Responses to Reviewers' Comments

Dear Editor,

We sincerely thank you for your constructive comments and valuable suggestions on our manuscript entitled "*A multi-decadal dataset of surface damage on Antarctic ice shelves (1999–2024)*" (Manuscript ID: *essd-2026-414*). We appreciate the time and effort you devoted to helping us improve the quality of our paper.

As detailed in our point-by-point response below, we have taken all of the comments and suggestions into account in the revised version of the manuscript. All revisions have been tracked. We sincerely hope that this manuscript will be acceptable for publication in the *Earth System Science Data*.

The comments are in **bold**, and the revised text is in *italics*.

Sincerely,

Dr. Gang Qiao

College of Surveying and Geo-Informatics

Tongji University

1239 Siping Road, Shanghai, China, 200092

Email: qiaogang@tongji.edu.cn

Phone: +86-21-13818426779

We sincerely appreciate the constructive comments and valuable suggestions, which have helped us improve the quality of our paper. As detailed in our point-by-point response below, we have carefully taken all comments and suggestions into account in the revised version of the manuscript. In the following responses, the comments are shown in "**bold**", our responses are shown in "non-bold", and the revised text in the manuscript is shown in "*italic*".

Reviewer 1:

This manuscript presents a valuable Antarctic ice shelf surface damage dataset. However, several parts of the methodology, model evaluation, validation and dataset description need to be explained more clearly. The following comments need to be addressed before the manuscript can be considered for publication.

Response:

We would like to thank you for spending precious time reviewing our manuscript. Your constructive comments and suggestions concerning our research contents are of great help to us in improving this manuscript. All of your comments and suggestions have been considered when revising the manuscript, and all revisions have been tracked.

1. Section 2.3.1 (Lines 188–197): "close in time" is not clear. Please specify the time gap range between the Landsat 8 images and the corresponding Sentinel-1 and ICESat-2 data used for annotation.

Response: Thank you for your constructive suggestion.

We have revised Section 2.3.1 to explicitly specify the temporal gaps between the Landsat 8 imagery and the corresponding Sentinel-1 and ICESat-2 data, as follows:

Specifically, for each Landsat 8 image used for annotation, Sentinel-1 SAR Interferometric Wide Swath (IW) Ground Range Detected (GRD) data covering the same region and with the temporally closest available acquisition date were selected. The acquisition-date difference between the Landsat 8 and selected Sentinel-1 images ranged from 0 to 3 days. The Sentinel-1 data were downloaded and processed in SNAP through a standard workflow, including thermal noise removal, radiometric calibration, speckle filtering, and terrain correction using the REMA DEM (Howat et al., 2019; Howat et al., 2022), to derive geocoded backscatter imagery. ICESat-2 ATL06 data acquired within a 91-day window centred on each Landsat 8 acquisition date (45 days before to 45 days after) were used. This window corresponds to one ICESat-2 repeat cycle. The data were downloaded and processed following Wang et al. (2021) to provide additional surface elevation information for identifying surface damage.

2. Section 2.3.1, page 8: Only Landsat 8 images from the 2020 austral summer are used for the training and test datasets? and the final dataset covers 1999–2024 and includes Landsat 7, Landsat 8 and Landsat 9 images. Please explain why this is sufficient to represent the full dataset or provide additional validation for the other sensors and time periods. This part needs to be explained more clearly in the manuscript.

Response: Thank you for your constructive suggestion.

We added a cross-sensor evaluation using 34 manually annotated 512×512 Landsat 7 SLC-on image tiles acquired before the SLC failure in 2003 from Brunt, Dotson and Thwaites ice shelves. None of these ice shelves were included in the training or validation sets. The trained model was applied to this independent cross-sensor test set for quantitative evaluation. Accordingly, Sect. 3.2.2 was revised as follows:

3.2.2 Performance of segmentation model on independent test sets

To further evaluate the generalizability of the image segmentation model, two independent test sets were used. First, Totten Ice Shelf, which was not included in either the training or validation set, was selected for an independent ice-shelf test. A Landsat 8 optical image acquired on 2021-01-23 was manually annotated following the same criteria as described in Sect. 2.3.1, and the prediction over the contiguous image extent was evaluated as a whole. As shown in Fig. 9, the prediction is overall consistent with the label, with sparse FN (blue) errors distributed along the edges of damage regions and in damage areas with relatively low image contrast. The evaluation metrics are presented in Table 3. The results indicate that the model maintained good segmentation performance on Totten Ice Shelf, demonstrating its ability to generalize to a geographically independent region.

Second, to evaluate cross-sensor generalization, the trained model was applied to an independent cross-sensor test set comprising Landsat 7 SLC-on imagery. This test set comprised 34 manually annotated 512 × 512 image tiles from Brunt, Dotson and Thwaites ice shelves (Fig. S1), none of which were included in the training or validation sets. The images were manually annotated following the same criteria, but without the assistance of Sentinel-1 and ICESat-2 data. As shown in Table 3 and Fig. 10, the segmentation model achieved comparable overall performance on the Landsat 7 cross-sensor test set. Together, the results from these two independent test sets provide further evidence for the geographic and cross-sensor generalizability of the segmentation model.

Fig. 10 was added to show representative prediction examples from this test set:

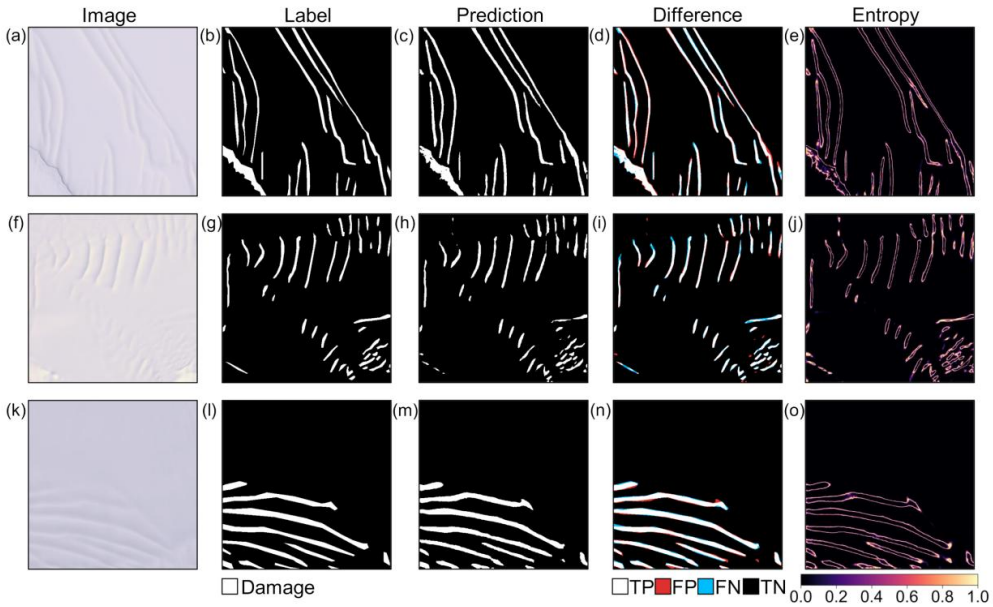

Figure 10. Examples of prediction performance on samples from the Landsat 7 SLC-on cross-sensor test set. (a), (f) and (k) show image samples from Brunt Ice Shelf, Dotson Ice Shelf and Thwaites Ice

Shelf, respectively. (b), (g) and (l) show the corresponding manually annotated labels. (c), (h) and (m) show the corresponding model predictions. (d), (i) and (n) show the pixel-wise difference maps between the corresponding labels and predictions, where TP denotes true positive, FP denotes false positive, FN denotes false negative and TN denotes true negative. (e), (j) and (o) show the corresponding prediction entropy maps.

Table 3 was correspondingly updated:

Table 3. Performance of the segmentation model on the validation set and independent test sets. (L8: Landsat 8; L7: Landsat 7; SLC-on: scan-line-corrector-on)

Evaluation set	*Sensor*	*Evaluation extent*	*mIoU*	*OA*	*Precision*	*Recall*	*F1*
Validation set	*L8*	*52 tiles*	*0.845*	*0.984*	*0.816*	*0.841*	*0.829*
Independent ice-shelf test	*L8*	*Contiguous scene*	*0.822*	*0.960*	*0.838*	*0.792*	*0.814*
Independent cross-sensor test	*L7 SLC-on*	*34 tiles*	*0.837*	*0.985*	*0.799*	*0.834*	*0.816*

And Fig. S1 has been updated accordingly:

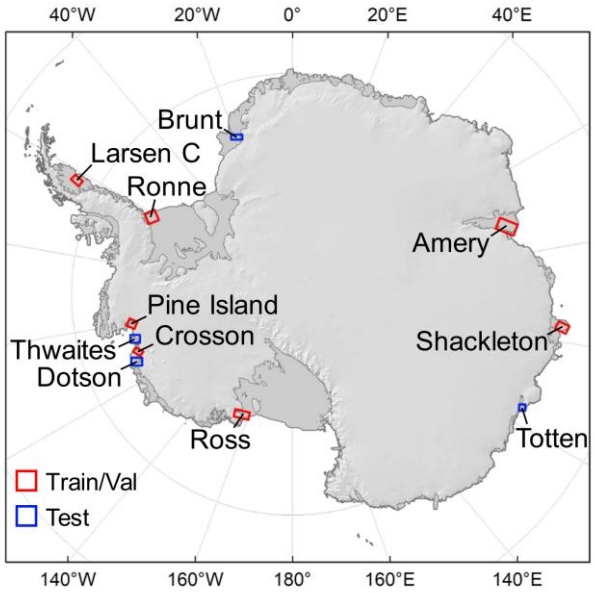

Figure S1. Spatial distribution of the source imagery used for the training/validation and test images. Red boxes indicate the regions from which the training and validation tiles were sampled, and blue boxes indicate the test regions. Training and validation tiles were split within the same ice-shelf regions. The background imagery and boundary data are from Bedmap2.

The figures showing temporal variations have also been updated using revised uncertainty bounds derived from the evaluation metrics of the newly added independent tests, as follows:

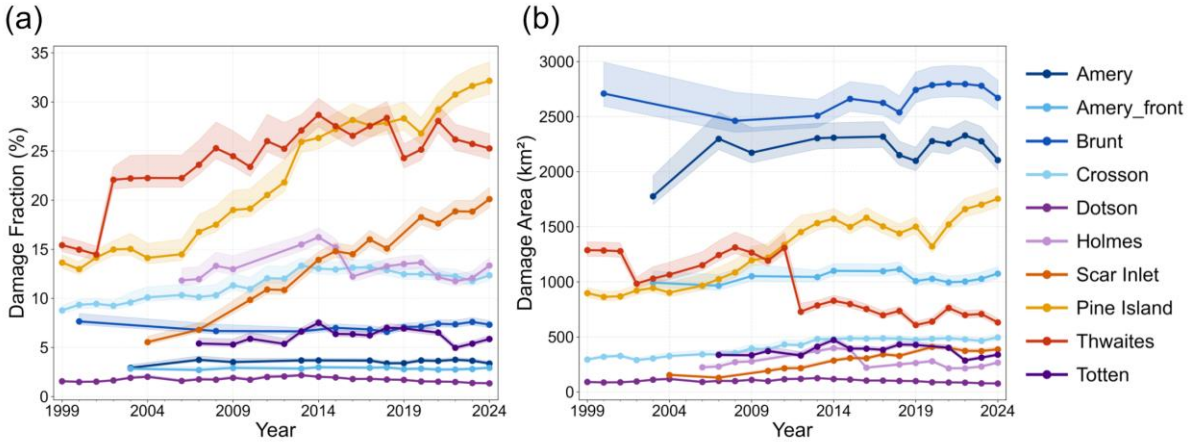

Figure 13. Temporal variations in damage fraction (a) and damage area (b) for the ice shelves in the dataset during 1999–2024. Colour-vision-deficiency-friendly versions of this figure are provided in Figs. S8 and S9.

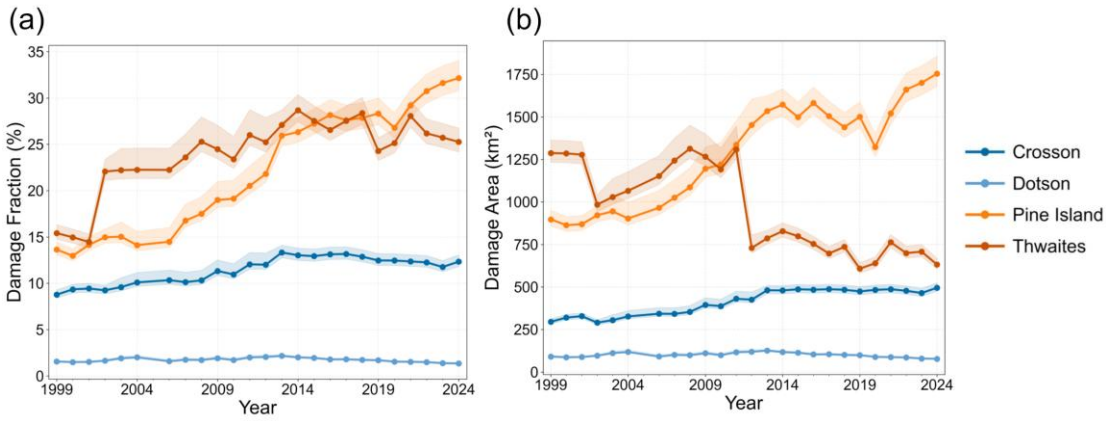

Figure S8. Temporal variations in damage fraction (a) and damage area (b) for Crosson, Dotson, Pine Island and Thwaites Ice Shelves during 1999–2024.

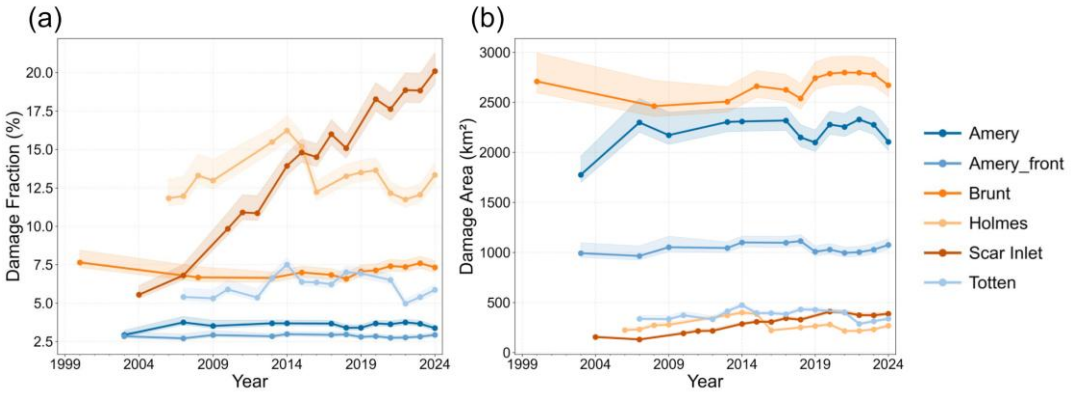

Figure S9. Temporal variations in damage fraction (a) and damage area (b) for Amery, Amery_front, Brunt, Holmes, Scar Inlet and Totten Ice Shelves during 1999–2024.

Considering the highly consistent spectral band characteristics of the sensors onboard Landsat 8 and Landsat 9, no additional evaluation was conducted for Landsat 9 imagery. The validation strategy was clarified in Sect. 2.5.1 as follows:

The above metrics were calculated for the validation set of the manually annotated dataset. The evaluation was further conducted on a contiguous Landsat 8 scene and a set of Landsat 7 SLC-on image tiles from ice shelves not represented in the training or validation data, to assess the geographic generalizability of the segmentation model and its performance across sensors. Considering the highly consistent spectral band characteristics of the sensors onboard Landsat 8 and Landsat 9, no additional evaluation was conducted for Landsat 9 imagery.

3. Section 2.3.2: It would be helpful to add a figure to show the segmentation model architecture, including example input images, the SegFormer model and the corresponding output surface damage map.

Response: Thank you for your constructive suggestion.

We added a schematic figure in Sect. 2.3.2 to illustrate the SegFormer-based surface damage segmentation, including an example RGB input, the SegFormer model and the corresponding binary surface damage prediction. As SegFormer is an established network architecture rather than a methodological contribution of this study, only its main encoder-decoder structure is shown in the schematic. Further architectural details are available in Xie et al. (2021). The added figure is as follows:

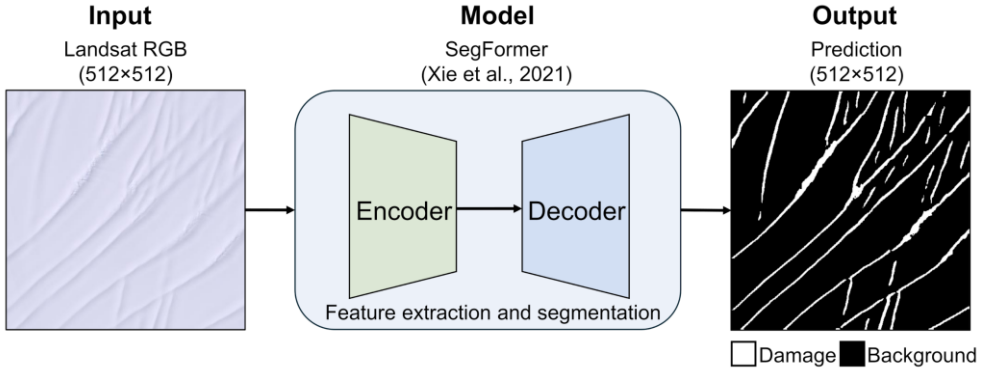

Figure 6. Schematic overview of SegFormer-based surface damage segmentation.

4. Section 2.3.2: Please explain why you used SegFormer model for this study. Did you compare it with other segmentation models such as U-Net, DeepLabV3 or other transformer based models? Need to briefly clarify the performance comparison and explain why you selected SegFormer model over the other models. This part should be explained more clearly in this manuscript.

Response: Thank you for your constructive suggestion.

We conducted additional comparisons between SegFormer-B2, U-Net and DeepLabV3+ with their respective commonly used implementations, under the same training strategy. The three models were evaluated on the same validation set, and SegFormer achieved the highest overall performance. The comparison results have been added as Table S2, and Sect. 2.3.2 has been revised as follows:

To further assess its suitability for this task, SegFormer-B2 was compared with U-Net (Ronneberger et al., 2015) using a ResNet-34 encoder and DeepLabV3+ (Chen et al., 2018) using a ResNet-50 encoder under the same training strategy. SegFormer achieved the best overall performance (Table S2) and was therefore selected for surface damage mapping.

Table S2. Performance of different segmentation models on the validation set.

Model	***Configuration***	***mIoU***	***OA***	***Precision***	***Recall***	***F1***
SegFormer	*B2*	*0.845*	*0.984*	*0.816*	*0.841*	*0.829*
U-Net	*ResNet-34 encoder*	*0.843*	*0.983*	*0.814*	*0.838*	*0.825*
DeeplabV3+	*ResNet-50 encoder*	*0.834*	*0.982*	*0.797*	*0.832*	*0.814*

5. Section 2.3.3: Please clarify how the final model was selected. Was a validation dataset used to choose the best model, or was the model from the last training epoch used?

Response: Thank you for your constructive suggestion.

We have clarified the model selection procedure and corrected the terminology used for the data split. The 276 manually annotated images were divided into 224 training images (80 %) and 52 validation images (20 %), and the validation set was used for model selection. The model checkpoint with the best performance on the validation set was selected as the final model. The selected checkpoint corresponded to epoch 190, rather than the last training epoch. We have therefore revised the original "test set" to "validation set" throughout the manuscript. The Landsat 8 contiguous scene and Landsat 7 SLC-on imagery described in Sect. 3.2.2 are used as the independent test data.

Sect. 2.3.3 has been revised accordingly:

The manually annotated damage dataset is used for model training and validation, with 224 images (80 %) used as the training set and 52 images (20 %) used as the validation set. Images from each region are randomly allocated to the training and validation sets to ensure that both subsets capture a representative range of surface damage characteristics. No spatial overlap exists between any images….

The model checkpoint with the best performance on the validation set was selected as the final model. The selected checkpoint corresponded to epoch 190.

6. Could you please briefly explain how the hyperparameters (learning rate, batch size, number of epochs, etc) were selected for training?

Response: Thank you for your constructive suggestion.

We have added a brief explanation of the hyperparameter selection in Sect. 2.3.3. The training hyperparameters were selected based on preliminary experiments, model convergence behaviour and computational constraints. The learning-rate schedule and number of epochs were determined to provide stable convergence, while the batch size was selected considering GPU memory availability. The corresponding description has been added as follows:

The training hyperparameters were selected based on preliminary experiments, model convergence behaviour and computational constraints. The learning-rate schedule and number of epochs were

determined according to model convergence, while the batch size was selected considering GPU memory availability.

7. Section 2.3.3: Please clarify which data augmentation methods were used during training and whether they were applied to all training images?

Response: Thank you for your constructive suggestion.

We have clarified the data augmentation procedure in Sect. 2.3.3. All training images were subject to this augmentation procedure. Random rotation between −45° and 45° was applied to each training sample, while brightness adjustment within ±25 % was applied with a probability of 50 %. The corresponding description has been revised as follows:

During training, all training images were subject to a random augmentation procedure comprising rotation between −45° and 45° and brightness adjustment within ±25 %. Rotation was applied to each training sample, whereas brightness adjustment was applied with a probability of 50 %. These augmentations were used to increase sample diversity and improve model robustness and generalization ability.

8. Please explain more clearly how the training and test datasets were split. To better demonstrate the model's generalizability, please consider using spatial cross-validation, for example, by training on ice shelves in some regions of Antarctica (e.g., western Antarctica) and testing on ice shelves in other regions (e.g., eastern Antarctica). This can provide a better evaluation of the model's performance in unseen regions.

Response: Thank you for your constructive suggestion.

We have clarified the data split in Sect. 2.3.3. The 276 manually annotated images were divided into 224 training images (80 %) and 52 validation images (20 %). Images from each region were randomly allocated to the training and validation sets, with no spatial overlap between any images. Sect. 2.3.3 has been revised accordingly:

The manually annotated damage dataset is used for model training and validation, with 224 images (80 %) used as the training set and 52 images (20 %) used as the validation set. Images from each region are randomly allocated to the training and validation sets to ensure that both subsets capture a representative range of surface damage characteristics. No spatial overlap exists between any images. Geographic generalizability is further evaluated using test data from ice shelves not represented in either the training or validation set (Sects. 2.5.1 and 3.2.2).

As noted in our response to Comment 2, to further evaluate the model in unseen regions, additional test data were selected from ice shelves not represented in either the training or validation set, including the contiguous Landsat 8 test scene over Totten Ice Shelf and the Landsat 7 SLC-on test data from Brunt, Dotson and Thwaites ice shelves described in Sect. 3.2.2. These geographically separate test data provide a direct evaluation of the model’s performance in previously unseen ice-shelf regions.

9. Section 2.4.2: Since the final products were manually refined, please explain how consistency was maintained throughout the dataset. For example, were the same criteria applied to all ice shelves and years?

Response: Thank you for your constructive suggestion.

We have further clarified the consistency of the manual extent definition in Sect. 2.4.2 by explicitly stating that the same criteria were applied throughout the dataset:

The same criteria were applied consistently across all ice shelves and years.

10. Section 2.4.2: Please describe the manual refinement process in more detail. What types of edits were made, and were only obvious errors corrected, or were the predicted damage boundaries also manually modified?

Response: Thank you for your constructive suggestion.

We have clarified that the manual adjustment in Sect. 2.4.2 refers only to the definition of the effective floating ice-shelf extent and does not involve manual modification of the damage predictions. The following description has been added:

The manual adjustment in this step is limited to defining the effective floating ice-shelf extent and no manual editing of predicted damage pixels or damage boundaries is performed during this process.

11. Section 3.2.3 (Lines 373–385): The DiffGF restoration is evaluated using only one simulated example. Since restored Landsat 7 images are used in many parts of the dataset, please consider testing the method on more regions and different surface conditions to better show that it works well. Also, Table 4 reports the results for the whole image. It would be more useful to also report the results only for the reconstructed SLC-off gap areas, where the restoration is actually applied.

Response: Thank you for your constructive suggestion.

We have expanded the evaluation in Section 3.2.3. In addition to the primary test on Scar Inlet Ice Shelf, we added simulated SLC-off tests using regional images from Amery, Crosson and Pine Island ice shelves to cover different regions and surface conditions. The corresponding restoration results and surface damage segmentation results are shown in Figs. S5–S7, and the quantitative segmentation performance is reported in Table S3. These additional tests show consistently high segmentation performance, providing further support for the applicability of restored Landsat 7 SLC-off imagery to surface damage mapping.

Table 4 has also been revised to report the image quality metrics for the Scar Inlet test both over the full image and specifically within the simulated SLC-off gap areas.

Section 3.2.3 has been revised accordingly to describe these additional evaluations and results, as follows:

3.2.3 Performance of segmentation on restored Landsat 7 SLC-off imagery

The restoration performance of DiffGF on Landsat 7 SLC-off imagery is also evaluated (see (Tang et al., 2026b)). Scar Inlet Ice Shelf is selected as the primary test case, as it contains multiple types of surface damage and was not included in the training data of either the DiffGF framework or the image segmentation model.

The test imagery is constructed from a Landsat 8 image acquired on 2021-12-21, to which an SLC-off mask extracted from a Landsat 7 SLC-off image acquired on 2012-02-20 at the same location is applied, thereby generating simulated SLC-off imagery with ground-truth reference. DiffGF is then used to restore the simulated SLC-off imagery, and the image segmentation model is applied to the restored imagery for surface damage extraction. The results are shown in Figs. 11, S3 and S4. Image quality metrics between the restored imagery and the ground-truth imagery are reported in Table 4. The restoration metrics are calculated both over the full image and the simulated SLC-off gap areas. The comparison between the segmentation result from the restored imagery and that from the ground-truth imagery is shown in Table 5.

To further assess the applicability of DiffGF across different regions and surface conditions, additional simulated SLC-off tests are conducted using regional test images from Amery, Crosson and Pine Island ice shelves. The corresponding restoration results and surface damage segmentation performance are provided in Figs. S5–S7 and Table S3.

The evaluation on Scar Inlet Ice Shelf indicates that DiffGF achieves good restoration performance in this test case, and that the surface damage segmentation result obtained from the restored imagery (Fig. 11e) is close to that from the ground-truth imagery (Fig. 11d). The additional regional tests also show consistently high segmentation performance on the restored imagery (Table S3). Together, these results provide further support for the use of restored Landsat 7 SLC-off imagery in surface damage mapping.

Table 4. Image quality metrics of the restored SLC-off test image over the full image and SLC-off gap areas. (LPIPS is calculated only on the full image)

Extent	*PSNR*	*UIQI*	*SSIM*	*CC*	*RMSE*	*LPIPS*
Full image	*35.070*	*0.994*	*0.966*	*0.994*	*0.018*	*0.060*
SLC-off gaps	*27.779*	*0.962*	*0.862*	*0.963*	*0.041*	–

The added Table S3 and Figs. S5–S7 are as follows:

Table S3. Segmentation evaluation of the additional regional restoration tests against the original complete image.

Site	*mIoU*	*OA*	*Precision*	*Recall*	*F1*
Amery Ice Shelf	*0.958*	*0.995*	*0.961*	*0.958*	*0.960*
Crosson Ice Shelf	*0.963*	*0.992*	*0.973*	*0.960*	*0.966*
Pine Island Ice Shelf	*0.956*	*0.978*	*0.986*	*0.962*	*0.974*

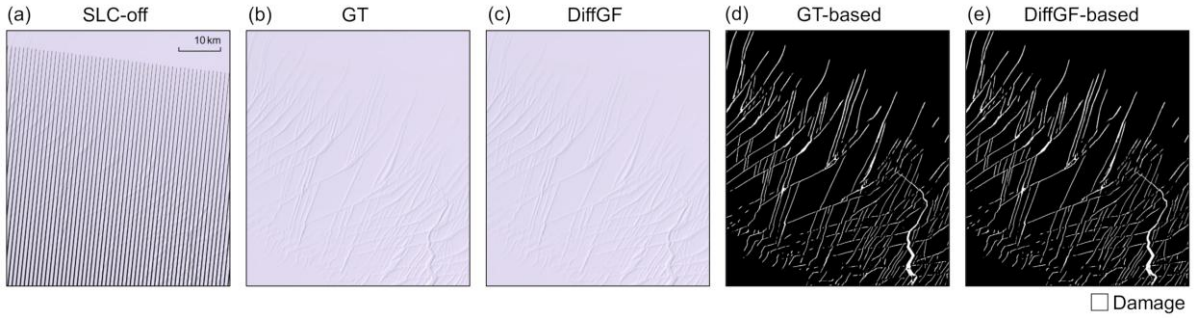

Figure S5. SLC-off restoration performance for a regional test image from Amery Ice Shelf. (a) Simulated SLC-off image generated by applying an SLC-off mask extracted from Landsat 7 imagery (2009-01-20) to Landsat 8 imagery (2016-11-15). (b) Ground truth (GT) image. (c) DiffGF-restored image. (d) Damage segmentation derived from GT. (e) Damage segmentation derived from DiffGF restoration.

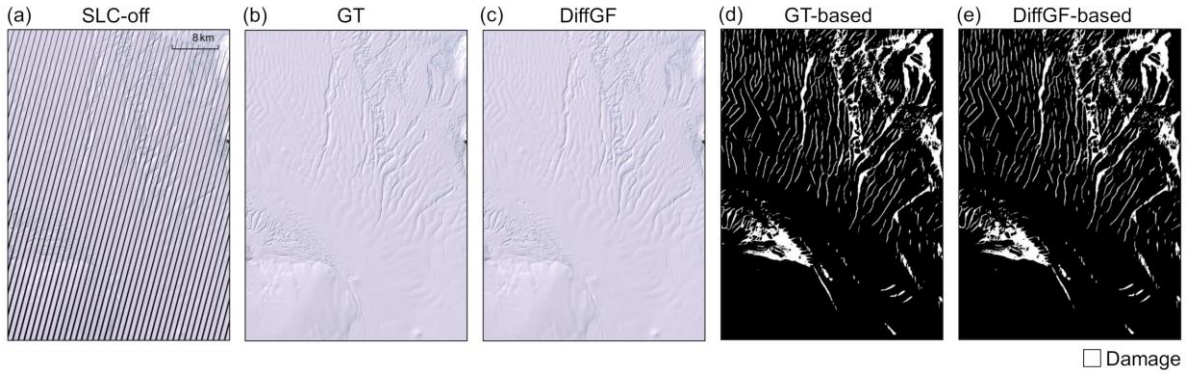

Figure S6. SLC-off restoration performance for a regional test image from Crosson Ice Shelf. (a) Simulated SLC-off image generated by applying an SLC-off mask extracted from Landsat 7 imagery (2008-02-21) to Landsat 8 imagery (2014-12-05). (b) Ground truth (GT) image. (c) DiffGF-restored image. (d) Damage segmentation derived from GT. (e) Damage segmentation derived from DiffGF restoration.

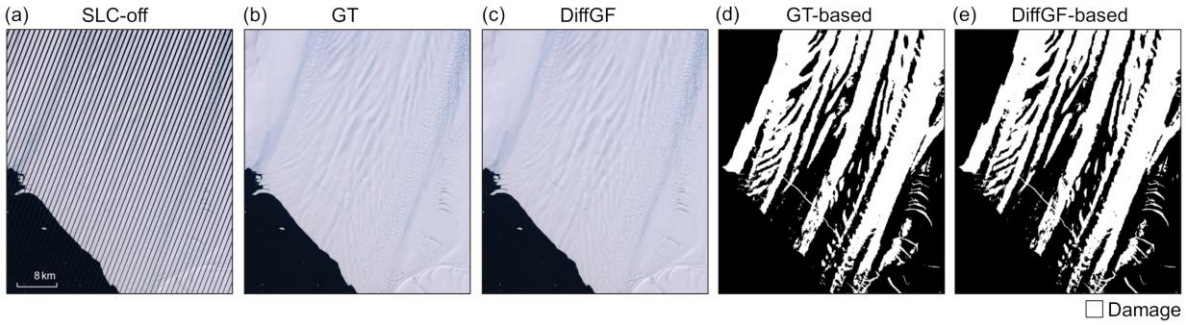

Figure S7. SLC-off restoration performance for a regional test image from Pine Island Ice Shelf. (a) Simulated SLC-off image generated by applying an SLC-off mask extracted from Landsat 7 imagery (2007-01-05) to Landsat 8 imagery (2017-01-08). (b) Ground truth (GT) image. (c) DiffGF-restored image. (d) Damage segmentation derived from GT. (e) Damage segmentation derived from DiffGF restoration.

12. Please define surface damage more clearly. It would be helpful to explain exactly which features are considered damage and which are not. Based on the example images, the mapped

features appear to include crevasses, rifts, and fracture networks, which are also related to surface roughness and deformation. Please provide a clearer definition of what is considered surface damage and explain how it is distinguished from normal surface roughness or naturally crevassed ice.

Response: Thank you for your constructive suggestion.

We have added a clearer description of the definition and annotation criteria of surface damage in Sect. 2.3.1, as follows:

These features were annotated as a single damage class. Surface damage was identified in the Landsat imagery based on distinct linear or curvilinear surface discontinuities and densely fractured areas characterized by clearly disrupted surface patterns. In contrast, general surface roughness and other gradual variations in surface appearance were not labelled as damage in the absence of distinct surface discontinuities or clearly disrupted patterns.

13. Section 3: The uncertainty analysis is useful; however, it would be much more informative to include spatial uncertainty maps (or confidence maps) along with the damage maps. For example prediction entropy or the standard deviation/variance of predictions from multiple runs with different random seeds or hyperparameters.

Response: Thank you for your constructive suggestion.

We have added prediction entropy maps to the representative examples shown in Figs. 8 and 10, so that spatial patterns of model uncertainty can be viewed alongside the damage maps:

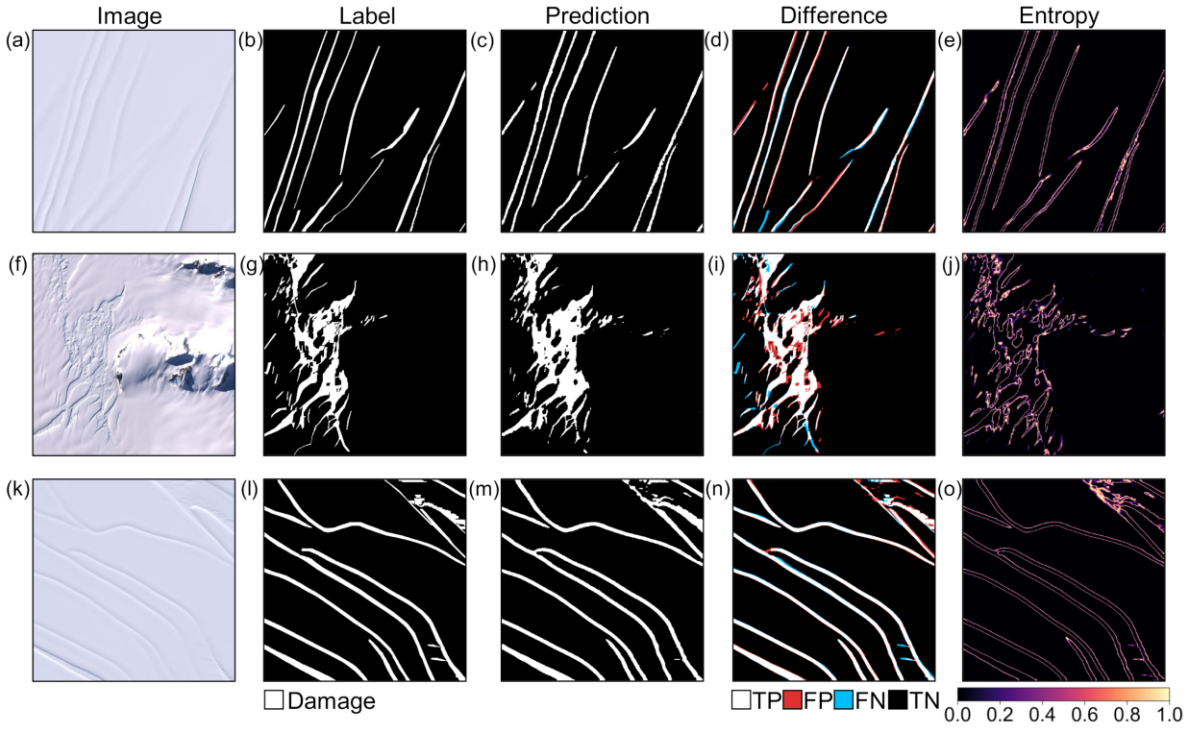

Figure 8. Examples of prediction performance on samples from the validation set. (a), (f) and (k) show image samples from Amery Ice Shelf, Crosson Ice Shelf and Larsen C Ice Shelf, respectively. (b), (g)

and (l) show the corresponding manually annotated labels. (c), (h) and (m) show the corresponding model predictions. (d), (i) and (n) show the pixel-wise difference maps between the corresponding labels and predictions, where TP denotes true positive, FP denotes false positive, FN denotes false negative and TN denotes true negative. (e), (j) and (o) show the corresponding prediction entropy maps.

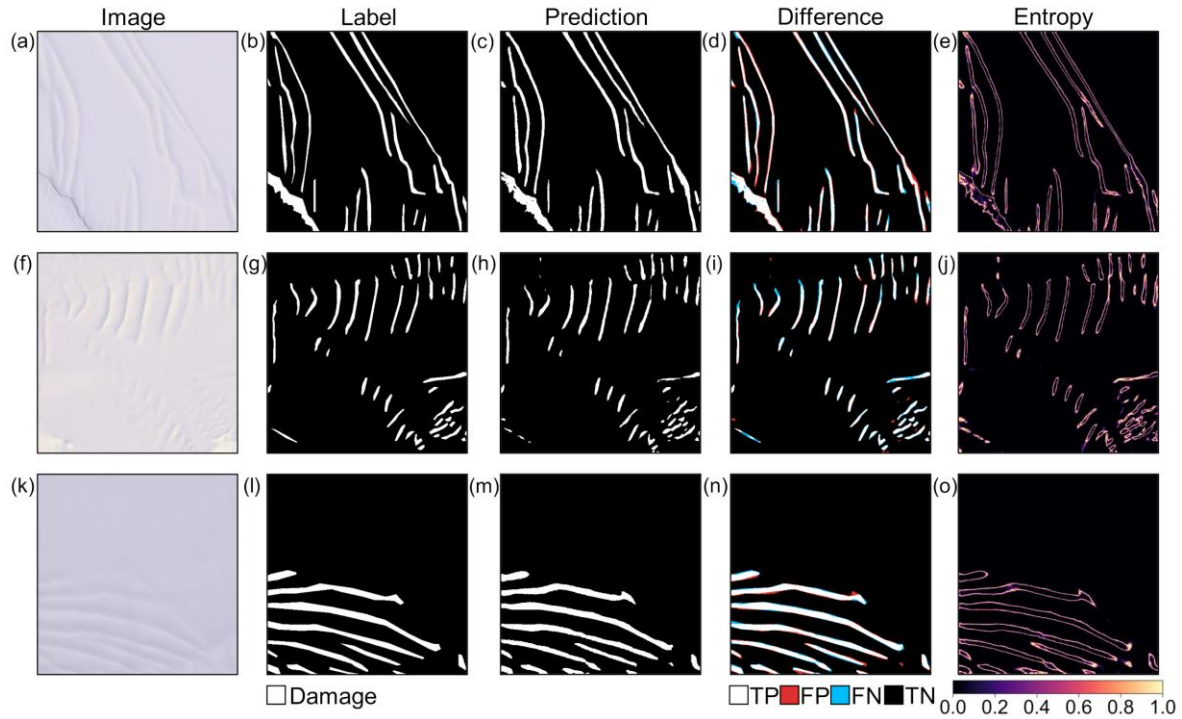

Figure 10. Examples of prediction performance on samples from the Landsat 7 SLC-on cross-sensor test set. (a), (f) and (k) show image samples from Brunt Ice Shelf, Dotson Ice Shelf and Thwaites Ice Shelf, respectively. (b), (g) and (l) show the corresponding manually annotated labels. (c), (h) and (m) show the corresponding model predictions. (d), (i) and (n) show the pixel-wise difference maps between the corresponding labels and predictions, where TP denotes true positive, FP denotes false positive, FN denotes false negative and TN denotes true negative. (e), (j) and (o) show the corresponding prediction entropy maps.

We have also added the following description to Sect. 3.2.4, as follows:

To complement these map-level estimates with spatial information, prediction entropy maps were generated for representative samples from the validation and cross-sensor test sets (Figs. 8 and 10). Prediction entropy was calculated from the pixel-level class probabilities and normalized to the range 0–1. The entropy maps show that higher entropy is mainly found along damage boundaries and within regions characterized by complex damage patterns, generally coinciding with the spatial distribution of FP and FN errors in the difference maps.

14. In this study, quantifying data uncertainty is particularly important because it can help identify label noise, inherent image noise and uncertainties in the manual annotations.

Section 3: Please provide validation results for additional ice shelves, other sensors and different years. If possible, please include quantitative evaluation. This is particularly important because there are differences among these sensors, and it would better show that the model performs consistently across different years, not only on the current test data.

Response: Thank you for your constructive suggestion.

This comment is closely related to Comment 2 and was addressed together with it.

As described in our response to Comment 2, we added quantitative evaluations on an independent Landsat 8 image of Totten Ice Shelf acquired in 2021 and on 34 manually annotated Landsat 7 SLC-on image tiles from Brunt, Dotson and Thwaites ice shelves acquired before 2003. These additional tests evaluate the model across geographically independent ice shelves, different acquisition periods and sensors. The corresponding results have been added to Sect. 3.2.2, Table 3 and Fig. 10.

15. Please provide some software and computing details used to train and run the model, such as the programming language, main libraries and GPU.

Response: Thank you for your constructive suggestion.

We have added the software and computing details in Sect. 2.3.2, including the programming language, main libraries and GPU used for model training:

The model was implemented in Python with PyTorch and the Transformers library. Training was conducted on an NVIDIA A100 GPU in a shared high-performance computing environment with non-exclusive GPU allocation.

16. It is required to add a map to show the spatial distribution of the training and test images using different colors. This can help readers understand their spatial distribution and check whether the test images are well separated from the training images. Please also show any validation images separately, if a validation set was used.

Response: Thank you for your constructive suggestion.

We revised Fig. S1 to show the spatial distribution of the source imagery used for the training/validation and independent test images using different colours. As already described in Sect. 2.3.3, the training and validation tiles were randomly split within the same source ice-shelf regions, they occupy the same regional footprints at the scale of Fig. S1. The test regions are shown separately in blue, highlighting their spatial separation from the training/validation regions.

The updated Fig. S1 is as follows:

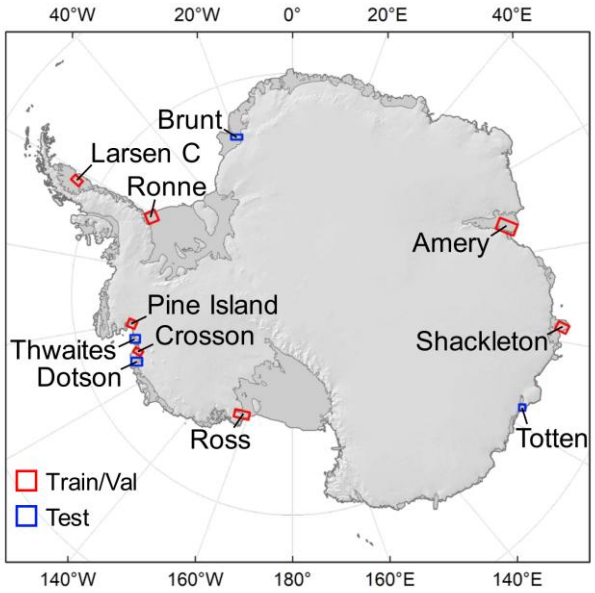

Figure S1. Spatial distribution of the source imagery used for the training/validation and test images. Red boxes indicate the regions from which the training and validation tiles were sampled, and blue boxes indicate the test regions. Training and validation tiles were split within the same ice-shelf regions. The background imagery and boundary data are from Bedmap2.

Reviewer 2:

The paper presents a potentially useful data set of ice shelf surface damage based on visible-NIR imagery and a machine-learning approach to detecting the surface features indicating damage. The paper is well-written, virtually zero issues with English or grammar, and could be considered publishable as it is, with proper qualification.

Response:

We would like to thank you for spending precious time reviewing our manuscript. Your constructive comments and suggestions concerning our research contents are of great help to us in improving this manuscript. All of your comments and suggestions have been considered when revising the manuscript, and all revisions have been tracked.

1. But. Ice shelf damage comes in three types: surface damage, (a) rifts and (b) crevasses, and (c) basal fractures. All three weaken shelves, but in differing ways that would be important to modelers seeking to use your data set to explore various ice shelf vulnerabilities. So, while the data set might be valid within the scope of what has been done, and the title, it would be hard to implement for the most logical application(s).

Response: Thank you for your constructive suggestion.

We agree that different types of damage can have different mechanical effects, and that distinguishing these types would provide additional information for applications concerned with their specific roles in ice-shelf weakening.

The present dataset is designed to provide a consistent, multi-decadal record of surface damage observable in Landsat imagery. Different damage types are therefore not distinguished, and surface expressions associated with basal crevasses may also be included in the mapped damage. Despite this limitation, the dataset remains useful for analyses of the overall evolution of surface damage and related ice-shelf changes. Further differentiation of damage types would provide additional information for investigating their potentially different mechanical effects, but would require additional evidence from other data sources.

We have clarified this point in the Discussion and added the following statement:

Further differentiation of damage types would provide additional information for investigating their potentially different mechanical effects, but requires evidence from other data sources. Future work could further develop damage-type classification by incorporating broader spatial context and auxiliary glaciological information, such as ice thickness.

2. The authors detect surface damage on the basis of sharp linear breaks in pixel brightness – what are the sources of false positives here? Frost patches can produce linear breaks in brightness, were these detected? Please include more of discussion of the false positives, negatives, etc. – what were the causes of these?

Response: Thank you for your constructive comment.

We have further examined the false-positive and false-negative patterns in the evaluation results. Inspection of the difference maps (Figs. 8–10) shows that most false-positive and false-negative pixels occur along damage boundaries, rather than as spatially isolated misclassifications. We added entropy maps for Figs. 8 and 10. These errors mainly reflect small differences in boundary delineation between the manual labels and model predictions, and this pattern is also consistent with the boundary-related uncertainty shown in the entropy maps. The updated Figs. 8 and 10 are as follows:

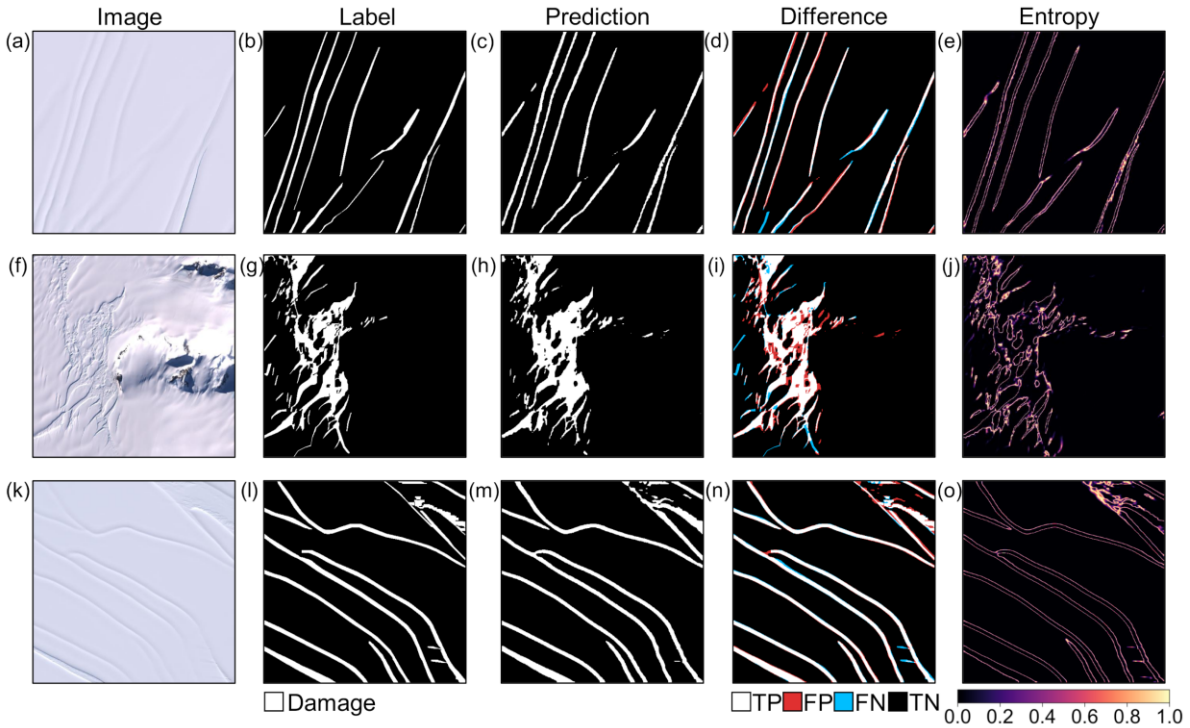

Figure 8. Examples of prediction performance on samples from the validation set. (a), (f) and (k) show image samples from Amery Ice Shelf, Crosson Ice Shelf and Larsen C Ice Shelf, respectively. (b), (g) and (l) show the corresponding manually annotated labels. (c), (h) and (m) show the corresponding model predictions. (d), (i) and (n) show the pixel-wise difference maps between the corresponding labels and predictions, where TP denotes true positive, FP denotes false positive, FN denotes false negative and TN denotes true negative. (e), (j) and (o) show the corresponding prediction entropy maps.

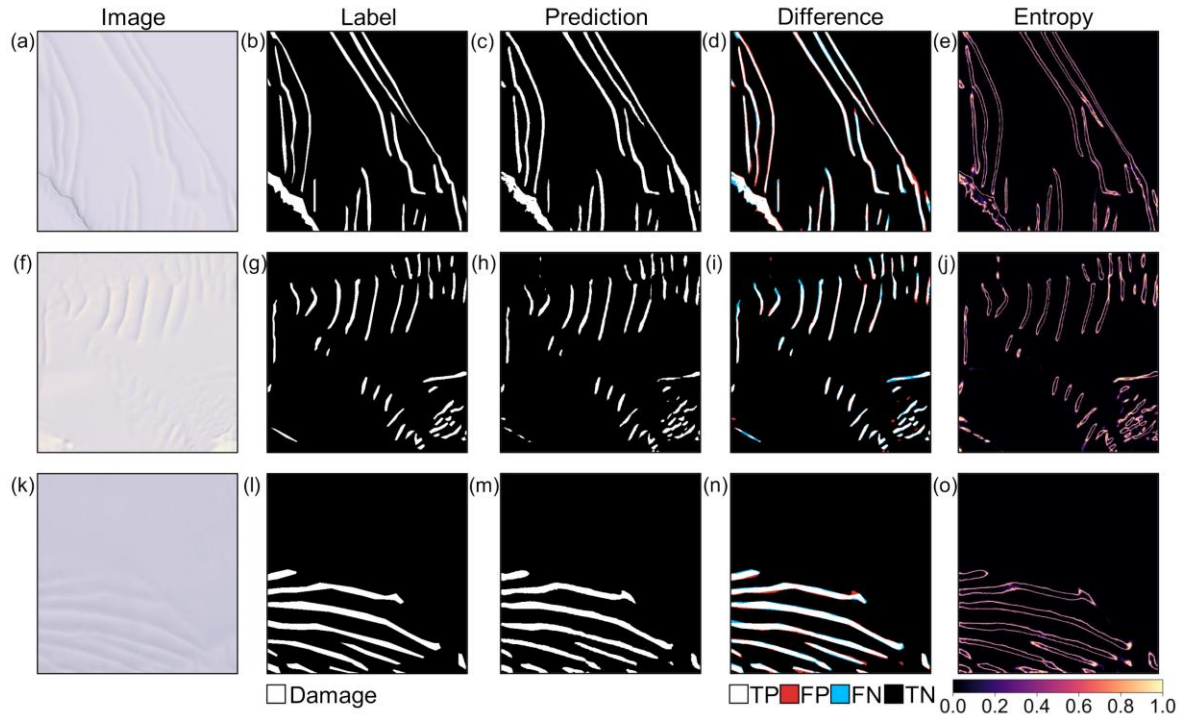

Figure 10. Examples of prediction performance on samples from the Landsat 7 SLC-on cross-sensor test set. (a), (f) and (k) show image samples from Brunt Ice Shelf, Dotson Ice Shelf and Thwaites Ice Shelf, respectively. (b), (g) and (l) show the corresponding manually annotated labels. (c), (h) and (m) show the corresponding model predictions. (d), (i) and (n) show the pixel-wise difference maps between the corresponding labels and predictions, where TP denotes true positive, FP denotes false positive, FN denotes false negative and TN denotes true negative. (e), (j) and (o) show the corresponding prediction entropy maps.

In the evaluated examples, we did not identify a distinct systematic false-positive pattern attributable to frost patches. Instead, boundary delineation was the dominant source of error. Other factors affecting damage extraction, including illumination variations and surface meltwater, are also discussed in the revised manuscript. We have expanded the Discussion by adding the following content:

However, the identification of individual damage features in the annual products may still be affected by seasonal snow bridging over crevasses and variations in imaging conditions such as illumination angle. The generally good multi-temporal consistency of the dataset suggests that these variations do not dominate the mapped temporal patterns……. Another aspect of uncertainty concerns the delineation of damage boundaries. Inspection of the difference maps (Figs. 8–10) shows that most false-positive and false-negative pixels occur along damage boundaries, which is also reflected by the higher prediction uncertainty along these boundaries in the entropy maps.

3. In several of the maps of results you _are_ showing surface damage above bottom crevasses, and this stems from surface ruptures initiated by deformation around bottom crevassing (there are many papers to cite here – one is by McGrath for the Larsen C). This is very prevalent on Scar Inlet (your 'Larsen B' test area) and the northeastern Larsen C, as well as in the floating parts of Thwaites Western Ice Tongue. This kind of damage is especially common where the ice is thick (in other words, where buoyancy and lateral deformation forces are large).

McGrath, D., Steffen, K., Scambos, T., Rajaram, H., Casassa, G. and Lagos, J.L.R., 2012. Basal crevasses and associated surface crevassing on the Larsen C ice shelf, Antarctica, and their role in ice-shelf instability. *Annals of glaciology*, *53*(60), pp.10-18.

Response: Thank you for your constructive comment.

We agree that some of the mapped surface damage may represent surface expressions associated with basal crevasses. As noted by McGrath et al. (2012), basal crevasses can induce surface deformation and associated surface crevassing through hydrostatic adjustment and the resulting bending stresses. These surface expressions can therefore be captured in our Landsat-based surface damage maps, but are not separately distinguished from other types of surface damage in this dataset.

We have clarified this point in the Discussion by adding the following text:

Surface expressions associated with basal crevasses may also be included in the maps, as basal crevasses can induce surface deformation and associated surface crevassing through hydrostatic adjustment and the resulting bending stresses (Mcgrath et al., 2012).

4. It would be interesting to consider how to modify the algorithm or segmentation pre-processing to facilitate bottom crevasse detection. This would likely involve a larger cluster of pixels in the segmentation regions and inclusion ice thickness mappings for scaling the detection of the surface undulations that mark bottom crevasses. There are also notable differences in bottom crevasse surface manifestations between warm-water cavities (Bellingshausen-Amundsen Sea coasts) having to do with basal erosion due to melting.

Response: Thank you for your constructive suggestion.

We agree that incorporating broader spatial context and auxiliary information such as ice thickness could be useful for specifically identifying surface expressions associated with basal crevasses, particularly given that their manifestations may vary under different ice-shelf and basal melt conditions.

However, developing a dedicated basal-crevasse detection framework would require additional data and methodological design beyond the scope of the present study, which focuses on consistent multi-decadal mapping of observable surface damage from Landsat imagery. We therefore regard this as an important direction for future work and have added a brief statement to the Discussion, as follows:

Future work could further develop damage-type classification by incorporating broader spatial context and auxiliary glaciological information, such as ice thickness.

5. A little more detail on what was manually identified as surface damage would be good to describe. This might only require a few sentences in Section 2.3.1.

Response: Thank you for your constructive suggestion.

We have added a clearer description of the definition and annotation criteria of surface damage in Sect. 2.3.1, as follows:

These features were annotated as a single damage class. Surface damage was identified in the Landsat imagery based on distinct linear or curvilinear surface discontinuities and densely fractured areas characterized by clearly disrupted surface patterns. In contrast, general surface roughness and other gradual variations in surface appearance were not labelled as damage in the absence of distinct surface discontinuities or clearly disrupted patterns.

6. In your temporal consistency checks (Section 3.2.5 and 3.3), did you look for seasonal variations in damage extent or detectability? Seasonal variations in snow bridging of crevasses should lead to some reduced evidence of damage in spring, greater evidence in winter (you are close to evaluating this in your Figure 10). ALSO, importantly, was the algorithm affected by solar illumination angle? Low angle autumn images may well reveal more extensive areas exceeding the damage threshold. Further to this --- areas of extremely rough sastrugi might trigger the algorithm to map surface damage. You could look in some key areas of east Antarctica where meter-scale perennial sastrugi form as a test of the algorithm.

Response: Thank you for raising these points. We agree that seasonal changes in surface conditions, solar illumination and rough snow surfaces may influence the detectability of surface damage in optical imagery.

The dataset is designed as an annual product based on Landsat imagery acquired during the sunlit season. We therefore acknowledge that the present dataset does not allow a systematic assessment of the full seasonal cycle of damage detectability, including possible seasonal effects associated with snow bridging. The original Fig. 10 (now Fig. 12 in the revised manuscript) compares our optical damage mapping with a SAR-based product. Some damage features are indeed observed in the SAR data but are not apparent in the optical imagery. However, this comparison cannot isolate seasonal effects because differences in sensor characteristics are also involved. The Landsat observations do not support a systematic seasonal comparison. We also agree that solar illumination angle may influence the appearance of surface damage in optical imagery. In the present study, we did not isolate solar illumination angle as an independent factor because the timing of usable cloud-free Landsat observations varies from year to year, resulting in differences in acquisition date, illumination and surface conditions across the multi-year record. We have therefore clarified this limitation in the Discussion as follows:

However, the identification of individual damage features in the annual products may still be affected by seasonal snow bridging over crevasses and variations in imaging conditions such as illumination angle. The generally good multi-temporal consistency of the dataset suggests that these variations do not dominate the mapped temporal patterns.

The multi-temporal consistency assessment in Sect. 3.2.5 evaluates the final damage maps across the range of actual acquisition conditions represented in the dataset. The generally good consistency between temporally adjacent maps, together with the improvement in consistency after Lagrangian correction for most ice shelves, indicates that the mapped temporal variations are physically consistent with ice advection. To better acknowledge the potential influence of acquisition and surface conditions, we have revised Sect. 3.2.5 as follows:

Overall, the dataset shows good multi-temporal consistency across temporally adjacent pairs. The improvement in consistency metrics after applying the Lagrangian correction further indicates that the temporal variations captured in the dataset are physically consistent with ice advection, providing additional evidence for the reliability of the mapped changes in surface damage. This multi-temporal consistency assessment therefore supports the use of the dataset for subsequent temporal analysis, although variations in image acquisition conditions and surface conditions may still affect the detectability of individual damage features.

We also appreciate the reviewer's suggestion regarding extremely rough sastrugi. Following this comment, we carefully examined the source imagery over the ice-shelf regions included in the dataset. We did not identify extensive fields of strongly developed sastrugi comparable to those reported over the interior East Antarctic Ice Sheet. Because the released dataset is restricted to ice shelves, we considered that an additional test over inland ice-sheet region would not provide a representative assessment of errors within the spatial domain of the dataset. Nevertheless, we agree that local variations in snow-surface roughness may affect the detectability of individual damage features, and this possibility is encompassed by the limitation noted above.

Minor comments

7. In several places, you refer to the remnant Larsen B ice shelf. Consider instead describing this as 'Scar Inlet Ice Shelf', as nearly all of the remnant Larsen B is within this ice-shelf-filled bay.

Response: Thank you for your constructive comment.

We've changed references to the remnant Larsen B Ice Shelf to "Scar Inlet Ice Shelf" throughout the manuscript, including the text and figures.

8. Line 52, cite Scambos et al. (2007) or Haran et al. (NSIDC) for the MOA mosaics. Note, you might also explore the LIMA mosaic (Bindschadler et al.) and RAMP (Jezek et al.) because these are uniformly processed they make a good template for your algorithm. More importantly – run your processing on the REMA mosaic. These different mosaics (esp LIMA and REMA) will establish better baselines for your surface damage processing time-series.

Response: Thank you for your constructive comment.

We've added the suggested citation for the MOA mosaic (Scambos et al., 2007).

We also considered the suggested LIMA, RAMP and REMA products. However, their image characteristics and spatial resolutions differ substantially from the Landsat imagery used for model training, making direct application of the current model inappropriate. We therefore did not include these mosaics in the present analysis.

9. Line 133 – picky, but don't say 'austral summer', say 'sunlit season' or similar.

Response: Thank you for your constructive comment.

We have replaced “austral summer” with “sunlit season” when referring to the Landsat image-acquisition period and revised the corresponding text and figure captions accordingly. We have also revised the description in Sect. 2.1.2, as follows:

Owing to polar night conditions in Antarctica, the acquisition window for usable optical imagery is limited to the sunlit season, extending from November to March of the following year, broadly corresponding to the austral summer period. In the following sections, imagery for a given year refers to imagery acquired during the sunlit season starting in that calendar year.

10. Figure 7 – what is the scale of these image chips? More generally, it would be good to see a close-up view of a test area to see -exactly- what the algorithm is mapping.

Response: Thank you for your constructive comment.

These image chips are 512×512 pixels at a spatial resolution of 30 m, providing close-up examples of the surface features mapped by the segmentation model. In addition, Fig. S17 provides an example using aerial imagery to illustrate the surface characteristics of an area mapped as extensive damage.

11. Very minor note – I found the lack of spacing or indentation on the paragraphs and the references to be an impediment to quick reading and checking. I suggest adopting some kind of structure to facilitate a quicker scan.

Response: Thank you for your constructive comment.

The manuscript formatting follows the ESSD manuscript template and formatting requirements. We have therefore retained the prescribed formatting in the revised manuscript.