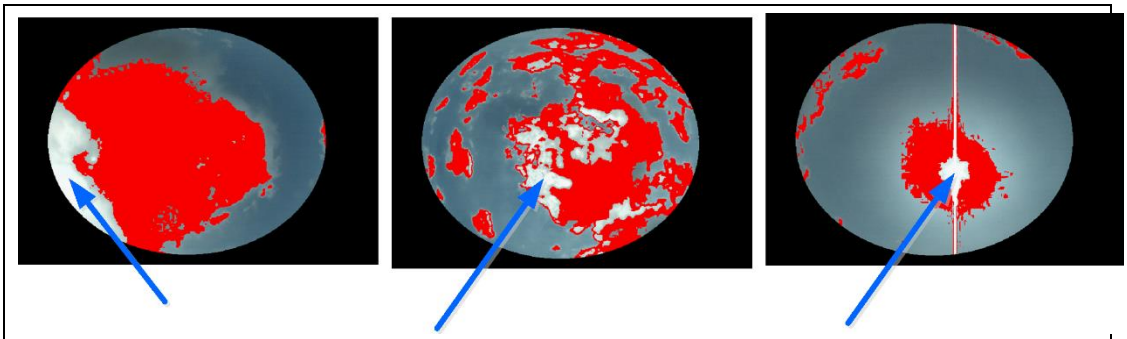
Dear Drs. Cheng and Lin, Thank you for allowing us to view your revised transcript, especially with the clearly marked red font indicating changes. This is a high quality paper, and I appreciate the chance to read and provide comments to this work prior to publication. I have one substantive comment, two requests for further clarification, and a number of very small comments about the text itself. All page numbers and line numbers refer to the "Revised Manuscript" submitted on Oct 27th.

Thank you for the positive comment. All comments are carefully addressed in this document.

1) Substantive comment. On Page 13, (lines 280-283) you discuss the choice of T hr_upper and Thr_lower when performing the RBR classification. It seems that any pixels that have RBR > Thr_upper are automatically marked as a non-cloud pixel. If so, this is incorrect. The pixels with RBR > Thr_upper should be removed from the analysis. This explains the results you show in Figure 4. If you decrease Thr_upper, then FN (the number of cloudy pixels incorrectly classified as non-cloud) will increase. This explains the dramatic reduction in Recall that is depicted in Figure 4 as the Thr_upper decreases from 1 to 0.9. It also is apparent in Figure 9c, where the RBR appears to capture the clouds quite well, except for the very bright regions where it fails, due to Thr_upper. Finally, upon reading more closely, it appears that you are treating the sun region and vertical lines in a similar fashion. On Line 127, you declare that pixels in the Hough line region or sun region "are determined as non-cloud pixels." Considering the sample image in Figure 2d, this also seems erroneous: these pixels should not be set to non-cloud. They should simply be removed from the analysis as contaminated pixels. This substantive comment certainly needs to be addressed, and may affect the results of the figures, since it should substantively change the Recall metric of the RBR procedure.

The reviewer stated that “The pixels with RBR > Thr_upper should be removed from the analysis.” Does it mean that these pixels should not be considered when calculating detection accuracy, precision, and recall?
However, in the works of cloud detection, all the pixels in an image are desired to be classified either as cloud or non-cloud. It is inappropriate to remove pixels from the analysis. The fact is, pixels with RBR> Thr_upper could be clouds in some images (as can be seen in the first and second row in Fig.9 (c)), and they could be non-cloud pixels in other images (the third row in Fig.9 (c)). It is a desired to distinguish these cloud and non-cloud pixels. We cannot say that we do not analyze the pixels with RBR > Thr_upper.
It is normal to have trade-off between precision and recall. As the thresholds become strict, it is normal that precision increases and recall drops.



The point is, RBR thresholding cannot distinguish if the pixels with RBR > Thr_upper are clouds or sun. Therefore, we propose to use sun region detection to distinguish them automatically. The automatic analysis procedure of Hough line region or sun region detection is what we propose to distinguish these pixels correctly. Otherwise, the computer have no knowledge if it belongs to sun or clouds. Therefore, for the comment that the pixels in the Hough line region or sun region should not be set to non-cloud and they should simply be removed from the analysis as contaminated pixels, we cannot agree with it. As we mentioned, all the pixels in an image are desired to be classified either as cloud or non-cloud automatically in research works of cloud detection.

1b) Note that point (r) below is also a substantive comment. What is the minimum spatial extent of a cloud that your method can resolve (in pixels). How does this affect your analysis?

The minimum spatial resolution for analysis is a pixel. The cloud detection is performed pixel by pixel. And the accuracy analysis is performed based on pixels.

2) First clarification request. I am not able to understand some aspects of Figure 3, and the text that describes it in lines 227-239. I do not understand what is the 3x3x3 neighborhood around a pixel p at level i. At first I wondered if these were the three color channels, so that Figure 3a shows the red, green, and blue channels stacked on top of each other for a 3x3 pixel region. However, if this is the case then the three levels in Figure 3b, showing a 5x5 region, 7x7 region, and 9x9 region don't make sense to me.

No, it is not referring to the three color channels. As stated in the manuscript, the 3×3 neighborhood is around a pixel *p* at level *i*, its previous level $i-1$ and its next level $i+1$. Note that the actual sizes of the image patches used to extract features are 5×5, 10×10, 15×15, 20×20, 25×25 as stated in the

experiments. But the numbers are too big to draw actual sizes of the image patches in the illustration figure.

Therefore, in line 233, I don't understand why a pixel would have 27 x 4 votes. If I try to understand this, it seems that at level i, with Li = 7, we actually consider the 21 x 21 pixel region around our pixel of interest: a 3x3 grid, where each grid has 7x7 pixels in it. Then all four of your classifiers are run on each of those 7x7 regions, and is determined as cloud or non-cloud for each classifier. Then it would be possible to add votes.

Please let me clarify this. For example, for each pixel at Li=10, we extract its feature vector using image patches of size 10x10. The size of the image patch only affects the dimensionality of the feature vector (before dimension reduction). As stated in 2.3, the dimension of the feature vector is $L_i \times L_i \times 10$ (all the color components and the red to blue ratio (RBR) of all the pixels in the patch are concatenated to form a feature vector.).

We can regard it as we create ℓ (denoting the total number of levels) images. The characteristics (features) of each pixel in an image of each level is represented by a feature vector (constructing using the patch size at its own level). And each pixel will have the classification results from the four classifiers (using its own feature vector). But to classify a pixel, it does not trust only the classification results using its own level. It tries to see the results from its neighbors.

As illustrated in Fig. 3, we consider the classifier results of itself and its 26 neighbors in the 3x3 regions at the current and adjacent levels. Therefore, there are 27 x 4 votes.

I think the wording that threw me off was "3x3 neighborhood" and "3x3x3 neighborhood," because these aren't really neighborhoods. I would suggest using "neighboring image patches" instead. Possibly adding light grey grid lines in the 3x3x3 squares in Figure 3a would help the reader understand that you are considering the image patch and all neighboring image patches.

The concept of multi-scale neighborhood in scale space is inspired by the works of scale invariant feature point detection. And the term neighborhood in scale space is commonly used in the image processing field. We have added two references in the revised manuscript. And we also added some explanations in the revised manuscript to avoid misunderstanding.

"We consider the classifier results of a pixel itself and its 26 neighbors in the 3x3 regions at the current and adjacent levels."
"The voting is performed in a multi-scale neighborhood, which is inspired by the works of Lowe (2004) and Bay et al. (2008). "
- Bay, H, Ess, A., Tuytelaars, T., Gool, L. V., 2008. SURF: Speeded Up Robust Features. Computer Vision and Image Understanding 110, 346-359.
- Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91-110.

3) Second clarification request. I also struggle to understand the rationale and math in lines 241-252, regarding P(x) ∈ cloud | Numi=1∼l(xleveli ∈ cloud) If the Latex formatter is not working, this is meant to be the left hand side of Eq. (4). I think lines 241-252 can be made much more clear.
If we are using 5 levels of classification, then presumably each pixel in the image will have 5 resulting decisions if it is a cloud or non-cloud. (If I understand Figure 3 correctly, then level i and level l don't make sense to me, because you can't have i − 1 and i + 1 in these two cases, respectively.) There must be something simple you do to obtain your final determination if this pixels is a cloud, given the results of the 5 levels of determination.

For the boundary conditions, when performing voting for pixels at level 1 and ℓ, as long as the total votes exceeds threshold (N_v=57), the pixel is still classified as cloud as that level in our implementation. At level 1, there is no level i − 1. At level ℓ, there is no level i+1. So there are 18*4=72 votes for the pixels at these two levels. We have added this explanation in the revised manuscript.

$\underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud)$ can be 0 (the pixel is not classified as clouds in any level) to ℓ (the pixel is classified as clouds in all levels). Therefore,

$P(x \in cloud \mid \underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud))$ represents the probability of a pixel belonging to cloud given the information $\underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud)$ = 0, 1, … , or ℓ. Since $P(x \in cloud \mid \underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud))$ cannot be directly obtained, it is re-expressed using Eq. (4) according to Bayesian rules of conditional probability.

By setting the prior probabilities $P(x \in cloud)$ and $P(x \in noncloud)$ as e

equal. The determining factor is $P(\underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud) \mid x \in cloud)$. As stated in

the manuscript, the likelihood term $P(\underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud) \mid x \in cloud)$ is learned

from the training dataset by constructing the normalized histogram of

$\underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud)$ using all ground truth cloud pixels. For example, if there

are 1000 cloud samples, and 200 of them have $\underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud)$ =1, then

$P(\underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud) = 1 \mid x \in cloud)$ =200/1000=0.2. If there are 1000 non-cloud

samples, and 500 of them have $\underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud)$ =1, then

$P(\underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud) = 1 \mid x \in noncloud)$ =500/1000=0.5. In this example,

$P(\underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud) = 1 \mid x \in noncloud)$ is larger than

$P(\underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud) = 1 \mid x \in cloud)$, which implies that

$P(x \in noncloud \mid \underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud) = 1)$ is larger than

$P(x \in cloud \mid \underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud) = 1)$. According to this, we should classify a pixel

with $\underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud)$ =1 as non-cloud.

Finally, here are some minor comments on the text.

a) line 39: please refrain from the use of "etc." in manuscripts.

b) line 50: "cloud" should be "could."

c) When citing the authors of a paper in a sentence, simply use a format like "The work by Long et al. (2006) suggested that ..." This recurs throughout the manuscript on lines 57, 66, 70, 73, and 201.

d) The review of literature of past attempts to detecting, classifying, and tracking clouds is quite nice, thank you. It seems that you should also consider the work by Zhuo et al. (2014), Isosalo et al. (2007), and Liu et al. (2015), all of whom considered

multiple spatial interrogations or segmented spatial regions of images as well.
e) Figure 1 is a helpful schematic. It seems that the training and manual labelling of the 250 images should factor into your system framework somewhere. Could you possibly add this? I would imagine a separate box that points into the "Classification via Multiple Classifiers" dashed box.

Thank you for the comments. We have revised the manuscript according to the comments. However, we are not sure which paper Liu et al. (2015) refers to. Instead, we found a work by Liu et al. which was published in Journal of Atmospheric and Oceanic Technology 2011 (Available from: Lei Liu, Feb 21, 2015) https://www.researchgate.net/publication/249605113_Cloud_Classification_Based_on_Structure_Features_of_Infrared_Images

Liu, L., Sun, X., Chen, F., Zhao, S., Gao, T., 2011. Cloud Classification Based on Structure Features of Infrared Images, Journal of Atmospheric and Oceanic Technology 28, 410-417.

f) On Line 149 you say that including multiple color spaces "provides the classi- fier more information that is beneficial to performing classification." Do you have any evidence to support this?

We have added the classification accuracy of using only single RGB color model as the feature vector in Fig. 5. The following explanation is added to the revised manuscript. "We also compare with the classification accuracy of using only single RGB color model as the feature vector to validate that adding other color models in the feature vector can yield better clasification results."

g) Line 154, change "the feature" to "each feature" for more clarity.
h) Line 171 change to REigenvalue for consistency with Eq. (1)
i) Line 179 change "And the tree is built recursively." To "The tree is then built recursively."

Thank you for the comments. We have revised the manuscript according to the comments.

j) In Eq. (3), what is the p in the denominator of the equation?

In Eq. (3), *p* is the number of dimensionality of *x* and μ_k, i.e. $x \in \Re^p$ and $\mu_k \in \Re^p$. We have added the explanation in the revised manuscript.

k) Line 234, it's not clear to me why we are referring to the feature vector, xLeveli .

Haven't the classifiers already independently determined cloud or non-cloud status for the feature vectors? I may be missing something here. I believe elsewhere you refer to this as pixel p at level i, can you just refer to the number of votes for this pixel as Vcloud(pi)?

We have revised the sentence. "Let $V_{cloud}(x_{Level_i})$ denotes the number of votes in the neighborhood classified as cloud for pixel *x* at level *i*."

l) Line 241, what is x (the second character in the line, in the P(x ∈ cloud . . . equation)?

We have revised the sentence. " $P(x \in cloud \mid \underset{i=1\sim\ell}{Num}(x_{Level_i} \in cloud))$ states the probability of pixel *x* belonging to cloud given the number of levels that the pixel is determined as cloud."

m) Line 264. Please indicate the geographical area where the images were collected.
n) Line 275. RGR should be RBR.

Thank you for the comments. We have revised the manuscript according to the comments.

o) In your discussion of Figure 4, can you please give a reason for the trends that we see? For example, Why does Accuracy increase as T hrlower increases? Why does Recall decrease as T hrlower increases?

We have added the following descriptions.
"In Fig. 4, we can observe the trade-off between precision and recall. As the thresholds become stricter, the precision increases and the recall drops. Precision rate and recall rate cannot be used alone to measure the accuracy since precision does not consider false negatives and recall does not consider false positives. Therefore accuracy defined in Eq. (5) is used as the conclusive metric to measure the performance."

p) Line 299. You state that T hrlower = 0.8 and T hrupper = 0.9 have the lowest detection accuracy (note the misspelled "detecction" Please use spell-check.) I presume you are only looking at the blue bar line. RBR's detection accuracy is about 0.78, while the 3 other methods appear to be 0.8 to 0.83. Is this really significant? Is there anything worth mentioning about the Precision and Recall? If not, why are they

in the figure? A simple sentence explaining why you choose Accuracy as your conclusive metric against the others would be appreciated.

Since precision and recall are not especially worth mentioned in Fig.5, we change Fig. 5 to display only accuracy. The following explanations are added when explaining Fig. 4.
"Precision rate and recall rate cannot be used alone to measure the accuracy since precision does not consider false negatives and recall does not consider false positives. Therefore accuracy defined in Eq. (5) is used as the conclusive metric to measure the performance."

q) Line 306. Note that the notation is not consistent with Line 152. I'm OK with this, although it's a bit sloppy.

We have revised the notation so that they are consistent.

r) Why don't you use L1 = 1 as the smallest level? It appears that you would have a small-scale limit on the size of clouds that you could resolve. What is that limit? Could you comment? You compare these results to other single-pixel results (e.g. Fig. 10). It seems to me that we can't quite compare them, since this method always operates on at least a 5x5 grid region. (And, in fact, the 8 neighboring 5x5 regions as well? So the small scale limitation could be quite substantial!)

Note that using the image patch of L_ixL_i to construct the feature vector does not mean classifying the entire L_ixL_i area as the same label. For each pixel, we construct its own feature vector. And each pixel is classified individually. Therefore, the classification is performed on each pixel. It is just that the feature vector of a pixel includes the information from its local image patch.

s) Line 308. Change "using feature" to "using the feature"
t) Line 309. This sentence didn't make sense to me. I suggest, "The value of T hrP CA is typically between 90
u) Line 314. Why do you say "cross-validated"? What does that mean?
v) In the caption to Fig. 6, refer the reader to Eq. 5.

Thank you for the comments. We have revised the manuscript according to the comments. Ten-fold cross validation means that the dataset is divided into 10 none-overlapping subsets. Nine subsets are used for training, and the remaining one subset is used for testing. Then the training subsets and testing subsets are rotated for 10 times. The average classification rate of these 10 experiments is the 10-fold cross validated accuracy.

w) Line 374. You say “fixed and dynamic RBR thresholding.” I don’t see that you did anything other than fixed thresholding, using 2 fixed threshold values of 0.8 and 0.9.

We also compared with HYTA (Li et al., 2011), which uses dynamic thresholding.

Many thanks for this fine manuscript, and your hard work. It is a contribution to our field.