Response to Reviewer 2 comments on Glacier algae accelerate melting on the south-western Greenland Ice Sheet by Cook et al.: Round 2

In this revision, Cook et al. provided more details on methods as suggested by reviewers. The fieldwork and radiative transfer modeling components are very important and valuable for the cryosphere community to understand better the impurity properties. However, some key/core information/figures are still lacking for the section using UAV and Sentinel-2 remote sensing data assisted by ASD field spectra to conduct image classification. In addition, there are some contradictory and inconsistent arguments throughout the text. These problems need to be resolved:

We thank the reviewer for the second set of comments. We have provided a point-by-point response below:

1) I asked the authors to provide the ASD spectra (resampled to UAV bands and Sentinel-2 bands) compared against the real UAV and Sentinel-2 spectra, but I didn't see them in this revision. I think the authors should at least provide a figure showing how consistent the ASD resampled spectra are with the UAV and Sentinel-2 spectra. This is very important for the image classification, because UAV spectra and Sentinel-2 spectra have uncertainties of atmospheric correction and other radiation correction procedures. If they don't match very well, using ASD-resampled spectra to train the RF or other classifiers is not convincing. Although the authors claim that the RF method is advanced and novel for studying the cryosphere, it is essentially an image classification problem and subject to basic image classification rules. The authors classified surface types into snow, clean ice, cryoconite, water, low biomass algae, and high biomass algae. Please provide a figure showing the comparison between the ASD reduced spectra and the UAV and Sentinel-2 spectra for each class, and label the bandwidth for each UAV and Sentinel-2 band. If the spectral curves of these classes are highly correlated, the whole classification problem will become a thresholding problem, and the derived 'high biomass algae' will be equivalent to 'dark ice'. Without providing this information (i.e. actual plots), the classification method is not solid, and the classification results are not convincing.

We provide below a series of five plots showing ASD reflectance plotted against UAV reflectance for each of the five UAV bands (Figure 1: UAV on the y-axis and ASD on the x-axis). These meaurements were made for the points in our UAV imagery where measurements were also made using the ASD field spectrometer. The blue markers show the ASD spectra plotted against uncorrected UAV spectra. The red markers show the same data after a systematic offset was corrected in order to calibrate the UAV against the ASD field spectra (the magnitude of the offset is given in the header of each plot).

From these plots it is clear that following subtraction of the systematic offset from each UAV band there is a close match between the resampled ASD and UAV reflectance that supports the validity of our classifier. We cannot co-locate centimeter scale ASD measurements on the 20 m resolution Sentinel-2 imagery so the *equivalent plots cannot be generated for ASD vs Sentinel-2. We also direct the reviewer to Figure 5C in Tedstone et al. (in review: https://www.the-cryosphere-discuss.net/tc-2019-131/tc-2019-131.pdf, copied below as Figure 2) where the albedo measured by Sentinel-2 and our offset-corrected UAV are compared for several flights made during the same field season at the same field site, with the black crosses representing the S2 albedo and yellow crosses representing UAV albedo.*

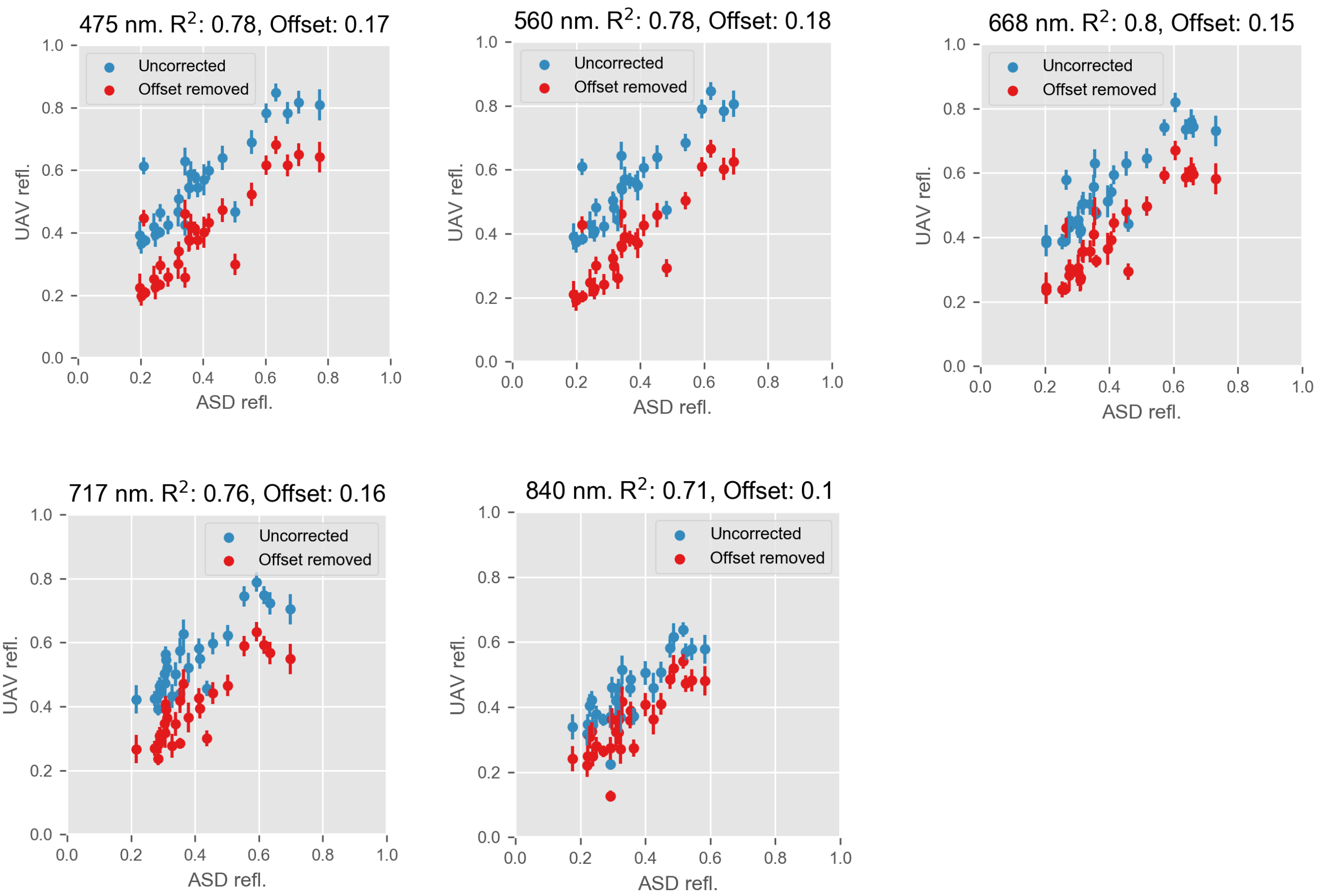

Figure 1: UAV reflectance plotted against ASD reflectance for each UAV band, in uncorrected (blue) and corrected (red) form.

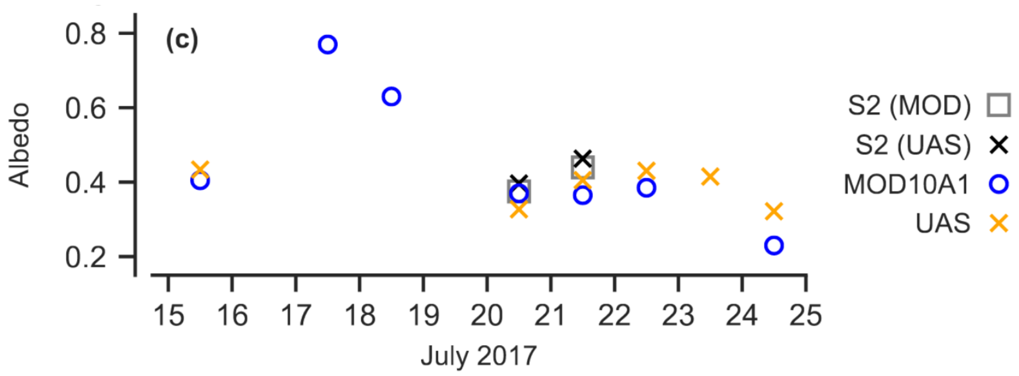

Figure 2: UAV and S2 albedo for days with coincident UAV flights and S2 overpasses during our 2017 field season from Tedstone et al. (in review).

As requested, we also provide a comparison of the reflectance values for each surface class in each waveband (Figure 3 and Figure 4). In Figure 3 close correspondence between the reflectance values measured by the UAV and the ASD field spectrometer is demonstrated, and there is clear separation between the classes, supporting the validity of the classifier. Figure 4 shows that there is a weaker correspondence between the reflectance measured by the ASD and Sentinel-2, but there is generally still separation of the classes. However, there is poor separation in the HA class at the short visible wavelengths. We highlight two key features of our approach:

1. The ambiguity in the short visible wavelengths demonstrates why information from all nine bands is required and therefore why our classifier is preferable to a threshold. Our classifier uses information from all 9 bands to separate the classes, which provides a degree of resilience to ambiguity in the short visible wavelengths, whereas a simple thresholding approach would be more likely to confuse the classes.

2. Regarding the Sentinel-2 classifier specifically, whilst in principle there is potential for misclassification of LA as HA, this is unlikely in practise because HA patches, as shown by our UAV observations, are not spatially expansive over metres+ scales and therefore all HA surfaces detected by Sentinel-2 are mixed with other, brighter surface types. The Sentinel-2 classifier therefore remains conservative. In the cases where HA is identified in Sentinel-2 imagery, we would still expect mean reflectance of these pixels to be higher than in the UAV and ASD spectra because the area-averaged reflectance is pushed up by mixing of HA with brighter surrounding surfaces, and indeed we can see this is the case in Figure 4 – the darkest end-members found at the 30 cm ASD scale are simply not found at the 20 m Sentinel-2 scale.

We have discussed this scale issue in the manuscript, and now add Figures 1, 3 and 4 to Supp Info 6. We also direct the reviewer to Tedstone et al (in review:https://www.the-cryosphere-discuss.net/tc-2019-131/) where these issues are discussed in more detail, including further evidence for coarser spatial scales leading to higher algal detection limits (i.e. a conservative algal classifier).

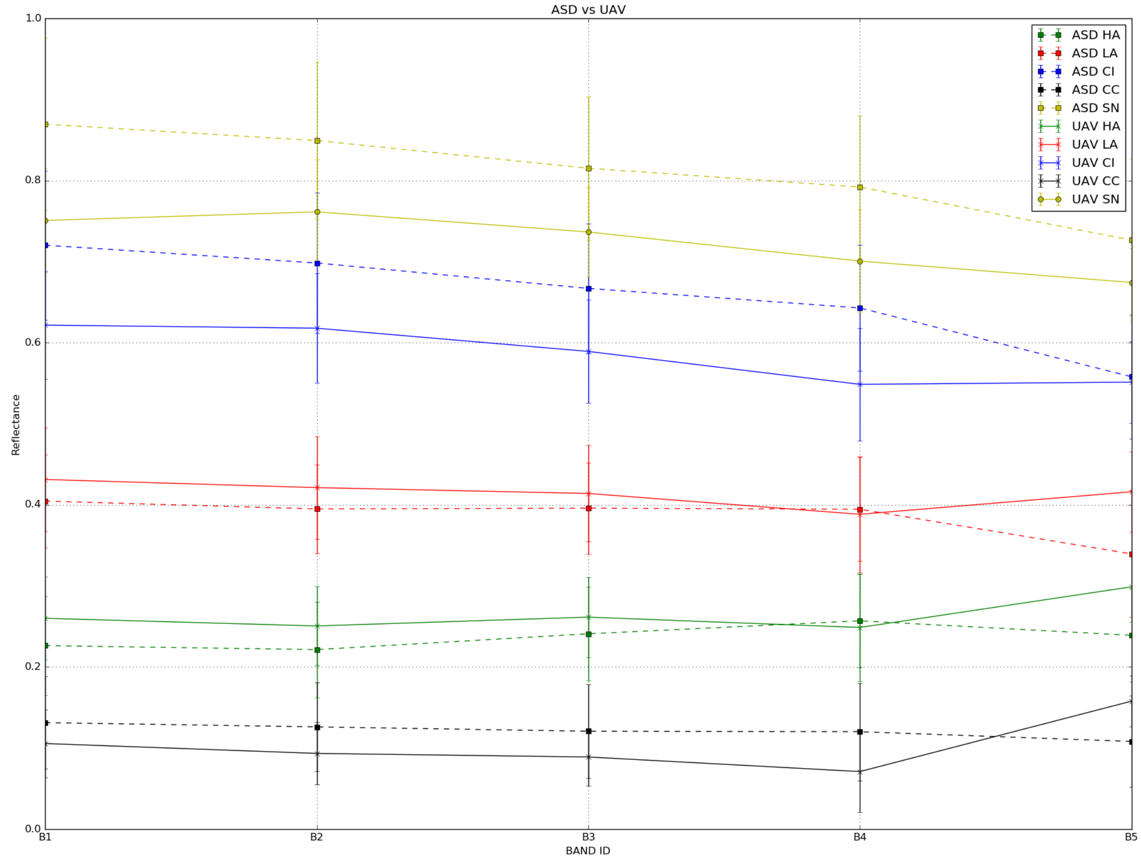

Figure 3: Comparison of mean reflectance values for each surface class from the ASD and UAV data.

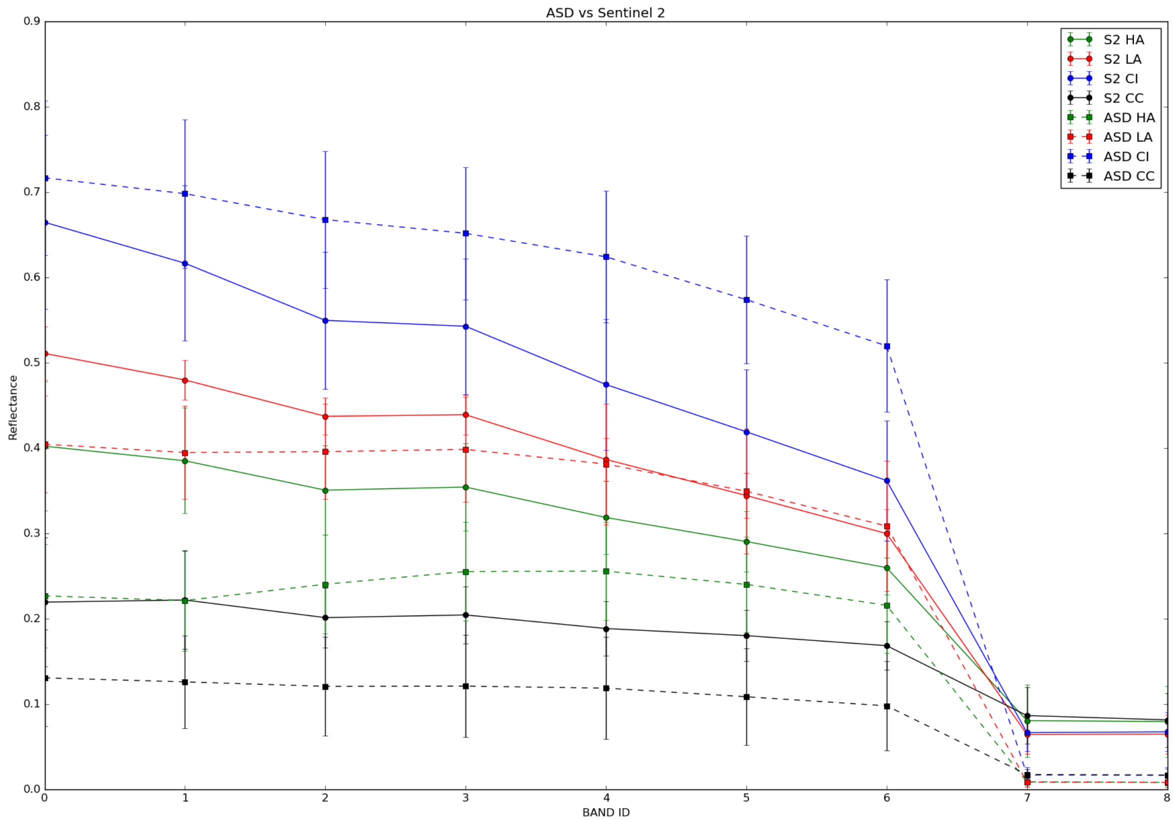

Figure 4: Comparison of mean reflectance values for each surface class from the ASD and Seninel-2 data.

2) The arguments regarding glacier algae and red mineral dust are contradictory. In the introduction section, when the authors mentioned the recent study by Wang et al. (2018) who proposed to use the reflectance ratio between 709 nm and 673 nm to map ice algae in southwest Greenland, it is claimed that 'the red-edge is potentially vulnerable to false positives due to red mineral dusts (Seager et al. 2005; Cook et al. 2017b)'. On page 8 line 260-265, the authors mentioned again that red dust has 'similarly shaped spectral albedo to glacier algae'. However, throughout the paper and the two referenced papers, I didn't find any convincing evidence (i.e. data) showing that the red dust can indeed mask out the chlorophyll signature at red-NIR spectrum. If the red dust is indeed very similar to glacier algae in spectral properties, then no practical methods would exist to separate them regardless of the spatial resolution of the remote sensing data, which means the RF method cannot work as well. The authors claim later in the manuscript (on page 14 line 500-510) that the chlorophyll-a signal between 670 and 710 nm (red edge) is uniquely biological and use this red-edge (Painter et al., 2001 and Wang et al., 2018) to support their results. On page 16 line 550-555, the authors again argue that red minerals occur only in very low concentrations and would have a negligible effect on the ice optics. These arguments are contradictory, with the later arguments supporting the red-edge method used for mapping algae in southwest Greenland. I suggest the authors should properly acknowledge previous studies.

There is precedent in the literature for the red-edge method being confused by red dusts. A famous example is that the red-edge was cited in the 1950s - 1960s as evidence for chlorophyll-containing green plants on the surface of Mars, but of course this was later revealed to be red dusts (see Seager et al. 2005; Sagan and Pollack, 1969: https://www.nature.com/articles/223791a0). Sparks et al. (2009) shows similarity in spectral edges between vegetation and iron oxide (https://www.pnas.org/content/pnas/early/2009/04/28/0810215106.full.pdf) that they state could lead to false positive bio-detection in a red-edge study.

The popularity of the red-edge method arises from its widespread use in vegetation mapping where low spatial and spectral resolution sensors can be applied with confidence because there is an abundance of a priori knowledge about the surface reflectance and end-member spectral libraries. This is not true in our case. It is only with our field samples and our radiative transfer model that we can now discount the effects of red dusts upon ice surface darkening in this region of the GrIS. Previous to our study, there has been scarce evidence to support the application of the red-edge method for detecting glacier algae. While the red-edge method does a fair job at classifying the ice surface in our case, it is outperformed by our RF classifier. The feature importances are highest at blue and green wavelengths for our classifier, suggesting that the red-edge classifier overlooks important spectral information for classifying the ice surface. We therefore see no reason to employ a lesser method when we have a better-performing one available to us.

Where we describe the algal and red-mineral spectra as similar we mean that both of these light-absorbing particles preferentially absorb light in the short visible wavelengths. Sensors with low spectral resolution that cannot resolve uniquely biological features such as the 680nm absorption feature or the "chlorophyll bump" at 550 nm, or the "flattened" spectrum across the blue-green wavelengths in the case of glacier-algae will struggle to separate them. However, at our field site we have shown that in this study the low abundance of red mineral phases means the red-edge method is applicable as a rudimentary life-detector. However, our RF classifier is more accurate as it uses information from additional wavelengths as well as the red-NIR. The red-edge is simply a presence/absence detector and is unable to separate the ice surface into multiple classes as our RF classifier does. The red-edge also relies on sufficient spectral resolution in the red-NIR part of the spectrum that is not available on many sensors, whereas our method of resampling hyperspectral data can be applied to any sensor by retraining on new subsets of the spectra. Furthermore again, with highly variable ice albedo that in itself can range from <0.3 to >0.7, the red-edge can be dampened or deepened independently of changes to the algal abundance – classifiers sensitive to shorter wavelengths are more robust to this as the majority of the light absorption by algae occurs at shorter visible wavelengths. Therefore, there are several compelling reasons to use our classifier over a simple red-edge classifier. We also note that the 680 nm feature is not a "red-edge". Rather, it is a discrete absorption feature centred on 680 nm whose presence is diagnostic for chlorophyll-a and whose area scales with chlorophyll-a concentration (Painter et al. 2001). It can only be reliably applied to hyperspectral data as the method relies upon interpolation between the shoulders of the feature and the calculation of the area beneath this interpolated line. We note that we did not describe the red-edge as "uniquely biological" - we only described the 680 nm absorption feature as uniquely biological. For the red-edge we simply state that is it widely used to detect green vegetation in other environments.

In response to the reviewer comments we have adjusted our description of the previous research in the revised manuscript (lines 100-110) so that it now reads:

"Recently, Wang et al. (2018) applied the vegetation red-edge (difference in reflectance between 673 and 709 nm) to map glacier algae over the south-western GrIS using Sentinel-3 OLCI data at 300 m ground resolution."

3) In the response, the authors said 'We attempted to use the vegetation red-edge on our spectra and achieved only ~80% accuracy for identifying algal ice compared to our random forest classifier that achieves >95% accuracy.' How many samples were used to calculate the accuracy? Were the training samples also used to obtain the 95% accuracy? If they were, it is not a fair comparison. Actually, for remote sensing classification problems, 80% accuracy is quite high. Why not use this information to cross-validate your method? It seems to me that the red-edge method actually supports your classification results.

We agree with the reviewer that the red-edge method does offer broad support for our classification method when applied to the full-resolution ASD spectra. It was applied to all of our algal spectra.
The RF classifier was also applied to all of our algal spectra, divided into the training set and test set. The ratio of training to test samples was 80:20, not 70:30 as was accidentally reported in the previous version of the manuscript, giving 45 samples in the test set. The performance of the classifier on the training and test sets were similar, suggesting the trained model generalises well to unseen instances. The accuracy, precision, recall and F1 scores for the RF classifier and several other classifiers that were trained for comparison are already available in Supp Info 5 for both the UAV and Sentinel-2 classifiers.

4) Page 7 line 227-228, please show simulation results in supplementary information.
We have added six simulations using different cell sizes to our supplementary information (Supp Info 2). The spectral and broadband albedo resulting from 10 – 1000 µg/g in the upper layer are presented as inFigure 5 below:

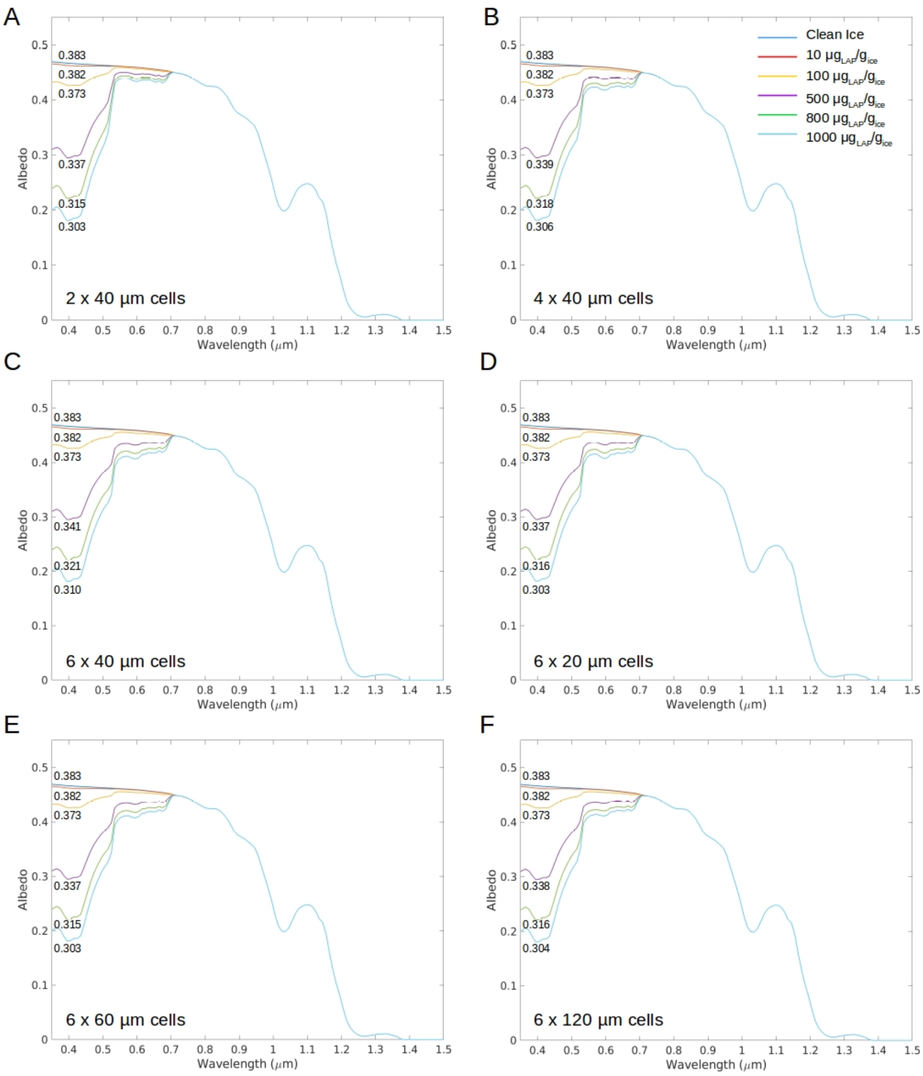

Figure 5: Spectral albedo for ice surfaces loaded with algae in mass-mixing ratios between 10 – 1000 µg/g. Broadband albedo are printed below each curve.

5) Page 7 line 376, does this mean that there are 231*0.7=162 training samples and 231*0.3=69 test samples? Were all accuracy tests based on 69 samples? In this case, how many samples for each class were available to test the classifiers?

The training-test split was actually 80:20 in the revised manuscript, we apologise for missing this update to the text. The sample instances in the test-set were as follows: Cryoconite: 6, Clean Ice: 5, Heavy Algae: 16, Light Algae: 7, Snow: 5, Water: 6. This gives a total of 45 instances (20% of 231).

6) Page 11 line 387-390, I don't quite understand the logic here, are the authors trying to say that they want to build an endmember library using ASD spectra? How can it be certain that using 'pure' spectra can get better accuracy for classifying images (with spatial heterogeneity within each pixel) rather than directly labeling the images (to obtain training sets)?

The reviewer is correct that we have built an end-member spectral library using ASD spectra. We consider this to be the only justifiable way to classify the ice surface. We deliberately avoided direct image-labelling because the labels in the training set would simply be guesses at the surface composition. Labelling images in this way may produce more data, but of lower quality since there will necessarily be label ambiguity. We measured the spectral reflectance of homogenous areas and analysed the ice composition in the laboratory, so we are sure of our labels. The field-of-view of the spectrometer was small so we can be confident that the spectra are representative of a single surface class rather than an area containing multiple surface types. The price of this is reduced data abundance as the reflectance data acquisition and sample processing is expensive and time-consuming, but the benefit is high data quality. With the current scarcity of knowledge/data on light absorbing particles and the highly variable physical structure of the ice in this area, directly-labelling images would suffer from high label uncertainty. We discuss these issues in lines 393 – 399.

7) Supp info 6: this figure is not easy to read, please directly give the correlation numbers in a table format, along with the number of test samples.

Unfortunately the specific instances used to train and test the Sentinel-2 classifier were not saved when we saved the model, the summary metrics and the figure, so we are unable to provide those values. However, we have provided the overall performance metrics (accuracy, precision, recall, F1: Supp Info 5) and the confusion matrix plots that clearly demonstrate that the classifier performs well and generalises well to new instances. An 80:20 ratio of training:test samples was used, meaning 45 instances were available in the test set, and multiple instances were available in the test set for every class.

8) I’m confused by your study area. On page 12 line 420, “The area of interest was the “common area” defined by Tedstone et al. (2017) bounded within the latitudinal range 65 – 70 N, and is equal to that used by Wang et al. (2018)”. Figure 1 and Figure 5 show an area between 67.615N and 67.599N, which is much smaller than 65 – 70 N. Which one is the study area? It seems that the UAV and Sentinel-2 only cover a very small portion.

The UAV image area was a 200 x 200 m area which was located inside a larger Sentinel-2 tile (65.615 – 67.599 N). We upscaled to 65-70N in our runoff modelling.We have clarified further in our methodology (lines 133-135 in the revised manuscript).

9) Section 2.10 and 3.7. Thanks to the authors for providing more details here. However, this part is still not very clear to me. When the authors generalized the runoff estimate (simulated based on three weather stations) over the entire region (65 – 70 N?), are the area percentages of low biomass algae and high biomass algae based on classification results over the 200*200 m UAV region and the 67.615-67.599N Sentinel-2 region to generalize to the entire region based on the bare ice area and dark ice area derived from MODIS data (500 m resolution)? If so, did the authors consider issues associated with different spatial scales? The authors pointed out that ‘The higher detection limit for algae with decreasing ground resolution makes our estimate of spatial coverage from Sentinel-2 conservative. We highlight that this will have a much larger effect on studies aiming to quantify cell abundance using Sentinel-3 where the ground resolution is 300 m.’ (page 19 line 669). Table 3 shows the area percentage of different classes, total algae is 78.5% on UAV, 57.99% on Sentinel-2 (2016) and 58.87% on Sentinel-2 (2017). Is this the evidence that supports the argument that Sentinel-2 estimate is conservative? If so, this doesn’t make sense since totally different areas are being compared (200*200m vs 10000*10000m). Please clarify. Given the spatial scale issues, it is incorrect to directly extrapolate the area percentage from centimeter resolution and 20m resolution to MODIS 500 m resolution. Please justify the methods.

We have provided a runoff range forced by our upper and lower estimate of algal surface coverage. The upper estimate is from our classified UAV image. We specifically caveat in the manuscript that this is an upper estimate because we cannot be sure that the coverage is representative across the upscaling area. The lower bound is from our classified Sentinel-2 image. Again, we specifically caveat that this is a lower bound because the coarser spatial scale introduces surface heterogeneity as discussed in the response to earlier questions, which makes detection of algae-laden surfaces at coarser scales increasingly conservative. This is true both at the scale of individual pixels (since patch sizes are generally much smaller than 20m) and at the scale of the entire image (since ice from outside the dark zone is included in the 100 x 100 km tile). These likely explain the differences in coverage presented in our Table 3. This is explained in section 3.7 as follows:

Line 680: "At the same time, our Sentinel-2 remote-sensing underestimates algal coverage because H_{bio} patches are often too small to be resolved at 20 m pixel resolution (Tedstone et al. in review). Therefore, we used the spatial coverage determined by our Sentinel-2 classification as a lower bound, and spatial coverage determined by our UAV classification as an upper bound on our estimate of total runoff attributed to the presence of algae."

The reviewer is correct that we have upscaled to the 65-70N ("south-western") region. We agreed with the reviewer in the initial round of review comments that upscaling to the entire western coast of the GrIS was excessive given the field measurements were all from the Kangerlussuaq region; however, upscaling over the south-western GrIS is justifiable, especially since there is literature precedent for upscaling over this region. We have delineated the extent of the dark zone using MODIS using protocols from the past-literature (Tedstone et al. 2017;Shimada et al. 2016) and applied our upper and lower algal coverage estimates to force the runoff model, over an area where we expect our field measurements and observations to be representative. We have therefore applied our classifier to UAV and Sentinel-2 images to determine realistic coverage estimates and applied them over the dark zone as defined using MODIS thresholds consistent with previous studies. We also point to the paper currently in open review in The Cryosphere by Tedstone et al. (https://www.the-cryosphere-discuss.net/tc-2019-131/), where inter-sensor comprisons and cross-scale comparisons between our field, UAV and satellite remote sensing data are quantified and examined in detail.

We have therefore considered the issues raised in the comment above and mitigated them by providing a range of estimates.