Enhancing existing Remote-sensing Datasets with weakly supervised Deep Learning: A Case Study on Antarctic Rock outcrops

Felix Dahle[1], Roderik Lindenbergh[1], and Bert Wouters[1]

[1]Dept. of Geoscience & Remote Sensing, Delft University of Technology, the Netherlands

Keywords: Rock outcrops, Cryosphere, Semantic Segmentation, U-Net, Machine learning, Antarctica

Abstract:

Accurate mapping of exposed rock is fundamental for cryospheric and geospatial analyses in Antarctica, yet existing products are of limited resolution and tend to underestimate true rock exposure. We present a weakly supervised deep-learning framework that refines existing rock masks by combining Sentinel-2 multispectral imagery with elevation and slope data from the Reference Elevation Model of Antarctica (REMA). A U-Net with eight input channels (six spectral bands, elevation, slope) is trained using imperfect Landsat- and GeoMap based labels. Trained on data from the Antarctic Peninsula, the model produces a 10 m rock mask that delineates small and shaded outcrops more effectively than existing datasets. While quantitative evaluation is constrained by imperfect reference data, qualitative inspection indicates improved rock–snow separation. The workflow is fully automated, requires no manual annotation, and scales efficiently to all rock-hosting regions of the continent reachable by Sentinel-2 multispectral coverage. Beyond rock mapping, the framework is transferable to other scenarios with incomplete or uncertain reference data, such as vegetation, snow, or water mapping. The resulting rock mask for complete Antarctica, together with the trained model and preprocessing scripts, will be released to support reproducible large-scale mapping and future cryospheric research.

1. Introduction

Accurate delineation of exposed rock in Antarctica is the basis for a wide range of geoscientific applications. Rock outcrops provide the only long-lived, stable land surfaces within a dynamic cryosphere and are essential for (i) geodetic and photogrammetric workflows (e.g., co-registration of historical imagery and Structure from Motion (SfM) via stable ground control) points (Child et al., 2021), (ii) geological mapping and stratigraphic synthesis (Cox et al., 2023), (iii) ecological analyses requiring potential ice-free habitats (Lee et al., 2017), and (iv) reference areas for glacier change detection, ice-velocity validation, and mass-balance estimation (Li et al., 2018). A consistent, higher-resolution rock mask is therefore valuable across disciplines from geomorphology to climate monitoring.

Several rock-mask products for Antarctica already exist (Table 1). Early continent-wide information was provided through the SCAR Antarctic Digital Database (ADD), where rock exposure polygons were manually digitized from published maps (Thomson and Cooper, 1993). Building on this foundation, Burton-Johnson et al. (2016) introduced an automated classification based on Landsat-8, using a semi-automated thresholding approach designed to minimize misclassification of shaded ice or snow pixels. This dataset remains widely used and is distributed, for example, through Quantarctica (Matsuoka et al., 2021). More recently, GeoMap (Cox et al., 2023) compiled exposed bedrock and surficial geology from multiple sources, updating many polygons from the ADD. Meanwhile, the British Antarctic Survey continues to maintain the ADD, most recently integrating heterogeneous data from the 1960s–2023 (Gerrish et al., 2024).

Existing products differ not only in resolution and data sources but also in their definition of rock exposure and show sometimes substantial variation in mapped extent and boundary interpretation (as shown in Figure 1). Some delineate rock conservatively, omitting smaller or shaded rock outcrops, whereas others generalize more broadly, extending into mixed or debris-covered terrain. In addition, the original ADD database has known problems with geo-referencing (Burton-Johnson et al., 2016), complicating spatial alignment and cross-product comparison. These inconsistencies introduce uncertainty on accurate rock–ice separation.

Dataset	Year	Method
Johnson	2016	Automated Landsat-8 classification using NDSI and spectral thresholds
GeoMap	2023	Compilation of geological maps and field data
ADD	2024	Digitized and manually edited outcrops from topographic and remote sensing sources

Table 1: Major Antarctic rock-exposure datasets with year of latest update.

While previous efforts have primarily sought to generate new classifications from raw imagery and maps, this study focuses on refining existing datasets using an automated weakly supervised deep-learning framework. Our objective is to produce a more complete rock mask by leveraging Sentinel-2 multispectral imagery and topographic data from the Reference Elevation Model of Antarctica (REMA) digital elevation model (DEM) (Howat et al., 2019). Advances in semantic segmentation architectures such as the development of U-Nets enable robust integration of these multi-source inputs (Zhang et al., 2023). When trained with strategies that account for label noise, such models can learn effectively from imperfect reference data and generate classifications that capture fine-scale outcrops while maintaining the conservative characteristics of the existing datasets.

The proposed pipeline is first trained and validated on the Antarctic Peninsula (AP), which hosts the highest diversity of outcrops, and is then applied continent-wide to produce a 10 m-resolution probabilistic rock-exposure map of Antarctica. The

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-671-2026

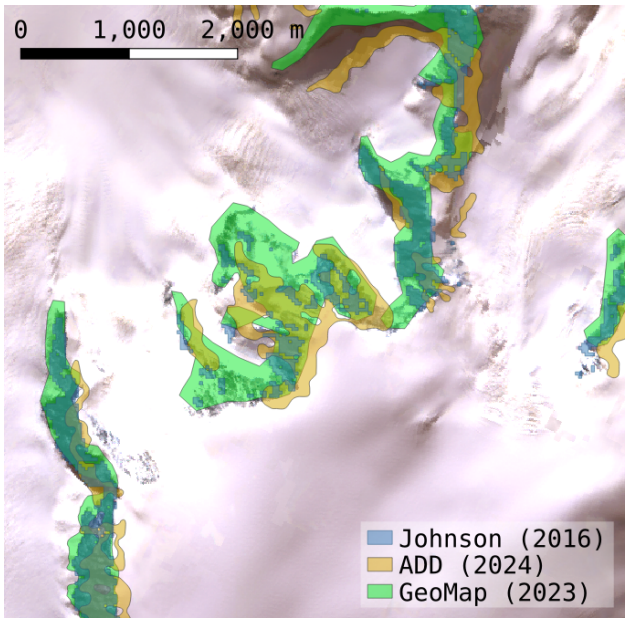

Figure 1: Example of differences in mapped rock extent between the GeoMap (green), ADD (orange), and Johnson (blue) rock masks.

trained model and accompanying preprocessing scripts are designed for reproducibility and adaptability, allowing users to apply the workflow to regional datasets or analogous mapping tasks in other regions.

2. Related work

The availability of higher-resolution satellite imagery (e.g., Sentinel-2) combined with advanced digital elevation models (e.g., REMA) has paved the way for more detailed land cover mapping in polar regions. At the same time, deep learning-based semantic segmentation has emerged as a powerful tool for pixel-wise classification in remote sensing (Zhu et al., 2017). The U-Net architecture, originally introduced by Ronneberger et al. (2015) for biomedical image segmentation, has proven particularly effective due to its ability to capture both contextual and fine-grained features through its encoder-decoder structure and skip connections. This design has led to its successful application in various remote sensing tasks, including general land cover classification (Deressu et al., 2025), semantic segmentation of historical images (Dahle et al., 2024), or detailed mapping of cryospheric features such as melt ponds (de Roda Husman et al., 2024) and sea ice types (Zhang et al., 2022). Regional deep-learning studies on other cryospheric landforms, such as automated rock-glacier detection by Robson et al. (2020), demonstrate that fusing multispectral, radar coherence, and DEM features significantly improves performance, underscoring the value of incorporating diverse data for robust classification in complex terrain.

While deep learning models have achieved remarkable success in remote sensing, their performance strongly depends on the availability of large, accurately annotated training datasets. Acquiring such dense and high-quality labelled data is often difficult and costly, particularly in polar and remote regions where manual annotation is time-consuming and ambiguous due to variable illumination and seasonal snow cover. To mitigate this dependency on perfectly clean reference data, weakly supervised learning (WSL) and robust training strategies have gained increasing attention. Approaches such as self-training (iteratively training with previous model results as new labels) and teacher–student consistency frameworks using the exponential moving average (EMA) (Tarvainen and Valpola, 2017) allow neural networks to learn effectively from noisy or incomplete labels. In this setup, the student network is updated through standard gradient descent, while the teacher network maintains an EMA of the student’s parameters. The teacher’s predictions serve as temporally smoothed pseudo-labels, stabilizing training and reducing sensitivity to local noise in the supervision data.

These methods refine imperfect annotations or exploit unlabelled data by enforcing consistent model predictions under different augmentations or model states. In remote sensing, WSL has been successfully applied to diverse tasks, including large-scale land cover mapping with sparse training data (Wang et al., 2020) and, within the cryosphere, to challenges such as cloud, shadow, and snow detection (Nambiar et al., 2022) and glacial lake mapping (Tan et al., 2025). Collectively, these studies demonstrate that weak supervision can yield high-quality, pixel-level classifications even when the available reference data are incomplete or contain systematic errors.

3. Input data

For training and validation of the updated Antarctic rock mask, we use three publicly available datasets: Sentinel-2 multispectral imagery, the REMA elevation, and the already existing continental-scale rock exposure masks. Together, these datasets provide high-quality spectral and topographic information that enable large-scale classification across the Antarctic continent (Table 2).

Dataset	Resolution	Temporal coverage
Sentinel-2	10 m-20 m	2015-2025
REMA	2 m, 8 m, 10 m, 32 m	2009-2017
Rock mask	30 m	2016 and 2023

Table 2: Primary input datasets used in this study, including their spatial resolution and temporal coverage.

Sentinel-2 imagery, provided by the European Space Agency (ESA) within the Copernicus program, offers multispectral coverage of Antarctica at 10–20 m spatial resolution (Drusch et al., 2012). We employ a median composite derived from all available Level-2A surface reflectance products between 2015 and 2025. From the 13 available spectral bands, we use Bands 2, 3, and 4 (visible), Band 8 (near-infrared), and Bands 11 and 12 (shortwave infrared). The visible and near-infrared bands are available at 10 m resolution, while the shortwave infrared bands (originally 20 m) are resampled to 10 m to ensure uniform input dimensions. This combination captures key spectral contrasts between rock, snow, and ice: rock surfaces typically exhibit lower reflectance in the visible range and higher reflectance in the shortwave infrared. All Sentinel-2 data were reprojected to a polar stereographic grid (EPSG:3031) for consistency with the other datasets. Due to the orbital inclination of the Sentinel-2 satellites, a ’pole hole’ exists where no optical imagery is acquired south of approximately 82.7°S. Consequently, the multispectral input—and thus the resulting rock mask—is geographically constrained to the area within the Sentinel-2 footprint.

Topographic information is provided by the REMA, a continent-wide digital surface model derived from stereoscopic

https://doi.org/10.5194/isprs-annals-XI-3-2026-671-2026

optical imagery between 2009 and 2017. REMA consists of individual DEM strips, each vertically registered to altimetric measurements from CryoSat-2 and ICESat, resulting in a vertical uncertainty below 1 m. We use the 10 m-resolution mosaicked product, in which elevation values represent the median of all overlapping DEM strips to minimize outlier influence and ensure a smooth, internally consistent surface. Elevations are expressed in meters relative to the WGS84 ellipsoid. Additionally, slope was derived from elevation and used as additional input data. Incorporating topographic data supports the discrimination between rock and snow-covered surfaces, as exposed bedrock commonly occurs on steep slopes and ridgelines, whereas ice and snow dominate lower, flatter terrain.

As reference data for training, we combined two complementary continent-wide rock-exposure datasets: the Landsat-based product of Johnson (Burton-Johnson et al., 2016) and the GeoMap compilation (Cox et al., 2023). The Johnson mask provides consistent coverage and conservative delineation, while GeoMap contributes more detailed and manually updated polygons. Both datasets are available as vector-based shape-files, however as the first mask is based on Landsat-8 with a pixel-size of 30m, it stills retains a pixelated appearance.

4. Methodology

The training follows a classical deep-learning approach for semantic segmentation using a U-Net architecture.

4.1 Automated data tiling and preprocessing workflow

To prepare the datasets for model training, the AP was subdivided into tiles of 5.2 km × 5.2 km, defined from the origin (0,0) in polar stereographic projection (EPSG:3031). This region was chosen because it contains the highest density and diversity of exposed rock outcrops across the continent, providing a representative and information-rich training area for model development. The tiling scheme ensures a regular grid that can be directly mapped to the input size of the U-Net architecture and enables large-scale, tile-based training. The selected tile size is a compromise between computational efficiency and spatial context, allowing the network to capture both local surface textures and regional topographic gradients. This procedure resulted in 5,980 tiles possible to use for training and validation.

For each tile, Sentinel-2 imagery was automatically extracted using a Python workflow connecting to the Google Earth Engine (GEE) API (Gorelick et al., 2017). The script gets all Level-2A surface reflectance scenes between 2015 and 2025, masks cloudy pixels, and computes the median reflectance of all remaining observations. The resulting cloud-free composite is then downloaded, reprojected to EPSG:3031, and clipped to the tile extent. This automated workflow enables reproducible, large-scale data preparation without manual intervention.

The REMA mosaics were downloaded from the Polar Geospatial Center and divided into matching tiles using the same grid definition as for the Sentinel-2 data, ensuring spatial alignment.

The reference rock mask shapefiles were rasterized to the same 10 m grid, producing a binary label raster in which rock pixels were assigned a value of 1 and all other surfaces pixels a value of 0. Although the polygons from the dataset of Johnson were derived from 30 m Landsat data, this upscaling is unproblematic for model training, as the coarser boundaries remain informative at higher resolution. To obtain a single, consistent reference layer, the Johnson and GeoMap masks were merged through a weighted average with equal weights (0.5 each), ensuring that both sources contributed equally to the final rock label while preserving their complementary characteristics. Both the Sentinel-2 reflectance bands and REMA elevation values were normalized prior to training: spectral bands, DEM and slope were all scaled to the 0–1 range.

To ensure meaningful training samples, only tiles containing at least one rock pixel were retained from all available tiles across the AP. Tiles were ranked by their proportion of rock coverage, and the 2400 tiles with the highest proportion were selected to focus the training on geologically relevant areas. Even within this subset, some tiles contained less than 0.1% rock coverage. In total, the selected tiles cover approximately 64,900 km² of the AP. Figure 2 illustrates representative examples of the selected tiles.

4.2 Network architecture

We employ a U-Net–based convolutional neural network for semantic segmentation of rock outcrops (Table 3). The architecture is well suited for mapping sharp terrain boundaries from multispectral and elevation data, while remaining computationally efficient. Conceptually, the network consists of an *encoder* that captures multi-scale contextual features and a *decoder* that reconstructs pixel-level predictions. Skip connections link corresponding layers to preserve fine details such as rock–snow edges. The overall network structure is illustrated in Figure 3, showing the symmetric encoder–decoder layout with convolutional blocks, max-pooling, bilinear upsampling, and skip connections between corresponding resolution levels.

Each input tile contains eight channels: six Sentinel-2 bands (B02, B03, B04, B08, B11, B12), REMA elevation and the derived slope. The model outputs a per-pixel probability between 0 and 1, which can be thresholded to produce binary rock masks. This probability output allows flexible use depending on the desired balance between omission and commission errors.

U-Net’s multi-scale representation is particularly effective for Antarctic landscapes, where small, high-contrast rock outcrops are embedded in large uniform snow or ice fields. Bilinear upsampling in the decoder ensures smooth reconstructions without artifacts. The model was implemented in `PyTorch` and trained on 520 × 520 pixel tiles (5.2 km at 10 m).

Parameter	Value
Input channels	8 (6 Sentinel-2 bands + elevation + slope)
Output channels	1 (Per pixel rock probability)
Architecture	U-Net encoder–decoder with skip connections
Upsampling	Bilinear interpolation
Framework	PyTorch

Table 3: Main architectural and implementation parameters of the U-Net segmentation model.

4.3 Training strategy

The network was trained on 520 × 520 pixel tiles (5.2 km × 5.2 km at 10 m resolution) containing the eight aforementioned input channels. A stratified split (80% training, 20% validation) preserved the distribution of rock coverage across both subsets, ensuring that tiles with varying degrees

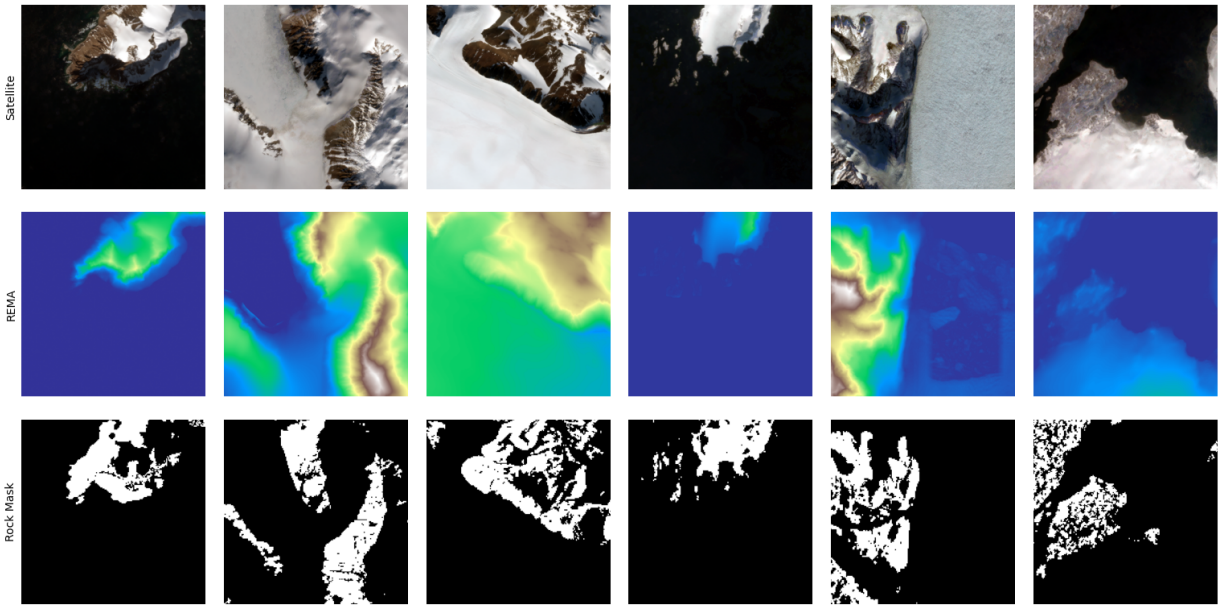

Figure 2: Example input data used for model training. Each column represents one training tile, showing the Sentinel-2 RGB composite (top), REMA elevation model (middle), and corresponding reference rock mask (bottom), where white denotes rock exposure and black indicates non-rock surfaces.

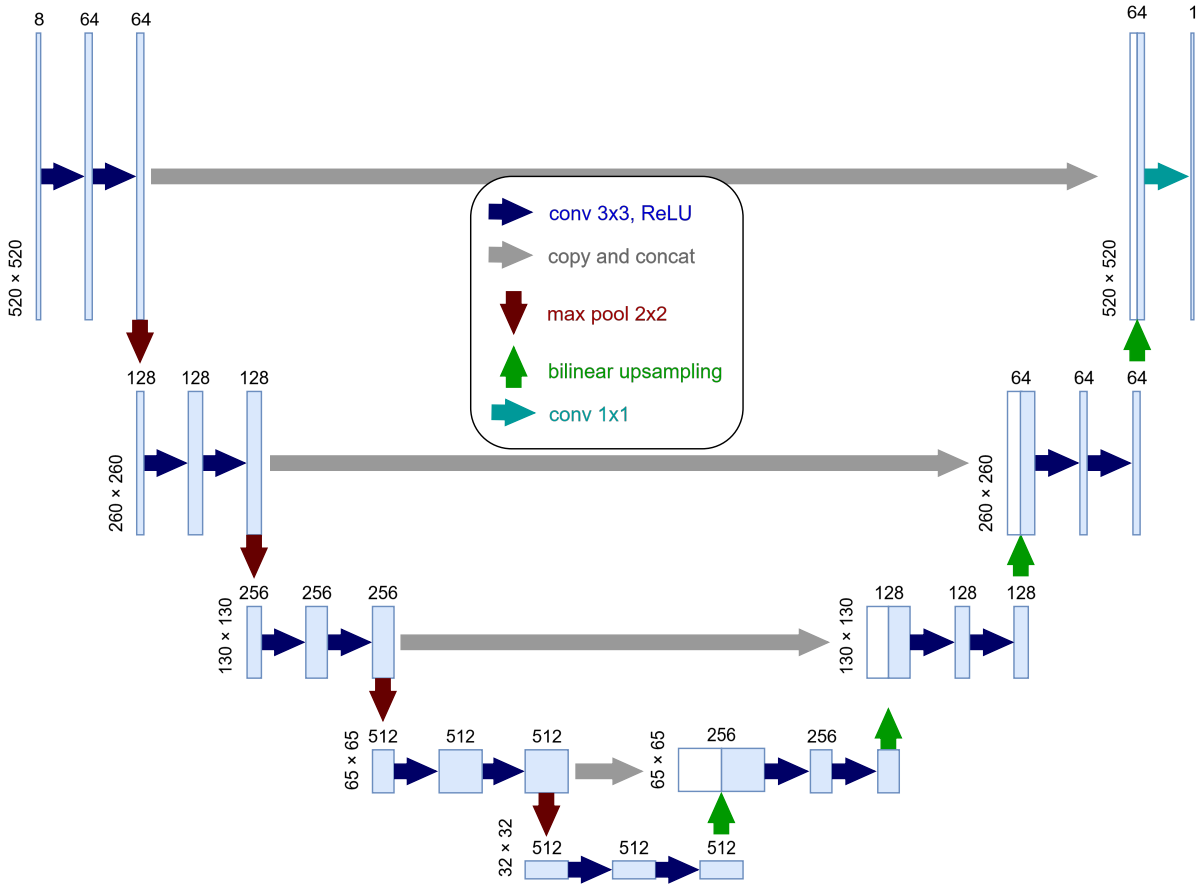

Figure 3: Architecture of the U-Net model used for rock-outcrop segmentation. The encoder (left) extracts multi-scale contextual features through convolution and pooling, while the decoder (right) reconstructs full-resolution predictions via bilinear upsampling and skip connections.

https://doi.org/10.5194/isprs-annals-XI-3-2026-671-2026

of rock exposure were proportionally represented in each subset. This strategy promotes balanced learning and reliable performance assessment across diverse surface conditions.

To enhance generalization and reduce overfitting to illumination conditions, data augmentations were applied using the `Albumentations` library (Buslaev et al., 2020). Each tile and its corresponding mask were randomly flipped horizontally or vertically and rotated by 90° (probability = 0.5). Additionally, moderate brightness and contrast adjustments were applied with a lower probability (0.2) to simulate varying illumination scenarios. These augmentations introduced rotational and reflectional invariance and improved the model's robustness to differences in lighting and acquisition geometry.

Because the rock class is strongly under-represented, an adaptive weighting was applied to the positive rock class in the loss function. This weight was derived from the ratio between negative and positive pixels in the training data and capped at 5 to avoid instability. The cap limits the maximum contribution of rock pixels to five times that of non-rock pixels, preventing excessive gradient imbalance while still compensating for class scarcity.

The loss function combined binary cross-entropy (80%) and Tversky loss (20%) to balance overall pixel accuracy and the delineation of small rock areas. Binary cross-entropy stabilized the optimization and favored consistent, conservative predictions, limiting false detections in uniform snow and ice regions. The Tversky term complemented this by emphasizing overlap between predicted and reference areas, which improved boundary precision and helped recover small or partially shaded rock outcrops that were often missing in the training data. A summary of the main training parameters and settings is provided in Table 4.

Category	Setting
Optimization	AdamW optimizer, initial learning rate: 1×10^{-4}, batch size: 4
Loss and weighting	Combined loss: 0.8 BCE + 0.2 Tversky; positive-class weight adaptive (capped at 5); FP > 0.95 down-weighted by 0.75
Regularization / stability	EMA decay $\alpha = 0.996$; consistency weight λ_c ramped $0 \rightarrow 5$ over first 5 epochs; early stopping based on (validation loss + consistency term)
Data augmentation	Random flips, 90° rotations, brightness + contrast adjustments via `Albumentations`
Training setup	Input size 520×520 px; 80/20 train–validation split; rock-tile selection ≥ 1 rock pixel

Table 4: Key training hyper-parameters of the U-Net model grouped by their functional role.

Adaptations for training with incomplete labelled datasets

The original rock mask by Burton-Johnson et al. (2016) provides high-quality supervision but is not error-free, as especially some outcrops are not completely labelled (see Figure 1). To address these errors, the training strategy was adapted to account for incomplete or uncertain labels.

The most important adaptation is a confidence-aware weighting scheme that reduces the penalization of potential false negatives in the reference data. Predictions classified as rock with high confidence but labelled as non-rock in the original dataset were down-weighted by 80% during loss computation. This adjustment prevents the model from being penalized for detecting true rock exposures that are absent from the existing labels, enabling the discovery of new, previously unmapped outcrops without destabilizing training.

To further mitigate label noise and improve temporal consistency, a teacher–student framework based on EMA weights was employed. After each epoch, both networks were evaluated, and early stopping was guided by a combined stability criterion (the sum of the validation loss and the mean student–teacher consistency term) ensuring selection of models that were both accurate and temporally stable rather than those merely minimizing loss on imperfect labels.

After the initial training round, high-confidence predictions from the teacher network were merged with the original rock mask to generate an updated set of soft labels. This pseudo-labelling step further reduced dependence on the original dataset and improved the completeness and spatial continuity of rock delineation. The model was then retrained using these updated labels, allowing it to iteratively refine its predictions and progressively improve segmentation quality.

4.4 Continent-wide rock mask

After training, the final model was applied to generate a high-resolution rock mask. For this purpose, the Antarctic continent was subdivided into tiles of 520×520 pixels at 10 m resolution. The tiling grid covered the full extent of the continent with an overlap of 60 pixels between adjacent tiles to minimize edge effects. While the REMA DEM provides true continent-wide coverage, the final classification is geographically constrained by the Sentinel-2 orbital inclination, which excludes the region south of approximately 82.7°S. To avoid unnecessary computation over ocean areas, only tiles with valid REMA data were processed. Since REMA extends slightly beyond the coastline, this approach ensures the inclusion of all nearshore regions without introducing gaps at the land–sea boundary. This procedure resulted in a total of 59,237 tiles. While it is generally preferable to evaluate models on independent datasets, the primary focus here was on producing the most complete rock mask possible rather than maximizing generalization performance. The workflow remains fully automated, from the on-demand retrieval of Sentinel-2 scenes to the final tile-based inference, requiring no manual intervention.

5. Results

Training was conducted on a workstation equipped with an NVIDIA Quadro RTX 5000 with 16 GB Video Memory and was performed in two stages as part of the iterative self-training strategy (Table 5). In total the training took around 9 h.

Stage	Label source	Time
1	Original (50% Landsat + 50% GeoMap)	4h
2	Mixed (25% Landsat based + 25% GeoMap + 50% Pseudo labels)	5h

Table 5: Summary of the two training stages in the iterative self-training process.

The final quantitative assessment evaluates the performance of the last training stage through pixel-wise comparison between probabilistic model predictions (thresholded at 0.9) and the two reference rock masks. Given the strong class imbalance and missing labels in the reference data, the evaluation emphasizes rock-detection completeness (recall) while also reporting accuracy, precision, F1-score, and IoU. As shown in Table 6, the model achieves higher recall against the Johnson mask, whereas precision is higher against GeoMAP.

Reference	Acc.	Prec.	Rec.	F1	IoU
Johnson (Landsat)	0.991	0.33	0.78	0.46	0.30
GeoMap	0.986	0.56	0.45	0.50	0.33

Table 6: Quantitative evaluation of the predicted rock mask against the masks of Johnson and GeoMap. Values are aggregated over all non-trained tiles of the AP.

To complement the quantitative evaluation, we visually inspected numerous tiles across the AP to assess the spatial and contextual reliability of the predictions. Figure 4 presents a selection of representative examples. The first row shows that shadowed areas are correctly identified as rock exposure, demonstrating robustness to illumination differences. The second row illustrates how the model distinguishes dark water bodies from shadowed or low-albedo rock surfaces, which often appear similar in visual imagery. The third row highlights the model’s capacity to capture small and fragmented outcrops with fine detail. The fourth row shows consistent performance over extensive rock exposures, confirming that large contiguous areas are handled effectively as well.

6. Discussion

The presented workflow produced a high-resolution rock-exposure mask for the AP, addressing limitations of previous datasets such as incomplete small outcrops, inconsistent boundaries, and local misalignments. By integrating Sentinel-2 multispectral imagery with topographic data from the REMA, the resulting product offers a more complete and spatially coherent depiction of exposed bedrock.

Although the model achieves only moderate quantitative scores when compared to the reference masks (Table 6), these values must be interpreted with care. Both existing datasets systematically underestimate true rock coverage, so many apparent false negatives actually represent valid small outcrops missing from the reference data. Conversely, in GeoMap, some polygons were drawn generously or exhibit slight spatial offsets, inflating apparent false positives. Consequently, the numerical metrics primarily reflect inconsistencies in the supervision data rather than genuine model errors, and the qualitative evaluation provides a more realistic measure of performance. Visual inspection across numerous tiles confirms that the predicted masks align closely with actual rock exposures, capturing fine-scale features and illumination variations far more reliably than the reference datasets.

Weakly supervised deep learning is central to this improvement. Starting from incomplete existing labels, the model progressively refined its understanding of rock signatures through self-training on confident predictions. This iterative process allowed it to generalize beyond the constraints of its supervision source and to recover previously unmapped rock exposures. A third self-training round was tested but did not further improve results. Instead, the model began to overfit, showing increased noise and exaggerated rock predictions.

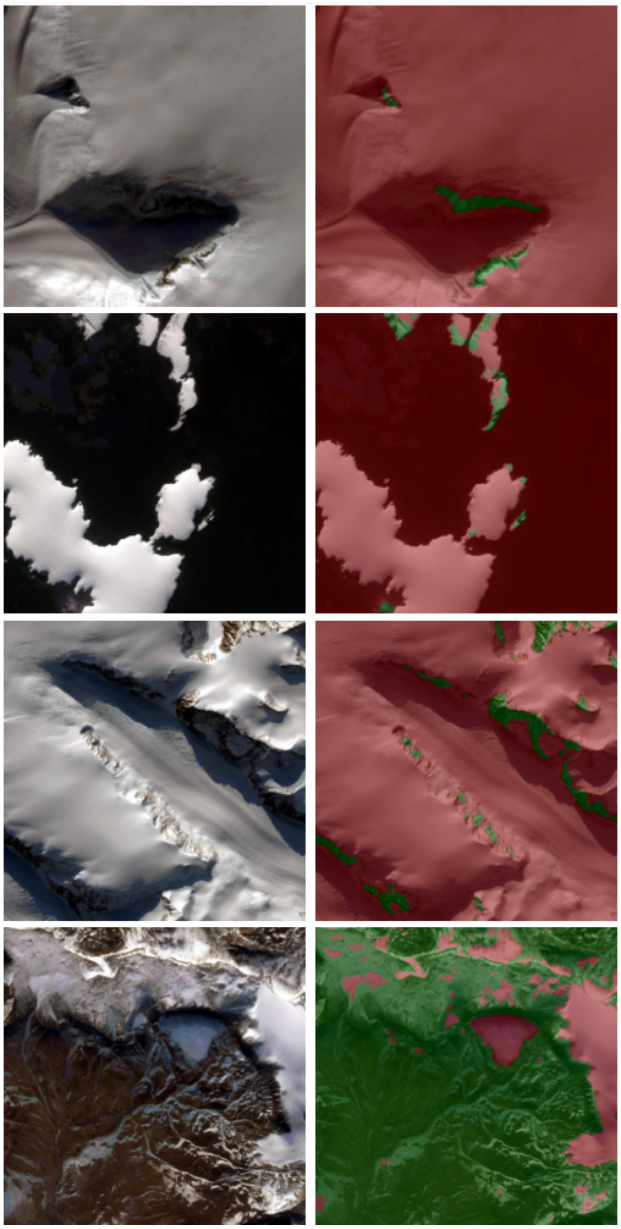

Figure 4: Representative visual evaluation examples. Each row shows a Sentinel-2 RGB composite (left) and the corresponding prediction from our model (right) using a probability threshold of 0.9. Green areas indicate rock exposure, while red areas denote non-rock surfaces.

Equally important was the confidence-aware weighting strategy: by reducing the loss contribution of highly confident pixels mislabelled as non-rock, the model was free to propose new rock detections while retaining stability. Without such selective down-weighting, the network would remain tied to the limitations of the input data.

The inclusion of the teacher–student framework based on EMA weights further stabilized learning by averaging model parameters over time. This temporal regularization suppressed local noise and reduced the influence of spurious predictions. As a result, the model achieved spatially smoother and more consistent delineations, especially in complex or shadowed terrain (see Figure 5).

Furthermore, a simple yet effective addition was the slope layer derived from REMA elevation. It improved the model’s ability

https://doi.org/10.5194/isprs-annals-XI-3-2026-671-2026

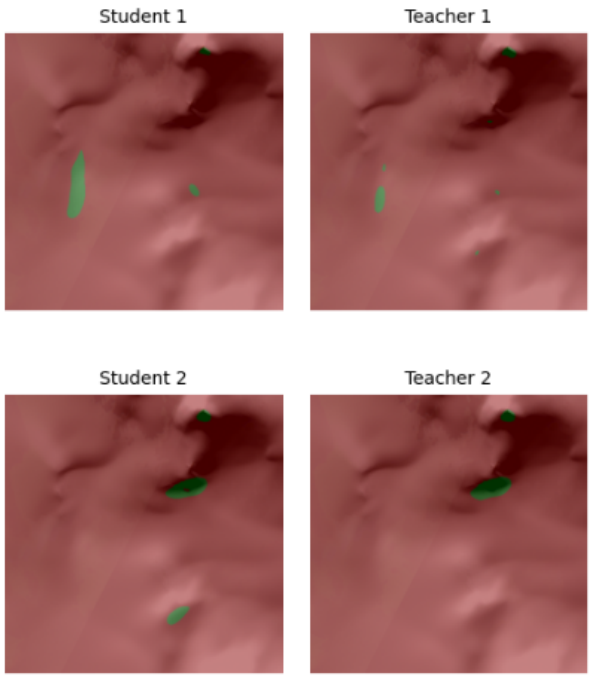

Figure 5: Top row: results from the first training stage; Bottom row: represents the subsequent stage using high-confidence pseudo-labels. In each pair, the left image shows the student and the right the teacher network. The teacher’s temporal averaging helps suppress local false detections, while the inclusion of pseudo-labels enables the detection of previously unmapped rock outcrops. Green areas indicate rock exposure, while red areas denote non-rock surfaces.

to discriminate rock from snow and ice in spectrally ambiguous regions. Because slope computation is computationally inexpensive and available for all DEMs, this represents a low-cost enhancement applicable to other segmentation tasks.

6.1 Limitations

Despite the improved spatial completeness, several limitations remain. Spatially, the model occasionally overestimates rock extent at the transition between rock and snow (Figure 6), likely reflecting uncertainty in the 30 m training masks where boundaries were coarsely digitized. Conversely, very small isolated outcrops may be missed where spectral or topographic contrast is weak. While tile-boundary artifacts were effectively suppressed using a 64-pixel overlap strategy, occasional false positives can still occur in regions with extreme cloud-shadow contamination that mimics the spectral signature of dark rock.

From a validation perspective, a key limitation is the inherent circularity of using ”noisy” reference datasets for quantitative assessment (Table 6). Because existing products systematically underestimate rock exposure, standard metrics like IoU and Precision are artificially penalized when the model correctly identifies true outcrops that are missing from the reference data. This suggests that qualitative inspection and manual verification remain essential for assessing performance in weakly supervised remote sensing tasks.

7. Conclusion

We present an improved, high-resolution rock-exposure mask for Antarctica derived through weakly supervised deep learning. By combining Sentinel-2 multispectral composites with REMA elevation and slope data, the model refines existing products and produces a 10 m rock mask that captures small outcrops and complex terrain boundaries more completely and consistently than previous datasets.

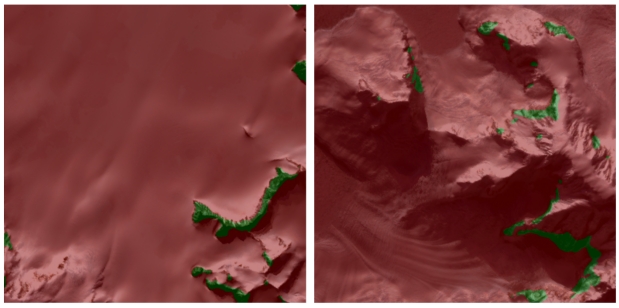

Figure 6: Examples of remaining limitations in the predicted rock masks. Both tiles omit certain small rock outcrops, while the right tile slightly overestimates rock extent along snow–rock boundaries. Green areas indicate rock exposure, while red areas denote non-rock surfaces.

A key strength of the proposed workflow lies in its scalability and efficiency. Because the framework relies on existing classification products as direct supervision sources, it entirely eliminates the bottleneck of manual annotation, making it highly adaptable to continent-wide applications.

To further mature this methodology and expand upon the current findings, subsequent research will prioritize three main objectives. First, to address the inherent challenges of evaluating against noisy reference data, we aim to develop a localized, manually digitized validation dataset. This will provide a rigorous, absolute performance baseline to better quantify the model’s true accuracy. Second, a comprehensive component analysis will be conducted to isolate and evaluate the specific performance drivers within the framework; this includes quantifying the individual contributions of the multi-modal input features (e.g., topographic versus spectral data) as well as the impact of specific architectural choices, such as the temporal ensembling. Finally, to achieve true continent-wide coverage, we plan to close the remaining observational data gap at the South Pole by integrating alternative satellite imagery and topography-only segmentation strategies.

The resulting 10 m Antarctic rock mask is publicly available via 4TU.ResearchData at `https://doi.org/10.4121/86b5f2b7-fb95-43be-893c-3b3f77f252f4`. Accompanying training models and preprocessing scripts are available at `https://github.com/fdahle/sfm-hist-rocks`, facilitating the application of this weakly supervised approach to other datasets.

Acknowledgments

This publication is part of the project Dutch Polar Climate and Cryosphere Consortium with project number ALWPP.2019.003 of the research programme NPP which is (partly) financed by the Dutch Research Council (NWO).

REFERENCES

Burton-Johnson, A., Black, M., Fretwell, P. T., Kaluza-Gilbert, J., 2016. An automated methodology for differentiating rock

from snow, clouds and sea in Antarctica from Landsat 8 imagery: a new rock outcrop map and area estimation for the entire Antarctic continent. *The Cryosphere*, 10(4), 1665–1677.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A. A., 2020. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2).

Child, S. F., Evans, A. J., Miles, V. V., Liu, L., Anderson, E., Braun, M. H., 2021. Structure-from-Motion Photogrammetry of Antarctic Historical Aerial Photographs Using Satellite-Derived Ground Control. *Remote Sensing*, 13(1), 21.

Cox, S., Lyttle, B., Elkind, S., Siddoway, C., Morin, P., Capponi, G., Abu-Alam, T., Ballinger, M., Bamber, L., Kitchener, B., Lelli, L., Mawson, J., Millikin, A., Dal Seno, N., Whitburn, L., White, T., Burton-Johnson, A., Crispini, L., Elliot, D., Wilson, G., 2023. A continent-wide detailed geological map dataset of Antarctica. *Scientific Data*, 10.

Dahle, F., Lindenbergh, R., Wouters, B., 2024. Revisiting the Past: A comparative study for semantic segmentation of historical images of Adelaide Island using U-nets. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 11, 100056.

de Roda Husman, S., Lhermitte, S., Bolibar, J., Izeboud, M., Hu, Z., Shukla, S., van der Meer, M., Long, D., Wouters, B., 2024. A high-resolution record of surface melt on Antarctic ice shelves using multi-source remote sensing data and deep learning. *Remote Sensing of Environment*, 301, 113950.

Deressu, T. F., Bojer, A. K., Debelee, T. G., Negera, W. G., Nadarajah, S., Gebissa, K. W., 2025. Enhancing land use and land cover classification with deep learning-based satellite imagery segmentation. *International Journal of Applied Earth Observation and Geoinformation*, 144, 104839.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P. et al., 2012. Sentinel-2: ESA’s optical high-resolution mission for GMES operational services. *Remote sensing of Environment*, 120, 25–36.

Gerrish, L., Fretwell, P. T., Cooper, P., 2024. High resolution vector polygons of antarctic rock outcrop – version 7.10. [Data set]. Accessed 2025-10-28.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*.

Howat, I. M., Porter, C., Smith, B. E., Noh, M.-J., Morin, P., 2019. The Reference Elevation Model of Antarctica. *The Cryosphere*, 13(2), 665–674.

Lee, J. R., Raymond, B., Bracegirdle, T. J. et al., 2017. Climate change drives expansion of Antarctic ice-free habitat. *Nature*, 547, 49–54.

Li, T., Sun, B., Zhang, N. et al., 2018. Antarctic Surface Ice Velocity Retrieval from MODIS-Based Mosaic of Antarctica (MOA). *Remote Sensing*, 10(7), 1045.

Matsuoka, K. et al., 2021. Quantarctica, an integrated mapping environment for Antarctica. *Environmental Modelling & Software*, 140, 105015.

Nambiar, K. G., Morgenshtern, V. I., Hochreuther, P., Seehaus, T., Braun, M. H., 2022. A Self-Trained Model for Cloud, Shadow and Snow Detection in Sentinel-2 Images of Snow- and Ice-Covered Regions. *Remote Sensing*, 14(8).

Robson, B. A., Bolch, T., MacDonell, S., Hölbling, D., Rastner, P., Schaffer, N., 2020. Automated detection of rock glaciers using deep learning and object-based image analysis. *Remote Sensing of Environment*, 250, 112033.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. 9351, 234–241.

Tan, H., Jiang, L., Liu, H., Zhang, T., Cheng, I., 2025. Prior knowledge-informed semantic segmentation framework for precise glacial lake mapping from multimodal imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 230, 630–643.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.

Thomson, J. W., Cooper, A. P. R., 1993. The SCAR Antarctic digital topographic database. *Antarctic Science*, 5(3), 239–244.

Wang, J., H. Q. Ding, C., Chen, S., He, C., Luo, B., 2020. Semi-Supervised Remote Sensing Image Semantic Segmentation via Consistency Regularization and Average Update of Pseudo-Label. *Remote Sensing*, 12(21).

Zhang, C., Chen, X., Ji, S., 2022. Semantic image segmentation for sea ice parameters recognition using deep convolutional neural networks. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102885.

Zhang, G., Roslan, S. N. A. B., Wang, C., Quan, L., 2023. Research on land cover classification of multi-source remote sensing data based on improved U-net network. *Scientific Reports*, 13(1), 16275.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8–36.