"An inventory of supraglacial lakes and channels across the West Antarctic Ice Sheet" by Diarmuid Corr et al.
Response to Referees

We thank the referees for the time and effort they have dedicated to reviewing our manuscript, and are grateful for their constructive comments, which have increased the quality of this manuscript substantially. We address each of the referee's comments in turn, with our responses given in blue text. In addition to improvements from the referees, we have addressed further issues throughout the manuscript, as is evident in the mark-up text.

Referee 1: https://doi.org/10.5194/essd-2021-257-RC1
Corr et al. present the first inventory of supraglacial lakes and channels across the West Antarctic Ice Sheet. The data quality and data presentation are excellent in the linked Zenodo repository (Corr et al., 2021). This paper is an excellent pair with Stokes et al. (2019), and, together, the papers provide a baseline dataset of supraglacial hydrology across the Antarctic (as highlighted in the author's Conclusion). The figures are clear overall, and I just have a few suggestions on some rough locations in the manuscript.

Thank you very much for this review. The authors agree with all suggestions and agree that these will improve the quality of the paper. Further details are provided in response to the specific comments below.

Specific Comments
L12: The sentence that begins with percentage is a comma splice. Suggest a re-write to not begin the sentence with a number and to split this sentence in two.
As requested, we have rewritten this sentence to improve the readability. The sentence now reads:

'We found 27.3% of feature area on grounded ice and 54.9% on floating ice shelves. In total, 17.8% of feature area crossed the grounding line.' (Page 1, Lines 12-13).

L34: Cite this inventory for the EAIS.
We have cited the appropriate paper (Stokes et al., 2019) as the inventory is downloaded from a link in the supplementary information (detailed in Stokes et al., 2019), and therefore no inventory citation has been included. (Page 3, Line 32).

Figure 1: Delineate WAIS and the EAIS on Figure 1.
We have delineated WAIS and EAIS on Figure 1 by adding the approximate boundary between the ice sheets that we have used in mapping. (Page 2, Figure 1).

Introduction: This section could use proofreading to identify a few typos here and there. Several sentences could be shortened to increase writing clarity (e.g., run-on sentence from L61–65).

We have proofread this section and the remainder of the paper, editing typos, and increasing the clarity of writing throughout. We have amended run-on sentences (with original line numbers given first and new line numbers in brackets at the end):

Original L22-25 now reads: 'The configuration of the supraglacial hydrological network is transient. It is determined both by the surface topography and the amount of water in the system; greater melt, for example, is likely to lead to deeper and more extensive lakes and channels (Tedesco et al., 2012; Luthje et al., 2006; Bell et al., 2018).' (Page 2, Line 21-23).

Original L57-59 now reads: 'Indeed, a recent study has shown evidence of five glaciers on the Antarctic Peninsula (Drygalski, Hektoria, Jorum, Crane and Cayley) undergoing near-synchronous speed-up events in March 2017, November 2017 and March 2018 (Tuckett et al., 2019). This suggests the surface meltwater may have entered the subglacial hydrological system.' (Page 3, Line 54-56).

Original L61–65 now reads: 'Supraglacial hydrology may exert a large effect on Antarctica's future evolution. For example, the UN Paris Agreement's limit on the rise in global temperatures of 1.5◦C (https://unfccc.int/sites/default/files/english_paris_agreement.pdf) will likely cause the Antarctic Peninsula to experience irreversible, dramatic change to glacial, terrestrial and ocean systems (Siegert et al., 2019). Under this warming (1.5◦C), ice shelves will experience a continued increase in meltwater production, and meltwater will therefore become more extensive (Siegert et al., 2019).' (Page 3, Line 58-62).

For the correction of various typos and changes see mark-up document.

Section 2.1.4. Lake vs channel classification: It would be great to have an additional figure that shows how the shape index metrics are used within the K-Means clustering approach. The figure cited in this section (Figure 4) only shows supraglacial channel and lake outlines, but not how the various metrics are combined. Additional details on how the K-Means algorithm was applied communicated via a descriptive figure would help better illustrate the "20 distinct clusters" (L195) identified. What do the top 3 or 4 clusters in the dataset look like? Suggest revise Figure 4 to illustrate the K-Means clustering approach, or add a new figure in addition to Figure 4.

As suggested, we've added examples of complex lakes, ringed lakes, standard small lakes, and thin ribbon lakes to Figure 4, now Figure 10 (which had a larger lake and channel already identified). These examples are from the clusters, which represent different supraglacial features clustered by the K-Means approach. (Page 19, Figure 10).

The caption on Figure 10 now reads:
'Figure 10. Outlines of supraglacial features over RGB composites from S2 and L8 imagery. These outlines demonstrate 6 of the distinct clusters from the K-Means approach. A) A large SGL covering 0.20 km^2 on Hull Glacier (Figure 1); b) complex SGL with area 0.35 km^2 on Bach IS (Figure 1); c) ring lake with area 0.05 km^2 on Bach IS; d) 11 'standard' SGLs on Bach IS with areas ranging from 300 m^2 to 0.02 km^2; e) ribbon lake on GVI IS (Figure 1) which spans 2.6 km and covers 0.19 km^2; and f) discontinuous supraglacial channel spanning 1.3 km and covering 0.04 km^2 near Hull Glacier. RGB composites formed from L8 tile LC08_L1GT_022114_20170111 (a,f) from 11 January2017 and S2 tiles T18DXF_20170129 (b,c,d) from 29 January 2017 and T19DEB_20170103 (e) from 3 January 2017.' (Page 19, Figure 10).

We've added a table with the shape index values corresponding to each feature in Figure 10, which has the caption:
'Table 1. Value of each individual shape index (Equations 5-10) for the feature type defined in Figure 10.' (Page 20, Table 2).

Finally, we've added more explanation in the text around the K-Means Clustering approach, referencing Figure 10 and Table 2. The additional text reads:
'Samples from 6 of the 20 clusters are shown in Figure 10, with the corresponding value for each shape index in Table 2. As expected, ribbon lakes are similar to channels in most metrics, as both take long, narrow forms. However, the A:P (Equation 9) value differs vastly between channels and ribbon SGLs. The values displayed for A:P, Fractal, Reock and W:L (Equations 9, 11, 12 and 14) demonstrate clear differences between channels and all other lake classes. While IPQ and Schwartzberg (Equations 10 and 13) are useful in delineating standard, smaller lakes from channels. Through this method, we identify 10,223 lakes and 255 channels to be present during January 2017 on the WAIS and AP (Figure 10). ' (Page 18, Line 318-323).

L310: Move this equation up closer to where it is cited in the text.
Equation 14, now Equation 8, has been moved as requested. It now follows the text:
'To estimate the volume of water contained within each feature, we use an area-volume (A-V, Equation 8) scaling relationship from literature (Stokes et al., 2019). Based on this relationship, the total volume of meltwater stored in supraglacial lakes and streams is estimated to be 0.085 km^3 across the entire WAIS and AP.' (Page 17, Line 276-279).

Referee 2: https://doi.org/10.5194/essd-2021-257-RC2

The manuscript presents an approach to map supraglacial lakes and channels for the year 2017 across the West Antarctic Ice Sheet on the basis of Sentinel-2 and Landsat-8 data and spectral thresholding. It complements the study of Stokes et al 2019 that used a similar approach to map supraglacial lakes around the margin of the East Antarctic Ice Sheet. The resulting inventory of mapped lakes and channels is suggested as benchmark datasets for Earth system science data. Although the manuscript is generally well written and well presented, I have some major concerns, especially regarding the fit to the SI.

Thank you very much for this review. We appreciate your concerns and have taken steps to remedy each comment accordingly as discussed below. However, we believe the manuscript is a good fit for this Special Issue, which aims to provide a platform to develop and share benchmark datasets. Machine learning algorithms have been employed to map supraglacial hydrological features in Antarctica (Dirscherl et al. 2020; Halberstadt et al. 2020), however no continental scale datasets exist to be used as a benchmark. This manuscript and accompanying datasets (which are a result of extensive manual enhancement and quality control beyond simple thresholding approach) aim to close this gap. We believe the manuscript fits the SI by providing the first large scale, continental (when combined with Stokes et al., 2019) dataset, using the most rigorous approach currently available, which provides new benchmark data to be used as training, validation and testing in further machine learning processes.

The authors have considered all further comments and suggestions and address the specific comments below.

<u>Specific Comments</u>

I don't think that the resulting inventory can be regarded as a benchmark dataset for machine learning. First, the inventory is the result of a modelling procedure (by spectral thresholding) and hence does not present reference data with "ground truth quality", which however would be required for a good benchmark dataset to be useful for algorithm and model comparison.

The authors agree that the results of NDWI thresholding approach alone would not present data with "ground truth quality", however the thresholding approach is just one part of our process. The datasets were enhanced by intensive manual post-processing, carried out by human experts, to ensure the high-quality dataset being established, which is of much higher quality than spectral thresholding alone. We carried out extensive quality control checks, where the output of the manually enhanced thresholding approach was compared to datasets generated manually by three human experts. As noted in the manuscript (Section 2.2. Accuracy Assessment, Pages 9-11, Lines 172-219), the manually post-processed methods resulted in scores of 85.3% and 77.6% for sensitivity, 99.1% and 99.7% for specificity and overall accuracy of 98.7% and 98.3% for Sentinel-2 and Landsat-8 sensors respectively. We believe this presents a dataset at large scale, that is as close to ground truth quality as feasibly possible.

We acknowledge that these aspects were not sufficiently clear within the original manuscript, and so have amended the text to clarify the points raised by the reviewer; specifically, we have added text to explain that the inventory is more detailed than the result of a modelling procedure alone:
'However, identifying supraglacial lake and channel pixels using NDWI alone is insufficient, because slush, rocks, clouds, and shadows can be spectrally similar to water (Moussavi et al.,2020). For this reason, they require additional processing steps to identify and mask these features in each image. Additional, manual post-processing steps described in Section 2.1.3 Post-processing, carried out by human experts, provide data which is of much higher quality than spectral thresholding alone.' (Page 5, Line 109-113)

In addition, systematic, continent-scale inventories of Antarctic supraglacial hydrology are rare, which was one of the primary motivations for this work. Only a single continent-scale study has been done previously, and this only covered East Antarctica (Stokes et al., 2019). Therefore, no continent-wide dataset currently exists for this type of exercise. As in-situ datasets at such large scale are not available, manually enhanced datasets using satellite imagery is the only credible solution for the creation of training data. However, the usage of our dataset in machine learning algorithms is one of many potential applications as is discussed in the manuscript (Page 3, Lines 35-38 and Page 21, Lines 363-365).

'Our inventory provides a baseline for monitoring future changes and also serves as a training/forcing dataset for other studies, such as those focused upon methodological development or climate and glaciological modelling. High quality training data are, for example, a vital component of machine learning methodologies, while accurate observations of melt features can act as both boundary conditions and validation for physical models.'

'Others can use the dataset produced in this study to assist approaches that utilise other types of satellite data, for example, those that exploit Synthetic Aperture Radar imagery but that require an a-priori lake distribution (Miles et al., 2017; Leeson et al., 2020).'

Second, the dataset is limited to the labelled locations of lakes and channels, but no predictor data are given so that the data cannot be used to test different machine learning approaches without extensive further data acquisition.
We agree that the addition of predictor data will result in a high-quality manuscript, we have now adapted the dataset to include all predictor data (i.e., the sensor data for each). Because the satellite sensor data are widely available from the sources listed, we believed it wasn't necessary to include the sensor data alongside shapefiles, as discussed in the manuscript:
'Additional L8 and S2 imagery are freely available at earthexplorer.usgs.gov and scihub.copernicus.eu respectively. Scripts for downloading the data were extracted from

GitHub (Hagolle, Olivier, since 2014) and (Hagolle, Olivier, since 2015), however with changes to data structure on both repositories these scripts may no longer be effective. Alternatively, imagery is available to download from Google Cloud Storage (https://cloud.google.com/storage/docs/public-datasets/) using Python scripting (Nunes, Vasco, since 2016).' (Page 22, Lines 379-383)

As discussed in the text (Section 3.4 Data Usage, Page 21, Lines 355-370), the final lake dataset doesn't contain predictor data as it is a combination of Sentinel-2 and Landsat-8 data and isn't intended for use in ML approaches. It gives a measure of the maximum extent of supraglacial hydrology across the region, throughout January 2017 (extended into February 2017 for full coverage at the cloudy Antarctic Peninsula). This dataset compliments, and is used for comparison with Stokes et al., 2019. With the source data-tile specified for each polygon, the polygons can be used in a water: not-water binary classification approach. The dataset consists of 23,389 individual polygons for S2 and, 17,571 individual polygons for L8 across the test area covering approximately half the Antarctic coastline, throughout 6 weeks of the 2017 melt-season. The data covers a large area and due to this spatial scale displays variation in the cycle of supraglacial hydrology from first formation to refreezing/draining of lakes. This is useful in training algorithms for use in within a single melt-season and throughout multiple melt-seasons, across the continent.

The relevant lines in Section 4. Code and data availability now read:
'The datasets consist of the final lake and channel polygon maps for both sensors combined (i.e. our final maximum extent map of supraglacial hydrology) https://doi.org/10.5281/zenodo.5589525 (Corr et al., 2021f), plus polygons for each sensor, L8 (https://doi.org/10.5281/zenodo.5589460 (Corr et al., 2021c) and https://doi.org/10.5281/zenodo.5588496 (Corr et al., 2021b)) and S2 (https://doi.org/10.5281/zenodo.5589494 (Corr et al., 2021d) and https://doi.org/10.5281/zenodo.5589522 (Corr et al., 2021e)) individually. In addition, predictor data for each sensor (i.e., the data tiles containing all bands for S2 and L8) are provided for each of the polygons.' (Page 22, Lines 371-384)

Further, I have concerns about the thresholding approach used to delineate water pixels. The authors applied a series of thresholds (see Fig. 2 and 3) to the spectral channels but it is unclear to me how these thresholds were derived. Is that based on try and error, and if so, how was the error assessed (visual? statistically by comparing to the manually digitized reference data?).
We agree that the choice of thresholds and their derivation in the manuscript is not explained clearly, we have revised the manuscript to include more details as below. The thresholds used were derived from existing literatures (Moussavi et al., 2020, Stokes et al., 2019; Williamson et al., 2017), however, Stokes et al., 2019 (Green-NIR) and Moussavi et al., 2020 (Blue-Red) use two different variations of NDWI in their approach. As we

aimed to compare and combine our outputs with Stokes et al., 2019, we decided on using a Green-NIR NDWI equation as the primary threshold. To determine the thresholds, we compared the output from two thresholds on the Green-NIR NDWI (0.3 as in Stokes et al., 2019) and a lower threshold of 0.175 to maximise the delineated lake area, with the Dual-NDWI thresholding approach. This analysis has now been added as supplementary material to the manuscript in Appendix A and demonstrates the rationale behind choosing the Dual-NDWI thresholding approach. (Appendix A, Pages 22-24, Lines 398-420)

Moussavi et al., 2020 carried out analysis on the distribution of pixel values from Landsat-8 and Sentinel-2 tiles (Figures 2 and 4, Moussavi et al., 2020) representing the different spectral properties of lakes, slush, snow, shaded snow, clouds, cloud shadows, sunlit rocks, and shaded rocks to determine the thresholds in their approach. The thresholds for rock and cloud masking were adapted from Moussavi et al., 2020, and were selected to ensure maximum lake delineation, with any additional false positives created removed in the manual post-processing stage. These points have been described in the revised manuscript which reads:
'Previous analysis on the distribution of pixel values from Landsat-8 and Sentinel-2 tiles (Figures 2 and 4 in Moussavi et al. (2020)) represent the different spectral properties of lakes, slush, snow, shaded snow, clouds, cloud shadows, sunlit rocks, and shaded rocks. This analysis was used to determine the thresholds in their approach. The thresholds we select for rock and cloud masking were adapted from this approach (Moussavi et al., 2020; and the source code from GitHub (Moussavi, Mahsa, since 2019)). Thresholds were selected to produce maximum lake delineation, with any additional false positives created removed in the manual post-processing stage. The analysis conducted to select the thresholds, the NDWI thresholding approach and the additional band filters is described in Appendix A.' (Page 8, Lines 143-150)

I also wonder why the authors did not use a supervised classification algorithm but decided for a series of manually selected thresholds instead.
We agree that supervised classification approaches could potentially be a valuable tool for mapping supraglacial hydrology from satellite imagery. However, machine learning algorithms are largely unexploited in this field, with a distinct lack of available training data. In addition, a large-scale study such as this requires widescale validation and testing for generalisation and transferability of the methods. It is hoped that this dataset will be used for such a task, and we hope to apply machine learning techniques using supervised classification in our future work; however, it is beyond the scope of the current study. Currently, Quality Control with NDWI thresholding is the best option to provide a benchmark dataset in such large scale. The authors believe this study to be an important starting point in the use of ML approaches in mapping supraglacial hydrology. For this reason and because NDWI thresholding approaches are the golden standard approach to mapping supraglacial hydrology in Greenland and Antarctica (Morriss et al., 2013;

Moussavi et al., 2016; Xu, 2006; Stokes et al., 2019; Williamson et al., 2017), we decided to use the thresholding approach coupled with extensive manual enhancement. We have amended the manuscript to make our rationale clear in choosing the approach, which now reads:

'Supervised classification algorithms are in their infancy in the supraglacial hydrology field (Dirscherl et al., 2020; Halberstadt et al., 2020) and large-scale, continental studies require validation and testing for generalisation and transferability of the methods. The aim of this study was to produce a dataset to assist such studies and, consequently NDWI (Normalised Difference Water Index) thresholding was selected as our approach. Currently, NDWI thresholding methods are the standard approach to mapping supraglacial hydrology in Greenland and Antarctica (Morriss et al., 2013; Moussavi et al., 2016, 2020; Stokes et al., 2019; Williamson et al., 2017; Xu, 2006).' (Page 5, Lines 99-104)

The validation is not complete. If I get it correctly, the accuracy/sensitivity/specificity values that are given in the manuscript refer to the water delineation. What is missing is the validation for the classification into channels and lakes.

We agree it would be valuable to provide validation for the classification of water into channels and lakes, however due to the lack of an objective definition as to what lakes and channels are, it is not possible to compute quantitative accuracy metrics. Instead, this distinction should be viewed more as a guide to the relative split between different shaped hydrological features. The partition of lakes and channels is not part of the core method, and, we have therefore revised and restructured the manuscript to reflect this distinction. Specifically, we have moved the section on the classification of water into channels and lakes (Section 2.1.4, Lakes vs channel classification) into the discussion section (Section 3.2, Lakes vs channel features) of the manuscript. Section 2.2, Accuracy assessment now directly follows the classification processes. (Page 17, Lines 284-331). We have also added additional text to explain this distinction, which now reads:

'The values reported for accuracy, sensitivity and specificity are for the thresholding approach, which consists of all water pixels, including channels and lakes. Although it would be valuable to also provide validation metrics for the classification of water into channels and lakes, due to the lack of an objective definition as to what lakes and channels are it is not possible to compute accuracy, sensitivity, or specificity metrics at present. Channels and lakes are defined from within the classification of surface water, based solely upon their shape. To concretely define channels, would require auxiliary data, such as water flow and topography at instances in close temporal proximity to the satellite imagery. The aim of our channel and lake discrimination is therefore not to provide a measure or definition of each, but rather it is an indicator that should be viewed more as a guide to the relative split between them.' (Page 14, Lines 260-267).

Referee 3: https://doi.org/10.5194/essd-2021-257-RC3

The authors present a methodology to create a benchmark dataset of supraglacial hydrology across the Antarctic. Their approach takes advantage of the large amount of multi-spectral satellite imagery provided by Landsat-8 and Sentinel-2. They propose a pipeline consisting of a cascade of spectral thresholding for the creation of a binary mask and subsequent K-means clustering to further distinguish lakes and channels. The manuscript is clear and the exposition is easy to follow and linear.

Thank you very much for this review. We appreciate your concerns and have taken steps to remedy the issues in the manuscript, as discussed below.

Specific Comments

I have only one specific comment related to the validation approach that was used, particularly thinking about future use by researchers in the area of machine learning. As stated by the authors already in the abstract, the dataset provides a partition of the supraglacial features into non-water, lakes and streams. However, from my understanding in Table 1 the authors are showing aggregated results (binary water/non-water) for the accuracy and sensitivity metrics. I think that for the future users to benefit the most from the proposed inventory, it would be beneficial to show a validation considering all the classes.

As is noted in our reply to Reviewer 2, we agree it would be valuable to provide validation for the classification of water into channels and lakes. However, due to the lack of an objective definition as to what lakes and channels are, it is not possible to compute quantitative accuracy metrics. Instead, this distinction should be viewed more as a guide to the relative split between different shaped hydrological features. The separation of surface water features into channels and lakes was performed only on the final lake dataset (the maximum extent of all supraglacial features from Sentinel-2 and Landsat-8 throughout January 2017 combined). This dataset, as highlighted in the revised manuscript (Page 21, Lines 355-370), is not to be used as a machine learning dataset. The machine learning datasets are provided alongside this final map of meltwater extent (not a machine learning resource) and contain meltwater polygons for each sensor (S2 and L8), alongside the respective source sensor/predictor data.

In addition, we have moved, renamed, and restructured Section 2.1.4, Lakes vs channel classification to Section 3.2, Lakes vs channel features, to distinguish this analysis from the pixel-based classification of surface water and not-surface water pixels. This change was carried out to make clear that the partition of lakes and channels is not part of the core method. Section 2.2, Accuracy assessment now directly follows the classification processes. (Page 17, Lines 284-331).