9, 8.5 and 7.0 million pixels of clear-sky, liquid water cloud, ice cloud, respectively from year 2017.

4.3 RF model training and configuration

RF model performance is determined by both its inputs (spectral or other information) and its configuration (N_{Tree} and N_{Depth}). Therefore, extensive testing must be conducted to find the optimal inputs and configuration. The 4-year collocated VIIRS-CALIOP dataset from 2013 to 2016 after quality control (see Section 4.2) is used for both training (75%) and testing (25%) purposes. The testing set, also known as cross-validation set, is used to tune and optimize the RF model parameters. Here we define an accuracy score to evaluate the overall model performance. The accuracy score is the ratio of pixels (samples) where both the CALIOP and RF model have the same categories to total pixels. In this study, we tested six groups of input variables for each RF model. The set of model input variables with a relatively high accuracy score and low memory/computing requirement will be selected.

Table 3 provides accuracy scores of the IR-based all-day model trained and tested with different inputs. It shows that with a fixed RF model configuration (N_{Tree} = 150 and N_{Depth} = 15), the RF all-day model with input #4 and #6 have the best overall accuracy scores for all surface types. Generally, by including surface skin temperature (T_s) and geolocation (i.e., latitude and longitude), the accuracy scores for all surface types increase by 2-3%. The surface emissivity

Deleted:)

Deleted: minimum selection rates (20%).

Deleted: which

Deleted: 33

Deleted: 0

Deleted: 24.6

Deleted: 3

Deleted: 5

Deleted: that the

Deleted:

Deleted: 2

vector $\boldsymbol{\varepsilon}_s$ is less important, likely because this information is highly correlated to surface type and geolocation. In this study, input #4 is selected mainly because with similar performance, it requires less memory and computing resources, and it is quite possible that more uncertainty is introduced with the use of a surface emissivity vector $\boldsymbol{\varepsilon}_s$ from another retrieval product.

A set of model configurations (N_{Tree} and N_{Depth}) are also tested based on the selected input #4. While the number of trees and the maximum depth of individual trees are important determinants for RF model performance, the overall accuracy scores for all surface types are less sensitive to these two model parameters when more than 100 trees and 10 maximum tree depths are used (not shown here). Therefore, we trained the RF all-day models with input #4 and the model configuration used in Table 3, i.e., N_{Tree} = 150 and N_{Depth}= 15.

Deleted: 2

Similar input variable tests for the RF daytime model (IR plus NIR and SWIR observations) showed that the optimal input includes reflectances in the 0.86, 1.24, 1.38, 1.64 and 2.25μm bands, BTs in the same 3 IR bands used in the all-day model, geolocation, and solar/satellite viewing zenith angles (See Table 4). The same model configuration used in the all-day model, e.g., 150 trees with the maximum depth 15, is used in the daytime model. The accuracy scores of the RF daytime model are higher than the RF all-day model by 2-3% over almost all surface types except high-latitude regions covered by snow and ice, where the daytime model accuracy score is higher by up to 6% than the all-day model due to the inclusion of the 1.38, 1.64 and 2.25μm SWIR bands.

Deleted: T_s,

Deleted: .

4.4 Evaluating the RF Models

The trained RF all-day and daytime models are validated using collocated CALIOP data in 2017. Existing VIIRS cloud products CLDMSK and CLDPROP (see Table 1) are included for direct comparison with the RF models and CALIOP reference. Several other products, such as the

MODIS CLDMSK and CLDPROP and standard MYD35 and MYD06, are also included for comparison although they could be different from the RF models due to other non-algorithm reasons, such as the VZA and pixel size differences mentioned before.

4.5.1 Cloud mask

Cloud mask from the two RF models and VIIRS/MODIS products are first compared with CALIOP lidar observations. For the two models, a cloudy pixel indicates a predicted label "liquid" or "ice". Here we define cloudy and clear pixels as "positive" and "negative" events, respectively. A true positive rate (TPR) and false positive rate (FPR) can then be used to evaluate model performance. The TPR and FPR are defined as:

$$\mathrm{TPR} = \frac{TP}{TP+FN}, \tag{2}$$

$$\mathrm{FPR} = \frac{FP}{FP+TN}, \tag{3}$$

where TP (True Positive) and TN (True Negative) are the number of lidar-labeled "cloudy" and "clear" pixels, respectively, that are correctly detected by the models; whereas FN (False Negative) and FP (False Positive) are the number of lidar-labeled "cloudy" and "clear" pixels incorrectly identified by the models. Therefore, TPR, also called model sensitivity, indicates the fraction of all positive events (i.e., lidar cloudy pixels) that are correctly detected by the models. Similarly, FPR, also called false alarm rate, indicates the fraction of all negative events (i.e., lidar clear pixels) that are incorrectly detected as positive (cloudy). TPR and FPR are two critical parameters in model evaluation. A perfect model is associated with a high TPR (close to 1) and a low FPR (close to 0).

Figure 6 shows daytime cloud mask TPR-FPR plots from the two RF models and the other products listed in Table 1. Globally, all products agree well with lidar observations (Figure 6a).

The overall TPRs are higher than 0.94 and FPRs are lower than 0.08. The RF daytime model (red circle), with a TPR of 0.97 and an FPR of 0.05, is slightly better than the RF all-day model (yellow circle) and other products. Figure 6b-6h show comparisons over different surface types. It is clear that the RF daytime model has a robust performance for all surface types. The MODIS MYD35 cloud mask algorithm (black circle) performs best over ocean but has a relatively high FPR (0.22) over forest and low TPR over snow/ice and barren (0.85) regions. As mentioned in Section 3, the "false cloudy" pixels from MYD35 and CLDMSK may increase the FPRs correspondingly.

The RF all-day model works fairly well and is comparable to other products for all surface types regardless of the fact that it only uses three IR window channels from VIIRS while all other products in the daytime models use VNIR observations. Nighttime (SZA > 85°) cloud mask comparisons are shown in Figure 7. The overall performances of all operational products decrease in particular for snow/ice regions. For example, the VIIRS/MODIS CLDMSK products over snow/ice surface have large fractions of missing "cloudy" pixels (e.g., TPRs < 0.7) and false alarm rates (FPRs > 0.2) over snow/ice surface. The decrease is more likely explained by the lack of SWIR bands and the small cloud-snow/ice surface temperature contrast during the nighttime of summer polar regions. However, the RF all-day model has the best performance for nighttime pixels, indicating the strong capability of ML based algorithm in capturing hidden spectral features and optimizing dynamic thresholds of clear and cloudy pixels.

4.5.2 Cloud thermodynamic phase

The RF cloud thermodynamic phase products are also compared with CALIOP lidar and existing VIIRS and MODIS products. For consistent nomenclature, we arbitrarily define ice clouds and liquid water clouds as "positive" and "negative" events, respectively. A low TPR indicates underestimation of ice cloud fraction, while a high FPR indicates a large fraction of liquid water

Deleted: model derived

cloud samples are identified as ice cloud. To focus on cloud thermodynamic phase classification, pixels detected as “clear” by either the lidar reference labels or by the RF models and existing products are excluded. The OP-Phase from both MYD06 and CLDPROP, and the IR-Phase from MYD06, have an “unknown phase” category, which is not included in the TPR-FPR analysis.

Deleted:

Figure 8 shows daytime cloud phase TPR-FPR plots from the two RF models and the MODIS/VIIIRS products. The two RF models and the MODIS MYD06 OP-Phase are the top 3 phase algorithms for all surface types. The MODIS MYD06 IR-Phase, MODIS/VIIRS CLDPROP OP-Phase, and CT-Phase have either relatively lower TPRs or higher FPRs over particular surface types, such as shrubland, snow/ice, and barren regions. Comparisons between nighttime phase algorithms are shown in Figure 9. For nighttime clouds, the RF all-day model works better than both CT-Phase and IR-Phase algorithms for all surface types. Overall, the performance of the hand-tuned algorithms decreases significantly over snow/ice or barren surfaces. For example, the TPR-FPR plot shows that over daytime snow/ice surface (Figure 8 g), the MODIS CLDPROP OP-Phase and MODIS MYD06 IR-Phase frequently predict liquid water cloud as ice cloud. Similar to the daytime plot, the MYD06 IR-Phase also shows a high FPR rate over snow/ice surface, indicating an overestimated (underestimated) ice (liquid water) cloud fraction. Possible reasons include strong surface reflection, low surface cloud contrast, relatively less training samples and high solar zenith angles. However, the two RF models work fairly well and show consistent accuracy rates across all surface types.

It is also important to note that the number of pixels used for cloud phase TPR-FPR comparisons in Figures 8 and 9 are different for products that have “unknown phase” categories, namely, MYD06 IR-Phase, MYD06 OP-Phase, and CLDPROP OP-Phase. As shown in Table 5, the MYD06 IR-Phase has a relatively large “unknown phase” phase fraction (15% for all surface

Deleted: Figure 10

types and 34% for snow/ice) in comparison to the OP-Phase products from both MYD06 and CLDPROP, which have 2~3% “unknown phase” fraction approximately.

As discussed in Section 2.2, recall that the RF model predicted pixel type is derived by setting thresholds on the probabilities for each classification type, e.g., an ice phase decision is reached if the probability of ice is greater than the probabilities of liquid and clear. Figure 10 shows the probability distribution functions of the RF all-day model for four scene types as determined by collocated CALIOP, namely, (a) clear, (b) liquid, (c) ice, and (d) multi-layer clouds with different thermodynamic phases (e.g., ice over liquid). As expected, for the first three types, which are included in the training/validation processes, the probability distributions have strong peaks close to either 0 or 1. For the multiple phase cases (panel d), the liquid and ice probabilities are more broadly distributed, indicating that the model may recognize signals from both liquid and ice and therefore provide ambiguous phase results. More nuanced thresholds can therefore be applied to the probabilities, for instance to create an “unknown” phase category following MYD06 and CLDPROP convention [*Marchant et al.*, 2016] that can indicate complicated cloud scenes. Furthermore, the probabilities themselves can provide a useful quality assurance metric for downstream cloud property retrievals that often must make an assumption on cloud phase. Nevertheless, assigning an appropriate phase for downstream imager-based cloud property retrievals is difficult for complex, multilayer cloud scenes, as such an assignment often depends on the optical/microphysical properties and vertical distribution of the cloud layers in the scene [*Marchant et al.*, 2020]. Further investigation is necessary to understand how to use the RF phase probabilities more quantitatively in complicated cases.

Figure 11 shows monthly mean daytime cloud and phase fractions from the VIIRS CLDMSK and CLDPROP OP-Phase products (top row), and those from the RF daytime model (second row), in January 2017. For the cloud mask comparison, cloud fractions (CF) from the two products have similar spatial patterns, while it is also clear that the VIIRS CLDMSK CFs are higher over tropical oceans by approximately 10% and lower over land by 5% (Figure 11 c). This is consistent with the cloud mask TPR-FPR analysis shown in Figure 6. Over the tropical ocean, the VIIRS CLDMSK is more “cloudy”, probably due to a fraction of sunglint pixels that are detected as liquid clouds, leading to a large FPR rate. Another reason for the relatively large cloud fraction (or liquid water cloud fraction) difference is that in regions covered by “broken” cumulus clouds, and or clouds with more complicated structures, the inherent viewing geometry differences in the training datasets may adversely affect the performance of the RF models. For example, CALIOP, with a nadir viewing geometry may observe clear gaps between two small cloud pieces, while VIIRS, with an oblique viewing angle, detects broken liquid clouds nearby or high clouds along its long line-of sight. Comparison between the VIIRS product and the RF daytime model shows more ice clouds from the RF daytime models over land, which is consistent with the cloud phase TPR-FPR plots as shown in Figure 8. The RF daytime model may have better performance due to the consideration of surface type. However, it is also important to notice that due to the lack of “aerosol” types in current training, in central Africa, the RF models may misidentify elevated smoke as ice cloudy pixels. For most land surface types except snow/ice, the CLDPROP OP-Phase has lower TPR rates than the RF daytime models by 0.1, in comparison with the CALIOP.

Deleted: 10

In addition to the higher CFs over low latitude ocean from the VIIRS CLDMSK product, more pronounced CF (liquid) differences can be found in northeast and northwest China. Cloud differences in the two regions are spatially correlated with locations that have heavy aerosol loadings or snow coverage. For example, heavy aerosol loadings due to pollution in Northeast China, and a wide land snow coverage in Northwest China are frequently observed in the winter.

The VIIRS CLDMSK may identify pixels with white surface and heavy aerosol loadings as "cloudy". Some of these pixels are expected to be restored to clear-sky category in the CLDPROP OP-Phase product (Figure 11 f and i). As evidence, Figure 12 shows comparisons between the VIIRS products and the RF daytime model in July 2017. The large cloud (liquid) fraction differences over North China vanish in the summer. This indicates that the RF models might be able to handle complicated (or unexpected) surface type and strong aerosol events better than the hand-tuned VIIRS algorithm. However, further investigation is required to understand the performances of both the VIIRS products and the RF models.

Deleted: 10

Deleted: 10

Deleted: 11

Deleted: traditional

5. Discussion

In this Section, we will review the strengths and potential limitations and weaknesses of the RF models.

5.1 Advantages

The above results show that, for the screened clear/cloudy samples, the two RF models have better and more consistent performance over different regions and surface types in comparison with the MODIS and VIIRS products, suggesting the potential to improve the overall performance in more global operational applications. In addition to better performance, it is convenient and efficient to apply the present RF models or other similar ML-based models to other instruments similar to VIIRS, such as the geostationary imagers Advanced Himawari Imager (AHI) on Himawari-8/9, the ABI on GOES-16/17, and the Spinning Enhanced Visible and Infrared Imager (SEVIRI) on Meteosat Second Generation, as long as reliable reference pixel labels are available. With hand-tuned methods, adjustment is always required in the case of calibration changes, algorithm porting to another similar instrument, or changes in solar/viewing geometries and surface conditions. Manual adjustments can be time-consuming (e.g., months or years), whereas

Deleted: .

Deleted: The

the two RF models used in this study were trained and tested for 7 surface types and using different input variables in 3 hours (on an HPC Platform using 32 Intel Xeon Gold 6126 Processors @ 2.60 GHz). More important, manual algorithm adjustment may not provide the best continuity between two instruments. For example, although the MODIS CLDPROP OP-Phase and VIIRS CLDPROP OP-Phase are designed for climate record continuity purpose, cloud thermodynamic phases from the two products are different by up to 4% for all surface pixels, and by up to 10% over surfaces covered by snow/ice (see Figure 8 light blue and light green dots). Further investigation is necessary to understand if, using ML approaches, a better climate record continuity will be achieved with a uniform training dataset. Besides providing a discrete category for each pixel, the RF models provide an ensemble of predictions and probabilities of individual categories, which are useful diagnostic variables in evaluating models in complicated scenarios.

5.2 Limitations and possible caveats

Although the evaluation demonstrates that the current RF models are highly consistent with CALIOP, the models may suffer some artifacts due to the quality of the training data and due to sampling issues.

5.2.1 Quality of the training/validation data

The RF models learn spectral structures of cloud/clear pixels according to the reference labels. As a consequence, the present model performance relies heavily on the quality of CALIOP Level-2 data. It is already known that the lidar signal has limitations in detecting the bottom of an optically thick cloud or lower level clouds underneath an opaque cloud [*Sassen and Cho*, 1992]. Some complicated multiple-phase scenes may be misidentified as simple single-phase scenes due to the penetration limit of CALIOP (e.g., the uppermost ice cloud optical thickness greater than 3). Using combined CALIOP and CloudSat data as reference in the future could be a better way to

Deleted: can be

Deleted: different

Deleted: a few

Deleted: . In contrast, traditional methods

Deleted: suffer from the change of instrument, solar/viewing geometries, and surface conditions.

Deleted: MYD06

Deleted: use similar input and strategies

Deleted: 5

Deleted: 20

Deleted: black and

Deleted: circles).

Deleted: Cloud

Deleted: classification from the RF daytime model, using both SW and IR observations,

Deleted: different from the

Deleted: at multi-layer scenes if

Deleted: top

Deleted: layer is optically thick enough for lidar (e.g.,

Deleted: and introducing a “multiple layer clouds” category could be a way to mitigate this impact improve the training/validation datasets [*Marchant et al.*, 2020]. However, as noted in that study, CloudSat observations cannot be used without careful filtering since a multilayer scene that is radiatively indistinct from the upper level cloud layer is not necessarily consistent with multilayer detection detected from a cloud radar.

Additional uncertainties may come from the inconsistency in view angles between the collocated CALIOP labels and VIIRS spectral observations. For instance, CALIOP always has a quasi-nadir viewing angle (e.g., 3°) whereas the collocated VIIRS observations have a wide VZA range (e.g., 0° to 50°). A wide VIIRS VZA range in the training dataset improves model performance, especially for predicting VIIRS pixels with large VZAs. However, the difference between the CALIOP and VIIRS viewing geometry could create undesirable artifacts in the training process. As shown in Figure 11, in the descending areas of the Hadley cell over low-latitude ocean, where marine boundary layer clouds are dominant, there are relatively large CF differences between the CLDMSK and the RF models. A reason for the large liquid cloud fraction differences is that the quality of training datasets decreases in regions covered by "broken" cumulus clouds, and or clouds with more complicated structures. Further investigation is required to check if the training dataset collection process introduces sampling bias into the training dataset.

Deleted: 10

Deleted: data screening

5.2.2 Sampling issue

Uneven sampling may also influence the training of RF models. Figure 13 shows the cloud fraction as a function of viewing geometry. Quasi-constant fractions of both liquid and ice clouds are found for all operational products and the RF models when VZAs are smaller than 45°, except the MODIS MYD06 IR-Phase, which has a strong VZA dependency. However, liquid (ice) cloud fractions from the two RF models increase (decrease) rapidly at high VZAs (greater than 50°), which is likely caused by the sampling issue. A significant fraction of the training data (greater

Deleted: 12

than 98%) is located in the region with VZA less than 50° (see the gray dashed distributions in Figure 13). It is difficult to mitigate this issue using collocated VIIRS-CALIOP data or observations from other similar instruments in the training process. One possible way is using model-generated synthetic training data and labels with reliable radiative transfer models. Results from the RF daytime model are not shown in Figure 13 since they are highly consistent with the RF all-day model.

Deleted: 12

Deleted:

Deleted: 12

5.2.3 Labeling strategy

For RF or other ML models, each pixel's classification is determined by prediction probabilities (P) of all potential types. Here we selected a regular strategy that labels a pixel using the class with the highest probability (see Eq. 1). This strategy is logical for problems with two categories (e.g., cloud mask only). For problems including 3 or more classes, however, the present strategy is not the only way to label pixels. For example, a pixel is labeled as "clear" if P_{clear} is larger than both P_{liquid} and P_{ice} according to the current labeling strategy. It is also possible that, for the same pixel (less than 0.5% for the two RF models), P_{clear} is lower than the sum of P_{liquid} and P_{ice}, making a "cloudy" label more appropriate. For the cloud mask and phase problem discussed in this paper, in addition to pixel labels, users must be aware of probabilities of the three types. Another possible way to avoid the ambiguous labeling is using two RF models, one for cloud masking and one for phase, such that a "clear" or "cloudy" label is given first by the cloud mask model, while a corresponding "liquid" or "ice" label is assigned to "cloudy" pixels in the cloud phase model. However, two RF models double the training process and require more computing resources in operational applications.

Deleted: Using multiple RF models becomes more impractical if both aerosol and cloudy pixels are considered.

6. Conclusions

Two Machine-Learning Random Forest (RF) models were trained to provide pixel types (i.e., clear, liquid water cloud, and ice cloud) using VIIRS 750-meter spectral observations. A daytime model that uses NIR, SWIR, and IR bands and an all-day model that only uses IR bands were trained separately. In the training processes, reference pixel labels are from collocated CALIOP Level 2 1 km cloud layer and 5 km aerosol layer products from 2013 to 2016. Careful tests were conducted to optimize model input and configuration. The two RF models were trained for 7 different surface types (i.e., ocean/water, forest, cropland, grassland, snow/ice, barren/desert, and shrubland) to improve model performance. In addition to geolocation and solar/satellite geometry information, we found that using 5 NIR and SWIR bands (0.86, 1.24, 1.38, 1.64 and 2.25 μm) and three IR bands (8.6, 11, and 12μm) in the daytime RF model and using the three IR bands and surface temperatures in the all-day RF model achieved great performances for all surface types.

Deleted: can achieve the best

The cloud mask and thermodynamic phase classifications from the two RF models were validated using the selected aerosol-free, homogeneous samples in 2017. For daytime cloud mask comparisons over all surface types, the RF daytime model, with a high TPR (0.93 and higher) and low FPR (0.07 and lower), performs best among all models evaluated, including MODIS MYD35 and MODIS/VIIRS CLDMSK products. The RF all-day model works fairly well and is comparable to other products for all surface types, even in daytime when all other products use shortwave observations and it does not. For the nighttime cloud mask, the RF all-day model has the best performance over all products, demonstrating the strong capability of ML-based algorithms for capturing hidden spectral features of clear and cloudy pixels. All nighttime products perform slightly weaker at snow/ice regions. The decline is likely explained by the lack of SWIR bands and the small thermal contrast between the clouds and the surface during the summer

Deleted: collocated CALIOP products

nighttime in polar regions. In this case, the ML-based algorithms are not able to compensate for the missing physical signatures.

For the daytime cloud thermodynamic phase comparison, we showed that the two RF models are comparable with the MODIS MYD06 OP-Phase product, and are among the top 3 phase algorithms for all surface types. The MODIS MYD06 IR-Phase, VIIRS/MODIS CLDPROP OP-Phase, and CT-Phase have either relatively lower TPRs or higher FPRs over certain surface types, such as shrubland, snow/ice, and barren regions. For nighttime clouds, the RF all-day model works better than both CLDPROP CT-Phase and MYD06 IR-Phase for all surface types.

In this study, we have demonstrated the advantages of using ML-based (specifically, RF) models in cloud masking and thermodynamic phase detection. In contrast with hand-tuned methods, the RF models can be efficiently trained and tested for different surface types and using different input variables. Meanwhile, for aerosol-free, homogeneous samples, the two RF models show better and more consistent performance over different regions and surface types in comparison with existing VIIRS and MODIS datasets. For more complicated scenes, RF probabilities are more informative than binary mask/phase designations. However, further investigation is required to understand how to use probabilities more quantitatively.

Deleted: detections

Deleted: to traditional

Deleted: manually-defined thresholds and matching conditions are no longer needed. The

Deleted: in a few hours.

Deleted: products.

In the future, more spectral bands and/or spatial patterns can be used to improve pixel classification skills, such as including more pixel types (e.g., dust and smoke). It is convenient to apply RF models or other similar ML-based models to other instruments similar to VIIRS with the help of active instruments. Most importantly, cloud mask and thermodynamic phase products from well-trained RF models could be used to train other instruments in the absence of active sensors. For example, the current RF model based VIIRS cloud mask/phase data could be used as reference to train ML-based models for other instruments, such as MODIS, ABI/AHI, SEVIRI, and airborne

Deleted: can

Deleted: can

instruments. It remains as future work to determine how such an approach might lead to improved consistency in cloud properties derived from different satellite imagers.

Deleted:

It is also important to emphasize that the model performance is highly reliant on the quality of the training samples and reference labels. For example, in this study, more than 98% of the training data have a VZA less than 50°, leading to more uncertain cloud phase fractions at large VZAs. Using synthetic training data generated with reliable radiative transfer models could be a possible way to mitigate this artifact.

Deleted: an unavoidable bias of

Acknowledgements

The authors are grateful for support from the NASA Radiation Sciences Program. C. Wang acknowledges funding support from NASA through the New (Early Career) Investigator Program in Earth Science (80NSSC18K0749) managed by Lin Chambers and Allison Leidner. The computations in this study were performed at the UMBC High Performance Computing Facility (HPCF). The facility is supported by the U.S. National Science Foundation through the MRI program (grants CNS-0821258 and CNS-1228778) and the SCREMS program (grant DMS 0821311), with additional substantial support from UMBC. The Collection 6.1 Aqua/MODIS cloud products (doi: dx.doi.org/10.5067/MODIS/MYD06_L2.061) and MODIS/VIIRS Continuity cloud products (Version 001) are publicly available from the NASA and Atmosphere Archive and Distribution System (LAADS) (http://ladsweb.nascom.nasa.gov). The CALIPSO Level 2 Cloud/Aerosol layer products (version 4) products are publicly available from the Atmospheric Science Data Center (https://eosweb.larc.nasa.gov/).

Reference:

Ackerman, S. A., Holz, R. E., Frey, R., Eloranta, E. W., Maddux, B. C., and McGill, M., Cloud detection with MODIS. Part II: Validation, *J. Atmos. Oceanic Technol.*, **25,** 1073–1086, doi: 10.1175/2007JTECHA1053.1, 2008.

Ackerman, S. A., Frey, R., Heidinger, A., Li, Y., Walther, A., Platnick, S., Meyer, K., Wind, G., Amarasinghe, N., Wang, C., Marchant, B., Holz, R. E., Dutcher, S., Hubanks, P., EOS MODIS and SNPP VIIRS Cloud Properties: User guide for climate data record continuity Level-2 cloud top and optical properties product (CLDPROP), version 1, 2019.

Baum, B. A., Menzel, W. P., Frey, R. A., Tobin, D. C., Holz, R. E., Ackerman, S. A., Heidinger, A. K., and Yang, P., MODIS cloud-top property refinements for Collection 6, *J. Appl. Meteor. Climatol.*, **51,** 1145-1163, doi: 10.1175/JAMC-D-11-0203.1, 2012.

Breiman, L., Random forests - random features. Technical report, University of California at Berkeley, Berkeley, California, 1999.

Moved (insertion) [1]

Brodzik M. J., and Stewart J. S., Near-Real-Time SSM/I-SSMIS EASE-Grid Daily Global Ice Concentration and Snow Extent, Version 5, doi:10.5067/3KB2JPLFPK3R, 2016.

Cao, C., Xiong, J., Blonski, S., Liu, Q., Uprety, S., Shao, X., Bai, Y., and Weng, F., Suomi NPP VIIRS sensor data record verification, validation, and long-term performance monitoring, *J. Geophys. Res. Atmos.*, **118,** 11,664-11,678, doi:10.1002/2013JD020418, 2013.

Cho, H., Nasiri, S. L., and Yang, P., Application of CALIOP Measurements to the Evaluation of Cloud Phase Derived from MODIS Infrared Channels, *J. Appl. Meteor. Climatol.*, **48,** 2169-2180, doi:10.1175/2009JAMC2238.1, 2009.

Dietterich, T. G., Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems, MCS 2000, Lecture Notes in Computer Science, **vol. 1857**, Springer, Berlin, Heidelberg, 2000.

Freund, Y., An Adaptive Version of the Boost by Majority Algorithm, in Machine Learning, **43,** 293-318, 2001.

Deleted: Erwan Scornet,

Moved down [2]: Tuning parameters in random forests. ESAIM: Procs, 60: 144–162, 2018.¶

Frey, R. A., Ackerman, S. A., Liu, Y., Strabala, K. I., Zhang, H., Key, J. R., and Wang, X.: Cloud detection with MODIS. Part I: Improvements in the MODIS cloud mask for Collection 5, *J. Atmos. Oceanic Technol.,* **25,** 1057–1072, doi:10.1175/2008JTECHA1052.1, 2008.

Friedman, J. H., Greedy function approximation: a gradient boosting machine, *Ann. Stat.*, **29,** 1189–1232, 2001.

Gelaro, R., et al., The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), *J. Climate*, **30,** 5419–5454, doi:10.1175/JCLI-D-16-0758.1, 2017.

Hall, D. K., and Riggs, G. A., MODIS/Aqua Snow Cover Daily L3 Global 500m SIN Grid, Version 6. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center, doi:10.5067/MODIS/MYD10A1.006, 2016.

Haynes, J. M., Noh, Y. J., Miller, S. D., Heidinger, A., and Forsythe, J. M., Cloud geometric thickness and improved cloud boundary detection with GEOS ABI, 15th Annual Symposium on New Generation Operational Environment Satellite Systems, Phoenix, AZ, 6 - 10 January, 2019.

Heidinger, A. K., Evan, A. T., Foster, M. J., and Walther, A., A naive bayesian cloud-detection scheme derived from CALIPSO and applied within PATMOS-x, *J. Appl. Meteor. Climatol.*, **51**, 1129–1144, doi:10.1175/JAMC-D-11-02.1, 2012.

Ho, T. K, The random subspace method for constructing decision forests, *IEEE Trans. Pattern Anal. Mach. Intell*. **20,** 832–844, 1998.

Holz, R. E., Ackerman, S. A., Nagle, F. W., Frey, R., Dutcher, S., Kuehn, R. E., Vaughan, M. A., and Baum, B., Global Moderate Resolution Imaging Spectroradiometer (MODIS) cloud detection and height evaluation using CALIOP, *J. Geophys. Res.*, **113,** D00A19, doi:10.1029/2008JD009837, 2008.

Hu, X. F., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., and Liu, Y., Estimating PM2.5 concentrations in the conterminous United States using the random forest approach, *Environmental Science & Technology,* **51,** 6936–6944, doi:10.1021/acs.est.7b01210, 2017.

Ji, C. and Ma, S., Combinations of weak classifiers, *IEEE Transactions on Neural Networks*, **8**, 32–42, 1997.

Joachims, T., Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning, 137–142, Springer-Verlag, 1998.

Justice C. O., Vermote, E., Privette J., and Sei, A., The Evolution of U.S. Moderate Resolution Optical Land Remote Sensing from AVHRR to VIIRS. Land Remote Sensing and Global Environmental Change, B. Ramachandran, C. Justice, and M. Abrams, Eds., Remote Sensing and Digital Image Processing, **11,** Springer, New York, NY., 781-806, 2011.

Kox, S., Bugliaro, L., and Ostler, A.: Retrieval of cirrus cloud optical thickness and top altitude from geostationary remote sensing, *Atmos. Meas. Tech.*, **7,** 3233–3246, doi:10.5194/amt-7-3233-2014, 2014.

Latinne, P., Debeir, O., Decaestecker, C., Limiting the number of trees in random forests, in Multiple Classifier Systems, Manchester, U.K. IEEE, **2013**, 178-187, 2001.

Lee, T. E., Miller, S. D., Turk, F. J., Schueler, C., Julian, R., Deyo, S., Dills, P., and Wang, S., The NPOESS VIIRS Day/Night Visible Sensor, *Bull. Amer. Meteor. Soc.*, **87,** 191–200, https://doi.org/10.1175/BAMS-87-2-191, 2006.

Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., Patadia, F., and Hsu, N. C., The Collection 6 MODIS aerosol products over land and ocean, *Atmos. Meas. Tech.*, **6,** 2989–3034, doi:10.5194/amt-6-2989-2013, 2013.

Liu, Y., Ackerman, S. A., Maddux, B. C., Key, J. R., and Frey, R. A., Errors in cloud detection over the Arctic using a satellite imager and implications for observing feedback mechanisms, *J. Climate*, **23,** 1894–1907, doi:10.1175/2009JCLI3386.1, 2010.

Maddux, B. C., Ackerman, S. A., and Platnick, S., Viewing geometry dependencies in MODIS cloud products, *J. Atmos. Oceanic Technol.*, **27,** 1519–1528, doi:10.1175/2010JTECHA1432.1, 2010.

Moved (insertion) [3]

Formatted: Font color: Auto

Moved up [1]: Random forests - random features. Technical report, University of California at Berkeley, Berkeley, California, 1999.¶

Deleted: Leo Breiman,

Martins, J. V., Tanré, D., Remer, L., Kaufman, Y., Mattoo, S., and Levy, R., MODIS cloud screening for remote sensing of aerosols over oceans using spatial variability, *Geophys. Res. Lett.*, **29**, doi:10.1029/2001GL013252, 2002.

Marchant, B., Platnick, S., Meyer, K. G., Arnold, G. T., and Riedi, J., MODIS Collection 6 shortwave-derived cloud phase classification algorithm and comparisons with CALIOP, *Atmos. Meas. Tech.*, **9,** 1587–1599, doi:10.5194/amt-9-1587-2016, 2016.

Marchant, B., Platnick, S., Meyer, K., and Wind, G.: Evaluation of the Aqua MODIS Collection 6.1 multilayer cloud detection algorithm through comparisons with CloudSat CPR and CALIPSO CALIOP products, *Atmos. Meas. Tech. Discuss.*, doi:10.5194/amt-2019-448, in review, 2020.

McGill, M. J., Yorks, J. E., Scott, V. S., Kupchock, A. W., and Selmer, P. A., The Cloud-Aerosol Transport System (CATS): A technology demonstration on the *International Space Station, Proc. SPIE* **9612**, Lidar Remote Sensing for Environmental Monitoring XV, 96120A, doi:10.1117/12.2190841, 2015.

Meyer, K. G., Platnick, S., Arnold, G. T., Holz, R. E., Veglio, P., Yorks, J. E., and Wang, C., Cirrus cloud optical and microphysical property retrievals from eMAS during SEAC4RS using bi-spectral reflectance measurements within the 1.88 µm water vapor absorption band, *Atmospheric Measurement Techniques*, **9** (4), 1743-1753, doi:10.5194/amt-9-1743-2016, 2016.

Noel, V., Chepfer, H., Chiriaco, M., and Yorks, J.: The diurnal cycle of cloud profiles over land and ocean between 51° S and 51° N, seen by the CATS spaceborne lidar from the International Space Station, *Atmos. Chem. Phys.*, **18,** 9457–9473, doi:10.5194/acp-18-9457-2018, 2018.

Oshiro T. M., Perez P. S., Baranauskas J. A., How many trees in a random forest, in Machine Learning and Data Mining in Pattern Recognition. MLDM 2012. Lecture Notes in Computer Science, **7376,** Springer, Berlin, Heidelberg, 2012.

Pavolonis, M. J., Heidinger, A. K., and Uttal, T., Daytime global cloud typing from AVHRR and VIIRS: Algorithm description, validation, and comparisons, *J. Appl. Meteor.*, **44,** 804–826, 2005.

Pedregosa, F. et al., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830, 2011.

Platnick, S., Meyer, K. G., King, M. D., Wind, G., Amarasinghe, N., Marchant, B., Arnold, G. T., Zhang, Z., Hubanks, P. A., Holz, R. E., Yang, P., Ridgway, W. L., Riedi, J.: The MODIS cloud optical and microphysical products: Collection 6 updates and examples from Terra and Aqua, *IEEE Transactions on Geoscience and Remote Sensing*, **55,** 502-525, doi: 10.1109/TGRS.2016.2610522, 2017.

Remer, L. A., Kaufman, Y. J., Tanré, D., Mattoo, S., Chu, D. A., Martins, J. V., Li, R., Ichoku, C., Levy, R. C., Kleidman, R. G., Eck, T. F., Vermote, E., and Holben, B. N., The MODIS aerosol algorithm, products, and validation, *J. Atmos. Sci.*, **62**, 947-973, doi:10.1175/JAS3385.1, 2005.

Sassen, K., and Cho, B. S., Subvisual-thin cirrus lidar dataset for satellite verification and climatological research, *American Meteorological Society*, **31,** 1275–1285. http://doi.org/10.1175/1520-0450(1992)031<1275:STCLDF>2.0.CO;2, 1992.

Sayer, A. M., Munchak, L. A., Hsu, N. C., Levy, R. C., Bettenhausen, C., and Jeong, M.-J., MODIS Collection 6 aerosol products: Comparison between Aqua's e-Deep Blue, Dark Target, and “merged” data sets, and usage recommendations, *J. Geophys. Res. Atmos.*, **119,** 13,965-13,989, doi:10.1002/2014JD022453, 2014.

Sayer, A. M., Hsu, N. C., Lee, J., Bettenhausen, C., Kim, W. V., and Smirnov, A., Satellite Ocean Aerosol Retrieval (SOAR) algorithm extension to S-NPP VIIRS as part of the “Deep Blue” aerosol project, *J. Geophys. Res. Atmos.*, **123,** doi:10.1002/2017JD027412, 2017.

Scornet, E., Tuning parameters in random forests. ESAIM: Procs, 60: 144–162, 2018.

Moved (insertion) [2]

Seemann, S. W., Borbas, E. E., Knuteson, R. O., Stephenson, G. R., and Huang, H., Development of a global infrared land surface emissivity database for application to clear sky sounding retrievals from multispectral satellite radiance measurements, *J. Appl. Meteor. Climatol.*, **47**, 108–123, 2008.

Stephens, G. L., et al., The CloudSat mission and the A-Train: A new dimension of space-based observations of clouds and precipitation, *Bull. Amer. Meteorol. Soc.*, **83,** 1771-1790, doi:10.1175/BAMS-83-12-1771, 2002.

Strandgren, J., Bugliaro, L., Sehnke, F., and Schröder, L.: Cirrus cloud retrieval with MSG/SEVIRI using artificial neural networks, *Atmos. Meas. Tech.*, **10,** 3547–3573, doi:10.5194/amt-10-3547-2017, 2017.

Stubenrauch, C. J., Rossow, W. B., Kinne, S., Ackerman, S., Cesana, G., Chepfer, H., Di Girolamo, L., Getzewich, B., Guignard, A., Heidinger, A., Maddux, B. C., Menzel, W. P., Minnis, P., Pearl, C., Platnick, S., Poulsen, C., Riedi, J., Sun-Mack, S., Walther, A., Winker, D., Zeng, S., and Zhao, G., Assessment of Global Cloud Datasets from Satellites: Project and Database Initiated by the GEWEX Radiation Panel, *Bull. Amer. Meteor. Soc.*, **94,** 1031–1049, doi:10.1175/BAMS-D-12-00117.1, 2013.

Sulla-Menashe, D., and Friedl, M. A., User Guide to Collection 6 MODIS Land Cover (MCD12Q1 and MCD12C1) Product; USGS: Reston, VA, USA, 2018.

Tanelli, S., Durden, S. L., Im, E., Pak, K., Reinke, D., Partain, P., Haynes, J., and Marchand, R., CloudSat's cloud profiling radar after two years in orbit: Performance, calibration, and processing, *IEEE Trans. Geosci. Remote Sens.*, **46,** 3560–3573, doi:10.1109/TGRS.2008.2002030, 2008.

Thampi, B. V., Wong, T., Lukashin, C., and Loeb, N. G., Determination of CERES TOA fluxes using machine learning algorithms. Part I: Classification and retrieval of CERES cloudy and clear scenes, *J. Atmos. Oceanic Technol.*, **34,** 2329–2345, doi:10.1175/JTECH-D-16-0183.1, 2017.

Tumer, K., and Ghosh, J., Error correlation and error reduction in ensemble classifiers, *Connection Science,* **8**, 385-403, doi:10.1080/095400996116839, 1996.

Wan, Z., Zhang, Y., Zhang, Q., and Li, Z.-L., Quality assessment and validation of the MODIS global land surface temperature, *Int. J. Remote Sens.*, **25**, 261–274, doi:10.1080/0143116031000116417, 2004.

Wang, C., Yang, P., Dessler, A., Baum, B. A., and Hu, Y., Estimation of the cirrus cloud scattering phase function from satellite observations, *Journal of Quantitative Spectroscopy and Radiative Transfer*, **138**, 36-49 doi:10.1016/j.jqsrt.2014.02.001, 2014.

Wang, C., Platnick, S., Zhang, Z., Meyer, K., and Yang, P., Retrieval of ice cloud properties using an optimal estimation algorithm and MODIS infrared observations: 1. Forward model, error analysis, and information content, *J. Geophys. Res. Atmos.*, **121,** 5809-5826 doi:10.1002/2015jd024526, 2016a.

Wang, C., Platnick, S., Zhang, Z., Meyer, K., Wind, G., and Yang, P., Retrieval of ice cloud properties using an optimal estimation algorithm and MODIS infrared observations: 2. Retrieval evaluation, *J. Geophys. Res. Atmos.*, **121**, doi:10.1002/2015jd024528, 2016b.

Wang, C., Platnick, S., Fauchez, T., Meyer, K., Zhang, Z., Iwabuchi, H., and Kahn, B. H., An assessment of the impacts of cloud vertical heterogeneity on global ice cloud data records from passive satellite retrievals, *Journal of Geophysical Research: Atmospheres*, **124,** 1578-1595. doi:10.1029/2018JD029681, 2019.

Winker, D. M., Tackett, J. L., Getzewich, B. J., Liu, Z., Vaughan, M. A., and Rogers, R. R., The global 3-D distribution of tropospheric aerosols as characterized by CALIOP, *Atmos. Chem. Phys.*, **13,** 3345-3361, doi:10.5194/acp-13-3345-2013, 2013.

Wolters, E. L., Roebeling, R. A., and Feijt, A. J., Evaluation of cloud-phase retrieval methods for SEVIRI on Meteosat-8 using ground-based lidar and cloud radar data, *J. Appl. Meteor. Climatol.*, **47,** 1723–1738, doi:10.1175/2007JAMC1591.1, 2008.

Wu, Y., de Graaf, M., and Menenti, M., Improved MODIS Dark Target aerosol optical depth algorithm over land: angular effect correction, *Atmos. Meas. Tech.*, **9,** 5575-5589, doi:10.5194/amt-9-5575-2016, 2016.

Yuan, T., Wang, C., Song, H., Platnick, S., Meyer, K., and Oreopoulos, L., Automatically finding ship tracks to enable large-scale analysis of aerosol-cloud interactions, *Geophysical Research Letters*, **46,** 7726– 7733, doi: 10.1029/2019GL083441, 2019.

Deleted: Zhou, Y., Levy, R., Remer, L., Mattoo, S., Espinosa, R., Dust detection and dust aerosol retrieval with non-spherical aerosol models over Oceans within MODIS Dark-Target algorithm, *Atmos.*

Moved up [3]: *Meas.*

Formatted: Justified, Space Before: 6 pt, After: 6 pt

Deleted: *Tech, to be submitted*, 2019.¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
---Page Break---

Formatted: Font color: Auto

Formatted: Font color: Auto

Table 1. Existing VIIRS and MODIS cloud mask and phase products used for comparison. Note that MYD35 and MYD06 are the standard MODIS Aqua products, and CLDMSK and CLDPROP are the MODIS Aqua and VIIRS common algorithm continuity products.

Instrument	Cloud Mask	Cloud Phase
MODIS	MYD35 V6.1	MYD06 IR-Phase V6.1
		MYD06 OP-Phase V6.1
	CLDMSK V1.0	CLDPROP CT-Phase V1.0
		CLDPROP OP-Phase V1.1
VIIRS	CLDMSK V1.0	CLDPROP CT-Phase V1.0
		CLDPROP OP-Phase V1.1

Table 2: Data collection strategies and the number of pixels for all surface types.

Deleted: Table 2

# of VIIRS 750m pixels (million)	Condition	Ocean	Forest	Cropland	Grass	Barren	Shrub	Snow/Ice	Total
All collocation	None	219.7	18.7	8.7	17.5	17.1	13.6	37.4	332.7
Aerosol Free	CALIOP Aerosol 5km column AOD < 0.05	142.6	13.0	3.7	10.0	10.5	9.3	34.3	223.2
Clear	Aerosol Free, Cloud 1km Layer = 0	17.7	2.5	1.5	1.8	2.9	3.1	13.1	42.5
Clear (homogeneous)	**Aerosol Free, Cloud 1km/5km Layer = 0**	**15.2**	**2.3**	**1.5**	**1.7**	**2.7**	**3.0**	**12.7**	**39.1**
Cloudy	Aerosol Free, Cloud 1km Layer > 0	124.9	10.5	2.1	8.1	7.7	6.2	21.2	180.7
Cloudy (homogeneous)	Aerosol Free, Cloud 1km/5km Layer > 0	115.5	9.5	1.8	7.4	6.6	5.3	15.8	162.0
Single Phase Cloud	Aerosol Free, Cloud 1km Liquid or Ice Phase	65.1	4.4	1.0	4.0	3.4	2.4	13.5	93.7
Single Phase Cloud (homogeneous)	**Aerosol Free, Cloud 1km/5km Liquid or Ice Phase**	**64.2**	**4.3**	**0.9**	**3.9**	**3.3**	**2.3**	**12.7**	**91.5**
Liquid Phase Cloud (homogeneous)	**Aerosol Free, Cloud 1km/5km Liquid Phase**	**40.5**	**1.8**	**0.3**	**1.7**	**1.3**	**1.0**	**3.2**	**49.7**
Ice Phase Cloud (homogeneous)	**Aerosol Free, Cloud 1km/5km Ice Phase**	**23.7**	**2.5**	**0.6**	**2.2**	**2.0**	**1.3**	**9.5**	**41.8**

Table 3: Accuracy scores of RF all-day models based on testing pixels with different inputs and a fixed model configuration (N_Trees = 150 and Max_TreeDepths = 15).

# Input	Model Input	Ocean	Forest	Shrubland	Crop	Grassland	Barren	Snow/Ice	All Surface*
1	$BT_{8.6}$, BT_{11}, BT_{12}, and VZA	90.3	89.9	88.7	88.4	88.2	88.0	87.4	89.4
2	$BT_{8.6}$, BT_{11}, BT_{12}, VZA, and Lat/Lon	92.1	90.1	89.8	90.7	89.5	90.1	88.0	90.9
3	$BT_{8.6}$, BT_{11}, BT_{12}, VZA, and T_S	93.1	90.9	89.9	91.4	90.2	90.3	88.5	91.7
4	$BT_{8.6}$, BT_{11}, BT_{12}, VZA, Lat/Lon, and T_S	93.2	91.7	90.0	91.8	91.2	90.8	88.9	92.0
5	$BT_{8.6}$, BT_{11}, BT_{12}, VZA, T_S, and ε_S	93.2	91.4	89.8	91.4	90.4	90.4	88.8	91.9
6	$BT_{8.6}$, BT_{11}, BT_{12}, VZA, Lat/Lon, T_S, and ε_S	93.2	91.8	90.1	91.8	91.3	90.6	88.9	92.0

*The all-surface accuracy scores are weighted by pixel numbers of individual surface types.

Formatted ... [1]

Deleted: 0.

Deleted: 0.

Deleted: 0.

Deleted: 0.

Deleted: 0.

Deleted: 0.

Deleted: 0.

Deleted: 0.

Deleted: 0.

Deleted: 0.

Deleted: 0.90

Deleted: 0.91

Deleted: 0.90

Deleted: 0.

Deleted: 0.

Deleted: 0.91

Deleted: 0.

Deleted: 0.91

Deleted: 0.90

Deleted: 0.

Deleted: 0.

Deleted: 0.

Deleted: 0.89

Deleted: 0.92

Deleted: 0.

Deleted: 0.92

Deleted: 0.

Deleted: 0.92

Deleted: 0.

Deleted: 0.91

Deleted: 0.89

Deleted: 0.

Deleted: 0.

Deleted: 0.

Deleted: 0.90

Deleted: 0.

Deleted: 0.

Deleted: 0.

Deleted: 0.89

Deleted: 0.92

Deleted: 0.

Deleted: 0.92

Deleted: 0.

Deleted: 0.92

Deleted: 0.

Deleted: 0.91

Deleted: 0.89

Deleted: 0.

Table 4: Accuracy scores of RF daytime models based on testing pixels with different inputs and a fixed model configuration (N_Trees = 150 and Max_TreeDepths = 15).

# Input	Model Input	Ocean	Forest	Shrubland	Crop	Grassland	Barren	Snow/Ice	All Surface*
1	$BT_{8.6}$, BT_{11}, BT_{12}, $R_{0.86}$, $R_{1.38}$, $R_{1.61}$, $R_{2.25}$, VZA, and SZA	95.47	93.71	93.25	93.86	92.82	94.04	94.94	94.97
2	$BT_{8.6}$, BT_{11}, BT_{12}, $R_{0.86}$, $R_{1.38}$, $R_{1.61}$, $R_{2.25}$, VZA, SZA, and RAA	95.47	93.72	93.22	93.84	92.81	94.02	94.94	94.97
3	$BT_{8.6}$, BT_{11}, BT_{12}, $R_{0.86}$, $R_{1.38}$, $R_{1.61}$, $R_{2.25}$, Lat/Lon, VZA, and SZA	95.47	93.74	93.36	93.95	92.95	94.16	94.95	94.99
4	$BT_{8.6}$, BT_{11}, BT_{12}, $R_{0.86}$, $R_{1.38}$, $R_{1.61}$, $R_{2.25}$, $R_{1.24}$, Lat/Lon, VZA and SZA	95.51	93.73	93.47	93.93	92.98	94.21	95.05	95.04
5	$BT_{8.6}$, BT_{11}, BT_{12}, $R_{0.86}$, $R_{1.38}$, $R_{1.61}$, $R_{2.25}$, Ts, Lat/Lon, VZA, SZA, and RAA	95.45	93.77	93.36	93.93	92.92	94.21	94.95	94.98
6	$BT_{8.6}$, BT_{11}, BT_{12}, $R_{0.86}$, $R_{1.38}$, $R_{1.61}$, $R_{2.25}$, $R_{0.48}$, $R_{0.67}$, $R_{1.24}$, VZA, and SZA	95.51	93.90	93.54	94.11	93.07	94.38	95.17	95.09

*The all-surface accuracy scores are weighted by pixel numbers of individual surface types.

Table 5: Fractions of the 2017 validation samples that have determined phases (i.e., liquid water or ice) in different surface types.

Determined Phase (%)	Ocean	Forest	Shrubland	Crop	Grassland	Barren	Snow/Ice	All
MODIS MYD06 IR-Phase	89	**75**	**74**	80	**79**	**75**	**66**	85
MODIS MYD06 OP-Phase	97	99	97	98	99	95	92	97
MODIS CLDPROP OP-Phase	98	99	98	99	99	97	99	98
VIIRS CLDPROP OP-Phase	98	99	97	99	98	96	99	98

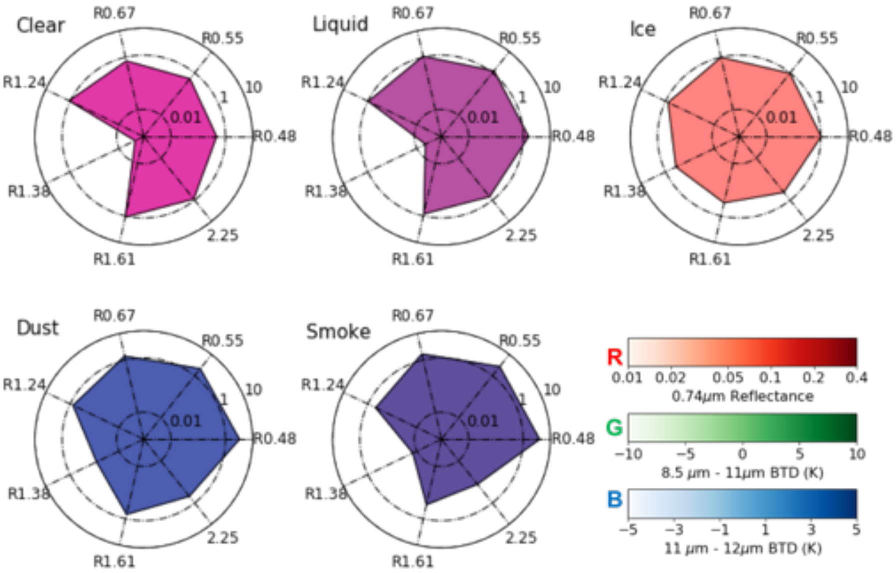


Figure 1. Spectral patterns of the five different pixel types (averaged over 1,000 pixels for each type). For each plot, an apex indicates reflectance ratio between a given VNIR/SWIR band and the 0.86-μm band, and the spread is filled by false RGB composite (Red: 0.74-μm reflectance; Green: 8.5-11μm brightness temperature difference (BTD); Blue: 11-12μm BTD). The spectral patterns are used in the machine learning algorithms.

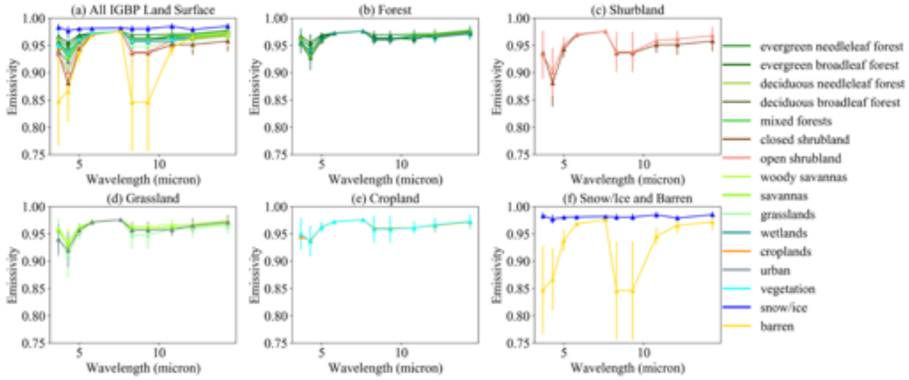


Figure 2. Climatology of the spectral surface emissivity data from the UW-Madison baseline fit land surface emissivity database [*Seemann et al.*, 2008] for different IGBP surface types. Error bars indicate the emissivity standard deviations at given wavelengths.

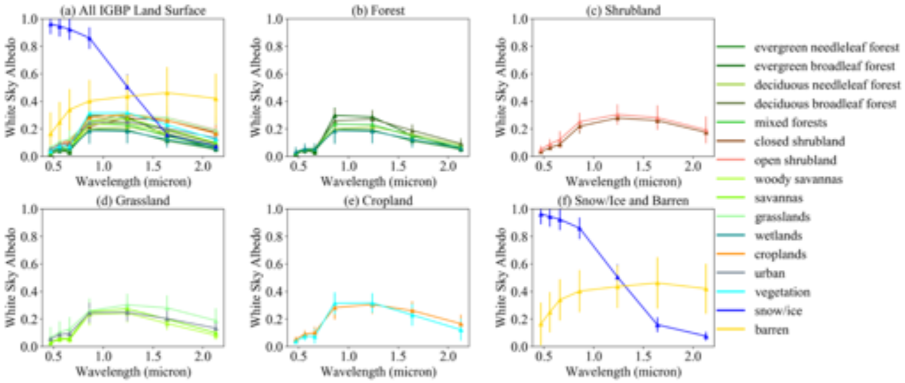


Figure 3. Climatology of the spectral surface white sky surface albedo data from MCD12C1 [*Sulla-Menashe and Friedl* 2018] for different IGBP surface types. Error bars indicate the albedo standard deviations at given wavelengths.

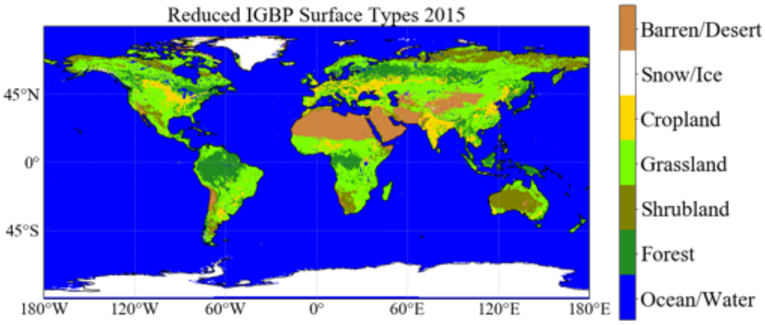


Figure 4. A global map of the seven reduced surface types chosen for the RF model training.

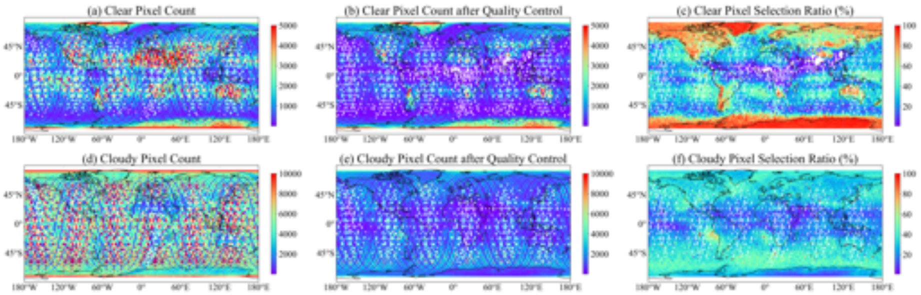
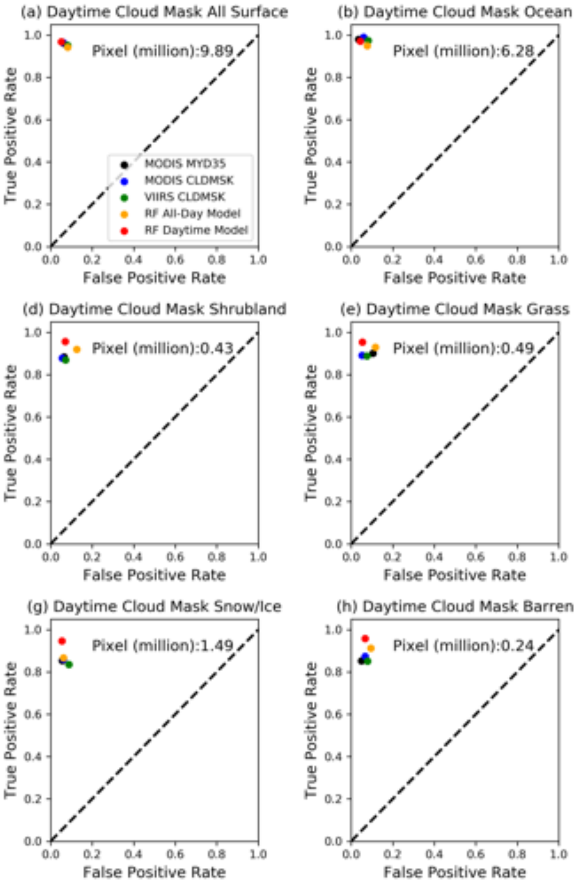

(a) Clear Pixel Count
(b) Clear Pixel Count after Quality Control
(c) Clear Pixel Selection Ratio (%)
(d) Cloudy Pixel Count
(e) Cloudy Pixel Count after Quality Control
(f) Cloudy Pixel Selection Ratio (%)
45°N
0°
45°S
180°W
120°W
60°W
0°
60°E
120°E
180°E
5000
4000
3000
2000
1000
100
80
60
40
20
10000
8000
6000
4000
2000

Figure 5. Global distributions of the of clear and cloudy pixels from collocated VIIRS and CALIOP data from 2013 to 2017. Panels a) and d) show the total clear and cloudy pixel counts, respectively. Panels b) and d) show the pixel counts after applying the quality control. The corresponding selection ratios are shown in panels c) and f).



Deleted:

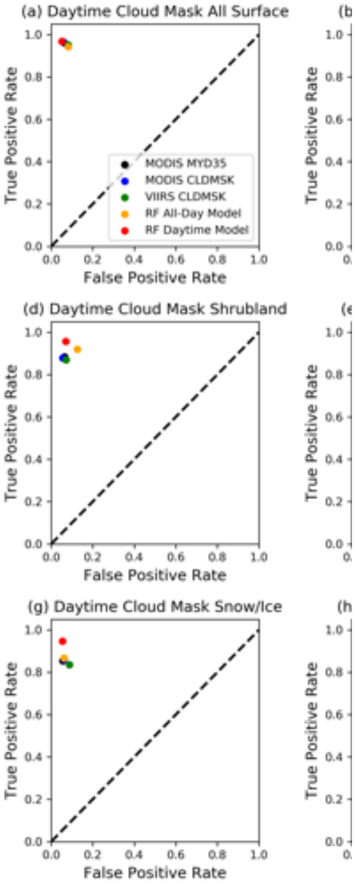


Figure 6. False Positive Rate (FPR) versus True Positive Rate (TPR) plots of daytime cloud mask from the two RF models and operational algorithms. Collocated CALIOP Level 2 products in 2017 are used as reference. Global comparisons are shown in panel (a), while panels (b) through (h) show comparisons for difference surface types. The total pixel number is shown in each panel.

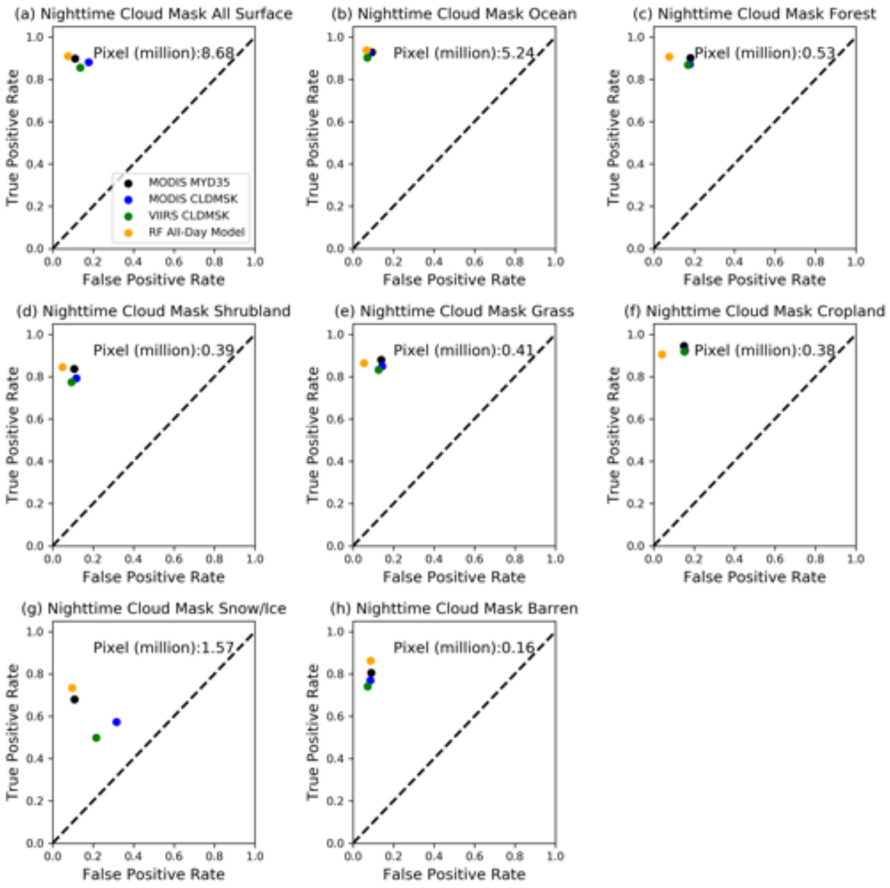
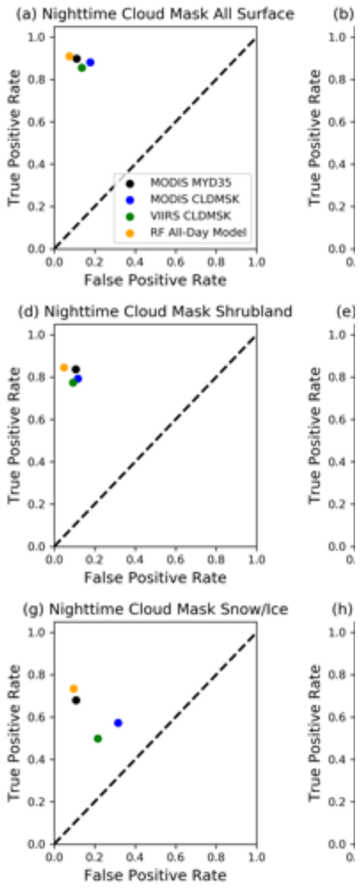


Figure 7. Similar to Figure 6, but for nighttime cloud mask comparisons. The total pixel number is shown in each panel.



Deleted:

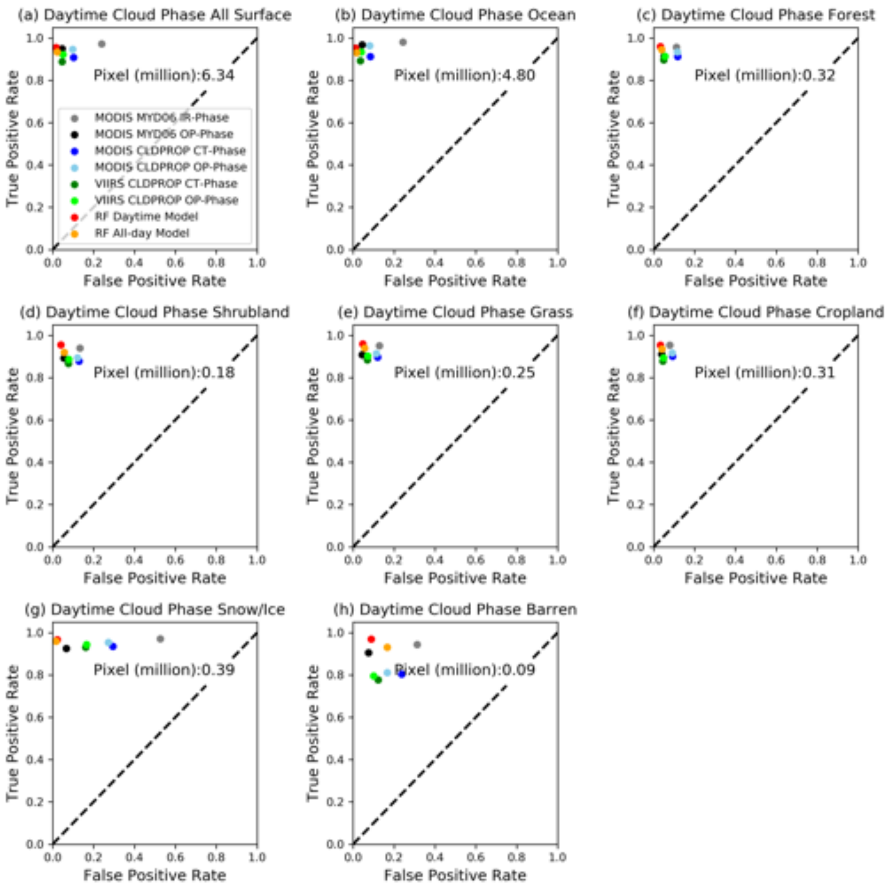
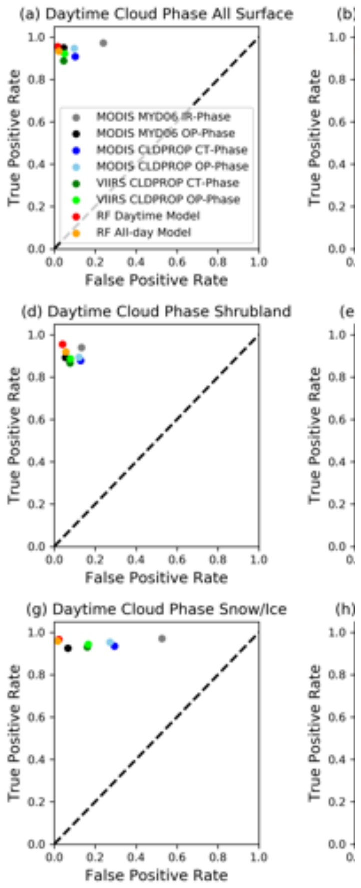


Figure 8. Similar to Figure 6, but for daytime cloud thermodynamic phase comparisons. The total pixel number is shown in each panel. Note that for specific products, the total pixel numbers are less because of the exclusion of “unknown phase” category (see text for more details).

Deleted:



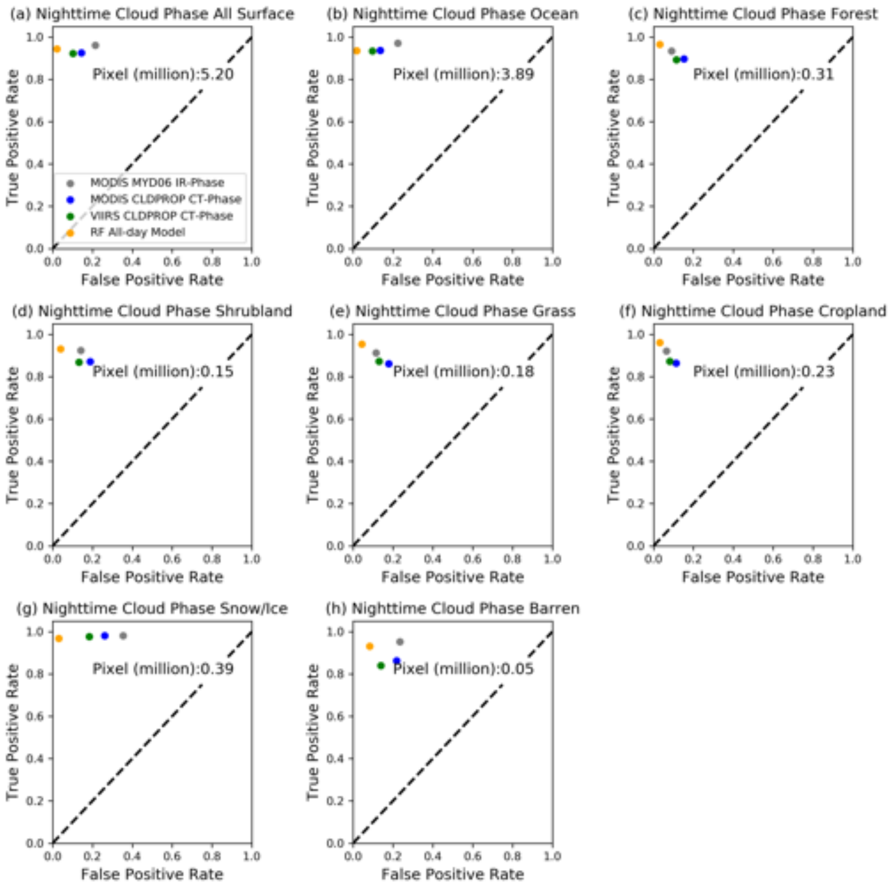
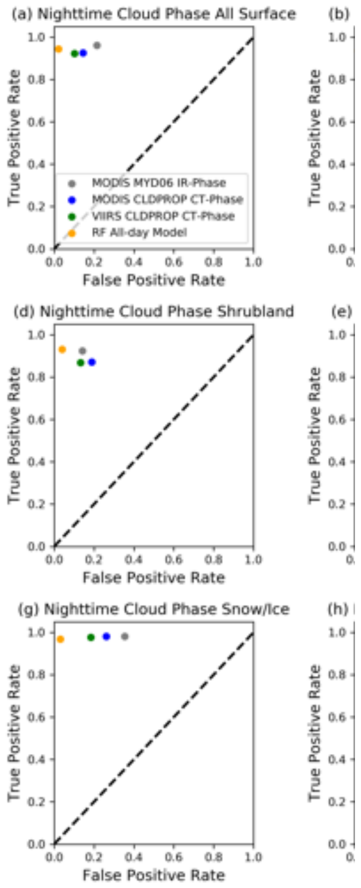


Figure 9. Similar to Figure 6, but for nighttime cloud thermodynamic phase comparisons. The total pixel number is shown in each panel. Note that for specific products, the total pixel numbers are less because of the exclusion of "unknown phase" category (see text for more details).

Deleted:



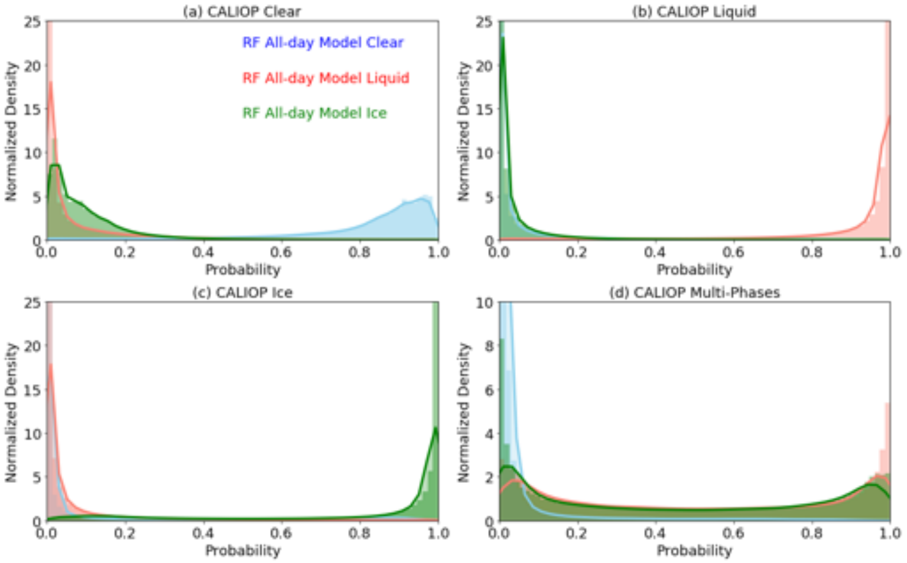


Figure 10. Normalized density functions of the clear (blue), liquid water cloud (red), and ice cloud (green) probabilities from the RF all-day model in four CALIOP detected aerosol-free scenes: (a) clear, (b) homogenous liquid, (c) homogenous ice, and (d) multi-layer cloud with different thermodynamic phases.

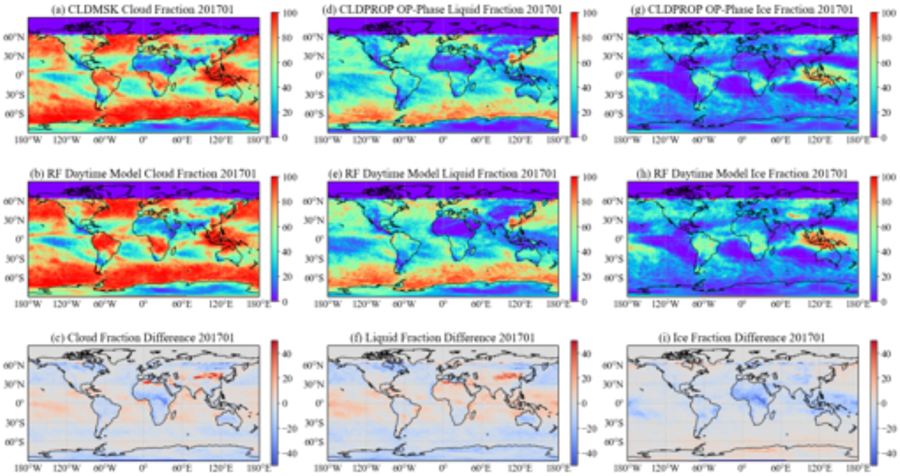

(a) CLDMSK Cloud Fraction 201701
(d) CLDPROP OP-Phase Liquid Fraction 201701
(g) CLDPROP OP-Phase Ice Fraction 201701
(b) RF Daytime Model Cloud Fraction 201701
(e) RF Daytime Model Liquid Fraction 201701
(h) RF Daytime Model Ice Fraction 201701
(c) Cloud Fraction Difference 201701
(f) Liquid Fraction Difference 201701
(i) Ice Fraction Difference 201701
60°N
30°N
0°
30°S
60°S
180°W
120°W
60°W
0°
60°E
120°E
180°E
100
80
60
40
20
0
40
20
0
−20
−40

Figure 11. Comparisons between one-month daytime cloud mask and thermodynamic phase products from the VIIRS CLDMSK and CLDPROP OP-Phase (top row) and the RF daytime model (second row), and their differences (VIIRS – RF daytime, bottom row) in January, 2017.

Deleted: 10

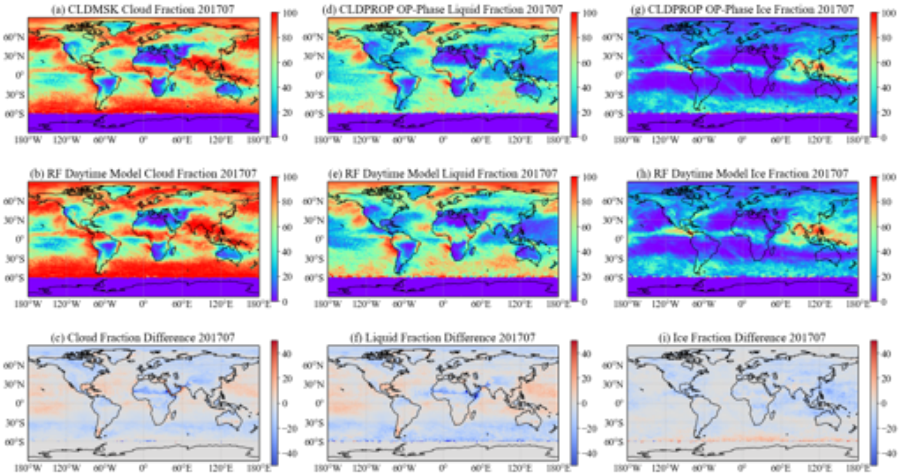

(a) CLDMSK Cloud Fraction 201707
(d) CLDPROP OP-Phase Liquid Fraction 201707
(g) CLDPROP OP-Phase Ice Fraction 201707
(b) RF Daytime Model Cloud Fraction 201707
(e) RF Daytime Model Liquid Fraction 201707
(h) RF Daytime Model Ice Fraction 201707
(c) Cloud Fraction Difference 201707
(f) Liquid Fraction Difference 201707
(i) Ice Fraction Difference 201707
60°N
30°N
0°
30°S
60°S
180°W
120°W
60°W
0°
60°E
120°E
180°E
100
80
60
40
20
0
40
20
0
−20
−40

Figure 12. Similar to Figure 11, but for comparisons in July, 2017.

Deleted: 11

Deleted: 10

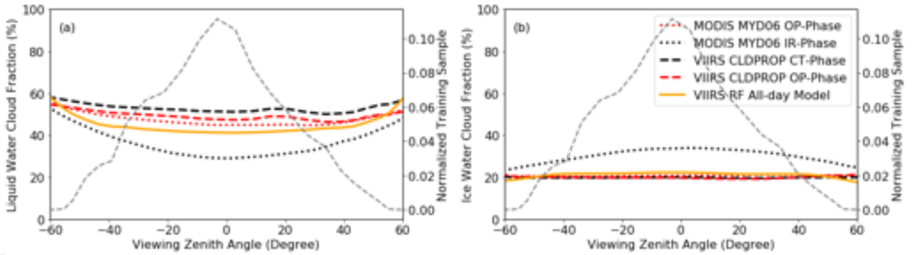
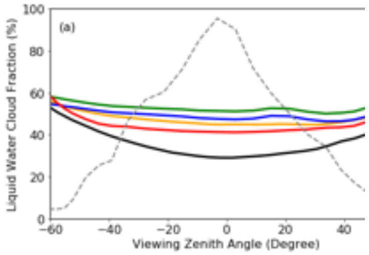


Figure 13. Liquid water (a) and ice (b) cloud fractions as a function of viewing zenith angle from the one-month daytime cloud mask/phase products in January 2017. The gray dashed curve is the probability density function of the 4-year VIIRS/CALIOP training samples (2013-2016).

Deleted:



Deleted: 12

Page 37: [1] Formatted WANG, CHENXI (GSFC-613.0)[UNIVERSITY OF MARYLAND BALTIMORE COUNTY] 2/20/20 9:38:00 PM

Font: 12 pt