https://doi.org/10.5194/essd-2025-758
Preprint. Discussion started: 2 March 2026

Open Access Earth System Science Data Discussions

Shelf-Bench: A benchmark dataset for Antarctic ice shelf front and coastline delineation from multi-sensor radar satellite data

Celia A. Baumhoer[1], Amy B. Morgan[2,3], Xinyu Hou[3], Jowan L. Fromentin[4], Thorsten Hoeser[1], Andreas J. Dietz[1], Andrew Markham[3], Laura A. Stevens[2] and Claudia Kuenzer[1,5]

[1] Land Surface Dynamics Department, German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Münchner Straße, Wessling, 82234, GER
[2] Department of Earth Sciences, University of Oxford, South Parks Road, Oxford, OX1 3AN, UK
[3] Department of Computer Science, University of Oxford, Wolfson Building, Parks Road, Oxford OX1 3QD, UK
[4] Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ
[5] Chair of Remote Sensing, Institute of Geography and Geology, University Wuerzburg, Wuerzburg, Germany

Correspondence to: Celia A. Baumhoer (celia.baumhoer@dlr.de)

Abstract. Continuous monitoring of Antarctic ice shelf fronts is essential for understanding ice sheet dynamics, detecting iceberg calving events, supporting operational logistics, and generating up-to-date continental maps. However, the automated and continuous delineation of ice shelf fronts has been held back by a lack of suitable training data for deep learning models. We present Shelf-Bench, a comprehensive benchmark dataset comprising 161 manually annotated SAR scenes from three satellite sensors (ERS, Envisat, and Sentinel-1), providing spatial coverage of the Antarctic coastline with multi-temporal seasonal acquisitions spanning 1992-2021. The dataset features manually delineated masks paired with pre-processed imagery at moderate spatial resolution. Through complexity analysis, we characterize delineation challenges, including fast ice, crevassed surfaces, dense iceberg mélange, and limited spatial context. We evaluate five state-of-the-art semantic segmentation architectures, establishing baseline performance metrics. Baseline models showed strongly contrasting behaviour on Shelf-Bench: architectures that achieved higher pixel-wise accuracy tended to produce larger boundary errors, while models with better geometric precision obtained lower semantic scores. This trade-off indicates that the dataset jointly challenges ice-ocean classification and fine-scale calving front delineation, revealing complementary challenges which make it a profound benchmark for automated ice front mapping. By providing this open-access, standardized benchmark, Shelf-Bench enables accelerated development of deep learning methodologies for automated Antarctic coastline detection and supports continuous monitoring across current and future SAR satellite missions. The Shelf-Bench dataset is available at https://doi.org/10.5281/zenodo.17610870.

1. Introduction

Ice shelves are extensions of the Antarctic ice sheet. Seventy-five percent of the Antarctic coastline is fringed by ice shelves (Rignot et al., 2013), creating a very dynamic coastline that is continuously influenced by the advance and retreat of ice shelf margins and glacier tongues (Baumhoer et al., 2021; Greene et al., 2022). Ice shelves serve as a safety band for the Antarctic

https://doi.org/10.5194/essd-2025-758
Preprint. Discussion started: 2 March 2026

ice sheet, providing in large regions a buttressing force that regulates ice discharge into the ocean (Fürst et al., 2016). Consequently, the current state of ice shelf extent is critical for understanding future ice sheet mass loss and its direct impact on global sea level rise (Alley et al., 2005; Dutton et al., 2015). Monitoring these dynamic changes is essential for predicting the potential impacts of ice shelf retreat on the stability of the Antarctic ice sheet and global sea levels.

Continuous monitoring of Antarctic coastlines serves multiple critical functions. Accurate and up-to-date front positions are vital for operational logistics, enabling safe unloading of supplies at ice shelf fronts, and for scientific applications including masking model outputs, processing satellite-derived datasets, and rapidly detecting iceberg calving events. Such monitoring is essential for understanding calving dynamics and ice sheet stability, yet remains challenging across the 40,000 km Antarctic coastline due to its remoteness and complexity (Baumhoer et al., 2018; Liu and Jezek, 2004a).

Although open-access satellite data now enables continuous monitoring through automated methods, existing deep learning models for calving front delineation exhibit spatially variable accuracy, with significant error margins persisting in certain regions (Baumhoer et al., 2019; Heidler et al., 2021). Developing robust deep learning models for Antarctic coastline delineation requires high-quality, representative benchmark datasets to support robust model training and enable standardized algorithm comparison. While benchmark datasets now exist for marine-terminating glaciers (Gourmelon et al., 2022; Lu et al., 2025), equivalent standardized training and validation datasets for the Antarctic coastline remain unavailable. The availability of benchmark datasets for Greenland has substantially accelerated model development in recent years (Gourmelon et al., 2025b; Wu et al., 2024; Zhao et al., 2025). In contrast, automated deep learning-based mapping of Antarctic ice shelf fronts remains limited (Baumhoer et al., 2019, 2023; Heidler et al., 2021), representing a significant gap in the field.

To address this gap, we present Shelf-Bench: a benchmark dataset for Antarctic ice shelf front and coastline delineation comprising 161 manually annotated Synthetic Aperture Radar (SAR) scenes from three satellite sensors (ERS, Envisat, and Sentinel-1). The applicability of Shelf-Bench extends across multiple research communities. Because the dataset is built on open-access satellite imagery, glaciologists can leverage models trained on Shelf-Bench to continuously monitor ice shelf front positions in future satellite acquisitions. Models trained on the dataset can be readily applied to current and future SAR satellite missions, particularly Sentinel-1 with its consistent global coverage, enabling sustained monitoring of Antarctic ice shelf fronts well into the future. For the computer vision and machine learning community, the curated nature and accessibility of Shelf-Bench provide an ideal benchmark for evaluating and adapting existing deep learning architectures to novel cryospheric applications. Furthermore, as foundation models become increasingly prevalent in Earth observation, large-scale annotated datasets like Shelf-Bench serve as valuable training resources to improve the representation of polar regions and cryospheric processes. The availability of this standardized, multi-purpose resource will drive progress toward continuous, automated monitoring of Antarctic ice sheet dynamics, positioning Shelf-Bench as an asset for advancing glaciological research, algorithm development, and operational monitoring in an era of continuous satellite data acquisition.

https://doi.org/10.5194/essd-2025-758
Preprint. Discussion started: 2 March 2026

2. Background and Related Work

The labour-intensive task of mapping the Antarctic coastline has historically relied on both national and international collaboration. Initiated by the United States Geological Survey (USGS) in the 1990s, the Coastal-Change and Glaciological Maps of Antarctica project, in partnership with the Scott Polar Research Institute, aimed to create a comprehensive series of 24 maps detailing coastlines, ice-shelf fronts, and glacier termini through the manual digitization of selected Landsat data (Ferrigno et al., 1998; Williams et al., 1995). Liu and Jezek (2004b) further contributed to this effort by developing a semi-automated method for coastline extraction using RADARSAT-1 data. Significant contributions from NASA and the National Snow and Ice Data Center (NSIDC) resulted in the creation of the Mosaic of Antarctica, featuring manually mapped coastlines for 2004, 2009, and 2014 (Scambos et al., 2007). During the International Polar Year, additional coastline products were generated using ALOS PALSAR and Envisat ASAR imagery for 2007-2009 (Rignot et al., 2013). Today, while manual, semi-automated, and automated techniques coexist in coastline mapping, manual methods still dominate, with valuable tools like GEEDiT simplifying manual front delineation (Lea, 2018). The coastline dataset of the Antarctic Digital Database (ADD) is updated manually by the British Antarctic Survey (BAS) on a regular basis (Gerrish et al., 2024). Moreover, Greene et al. (2022) presented significant efforts in Antarctic coastline mapping and derived annual coastlines from multi-sensor satellite imagery with a semi-automated approach. Some recent studies have continued to rely on manual digitalization from Landsat imagery (Pritchard et al., 2025). Automated delineation for monitoring was introduced by Baumhoer et al. (2023) employing deep learning (DL) to detect calving front positions from Sentinel-1 imagery. As both national and international mapping initiatives diminish whilst the amount of available satellite data increases, it becomes crucial to advance the methodological development for the automated delineation of calving front boundaries.

2.1. Antarctic calving front and coastline datasets

At present, there is no comprehensive, large-scale benchmark dataset for Antarctica that supports training and evaluating deep learning models for ice shelf front and coastline delineation. A benchmark dataset does exist for glacier calving front detection, known as CaFFe, and it includes five Antarctic glaciers (Gourmelon et al., 2022). However, while this dataset is of high quality and utility for glacier studies, it covers only a limited coastal segment of the Antarctic Peninsula and does not extend to the broader Antarctic ice-shelf domain. This lack of a standardized and widely adopted dataset limits the reproducibility of methods and makes it difficult to directly compare the performance of different approaches.

Nevertheless, several valuable geospatial datasets on calving front and coastline positions are already available, many of which have been produced through manual delineation, semi-automated processing or fully automated approaches. When used with the original imagery, these datasets could serve as an important resource for DL training and validation tasks. However, their use typically requires additional preprocessing, including harmonization of spatial resolution, alignment of temporal coverage, and correction for differences in mapping conventions. Despite these challenges, such datasets represent a foundation for advancing automated monitoring efforts and could be leveraged to develop regional or problem-specific training and validation datasets. In Table 1, we provide a summary of the existing datasets on Antarctic calving fronts and coastline positions and the satellite modality they were derived from. Shelf-Bench offers several key advantages over existing datasets listed in Table 1. Shelf-Bench features manually curated and precisely delineated training masks paired with pre-processed satellite imagery stored in two file formats ready to use for training deep neural networks. Moreover, it has a comprehensive spatial coverage spanning the most complex regions of the Antarctic coastline with multi-temporal acquisitions in both summer and winter seasons to cover all possible backscatter values.

Table 1 Overview of open-access datasets mapping Antarctic glacier and ice-shelf fronts and the Antarctic coastline. Datasets are grouped by type and listed in reverse chronological order, with the most recent datasets first

Type	Dataset	Modality	Satellite	Coverage	Time Span	Annotation	Temporal Resolution	Spatial Resolution [m]	Available at
glacier	Loebel et al. 2024	optical	Landsat-8/9	AP, 42 glaciers	2013-2023	automatic	irregular, up to weekly	30	https://doi.org/10.1594/PANGAEA.963725
	Gourmelon et al. 2022	SAR	ERS, Envisat, Radarsat, ALOS, TerraSAR-X, Sentinel-1	AP, 43 glaciers	1996-2021	manually	irregular	depending on sensor	https://doi.org/10.1594/PANGAEA.940950
shelf	Baumhoer et al. 2023	SAR	Sentinel-1	Antarctica, 36 shelves	2015-2025	automatic	monthly	40	https://doi.org/10.15489/btc4qu75gr92.
	Wuite et al. 2019	SAR	CryoSat-2	Ronne-Filchner Ice Shelf	2011-2018	automatic	bi-annual	200	http://cryoportal.enveo.at
coastline	ADD	optical/ SAR	Worldview, Sentinel-1, Landsat	Antarctica	1993-2024	manually	updated regularly	< 30	https://add.scar.org/
	BedMap3	optical	Landsat 8	Antarctica	2022	manually	Jan-Mar	500	https://doi.org/10.5285/2d0e4791-8e20-46a3-80e4-f5f6716025d2
	Greene et al. 2022	optical/ SAR	various datasets and sensors	Antarctica	1997-2021	semi-automatic	annual	240	https://doi.org/10.5281/zenodo.5903643
	Andreasen et al. 2023	optical	MODIS	Antarctica, 34 shelves	2009-2019	manually	annual	250	https://doi.org/10.5281/zenodo.7830051
	MOA	optical	MODIS	Antarctica	2004, 2009, 2014	manually, semi-automatic	annual	125/750	https://doi.org/10.5067/68TBT0CGJSOJ
	MEaSUREs v2	SAR	ALOS, Enivsat	Antarctica	2008/ 2009	manually	two years	~ 150	https://doi.org/10.5067/AXE4121732AD
	Cook et al. 2021	optical/ SAR	maps, aerial & satellite imagery	AP	1843-2008	manually	irregular	30	https://doi.org/10.5285/07727663-9b94-4069-a486-67e4d82177d3
	RAMP	SAR	Radarsat	Antarctica	1997, 2000	semi-automaitc	annual	25	http://research. bpcrc.osu.edu/rsl/ radarsat/data/
	This study	SAR	Sentinel-1, ERS, Envisat	Antarctica,	1992-2021	manually	irregular	30/40	https://doi.org/10.5281/zenodo.17610870

2.2. Algorithms for Antarctic calving front delineation

DL methods are increasingly applied to delineate calving fronts and ice sheet coastlines in satellite imagery, yet the task is complicated by the need to distinguish between several visually similar boundaries in polar regions. Marine-terminating glaciers end in the ocean with narrow, often rapidly changing calving fronts, typically surrounded by ice mélange and smaller icebergs. In contrast, ice shelves are floating extensions of the ice sheet that can span hundreds of kilometres, with calving fronts that are broader, more stable over short timescales, and shaped by often longer calving cycles of tabular icebergs over several decades (Fricker et al., 2002). Delineating the Antarctic coastline is particularly challenging because the coastline is not a simple land-water boundary, but a composite of ice-water and rock-water margins. Large coastline segments consist of calving fronts from marine-terminating glaciers along the Antarctic Peninsula and extensive ice shelves in West and East Antarctica, where ice flows into the ocean as floating ice platforms. In these regions, the coastline coincides with the seaward edge of floating ice, which can shift seasonally or over longer periods following tabular iceberg calving. Other sections trace the grounded ice sheet margin, where ice rests on bedrock, and exposed rock outcrops form stable boundaries within the otherwise dynamic ice-dominated coastline boundary. Recognizing that each boundary type exhibits distinct geometric patterns, backscatter intensity, and temporal variability is critical for developing benchmark datasets and training algorithms.

Recent advances in machine learning have enabled automated mapping of calving fronts, including those of both glaciers and ice shelves. We briefly summarise existing DL-based methods for the automated extraction of calving fronts in polar regions, and organise the studies according to their design and release year (Figure 1). However, the absence of a standardized benchmark dataset for Antarctic ice shelves prevents objective performance comparison across algorithms. Whereas, standardized model intercomparison using the CaFFe benchmark dataset has enabled identification of optimal algorithms for outlet glacier applications (Gourmelon et al., 2025a), determining the most effective approach for ice shelf front delineation remains unclear without a standardized benchmark dataset.

The use of DL for automatically identifying calving fronts from satellite images started in 2019, mainly using variations of the U-Net architecture (Baumhoer et al., 2019; Mohajerani et al., 2019; Zhang et al., 2019). This initial work led to the development of various DL techniques for extracting calving fronts from both optical and SAR satellite data. Most of these techniques rely on Convolutional Neural Networks (CNNs), with the U-Net framework being the most commonly used for segmentation (Baumhoer et al., 2019; Hartmann et al., 2021; Loebel et al., 2022, 2025; Mohajerani et al., 2019; Zhang et al., 2019).

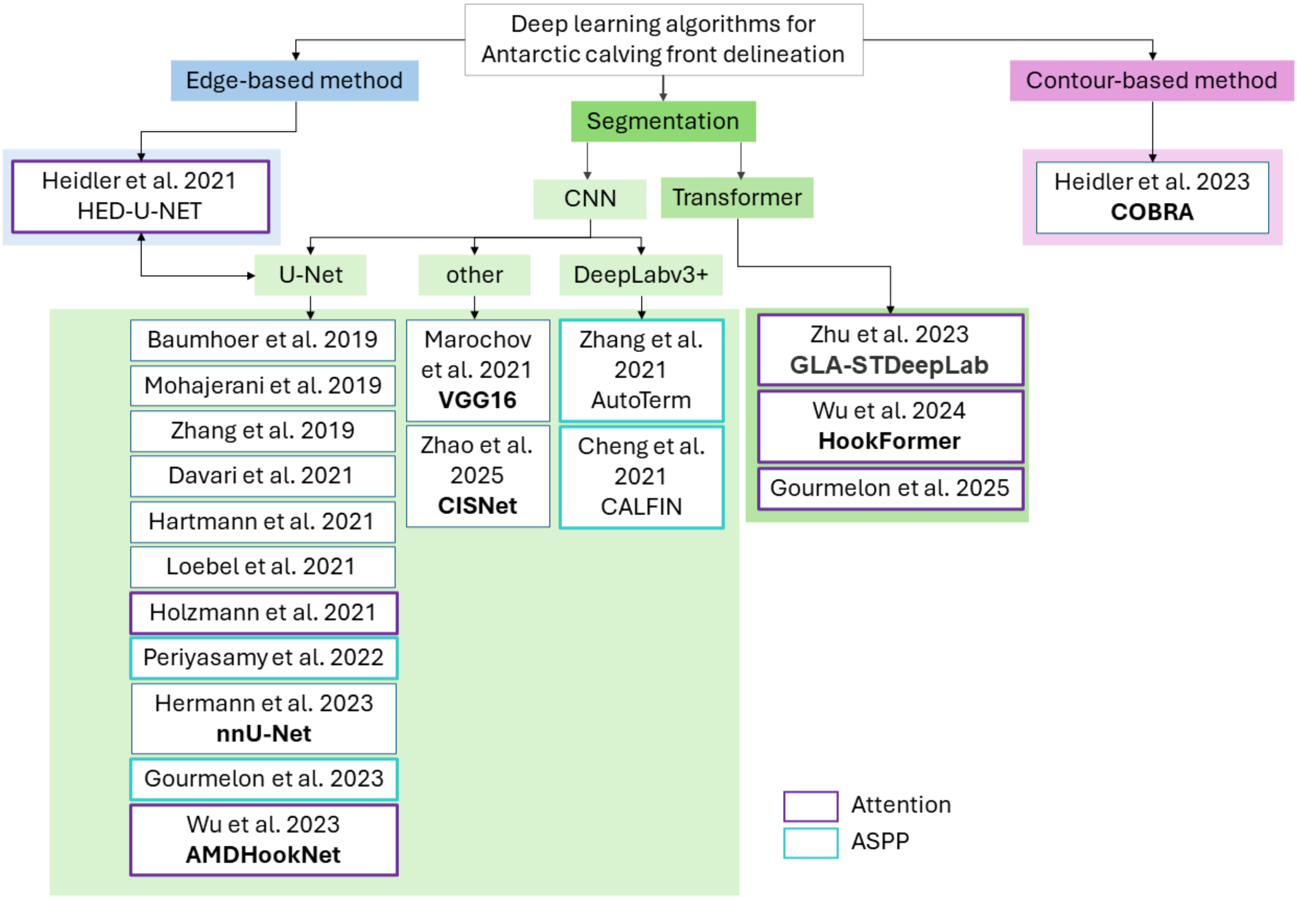


Figure 1 Tree diagram of studies applying deep learning (DL) methods for calving front delineation divided into edge detection (blue), segmentation (green) and contour-based methods (pink). The model's name is shown in bold if a model architecture has a specific name.

Additionally, Marochov et al. (2021) investigated a VGG16 based architecture for classifying satellite images of glaciers, with calving front extraction as a secondary focus. The CISNet is another recent CNN design that employs a U-ConvNextV2 to explore semantic relationships in glacier images by linking semantic segmentation with change detection tasks (Zhao et al., 2025). Subsequent research continued to use U-Net architectures, but with modifications that incorporate architectural advancements like Atrous Spatial Pyramid Pooling (ASPP) (Gourmelon et al., 2023; Periyasamy et al., 2022) and attention mechanisms (Heidler et al., 2021; Holzmann et al., 2021; Wu et al., 2023). By integrating ASPP modules into the U-Net's bottleneck stage, multi-scale features can be extracted. Attention-enhanced U-Nets add spatial, channel-wise, or self-attention blocks into skip connections or decoder layers, enabling the network to focus on important areas while minimizing distractions from irrelevant background features. Another enhancement to the original U-Net architecture is the concept of deep supervision, which involves adding auxiliary loss functions to intermediate layers rather than depending solely on the final output layer for training supervision (Heidler et al., 2021; Herrmann et al., 2023; Wu et al., 2023).

Alongside these advancements, DeepLabv3+ has emerged as an alternative to traditional U-Net designs (Cheng et al., 2021; Zhang et al., 2021). The network combines ASPP for comprehensive multi-scale context capture with an encoder-decoder refinement module to enhance object boundary recovery. A recent study has also begun exploring contour-based methods for calving front delineation with the development of the COBRA architecture (Heidler et al., 2023). The aim is to trace the boundary between the ocean and the ice sheet, rather than segmenting these features within the image. The COBRA architecture outperformed previous CNN-based architectures for glacier front detection and achieved competitive accuracy for ice shelf front delineation.

Finally, while transformer architectures are quite new in this field (Gourmelon et al., 2025b; Wu et al., 2024), they have demonstrated superior performance in large-scale comparisons on the CaFFe benchmark dataset for glacier calving front delineation (Gourmelon et al., 2022). Transformer models have achieved an average accuracy of 221 meters for post-processed results on glacier calving fronts (Wu et al., 2024). The latest model, SSL4SAR, uses a transformer architecture with self-supervised learning, achieving a mean distance error (MDE) of 75 meters on the CaFFe dataset, getting closer to the human annotation MDE of 38 meters (Gourmelon et al., 2025b). In summary, the development of deep learning techniques for calving front delineation indicates a transition from conventional CNN frameworks to novel transformer architectures, resulting in a significant improvement in accuracy that more closely matches human annotations.

3. The Shelf-Bench dataset

3.1. Temporal and Geographical Coverage

The Shelf-Bench dataset encompasses various segments of the Antarctic coastline and is made up of 161 satellite images based on open access C-band SAR data obtained through the satellite missions Sentinel-1, Envisat, and ERS-1/2 with main specifications noted in Table 2. The test set was defined after expert evaluation of the images. It includes a representative selection of typical ice-shelf boundaries with varying levels of complexity to ensure comprehensive coverage of the problem space acquired by different sensors. The training and validation set is based on 81 scenes of varying spatial resolution and size from Sentinel-1, 10 scenes from Envisat, and additional 32 scenes from ERS-1/2. The test set is based on 38 scene subsets especially focusing on the front, consisting of 27 from Sentinel-1, 3 from Envisat, and 8 from ERS. The geographical distribution of both the training and test datasets is illustrated in Figure 2.

Table 2 Summary of the SAR sensor specifications of the imagery used in the benchmark dataset. Only multi looked imagery was used in the IMP (ERS/Envisat) and GRD (Sentinel-1) formats.

Platform	Sensor	Launch	Mission end	Mode	SAR freq. band	Pol.	Repeat cycle [d]	Pixel size (multi-looked)	Pixel size (pre-processed)	Swath width [km]
ERS-1/2	AMI	1991/ 1995	2000/ 2011	IMP	C-band	VV	35/1	12.5 m	30.0 m	100
Envisat	ASAR	2002	2012	IMP	C-band	VV	35	30.0 m	30.0 m	100
Sentinel-1 A/B	SAR	2014	ongoing	EW	C-band	HH	6/12	40.0 m	40.0 m	410

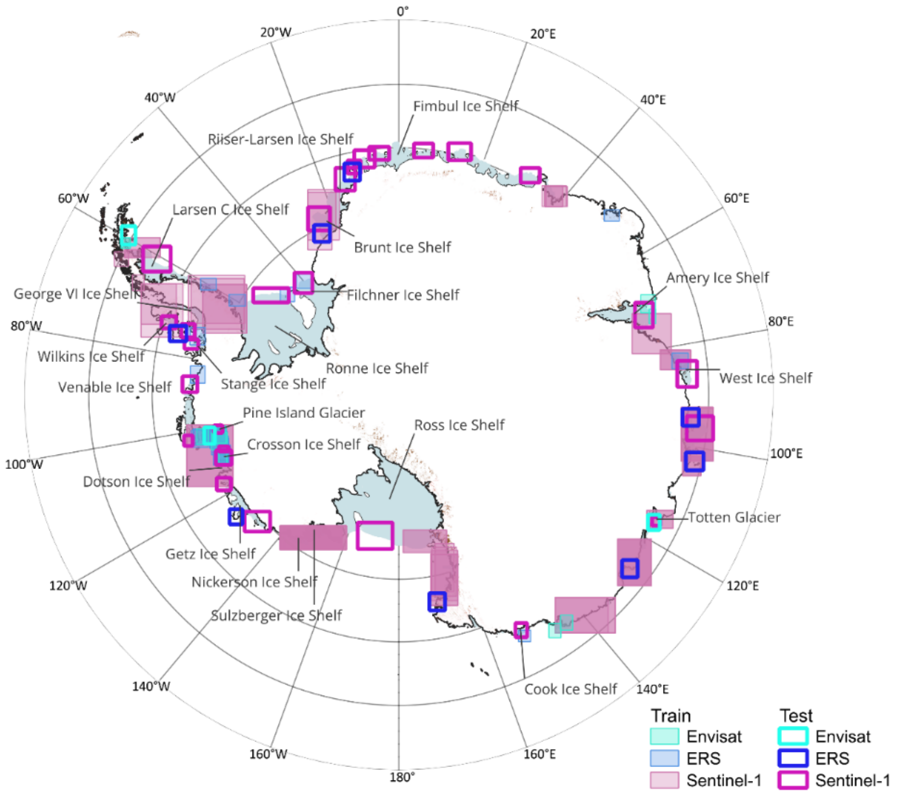


Figure 2 Geographical distribution of the training and test datasets. Filled scene extents mark training imagery, and boundary extents mark test imagery for ERS-1/2 (blue), Envisat (turquoise) and Sentinel-1 (pink). Larger extents of Sentinel-1 reflect the larger swath width compared to Envisat and ERS-1/2 data.

The uneven geographical distribution of the dataset stems from the available satellite data, the use of legacy datasets from earlier studies (Baumhoer et al., 2019, 2023; Heidler et al., 2021; Wagner, 2023), and a strategic focus on morphologically varied ice shelves and coastal areas. For instance, increased data availability at the highly dynamic Pine Island Bay coastal region is due to ERS and Envisat data created in a prior study (Wagner, 2023). Moreover, the sampling strategy emphasizes data quality, quantity, and relevance over data uniformity in coverage, making the dataset particularly suitable for evaluating models across diverse coastal environments. The temporal composition of the benchmark dataset (Figure 3) is determined by the availability of satellite imagery from each satellite and spans the period from 1991 to 2022. Training data for Sentinel-1 is accessible for the years 2017 to 2019, while test data is available for the years 2014 and 2016, as well as from 2020 to 2022. Envisat training data encompasses the years 2004 to 2011, with test data collected in 2008 and 2009 during a period without training data. For ERS-1/2, training data covers a longer period from 1991 to 2011, and the test data is from the year 1996. This year is significant as it is the sole year offering pan-Antarctic coverage during the ERS-1/2 era, facilitating a geographically diverse test set for ERS. None of the three satellite sensors provide data between 2012 and early 2014, resulting in a short time series gap. The entire benchmark dataset guarantees that there exists either a temporal or geographical distinction between the test and training datasets to prevent overlaps. A detailed list of temporal and geographical coverage by each satellite mission can be found in Table A 1in Appendix A.

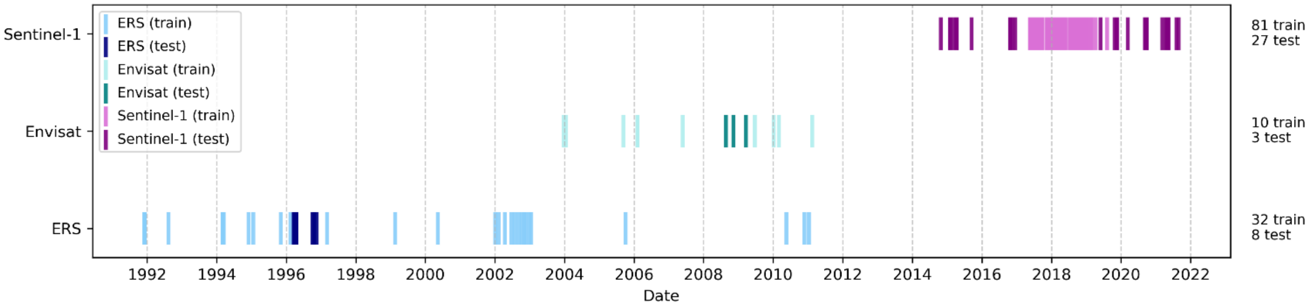


Figure 3 Temporal distribution of the training and test datasets for Sentinel-1, Envisat and ERS-1/2.

3.2. Dataset generation

3.2.1. Pre-processing ERS-1/2 & Envisat data

ERS-1, launched by the European Space Agency (ESA) in July 1991, was the first European SAR mission, followed by ERS-2 in April 1995. Together, they provided continuous SAR data from 1991 to 2011, creating one of the longest time series of its kind. Both satellites were equipped with an Active Microwave Instrument (AMI) operating at C-band frequency and Vertical-Vertical (VV) polarization, capturing scenes sized 100 by 102.5 km with a spatial resolution of 12.5 m, and a revisiting time of 35 days. The dataset includes ERS-1/2 SAR IM Precision L1 products (SAR_IMP_1P), which are multi-look, ground range amplitude products corrected for radar gain and range spreading loss (Jensen, 1995; van 't Klooster, 2011).

Envisat, launched in March 2002 as the successor to the ERS mission, operated until its failure in April 2012. It featured the Advanced Synthetic Aperture Radar (ASAR) sensor, which captured data in both single-polarization and co- and cross-polarization modes. This study utilizes the ASAR IM Precision L1 product (ASA_IMP_1P), a continuation of the ERS IM Precision L1 product with VV polarization. The data offers a resolution of 30 meters, covering scenes of 100 by 55-100 kilometres, with a revisit interval of 35 days (ERS-products-specification-with-Envisat-format, 2025; Rignot and van Zyl, 1993). Using the Graph Processing tool of the ESA SNAP Toolbox, the initial step involved applying the Precise Orbit files from the Delft Institute for Earth-Oriented Space Research (DEOS) for each ERS and Envisat scene (DEOS, 2016). These orbit files provide accurate satellite position and velocity information and can significantly enhance the SAR images' geolocation accuracy. Both the ERS and Envisat images were adjusted from digital numbers (DN) to the backscattering coefficient σ0, using techniques from Laur et al. (2002) for ERS and Rosich (2004) for Envisat. This adjustment accounts for elements like antenna patterns, incidence angles, and absolute calibration constants. Next, to lessen the variation in pixel intensity caused by speckle, a low pass filter was used to reduce speckle while keeping structure of the features of interest intact. The Refined Lee filter was chosen for its simplicity and better edge preservation compared to other filters. Due to the side-looking nature of SAR, images can become geometrically distorted, particularly over uneven landscapes such as the Antarctic coast. To correct for this, geometric terrain correction was used, which mimics radiometric effects caused by terrain with the help of the PolarDEM (90 m resolution) (Wessel et al., 2021). To achieve consistency across sensors, all images were resampled to 30 m resolution using bilinear interpolation and were reprojected to the Antarctic Polar Stereographic projection (EPSG:3031). To focus on the ice front areas, the images were clipped using the RADARSAT-1 Antarctic coastline product from Liu and Jezek (2004a), with an added 50 km buffer to both sides.

3.2.2. Pre-processing Sentinel-1 data

The Sentinel-1 mission, part of the European Union's Copernicus Programme, consists of a constellation of SAR satellites designed for all-weather, day-and-night Earth observation. For the benchmark dataset, we selected data with the imaging mode Extra Wide (EW) swath with a spatial resolution of 40 m, which is specifically tailored for monitoring large ocean and sea ice areas such as the Antarctic coastline (Nagler et al., 2015; Sun and Li, 2021). In EW mode, Sentinel-1 provides SAR data with a swath width of around 400 km and operates in single and dual polarization. To keep the benchmark dataset uniform, we selected only the HH polarization to maintain a fixed number of channels in line with the ERS and Envisat data. We selected the HH polarization because it is more sensitive to the double bounce effect at edges such as the ice shelf front. The pre-processing is performed with the Graph Processing Tool of the ESA SNAP Toolbox 8.0 on a Hadoop Cluster with 63 nodes (32/64 GB RAM). For single-polarized scenes, the pre-processing includes applying the orbit file, thermal noise removal, radiometric calibration and geometric terrain correction with the TanDEM-X PolarDEM (Wessel et al., 2021). To create a dataset focused on the coastline, we clipped the pre-processed scenes with a more recent 50 km buffered coastline based on the MODIS coastline from the year 2014 (Scambos et al., 2007). The coastline buffer significantly influences the Sentinel-1 data because its swath width is 400 km, compared to 100 km for ERS and Envisat, with most image data falling within the buffer.

3.2.3. Data Annotation and Label Production

Labels were produced through a multi-expert annotation protocol. Initially, one expert labeled a subset of the available Sentinel-1 scenes. Entire scenes were annotated in QGIS by delineating the coastline as the border between ocean and ice.

These annotations were then refined using high-resolution optical imagery (e.g., WorldView), particularly over small outlet glaciers and mélange-dominated areas. Elevation data from the TanDEM-X PolarDEM 90 (Wessel et al., 2021) were additionally employed to validate and adjust annotations in regions of fast ice, where elevation provides a more reliable indicator than backscatter alone.

The refined annotations were compared against established coastline products (Liu and Jezek, 2004a; Scambos et al., 2007), which exhibit notable methodological differences in coastline definition (Baumhoer et al., 2021). The expert determined the final front positions by balancing correctness to the Sentinel-1 imagery with consistency relative to these existing products and labeled all remaining Sentinel-1 scenes based on this protocol. The annotated Sentinel-1 scenes were subsequently provided to two additional experts, who labeled the corresponding ERS and Envisat data. Their annotations were then reviewed and corrected by the initial expert to ensure consistency and quality. Finally, the line-based front positions were converted to polygons and rasterized to produce the final binary label set.

4. The Baseline

4.1. Baseline Models

We employed five baseline models to establish initial accuracy metrics for our benchmark dataset. These baselines provide a foundational performance reference, allowing future studies to compare and assess new methods against consistent, well-defined standards. By reporting the results of multiple baselines, we ensure a comprehensive evaluation of the dataset's difficulty and characteristics, thereby supporting reproducibility and facilitating fair benchmarking within the research community. Shelf-Bench is trained on the following five deep learning models to evaluate image segmentation performance and provide a baseline for future model developments:

- U-Net: (Ronneberger et al., 2015)
- Feature Pyramid Networks (FPN) (Lin et al., 2017)
- DeepLabV3 (Chen et al., 2017)
- Vision Transformer (Dosovitskiy et al., 2020)
- DINOv3 7B satellite model (Siméoni et al., 2025)

The selection of the five baseline models was motivated by their different architectures and strong performance across various image segmentation and representation learning tasks. U-Net was chosen for its effectiveness in pixel-level segmentation, particularly in remote sensing and for calving front segmentation (see Figure 1) (Ronneberger et al., 2015), and is widely used as reference architecture, in Earth observation (Hoeser et al., 2020). Feature Pyramid Networks (FPN) were included for their ability to capture multi-scale contextual information, which is essential for handling objects of varying sizes in satellite imagery along the Antarctic coastline (e.g. varying iceberg sizes) (Lin et al., 2017). The DeepLabV3 architecture is based on Deep Neural Networks (DNN) suitable for semantic segmentation, and utilises Atrous Dilated Convolutions which enlarge the convolution, enhancing feature detection without increasing computation or the number of weights (Chen et al., 2017). The

Vision Transformer (ViT) represents a shift toward transformer-based architectures that model global dependencies effectively, offering an alternative to established convolutional methods. Transformers use self-attention to model relationships between input tokens. In ViT based models, images are typically split into patches, embedded as tokens, and treated as a sequence so the Transformer can process them (Dosovitskiy et al., 2020). Finally, the state-of-the-art vision foundation model DINOv3 7B was incorporated as a self-supervised vision transformer, pre-trained specifically on satellite data, where the DINOv3-sat fine-tuned variant is a relevant and high-capacity reference for modern geospatial analysis. The application of foundation models in cryosphere research is still in its initial stages, with only a limited number of studies available to date (Jiang et al., 2025; Kaushik et al., 2025; Shankar et al., 2023). Therefore, to the authors' knowledge, this study is the first glaciology study to utilise DINOv3 7B for image segmentation in this specific domain. We use the DINOv3 7B ViT model pretrained on the optical dataset SAT-493M with a U-Net head, and input each Shelf-Bench SAR image into the RGB channels three times to overcome mismatches in optical and SAR channel sizes. Table 3 summarises the configurations of all five baseline models, including information on the pre-trained weights used. ImageNet weights come from a dataset totalling 3.2 million images (Deng et al., 2009) and SAT-493M weights for DINOv3 refer to an optical satellite dataset of 493 million 512 x 512 RGB images (Siméoni et al., 2025).

Table 3 Model architectures trained on Shelf-Bench with model specifics detailed. Training time was estimated based on training for 1000 epochs with batch size 32, Adam optimizer, and a learning rate of 0.0001. Trainable parameters given in millions (M). Inference time for GPU, given in milliseconds (ms).

Model	Model type	Encoder	Model Head	Weights	Trainable Parameters	Inference Time per patch (ms)	Reference
U-Net	Convolutional encoder–decoder with skip connections	ResNet50	n/a	ImageNet	23 M	0.80	Ronneberger et al., 2015
FPN	CNN-based multi-scale feature extractor / segmentation	ResNet50	n/a	ImageNet	23 M	0.89	Lin et al., 2017
DeepLabV3	CNN segmentation model using atrous/dilated convolutions	ResNet50	n/a	ImageNet	23 M	1.51	Chen et al., 2017
ViT	Vision Transformer with multi-head self-attention	ViT-L_16	U-Net	ImageNet	303 M	7.29	Dosovitskiy et al., 2020
DINOv3	Self-supervised Vision Transformer (ViT backbone)	dinov3_vitl16	U-Net	SAT-493M	300 M	11.1	Siméoni et al., 2025

4.2. Baseline Model Training

To establish a strong performance benchmark, the five baseline models were trained and evaluated using a consistent experimental setup. The dataset is split into training and testing subsets with an approximate 75/25 scene-based ratio and scenes tiled into 256 x 256 pixel. This size preserves multiscale features necessary for ice shelf front delineation and common DL architectures across different compute platforms also enabling research teams with limited GPU access. The patches are extracted using a sliding-window, generating 42,424 patches for training, 6,974 for validation and 6,545 for test. An overlap can be set to increase patches extracted, however an overlap of zero was chosen for the training and test data. Patches are

https://doi.org/10.5194/essd-2025-758
Preprint. Discussion started: 2 March 2026

normalized locally using a percentile clip method on individual patches improving contrast in glacial environments (Loebel et al., 2024). Each patch is matched with its corresponding label. For simplicity, class labels are named `Ice’ and `Ocean’. The `Ice’ class comprises all land ice from the ice sheet and all attached floating ice beyond the grounding line (ice shelf). The `Ocean’ class refers to sea ice, including fast ice, icebergs and the ocean itself, as well as no data areas. We discard patches which contain >80% no data. The two `Ice’ and `Ocean’ classes are roughly balanced. The validation set enabled monitoring of the model performance during training, before final evaluation on the test set.

Training was carried out using JASMIN, operated by the Natural Environment Research Council (NERC), on the Orchid GPU cluster with 5 Nvidia A100 GPU 40GB SXM4 cards. All five models ran for 150 epochs, with a learning rate of 0.0001 with the Adam optimizer (Kingma and Ba, 2014) and early stopping with a patience of 60 epochs imposed on the validation loss. We selected 150 epochs to ensure stable fine-tuning of model heads. The largest batch size that was computationally possible was batch size 32 per GPU, as there is a trade-off between image size and batch size. The inference time per patch varied between 0.80 - 11.1 ms between models run on GPU, which was ~60x faster than running on CPU for all models. During training only, data augmentation was applied to increase the robustness of the model. Augmentations included random rotations, horizontal flips, random resized crops, random brightness contrasts and gaussian noise. The Adam optimiser was applied for training due to its relative robustness compared with alternative optimisers. The loss function used for all five models was a specialised weighted combination of Dice Loss (Sudre et al., 2017) and Focal Loss (Lin et al., 2018), both commonly used in computer vision tasks and both contribute to tackling different aspects of the difficulties of glacier delineation. This combined loss function approach was inspired by Gourmelon et al. (2022), where the authors found this combination of loss functions improved performance. The Dice loss component measures the overlaps between true and predicted class boundaries, which is useful for learning the often complex and irregular shapes that calving fronts have. Alternatively, the Focal Loss component addresses class imbalance.

4.3. Baseline Results

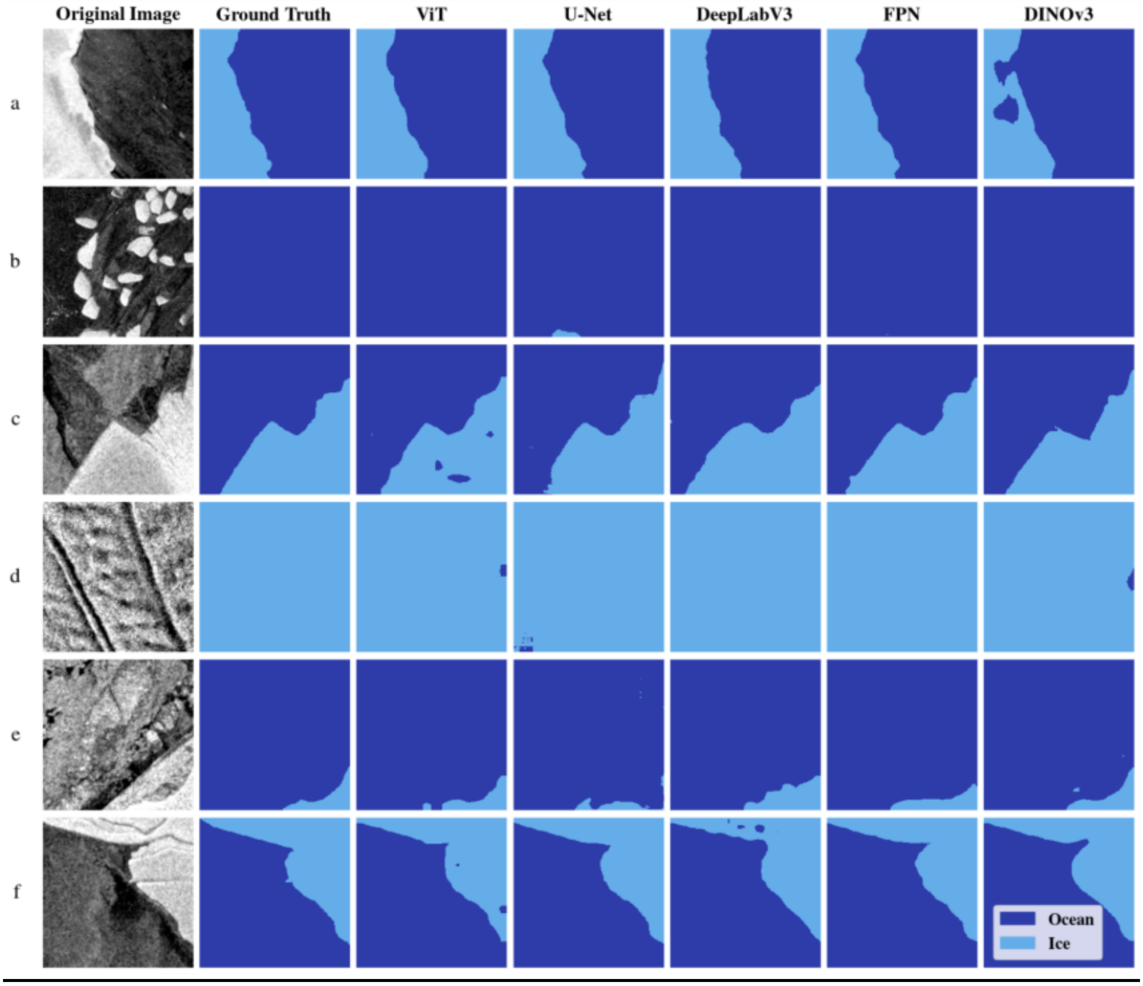
4.3.1. Model Evaluation: Segmentation

In this section, we assess the five different segmentation models on the test set of Shelf-Bench. Model performance was continuously evaluated on the training and validation datasets using the evaluation metrics given in Table 4, which provide information on how well the segmentation model performs compared to the human-annotated label. The epoch run with best performance was determined using the highest mean Intersection over Union (IoU) score on the validation dataset. The test dataset underwent the same pre-processing (6,545 patches) and normalisation as the training data, and is independent and geographically and temporally distinct from the training and validation datasets (see Figure 2, Figure 3), allowing a fair evaluation of the models’ performances.

Table 4 Evaluation metrics definitions commonly used in segmentation studies. IoU - Intersection over Union. All values range between 0 and 1.

Metric	Formula	Definition
Precision	TP / (TP + FP)	Proportion of correctly predicted positives out of all positive predictions; reflects how well the model avoids false positives.
Recall	TP / (TP + FN)	Proportion of actual positives correctly identified; measures ability to find all relevant instances.
IoU	TP / (TP + FP + FN)	Intersection over Union; quantifies overlap between predicted and ground-truth regions (localisation accuracy).
F1-score	2 × (Precision × Recall) / (Precision + Recall)	Harmonic mean of precision and recall; balances false positives and false negatives in one metric.

335



340 **Figure 4 Model comparison of segmentation performance (from left to right) on the original SAR image, ground truth label, ViT, U-Net, DeepLabV3, FPN and DINOv3. The two classes 'Ocean' and 'Ice' are dark blue and light blue respectively. A range of glacial scenes are shown to demonstrate varying model performance. The original satellite scenes were acquired by (a) Envisat, 2008-08-24, (b) Envisat, 2009-03-23, (c) ERS, 1996-10-05, (d) Sentinel-1A, 2015-04-08, (e) Sentinel-1B, 2016-10-25, and (f) Sentinel-1B, 2020-09-21.**

https://doi.org/10.5194/essd-2025-758
Preprint. Discussion started: 2 March 2026

Firstly, results are interpreted visually by comparing the outputs from the different baselines with examples provided in Figure 4. All five models display good capability to distinguish the two classes ‘Ocean’ and ‘Ice’. In particular, the interface between ice and ocean is delineated accurately by all models as long as there is a clear ice-ocean boundary (Figure 4a). Only DINOv3 introduces holes in the ice prediction. The U-Net delineated the front closest to the ground truth whereas especially ViT and DeepLabV3 detect the boundary but generalize the coastline and miss smaller details. All models were able to detect icebergs and classify them correctly as ocean even though having the same backscatter signature as the ‘Ice’ class (Figure 4b). The presence of sea ice and fast ice challenges all models leading to slight discrepancies compared to the ground truth (Figure 4c,e,f). Ice surface features such as crevasses, fractures, and shadows are also successfully labelled as ‘Ice’ by all models (Figure 4d).

The quantitative evaluation of the five segmentation models on ocean and ice classification tasks revealed more distinct performance characteristics across different evaluation metrics (Table 5). Overall pixel accuracy exceeds 0.906 across all models suggesting that the Shelf-Bench dataset supports stable convergence and reproducible segmentation performance. Class-wise Intersection over Union (IoU) reveals an intrinsic asymmetry in dataset difficulty between ‘Ocean’ and ‘Ice’ regions. For every model, IoU for the ocean class is systematically higher than for the ice class. Ocean IoU ranges from 0.839 to 0.889, whereas ice IoU ranges from 0.815 to 0.871. This consistent gap suggests that the dataset contains greater intra-class variability and boundary ambiguity within ice regions. The persistence of this pattern across architectures indicates that it is a dataset-driven property rather than an artifact of a particular model design. Precision and recall metrics further characterize annotation and class distribution properties. Ocean recall is uniformly high (0.899–0.944), demonstrating that ocean pixels are rarely missed in the classification. In contrast, ice recall exhibits a wider spread (0.882–0.93), reflecting that only the best performing model on ice recall (ViT) finds most of the relevant ice instances. Precision is higher for ocean than ice except for DINOv3 performing higher on ice precision with 0.930 compared to 0.918 for ocean precision. This means, that most models are better in avoiding false positive ocean classifications than ice. The lower precision for ice indicates that the Shelf-Bench dataset includes a sufficient number of difficult ‘Ice’ samples to challenge model performance.

F1 scores show close agreement between classes, with an F1-score for ‘Ocean’ between 0.912 and 0.941 and ‘Ice’ between 0.898 and 0.931. The small but persistent ocean–ice performance gap aligns with the IoU analysis and reinforces the interpretation that ‘Ice’ segmentation represents the primary source of dataset difficulty.

Table 5 Model performance evaluated on the Shelf-Bench test set. Best results are in bold and underlined values refer to the second-best result. Upwards arrows indicate higher model performance with increasing value.

	Pixel	IoU	IoU	Precision		Recall		F1	
Model	Accuracy ↑	Ocean ↑	Ice ↑	Ocean ↑	Ice ↑	Ocean ↑	Ice ↑	Ocean ↑	Ice ↑
DeepLabV3	0.926	0.874	0.848	0.926	0.925	0.939	0.91	0.933	0.918
U-Net	0.906	0.839	0.815	0.926	0.882	0.899	0.914	0.912	0.898
DINOv3	0.923	0.87	0.842	0.918	**0.930**	**0.944**	0.898	0.931	0.914
FPN	0.921	0.864	0.842	0.936	0.903	0.918	0.925	0.927	0.914
ViT	**0.937**	**0.889**	**0.871**	**0.95**	0.921	0.933	**0.941**	**0.941**	**0.931**

4.3.2. Model Evaluation: Mean Distance Error

Overall segmentation accuracies common in computer vision can easily overlook important details at the boundary between the segmentation classes ‘Ice’ and ‘Ocean’. However, a clear boundary between the two classes is essential for delineating clear ice shelf front boundaries and create an accurate representation of the Antarctic coastline. Therefore, providing an additional accuracy metric to account for differences in detected front position is crucial. We provide accuracies of front prediction as the Mean Distance Error (MDE) between the predicted and ground truth front (Table 6). MDE is commonly used in calving front delineation studies (e.g. Gourmelon et al. 2022, Loebel et al. 2025) and is defined as the average of the mean closest-point distances between the ground truth and predicted ice fronts, for all images. To extract the front position between the class ‘Ice’ and ‘Ocean’, we binarized model predictions (threshold 0.5) and converted the raster into a line shapefile. Satellite scene boundary artifacts were removed by binary erosion based on the satellite scene data extent. Discontinuities and features which are not part of the front (e.g. icebergs) where removed by morphological filtering to create only two main connected segments for each class. The distance between the front lines extracted from the ground truth and the line generated from the model prediction is then calculated by converting the pixel distances to metres. The pixel size is 30 m for ERS and Envisat and 40 m for Sentinel-1. It is important to mention that the MDE calculation was done on each patch and not on merged scenes. This means strong outliers can occur if only small parts of the front or rifts are present in the patch margins due to reduced context.

Visual examples of front predictions are illustrated in Figure 5 for each baseline and different coastline areas. Figure 5a shows an example of a grounded coastline boundary. Most models achieve MDE accuracies of 1-2 pixels (30 m ERS/Envisat, 40 m Sentinel-1) except FPN and ViT which match the ground truth less accurately with a MDE of 81.9 m and 135.8 m, respectively. The ViT delineates the coastline to be further inland than the ground truth and other models, highlighted by the higher MDE (Figure 5a). A clear boundary is also given in Figure 5d where the U-Net achieves best results with MDE 29.7 m. Models like the ViT or DeepLabV3 mistake offshore icebergs as the ice-ocean boundary resulting in an MDE of 121.2 m and 121.8 m, respectively. Varying sea ice cover as shown in Figure 5c and Figure 5e resulted in high MDE for the ViT and partly for DinoV3 and UNet. DeepLabV3 and FPN produced more stable front delineations with 1-2 pixels in difference to the ground truth. The example of a heavily crevassed ice shelf front in Figure 5b poses challenges for every baseline model, resulting in strongly varying front positions. The ViT achieved the best performance on this difficult example, with an MDE of 273.2 m, but the ViT positioned the front oceanward whereas other models created delineations that would otherwise match differences in human annotations.

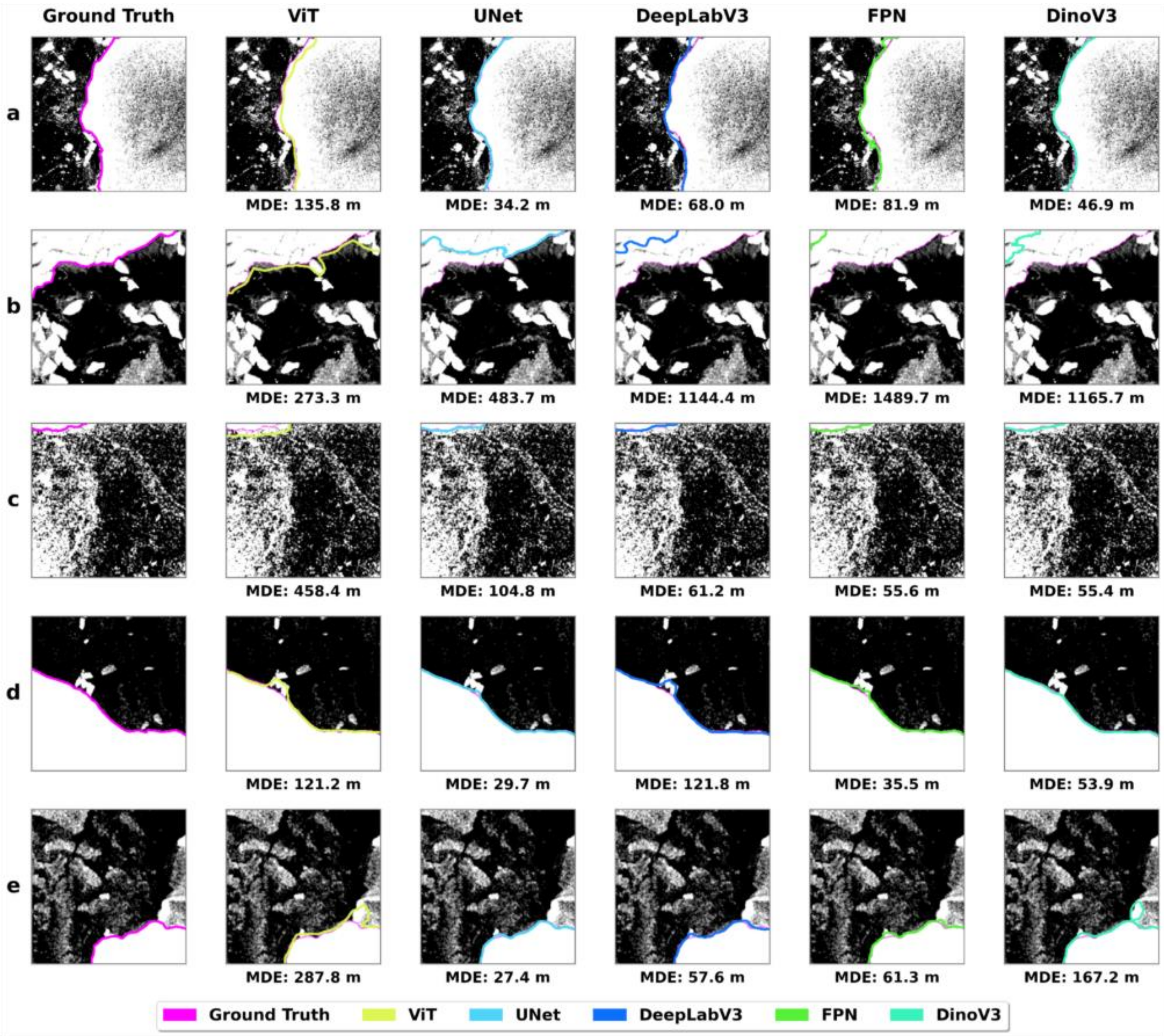


Figure 5 Visual examples of differences in front delineation based on the five baselines. Ground truth is highlighted in pink. The satellite scenes shown refer to (a) ERS, 1996-03-22, (b) Envisat, 2009-03-23, (c) Sentinel-1A, 2015-04-08, (d) Sentinel-1A, 2015-09-14, and (e) Sentinel-1A, 2021-04-26.

In addition to the visual inspections, the MDE analysis for the entire test set revealed substantial variations in segmentation accuracy across models and satellite sensors (Table 6). Lower MDE values indicate better alignment between predicted and ground truth ice front boundaries. We report both mean and median MDE to enable comparison with prior literature while providing a more robust performance indicator. Because the patch-based evaluation produces strong outliers that skew the mean, the median MDE offers a more reliable summary of typical boundary error.

DeepLabV3 achieved the lowest mean MDE across all satellites (514.3 m) followed by the U-Net (559.7 m Despite achieving the highest pixel-based accuracy, ViT yields the largest boundary errors, with a mean MDE of 692.5 m and a median MDE of

127.7 m, indicating strong within-class classification but poor front localization, as also illustrated in Figure 5. This contrast shows that Shelf-Bench is challenging both in semantic discrimination and in precise boundary delineation.

Median MDE values are substantially smaller (49.4–127.7 m) because they are less affected by extreme outliers. U-Net achieves the lowest median MDE across sensors (49.4 m), followed by DINOv3 (63.5 m) and DeepLabV3 (83.6 m), indicating generally accurate front localization despite occasional large errors, whereas ViT (127.7 m) and FPN (109.5 m) show poorer median boundary precision and primarily capture the overall ‘Ocean’ and ‘Ice’ class structure. The overall mean MDE across models is 604.3 m, while the median of 86.7 m corresponds to an average front offset underscoring that boundary prediction remains challenging even for the strongest methods. Full MDE statistics with confidence intervals are reported in Appendix Table A 2.

Table 6 The mean and median MDE metric for each segmentation model calculated on the test dataset and MDE metrics by satellite type and model. All values given in metres. The downwards arrow indicates smaller results are desirable.

	mean				median			
Model	**Envisat MDE (m) ↓**	**ERS MDE (m) ↓**	**Sentinel-1 MDE (m) ↓**	**Average (m)**	**Envisat MDE (m) ↓**	**ERS MDE (m) ↓**	**Sentinel-1 MDE (m) ↓**	**Average (m)**
DeepLabV3	**365.20**	627.91	**488.96**	**514.29**	91.87	101.49	76.77	83.57
U-Net	520.50	548.06	569.26	559.65	**63.14**	**46.22**	**49.41**	**49.40**
DINOv3	595.39	814.84	592.06	648.28	99.50	93.79	53.44	63.46
FPN	510.20	**442.26**	687.02	606.64	124.53	115.49	101.39	109.51
ViT	591.97	734.08	689.87	692.51	209.84	136.76	118.96	127.71
Average (m)	516.65	633.43	605.43	604.27	117.27	98.75	80.00	86.73

4.3.3. Hard Example Evaluation

We conducted a complexity analysis of the Shelf-Bench dataset to characterize its diversity and identify inherent challenges for ice shelf front delineation. The 16 most challenging patches in the dataset were identified based on consistently high delineation difficulty across all baseline models, with detailed error metrics provided in Table A 3. Visual inspection (Figure 6) reveals that these challenging patches share common characteristics: complex surface textures including fast ice with similar backscatter properties to ice shelves, crevassed surfaces and rifts, dense concentrations of icebergs in proximity to the ice shelf front, and critically, narrow delineation zones where only a small portion of the calving front is visible. These morphological features inherently complicate coastline identification, even for human observers, as they create visual ambiguity between iceberg edges and ice shelf boundaries without sufficient spatial context.

https://doi.org/10.5194/essd-2025-758
Preprint. Discussion started: 2 March 2026

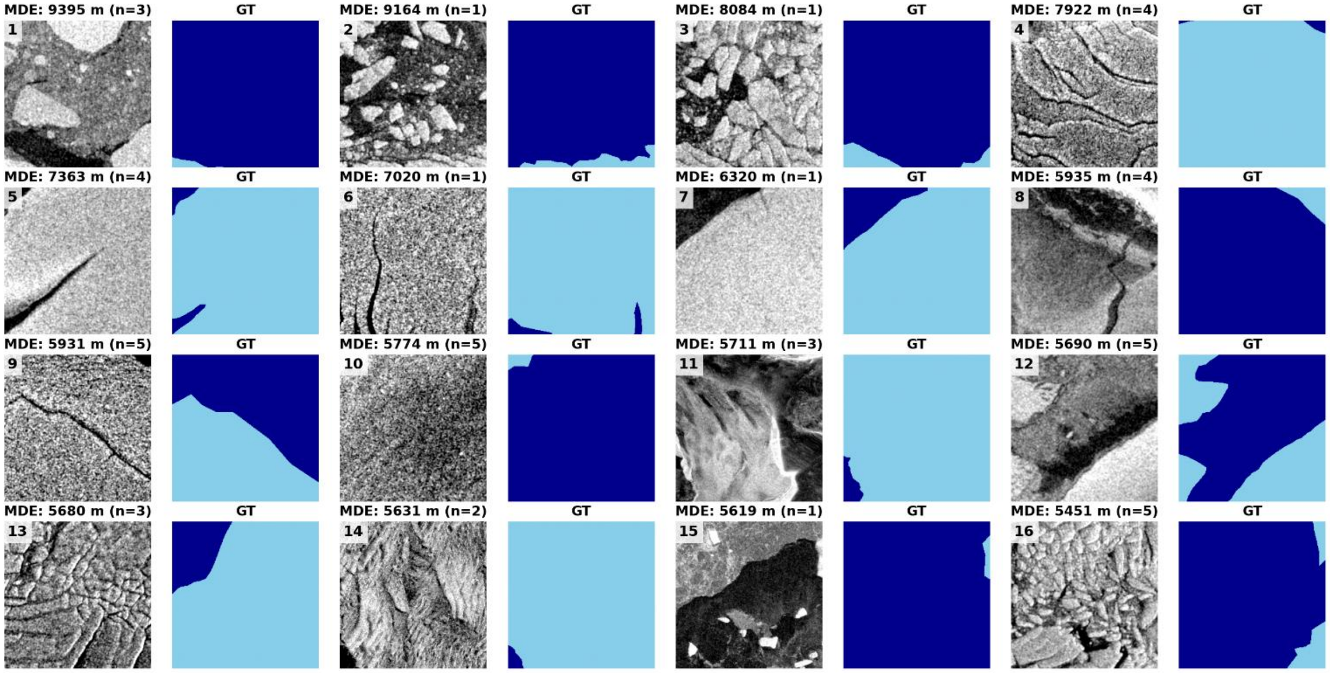


Figure 6 Visual examples of the 16 patches (256 x 256 pixel) ranked as the most difficult across all five baseline models. The MDE is given in metres as the median value across all five baselines. The ground truth is shown next to each patch, with ice marked in light blue and the ocean in navy. n indicates the number of models that detected an ice-ocean boundary in each patch.

In the second part of the analysis, we computed the mean MDE of all patches within each test scene to identify the four hardest and four easiest scenes across all baseline models (Figure 7). The three most difficult scenes (Figure 7, top row) share a common characteristic: the ice-shelf surface is heavily fractured, and numerous icebergs are present in front of the shelf, increasing visual complexity. The hard example over the Larsen-C Ice Shelf, although not exhibiting surface damage, also contains several icebergs near the shelf front. Additionally, the buffer used to clip the scene around the current coastline lies close to the image border and follows the calving of iceberg A-68, which may have further influenced model performance. The mean MDE over the hardest four satellite scenes across all models is 1437.3 m.

In contrast, the easiest test scenes (Figure 7, bottom row) share several visual characteristics that contribute to the strong performance of all baseline models. In these scenes, the ice-shelf front is sharply defined and exhibits a high contrast against the adjacent ocean, allowing for clear delineation of the calving front. The ocean surface is largely free of icebergs or sea ice fragments, reducing visual ambiguity near the boundary. Furthermore, the ice-shelf surfaces appear smooth and continuous, with minimal fracturing or crevassing. Although minor flowlines and localized signs of surface deformation are present, they occur at small scales and do not appear to interfere with model predictions. Overall, these conditions provide a clean and well-structured visual context, enabling the models to accurately distinguish between ice and open-water regions. The median delineation difficulty across the four simplest scenes is 155.1 m, representing approximately a 9-fold difference from the most challenging scenes and demonstrating the substantial range of complexity captured within Shelf-Bench.

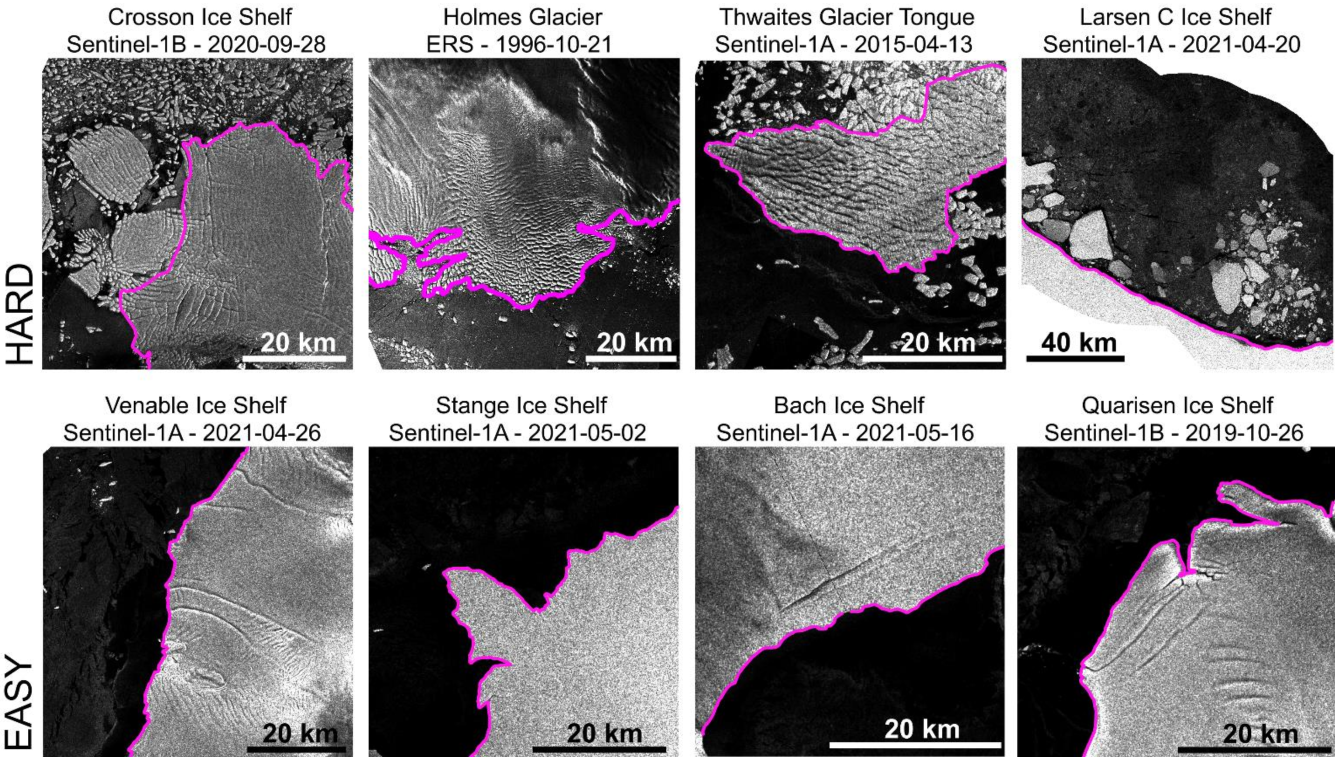


Figure 7 Visual comparison of test scene difficulty, with the top row showing the most difficult samples (from left to right) and the bottom row showing the easiest samples (left to right). The pink line shows the ground truth front position. Copernicus Sentinel-1 Data 2022.

5. Discussion

5.1. Understanding Dataset Characteristics Through Baseline Performance

The baseline models behave differently on Shelf-Bench, indicating that the dataset contains complementary challenges that no single architecture captures equally well. DeepLabV3's lowest mean MDE (514.3 m) and balanced class performance suggest that the dataset spans multiple spatial scales, while U-Net's low median MDE (49.4 m) shows that it also includes sufficiently clear fronts to evaluate pixel-level delineation accuracy. Large spreads in mean MDE across SAR sensors (up to 372.6 m for Envisat) reveal substantial heterogeneity in image characteristics, meaning that Shelf-Bench presents sensor-dependent difficulty. Even median MDE errors remain sizeable (69.6-146.7 m), corresponding to front offsets of roughly 2–5 pixels. ViT shows systematic misclassifications reflected in its high median MDE, particularly an inability to delineate the calving front accurately at the pixel level, which may stem from patch tokenization in the embedding process.

The lower performance of all models based on IoU for the class 'Ice' compared to 'Ocean' suggests Shelf-Bench contains scenes where intensity-based classification strategies are insufficient and spatial morphology needs to be captured by the model for accurate classifications. These failures highlight the dataset's complexity in representing morphologically distinct but spectrally similar ice features (e.g. icebergs and shelf ice), which is a realistic challenge that Shelf-Bench successfully captures and that future users must account for.

DINOv3's average performance on mean and median MDE compared to DeepLabV3 and U-Net despite advanced pre-training reveals important characteristics of Shelf-Bench. The modest performance gains suggest that SAR-based ice shelf delineation represents a sufficiently specialized task that general remote sensing pre-training provides limited benefit. This finding indicates that Shelf-Bench effectively captures domain-specific complexity that cannot be overcome through general-purpose foundation model pre-training which is lacking samples from the cryosphere (Siméoni et al., 2025). This underscores the dataset's value for training task-specific models. The dataset's specialized nature may require researchers to either employ domain-specific pre-training approaches (such as SSL4SAR) (Gourmelon et al., 2025b) or train models from scratch on Shelf-Bench data.

The strong contrast between models, reflected in the large discrepancies between IoU and MDE rankings, indicates that no single architecture can be considered universally best on this benchmark. Methods that perform well in region-based overlap metrics do not necessarily achieve accurate front localization, underscoring that ice front delineation is inherently a multi-objective and highly challenging task. This suggests that future work may benefit from ensemble strategies that combine complementary model strengths, potentially improving both semantic consistency and geometric precision beyond what individual models can achieve.

Overall, the mean MDE of 604.3 m across all baseline models reflects the inherent complexity captured in Shelf-Bench. The strong difference between mean and median MDE highlights the existence of few very difficult patches introducing strong outliers as discussed in the hardness analysis. Nevertheless, this benchmark performance sits within the range reported by Gourmelon et al. (2022) for outlet glaciers, yet Shelf-Bench represents fundamentally different delineation challenges: Antarctic ice shelves are orders of magnitude larger (hundreds of square kilometres) and exhibit morphological complexity distinct from outlet glaciers. The variety of performance outcomes across the five evaluated architectures demonstrates that Shelf-Bench successfully presents sufficiently complex and diverse scenarios to differentiate between model capabilities.

5.2. Sample Hardness in the Shelf-Bench Dataset

The examination of the hardest and easiest scenes in Shelf-Bench reveals the complexity factors that challenge ice shelf front delineation from SAR imagery, independent of the algorithms used. A primary source of complexity arises from limited spatial context, where narrow visible sections of the calving front create visual ambiguity that is difficult to resolve (see Figure 6). Both human interpreters and automated systems struggle to distinguish iceberg edges from true ice shelf boundaries when confronted with such limited contextual information, highlighting a fundamental limitation of patch-based analysis for boundary detection in this application. Increasing the tile size and introducing overlap tiling with centred predictions would provide broader spatial context and help reduce such ambiguities. Additionally, incorporating overlapping tiles during inference could further enhance performance by smoothing predictions in overlapping regions and mitigating local inconsistencies especially at patch edges.

A critical dataset characteristic revealed through the analysis is variability in boundary representation, particularly regarding rift mapping conventions. Different experts naturally interpret and delineate rift structures with varying levels of detail, and this variability is reflected in the reference annotations of Shelf-Bench. These annotation variations represent genuine ambiguity in the delineation task itself and underscore the subjective nature of boundary definition in complex ice shelf environments. Such variability is particularly pronounced in regions with dense or fine-scale rift structures, contributing substantial complexity to these scenes.

Iceberg mélange adjacent to ice shelf fronts represents perhaps the most challenging morphological context captured in Shelf-Bench. Dense concentrations of icebergs positioned in front of ice shelves create scenes with inherent visual complexity, as icebergs and shelf ice exhibit similar scattering behaviour in SAR imagery. The spatial proximity of mélange to the true ice shelf front makes accurate boundary identification subjective and ambiguous, even for experienced human annotators. This is exemplified in the Crosson Ice Shelf scene, where median delineation uncertainty reaches 1925.7 m, reflecting the fundamental difficulty of distinguishing shelf fronts in highly fractured ice shelves with extensive mélange. These scenes represent some of the most realistic and challenging conditions encountered along the Antarctic coastline, highlighting the diversity and representativeness of Shelf-Bench for capturing true operational complexity in ice shelf front mapping.

5.3. Limitations of Shelf-Bench

Even though we tried to provide the best possible benchmark dataset for ice shelf front delineations there are some remaining limitations that warrant discussion. The dataset exhibits uneven spatiotemporal distribution, with denser training coverage in some regions (e.g. Pine Island Bay) (see Figure 2) and uneven distribution of satellite acquisitions over time (see Figure 3). This may introduce geospatial and seasonal biases in models trained on the dataset. Even though, this did not result in lower prediction accuracies for specific regions because they weren't represented within the training set. Three out of the four hardest test scenes are also covered spatially by the training data at another point in time whereas the easiest test samples weren't spatially covered by the training data. Additionally, Shelf-Bench includes multiple scenes of the same ice shelves acquired at different times, introducing temporal redundancy where impact on model training remains unclear. Further analysis would be necessary to assess whether temporal repetition improves model generalization or merely increases computational demands and storage requirements without meaningful performance gains. Moreover, dataset annotation was performed by a limited number of expert interpreters, which introduces subjectivity in coastline delineation. Despite annotation protocols and consultation of auxiliary data sources, reference annotations may contain errors in particularly challenging regions where ice shelf fronts lack clear, unambiguous boundaries. These ambiguities reflect general difficulties in ice shelf front definition rather than annotation failures, and are discussed in Section 3.2.3.

The dataset is formulated as a two-class problem distinguishing 'Ice' from class 'Ocean'. However, no-data regions occur along the edges of the pre-processed satellite scenes, and these areas are currently grouped within the ocean class. This effectively increases the separability of the 'Ocean' class and may partially explain the higher precision and recall observed for that category. To mitigate this effect, we restricted the dataset to patches containing at least 80% valid data. A limited proportion of no-data pixels was intentionally retained to expose the model to realistic scene boundaries; nevertheless, their inclusion may have introduced a slight positive bias in the reported 'Ocean' class performance metrics.

Moreover, the dataset currently incorporates only one polarization per satellite sensor to maintain consistency across the heterogeneous SAR sources, even though Sentinel-1 data in many regions includes dual-polarization (HH and HV) measurements. Dual-polarization SAR data is known to provide improved ice type classification capabilities compared to single-polarization data (Baumhoer et al., 2023), suggesting that future versions incorporating multi-polarization information could enhance the dataset's utility for detailed ice shelf characterization. Finally, while the analysis of scene hardness presented here identifies key morphological sources of delineation complexity, deeper statistical investigation of specific challenge types (e.g., rift patterns, mélange density, surface roughness) should be performed (Seedat et al., 2024). Those patterns combined with targeted synthetic label generation could substantially expand the dataset's coverage of difficult scenarios and provide stronger training signals for underrepresented complexity classes (Hoeser and Kuenzer, 2022). These limitations should be taken into account when using the Shelf-Bench dataset.

6. Conclusions

Shelf-Bench is a comprehensive benchmark dataset comprising 161 manually annotated SAR scenes for Antarctic ice shelf front and coastline delineation, sourced from three satellite sensors (ERS, Envisat, and Sentinel-1). The dataset's foundation in open-access satellite imagery enables broad applicability across glaciological research, computer vision algorithm development, and foundation model training for Earth observation. Shelf-Bench offers key advantages over existing datasets, including vast spatial coverage of the Antarctic coastline with multi-temporal seasonal acquisitions, precisely curated delineation masks paired with pre-processed imagery, and moderate spatial resolution that balances analytical detail with computational accessibility. By providing a standardized, open-access benchmark, Shelf-Bench enables objective model intercomparison, facilitates the development of custom deep learning architectures for ice shelf front detection, and supports continuous monitoring of Antarctic ice sheet dynamics across current and future satellite missions. We provide five robust benchmarks that establish a consistent reference for future comparisons and highlight the intrinsic challenges of the Shelf-Bench dataset. The evaluation of the baseline models revealed strongly contrasting behavior: while ViT achieved the highest pixel accuracy (93.7%) and F1-score (93.1%), it produced the largest mean MDE (692.5 m). Conversely, DeepLabV3 yielded the lowest mean MDE (514.3 m) and U-Net the lowest median MDE (49.4 m) whilst achieving lower pixel accuracies (DeepLabV3: 92.6%; U-Net: 90.6 %) and F1-scores (DeepLabV3: 91.8%; U-Net: 89.8%). These discrepancies show that the dataset simultaneously stresses large-scale classification and precise edge detection, with no model performing equally well on both tasks. The gap between semantic accuracy and geometric precision emphasizes Shelf-Bench's sensitivity to fine-scale boundary structure. Substantial performance variations across sensors and under challenging conditions, as reflected in the sample hardness analysis, confirm that Shelf-Bench preserves realistic heterogeneity in acquisition settings and ice front morphology. This variability is a defining property of the dataset and reflects the complexity of automated ice front delineation.

Consequently, Shelf-Bench serves not only as a benchmark for ranking models, but as a dataset that poses specific challenges and guides future methodological improvements but also to build frameworks for automated calving front monitoring in view of the increasing availability of SAR satellite images. Moving forward, we envision Shelf-Bench supporting several key research directions, including the development of multi-modal models capable of generalizing across SAR platforms, the evaluation of pre-trained foundation models, and the use of ensemble approaches to produce more robust predictions in challenging regions. In addition, incorporating models with calibrated uncertainty can help quantify confidence in areas where predicted front positions are less reliable. Finally, integrating temporal information for improved consistency and change detection, as well as embedding physical constraints informed by glaciological knowledge, could further enhance model performance. By providing open access to both the dataset and baseline model implementations, we aim to lower barriers to entry for future researchers and to foster collaborative progress in this important application domain.

Data availability

The Shelf-Bench dataset is available at https://doi.org/10.5281/zenodo.17610870 (Baumhoer and Morgan, 2025). Shelf-Bench includes 161 satellite scenes from three different SAR sensors. All scenes are stored in the scenes folder, along with matching masks in the masks folder. The 38 test scenes and their corresponding labels are stored in the 'test' folder. To ensure interoperability across the three SAR missions (Sentinel-1, ERS-1/2 and Envisat), a unified naming convention has been applied to all satellite scenes and masks. The raw input filenames varied greatly due to the different naming conventions of each mission. Processing all file names into a uniform structure results in a dataset in which platform, polarisation, and temporal information can be accessed from the file names. Scenes are saved as georeferenced GeoTIFFs in polar stereographic projection (EPSG:3031) with one channel. For Sentinel-1 data with more than one polarisation, only the HH polarisation was included to make the benchmark dataset uniform. To provide a quick overview, we offer the GeoPackage 'shelfbench_extents.gpkg’ that includes the extents of all training and test scenes. All scenes and masks were reformatted according to the following:

[SATELLITE-ID]_[YYYYMMDD]_[POLARISATION]_[SCENE_ID].tif

- SATELLITE-ID: ‘S1A‘, ‘S1B‘, ‘ERS‘, or ‘ENV‘. S1A and SLB refers to Sentinel-1a and Sentinel-1b respectively.
- YYYYMMDD: Acquisition date.
- POLARISATION: e.g., ‘HH‘, ‘VV‘
- SCENE_ID: Product-specific identifiers (e.g., scene code, crop number, or location prefix).

The pre-processed georeferenced satellite scenes can be tiled to custom patch sizes with the script data-preprocessing.py in the Shelf-Bench GitHub repository. Alongside the georeferenced GeoTiff files, we also provide pre-processed 256 x 256 pixel patches stored as PNG images as they were used for the baselines presented here. Patches are named by the scene name and the patch number. Patches were additionally converted to decibel (dB) and locally z-score normalized. The total size of the zipped Shelf-Bench dataset including labels is ~ 12.5 GB for the georeferenced satellite scene version, ~2.9 GB for the tiled PNG patches and ~2.7 GB for the trained weights.

The raw satellite scenes of ERS and Envisat data are freely available via https://eoiam-idp.eo.esa.int/ and can be pre-processed with the ESA SNAP Toolbox available at https://step.esa.int/. Open access Sentinel-1 data is available via the Copernicus Browser https://browser.dataspace.copernicus.eu/.

Code availability

Code produced for this study is available at https://github.com/amymorgan01/Shelf-Bench. It includes code for patching satellite scenes and labels into tiles. The repository also includes model architectures, training and validation set up. For the model implementations we used the PyTorch versions of FPN, U-Net, and DeepLabV3. The ViT-L_16 model was downloaded from https://storage.googleapis.com/vit_models/imagenet21k+imagenet2012/ViT-L_16.npz. The DINOv3 model is specifically the ViT-L/16 distilled 300M SAT-493M model downloaded from https://github.com/facebookresearch/dinov3. The trained weights can be downloaded with the Shelf-Bench dataset at https://doi.org/10.5281/zenodo.17610870.

Author contributions

Conceptualization: C.B.; Data curation; Formal Analysis; Investigation; Validation; Visualization; Writing - original draft: C.B., A.M; Software: A.M., X.H., J.F.; Writing - review & editing: All authors read, reviewed, edited, and approved the final manuscript.

Acknowledgements

The authors acknowledge the European Space Agency and the European Commission for the acquisition and availability of Copernicus Sentinel-1 and Sentinel-2 data as well as Envisat, ERS-1, and ERS-2 data. Furthermore, the authors acknowledge the support of JASMIN, the UK's collaborative data analysis environment, which facilitated model training (https://www.jasmin.ac.uk). The authors used DeepL Write and Claude Sonnet 4.5 to enhance the style of pre-written text in this manuscript and to correct grammatical errors. Claude and Microsoft Copilot were used for improving and cleaning code. These tools were not used to generate content or citations, nor to summarize other existing works. The authors take full responsibility for the final version of the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Financial support

A.B. Morgan's work was conducted as part of the Intelligent Earth CDT supported by funding from the UK Research and Innovation Council (UKRI) grant number EP/Y030907/1. C. Baumhoer received funding from DLR's Polar Monitor Project.

References

Alley, R. B., Clark, P. U., Huybrechts, P., and Joughin, I.: Ice-sheet and sea-level changes, science, 310, 456–460, 2005.

ERS-products-specification-with-Envisat-format: https://earth.esa.int/eogateway/documents/20142/37627/ERS-products-specification-with-Envisat-format.pdf, last access: 25 May 2025.

Baumhoer, C. and Morgan, A.: The Shelf-Bench Dataset: A benchmark dataset for Antarctic ice shelf front and coastline delineation from multi-sensor radar satellite data, https://doi.org/10.5281/zenodo.17610870, 2025.

Baumhoer, C. A., Dietz, A. J., Dech, S., and Kuenzer, C.: Remote Sensing of Antarctic Glacier and Ice-Shelf Front Dynamics—A Review, Remote Sens., 10, 1445, https://doi.org/10.3390/rs10091445, 2018.

Baumhoer, C. A., Dietz, A. J., Kneisel, C., and Kuenzer, C.: Automated Extraction of Antarctic Glacier and Ice Shelf Fronts from Sentinel-1 Imagery Using Deep Learning, Remote Sens., 11, 2529, https://doi.org/10.3390/rs11212529, 2019.

Baumhoer, C. A., Dietz, A. J., Kneisel, C., Paeth, H., and Kuenzer, C.: Environmental drivers of circum-Antarctic glacier and ice shelf front retreat over the last two decades, The Cryosphere, 15, 2357–2381, https://doi.org/10.5194/tc-15-2357-2021, 2021.

Baumhoer, C. A., Dietz, A. J., Heidler, K., and Kuenzer, C.: IceLines – A new data set of Antarctic ice shelf front positions, Sci. Data, 10, 138, https://doi.org/10.1038/s41597-023-02045-x, 2023.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation, https://doi.org/10.48550/arXiv.1706.05587, 5 December 2017.

Cheng, D., Hayes, W., Larour, E., Mohajerani, Y., Wood, M., Velicogna, I., and Rignot, E.: Calving Front Machine (CALFIN): glacial termini dataset and automated deep learning extraction method for Greenland, 1972–2019, The Cryosphere, 15, 1663–1675, https://doi.org/10.5194/tc-15-1663-2021, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, and Li Fei-Fei: ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), Miami, FL, 248–255, https://doi.org/10.1109/CVPR.2009.5206848, 2009.

DEOS: Delft Institute for Earth-Oriented Space Research. Orbit Files, 2016.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, https://doi.org/10.48550/arXiv.2010.11929, 3 June 2020.

Dutton, A., Carlson, A. E., Long, A. J., Milne, G. A., Clark, P. U., DeConto, R., Horton, B. P., Rahmstorf, S., and Raymo, M. E.: Sea-level rise due to polar ice-sheet mass loss during past warm periods, Science, 349, https://doi.org/10.1126/science.aaa4019, 2015.

Ferrigno, J. G., Williams Jr, R. S., Rosanova, C. E., Lucciiitta, B. K., and Swithinbank, C.: Analysis of coastal change in Marie Byrd Land and Ellsworth Land, West Antarctica, using Landsat imagery, Ann. Glaciol., 27, 33–40, 1998.

Fricker, H. A., Young, N. W., Allison, I., and Coleman, R.: Iceberg calving from the Amery Ice Shelf, East Antarctica, Ann. Glaciol., 34, 241-246-241–246, https://doi.org/10.3189/172756402781817581, 2002.

Fürst, J. J., Durand, G., Gillet-Chaulet, F., Tavard, L., Rankl, M., Braun, M., and Gagliardini, O.: The safety band of Antarctic ice shelves, Nat. Clim. Change, 6, 479–482, 2016.

Gerrish, L., Ireland, L., Fretwell, P. T., and Cooper, P.: High resolution vector polylines of the Antarctic coastline - VERSION 7.10 (7.10), https://doi.org/10.5285/567C0911-83A0-493C-9DC7-15245C1C5F5E, 2024.

Gourmelon, N., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: Calving fronts and where to find them: a benchmark dataset and methodology for automatic glacier calving front extraction from synthetic aperture radar imagery, Earth Syst. Sci. Data, 14, 4287–4313, https://doi.org/10.5194/essd-14-4287-2022, 2022.

Gourmelon, N., Klink, J., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: Conditional Random Fields for Improving Deep Learning-Based Glacier Calving Front Delineations, in: IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, 4939–4942, https://doi.org/10.1109/IGARSS52108.2023.10282915, 2023.

Gourmelon, N., Heidler, K., Loebel, E., Cheng, D., Klink, J., Dong, A., Wu, F., Maul, N., Koch, M., Dreier, M., Pyles, D., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: Comparison Study: Glacier Calving Front Delineation in Synthetic Aperture Radar Images With Deep Learning, https://doi.org/10.48550/arXiv.2501.05281, 9 January 2025a.

Gourmelon, N., Dreier, M., Mayr, M., Seehaus, T., Pyles, D., Braun, M., Maier, A., and Christlein, V.: SSL4SAR: Self-Supervised Learning for Glacier Calving Front Extraction From SAR Imagery, IEEE Trans. Geosci. Remote Sens., 63, 1–12, https://doi.org/10.1109/TGRS.2025.3580945, 2025b.

Greene, C. A., Gardner, A. S., Schlegel, N.-J., and Fraser, A. D.: Antarctic calving loss rivals ice-shelf thinning, Nature, 2022.

Hartmann, A., Davari, A., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: Bayesian U-Net for Segmenting Glaciers in Sar Imagery, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 3479–3482, https://doi.org/10.1109/IGARSS47720.2021.9554292, 2021.

Heidler, K., Mou, L., Baumhoer, C., Dietz, A., and Zhu, X. X.: HED-UNet: Combined Segmentation and Edge Detection for Monitoring the Antarctic Coastline, IEEE Trans. Geosci. Remote Sens., 60, 1–14, https://doi.org/10.1109/TGRS.2021.3064606, 2021.

Heidler, K., Mou, L., Loebel, E., Scheinert, M., Lefèvre, S., and Zhu, X. X.: A Deep Active Contour Model for Delineating Glacier Calving Fronts, IEEE Trans. Geosci. Remote Sens., 61, 1–12, https://doi.org/10.1109/TGRS.2023.3296539, 2023.

Herrmann, O., Gourmelon, N., Seehaus, T., Maier, A., Fürst, J. J., Braun, M. H., and Christlein, V.: Out-of-the-box calving-front detection method using deep learning, The Cryosphere, 17, 4957–4977, https://doi.org/10.5194/tc-17-4957-2023, 2023.

Hoeser, T. and Kuenzer, C.: SyntEO: Synthetic dataset generation for earth observation and deep learning – Demonstrated for offshore wind farm detection, ISPRS J. Photogramm. Remote Sens., 189, 163–184, https://doi.org/10.1016/j.isprsjprs.2022.04.029, 2022.

Hoeser, T., Bachofer, F., and Kuenzer, C.: Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review—Part II: Applications, Remote Sens., 12, 3053, https://doi.org/10.3390/rs12183053, 2020.

Holzmann, M., Davari, A., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: Glacier Calving Front Segmentation Using Attention U-Net, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium, Brussels, Belgium, 3483–3486, 2021.

Jensen, W.: ERS-1/2 and its data used in operational systems, in: 1995 International Geoscience and Remote Sensing Symposium, IGARSS '95. Quantitative Remote Sensing for Science and Applications, 1995 International Geoscience and Remote Sensing Symposium, IGARSS '95. Quantitative Remote Sensing for Science and Applications, 1044–1046 vol.2, https://doi.org/10.1109/IGARSS.1995.521133, 1995.

Jiang, D., Li, S., Hajnsek, I., Siddique, M. A., Hong, W., and Wu, Y.: Glacial lake mapping using remote sensing Geo-Foundation Model, Int. J. Appl. Earth Obs. Geoinformation, 136, 104371, https://doi.org/10.1016/j.jag.2025.104371, 2025.

Kaushik, S., Maurya, L., Tellman, E., Zhang, G., and Dharpure, J. K.: Debris covered glacier mapping using newly annotated multisource remote sensing data and geo-foundational model, Sci. Remote Sens., 12, 100319, https://doi.org/10.1016/j.srs.2025.100319, 2025.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, CoRR, 2014.

van 't Klooster, K.: ERS-1, European remote-sensing satellite was launched 20 years ago, in: 2011 21st International Crimean Conference "Microwave & Telecommunication Technology," 2011 21st International Crimean Conference "Microwave & Telecommunication Technology," 117–118, 2011.

Laur, H., Bally, P., Meadows, P., Sanchez, J., Schaettler, B., Lopinto, E., and Esteban, D.: Derivation of the backscattering coefficient σ0 in ESA ERS SAR PRI products, in: Proc. of the Second International Workshop on ERS Applications, 139, 2002.

Lea, J. M.: The Google Earth Engine Digitisation Tool (GEEDiT) and the Margin change Quantification Tool (MaQiT); simple tools for the rapid mapping and quantification of changing Earth surface margins, Earth Surf. Dyn., 6, 551–561, https://doi.org/10.5194/esurf-6-551-2018, 2018.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S.: Feature Pyramid Networks for Object Detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 936–944, https://doi.org/10.1109/CVPR.2017.106, 2017.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P.: Focal Loss for Dense Object Detection, https://doi.org/10.48550/arXiv.1708.02002, 7 February 2018.

Liu, H. and Jezek, K. C.: A complete high-resolution coastline of Antarctica extracted from orthorectified Radarsat SAR imagery, Photogramm. Eng. Remote Sens., 70, 605–616, https://doi.org/10.14358/pers.70.5.605, 2004a.

Liu, H. and Jezek, K. C.: Automated extraction of coastline from satellite imagery by integrating Canny edge detection and locally adaptive thresholding methods, Int. J. Remote Sens., 25, 937–958, https://doi.org/10.1080/0143116031000139890, 2004b.

Loebel, E., Scheinert, M., Horwath, M., Heidler, K., Christmann, J., Phan, L. D., Humbert, A., and Zhu, X. X.: Extracting glacier calving fronts by deep learning: the benefit of multi-spectral, topographic and textural input features, IEEE Trans. Geosci. Remote Sens., 1–1, https://doi.org/10.1109/TGRS.2022.3208454, 2022.

Loebel, E., Baumhoer, C. A., Dietz, A. J., Scheinert, M., and Horwath, M.: Glacier calving front locations for the Antarctic Peninsula Ice Sheet derived from remote sensing and deep learning from 2013 to 2023, https://doi.org/10.1594/PANGAEA.963725, 2024.

Loebel, E., Baumhoer, C. A., Dietz, A., Scheinert, M., and Horwath, M.: Calving front positions for 42 key glaciers of the Antarctic Peninsula Ice Sheet: a sub-seasonal record from 2013 to 2023 based on deep-learning application to Landsat multi-spectral imagery, Earth Syst. Sci. Data, 17, 65–78, https://doi.org/10.5194/essd-17-65-2025, 2025.

Lu, X., Jiang, L., Li, D., Liu, Y., Sole, A., and Livingstone, S. J.: Calving front positions for Greenland outlet glaciers (2002&ndash;2021): a spatially extensive seasonal record and benchmark dataset for algorithm validation, Earth Syst. Sci. Data Discuss., 1–25, https://doi.org/10.5194/essd-2025-304, 2025.

Marochov, M., Stokes, C. R., and Carbonneau, P. E.: Image classification of marine-terminating outlet glaciers in Greenland using deep learning methods, The Cryosphere, 15, 5041–5059, https://doi.org/10.5194/tc-15-5041-2021, 2021.

Mohajerani, Y., Wood, M., Velicogna, I., and Rignot, E.: Detection of Glacier Calving Margins with Convolutional Neural Networks: A Case Study, Remote Sens., 11, 1–13, https://doi.org/10.3390/rs11010074, 2019.

Nagler, T., Rott, H., Hetzenecker, M., Wuite, J., and Potin, P.: The Sentinel-1 Mission: New Opportunities for Ice Sheet Observations, Remote Sens., 7, 9371–9389, https://doi.org/10.3390/rs70709371, 2015.

Periyasamy, M., Davari, A., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: How to Get the Most Out of U-Net for Glacier Calving Front Segmentation, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 15, 1712–1723, https://doi.org/10.1109/JSTARS.2022.3148033, 2022.

Pritchard, H. D., Fretwell, P. T., Fremand, A. C., Bodart, J. A., Kirkham, J. D., Aitken, A., Bamber, J., Bell, R., Bianchi, C., Bingham, R. G., Blankenship, D. D., Casassa, G., Christianson, K., Conway, H., Corr, H. F. J., Cui, X., Damaske, D., Damm, V., Dorschel, B., Drews, R., Eagles, G., Eisen, O., Eisermann, H., Ferraccioli, F., Field, E., Forsberg, R., Franke, S., Goel, V., Gogineni, S. P., Greenbaum, J., Hills, B., Hindmarsh, R. C. A., Hoffman, A. O., Holschuh, N., Holt, J. W., Humbert, A., Jacobel, R. W., Jansen, D., Jenkins, A., Jokat, W., Jong, L., Jordan, T. A., King, E. C., Kohler, J., Krabill, W., Maton, J., Gillespie, M. K., Langley, K., Lee, J., Leitchenkov, G., Leuschen, C., Luyendyk, B., MacGregor, J. A., MacKie, E., Moholdt, G., Matsuoka, K., Morlighem, M., Mouginot, J., Nitsche, F. O., Nost, O. A., Paden, J., Pattyn, F., Popov, S., Rignot, E., Rippin, D. M., Rivera, A., Roberts, J. L., Ross, N., Ruppel, A., Schroeder, D. M., Siegert, M. J., Smith, A. M., Steinhage, D., Studinger, M., Sun, B., Tabacco, I., Tinto, K. J., Urbini, S., Vaughan, D. G., Wilson, D. S., Young, D. A., and Zirizzotti, A.: Bedmap3 updated ice bed, surface and thickness gridded datasets for Antarctica, Sci. Data, 12, 414, https://doi.org/10.1038/s41597-025-04672-y, 2025.

Rignot, E., Jacobs, S., Mouginot, J., and Scheuchl, B.: Ice-Shelf Melting Around Antarctica, Science, 341, 266–270, https://doi.org/10.1126/science.1235798, 2013.

Rignot, E. J. M. and van Zyl, J. J.: Change detection techniques for ERS-1 SAR data, IEEE Trans. Geosci. Remote Sens., 31, 896–906, https://doi.org/10.1109/36.239913, 1993.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, vol. 9351, edited by: Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., Springer International Publishing, Cham, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

Rosich, B.: Absolute calibration of ASAR level 1 products generated with PF-ASAR, Ftpftp Esrin Esa ItpubESA-DOCENVISATASARASAR-Prod.-Absol.-Calibration-V1 4 Pdf, 2004.

Scambos, T. A., Haran, T. M., Fahnestock, M. A., Painter, T. H., and Bohlander, J.: MODIS-based Mosaic of Antarctica (MOA) data sets: Continent-wide surface morphology and snow grain size, Remote Sens. Environ., 111, 242–257, https://doi.org/10.1016/j.rse.2006.12.020, 2007.

Seedat, N., Imrie, F., and Schaar, M. van der: Dissecting Sample Hardness: A Fine-Grained Analysis of Hardness Characterization Methods for Data-Centric AI, https://doi.org/10.48550/arXiv.2403.04551, 7 March 2024.

Shankar, S., Stearns, L. A., and Veen, C. J. van der: Semantic segmentation of glaciological features across multiple remote sensing platforms with the Segment Anything Model (SAM), J. Glaciol., 1–10, https://doi.org/10.1017/jog.2023.95, 2023.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., and Bojanowski, P.: DINOv3, https://doi.org/10.48550/arXiv.2508.10104, 13 August 2025.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M.: Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Cham, 240–248, https://doi.org/10.1007/978-3-319-67558-9_28, 2017.

Sun, Y. and Li, X.-M.: Denoising Sentinel-1 Extra-Wide Mode Cross-Polarization Images Over Sea Ice, IEEE Trans. Geosci. Remote Sens., 59, 2116–2131, https://doi.org/10.1109/TGRS.2020.3005831, 2021.

Wagner, L.: Analysis of ice shelf front dynamics in Pine Island Bay (Antarctica) based on long-term SAR time series and deep learning, masters, Universität Würzburg, 103 pp., 2023.

Wessel, B., Huber, M., Wohlfart, C., Bertram, A., Osterkamp, N., Marschalk, U., Gruber, A., Reuß, F., Abdullahi, S., Georg, I., and Roth, A.: TanDEM-X PolarDEM 90 of Antarctica: generation and error characterization, The Cryosphere, 15, 5241–5260, https://doi.org/10.5194/tc-15-5241-2021, 2021.

Williams, R. S., Ferrigno, J. G., Swithinbank, C., Lucchitta, B. K., and Seekins, B. A.: Coastal-change and glaciological maps of Antarctica, Ann. Glaciol., 21, 284–290, 1995.

Wu, F., Gourmelon, N., Seehaus, T., Zhang, J., Braun, M., Maier, A., and Christlein, V.: AMD-HookNet for Glacier Front Segmentation, IEEE Trans. Geosci. Remote Sens., 61, 1–12, https://doi.org/10.1109/TGRS.2023.3245419, 2023.

Wu, F., Gourmelon, N., Seehaus, T., Zhang, J., Braun, M., Maier, A., and Christlein, V.: Contextual HookFormer for Glacier Calving Front Segmentation, IEEE Trans. Geosci. Remote Sens., 62, 1–15, https://doi.org/10.1109/TGRS.2024.3368215, 2024.

Zhang, E., Liu, L., and Huang, L.: Automatically delineating the calving front of Jakobshavn Isbræ from multitemporal TerraSAR-X images: a deep learning approach, The Cryosphere, 13, 1729–1741, https://doi.org/10.5194/tc-13-1729-2019, 2019.

Zhang, E., Liu, L., Huang, L., and Ng, K. S.: An automated, generalized, deep-learning-based method for delineating the calving fronts of Greenland glaciers from multi-sensor remote sensing imagery, Remote Sens. Environ., 254, 112265, https://doi.org/10.1016/j.rse.2020.112265, 2021.

Zhao, J., Tong, J., Li, T., Sun, Y., Shao, C., and Dong, Y.: CISNet: Change information guided semantic segmentation network for automatic extraction of glacier calving fronts, ISPRS J. Photogramm. Remote Sens., 228, 666–678, https://doi.org/10.1016/j.isprsjprs.2025.08.001, 2025.

https://doi.org/10.5194/essd-2025-758
Preprint. Discussion started: 2 March 2026

Open Access Earth System Science Data Discussions

Appendix A

Table A 1 Distribution of training and test data per coastal section. Note that larger ice shelves (e.g. Ross, Ronne-Filchner) and scene extents may be located between coastal sections. For simplicity, they were added to only one section.

Main Region	Coastal Region	Major Ice Shelves	Train			Test		
			ERS	Envisat	S-1	ERS	Envisat	S-1
AP	Graham Land	Larsen A-C		2	4		1	1
AP	Palmer Land	George VI, Wilkins, Bach, Stange, Larsen D	4		6	1		3
WAIS	Ellsworth Land	Abbot, Cosgrove, Pine Island, Thwaites, Crosson, Dotson, Getz,	23	4	4	1	1	7
WAIS	Marie Byrd Land	Nickerson, Sulzberger			20			
EAIS	Victoria Land	Ross, Drygalski Ice Tongue			9	1		1
EAIS	Oats Land	Rennick Glacier						
EAIS	George V Land	Ninnis, Mertz, Cook	1	2				1
EAIS	Terre Adélie	-			4			
EAIS	Wilkes Land	Totten, Moskow			9	1	1	1
EAIS	Queen Mary Land	Conger, Schackleton			10	2		1
EAIS	Princess Elizabeth Land	West Ice Shelf	1		4			1
EAIS	Mac.Robertson Land	Amery, Publications		2	2			1
EAIS	Kemp Land	Edward VIII						
EAIS	Enderby Land	Shirase	1		4			
EAIS	Dronning Maud Land	Baudouin, Riiser-Larsen, Nivlisen, Quarisen, Ekstromisen				1		7
EAIS	Coats Land	Brunt, Ronne-Filchner	2		3	1		3
TOTAL	-	-	32	10	79	8	3	

Table A 2 MDE values with confidence intervals (CI) for the mean and median for each model and satellite sensor as addition to Table 6. Confidence intervals are calculated using bootstrap 95% percentile confidence intervals (with 1000 bootstrap draws) on the mean and median MDE, respectively.

		mean			median		
model	**satellite**	**mean**	**lower CI**	**Upper CI**	**median**	**lower CI**	**upper IC**
DeepLabV3	Envisat	365.2	206.53	551.82	91.87	71.18	160.42
	ERS	627.91	464.86	800.05	101.49	82.17	160.97
	Sentinel-1	488.96	381.04	606.4	76.77	65.14	88.31
DinoV3	Envisat	595.39	359.78	891.13	99.5	66.65	183.21
	ERS	814.84	621.41	1004.86	93.79	61.48	170.08
	Sentinel-1	592.06	477.86	716.33	53.44	47.38	62.14
FPN	Envisat	510.2	285.2	785.56	124.53	64.12	188.83
	ERS	442.26	324.69	577.87	115.49	64.47	173.69
	Sentinel-1	687.02	537.71	847.71	101.39	84.2	136.51
U-Net	Envisat	520.5	263.93	848.49	63.14	29.99	115.94
	ERS	548.06	393.33	716.39	46.22	29.55	89.59
	Sentinel-1	569.26	444.83	698.95	49.41	42.12	63.52
ViT	Envisat	591.97	370.4	848.52	209.84	98.44	423.59
	ERS	734.08	577.64	905.29	136.76	110.75	213.77
	Sentinel-1	689.87	556.59	832.3	118.96	111.01	143.19

Table A 3 MDE values for the most difficult patches in the test set. Visual examples are given in Figure 6.

Filename	Rank	Median distance	DeeplabV3	DINOv3	FPN	UNET	ViT
S1A_20210420_HH_EW_GRDM_A997__33_241_417_4096_4352.png	1	9395.2			10645.7	7199.3	9395.2
S1A_20150413_HH_EW_GRDM_814A__129_178_1_0_256.png	2	9164.2					9164.2
S1A_20150413_HH_EW_GRDM_814A__129_178_2_0_512.png	3	8083.9					8083.9
S1B_20161025_HH_EW_GRDM_E643__25_170_158_1280_768.png	4	7922.3	6854.9	8867.1	8300.1	7544.6	
S1A_20150131_HH_EW_GRDM_99B3__39_33_165_2048_1280.png	5	7363.1	7465.1		7376.9	7349.2	110.2
S1A_20200907_HH_EW_GRDM_7A4D__103_113_293_4608_1280.png	6	7019.8			7019.8		
S1B_20190529_HH_EW_GRDM_1EE2__61_249_121_1536_3328.png	7	6320.5				6320.5	
S1B_20210807_HH_EW_GRDM_3E5B__30_158_69_1280_1024.png	8	5934.9	82.3	7924.2	7049.7		4820.0
S1A_20200907_HH_EW_GRDM_7A4D__103_113_221_3328_3328.png	9	5931.2	5953.2	5964.8	5931.2	5910.7	5773.4
S1A_20191115_HH_EW_GRDM_8217__106_207_33_1024_256.png	10	5774.3	10573.3	5902.2	119.5	3639.0	5774.3
ERS_19960415_VV_202845__74_113_51_1024_1792.png	11	5710.6	5762.9		28.7	5710.6	
S1A_20150131_HH_EW_GRDM_99B3__39_33_195_2304_3840.png	12	5690.1	5745.4	5623.7	5705.2	5690.1	5526.1
S1B_20200928_HH_EW_GRDM_34F1__159_76_89_1536_1280.png	13	5679.8		5679.8	6266.8	2494.6	
ERS_19961115_VV_085115__127_180_332_4352_2304.png	14	5631.2		6706.0			4556.5

ERS_19960322_VV_094606__155_10_108_1792_768.png	15	5619.3					5619.3
S1B_20200928_HH_EW_GRDM_34F1__159_76_51_768_2304.png	16	5451.2	5195.8	5538.8	6440.6	5451.2	4611.9